LERNEN LEICHTER GEMACHT

2. Auflage

Statistik mit Excel

für dummies®

Grundlagen der Statistik verstehen

Mit Excel Daten analysieren

Statistiken richtig interpretieren und Hypothesen testen

Joseph Schmuller

Schummelseite

Excel stellt eine Vielzahl statistischer Funktionen bereit, mit denen Sie in Ihren Arbeitsblättern einen einzelnen Wert oder ein Array von Werten berechnen können. Bei den Excel-Analyse-Funktionen handelt es sich um ein Add-In, das Ihnen weitere Werkzeuge für statistische Analysen zur Verfügung stellt. Schauen Sie sich diese Werkzeuge an, um das Beste aus Ihren statistischen Analysen herauszuholen.

TABELLENBLATTFUNKTIONEN FÜR STATISTISCHE ANALYSEN

Mit Excel können Sie alle möglichen Berechnungen durchführen. Hier finden Sie eine Auswahl an statistischen Tabellenfunktionen von Excel. Jede dieser Funktionen zeigt als Ergebnis einen Wert in einer markierten Zelle an.

Die folgenden Funktionen kümmern sich um die zentrale Tendenz und die Streuung.

Funktion	Was die Funktion berechnet
MITTELWERT	Mittelwert einer Gruppe von Zahlen
MITTELWERTWENN	Mittelwert einer Gruppe von Zahlen, die eine bestimmte Bedingung erfüllen
MITTELWERTWENNS	Mittelwert einer Gruppe von Zahlen, die mehrere Bedingungen erfüllen
GEOMITTEL	Geometrischer Mittelwert einer Gruppe positiver Zahlen
HARMITTEL	Harmonischer Mittelwert einer Gruppe positiver Zahlen
MODUS.EINF	Häufigster Wert (Modalwert) einer Gruppe von Zahlen
MEDIAN	Median einer Gruppe von Zahlen
VAR.P	Varianz einer Gruppe von Zahlen, die eine Grundgesamtheit darstellt
VAR.S	Varianz einer Gruppe von Zahlen, die eine Stichprobe darstellt
STABW.N	Standardabweichung einer Gruppe von Zahlen, die eine Grundgesamtheit darstellt
STABW.S	Standardabweichung einer Gruppe von Zahlen, die eine Stichprobe darstellt
STANDARDISIERUNG	Standardwert, der auf einem gegebenen Mittelwert und einer gegebenen Standardabweichung beruht

Die folgenden Funktionen für die Ermittlung der relativen Position sind ebenfalls sehr nützlich.

Funktion	Was die Funktion berechnet
RANG.GLEICH	Rang einer Zahl in einer Zahlengruppe. Wenn mehr als eine Zahl den gleichen Rang besitzt, gibt die Funktion den obersten Rang der Zahl zurück.
RANG.MITTELW	Rang einer Zahl in einer Zahlengruppe. Wenn mehr als eine Zahl den gleichen Rang besitzt, wird die durchschnittliche Rangzahl zurückgegeben.
QUANTILSRANG.INKL	Gibt den prozentualen Rang eines Wertes in einer Zahlengruppe zurück. Diese Funktion berücksichtigt die beiden Ergebnisse 0 und 1.

Schummelseite

Funktion	Was die Funktion berechnet
QUANTILSRANG.EXKL	Gibt den prozentualen Rang eines Wertes in einer Zahlengruppe zurück. Diese Funktion berücksichtigt die beiden Ergebnisse 0 und 1 nicht.
QUANTIL.INKL	Das angegebene Quantil in einer Gruppe von Zahlen unter Verwendung von »größer als oder gleich«.
QUANTIL.EXKL	Das angegebene Quantil in einer Gruppe von Zahlen unter Verwendung von »größer als«.
QUARTILE.INKL	Gibt den Minimalwert, das untere Quartil (25 %-Quantil), den Median (50 %-Quantil), das obere Quartil (75 %-Quantil) oder den Maximalwert einer Zahlengruppe unter Verwendung von »größer als oder gleich« zurück.
QUARTILE.EXKL	Gibt das untere Quartil (25 %-Quantil), den Median (50 %-Quantil) oder das obere Quartil (75 %-Quantil) einer Zahlengruppe unter Verwendung von »größer als« zurück.

Es ist nicht verkehrt, auch die Funktionen zu kennen, die Korrelation und Regression berechnen.

Funktion	Was die Funktion berechnet
KORREL	Korrelationskoeffizienten zwischen zwei Zahlengruppen
PEARSON	Dasselbe wie mit KORREL (probieren Sie es aus!)
BESTIMMTHEITSMASS	Bestimmtheitsmaß zwischen zwei Zahlengruppen (Quadrat des Korrelationskoeffizienten)
STEIGUNG	Steigung einer Regressionsgeraden durch zwei Zahlengruppen
ACHSENABSCHNITT	Achsenabschnitt einer Regressionsgeraden durch zwei Zahlengruppen
STFEHLERYX	Standardfehler der Schätzung für eine Regressionsgerade durch zwei Zahlengruppen

ARRAYFUNKTIONEN FÜR STATISTISCHE ANALYSEN

Eine Arrayfunktion berechnet eine Gruppe von Werten und nicht nur einen Wert. Hier sind die statistischen Arrayfunktionen, die Excel zur Verfügung stellt. Jede Funktion fügt die berechneten Werte in einen markierten Zellbereich ein.

Drücken Sie ⇧ + Strg + ↵, um eine Arrayfunktion einzugeben.

Funktion	Was die Funktion berechnet
HÄUFIGKEIT	Häufigkeiten von Werten in einer Wertegruppe
MODUS.VIELF	Modalwert einer Zahlengruppe
RGP	Regressionsstatistiken, die auf der linearen Regression durch zwei oder mehr Zahlengruppen basieren

Schummelseite

Funktion	Was die Funktion berechnet
RKP	Regressionsstatistiken, die auf der nichtlinearen Regression durch zwei oder mehr Zahlengruppen basieren
TREND	Zahlen in einem linearen Trend, basierend auf bekannten Datenpunkten
VARIATION	Zahlen in einem nichtlinearen Trend, basierend auf bekannten Datenpunkten

DATENANALYSETOOLS VON EXCEL

Hierbei handelt es sich um die Analyse-Funktionen von Excel. Diese werden als Add-Ins geladen.

Analyse-Funktion	Das können Sie mit dieser Analyse-Funktion tun
Anova: Einfaktoriell	Varianzanalyse für mindestens zwei Stichproben
Anova: Zweifaktoriell mit Messwiederholung	Varianzanalyse mit zwei unabhängigen Variablen und mehreren Beobachtungen in jeder Kombination der Variablenausprägungen
Anova: Zweifaktoriell ohne Messwiederholung	Varianzanalyse mit zwei unabhängigen Variablen und einer Beobachtung in jeder Kombination der Variablenausprägungen
Korrelation	Berechnet aus den Messungen einer Stichprobe von mindestens drei Objekten eine Matrix der Korrelationskoeffizienten für alle möglichen Messpaare
Kovarianz	Berechnet aus den Messungen einer Stichprobe von mindestens drei Objekten eine Matrix der Kovarianz für alle möglichen Messpaare
Populationskenngrößen	Erstellt einen Bericht über zentrale Tendenz, Streuung der Daten und andere Eigenschaften der Werte im markierten Zellbereich
Exponentielles Glätten	Berechnet in einer Wertefolge eine Vorhersage basierend auf vorherigen Werten und auf einer Vorhersage für diese Werte
Zwei-Stichproben *F*-Test	Führt einen *F*-Test zum Vergleich zweier Varianzen durch
Histogramm	Berechnet individuelle und kumulierte Häufigkeiten für Werte im markierten Zellbereich
Gleitender Durchschnitt	Berechnet in einer Wertefolge eine Vorhersage, die den Mittelwert einer bestimmten Anzahl vorhergehender Werte darstellt

Schummelseite

Analyse-Funktion	Das können Sie mit dieser Analyse-Funktion tun
Zufallszahlengenerierung	Stellt eine bestimmte Menge an Zufallszahlen bereit, die aus einer von sieben möglichen Verteilungen generiert werden
Rang und Quantil	Erstellt eine Tabelle, die den Rang sowie den Quantilsrang jedes Werts in einem Datensatz enthält
Regression	Erstellt einen Bericht über die Regressionsstatistik basierend auf einer linearen Regression über einen Datensatz, der eine abhängige und mindestens eine unabhängige Variable enthält
Stichprobenziehung	Erstellt eine Stichprobe aus den Werten in einem festgelegten Zellbereich

Joseph Schmuller

Statistik mit Excel

für dummies®

Übersetzung aus dem Amerikanischen
von Rainer G. Haselier, Elke Jauch
und Calinka Gonschorek

WILEY

WILEY-VCH Verlag GmbH & Co. KGaA

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2016

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Original English language edition Statistical Analysis with Excel For Dummies 4th Edition © 2016 by Wiley Publishing, Inc.

All rights reserved including the right of reproduction in whole or in part in any form. This translation published by arrangement with John Wiley and Sons, Inc.

Copyright der englischsprachigen Originalausgabe Statistical Analysis with Excel For Dummies 4th Edition © 2016 by Wiley Publishing, Inc. [or the name of the copyright holder as it appears in the original Eng lang edition.]

Alle Rechte vorbehalten inklusive des Rechtes auf Reproduktion im Ganzen oder in Teilen und in jeglicher Form. Diese Übersetzung wird mit Genehmigung von John Wiley and Sons, Inc. publiziert.

Wiley, the Wiley logo, Für Dummies, the Dummies Man logo, and related trademarks and trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries. Used by permission.

Wiley, die Bezeichnung »Für Dummies«, das Dummies-Mann-Logo und darauf bezogene Gestaltungen sind Marken oder eingetragene Marken von John Wiley & Sons, Inc., USA, Deutschland und in anderen Ländern.

Das vorliegende Werk wurde sorgfältig erarbeitet. Dennoch übernehmen Autoren und Verlag für die Richtigkeit von Angaben, Hinweisen und Ratschlägen sowie eventuelle Druckfehler keine Haftung.

Coverfoto: alexaldo-iStock

Korrektur: Isolde Kommer

Satz: Beltz Bad Langensalza GmbH, Bad Langensalza
Print ISBN: 978-3-527-71287-8
ePub ISBN: 978-3-81170-0
mobi ISBN: 978-3-81169-4

Kapitel 1

Über den Autor

Joseph Schmuller hat langjährige Erfahrung im Unterrichten von Statistik an der Universität und ist Mitglied der American Statistical Association. Zurzeit arbeitet er als Professor an der University of North Florida. Er hat auch mehrere Bücher zu Informatik geschrieben, unter anderem »Teach Yourself UML in 24 Hours« und natürlich auch diesen Titel in den letzten drei Auflagen. Er erstellte Online-Kurse für das Bildungsunternehmen Lynda.com, das inzwischen zu Microsoft gehört. Zudem war er sechs Jahre lang Chefredakteur des PC AI Magazins.

Joseph Schmuller lebt mit seiner Familie in Jacksonville, Florida.

Inhaltsverzeichnis

- Cover
- Schummelseite
- Titel
- Impressum
- Über den Autor
- Einführung
- Teil I: Statistik und Excel: Wie füreinander geschaffen
  - Kapitel 1: Echte Daten auswerten
    - Die statistischen (und damit verwandten) Begriffe, die Sie einfach kennen müssen
    - Inferenzstatistik: Testen von Hypothesen
    - Was ist neu in Excel 2016?
    - Was ist alt in Excel 2016?
    - Die Grundlagen kennen
  - Kapitel 2: Grundlegendes zu den statistischen Funktionen von Excel
    - Erste Schritte
    - Machen Sie sich bereit für die Statistik
    - Auf oft verwendete Funktionen zugreifen
- Teil II: Daten beschreiben
  - Kapitel 3: Präsentieren: Daten in Diagrammen darstellen
    - Warum Diagramme verwenden?
    - Einige grundlegende Dinge
    - Die Diagramm-Features von Excel
    - So werden Sie ein Säulenheiliger
    - Segmente eines Kreisdiagramms
    - Eine Linie ziehen
    - Minidiagramme mit Sparklines

Balken über Balken
Punkt, Punkt, Komma, Strich …
Ein weiterer Einsatzzweck für das Punktdiagramm
Die Blasen schmecken
Alles auf Kurs
An der Oberfläche kratzen
Im Netz gefangen
Eine Treemap blühen und die Sonne hervorbrechen lassen
Ein Histogramm erstellen
Säulen sortieren: Pareto
Kastengrafik: Kästchen und Antennen
3D-Karten
Kapitel 4: Suchen Sie Ihre Mitte
Mittelwert: Die Lehre vom Durchschnitt
Mediane: Auf halber Strecke erwischt
Der Modalwert
Kapitel 5: Abweichungen vom Durchschnitt
Die Streuung berechnen
Zurück zu den Wurzeln: Standardabweichung
Verwandte Funktionen
Kapitel 6: Standards und Wertungen kennenlernen
Z-Werte einfangen
Wo stehen Sie?
Kapitel 7: Alles zusammenfassen
Auszählen
Groß und klein
Jetzt wird es esoterisch
Nun kommt die Häufigkeit ins Spiel
Können Sie mir eine Beschreibung geben?
Schnelle Analysen
Schnelle Kenngrößen
Kapitel 8: Was ist normal?

So kratzen Sie die Kurve
Eine ganz besondere Verteilung
Eine Standardnormalverteilung als Diagramm darstellen
Teil III: Rückschlüsse aus Daten ziehen
Kapitel 9: Die Sache mit dem Vertrauen: Schätzung
Stichprobenverteilungen verstehen
Ein BESONDERS wichtiges Konzept: der zentrale Grenzwertsatz
Die Vertrauensgrenzen
Passend für ein t
Kapitel 10: Ein-Stichproben-Hypothesentest
Hypothesen, Tests und Fehler
Noch einmal z-Werte
t-Test für eine Stichprobe
Eine t-Verteilung visualisieren
Testen einer Varianz
Eine Chi-Quadrat-Verteilung visualisieren
Kapitel 11: Zwei-Stichproben-Hypothesentest
Hypothesen für zwei
Noch einmal Stichprobenverteilungen
t-Test für zwei Stichproben
Ein passendes Paar: Hypothesentest für abhängige Stichproben
Zwei Varianzen testen
Die F-Verteilung visualisieren
Kapitel 12: Mehr als zwei Stichproben testen
Mehr als zwei Stichproben testen
Eine andere Art Hypothese, eine andere Art Test
Kapitel 13: Etwas komplexere Tests

Die Kombinationen knacken
Noch einmal Kombinationen knacken
Zwei Arten von Variablen, und zwar gleichzeitig
Excel mit einem gemischten Design verwenden
Die Ergebnisse als Diagramm darstellen
Nach der ANOVA

Kapitel 14: Lineare und multiple Regression

Das Streudiagramm
Geraden zeichnen
Regression: Was für eine Gerade!
Tabellenblattfunktionen für die Regression
Datenanalysetool: Regression
Irrsinnig viele Zusammenhänge auf einmal: multiple Regression
Excel-Tools für multiple Regression

Kapitel 15: Korrelation: Aufstieg und Fall von Zusammenhängen

Noch einmal Streudiagramme
Korrelation und Regression
Hypothesen über Korrelationen testen
Tabellenblattfunktionen für die Korrelation
Datenanalysetool: Korrelation
Datenanalysetool: Kovarianz
Hypothesen über Korrelationen testen

Kapitel 16: Es wird höchste Zeit

Eine Zeitreihe und ihre Komponenten
Sanftes Gleiten
Exponentielles Glätten
Prognosen mit einem Klick

Kapitel 17: Parameterfreie Statistik

Unabhängige Stichproben
Parallelstichproben
Spearmans Rangkorrelationskoeffizient
Ein paar Hinweise zum Schluss

Teil IV: Wahrscheinlichkeit

Kapitel 18: Einführung in die Wahrscheinlichkeit

- Was ist Wahrscheinlichkeit?
- Zusammengesetzte Ereignisse
- Bedingte Wahrscheinlichkeit
- Große Wahrscheinlichkeitsräume
- Tabellenblattfunktionen
- Zufallsvariablen: diskret und stetig
- Wahrscheinlichkeitsverteilungen und Dichtefunktionen
- Die Binomialverteilung
- Tabellenblattfunktionen
- Hypothesen mit der Binomialverteilung testen
- Weitere Informationen zum Testen von Hypothesen
- Die hypergeometrische Verteilung

Kapitel 19: Mehr zur Wahrscheinlichkeit

- Betaverteilung
- BETA.INV
- Poissonverteilung
- Gammaverteilung
- Exponentialverteilung

Kapitel 20: Modelle

- Die Modellierung einer Verteilung
- Prüfen, ob das Modell passt
- Simulationen
- Den Würfel »zinken«

Teil V: Der Top-Ten-Teil

Kapitel 21: Zehn Tipps und Fallstricke bei der Arbeit mit Statistiken und Grafiken

- Signifikant bedeutet nicht immer wichtig

Der Versuch, eine Nullhypothese nicht zu verwerfen, hat einige Auswirkungen
Regressionen sind nicht immer linear
Extrapolieren über die Grenzen eines Streuungsdiagramms hinaus ist eine ganz schlechte Idee
Untersuchen Sie die Variabilität um eine Regressionsgerade herum
Eine Stichprobe kann auch zu groß sein
Achse ist nicht gleich Achse
Eine kategorische Variable grafisch so darzustellen, als sei sie eine quantitative Variable, ist ganz einfach falsch
Wann immer möglich, sollten Sie Variabilität in Ihr Diagramm einschließen
Seien Sie vorsichtig, wenn Sie Konzepte aus Statistikbüchern in Excel übertragen

Kapitel 22: Zehn Dinge (eigentlich zwölf), die einfach nicht in die anderen Kapitel passten

Diagrammdarstellung des Standardfehlers des Mittelwerts
Wahrscheinlichkeiten und Verteilungen
Stichproben ziehen
Unabhängigkeit testen: die wahre Verwendung von CHIQU.TEST
Logarithmica Esoterica
Daten sortieren

Anahng A: Wenn Ihr Arbeitsblatt eine Datenbank ist

Einführung in Excel-Datenbanken
Die Trabanten-Datenbank
Der Kriterienbereich
Das Format einer Datenbankfunktion
Datensätze zählen und Auszüge aus der Datenbank
DBANZAHL und DBANZAHL2

DBAUSZUG
Arithmetik
DBMAX und DBMIN
DBSUMME
DBPRODUKT
Statistiken
DBMITTELWERT
DBVARIANZ und DBVARIANZEN
DBSTDABW und DBSTDABWN
Datenmasken verwenden
PivotTable-Berichte
Anahng B: Die Kovarianzanalyse
Kovarianz unter die Lupe genommen
Warum Sie die Kovarianz analysieren
Wie Sie die Kovarianz analysieren
ANCOVA mit Excel
Methode 1: ANOVA
Methode 2: Regression
Nach der ANCOVA
Eins hab ich noch
Stichwortverzeichnis
Eula

Kapitel 1

Einführung

Was? Noch ein Buch über Statistik? Nun … dies ist ein Buch über Statistik. Okay. Aber meiner bescheidenen (und absolut parteiischen) Meinung nach ist dies *nicht* einfach nur ein weiteres Buch über Statistik.

Was? Noch ein Buch über Excel? Dieselbe absolut parteiische Meinung: Es ist nicht einfach noch ein Buch über Excel.

In diesem Buch geht es vielmehr um Folgendes: In vielen Büchern über Statistik lernen Sie die Grundzüge der Statistik kennen, erfahren aber nicht, wie Sie diese anwenden können. Das führt häufig dazu, dass die Grundzüge nicht wirklich verstanden werden. Mit Excel steht Ihnen ein fix und fertiges Paket für die Anwendung dieser Grundzüge zur Verfügung.

Wenn wir das Ganze von der anderen Seite her betrachten, haben wir dasselbe: In Büchern über Excel erfahren Sie etwas über die Funktionen von Excel, aber über die diesen Funktionen zugrunde liegenden Begriffe erfahren Sie nichts. Bevor ich Ihnen etwas über eine Statistikfunktion von Excel erzähle, erkläre ich Ihnen die Begriffe, auf denen diese Funktion beruht. So lernen Sie die Funktion kennen, während Sie sie nutzen. Und Sie werden sie daher effektiver einsetzen können.

Ich wollte kein Buch schreiben, in dem nur beschrieben wird: »Wählen Sie dieses Menü aus« und »Klicken Sie auf diese Schaltfläche«. Klar, manchmal muss das sein, wenn in einem Buch die Verwendung eines Software-Pakets beschrieben wird. Aber ich wollte mich darauf nicht beschränken.

Ich wollte auch kein statistisches »Kochbuch« schreiben: Wenn Problem Nr. 310 auftritt, verwenden Sie das statistische Verfahren Nr. 214. Ich wollte mich auch darauf nicht beschränken.

Fazit: In diesem Buch geht es nicht nur um Statistik oder nur um Excel, sondern um beides. Während ich Ihnen etwas über Statistik erzähle, beschreibe ich alle Statistikfunktionen von Excel. (Na ja, *fast* alle. Eine werde ich nicht beschreiben. Nämlich die Fourier-Analyse. Schon allein die zum Erklären der Fourier-Analyse erforderlichen mathematischen Grundkenntnisse würden ein ganzes Buch füllen. Außerdem werden Sie diese Funktion wahrscheinlich sowieso nie brauchen.)

Über dieses Buch

Obwohl bei der Statistik die Themen logisch aufeinander aufbauen, habe ich dieses Buch so strukturiert, dass Sie ein beliebiges Kapitel aufschlagen und lesen können. Mir ist wichtig, dass Sie das Gesuchte schnell finden und sofort anwenden können, und das unabhängig davon, ob es sich um einen Begriff aus der Statistik oder um eine Funktion von Excel handelt.

Wenn es Sie jedoch interessiert, können Sie dieses Buch selbstverständlich auch von vorn bis hinten durchlesen. Wenn Sie sich mit Statistik noch nicht auskennen und Excel für statistische Analysen benötigen, empfehle ich Ihnen, vorne im Buch zu beginnen, auch wenn Sie sich mit Excel bereits ganz gut auskennen.

Was Sie nicht lesen müssen

In jedem Lehrbuch finden Sie jede Menge Informationen. Da ist auch dieses Buch keine Ausnahme. Ich habe versucht, nur Nützliches in das Buch aufzunehmen. Das ist mir jedoch nicht immer gleich gut gelungen. Wenn Sie also an einem Thema nicht so besonders interessiert sind, brauchen Sie die Abschnitte, die mit dem Symbol »Vorsicht Technik!«, das ist das Symbol mit dem Kopf vom Dummies-Mann, gekennzeichnet sind, nicht zu lesen.

Gelegentlich werden Sie auf kurze Beschreibungen in einem grauen Kasten treffen. Diese enthalten ausführlichere Informationen zu einem Thema, gehören aber nicht zum Hauptthema. Wenn Sie wenig Zeit haben, können Sie diese überspringen.

Da ich dieses Buch so geschrieben habe, dass Sie es an einer beliebigen Stelle aufschlagen und verwenden können, finden Sie im gesamten Buch schrittweise Anleitungen. Bei vielen Verfahren sind einige der Schritte gleich. Wenn Sie ein paar der Verfahren durchgearbeitet haben, werden Sie bei einem neuen Verfahren die ersten Schritte überspringen können.

Törichte Annahmen über den Leser

Bei diesem Buch handelt es sich nicht um eine Einführung in Excel oder Windows. Daher setze ich Folgendes voraus:

- Sie kennen sich mit Windows oder dem Mac aus. Ich werde nicht erläutern, wie man auf Elemente zeigt, klickt, etwas auswählt und so weiter.
- Auf Ihrem Windows-PC oder Ihrem Mac ist Excel installiert und Sie können die Beispiele nachvollziehen. Ich werde nicht erläutern, wie Excel installiert wird.
- Sie haben bereits mit Excel gearbeitet und Sie kennen sich im Wesentlichen mit Arbeitsblättern und Formeln aus.

  Wenn Sie sich mit Excel noch nicht so gut auskennen, helfen Ihnen sicherlich die verschiedenen Excel-Bücher aus der Dummies-Reihe weiter.

Wie dieses Buch aufgebaut ist

Ich habe dieses Buch in fünf Teile gegliedert. Außerdem finden Sie am Ende des Buches zwei spannende Anhänge.

Teil I: Statistik und Excel: Wie füreinander geschaffen

In Teil I finden Sie eine allgemeine Einführung in die Statistik und in die statistischen Funktionen von Excel. Es werden wichtige Begriffe der Statistik beschrieben und nützliche Excel-Techniken erläutert. Wenn Ihr letzter Kurs in Statistik schon eine Weile her ist oder wenn Sie noch nie an einem Kurs in Statistik teilgenommen haben, beginnen Sie am besten hier. Wenn Sie noch nie mit den in Excel integrierten Funktionen (gleich welcher Art) gearbeitet haben, müssen Sie auf jeden Fall mit diesem Teil beginnen.

Teil II: Daten beschreiben

Ein Teil der Statistik befasst sich mit dem Zusammenfassen von Zahlen, um damit bestimmte Aussagen machen zu können. In diesem Teil erfahren Sie, wie Sie dabei vorgehen müssen. Wir alle kennen Mittelwerte und wissen, wie wir diese berechnen. Aber das ist noch nicht alles. In diesem Teil werden Sie weitere Statistikfunktionen zum Füllen der Lücken kennenlernen. Und ich werde Ihnen zeigen, wie Sie in Excel mit diesen Statistikfunktionen arbeiten können. In diesem Teil finden Sie außerdem eine Einführung in die Diagrammfunktionen von Excel.

Teil III: Schlüsse aus Daten ziehen

In Teil III geht es um die eigentliche Aufgabe der statistischen Analyse: die Bedeutung der Zahlen zu erkennen und Entscheidungsträgern beim Treffen von Entscheidungen zu helfen. Normalerweise sind die Daten Messungen einer Stichprobe aus einer Grundgesamtheit. Ziel ist es, mit diesen Daten Entwicklungen in der Grundgesamtheit zu ermitteln.

Dies wirft eine Vielzahl von Fragen auf: Welche Bedeutung hat ein Mittelwert? Was bedeutet die Differenz zwischen zwei Mittelwerten? Haben zwei Dinge etwas miteinander zu tun? Das sind nur einige wenige der Fragen, um die es in Teil III geht. In diesem Teil werden die Funktionen und Tools von Excel vorgestellt, mit deren Hilfe Sie diese Fragen beantworten können.

Teil IV: Umgang mit der Wahrscheinlichkeit

Die Wahrscheinlichkeit ist die Grundlage für statistische Analysen und Entscheidungsfindungen. Teil IV hat die Wahrscheinlichkeit zum Thema. Hier erfahren Sie, wie Sie die Wahrscheinlichkeit insbesondere im Bereich der Modellierung einsetzen können. Excel enthält zahlreiche integrierte Funktionen, die Ihnen dabei helfen, Wahrscheinlichkeiten zu verstehen und anzuwenden. Und in diesem Teil werden diese Funktionen erläutert.

Teil V: Der Top-Ten-Teil

Teil V erfüllt zwei Aufgaben. Zum einen werde ich Sie ausführlich über statistische Fallstricke sowie hilfreiche Tipps informieren. Insgesamt sind das zehn. Ich werde außerdem zehn (na gut, zwölf) Excel-Dinge beschreiben, die sonst nirgendwo hingepasst haben. Sie kommen aus allen Bereichen der Statistik. Wenn es um Excel und Statistik geht und es nirgendwo im Buch zu finden ist, dann finden Sie es hier.

Ganz schön praktisch, der Top-Ten-Teil!

Anhang A: Wenn Ihr Arbeitsblatt eine Datenbank ist

Außer Berechnungen durchführen kann Excel noch etwas anderes: Datensätze anlegen. Auch wenn Excel nicht speziell ein Datenbankprogramm ist, so verfügt es dennoch über einige Datenbankfunktionen. Einige davon sind statistische Funktionen. Eine Einführung in die Datenbankfunktionen von Excel sowie in die Arbeit mit Pivot-Tabellen, mit deren Hilfe Sie Ihre Datenbank auf den Kopf stellen und Daten auf unterschiedliche Art und Weise betrachten können, finden Sie in Anhang A.

Anhang B: Die Kovarianzanalyse

Die Kovarianzanalyse (Englisch: Analysis of Covariance, ANCOVA) ist ein statistisches Verfahren, das zwei andere Verfahren verbindet:

die Varianzanalyse und die lineare Regressionsanalyse. Wenn Sie wissen, in welcher Beziehung zwei Variablen stehen, können Sie dieses Wissen auf raffinierte Weise einsetzen. Die Kovarianzanalyse ist eine dieser Möglichkeiten. Das Problem ist, dass Excel für ANCOVA kein eingebautes Werkzeug bereitstellt – jedoch zeige ich, wie Sie das nutzen können, was Excel anbietet, um diese Aufgabe erledigt zu bekommen.

Symbole, die in diesem Buch verwendet werden

Wie in allen Büchern der Dummies-Reihe finden Sie auch in diesem Buch überall Symbole. Dabei handelt es sich um kleine Bildchen am Seitenrand, anhand derer Sie erkennen können, worum es in dem Abschnitt daneben geht.

Dieses Symbol ist ein Zeichen für einen Hinweis oder eine einfache Lösung, um Ihnen die Arbeit zu erleichtern.

Dieses Symbol ist ein Zeichen für zeitlose Weisheiten, die Sie auch lange, nachdem Sie das Buch gelesen haben, noch brauchen werden.

Achten Sie auf dieses Symbol. Es weist darauf hin, dass Sie etwas besser nicht tun sollten, wenn Sie Ihre Arbeit nicht zunichtemachen wollen.

Wie bereits erwähnt, steht dieses Symbol für Abschnitte, die Sie überspringen können, wenn Statistik und Excel nicht gerade Ihre Leidenschaft sind.

Die Beispieldateien

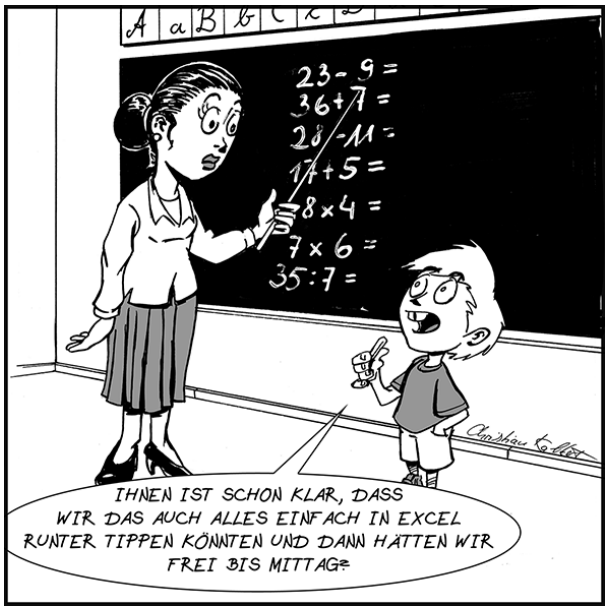
Alle Arbeitsmappen mit den Beispielen, die Sie in den Abbildungen dieses Buches sehen, können Sie von der Website zum Buch herunterladen. Die Adresse lautet: www.wiley-vch.de/publish/dt/books/ISBN3-527-71287-9. Laden Sie die ZIP-komprimierte Datei auf Ihren Computer herunter, entpacken Sie sie. Sie erhalten dann für jedes Kapitel einen Unterordner mit den Beispieldateien des jeweiligen Kapitels.

Wie es weitergeht

Sie können mit dem Lesen des Buchs an jeder beliebigen Stelle beginnen, aber hier sind noch ein paar Tipps. Sie möchten die Grundlagen der Statistik kennenlernen? Dann blättern Sie eine Seite weiter. Sie möchten sich die statistischen Funktionen von Excel selbst aneignen? Alles, was Sie dazu brauchen, finden Sie in Kapitel 2. Sie möchten lieber mit der Erstellung von Diagrammen beginnen? Dann schlagen Sie Kapitel 3 auf. Alles andere finden Sie im Inhaltsverzeichnis oder im Stichwortverzeichnis.

Teil I

Statistik und Excel: Wie füreinander geschaffen



In diesem Teil …

Die Fähigkeiten von Excel im Bereich Statistik kennenlernen

Erkunden, wie Sie mit Grundgesamtheiten und Stichproben arbeiten

Ihre Hypothesen testen

Fehler beim Fällen von Entscheidungen verstehen

Abhängige und unabhängige Variablen unterscheiden

Kapitel 1

Echte Daten auswerten

In diesem Kapitel

- Einführung in Begriffe der Statistik
- Von Stichproben Rückschlüsse auf Grundgesamtheiten ziehen
- Wahrscheinlichkeiten kennenlernen
- Entscheidungen treffen
- Neue und alte Features in Excel 2016
- Wichtige Excel-Grundlagen verstehen

Im Rahmen der Statistik geht es immer darum, Entscheidungen zu treffen, die auf Zahlengruppen beruhen. Statistiker stellen ständig Fragen: Was sagen uns die Zahlen? Welche Trends zeichnen sich ab? Welche Vorhersagen können wir treffen? Welche Schlüsse können wir ziehen?

Um diese Fragen zu beantworten, haben Statistiker eine beeindruckende Menge an Analysetools entwickelt. Mit diesen Tools wird den Bergen an Daten, die darauf warten, dass wir uns eingehend mit ihnen beschäftigen, eine Bedeutung zugeschrieben. Und mit diesen Tools können wir die Zahlen verstehen, die wir bei unserer Arbeit generieren.

Die statistischen (und damit verwandten) Begriffe, die Sie

einfach kennen müssen

Da intensives Rechnen häufig fester Bestandteil der Arbeit eines Statistikers ist, haben viele Leute die falsche Vorstellung, dass es bei der Statistik ausschließlich um die Verarbeitung großer Zahlenmengen geht. Das Rechnen ist jedoch nur ein kleiner Teil auf dem Weg hin zu einer vernünftigen Entscheidung.

Die Software nimmt uns diese Arbeit ab, so dass wir auf unserem Weg schneller vorankommen. Einige Software-Pakete sind auf die statistische Analyse spezialisiert und enthalten viele der Tools, die Statistiker verwenden. Excel wird zwar nicht explizit als Statistikpaket verkauft, enthält aber dennoch eine Reihe dieser Tools. Daher habe ich auch dieses Buch geschrieben.

Ich schrieb, Rechnen sei nur ein kleiner Teil auf dem Weg hin zu einer vernünftigen Entscheidung. Der wichtigste Teil sind die Konzepte, mit denen Statistiker arbeiten, und um diese geht es in diesem Kapitel in erster Linie.

Stichproben und Grundgesamtheiten

An Wahlabenden sagen Fernsehkommentatoren regelmäßig noch vor Schließung der Wahllokale das Ergebnis der Wahlen voraus. Meist liegen sie richtig. Wie geht das?

Ganz einfach: Eine Stichprobe von Wählern wird nach Abgabe ihrer Stimme befragt. Unter der Voraussetzung, dass die Wähler ehrlich sagen, wen sie gewählt haben, und vorausgesetzt, die Stichprobe ist für die Grundgesamtheit (oder Population) repräsentativ, können die Analysten in den Fernsehanstalten aufgrund der Stichprobendaten Rückschlüsse auf die Grundgesamtheit der Wähler ziehen.

Das ist die Aufgabe von Statistikern: aufgrund der Ergebnisse einer Stichprobe Rückschlüsse auf die Grundgesamtheit zu ziehen, aus der die Stichprobe entnommen wurde. Manchmal erweisen sich jedoch die anhand der Zahlen gezogenen Rückschlüsse als falsch. Das falsche Ergebnis einer Wahlumfrage führte zu dem

denkwürdigen Bild von US-Präsident Harry Truman mit einer Ausgabe der *Chicago Daily Tribune* in der Hand mit der berühmten, aber falschen Schlagzeile »Dewey Defeats Truman« (Dewey schlägt Truman) nach der Wahl 1948. Zu der Aufgabe eines Statistikers gehört es mitzuteilen, für wie realistisch er die Schlussfolgerung hält. Ein anderes Beispiel ebenfalls aus dem Bereich der Wahlforschung zeigt, dass derartige Schlussfolgerungen durchaus realistisch sein können. Das Ergebnis einer Wahlumfrage (wir gehen wieder von einer repräsentativen Stichprobe von Wählern aus) gibt an, wie viel Prozent der Wähler aus der Stichprobe die einzelnen Kandidaten favorisieren. Das Meinungsforschungsinstitut gibt an, für wie genau das Umfrageergebnis eingeschätzt wird. Wenn ein Nachrichtensprecher so etwas wie »auf 3 % genau« sagt, hören Sie eine Beurteilung der Glaubwürdigkeit.

Noch ein Beispiel. Nehmen wir einmal an, Sie haben die Aufgabe, die durchschnittliche Lesegeschwindigkeit aller Fünftklässler herauszufinden, Sie verfügen jedoch weder über die Zeit noch über die finanziellen Mittel, alle Fünftklässler zu testen. Was würden Sie tun?

Am besten nehmen Sie eine Stichprobe von Fünftklässlern, messen deren Lesegeschwindigkeit (in Wörtern pro Minute) und berechnen den Mittelwert dieser Lesegeschwindigkeit der Stichprobe. Sie können dann den Mittelwert der Stichprobe zur Schätzung des Mittelwerts der Grundgesamtheit heranziehen.

Das Schließen auf den Mittelwert einer Grundgesamtheit ist eine Art *Inferenz*, die Statistiker aus Stichprobendaten ziehen. Die Inferenz wird im Abschnitt »Inferenzstatistik: Testen von Hypothesen« ausführlicher beschrieben.

Einige Begriffe, die Sie kennen sollten: Die Eigenschaften einer Grundgesamtheit (wie der Mittelwert einer Grundgesamtheit) werden als *Parameter* bezeichnet und die Eigenschaften einer Stichprobe (wie der Mittelwert einer Stichprobe) als *Statistiken*. Wenn Sie sich bei Ihren

Betrachtungen auf Stichproben beschränken, sind Ihre Statistiken *deskriptiv* oder *beschreibend*. Wenn Sie Ihren Horizont erweitern und sich mit Grundgesamtheiten beschäftigen, sind Ihre Statistiken *inferenziell*.

Einige Schreibweisen, die Sie kennen sollten: Statistiker verwenden griechische Buchstaben (μ, σ, ρ) für Parameter und lateinische Buchstaben ($\bar{x}$, *s*, *r*) für Statistiken. In Abbildung 1.1 sind die Beziehungen zwischen Grundgesamtheiten und Stichproben sowie zwischen Parametern und Statistiken dargestellt.

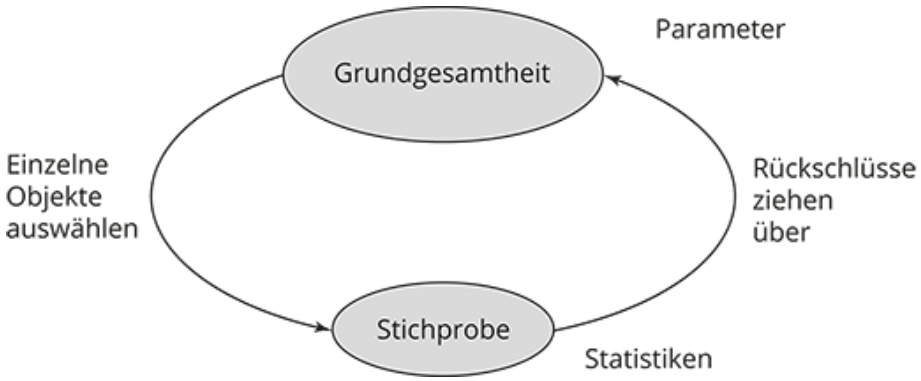


Abbildung 1.1: Die Beziehung zwischen Grundgesamtheiten, Stichproben, Parametern und Statistiken

Abhängige und unabhängige Variablen

Einfach ausgedrückt, ist eine *Variable* etwas, das mehrere Werte annehmen kann. (Etwas, das nur einen Wert annehmen kann, wird *Konstante* genannt.) Einige Variablen, die Sie bereits kennen, sind Tagestemperatur, Dow-Jones-Index, Ihr Alter und der Wert des Dollar in Euro.

Für Statistiker sind zwei Arten von Variablen wichtig: *unabhängige Variablen* und *abhängige Variablen*. Beide Variablen tauchen in jeder

Studie und Untersuchung auf und Statistiker bewerten die Beziehung zwischen beiden.

Stellen Sie sich beispielsweise vor, es gebe eine neue Möglichkeit, Lesen so zu lehren, dass Fünftklässler schneller lesen können. Bevor diese neue Methode an Schulen eingeführt wird, soll sie getestet werden. Dazu müsste ein Forscher eine Stichprobe von Fünftklässlern nach dem Zufallsprinzip in zwei Gruppen teilen. Eine Gruppe wird nach der neuen Methode unterrichtet, die andere nach der herkömmlichen. Der Forscher misst vor und nach dem Unterricht die Lesegeschwindigkeit aller Kinder, die an dieser Studie teilnehmen. Was dann geschieht, erfahren Sie in einem der nächsten Abschnitte (»Inferenzstatistik: Testen von Hypothesen«).

Hier geht es zunächst darum, dass Sie wissen, dass die unabhängige Variable in diesem Beispiel die *Unterrichtmethode* ist. Die beiden möglichen Werte dieser Variablen sind *Neu* und *Herkömmlich*. Die abhängige Variable ist die *Lesegeschwindigkeit*.

Grundsätzlich geht es darum herauszufinden, ob Änderungen der unabhängigen Variablen mit Änderungen der abhängigen Variablen zusammenhängen.

In den Beispielen in diesem Buch erfahren Sie, wie Sie verschiedene Eigenschaften von Wertegruppen mit Excel berechnen können. Denken Sie immer daran, dass ich mit einer Wertegruppe stets die Werte einer abhängigen Variablen meine.

Arten von Daten

Es gibt vier verschiedene Arten von Daten. Wenn Sie mit einer Variablen arbeiten, hängt es von der Datenart ab, wie Sie mit der Variablen arbeiten.

Die erste Art wird als *nominalskalierte* oder *nominale* Daten bezeichnet. Wenn eine Zahl eine nominale Variable ist, handelt es sich lediglich um einen Namen. Der Zahlenwert bedeutet nichts. Ein gutes Beispiel hierfür ist die Zahl auf dem Trikot eines Sportlers. Sie dient lediglich der Identifizierung des Sportlers, um ihn von den anderen Mitgliedern seines Teams unterscheiden zu können. Die Zahl ist kein Hinweis auf das Können des Sportlers.

Als Nächstes kommen die *ordinalskalierten* oder *ordinalen* Daten. Bei ordinalen Daten geht es um Ordnung. Die Zahlen erhalten eine Bedeutung, die über die bloße Identifizierung hinausgeht. Eine höhere Zahl bedeutet, dass eine Eigenschaft in einem höheren Maß vorhanden ist als bei einer niedrigeren Zahl. Ein Beispiel hierfür ist die Mohssche Härteskala. Diese Skala wird seit 1822 verwendet und gibt Werte zwischen 1 und 10 an. Mit dieser Skala geben Mineralogen den Härtegrad von Mineralen an. Diamant ist mit dem Härtegrad 10 das härteste Mineral und Talk mit dem Härtegrad 1 das weichste. Mit einem Mineral einer bestimmten Härte lässt sich jedes Mineral mit einer geringeren Härte ritzen.

Was bei der Mohsschen Skala (und allen Ordinalskalen) fehlt, ist das Konzept von gleichen Intervallen oder gleichen Differenzen. Die Differenz zwischen dem Härtegrad 10 und dem Härtegrad 8 ist nicht dieselbe wie zwischen dem Härtegrad 6 und dem Härtegrad 4.

Intervallskalierte Daten geben Differenzen an. Temperaturangaben in Celsius und Fahrenheit sind ein Beispiel für intervallskalierte Daten. Die Differenz zwischen 10 °C und 20 °C ist dasselbe wie die Differenz zwischen 30 °C und 40 °C.

Eine Tatsache bei den Temperaturangaben in Celsius oder Fahrenheit wird Sie überraschen: 20 °C ist nicht doppelt so warm wie 10 °C. Um eine Aussage hinsichtlich der Relation (doppelt so viel wie, halb so viel wie) machen zu können, muss null bedeuten, dass vom gemessenen Attribut absolut nichts vorhanden ist. Ein Temperaturwert von 0 °C bedeutet jedoch nicht, dass keine Wärme vorhanden ist. 0 °C ist lediglich ein willkürlicher Punkt auf der Celsius-Skala.

Zum letzten Datentyp zählen *verhältnisskalierte* Daten. Hier ist ein sinnvoll interpretierbarer Nullpunkt vorhanden. Bei Temperaturangaben liefert die Kelvin-Skala verhältnisskalierte Daten. 100 °K ist doppelt so warm wie 50 °K. Dies liegt daran, dass der Nullpunkt der Kelvin-Skala ein *absoluter Nullpunkt* ist, bei dem es keine molekulare Bewegung (die Voraussetzung für Wärme) mehr gibt. Ein weiteres Beispiel ist das Lineal. 8 cm ist doppelt so lang wie 4 cm. Der Wert 0 bedeutet, dass keine Länge vorhanden ist.

All diese Datenarten können die Basis einer unabhängigen oder einer abhängigen Variablen bilden. Welche Analysetools Sie verwenden, hängt von der Art der Daten ab, mit denen Sie zu tun haben.

Ein bisschen Wahrscheinlichkeit

Wenn Statistiker Rückschlüsse ziehen, drücken sie ihre Einschätzung der Glaubwürdigkeit dieser Rückschlüsse in Form von Wahrscheinlichkeiten aus. Sie können sich ihrer Rückschlüsse nie sicher sein. Sie können nur sagen, für wie wahrscheinlich sie ihre Rückschlüsse halten.

Was also ist Wahrscheinlichkeit? Das erläutere ich am besten anhand von ein paar Beispielen. Wie groß ist die Wahrscheinlichkeit, dass beim Werfen einer Münze Kopf geworfen wird? Intuitiv wissen Sie, dass die Chancen für Kopf ebenso wie für Zahl 50:50 stehen. Im Hinblick auf die zur Wahrscheinlichkeit gehörende Art der Zahlen ist das 1/2.

Und wie ist das beim Würfeln? Wie groß ist die Wahrscheinlichkeit, dass Sie eine 3 würfeln? Hmmm … Ein Würfel hat sechs Flächen und eine davon zeigt die 3, also sollte die Wahrscheinlichkeit bei 1 zu 6 liegen, richtig? Richtig.

Noch ein Beispiel. Sie ziehen aus einem Stapel Spielkarten wahllos eine Karte heraus. Wie groß ist die Wahrscheinlichkeit, dass Sie

Kreuz ziehen? Nun, ein Kartenspiel hat vier Farben, also lautet die Antwort 1 zu 4.

Ich glaube, Sie verstehen, worum es geht. Wenn Sie ermitteln möchten, mit welcher Wahrscheinlichkeit ein Ereignis eintritt, müssen Sie herausfinden, wie häufig dieses Ereignis eintreten kann, und Sie müssen diese Anzahl durch die Gesamtzahl aller möglichen Ereignisse teilen. Bei unseren drei Beispielen tritt das fragliche Ereignis (Kopf, 3 bzw. Kreuz) nur einmal ein.

Das Ganze kann jedoch noch etwas komplexer werden. Wie groß ist die Wahrscheinlichkeit, dass beim Würfeln eine 3 oder eine 4 gewürfelt wird? Nun kann das fragliche Ereignis zweimal eintreten, d. h. (1 + 1)/6 = 2/6 = 1/3. Und wie groß ist die Wahrscheinlichkeit, dass eine gerade Zahl gewürfelt wird? Das bedeutet, dass eine 2, 4 oder 6 gewürfelt wird und die Wahrscheinlichkeit somit (1 + 1 + 1)/6 = 3/6 = 1/2 beträgt.

Hinsichtlich der Wahrscheinlichkeit stellen sich noch weitere Fragen. Nehmen wir einmal an, Sie würfeln und werfen gleichzeitig eine Münze. Wie groß ist die Wahrscheinlichkeit, dass Sie eine 3 würfeln und Kopf werfen? Berücksichtigen Sie alle möglichen Ereignisse, die eintreten können, wenn Sie würfeln und gleichzeitig eine Münze werfen. Sie können Kopf und die Zahlen 1 bis 6 oder Zahl und die Zahlen 1 bis 6 werfen. Das ergibt insgesamt 12 Möglichkeiten. Für Kopf und 3 gibt es nur eine Möglichkeit. Also lautet die Lösung 1/12.

Die Formel für die Wahrscheinlichkeit, mit der ein bestimmtes Ereignis eintritt, lautet wie folgt:

$$\mathrm{P(Ereignis)} = \frac{\text{Anzahl der Möglichkeiten, mit denen ein Ereignis eintreten kann}}{\text{Gesamtzahl der möglichen Ereignisse}}$$

Ich habe diesen Abschnitt mit der Feststellung begonnen, dass Statistiker ihre Einschätzung der Glaubwürdigkeit von Rückschlüssen in Form von Wahrscheinlichkeiten ausdrücken, weshalb ich eigentlich auf dieses Thema gekommen bin. Wenn wir in diese Richtung weiterdenken, stoßen wir auf den Begriff der *bedingten* Wahrscheinlichkeit, also die Wahrscheinlichkeit, mit der

ein Ereignis eintritt, wenn ein anderes Ereignis eintritt. Nehmen wir einmal an, ich würfle, schaue mir das Ergebnis an (so dass Sie es nicht sehen können) und sage Ihnen, dass ich eine gerade Zahl gewürfelt habe. Wie groß ist die Wahrscheinlichkeit, dass ich eine 2 gewürfelt habe? Eigentlich beträgt die Wahrscheinlichkeit einer 2 1/6, aber ich habe die Auswahl begrenzt. Ich habe die drei ungeraden Zahlen (1, 3 und 5) als Möglichkeiten ausgeschlossen. Somit sind nur noch die drei geraden Zahlen (2, 4 und 6) möglich, so dass die Wahrscheinlichkeit, dass eine 2 gewürfelt wird, nun 1/3 beträgt.

Was hat nun die bedingte Wahrscheinlichkeit mit statistischer Analyse zu tun? Lesen Sie weiter.

Inferenzstatistik: Testen von Hypothesen

Vor dem Durchführen einer Studie formuliert ein Statistiker eine *Hypothese*, das heißt, er stellt eine vorsichtige Prognose auf, welches bestimmte Ergebnis zu erwarten ist. Wenn nach Abschluss der Untersuchung die Stichprobendaten in einer Tabelle erfasst sind, trifft er die zentrale Entscheidung, die ein Statistiker treffen muss: Er entscheidet, ob die Hypothese verworfen oder nicht verworfen wird.

Diese Entscheidung hängt von der Frage nach der bedingten Wahrscheinlichkeit ab: Wie groß ist die Wahrscheinlichkeit, dass sich diese Daten unter der Voraussetzung ergeben, dass die Hypothese zutrifft? Die statistische Analyse stellt Tools zum Berechnen der Wahrscheinlichkeit bereit. Wenn sich die Wahrscheinlichkeit als gering erweist, verwirft der Statistiker die Hypothese.

Ein Beispiel: Nehmen wir einmal an, Sie möchten wissen, ob eine bestimmte Münze symmetrisch ist, das heißt, ob Kopf ebenso häufig geworfen wird wie Zahl. Um diese Frage zu klären, werfen Sie die Münze beispielsweise hundert Mal. Diese 100 Würfe stellen Ihre Stichprobendaten dar. Wenn Sie von der Hypothese ausgehen, dass

die Münze symmetrisch ist, erwarten Sie, dass die Daten in Ihrer Stichprobe mit 100 Würfen 50 Mal Kopf und 50 Mal Zahl ergeben.

Wenn sich herausstellt, dass Sie 99 Mal Kopf und 1 Mal Zahl werfen, werden Sie die Hypothese von der symmetrischen Münze zweifellos verwerfen. Warum? Die bedingte Wahrscheinlichkeit, dass mit einer symmetrischen Münze 99 Mal Kopf und 1 Mal Zahl geworfen wird, ist sehr gering. Aber einen Moment. Die Münze kann symmetrisch sein und Sie können dennoch 99 Mal Kopf und 1 Mal Zahl werfen, richtig? Absolut. Das weiß man nie so genau. Sie müssen Stichprobendaten sammeln (das Ergebnis aus 100 Würfen) und Rückschlüsse ziehen. Die Rückschlüsse können richtig sein oder auch nicht.

Geschworene stehen ständig vor dieser Frage. Sie müssen zwischen widersprüchlichen Hypothesen entscheiden und die Indizien vor Gericht begründen. (Stellen Sie sich die Indizien als Daten vor.) Eine Hypothese lautet, dass der Angeklagte schuldig ist. Die andere Hypothese lautet, dass der Angeklagte unschuldig ist. Die Geschworenen müssen unter Berücksichtigung der Indizien im Prinzip die Frage nach der bedingten Wahrscheinlichkeit beantworten. Wie groß ist die Wahrscheinlichkeit des Indizes, vorausgesetzt, der Angeklagte ist nicht schuldig? Diese Frage wird durch den Urteilsspruch beantwortet.

Nullhypothese und Alternativhypothese

Betrachten wir noch einmal das eben beschriebene Experiment mit dem Münzenwerfen. Die Ergebnisse aus 100 Würfen stellen die Stichprobendaten dar. Vor dem Werfen der Münze formulieren Sie die Hypothese, dass die Münze symmetrisch ist, das heißt, Sie erwarten, dass Kopf und Zahl gleich häufig geworfen werden. Dieser Ausgangspunkt wird als *Nullhypothese* bezeichnet. In der Statistik wird für die Nullhypothese die Notation H_0 verwendet. Nach dieser Hypothese ist jede Kopf-Zahl-Verteilung in den Daten mit einer symmetrischen Münze vereinbar. Stellen Sie sich das Ganze so vor,

dass nichts in den Ergebnissen der Untersuchung außer der Reihe ist.

Eine alternative Hypothese ist möglich, nämlich dass die Münze nicht symmetrisch ist und daher Kopf und Zahl nicht gleich häufig geworfen werden. Diese Hypothese besagt, dass mit einer nicht symmetrischen Münze jede Kopf-Zahl-Verteilung mit einer nicht symmetrischen Münze vereinbar ist. Ob Sie es glauben oder nicht: Die alternative Hypothese wird *Alternativhypothese* oder *Gegenhypothese* genannt. In der Statistik wird für die Alternativhypothese die Notation H_1 oder H_A verwendet.

Werfen Sie, diese Hypothesen vorausgesetzt, die Münze 100 Mal und notieren Sie die Anzahl der Kopf- und Zahl-Würfe. Wenn sich dabei ergibt, dass etwa 90 Mal Kopf und 10 Mal Zahl geworfen wird, sollten Sie H_0 verwerfen. Wenn sich ergibt, dass Kopf und Zahl jeweils etwa 50 Mal geworfen werden, sollten Sie H_0 nicht verwerfen.

Ähnliches gilt für das Beispiel mit der Lesegeschwindigkeit weiter vorne in diesem Kapitel. Eine Stichprobe von Kindern lernt nach einer neuen Methode lesen, mit der die Lesegeschwindigkeit erhöht werden soll, während die andere Stichprobe nach der herkömmlichen Methode lesen lernt. Die Lesegeschwindigkeit der Kinder wird vor und nach dem Unterricht gemessen und der Fortschritt der einzelnen Kinder wird tabellarisch erfasst. Die Nullhypothese H_0 besagt, dass sich die beiden Methoden nicht voneinander unterscheiden. Wenn der Fortschritt mit der neuen Methode größer ist als mit der herkömmlichen Methode, so viel größer, dass es unwahrscheinlich ist, dass sich die Methoden nicht voneinander unterscheiden, verwerfen Sie H_0. Wenn nicht, dann verwerfen Sie H_0 nicht.

Ist Ihnen aufgefallen, dass ich *nicht* gesagt habe: »Nehmen Sie H_0 an«? So, wie die Logik nun mal funktioniert, können Sie eine Hypothese *niemals* annehmen. Sie können H_0 verwerfen oder Sie können H_0 nicht verwerfen.

Hier noch ein Beispiel aus dem echten Leben, das dabei helfen soll, dieses Konzept zu verstehen. Wenn ein Angeklagter vor Gericht erscheint, gilt die Unschuldsvermutung, bis das Gegenteil bewiesen ist. Sie können sich »nicht schuldig« als H_0 vorstellen. Die Aufgabe des Staatsanwalts besteht darin, das Gericht dazu zu bringen, H_0 zu verwerfen. Wenn das Gericht die Unschuldsvermutung ablehnt, wird der Angeklagte für schuldig befunden. Wenn es diese Vermutung verwirft, dann ist der Angeklagte schuldig. Wenn es die Unschuldsvermutung nicht verwirft, dann ist der Angeklagte nicht schuldig. Der Angeklagte ist niemals unschuldig. Dies wäre gleichbedeutend damit, H_0 zu akzeptieren.

Lassen Sie uns zum Beispiel mit dem Münzenwerfen zurückkehren. Ist Ihnen aufgefallen, dass ich beim Beispiel mit dem Münzenwerfen »etwa 50 Mal« gesagt hatte? Was bedeutet dieses »etwa«? Außerdem habe ich gesagt, dass Sie H_0 verwerfen sollen, wenn Kopf und Zahl im Verhältnis 90:10 geworfen wird. Aber was ist, wenn 85:15 geworfen wird? 80:20? 70:30? Wie groß muss die Differenz zur Verteilung 50:50 sein, damit H_0 verworfen wird? Um wie viel größer muss beim Beispiel mit der Lesegeschwindigkeit der Fortschritt sein, damit H_0 verworfen wird?

Ich werde diese Fragen hier nicht beantworten. Statistiker haben Entscheidungsregeln für Situationen wie diese entwickelt, und Sie werden diese Regeln im Verlauf dieses Buches kennenlernen.

Zwei Arten von Fehlern

Beim Auswerten der Daten aus einer Untersuchung und beim Entscheiden, ob H_0 verworfen werden soll oder nicht, können Sie nie absolut sicher sein. Sie wissen nie, wie die Realität wirklich aussieht. Im Zusammenhang mit dem Münzwurfbeispiel bedeutet das, dass Sie nie sicher wissen, ob die Münze symmetrisch ist. Ihnen bleibt nur, eine Entscheidung anhand der gesammelten Stichprobendaten zu treffen. Wenn Sie, was die Münze betrifft, sicher gehen möchten, müssen Sie alle Daten für die gesamte Grundgesamtheit der Würfe sammeln. Das bedeutet, Sie müssten die Münze bis ans Ende aller Tage werfen.

Da Ihre Entscheidung nie sicher ist, ist es möglich, dass Sie einen Fehler machen, gleichgültig, wie Sie entscheiden. Wie bereits erwähnt, kann die Münze symmetrisch sein und Sie können bei 100 Würfen dennoch ein Ergebnis von 99:1 erhalten. Das ist nicht wahrscheinlich, weshalb Sie H_0 verwerfen. Es ist außerdem möglich, dass die Münze nicht symmetrisch ist, und bei 100 Würfen dennoch 50 Mal Kopf geworfen wird. Auch das ist nicht wahrscheinlich, weshalb Sie H_0 in diesem Fall nicht verwerfen.

Obwohl diese Fehler nicht wahrscheinlich sind, sind sie dennoch möglich. Sie kommen in jeder Untersuchung vor, bei der Inferenzstatistik im Spiel ist. Statistiker nennen diese Fehler *Fehler 1. Art* (oder auch Alphafehler) und *Fehler 2. Art* (oder auch Betafehler).

Wenn Sie H_0 verwerfen, obwohl Sie das nicht sollten, dann ist das ein Fehler 1. Art. Das wäre bei dem Beispiel mit der Münze das Verwerfen der Hypothese, die besagt, dass die Münze symmetrisch ist, obwohl die Münze tatsächlich symmetrisch ist.

Wenn Sie H_0 nicht verwerfen, obwohl Sie das sollten, dann ist das ein Fehler 2. Art. Das ist dann der Fall, wenn Sie die Hypothese, die besagt, dass die Münze symmetrisch ist, nicht verwerfen, obwohl die Münze tatsächlich nicht symmetrisch ist.

Woher wissen Sie, ob Sie einen dieser Fehler gemacht haben? Das können Sie nicht wissen, zumindest nicht gleich, nachdem Sie entschieden haben, ob Sie H_0 verwerfen oder nicht. (Wenn es möglich wäre, das zu wissen, würden Sie den Fehler erst gar nicht machen!) Ihnen bleibt nur, weitere Daten zu sammeln und zu prüfen, ob die zusätzlichen Daten mit Ihrer Entscheidung vereinbar sind.

Wenn Sie meinen, H_0 neige dazu, den Status quo zu erhalten, und nichts als außergewöhnlich interpretieren (gleichgültig, wie es aussieht), bedeutet ein Fehler 2. Art, dass Sie etwas Wichtiges übersehen haben. So betrachtet, basieren viele ironische Ereignisse in der Geschichte auf Fehlern 2. Art.

Und das meine ich damit: In den 50er-Jahren bekamen talentierte junge Entertainer in einer amerikanischen Fernsehsendung die

Gelegenheit, ein paar Minuten lang auf der Bühne zu zeigen, was sie können. Für ihre Darbietungen konnten sie einen Preis gewinnen. Das Publikum stimmte ab, wer der Gewinner sein sollte. Die Produzenten rekrutierten in ganz Amerika Interessenten für die Show. Viele Jahre nachdem die Show nicht mehr gesendet wurde, wurde einer der Produzenten interviewt. Der Interviewer fragte ihn, ob er jemals jemanden beim Casting abgelehnt habe, den er besser nicht hätte ablehnen sollen.

»Nun«, sagte der Produzent, »einmal hat ein junger Sänger vorgesungen und er schien wirklich schlecht zu sein.«

»In welcher Hinsicht?«, fragte der Interviewer.

»In vielerlei Hinsicht«, antwortete der Produzent. »Er sang viel zu laut, wirbelte beim Gitarrespielen seinen Körper und seine Beine herum und er trug diese langen Koteletten. Wir waren der Ansicht, dieser Junge würde es nie schaffen, dankten ihm für seine Vorführung und schickten ihn nach Hause.«

»Moment mal, möchten Sie mir sagen, dass Sie …«

»Ja, ganz recht. Wir haben … Elvis Presley nach Hause geschickt!«

Das war in der Tat ein Fehler 2. Art.

Was ist neu in Excel 2016?

Microsoft hat im Menüband (den Registerkarten, die Sie oben im Fenster sehen) einige Änderungen vorgenommen, die Änderungen in Excel widerspiegeln. Die auffälligste Neuerung ist die Glühlampe rechts von der Registerkarte ADD-INS. Das Feld neben der Glühlampe ist mit *Was möchten Sie tun?* beschriftet. Hierbei handelt es sich um eine neue Möglichkeit, die Excel-Hilfe aufzurufen. Geben Sie in das Feld einen Satz wie **Diagramm einfügen** ein, damit Excel ein Menü öffnet, in dem Sie Schaltflächen zu den Befehlen finden, mit denen Sie Diagramme einfügen können. Außerdem enthält das Menü Befehle, mit denen Sie die Hilfe zum Thema »Diagramm einfügen« anzeigen lassen können. Ein Beispiel dafür sehen Sie in Abbildung 1.2.

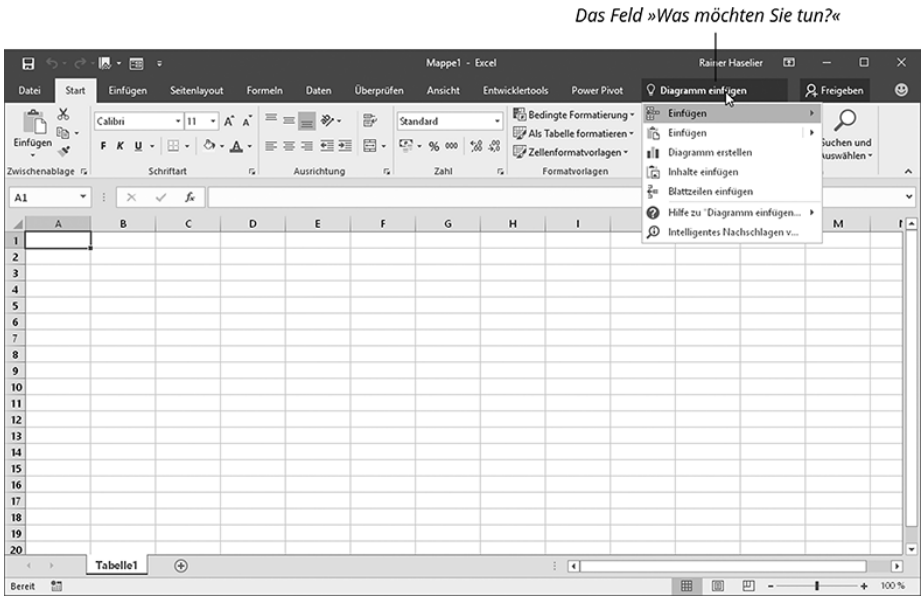


Abbildung 1.2: Die Benutzeroberfläche von Excel 2016 mit dem Feld WAS MÖCHTEN SIE TUN?

Leider steht dieses Feature in Excel 2016 für Mac nicht zur Verfügung. Dies trifft auch auf einige andere Features zu, von denen ich ein paar im folgenden Absatz erwähne. Im Großen und Ganzen wurde die Konsistenz zwischen der Windows- und der Mac-Version von Excel im Vergleich zu den vorherigen Versionen deutlich verbessert.

Abbildung 1.2 zeigt die Registerkarte EINFÜGEN, auf der sich für den Bereich DIAGRAMME einige Änderungen befinden. Eine Neuerung sind die statistischen Diagramme (die in der Mac-Version nicht vorhanden sind). Eine weitere sind die 3D-Karten, die neuen und verbesserten Power-View-Berichte (die in Excel 2013 eingeführt wurden und die in der MacOS-Version von Excel nicht zur Verfügung stehen). Diese Features werden in Kapitel 3 beschrieben.

Was ist alt in Excel 2016?

Jede Registerkarte im Menüband enthält Befehlsgruppen, in denen sich die zu einer bestimmten Kategorie gehörenden Schaltflächen befinden. Um herauszufinden, welche Aktion eine bestimmte Schaltfläche auslöst, bewegen Sie den Mauszeiger auf diese Schaltfläche (ohne zu klicken); ein kleines Fenster zeigt dann hilfreiche Informationen an.

Bestimmte Schaltflächen lösen nicht sofort eine Aktion aus; stattdessen zeigt ein Klick auf die Schaltfläche ein Menü an, in dem Sie den gewünschten Befehl auswählen können.

Microsoft verwendet eine Kurzschreibweise, um einen Mausklick auf eine Befehlsschaltfläche im Menüband anzugeben. Diese Schreibweise verwende ich auch in diesem Buch. Sie lautet:

REGISTERKARTE | NAME DER SCHALTFLÄCHE

Um beispielsweise anzugeben, dass auf der Registerkarte EINFÜGEN die Schaltfläche EMPFOHLENE DIAGRAMME angeklickt werden soll, schreibe ich

EINFÜGEN | EMPFOHLENE DIAGRAMME

Wenn ich diese Schaltfläche anklicke (nachdem vorher Zellen mit Daten markiert wurden), wird das Dialogfeld DIAGRAMM EINFÜGEN angezeigt, das Sie in Abbildung 1.3 sehen.

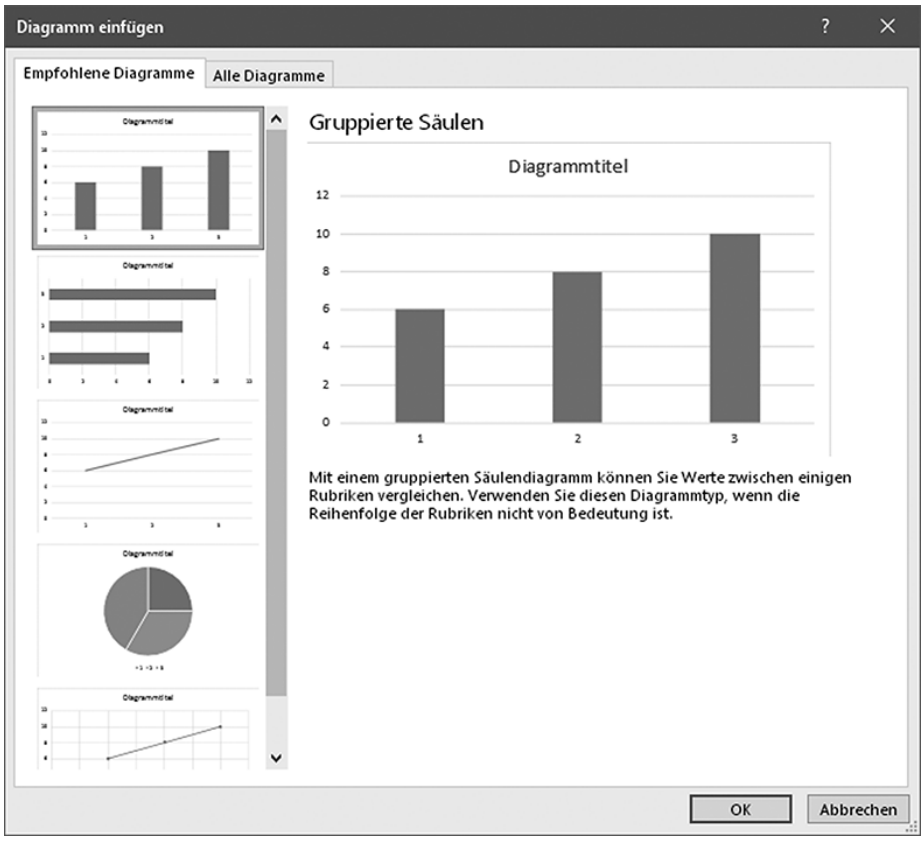


Abbildung 1.3: Wenn Sie EINFÜGEN | EMPFOHLENE DIAGRAMME anklicken, wird dieses Dialogfeld geöffnet.

Beachten Sie, dass im Dialogfeld die Registerkarte EMPFOHLENE DIAGRAMME geöffnet ist. Wenn Sie die Registerkarte ALLE DIAGRAMME anklicken (die in der Mac-Version nicht vorhanden ist), wird im Dialogfeld ein Katalog aller von Excel unterstützten Diagramme angezeigt (siehe Abbildung 1.4).

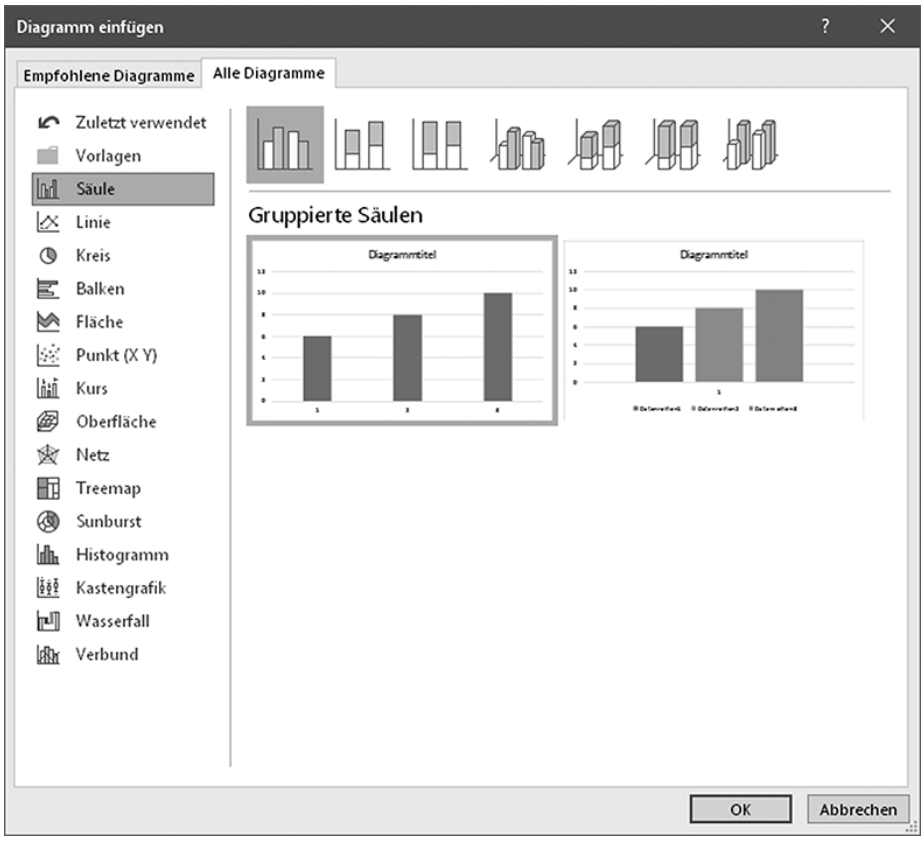


Abbildung 1.4: Die Registerkarte ALLE DIAGRAMME des Dialogfeldes DIAGRAMM EINFÜGEN

In der Kategorienliste an der linken Seite werden fünf der sechs neuen Diagrammtypen von Excel 2016 angezeigt: Wasserfall, Treemap, Sunburst, Histogramm und Kastengrafik. (Der sechste neue Diagrammtyp, Pareto, wird als Untertyp der Kategorie Histogramm aufgeführt). Die letzten drei Typen werden statistische Diagramme genannt. Die statistischen (sowie weitere) Diagramme werden in Kapitel 3 beschrieben.

Den größten Teil der statistischen Funktionalität finden Sie, wenn Sie folgenden Befehl wählen:

FORMELN | MEHR FUNKTIONEN | STATISTIK

Dies ist eine Erweiterung der Kurzschreibweise. Sie bedeutet: »Öffnen Sie die Registerkarte FORMELN, klicken Sie auf die Schaltfläche MEHR FUNKTIONEN und wählen Sie dann im Menü der Schaltfläche den Befehl STATISTIK aus. Abbildung 1.5 zeigt, was ich damit meine.

In Kapitel 2 zeige ich Ihnen, wie Sie das Menü mit den statistischen Funktionen einfacher zugänglich machen können.

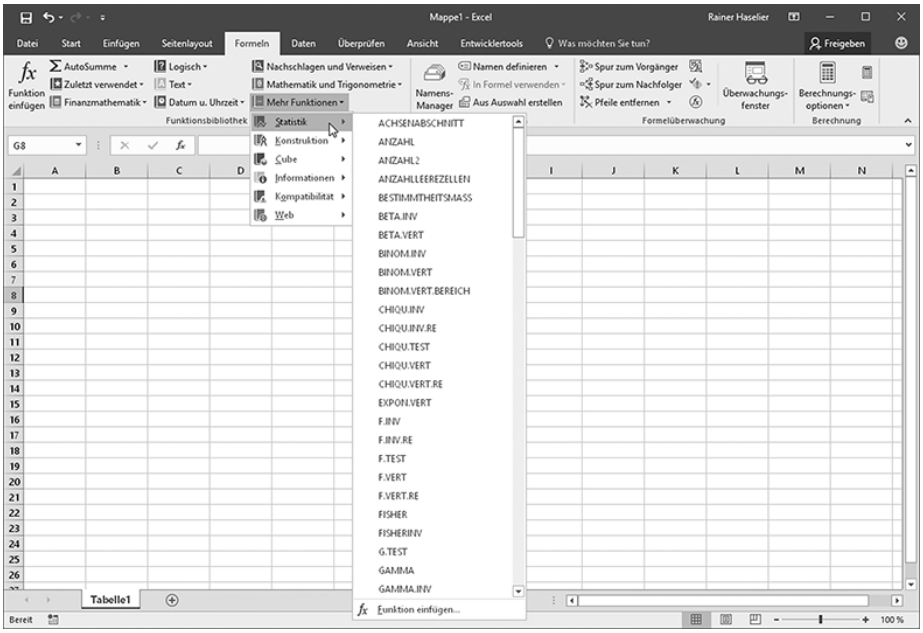


Abbildung 1.5: Auf das Menü mit den statistischen Funktionen zugreifen

In Excel 2010 hat Microsoft das Prinzip der Namensgebung der Funktionen geändert. Das Ziel hierbei war, Namen zu verwenden, die den Zweck und die Aufgabe der Funktion so gut wie möglich beschreiben. Außerdem hat Microsoft die Programmierung hinter einigen Funktionen geändert, damit diese genauer rechnen.

Auch Excel 2016 verwendet diese neue Namensgebung. Aus Kompatibilitätsgründen mit vorherigen Excel-Versionen stehen auch die alten Funktionsnamen (die meisten aus Excel vor 2010 und die Funktion `SCHÄTZER` aus Excel 2013) weiterhin zur Verfügung. Falls Sie also eine Arbeitsmappe für Benutzer von älteren Excel-Versionen erstellen, sollten Sie die alten Funktionsnamen verwenden.

Sie finden die alten Funktionen nicht im Menü mit den statistischen Funktionen, sondern in einem eigenen Menü.

Um es zu öffnen, wählen Sie FORMELN | MEHR FUNKTIONEN | KOMPATIBILITÄT.

Tabelle 1.1 hilft Ihnen beim Umsteigen von älteren Excel-Versionen. Die Tabelle führt den alten und den neuen Funktionsnamen auf sowie das Kapitel, in dem ich die neue Funktion vorstelle.

In der Tabelle können Sie erkennen, dass anstelle der alten Funktion SCHÄTZER nun fünf verschiedene Funktionen zur Verfügung stehen: PROGNOSE.LINEAR, PROGNOSE.ETS, PROGNOSE.ETS.KONFINT, PROGNOSE.ETS.SAISONALITÄT und PROGNOSE.ETS.STAT. Ich stelle diese Funktionen gemeinsam mit den neuen Prognosefeatures, die über einen Mausklick erreichbar sind, in Kapitel 16 vor.

Alte Funktion	Neue Funktion(en)	Kapitel
BETAINV	BETA.INV	19
BETAVERT	BETA.VERT	19
BINOMVERT	BINOM.VERT	18
CHIINV	CHIQU.INV.RE	10
CHITEST	CHIQU.TEST	20
CHIVERT	CHIQU.VERT.RE	10
EXPONVERT	EXPON.VERT	19
FINV	F.INV	11
FTEST	F.TEST	11
FVERT	F.VERT, F.VERT.RE	11
GAMMAINV	GAMMA.INV	19
GAMMAVERT	GAMMA.VERT	19
GTEST	G.TEST	10
HYPGEOMVERT	HYPGEOM.VERT	18
KONFIDENZ	KONFIDENZ.NORM, KONFIDENZ.T	9
KOVAR	KOVARIANZ.P, KOVARIANZ.S	15

Alte Funktion	Neue Funktion(en)	Kapitel
KRITBINOM	BINOM.INV	18
LOGINV	LOGNORM.INV	22
LOGNORMVERT	LOGNORM.VERT	22
MODUS	MODUS.EINF, MODUS.VIELF	4
NEGBINOMVERT	NEGBINOM.VERT	18
NORMINV	NORM.INV	8
NORMSINV	NORM.S.INV	8
NORMSVERT	NORM.S.VERT	8
NORMVERT	NORM.VERT	8
POISSON	POISSON.VERT	19
QUANTIL	QUANTIL.INKL, QUANTIL.EXKL	6
QUANTILRANG	QUANTILSRANG.INKL	6
QUARTILE	QUARTILE.INKL, QUARTILE.EXKL	6
RANG	RANG.GLEICH	6
SCHÄTZER	PROGNOSE.LINEAR, PROGNOSE.ETS, PROGNOSE.ETS.KONFINT, PROGNOSE.ETS.SAISONALITÄT, PROGNOSE.ETS.STAT	16
STABW	STABW.S	5
STABWN	STABW.N	5
TINV	T.INV, T.INV.2S	9
TTEST	T.TEST	11
TVERT	T.VERT.2S	10
TVERT	T.VERT.RE	10
VARIANZ	VAR.S	5
VARIANZEN	VAR.P	5
WEIBULL	WEIBULL.VERT	22

Tabelle 1.1: Alte Namen der statistischen Excel-Funktionen, deren Ersatz und das Kapitel, in dem die neue Funktion beschrieben wird

Die wichtigste Neuerung in Excel 2016 finden Sie in der Version für den Mac. Nach langer Abwesenheit steht nun wieder das Add-On Analyse-Funktionen zur Verfügung. Das Add-On Analyse-Funktionen, das in allen Windows-Versionen zur Verfügung steht, ist ein kostenloses Add-On mit Werkzeugen für komplexe statistische Analysen, die oft nur in speziellen Statistikanwendungen zur Verfügung stehen. In vorherigen Mac-Versionen konnten furchtlose Anwender eine ähnliche Sammlung von Werkzeugen verwenden, indem Sie eine Anwendung von Drittanbietern herunterluden, die jedoch nicht so gut in Excel integriert war wie das Add-On Analyse-Funktionen.

Es gibt jedoch eine Menge Mac-Anwender; diese werden über diese Änderung in Excel 2016 sehr erfreut sein.

Ich stelle das Add-On Analyse-Funktionen in Kapitel 2 vor.

Die Grundlagen kennen

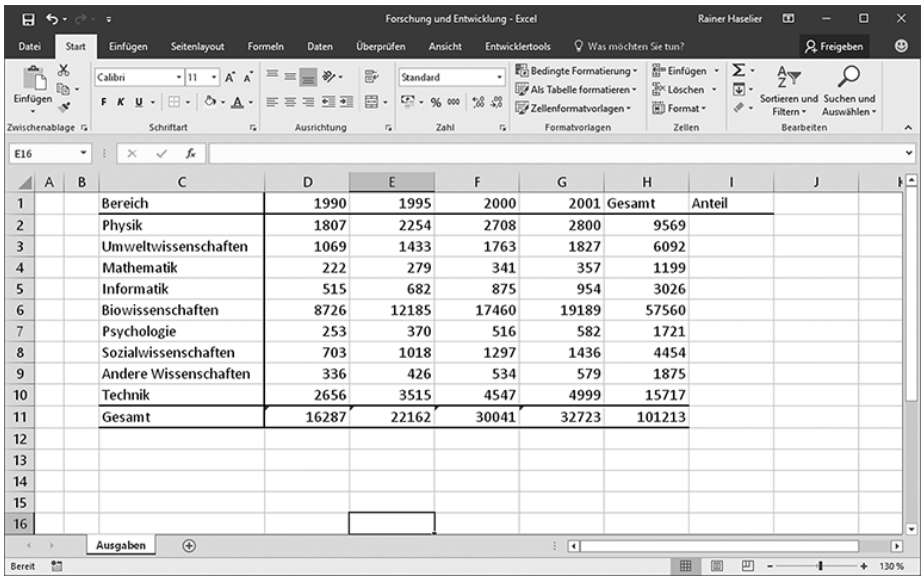
Ich gehe mal davon aus, dass Sie sich mit Excel schon ein bisschen auskennen. Dennoch sollten Sie sich etwas Zeit nehmen, damit wir einige grundlegende Dinge zu Excel klären können, die für den Rechenteil der statistischen Arbeit wichtig sind. Wenn Sie diese grundlegenden Dinge kennen, können Sie effizienter mit Excel-Formeln arbeiten.

Automatisches Ausfüllen von Zellen

Die erste wichtige Funktion ist die *AutoAusfüllen*-Funktion, mit deren Hilfe Excel eine Berechnung im ganzen Arbeitsblatt wiederholen kann. Wenn Sie eine Formel in eine Zelle einfügen, können Sie diese in angrenzende Zellen ziehen.

In Abbildung 1.6 sehen Sie ein Arbeitsblatt mit den Ausgaben für Forschung und Entwicklung im Bereich Wissenschaft und Technik an Hochschulen und Universitäten für die angegebenen Jahre. Die Daten, die aus dem Bericht der U.S. National Science Foundation

stammen, sind in Millionen US-Dollar angegeben. In Spalte H sehen Sie eine Spalte für den Gesamtbetrag der verschiedenen Bereiche und in Zeile 11 befinden sich die Gesamtbeträge für die einzelnen Jahre. (Auf die Bedeutung von Spalte I komme ich in Bälde zu sprechen.)



	A	B	C	D	E	F	G	H	I	J
1			Bereich	1990	1995	2000	2001	Gesamt	Anteil	
2			Physik	1807	2254	2708	2800	9569		
3			Umweltwissenschaften	1069	1433	1763	1827	6092		
4			Mathematik	222	279	341	357	1199		
5			Informatik	515	682	875	954	3026		
6			Biowissenschaften	8726	12185	17460	19189	57560		
7			Psychologie	253	370	516	582	1721		
8			Sozialwissenschaften	703	1018	1297	1436	4454		
9			Andere Wissenschaften	336	426	534	579	1875		
10			Technik	2656	3515	4547	4999	15717		
11			Gesamt	16287	22162	30041	32723	101213		
12										
13										
14										
15										
16										

Abbildung 1.6: Ausgaben für Forschung und Entwicklung im Bereich Wissenschaft und Technik

Nach dem Einfügen der Daten waren Spalte H und Zeile 11 noch ohne Werte. Wie habe ich die Gesamtbeträge in Spalte H und Zeile 11 bekommen?

Wenn Sie eine Formel zum Berechnen des Gesamtbetrags in der ersten Zeile (für Physik) erstellen möchten, besteht eine Möglichkeit (von vielen) darin,

```
=D2+E2+F2+G2
```

in Zelle H2 einzugeben. (Eine Formel beginnt immer mit »=«.) Wenn Sie die Enter-Taste drücken, wird in H2 der Gesamtbetrag angezeigt.

Um nun diese Formel in die Zellen H3 bis H10 einzufügen, zeigen Sie mit dem Mauszeiger auf die untere rechte Ecke von H2, bis aus dem Pfeil ein »+« wird. Halten Sie nun die linke Maustaste gedrückt und ziehen Sie die Maus über die Zellen. Dieses +-Zeichen wird als *Ausfüllkästchen* bezeichnet.

Wenn alle gewünschten Zellen ausgefüllt sind, lassen Sie die Maustaste los. Daraufhin werden in der Zeile die Gesamtbeträge angezeigt. Damit sparen Sie eine Menge Zeit, da Sie die Formel nicht acht Mal neu eingeben müssen.

Dasselbe gilt auch für die Gesamtbeträge der Spalten. Eine Möglichkeit, die Formel zum Addieren der Zahlen in der ersten Spalte (1990) zu erstellen, besteht darin,

```
=D2+D3+D4+D5+D6+D7+D8+D9+D10
```

in Zelle D11 einzugeben. Zeigen Sie mit dem Cursor auf das Ausfüllkästchen von D11, ziehen Sie das Ausfüllkästchen entlang der Zeile 11 und lassen Sie die Maustaste in Spalte H los. So werden die Gesamtbeträge automatisch in die Zellen E11 bis H11 gefüllt.

Ziehen ist nicht die einzige Möglichkeit. Eine andere Möglichkeit besteht darin, die gewünschten Zellen der Zeile oder Spalte (einschließlich der Zelle, die die Formel enthält) zu markieren und dann den Befehl START | FÜLLBEREICH zu wählen.

Wo ist der Befehl FÜLLBEREICH? In der Gruppe BEARBEITEN der Registerkarte START sehen Sie eine Schaltfläche mit einem Pfeil, der nach unten weist. Wenn Sie diese Schaltfläche anklicken, wird das zugehörige Menü geöffnet (siehe Abbildung 1.7). Wählen Sie dort den Befehl UNTEN aus; dieser Befehl hat die gleiche Wirkung wie das Ziehen des Ausfüllkästchens.

Abbildung 1.7: Das Menü der Schaltfläche FÜLLBEREICH

Eine weitere Vorgehensweise besteht darin, im Menü der Schaltfläche FÜLLBEREICH den Befehl REIHE auszuwählen. Hierdurch wird das gleichnamige Dialogfeld (siehe Abbildung 1.8) geöffnet. Schalten Sie in diesem Dialogfeld das Optionsfeld AUTOAUSFÜLLEN ein, klicken Sie auf OK und Sie sind fertig. Diese Vorgehensweise erfordert zwar einen zusätzlichen Schritt, jedoch ist sie ein wenig mehr mit früheren Excel-Versionen kompatibel.

Ich erkläre das alles, weil es in der statistischen Analyse häufig vorkommt, dass eine Formel in mehrere Zellen eingegeben werden muss. Die Formeln sind oft komplexer als die in diesem Abschnitt und Sie werden sie nicht zig Mal neu eingeben wollen. Daher lohnt es sich zu wissen, wie das automatische Ausfüllen funktioniert.

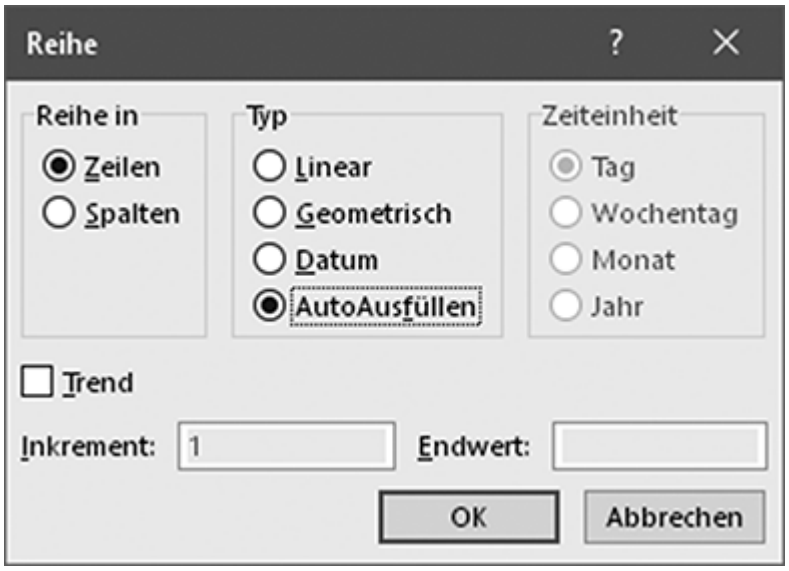


Abbildung 1.8: Das Dialogfeld REIHE

Ein schneller Weg zum AutoAusfüllen besteht darin, die erste Zelle der Reihe anzuklicken, den Mauszeiger auf die untere rechte Ecke der Zelle zu bewegen, bis das Ausfüllkästchen erscheint und dann doppelzuklicken. Dies funktioniert sowohl unter Windows als auch auf dem Mac.

Auf Zellen verweisen

Eine weitere wichtige Grundlage ist die Art und Weise, wie Excel Zellbezüge in Arbeitsblättern herstellt. Betrachten wir noch einmal das Arbeitsblatt in Abbildung 1.6. Jede automatisch ausgefüllte Formel unterscheidet sich etwas von der Anfangsformel. Die Formel in Zelle H2 lautet:

```
=D2+E2+F2+G2
```

Nach dem automatischen Ausfüllen lautet die Formel in H3:

```
=D3+E3+F3+G3
```

und die Formel in H4 lautet … nun, Sie können es sich bereits denken.

Das passt perfekt. Sie möchten den Gesamtbetrag in jeder Zeile berechnen, also passt Excel die Formel beim automatischen Ausfüllen der einzelnen Zellen entsprechend an. Diese Art Zellbezüge werden als *relative Zellbezüge* bezeichnet. Hier werden die Zellbezüge (die Zellbezeichnungen) entsprechend der Position im Arbeitsblatt angepasst. In diesem Beispiel bewirkt die Formel, dass die Zahlen in den Zellen in den vier Spalten links addiert werden.

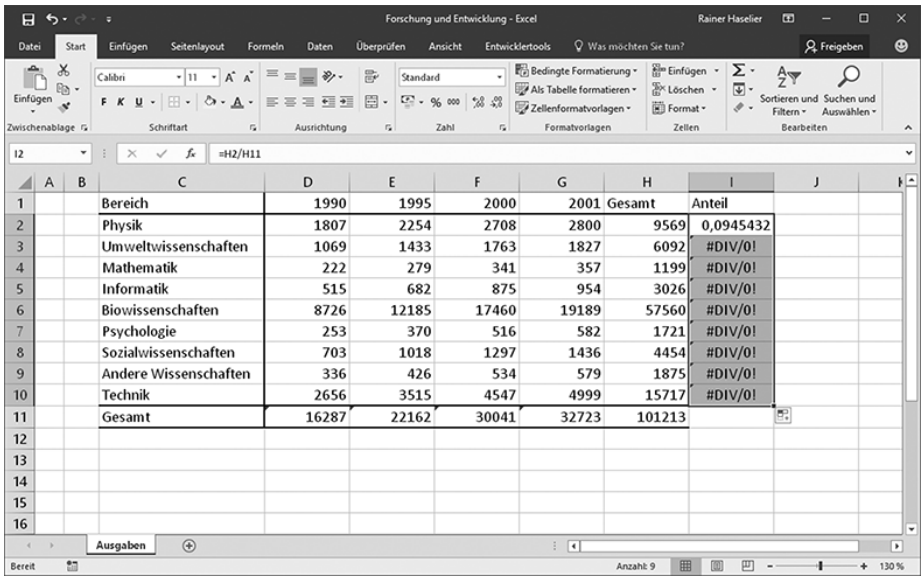
Es gibt noch eine andere Möglichkeit: Nehmen wir an, wir möchten wissen, wie groß der Anteil des Gesamtbetrags einer Zeile am Gesamtergebnis (der Zahl in H11) ist. Das müsste ganz einfach sein, nicht? Sie brauchen nur eine Formel für I2 zu erstellen und dann die Zellen I3 bis I10 automatisch auszufüllen.

Ähnlich wie beim ersten Beispiel geben Sie zunächst die folgende Formel in I2 ein:

```
=H2/H11
```

Wenn Sie die Enter-Taste drücken, wird in I2 der Anteil angezeigt.

Zeigen Sie mit dem Cursor auf das Ausfüllkästchen, ziehen Sie es über die Spalte I hinweg, lassen Sie die Maustaste in I10 los und … o je, o je! In Abbildung 1.9 ist das misslungene Ergebnis dargestellt. In den Zellen I3 bis I10 befindet sich dieses extrem hässliche #DIV/0! Was ist hier nur passiert?



	A	B	C	D	E	F	G	H	I	J
1			Bereich	1990	1995	2000	2001	Gesamt	Anteil	
2			Physik	1807	2254	2708	2800	9569	0,0945432	
3			Umweltwissenschaften	1069	1433	1763	1827	6092	#DIV/0!	
4			Mathematik	222	279	341	357	1199	#DIV/0!	
5			Informatik	515	682	875	954	3026	#DIV/0!	
6			Biowissenschaften	8726	12185	17460	19189	57560	#DIV/0!	
7			Psychologie	253	370	516	582	1721	#DIV/0!	
8			Sozialwissenschaften	703	1018	1297	1436	4454	#DIV/0!	
9			Andere Wissenschaften	336	426	534	579	1875	#DIV/0!	
10			Technik	2656	3515	4547	4999	15717	#DIV/0!	
11			Gesamt	16287	22162	30041	32723	101213		
12										
13										
14										
15										
16										

Abbildung 1.9: Ups! Fehler beim automatischen Ausfüllen!

Folgendes: Wenn Sie nichts anderes angeben, verwendet Excel beim automatischen Ausfüllen relative Zellbezüge. Somit lautet die in I3 eingegebene Formel nicht

```
=H3/H11
```

sondern

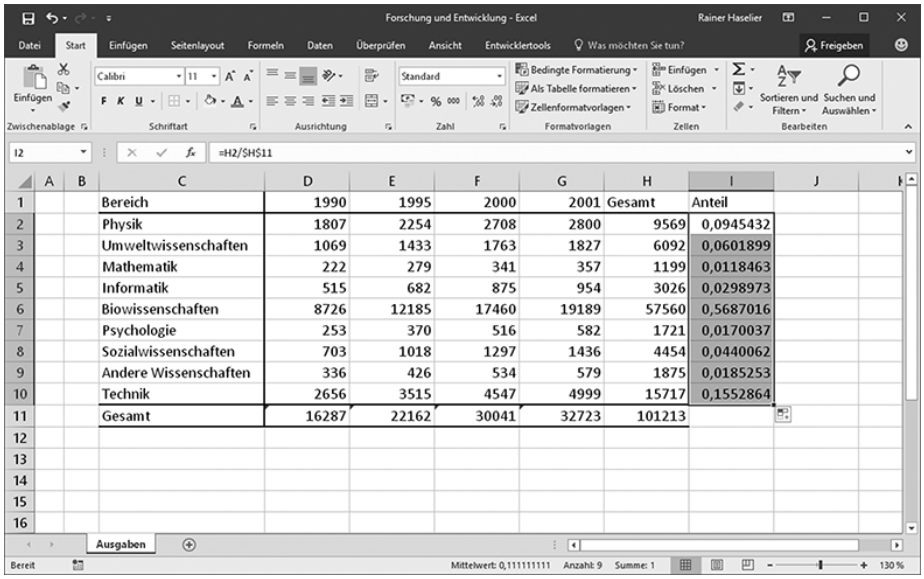
```
=H3/H12
```

Warum wird aus H11 H12? Bei relativen Zellbezügen wird davon ausgegangen, dass mit der Formel die Zahl in der Zelle durch die Zahl geteilt werden soll, die sich in derselben Spalte neun Zellen unterhalb befindet. Da H12 leer ist, verlangt die Formel eine Division durch null, was nicht geht.

Sie müssen Excel also mitteilen, dass alle Zahlen durch die Zahl in H11 geteilt werden sollen und nicht durch die Zahl, die sich neun Zellen weiter unten befindet. Dazu müssen Sie *absolute Zellbezüge* verwenden. Absolute Zellbezüge werden durch Einfügen von Dollarzeichen ($) in den Zellnamen gekennzeichnet. Die richtige Formel für I2 lautet

```
=H2/$H$11
```

So weiß Excel, dass beim automatischen Ausfüllen weder die Spalte noch die Zeile angepasst werden soll. In Abbildung 1.10 ist das Arbeitsblatt mit den jeweiligen Anteilen dargestellt. In der Abbildung ist Zelle I2 markiert. Beachten Sie die Formel in der Bearbeitungsleiste (dem Bereich oberhalb des Arbeitsblatts und unter dem Menüband).



	A	B	C	D	E	F	G	H	I	J
1			Bereich	1990	1995	2000	2001	Gesamt	Anteil	
2			Physik	1807	2254	2708	2800	9569	0,0945432	
3			Umweltwissenschaften	1069	1433	1763	1827	6092	0,0601899	
4			Mathematik	222	279	341	357	1199	0,0118463	
5			Informatik	515	682	875	954	3026	0,0298973	
6			Biowissenschaften	8726	12185	17460	19189	57560	0,5687016	
7			Psychologie	253	370	516	582	1721	0,0170037	
8			Sozialwissenschaften	703	1018	1297	1436	4454	0,0440062	
9			Andere Wissenschaften	336	426	534	579	1875	0,0185253	
10			Technik	2656	3515	4547	4999	15717	0,1552864	
11			Gesamt	16287	22162	30041	32723	101213		
12										
13										
14										
15										
16										

Abbildung 1.10: Automatisches Ausfüllen mit absoluten Zellbezügen

Um aus einem relativen einen absoluten Zellbezug zu machen, wählen Sie die gewünschte Zelladresse (oder Zelladressen) aus und drücken die Taste F4. Die F4-Taste funktioniert wie ein Schalter, mit dem zwischen relativen Zellbezügen (beispielsweise H11), absoluten Zellbezügen sowohl für die Zeile als auch für die Spalte in der Adresse ($H$11), absoluten Zellbezügen nur für den Zeilenteil (H$11) und absoluten Zellbezügen nur für den Spaltenteil ($H11) hin- und hergeschaltet werden kann.

Um in Excel 2016 für Mac zwischen relativen und absoluten Zellbezügen hin- und herzuschalten, halten Sie die fn-Taste

gedrückt, während Sie F4 drücken. Eine weitere Mac-Tastenkombination für diese Aktion ist MacBef+T.

Kapitel 2

Grundlegendes zu den statistischen Funktionen von Excel

In diesem Kapitel

- Mit Tabellenblattfunktionen arbeiten
- Schnell auf statistische Funktionen zugreifen
- Ein Array mit Ergebnissen erhalten
- Arrays mit Namen versehen
- Tools für die Analyse einsetzen
- Das Excel-Feature Schnellanalyse verwenden

In diesem Kapitel lernen Sie die statistischen Funktionen und die Datenanalysetools von Excel kennen. Wenn Sie bereits mit Excel gearbeitet haben, und davon gehe ich aus, dann wissen Sie, dass Excel eine ganze Fülle von Funktionen enthält, von denen einige statistische Funktionen sind. Sie können in jede Arbeitsblattzelle Daten eingeben, Excel anweisen, dass an Daten in einigen Zellen Berechnungen durchgeführt werden, oder Sie können die Daten mit den Tabellenblattfunktionen von Excel bearbeiten. Bei jeder Tabellenblattfunktion handelt es sich um eine integrierte Formel, mit der Sie sich den Aufwand sparen können, Excel anweisen zu müssen, eine Reihe von Berechnungen durchzuführen. Sowohl Einsteiger als auch Excel-Kenner wissen, dass Formeln ein zentraler Bestandteil von Excel sind. Die Datenanalysetools gehen über die

Formeln hinaus. Jedes Tool stellt eine Reihe informativer Ergebnisse bereit.

Erste Schritte

Viele der statistischen Funktionen von Excel werden in Form von Tabellenblattfunktionen zur Verfügung gestellt. In früheren Versionen (vor Excel 2003) war die Schaltfläche FUNKTION EINFÜGEN, die mit *fx* beschriftet ist, die einzige Möglichkeit, um auf die Tabellenblattfunktionen zuzugreifen. Ein Klick auf diese Schaltfläche öffnet das Dialogfeld FUNKTION EINFÜGEN, in der sich eine Liste der Excel-Funktionen befindet. Außerdem können Sie in diesem Dialogfeld nach Funktionen suchen. (Auf dem Mac öffnet diese Schaltfläche den Formel-Generator, bei dem es sich um einen anderen Namen für die gleiche Funktionalität handelt.) Obwohl Excel 2016 einfachere Möglichkeiten zur Verfügung steht, um auf die Tabellenblattfunktionen zuzugreifen, ist diese Schaltfläche weiterhin in der Bearbeitungsleiste vorhanden. Gleichzeitig stehen andere Verfahren zur Verfügung, um das Dialogfeld FUNKTION EINFÜGEN zu öffnen. Ich komme hierauf etwas weiter hinten ausführlich zurück.

Abbildung 2.1 zeigt die beiden Stellen, an denen sich die Schaltfläche FUNKTION EINFÜGEN und die Bearbeitungsleiste befinden. Die Bearbeitungsleiste befindet sich rechts neben einer der FUNKTION EINFÜGEN-Schaltflächen und dem Namenfeld. Alle drei befinden sich unter dem Menüband.

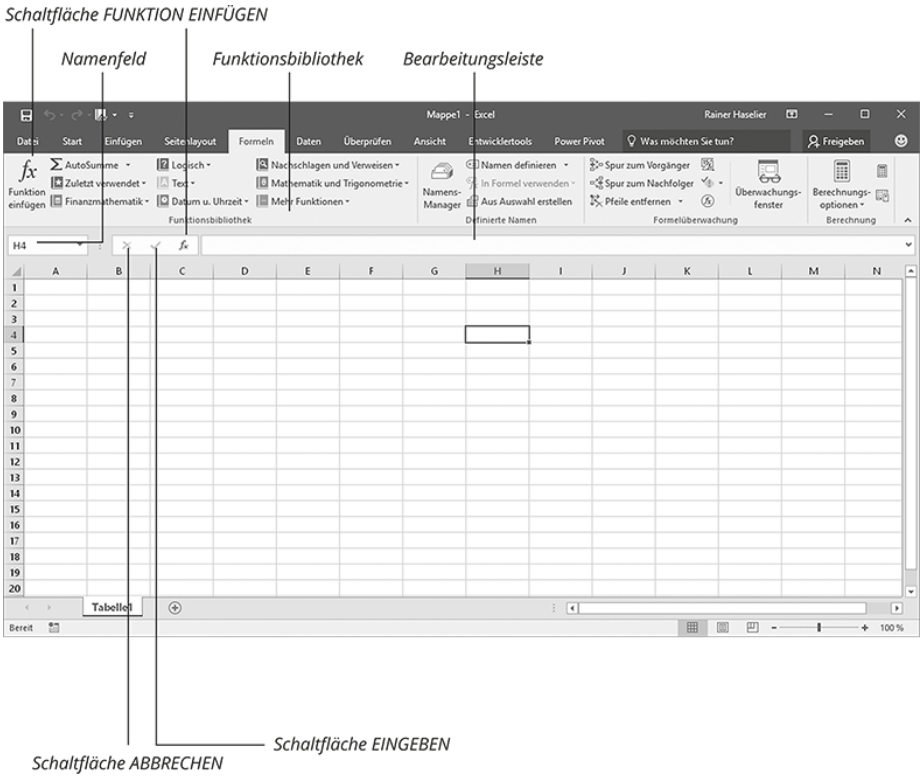


Abbildung 2.1: Wichtige Elemente für das Eingeben und Bearbeiten von Formeln

Neben dem Namenfeld und links von der Schaltfläche FUNKTION EINFÜGEN sehen Sie ein X und ein Häkchen. Beim X handelt es sich um die Schaltfläche ABBRECHEN und beim Häkchen um die Schaltfläche EINGEBEN. Das Anklicken von EINGEBEN entspricht dem Drücken der Enter-Taste auf der Tastatur: Excel wird hierdurch angewiesen, eine Berechnung durchzuführen, die Sie in eine Zelle eingegeben haben. Wenn Sie auf ABBRECHEN klicken, werden alle Änderungen, die Sie an der Zelle vorgenommen haben, verworfen, falls Sie noch nicht auf EINGEBEN geklickt haben.

Auf der Registerkarte FORMELN des Menübands finden Sie die Funktionsbibliothek. Mac-User finden in Excel 2016 für Mac ein ähnliches Layout vor.

Die Bearbeitungsleiste ist wie eine Kopie der Zelle, in der Sie sich befinden. Informationen, die in die Bearbeitungsleiste eingegeben werden, werden in die markierte Zelle eingegeben und umgekehrt.

In Abbildung 2.1 sehen Sie Excel mit geöffneter Registerkarte FORMELN. Dort sehen Sie die zweite Stelle, an der sich die Schaltfläche FUNKTION EINFÜGEN befindet. Die Schaltfläche ist ebenfalls mit *fx* beschriftet und befindet sich ganz links im Menüband in der Gruppe FUNKTIONSBIBLIOTHEK. Wie bereits weiter oben erwähnt, öffnet das Anklicken der Schaltfläche FUNKTION EINFÜGEN das gleichnamige Dialogfeld (siehe Abbildung 2.2).

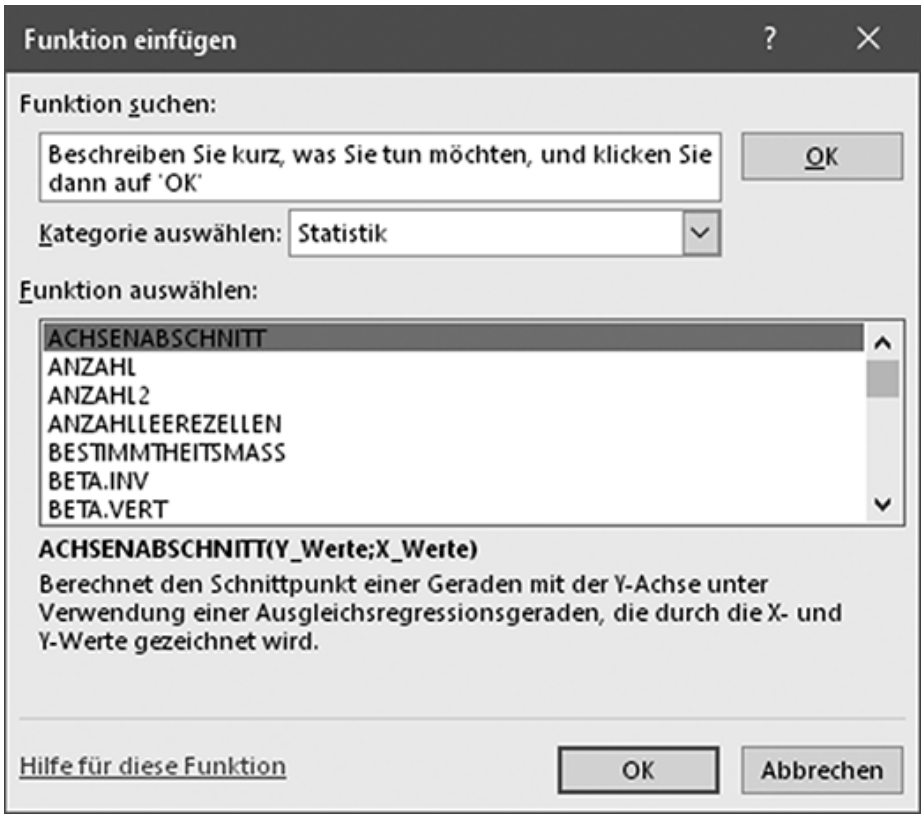


Abbildung 2.2: Das Dialogfeld FUNKTION EINFÜGEN

In diesem Dialogfeld können Sie die Funktion finden, die Sie benötigen, indem Sie entweder einen Suchbegriff eingeben oder durch die Liste der Excel-Funktionen blättern, nachdem Sie vorher im Listenfeld KATEGORIE AUSWÄHLEN die gewünschte Kategorie ausgewählt haben.

Neben dem Anklicken der Schaltfläche FUNKTION EINFÜGEN in der Bearbeitungsleiste können Sie das Dialogfeld FUNKTION EINFÜGEN öffnen, indem Sie den Befehl FORMELN | FUNKTION EINFÜGEN wählen.

Aufgrund der Anordnung der Befehle in den Excel-Versionen vor der Einführung des Menübands war damals das Dialogfeld FUNKTION EINFÜGEN extrem nützlich. In Excel 2016 hingegen ist es vor allem dann hilfreich, wenn Sie nicht genau wissen, welche Funktion Sie brauchen, oder wenn Sie nicht wissen, wo Sie die benötigte Funktion finden.

In der Gruppe FUNKTIONSBIBLIOTHEK sind die Funktionen nach Kategorien gruppiert; hierdurch ist es einfach, auf die benötigte Funktion zuzugreifen. Wenn Sie in der Gruppe FUNKTIONSBIBLIOTHEK eine der Kategorieschaltflächen anklicken, zeigt Excel ein Menü der Funktionen an, die zu dieser Kategorie gehören.

In den meisten Fällen arbeite ich mit den statistischen Funktionen, die über den Befehl MEHR FUNKTIONEN | STATISTIK erreichbar sind. Manchmal benötige ich auch mathematische Funktionen, die im Menü der Schaltfläche MATHEMATIK UND TRIGONOMETRIE zur Verfügung stehen. (Ein paar dieser Funktionen lernen Sie im Verlauf dieses Kapitels kennen.) In Kapitel 5 stelle ich den Einsatz von einigen logischen Funktionen vor.

Der letzte Befehl in allen Menüs, die nach dem Anklicken einer der Kategorien geöffnet werden, ist wiederum FUNKTION EINFÜGEN. Dieser Befehl ist damit eine weitere Option, die Ihnen zum Öffnen des Dialogfeldes FUNKTION EINFÜGEN zur Verfügung steht. (In der Mac-Version wird dieses Dialogfeld FORMEL-GENERATOR genannt.)

Das Namenfeld ist so etwas wie ein aktueller Bericht über Ihre Aktionen im Arbeitsblatt. Wenn Sie eine Zelle markieren, wird die Adresse dieser Zelle im Namenfeld angezeigt. Wenn Sie auf die Schaltfläche FUNKTION EINFÜGEN klicken, wird im Namenfeld der Name der Funktion angezeigt, die Sie in letzter Zeit am häufigsten verwendet haben.

Neben den statistischen Funktionen stellt Excel eine Reihe von Datenanalysetools bereit, die Sie über die Gruppe ANALYSE auf der Registerkarte DATEN aufrufen können.

Machen Sie sich bereit für die Statistik

In diesem Abschnitt erfahren Sie, wie Sie die Tabellenblattfunktionen und Datenanalysetools nutzen können.

Tabellenblattfunktionen in Excel 2016

Wie ich bereits weiter vorne erwähnt habe, zeigt die Gruppe FUNKTIONSBIBLIOTHEK der Registerkarte FORMELN alle Kategorien der Tabellenblattfunktionen an.

Im Allgemeinen werden die Tabellenblattfunktionen wie folgt verwendet:

1. **Geben Sie die Daten in ein Datenfeld ein, und markieren Sie eine Zelle für das Ergebnis.**

2. **Wählen Sie FORMELN | FUNKTIONSBIBLIOTHEK, klicken Sie die Schaltfläche mit der gewünschten Kategorie an und wählen Sie aus dem Menü die Funktion aus, die Sie verwenden wollen.**

   Hierdurch wird das Dialogfeld FUNKTIONSARGUMENTE geöffnet.

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente der Funktion ein.**

   Argument ist ein Ausdruck aus der Mathematik. Er hat nichts mit Debatten, Streit oder Auseinandersetzungen zu tun. In der Mathematik ist ein Argument ein Wert, der von einer Funktion bearbeitet wird.

4. **Klicken Sie auf OK, damit das Ergebnis in der markierten Zelle angezeigt wird.**

Um Ihnen ein Beispiel zu geben, beschreibe ich eine Funktion, mit der die Funktionsweise der Tabellenblattfunktionen von Excel deutlich wird. Diese Funktion (SUMME) addiert die Zahlen in Zellen, die Sie angeben, und zeigt die Summe in einer anderen Zelle an, die Sie ebenfalls angeben. Obwohl die Summenbildung ein integraler Bestandteil bei der Bearbeitung von Zahlen in der Statistik darstellt, befindet sich SUMME nicht in der Kategorie STATISTIK. Es handelt sich hierbei um eine typische Tabellenblattfunktion, die eine Operation durchführt, die Ihnen bekannt ist.

Und so gehen Sie vor, um die Funktion SUMME zu verwenden:

1. **Geben Sie Ihre Zahlen in einen Zellbereich ein, und markieren Sie eine Zelle für das Ergebnis.**

   In diesem Beispiel habe ich die Zahlen **45**, **33**, **18**, **37**, **32**, **46** und **39** in die Zellen C2 bis C8 eingegeben und C9 als Zelle für das Ergebnis festgelegt.

2. **Klicken Sie die Schaltfläche mit der gewünschten Kategorie an und wählen Sie im Menü der Schaltfläche die Funktion aus.**

   Hierdurch wird das Dialogfeld FUNKTIONSARGUMENTE geöffnet.

   Ich habe den Befehl FORMELN | MATHEMATIK UND TRIGONOMETRIE gewählt und nach unten gescrollt, bis die Funktion SUMME sichtbar war.

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente der Funktion ein.**

   Excel nimmt an, dass Sie die Zahlen in den Zellen C2 bis C8 addieren möchten, und zeigt diesen Zellbereich im Feld ZAHL1 an. Im Dialogfeld wird die Summe der Zahlen dieses

Zellbereichs angezeigt. In diesem Beispiel beträgt die Summe 250 (siehe Abbildung 2.3).

4. **Klicken Sie auf OK, damit die Summe in die markierte Zelle eingefügt wird.**

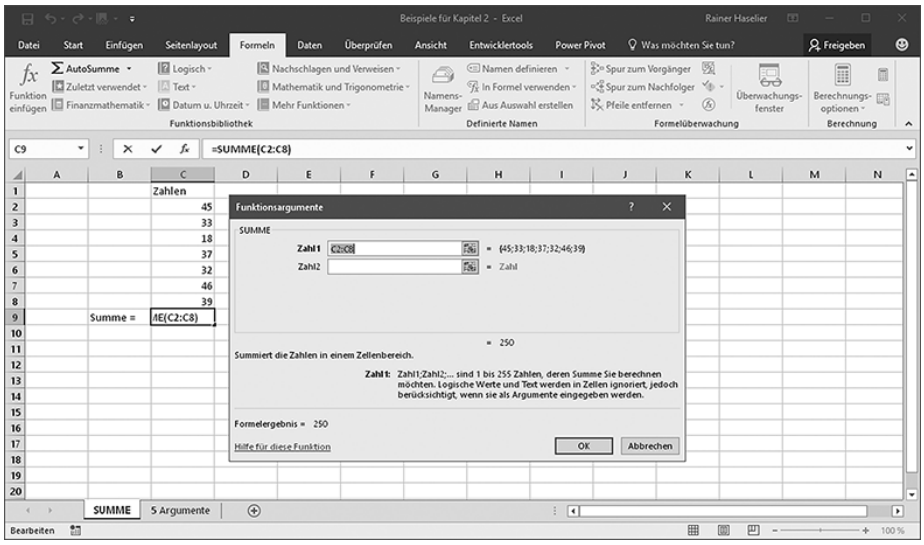


Abbildung 2.3: Die Funktion *SUMME* verwenden

Hier sind einige Dinge zu beachten. Wie in Abbildung 2.3 zu sehen ist, steht in der Bearbeitungsleiste

```
=SUMME(C2:C8)
```

Diese Formel gibt an, dass der Wert in der markierten Zelle der Summe aus den Zahlen in den Zellen C2 bis C8 entspricht.

Wenn Sie sich erst einmal ein bisschen besser mit einer Tabellenblattfunktion und deren Argumenten auskennen, können Sie das Menü überschlagen und den Namen der Funktion direkt in die Zelle oder in die Bearbeitungsleiste eingeben. Geben Sie zuerst ein Gleichheitszeichen ein. Wenn Sie die ersten Buchstaben des Funktionsnamens eingeben, öffnet Excel ein Menü, in dem die Funktionen angezeigt werden, deren Namen mit den eingetippten

Buchstaben beginnen (siehe Abbildung 2.4). Um eine der Funktionen auszuwählen, doppelklicken Sie auf deren Namen.

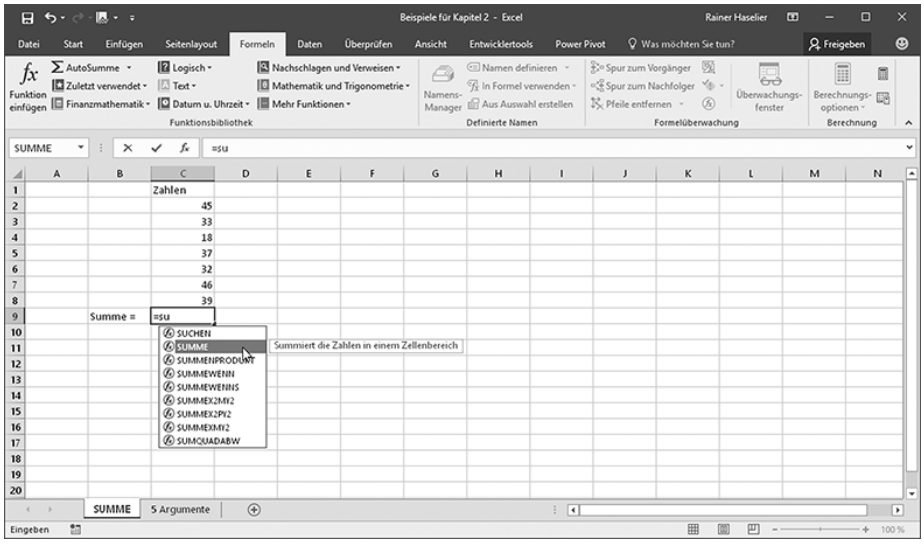


Abbildung 2.4: Während Sie eine Formel eingeben, zeigt Excel ein Menü an, in dem Sie die gewünschte Funktion auswählen können.

Beachten Sie außerdem die Felder im Dialogfeld FUNKTIONSARGUMENTE in Abbildung 2.3. In der Abbildung sind lediglich zwei Felder (ZAHL1 und ZAHL2) zu sehen. Der Zellbereich Datenfeld wird im Feld ZAHL1 angezeigt. Wozu dient dann aber das Feld ZAHL2?

Mit dem Feld ZAHL2 können Sie zur Summe ein weiteres Argument hinzufügen. Und damit noch nicht genug. Wenn Sie in das Feld ZAHL2 klicken, wird das Feld ZAHL3 angezeigt. Wenn Sie in das Feld ZAHL3 klicken, wird das Feld ZAHL4 angezeigt usw. Maximal werden 30 Felder angezeigt, wobei jedes Feld einem Argument entspricht. Ein Wert kann ein Zellbereich an einer beliebigen Stelle im Arbeitsblatt, eine Zahl, ein arithmetischer Ausdruck, der eine Zahl ergibt, eine Zelladresse oder ein Name sein, den Sie einer Zelle oder einem Zellbereich zugewiesen haben. Beim Eingeben von Werten wird im Dialogfeld FUNKTIONSARGUMENTE die aktualisierte Summe angezeigt. Wenn Sie auf OK klicken, wird die Summe in der markierten Zelle angezeigt.

Nicht für jede Tabellenblattfunktion können mehrere Argumente angegeben werden. Manche Tabellenblattfunktionen haben nur ein Argument. Bei denjenigen, die mehrere Argumente haben können, können Sie Daten angeben, die sich an beliebigen Stellen im Arbeitsblatt befinden. In Abbildung 2.5 ist ein Arbeitsblatt mit dem Dialogfeld FUNKTIONSARGUMENTE mit Daten aus zwei Zellbereichen, zwei arithmetischen Ausdrücken und einer Zelle dargestellt. Beachten Sie das Format der Formel in der Bearbeitungsleiste. Die Argumente sind durch Strichpunkte voneinander getrennt.

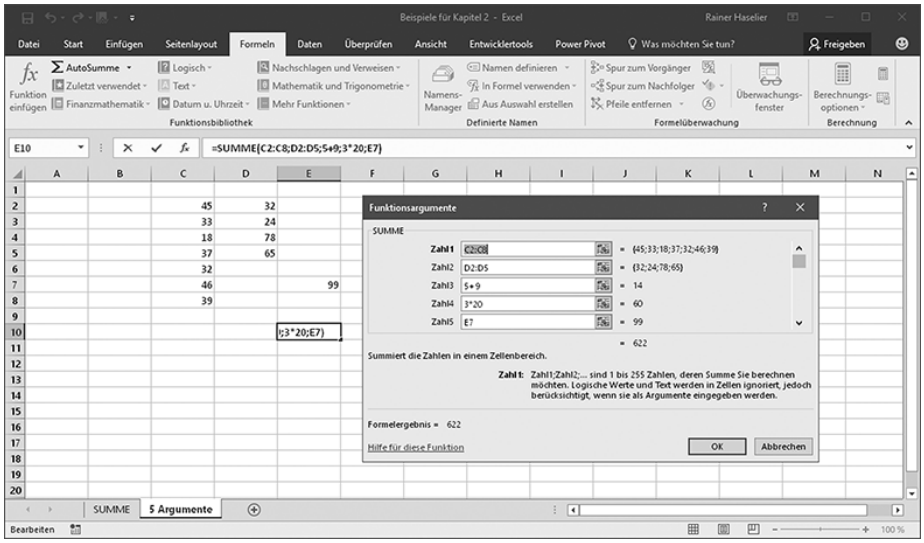


Abbildung 2.5: Die Funktion *SUMME* mit fünf Argumenten

Wenn Sie eine Zelle markieren, die in derselben Spalte, in der sich die Daten befinden, direkt unter der letzten Datenzelle steht, schließt Excel auf den Zellbereich, mit dem Sie arbeiten möchten. Excel errät jedoch nicht immer, was Sie wollen. Und manchmal liegt Excel einfach falsch. Wenn Excel entweder nicht errät, was Sie tun möchten, oder falsch rät, müssen Sie im Dialogfeld FUNKTIONSARGUMENTE die richtigen Werte eingeben.

Schnell auf statistische Funktionen zugreifen

In diesem Abschnitt zeige ich Ihnen, wie Sie schnell auf die statistischen Funktionen von Excel zugreifen können.

Sie wählen eine der statistischen Funktionen, in dem Sie FORMELN | MEHR FUNKTIONEN | STATISTIK wählen und im dann angezeigten Menü die Funktion anklicken, die Sie verwenden möchten (siehe Abbildung 2.6).

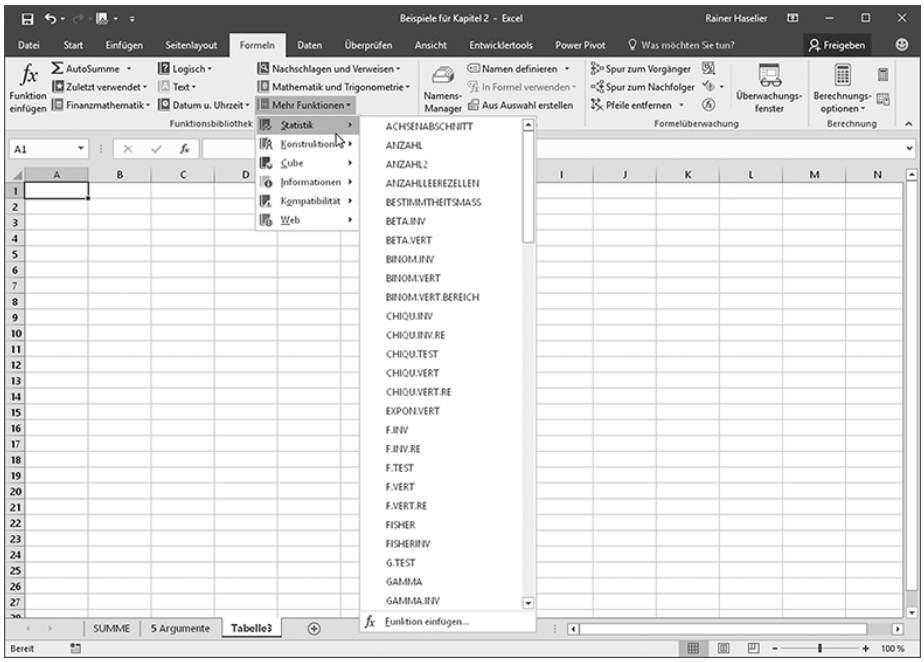


Abbildung 2.6: Auf die statistischen Funktionen von Excel zugreifen

Obwohl Excel die statistischen Funktionen ein wenig verborgen hat, können Sie eine nützliche Technik verwenden, um diese genauso leicht zugänglich zu machen wie die andere Kategorien: Fügen Sie die statistischen Funktionen in die Symbolleiste für den Schnellzugriff ein, die sich oben links im Excel-Fenster befindet. (Jede Office-Anwendung besitzt eine Symbolleiste für den Schnellzugriff, es sei denn, Sie sind ein Mac-Anwender. Dort hat keine Office-Anwendung diese Symbolleiste. Daher stehen die folgenden Schritte auf dem Mac nicht zur Verfügung. Es tut mir leid.)

Um die statistischen Funktionen in die Symbolleiste für den Schnellzugriff einzufügen, wählen Sie FORMELN | MEHR FUNKTIONEN und klicken dann den Eintrag STATISTIK mit der rechten Maustaste an. Wählen Sie im Menü den ersten Befehl aus, ZU SYMBOLLEISTE FÜR DEN SCHNELLZUGRIFF HINZUFÜGEN (siehe Abbildung 2.7). Hierdurch wird in die Symbolleiste für den Schnellzugriff eine weitere Schaltfläche eingefügt. Wenn Sie diese anklicken, wird das Menü mit den statistischen Funktionen geöffnet (siehe Abbildung 2.8).

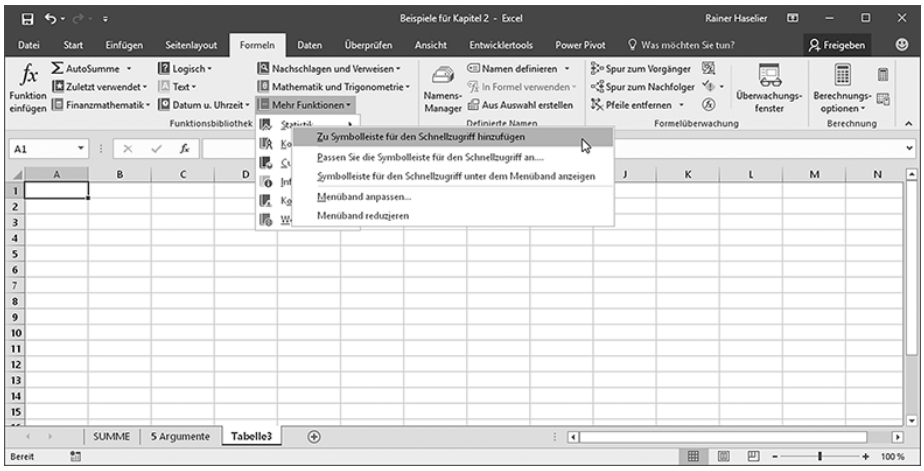


Abbildung 2.7: Die statistischen Funktionen in die Symbolleiste für den Schnellzugriff einfügen

Wenn ich im weiteren Verlauf des Buches die Verwendung der statistischen Funktionen beschreibe, gehe ich davon aus, dass Sie das Menü mit diesen Funktionen in die Symbolleiste für den Schnellzugriff eingefügt haben und Sie die benötigte Funktion schnell auswählen können. Im folgenden Abschnitt finden Sie ein Beispiel.

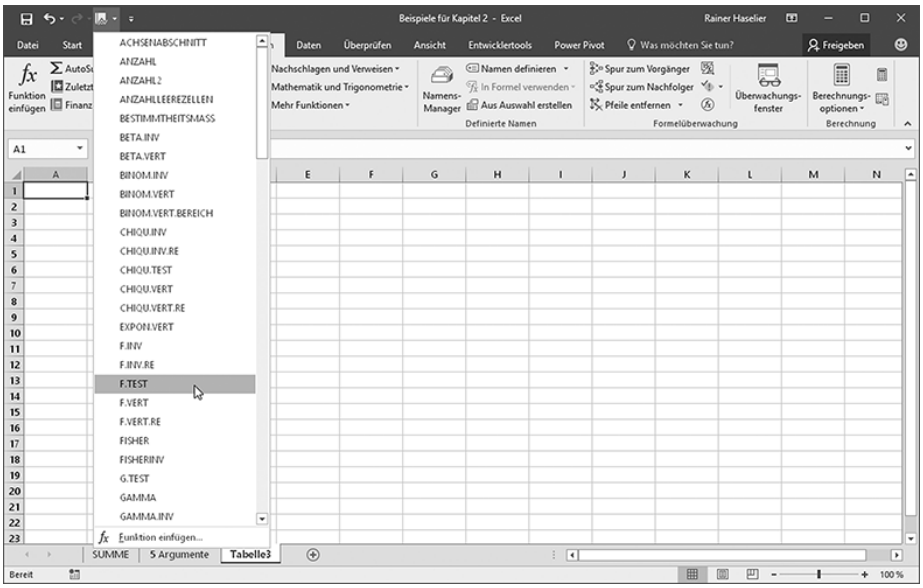


Abbildung 2.8: Über die Symbolleiste für den Schnellzugriff auf das Menü mit den statistischen Funktionen zugreifen

Arrayfunktionen

Bei den meisten in Excel integrierten Funktionen handelt es sich um Formeln, mit denen ein Einzelwert (wie etwa eine Summe) berechnet und der Wert in einer Zelle im Arbeitsblatt gespeichert wird. Bei Excel gibt es einen weiteren Funktionstyp. Dieser wird *Arrayfunktion* oder *Matrixfunktion* genannt. Damit werden mehrere Werte berechnet und nicht in einer einzelnen Zelle, sondern in einem Zellbereich gespeichert.

Ein gutes Beispiel für eine Arrayfunktion ist `HÄUFIGKEIT`, bei der es sich um eine statistische Funktion handelt. Mit dieser Funktion werden mehrere Werte so zusammengefasst, dass deutlich wird, wie sie sich auf Intervalle verteilen, die Sie festlegen. Nehmen wir beispielsweise die folgenden Werte:

77, 45, 44, 61, 52, 53, 68, 55

und die folgenden Intervalle:

50, 60, 70, 80

Mit der Funktion `HÄUFIGKEIT` können Sie ermitteln, wie viele Werte kleiner als oder gleich 50 (zwei in diesem Beispiel), wie viele größer

als 50 und wie viele kleiner als oder gleich 60 sind (drei in diesem Beispiel) und so weiter. Die Anzahl der Werte in jedem Intervall wird als *Häufigkeit* bezeichnet. Eine Tabelle mit den Intervallen und Häufigkeiten wird *Häufigkeitsverteilung* genannt.

Im Folgenden ein Beispiel für die Verwendung der Funktion `HÄUFIGKEIT`:

1. **Geben Sie die Werte in einen Zellbereich ein.**

   In Abbildung 2.9 befinden sich die Werte in den Zellen B2 bis B16.

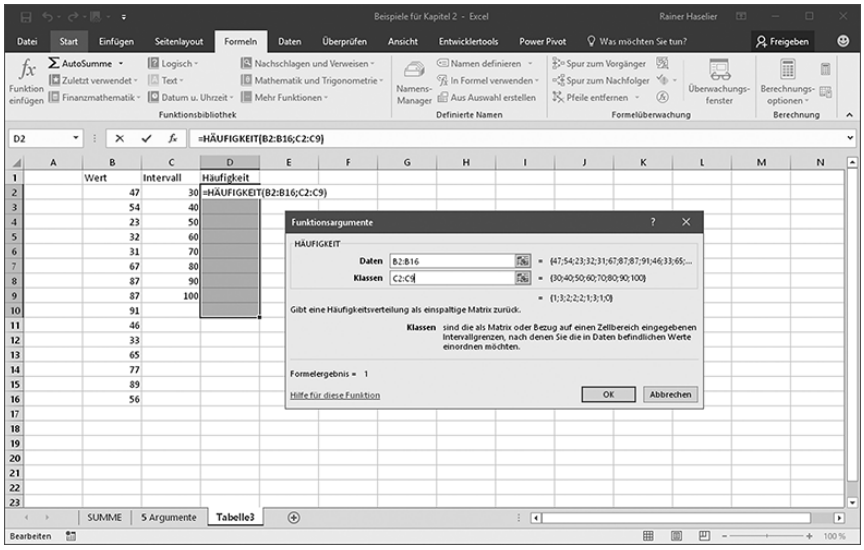


Abbildung 2.9: Die Funktion *HÄUFIGKEIT* verwenden

2. **Geben Sie die Intervalle in einen Zellbereich ein.**

   Ich habe die Intervalle in die Zellen C2 bis C9 eingegeben.

3. **Wählen Sie einen Zellbereich für die Häufigkeiten aus.**

   Ich habe Spalte D mit *Häufigkeit* beschriftet und die Zellen D2 bis D10 für die sich ergebenden Häufigkeiten ausgewählt. Wozu die zusätzliche Zelle? Die Funktion `HÄUFIGKEIT` gibt eine Spalte mit Daten zurück, die eine Zelle mehr enthält als der Zellbereich mit den Häufigkeiten.

4. **Wählen Sie im Menü mit den statistischen Funktionen HÄUFIGKEIT aus, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

   Ich habe die Schaltfläche verwendet, die ich der Symbolleiste für den Schnellzugriff hinzugefügt habe, und dort HÄUFIGKEIT gewählt.

5. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente der Funktion ein.**

   Ich verwende zuerst das Feld DATEN und gebe dort die Zellen ein, in denen sich die Werte befinden. In diesem Beispiel ist das B2:B16. Ich gehe davon aus, dass Sie Excel gut genug kennen, um zu wissen, dass es dafür mehrere Möglichkeiten gibt.

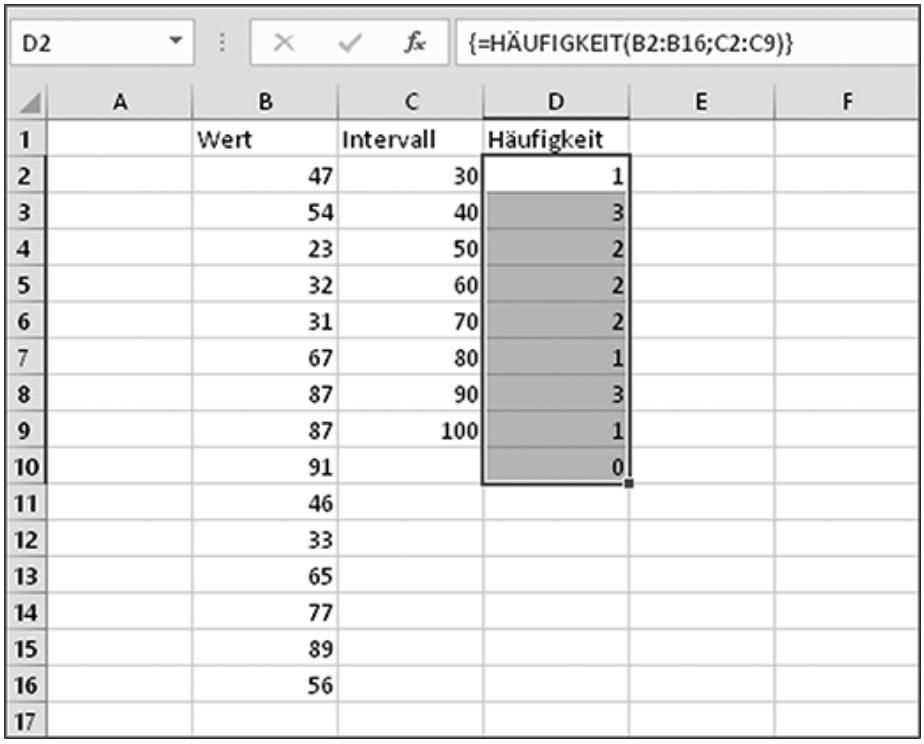
   Anschließend gebe ich den Zellbereich mit den Intervallen an. Bei der Funktion `HÄUFIGKEIT` werden Intervalle als Klassen bezeichnet und daher werden die Intervalle in das Feld KLASSEN eingegeben. In diesem Beispiel steht im Feld KLASSEN C2:C9. Sobald ich beide Felder angegeben habe, werden im Dialogfeld FUNKTIONSARGUMENTE die Häufigkeiten in geschweiften Klammern {} angezeigt.

6. **Drücken Sie die Tastenkombination Strg+Shift+Enter, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen und um die Werte in das markierte Array/den ausgewählten Zellbereich einzufügen. Auf dem Mac drücken Sie Ctrl+Shift+Enter oder MacBef+Shift+Enter.**

Das ist *sehr* wichtig. Da das Dialogfeld eine OK-Schaltfläche (und auf dem Mac die Schaltfläche FERTIG) besitzt, ist man versucht, auf OK zu klicken, um so die Ergebnisse in das Arbeitsblatt einzufügen. Jedoch führt das Anklicken von OK

bei Arrayformeln nicht zum gewünschten Effekt. Verwenden Sie immer die Tastenkombination Strg+Shift+Enter (beziehungsweise Ctrl+Shift+Enter oder MacBef+Shift+Enter auf dem Mac, siehe auch den nachfolgenden Tipp), um bei einer Arrayfunktion das Dialogfeld FUNKTIONSARGUMENTE zu schließen.

Nachdem Sie das Dialogfeld geschlossen haben, werden die Häufigkeiten in den gewünschten Zellen angezeigt (siehe Abbildung 2.10).



	A	B	C	D	E	F
1		Wert	Intervall	Häufigkeit		
2		47	30	1		
3		54	40	3		
4		23	50	2		
5		32	60	2		
6		31	70	2		
7		67	80	1		
8		87	90	3		
9		87	100	1		
10		91		0		
11		46				
12		33				
13		65				
14		77				
15		89				
16		56				
17						

Abbildung 2.10: Das Arbeitsblatt mit den Häufigkeiten

Beachten Sie die Formel in der Bearbeitungsleiste:

```
{= HÄUFIGKEIT(B2:B16,C2:C9)}
```

Excel kennzeichnet Feldfunktionen mit geschweiften Klammern.

Eigentlich bin ich nicht jemand, der Dinge ständig wiederholt. Aber hier muss ich eine Ausnahme machen. In Schritt 6 habe ich geschrieben, dass Sie bei Arrayfunktionen immer die Tastenkombination Strg+Shift+Enter (Ctrl+Shift+Enter oder MacBef+Shift+Enter auf dem Mac) drücken sollen. Merken Sie sich das gut, denn im Dialogfeld FUNKTIONSARGUMENTE gibt es keinerlei Hinweis darauf. Wenn Sie nach dem Eingeben der Argumente in eine Arrayfunktion einfach auf OK klicken, werden Sie sehr enttäuscht sein. Glauben Sie mir.

Eine Merkwürdigkeit auf dem Mac: Bevor ich die Tastenkombination Ctrl+Shift+Enter oder MacBef+Shift+Enter eingebe, muss ich die Bearbeitungsleiste anklicken. Anderenfalls werden die Werte nicht in das markierte Array eingefügt.

Hier ist noch ein einfacherer Weg, um all dies auf dem Mac zu erledigen: Anstelle Ctrl+Shift+Enter oder MacBef+Shift+Enter einzugeben, halten Sie die Tasten Ctrl+Shift gedrückt und klicken Sie dann auf die Schaltfläche EINGEBEN (die Schaltfläche mit dem Häkchen in der Bearbeitungsleiste). Die Aktion ist also: Ctrl+Shift +EINGEBEN anklicken. Wenn Sie die Taste MacBef gedrückt halten und dann auf EINGEBEN klicken, kommen Sie ebenfalls zum Ziel.

Was bedeutet schon ein Name? Ein Array voller Möglichkeiten

Je tiefer Sie in die statistischen Features von Excel eintauchen, desto öfter werden Sie mit Formeln arbeiten, die mehrere Argumente aufweisen. Diese Argumente verweisen, wie in den obigen Beispielen gezeigt, oft auf Zellbereiche.

Wenn Sie diese Arrays beziehungsweise Zellbereiche mit sinnvollen Namen versehen, behalten Sie besser den Überblick darüber, was genau Sie tun. Wenn Sie nach längerer Zeit wieder mal ein Arbeitsblatt öffnen, um es weiter zu bearbeiten, kommen Sie mit sinnvollen Arraynamen schneller wieder ins Thema. Noch ein Vorteil: Wenn Sie Ihr Arbeitsblatt und die enthaltenen Formeln anderen erklären wollen, sind sinnvolle Arraynamen äußerst hilfreich.

In Excel ist es ganz einfach, Zellbereiche mit Namen zu versehen. Spalte C in Abbildung 2.11 ist mit *Umsatz_Millionen* beschriftet, was für »Umsatz in Millionen Euro« steht. Mit dieser Überschrift wird die Bedeutung dieser Spalte (und damit auch der gesamten Tabelle) besser deutlich. Wenn ich Excel anweise, *Umsatz_Millionen* als Namen für den Zellbereich C2 bis C13 zu verwenden, kann ich diesen Namen überall dort verwenden, wo ich auf diesen Zellbereich Bezug nehme.

Warum habe ich *Umsatz_Millionen* verwendet und nicht *Umsatz in Millionen* oder *Umsatz (Millionen)* oder *Umsatz: Millionen*? Ganz einfach: Excel erlaubt in diesen Namen keine Leerzeichen und keine Symbole oder andere Sonderzeichen. Wenn Sie einem Zellbereich einen Namen zuweisen, müssen Sie die folgenden vier Regeln beachten. Der Name

- muss mit einem Buchstaben beginnen, es darf also keine Zahl und kein Satzzeichen verwendet werden,
- muss im Arbeitsblatt eindeutig sein,
- darf (wie bereits erwähnt) keine Leerzeichen oder Symbole enthalten; verwenden Sie einen Unterstrich, um im Namen

einzelne Wörter voneinander zu trennen,

- darf nicht mit einem Zellbezug identisch sein.

	A	B	C	D
1	Jahr	Region	Umsatz_Millionen	
2	2006	Nord	20	
3	2006	Süd	22	
4	2006	Ost	19	
5	2006	West	25	
6	2007	Nord	26	
7	2007	Süd	28	
8	2007	Ost	21	
9	2007	West	27	
10	2008	Nord	32	
11	2008	Süd	29	
12	2008	Ost	25	
13	2008	West	31	
14				

Abbildung 2.11: Namen für Zellbereiche

Gehen Sie folgendermaßen vor, um einen Namen zu definieren:

1. **Fügen Sie oben in der Spalte (oder links in einer Zeile), die Sie benennen wollen, einen beschreibenden Namen ein.**

   Ein Beispiel hierfür sehen Sie in Abbildung 2.11.

2. **Markieren Sie den Zellbereich, dem Sie einen Namen zuweisen wollen.**

   In unserem Beispiel ist dies der Zellbereich C2:C13. Warum darf C1 nicht eingeschlossen werden? Das erkläre ich umgehend.

3. **Klicken Sie den markierten Zellbereich mit der rechten Maustaste an.**

   Hierdurch wird das Kontextmenü geöffnet, das Sie in Abbildung 2.12 sehen.

Abbildung 2.12: Wenn Sie eine markierte Zelle mit der rechten Maustaste anklicken, wird dieses Kontextmenü geöffnet.

4. **Klicken Sie in diesem Kontextmenü auf NAMEN DEFINIEREN.**

   Dieser Befehl öffnet das Dialogfeld NEUER NAME, das Sie in Abbildung 2.13 sehen. Wie Sie erkennen können, nimmt Excel an, dass als Name für den Zellbereich C2:C13 *Umsatz_Millionen* verwendet werden soll. Wenn Sie Excel einen Zellbereich anbieten, dem Sie einen Namen zuweisen wollen, sucht Excel nach einem brauchbaren Namen in der Nähe der Auswahl, und zwar oberhalb der Markierung, wenn Zellen in einer Spalte markiert sind, oder links von der Markierung, wenn

Zellen in einer Zeile markiert sind. Falls kein Name vorhanden ist, müssen Sie im Dialogfeld NEUER NAME selbst einen eingeben. (Sie können das Dialogfeld NEUER NAME auch mit dem Befehl FORMELN | NAMEN DEFINIEREN aufrufen.)

Abbildung 2.13: Das Dialogfeld NEUER NAME

Wenn Sie einen Zellbereich, wie eine Spalte, markieren, bei dem sich oben eine Beschriftung befindet, können Sie diese Beschriftung ebenfalls markieren; Excel weist dann den Namen dem gesamten Zellbereich zu. *Ich rate jedoch dringend davon ab, dies zu tun. Warum?* Wenn ich die Zellen von C1 bis C13 markiere, verweist der Name *Umsatz_Millionen* auf die Zellen C1 bis C13 und nicht nur auf C2 bis C13. In diesem Fall handelt es sich beim ersten Wert des Bereichs um Text und nicht um Zahlen.

Für Funktionen wie `SUMME` (oder `SUMMEWENN` beziehungsweise `SUMMEWENNS`, die ich anschließend vorstelle) macht dies keinen Unterschied. In diesen Formeln ignoriert Excel Werte, die keine Zahlen sind. Wenn Sie

jedoch den gesamten Zellbereich in eine Berechnung einschließen, macht es wohl einen Unterschied, dass der definierte Name sowohl Text als auch Zahlen umfasst. Excel geht dann davon aus, das der Text Teil des Zellbereichs ist und versucht, diesen bei der Berechnung zu verwenden. Mehr hierzu finden Sie im folgenden Abschnitt, in dem es um das Erstellen eigener Arrayformeln geht.

1. **Klicken Sie auf OK.**

   Excel weist dem Zellbereich den Namen zu.

Ich kann nun diesen Namen in einer Formel verwende. Das Anklicken einer Zelle (wie C14) und das Eingeben der Formel `SUMME` direkt in die Zelle öffnet die Kästchen, die Sie in Abbildung 2.14 sehen.

Diese Kästchen erscheinen, während Sie die Formel eintippen. Wenn Sie UMSATZ_MILLIONEN auswählen und dann die Tab-Taste drücken, wird die Formel so erstellt, dass Excel sie versteht. Sie müssen dann noch die schließende runde Klammer ergänzen (siehe Abbildung 2.15) und Enter drücken, damit Sie das Ergebnis sehen.

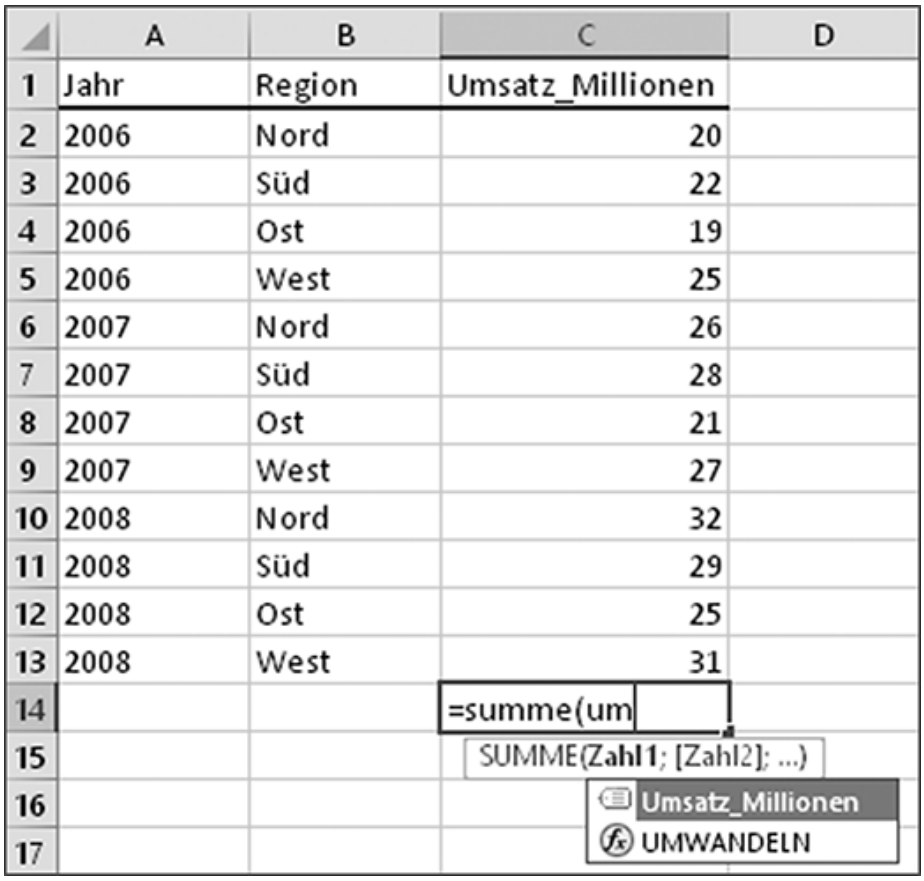


Abbildung 2.14: Wenn Sie eine Formel direkt in eine Zelle eingeben, werden diese Kästchen geöffnet.

Diese Formel

```
=SUMME(Umsatz_Millionen)
```

ist dank des Namens für den Zellbereich viel aussagekräftiger als diese:

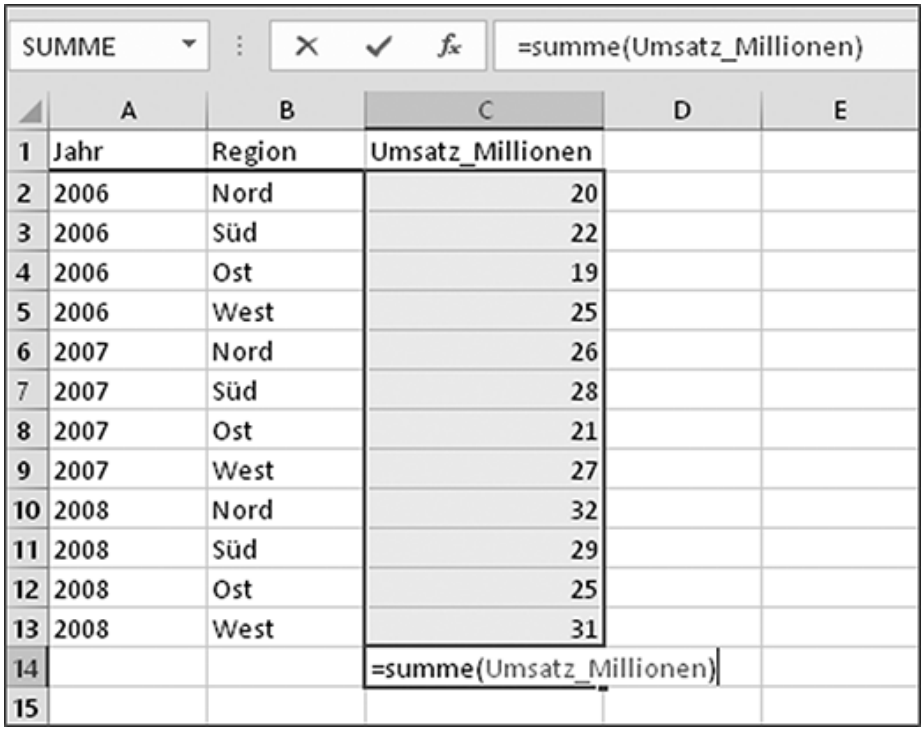
```
=SUMME(C2:C13)
```

Ein paar andere Funktionen zeigen, wie nützlich die Möglichkeit ist, Zellbereiche mit Namen versehen zu können. Diese Funktionen, `SUMMEWENN` und `SUMMEWENNS`, addieren die Werte von Zellen, wenn bestimmte Bedingungen in einem Zellbereich (`SUMMEWENN`) oder in mehr als einem Zellbereich (`SUMMEWENNS`) erfüllt sind.

Um die definierten Namen optimal verwenden zu können, habe ich sowohl Spalte A (Jahr) als auch Spalte B (Region) mit einem Namen versehen, und zwar so, wie ich es auch für Spalte C gemacht habe.

Wenn Sie für einen Zellbereich wie B2:B13 in diesem Beispiel einen Namen definieren, müssen Sie ein wenig aufpassen. Excel verhält sich ein wenig sonderbar, wenn sich in den Zellen Namen befinden. Excel vermutet dann, dass der Name, der sich in der obersten Zelle befindet, der Name ist, den Sie dem Zellbereich zuweisen wollen. In unserem Beispiel schlägt Excel daher *Nord* als Name vor und nicht *Region*. Sollte dies passieren, nehmen Sie die Änderung im Dialogfeld NEUER NAME vor.

Um sich einen Überblick über alle Namen in der Arbeitsmappe zu verschaffen, wählen Sie den Befehl FORMELN | NAMENS-MANAGER. Hierdurch wird das Dialogfeld NAMENS-MANAGER geöffnet, das Sie in Abbildung 2.16 sehen. Die weiteren Schaltflächen in der Gruppe DEFINIERTE NAMEN auf der Registerkarte FORMELN sind ebenfalls nützlich.



	A	B	C	D	E
1	Jahr	Region	Umsatz_Millionen		
2	2006	Nord	20		
3	2006	Süd	22		
4	2006	Ost	19		
5	2006	West	25		
6	2007	Nord	26		
7	2007	Süd	28		
8	2007	Ost	21		
9	2007	West	27		
10	2008	Nord	32		
11	2008	Süd	29		
12	2008	Ost	25		
13	2008	West	31		
14			=summe(Umsatz_Millionen)		
15					

Abbildung 2.15: Die Formel vervollständigen

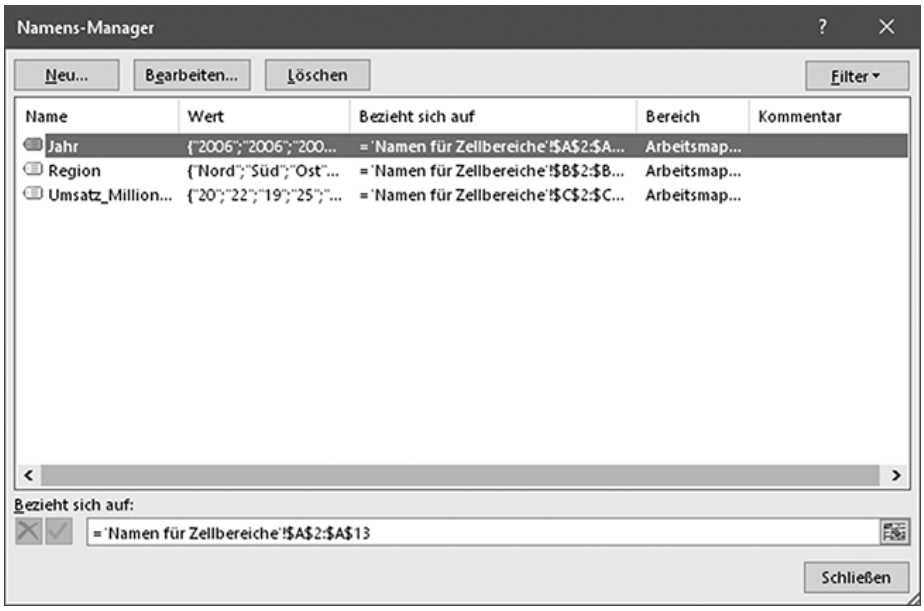


Abbildung 2.16: Die in einer Arbeitsmappe definierten Namen verwalten

Nun möchte ich die Daten in Spalte C addieren, aber nur für die Region Nord. Der Wert einer Zelle in Spalte C soll also nur dann berücksichtigt werden, wenn in der benachbarten Zelle in Spalte C *Nord* enthalten ist. Um diese Aufgabe zu erledigen, gehe ich folgendermaßen vor:

1. **Wählen Sie die Zelle aus, in der das Ergebnis stehen soll.**

   Ich habe Zelle C15 ausgewählt.

2. **Wählen Sie die geeignete Funktionskategorie und klicken Sie im Menü der Schaltfläche eine Funktion an.**

   Hierdurch wird das Dialogfeld FUNKTIONSARGUMENTE geöffnet. Ich habe FORMELN | MATHEMATIK UND TRIGONOMETRIE ausgewählt, im Menü nach unten gescrollt und SUMMEWENN angeklickt.

   SUMMEWENN verwendet drei Argumente. Das erste, BEREICH, ist der Zellbereich, der ausgewertet werden soll, um zu prüfen, ob der Wert in der Summe enthalten sein soll (in unserem Beispiel ist dies der Zellbereich, in dem *Nord, Süd, Ost* oder *West* steht). Das zweite Argument, SUCHKRITERIEN, ist der Wert, der in BEREICH

enthalten sein muss (in unserem Beispiel also *Nord)*. Das dritte Argument, SUMME_BEREICH, enthält die Werte, die ich addieren möchte.

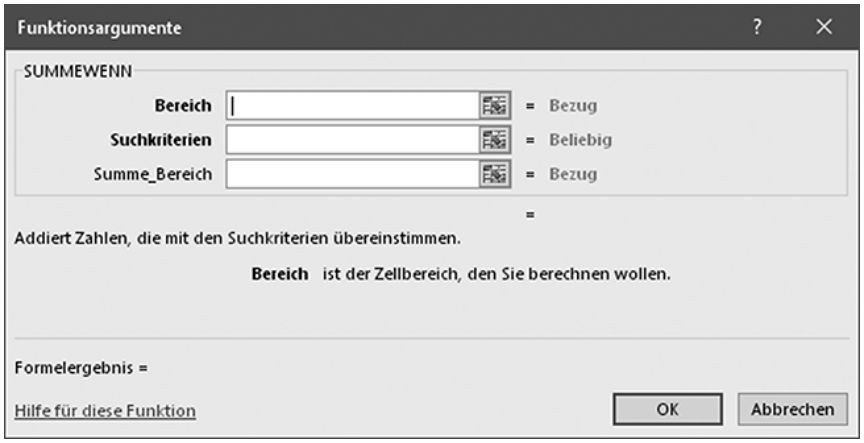


Abbildung 2.17: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *SUMMEWENN*

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte der Argumente ein.**

   An dieser Stelle wird eine weitere Schaltfläche der Gruppe DEFINIERTE NAMEN auf der Registerkarte FORMELN nützlich. Klicken Sie dort auf die Schaltfläche IN FORMEL VERWENDEN, um die Dropdownliste zu öffnen, die Abbildung 2.18 zeigt.

   Sie können diese Dropdownliste verwenden, um die Argumente BEREICH und SUMME_BEREICH anzugeben, wie es Abbildung 2.19 zeigt. In das Feld SUCHKRITERIEN habe ich **Nord** eingetippt. Excel ergänzt die doppelten Anführungszeichen.

4. **Klicken Sie auf OK.**

   Das Ergebnis (78) wird in der ausgewählten Zelle angezeigt.

In der Bearbeitungsleiste wird diese Formel angezeigt:

```
=SUMMEWENN(Region;"Nord";Umsatz_Millionen)
```

Ich kann diese Formel auch direkt in die Bearbeitungsleiste eingeben, ohne das Dialogfeld oder die Dropdownliste zu

verwenden. Wenn ich das Dialogfeld nicht verwende, muss ich die doppelten Anführungszeichen beim Suchkriterium angeben.

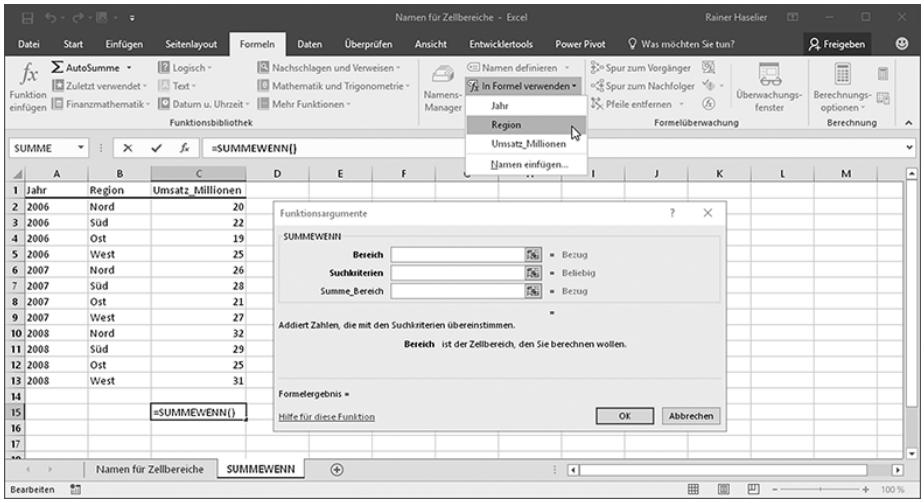


Abbildung 2.18: Die Dropdownliste IN FORMEL VERWENDEN

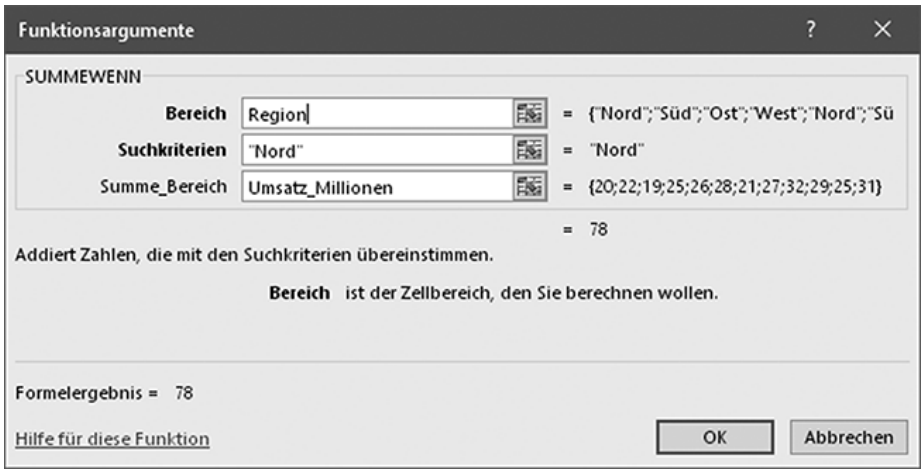


Abbildung 2.19: Das ausgefüllte Dialogfeld FUNKTIONSARGUMENTE für die Funktion *SUMMEWENN*

Die Formel in der Bearbeitungsleiste ist einfacher zu verstehen als diese:

```
=SUMMEWENN(B2:B13;"Nord";C2:C13)
```

Stimmen Sie mir zu?

Es kann Fälle geben, in denen Sie für BEREICH und für SUMME_BEREICH denselben Zellbereich eingeben. Um beispielsweise nur die Zellen in

Umsatz_Millionen zu addieren, deren Werte kleiner sind als 25, wird diese Formel verwendet:

```
=SUMMEWENN(Umsatz_Millionen;"
<25";Umsatz_Millionen)
```

Das zweite Argument, SUCHKRITERIEN, muss immer in doppelten Anführungszeichen eingegeben werden.

Was ist mit SUMMEWENNS? Diese Funktion können Sie verwenden, um die Summe der Umsätze der Region Nord zu ermitteln, aber lediglich für die Jahre 2006 und 2007. Führen Sie diese Schritte durch, um mit SUMMEWENNS diesen Wert zu ermitteln:

1. **Wählen Sie die Zelle aus, in der das Ergebnis stehen soll.**

   Ich habe Zelle C17 ausgewählt.

2. **Wählen Sie die geeignete Funktionskategorie und klicken Sie im Menü der Schaltfläche eine Funktion an.**

   Hierdurch wird das Dialogfeld FUNKTIONSARGUMENTE geöffnet.

   In diesem Beispiel habe ich im Menü der Schaltfläche MATHEMATIK UND TRIGONOMETRIE auf SUMMEWENNS geklickt, um das Dialogfeld in Abbildung 2.20 zu öffnen.

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte der Argumente ein.**

   Beachten Sie, dass bei SUMMEWENNS SUMME_BEREICH das erste und bei SUMMEWENN das letzte Argument ist. Die Werte der Argumente können Sie Abbildung 2.20 entnehmen.

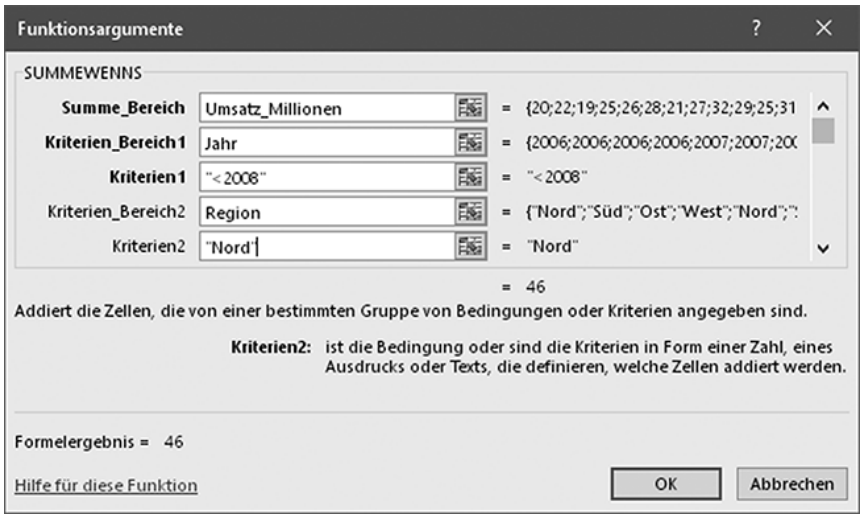


Abbildung 2.20: Das ausgefüllte Dialogfeld FUNKTIONSARGUMENTE für die Funktion *SUMMEWENNS*

4. **Die Formel in der Bearbeitungsleiste lautet:**

   ```
   =SUMMEWENNS(Umsatz_Millionen;Jahr;"
   <2008";Region;"Nord")
   ```

5. **Klicken Sie auf OK.**

   Das Ergebnis 46 wird in die ausgewählte Zelle eingefügt.

Wenn für die Zellbereiche keine Namen definiert wären, würde die Formel so aussehen:

```
=SUMMEWENNS(C2:C13;A2:A13;"<2008";B2:B13;"Nord")
```

Diese Formel ist viel schwieriger zu verstehen.

Ein definierter Name verwendet absolute Zellbezüge (siehe Kapitel 1). Wenn Sie für einen Zellbereich, für den ein Name definiert ist, das Feature Automatisch ausfüllen verwenden, wartet eine unangenehme Überraschung auf Sie: Anstatt eine Gruppe von Zellen automatisch auszufüllen, kopieren Sie einen Wert.

Hier ein Beispiel, das verdeutlicht, was ich damit meine. Angenommen, Sie weisen dem Zellbereich A2:A11 den Namen `Datenreihe_1` und dem Zellbereich B2:B11 den Namen `Datenreihe_2` zu. In A12 berechnen Sie mit `SUMME(Datenreihe_1)` die Summe dieses Zellbereichs. Weil Sie clever sind, probieren Sie, das Ergebnis aus A12 nach B12 zu ziehen, um `SUMME(Datenreihe_2)` zu berechnen. Was steht nach dem Ausfüllen in B12? `SUMME(Datenreihe_1)`.

Excel unterscheidet bei den definierten Namen nicht zwischen Groß- und Kleinschreibung. Wenn Sie einem Zellbereich den Namen `Test` zugewiesen haben, erhalten Sie mit `SUMME(Test)`, `SUMME(test)` und `SUMME(tEST)` immer das gleiche Ergebnis.

Eigene Arrayformeln erstellen

Sie können nicht nur die in Excel integrierten Arrayformeln verwenden, sondern auch eigene Arrayformeln erstellen. Um sich die Dinge zu vereinfachen, können Sie auch hier benannte Zellbereiche verwenden.

Abbildung 2.21 zeigt zwei benannte Zellbereiche X und Y, die in Spalte C beziehungsweise Spalte D definiert wurden. X verweist auf C2 bis C5 (*nicht* auf C1 bis C5) und Y verweist auf D2 bis D5 (*nicht* auf D1 bis D5). Die Spaltenüberschrift in Spalte F lautet XY. Jede Zelle in Spalte F soll das Produkt der korrespondierenden Zellen in Spalte C und Spalte D enthalten.

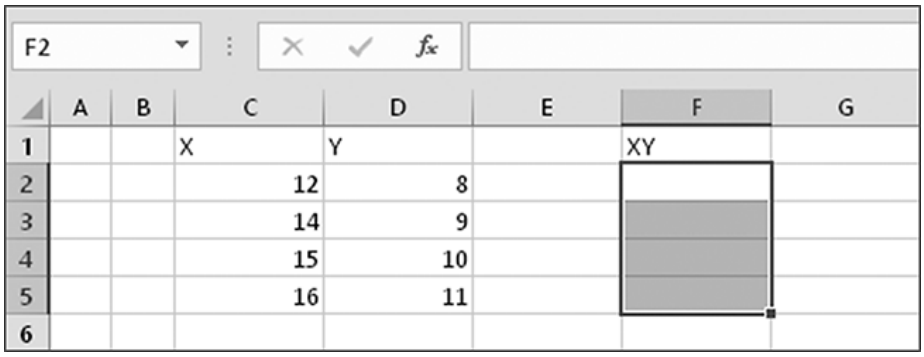


Abbildung 2.21: Zwei benannte Zellbereiche

Ein einfacher Weg, um die Produkte zu erhalten, ist natürlich, in F2 **= C2*D2** einzugeben und dann die weiteren Zellen in Spalte F automatisch auszufüllen.

Um jedoch die Verwendung eigener Arrayformeln zu verdeutlichen, führen Sie diese Schritte durch, um die Daten zu bearbeiten (siehe Abbildung 2.21).

1. **Wählen Sie den Zellbereich aus, der die Ergebnisse der Arrayformel erhalten soll.**

   Dies wäre hier der Zellbereich F2 bis F5, oder in der Excel-Terminologie: F2:F5. (In Abbildung 2.21 ist dieser Zellbereich ausgewählt.)

2. **Geben Sie in den ausgewählten Zellbereich die Formel ein.**

   Die Formel in diesem Beispiel lautet: **= X * Y**.

3. **Drücken Sie** Strg+Shift+Enter **(und nicht nur** Enter**). Auf dem Mac drücken Sie** Ctrl+Shift+Enter **oder** MacBef+Shift+Enter**.**

   Das Ergebnis wird, wie in Abbildung 2.22 zu sehen, in F2 bis F5 angezeigt. Beachten Sie, dass in der Bearbeitungsleiste die Formel `{=X*Y}` angezeigt wird. Wie ich bereits weiter vorne

erwähnt habe, markiert Excel Arrayformeln durch geschweifte Klammern.

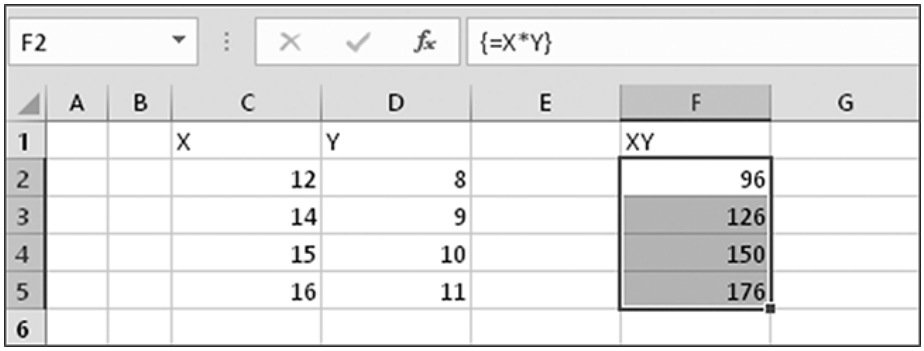


Abbildung 2.22: Die Ergebnisse der Arrayformel {=X*Y}

Noch ein weiterer Punkt, auf den ich weiter vorne bereits hingewiesen habe: Wenn Sie einen Zellbereich mit einem Namen versehen, achten Sie darauf, die Zelle, die den Namen/die Spaltenüberschrift enthält, *nicht* mit einzubeziehen. Sollte die Zelle mit der Spaltenüberschrift Teil des definierten Namens sein, versucht eine Arrayformel wie `{=X*Y}`, den Buchstaben X mit dem Buchstaben Y zu multiplizieren, um das erste Ergebnis zu ermitteln. Da dies nicht möglich ist, erhalten Sie hier stattdessen den unschönen `#WERT!`-Fehler.

Das merkwürdige Verhalten auf dem Mac, das ich im Zusammenhang mit der Arrayfunktion `HÄUFIGKEIT` erwähnte, taucht bei eigenen Arrayformeln nicht auf. Es ist also nicht erforderlich, in die Bearbeitungsleiste zu klicken, bevor Sie die Tastenkombination drücken.

Bei gedrückter MacBef-Taste die Schaltfläche EINGEBEN anzuklicken (das Häkchen in der Bearbeitungsleiste), funktioniert auch bei eigenen Arrayformeln.

Die Datenanalysetools verwenden

Excel verfügt über eine Reihe raffinierter Tools für die Datenanalyse. In Tabelle 2.1 sind die Tools aufgeführt, die ich beschreiben werde. (Das eine Tool, das ich nicht beschreibe, die Fourier-Analyse, ist extrem technisch.) Einige der Begriffe in der Tabelle werden Sie vielleicht noch nicht kennen, daher werde ich diese im weiteren Verlauf des Buches erläutern.

Tool	Was es macht
Anova: Einfaktoriell	Varianzanalyse für mindestens zwei Stichproben
Anova: Zweifaktoriell mit Messwiederholung	Varianzanalyse mit zwei unabhängigen Variablen und mehreren Beobachtungen in jeder Kombination der Variablenausprägungen
Anova: Zweifaktoriell ohne Messwiederholung	Varianzanalyse mit zwei unabhängigen Variablen und einer Beobachtung in jeder Kombination der Variablenausprägungen
Korrelation	Berechnet aus den Messungen einer Stichprobe von mindestens drei Objekten eine Matrix der Korrelationskoeffizienten für alle möglichen Messpaare
Kovarianz	Berechnet aus den Messungen einer Stichprobe von mindestens drei Objekten eine Matrix der Kovarianz für alle möglichen Messpaare
Populationskenngrößen	Erstellt einen Bericht über zentrale Tendenz, Streuung der Daten und andere Eigenschaften der Werte im markierten Zellbereich
Exponentielles Glätten	Berechnet in einer Wertefolge eine Vorhersage basierend auf vorherigen Werten und auf einer Vorhersage für diese Werte
Zwei-Stichproben F-Test für Varianzen	Führt einen F-Test zum Vergleich zweier Varianzen durch
Histogramm	Berechnet individuelle und kumulierte Häufigkeiten für Werte im markierten Zellbereich

Tool	Was es macht
Gleitender Durchschnitt	Berechnet in einer Wertefolge eine Vorhersage, die den Mittelwert einer bestimmten Anzahl vorhergehender Werte darstellt
Zufallszahlengenerierung	Stellt eine bestimmte Anzahl an Zufallszahlen bereit, die aus einer von sieben möglichen Verteilungen generiert werden
Rang und Quantil	Erstellt eine Tabelle, die den Rang sowie den Quantilsrang jedes Werts in einem Datensatz enthält
Regression	Erstellt einen Bericht über die Regressionsstatistik basierend auf einer linearen Regression über einen Datensatz, der eine abhängige und mindestens eine unabhängige Variable enthält
Stichprobenziehung	Erstellt eine Stichprobe aus den Werten in einem festgelegten Zellbereich
Zwei-Stichproben t-Test	Drei Tools für t-Tests testen die Differenz zwischen zwei Mittelwerten. Ein Test basiert auf der Annahme gleicher Varianzen in den beiden Stichproben. Ein anderer Test basiert auf der Annahme ungleicher Varianzen in den beiden Stichproben. Der dritte Test setzt abhängige Stichproben voraus.
Zwei-Stichproben-Test bei bekannten Varianzen	Führt einen Zwei-Stichproben-Test zum Vergleichen zweier Mittelwerte durch, wobei die Varianzen bekannt sind

Tabelle 2.1: Die Datenanalysetools von Excel

Bei den Datenanalysewerkzeugen handelt es sich um ein Add-In. Um das Add-In verwenden zu können, müssen Sie es zuerst laden. In der Windows-Version wählen Sie dazu DATEI | OPTIONEN.

Excel zeigt das Dialogfeld EXCEL-OPTIONEN an. Führen Sie dann diese Schritte durch:

1. **Klicken Sie an der linken Seite des Dialogfeldes EXCEL-OPTIONEN auf ADD-INS.**

   Excel zeigt eine Liste von Add-Ins an. Unter der Liste befindet sich eine Dropdownliste, die mit VERWALTEN beschriftet ist.

2. **Wählen Sie in dieser Liste EXCEL ADD-INS aus (falls diese Option nicht bereits ausgewählt ist).**

3. **Klicken Sie auf LOS.**

   Excel zeigt das Dialogfeld ADD-INS an (siehe Abbildung 2.23).

4. **Schalten Sie die Kontrollkästchen neben ANALYSE-FUNKTIONEN und neben SOLVER ein. Klicken Sie dann auf OK.**

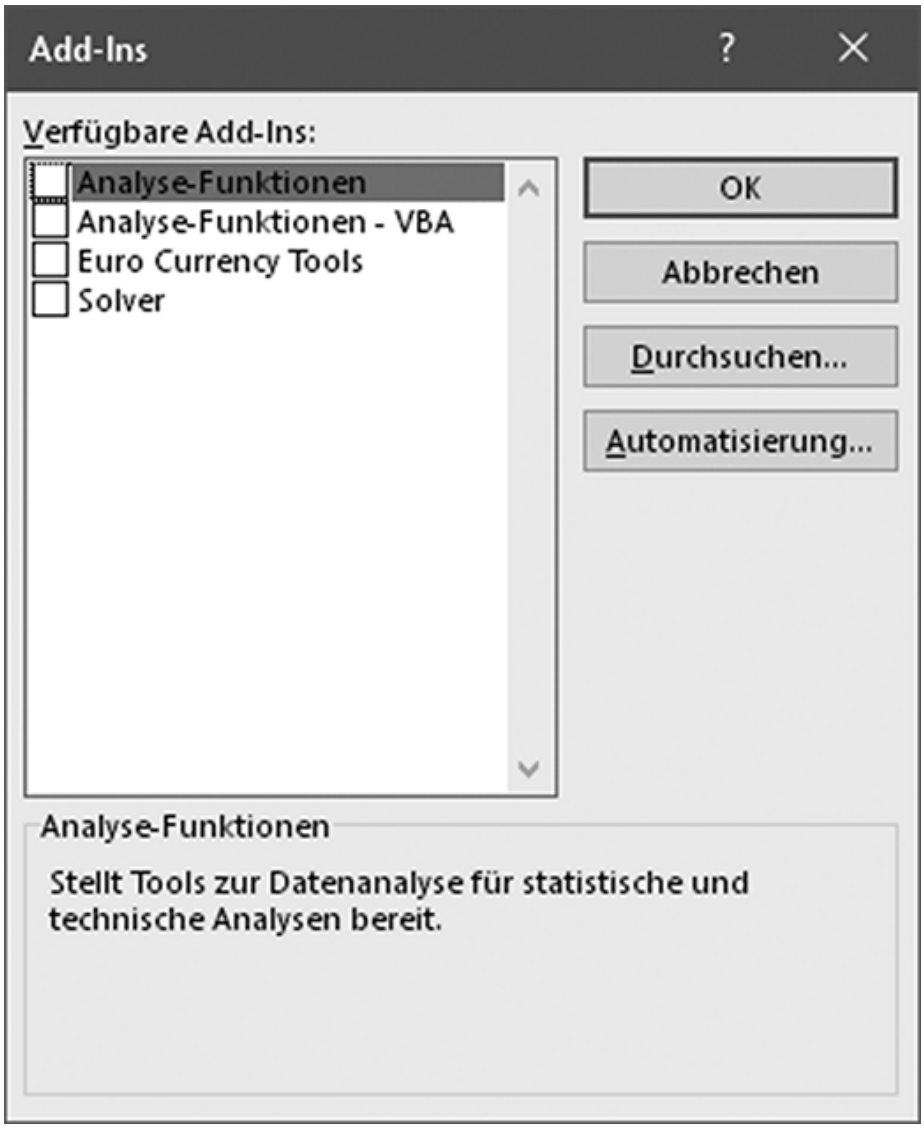


Abbildung 2.23: Das Dialogfeld ADD-INS

Nachdem Excel die Analysefunktionen und den Solver geladen hat, werden auf der Registerkarte DATEN in der Gruppe ANALYSE die Schaltflächen DATENANALYSE und SOLVER angezeigt.

Auf dem Mac ist die Installation der Analysefunktionen und des Solvers in weniger Schritten erledigt als unter Windows.

Wählen Sie zuerst im Menü EXTRAS | EXCEL-ADD-INS. Abbildung 2.24 zeigt diesen Menübefehl.

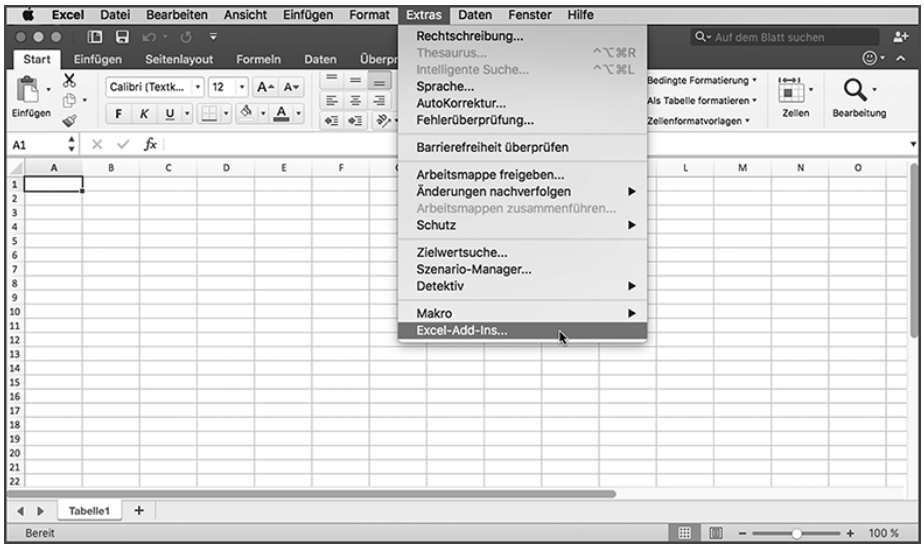


Abbildung 2.24: Das Menü EXTRAS auf dem Mac

Excel öffnet das Dialogfeld ADD-INS (siehe Abbildung 2.25). Schalten Sie die Kontrollkästchen neben ANALYSE-FUNKTIONEN und neben SOLVER ADD-IN ein. Klicken Sie auf OK und Sie sind fertig. Im Menü EXTRAS tauchen nun zwei neue Befehle auf: DATENANALYSE und SOLVER.

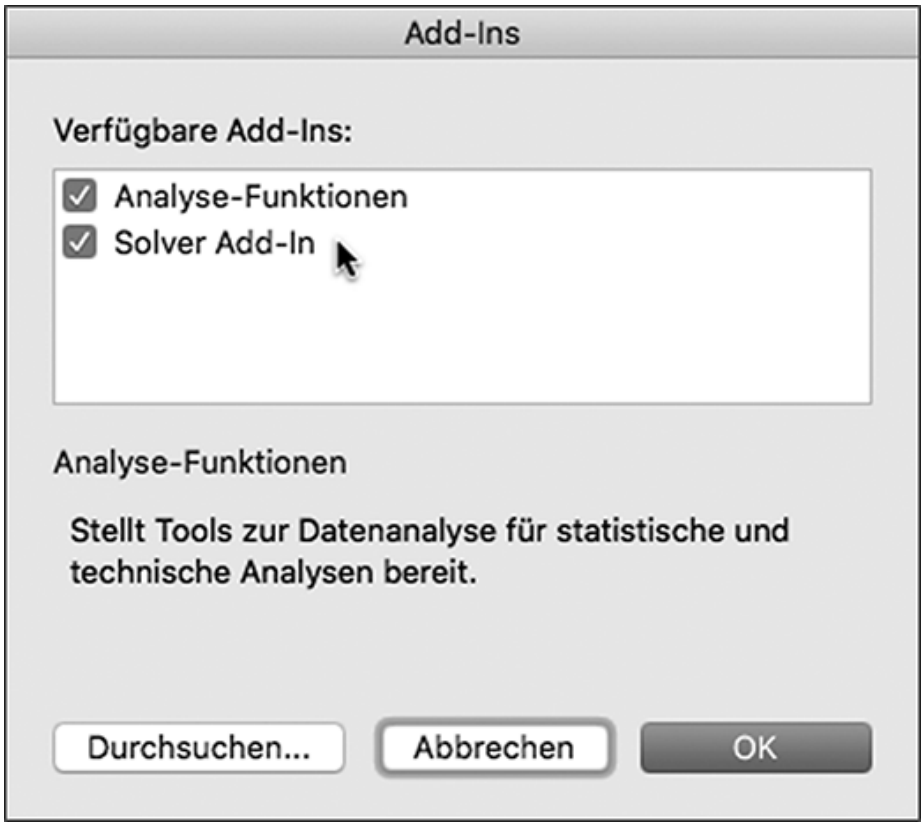


Abbildung 2.25: Das Dialogfeld ADD-INS auf dem Mac

Hier folgenden die allgemeinen Schritte, um eines der Datenanalysetools zu verwenden:

1. **Geben Sie die Daten in einen Zellbereich ein.**
2. **Wählen Sie DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**
3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN das Datenanalysetool aus, das Sie verwenden möchten.**
4. **Klicken Sie auf OK (oder doppelklicken Sie auf die Auswahl), um das Dialogfeld für das ausgewählte Tool zu öffnen.**
5. **Geben Sie im Dialogfeld die angeforderten Informationen ein.**

Ich weiß, dieser Schritt ist etwas vage, aber jedes Tools ist anders.

6. **Klicken Sie auf OK, um das Dialogfeld zu schließen und die Ergebnisse anzeigen zu lassen.**

Hier ein Beispiel, damit Sie sich mit diesen Tools vertraut machen können. In diesem Beispiel werde ich das Tool für Populationskenngrößen (die auch als beschreibende oder deskriptive Statistiken bezeichnet werden) verwenden. Dieses Tool berechnet eine Reihe von Kenngrößen, die mehrerer Werte zusammenfassen.

1. **Geben Sie die Daten in einen Zellbereich ein.**

   In Abbildung 2.26 ist eine Reihe mit Zahlen in den Zellen B2 bis B9 dargestellt. Die Spaltenüberschrift befindet sich in der Zelle B1.

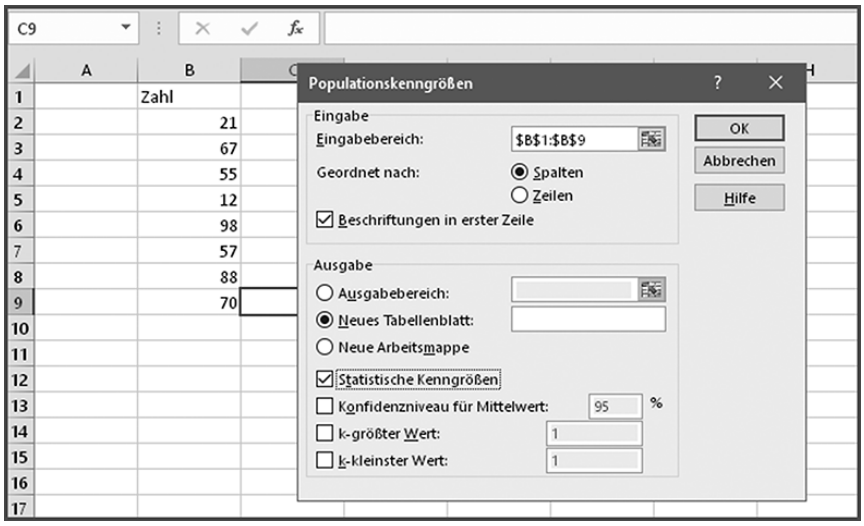


Abbildung 2.26: Das Analysetool für Populationskenngrößen

2. **Klicken Sie auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Klicken Sie auf POPULATIONSKENNGRÖSSEN, um das Dialogfeld POPULATIONSKENNGRÖSSEN zu öffnen.**

4. **Geben Sie den Zellbereich mit den Eingabedaten an.**

   Geben Sie im Feld EINGABEBEREICH die Zellen an, in denen sich die Werte befinden. In diesem Beispiel ist das B1:B9. Hierzu setzen Sie am besten den Cursor in die obere Zelle (B1), halten die Shift-Taste gedrückt und klicken auf die untere Zelle (B9). Damit wird der Eingabebereich mit absoluten Zellbezügen ($B$1:$B$9) in das Feld EINGABEBEREICH eingegeben.

5. **Aktivieren Sie das Optionsfeld SPALTEN, um anzugeben, wie die Daten geordnet sind.**

6. **Aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE, da der Eingabebereich eine Spaltenüberschrift enthält.**

7. **Aktivieren Sie das Optionsfeld NEUES TABELLENBLATT, falls dieses nicht bereits aktiviert ist.**

   Damit wird Excel mitgeteilt, dass innerhalb der aktuellen Arbeitsmappe ein neues Tabellenblatt erstellt werden soll und dass die Ergebnisse in diesem neu erstellten Tabellenblatt angezeigt werden sollen.

8. **Aktivieren Sie das Kontrollkästchen STATISTISCHE KENNGRÖSSEN. Die anderen Kontrollkästchen lassen Sie deaktiviert.**

9. **Klicken Sie auf OK.**

   Das neue Tabellenblatt wird geöffnet und zeigt die Kenngrößen der Daten an. In Abbildung 2.27 ist das neue Tabellenblatt dargestellt, nachdem ich die Spalte A verbreitert habe.

	A	B	C
1	*Zahl*		
2			
3	Mittelwert	58,5	
4	Standardfehler	10,53734854	
5	Median	62	
6	Modus	#NV	
7	Standardabweichung	29,80412244	
8	Stichprobenvarianz	888,2857143	
9	Kurtosis	-0,612417016	
10	Schiefe	-0,459653925	
11	Wertebereich	86	
12	Minimum	12	
13	Maximum	98	
14	Summe	468	
15	Anzahl	8	
16			

Abbildung 2.27: Die Ausgabe des Tools für Populationskenngrößen

Die Bedeutung der einzelnen Elemente in der Zusammenfassung werde ich Ihnen nicht erklären. Damit warte ich bis Kapitel 7, wenn ich mich eingehender mit Populationskenngrößen beschäftige.

Auf oft verwendete Funktionen zugreifen

Müssen Sie schnell auf einige der häufig verwendeten statistischen Funktionen zugreifen? Die Funktionen `MITTELWERT`, `MIN` (kleinster Wert im ausgewählten Zellbereich) und `MAX` (größter Wert im ausgewählten Zellbereich) stehen Ihnen zur Verfügung, indem Sie auf den Pfeil der Schaltfläche SUMME beziehungsweise AUTOSUMME

klicken, die Sie auf den Registerkarten START beziehungsweise FORMELN finden. Über diese Schaltflächen können Sie ebenfalls auf die mathematischen Funktionen SUMME und ANZAHL (zählt die Anzahl der numerischen Werte in einem Zellbereich) zugreifen.

Die Schaltfläche SUMME beziehungsweise AUTOSUMME ist mit Σ beschriftet. Abbildung 2.28 zeigt, wo sich diese Schaltfläche auf der Registerkarte START befindet und welche Befehle im Menü der Schaltfläche enthalten sind.

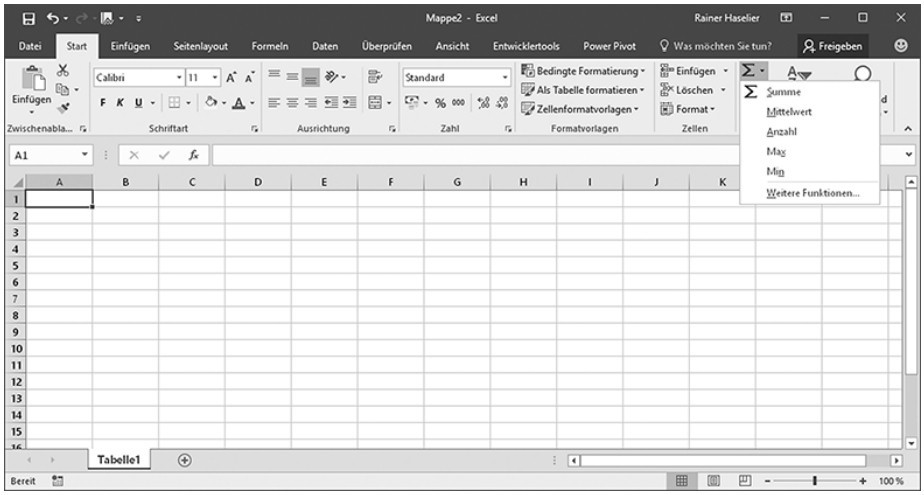


Abbildung 2.28: Das Menü der Schaltfläche SUMME auf der Registerkarte START

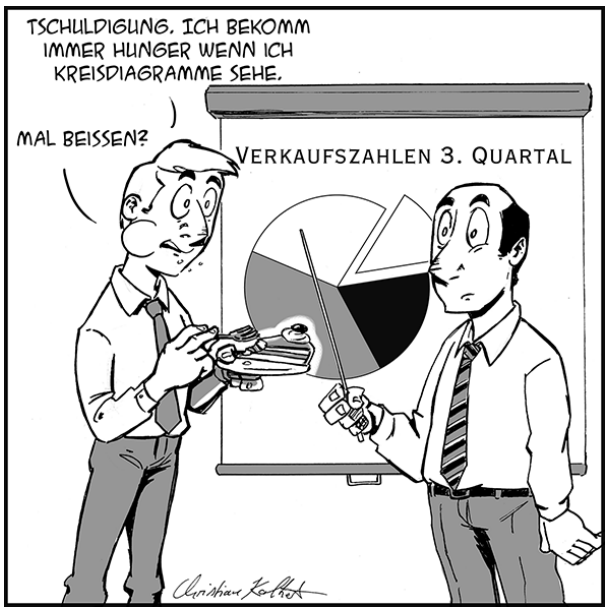
Wenn Sie nicht den Pfeil, sondern die Schaltfläche START | SUMME oder FORMELN | AUTOSUMME selbst anklicken, erhalten Sie die Funktion SUMME.

Der letzte Befehl im Menü ist eine weitere Möglichkeit, um das Dialogfeld FUNKTION EINFÜGEN zu öffnen.

Ein weiterer Vorteil der Verwendung dieses Menüs: Sie sparen einen Schritt. Wenn Sie eine der Funktionen wählen, brauchen Sie keine Zelle für das Ergebnis zu markieren. Wählen Sie einfach einen Zellbereich aus und die Funktion fügt das Ergebnis in eine Zelle direkt nach dem ausgewählten Zellbereich ein.

Teil II

Daten beschreiben



In diesem Teil …

Daten zusammenfassen und beschreiben

Die Diagrammfeatures von Excel verwenden

Die zentrale Tendenz und Streuung ermitteln

Mit Standardwerten arbeiten

Normalverteilungen verstehen und visualisieren

Kapitel 3

Präsentieren: Daten in Diagrammen darstellen

In diesem Kapitel

- Einführung in Diagramme
- Die Diagramm-Features von Excel nutzen
- Diagramme für die Statistik erstellen

Die grafische Darstellung von Daten ist ein extrem wichtiges Thema in der Statistik. In Diagrammen können Sie Beziehungen und Trends erkennen, die Ihnen bei der Betrachtung der bloßen Zahlen möglicherweise nicht auffallen würden. Grafische Darstellungen sind auch in anderer Hinsicht hilfreich: Sie können damit Ihre Ideen präsentieren und anderen Ihre Sicht der Dinge näher bringen.

Es gibt viele unterschiedliche Arten von Diagrammen. In diesem Kapitel erläutere ich die Arten von Diagrammen, die im Rahmen der Statistik verwendet werden, und erkläre, wann Sie welchen Diagrammtyp am besten verwenden. Ich verrate Ihnen außerdem, wie Sie diese Diagramme mit Excel erstellen können.

Warum Diagramme verwenden?

Nehmen wir einmal an, Sie müssen für einen Tagungsausschuss eine Statistik über die Einnahmen aus dem kommerziellen Weltraumsektor in den frühen 1990er-Jahren erstellen.

Was würden Sie eher präsentieren – die Daten in Tabelle 3.1 oder das Diagramm in Abbildung 3.1, in dem dieselben Daten dargestellt

sind? (Die Daten stammen übrigens aus dem statistischen Jahrbuch der Vereinigten Staaten [Statistical Abstract of the U.S.] vom US-amerikanischen Handelsministerium.)

Sektor	1990	1991	1992	1993	1994
Gelieferte kommerzielle Satelliten	1.000	1.300	1.300	1.100	1.400
Satellitendienste	800	1.200	1.500	1.850	2.330
Satelliteneinrichtungen am Boden	860	1.300	1.400	1.600	1.970
Kommerzielle Satellitenstarts	570	380	450	465	580
Fernerkundungsdaten	155	190	210	250	300
Infrastruktur der kommerziellen Forschung und Entwicklung	0	0	0	30	60
Gesamt	3.385	4.370	4.860	5.295	6.640

Tabelle 3.1: Einnahmen der USA aus dem kommerziellen Weltraumsektor von 1990 bis 1994 (in Mio. US-Dollar)

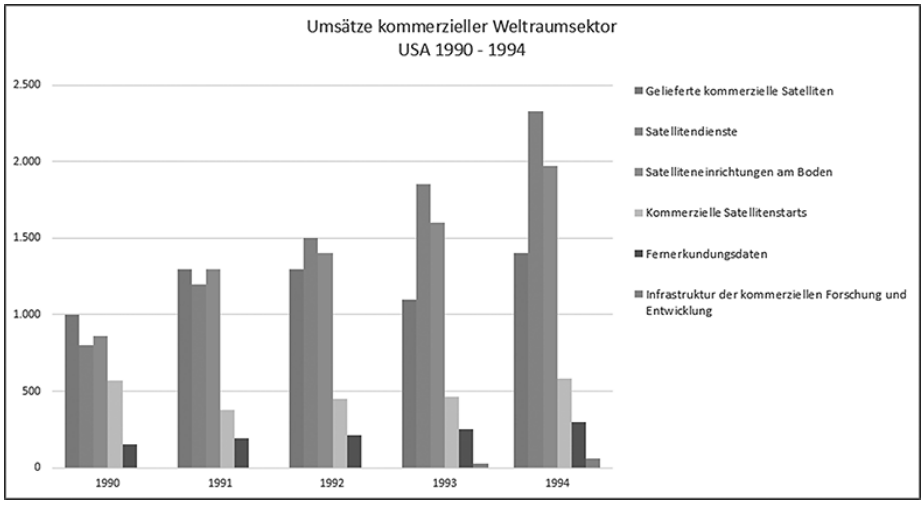


Abbildung 3.1: Die Daten aus Tabelle 3.1 in einem Diagramm

Welche Präsentation würde einen stärkeren und dauerhafteren Eindruck hinterlassen? Obwohl die Tabelle mit Sicherheit informativ ist, würden die meisten Leute behaupten, dass das Diagramm die Inhalte besser vermittelt und dass diese so leichter zu behalten sind.

(Augen, die beim Betrachten von Zahlen leicht glasig werden, leuchten häufig auf, wenn sie Bilder sehen.)

Anhand des Diagramms werden Trends deutlich, die in der Tabelle nicht so schnell erkannt werden. (Die Satellitendienste sind schnell angewachsen. Die kommerziellen Satellitenstarts nicht so sehr.) Fazit: Tabellen sind gut, Diagramme sind besser.

Diagramme machen Ideen lebendig, die ohne sie möglicherweise nur schwer zu verstehen sind. Ich mache das übrigens im ganzen Buch. Ich veranschauliche Dinge durch … nun … die Veranschaulichung von Dingen!

Einige grundlegende Dinge

Die meisten Diagramme besitzen wie das in Abbildung 3.1 eine horizontale und eine vertikale Achse. Andere Diagrammtypen (Kreisdiagramm, Treemap und Sunburst), die ich weiter hinten in diesem Kapitel vorstelle, besitzen diese Achsen nicht. Weder das Netzdiagramm noch das Kastendiagramm (die ebenfalls weiter hinten beschrieben werden), besitzen eine horizontale Achse.

Die horizontale Achse wird üblicherweise als *x-Achse* und die vertikale Achse als *y-Achse* bezeichnet.

Es ist außerdem üblich, an derhorizontalen Achse die *unabhängige Variable* und an der vertikalen Achse die *abhängige Variable* abzutragen. Bei einem Diagrammtyp von Excel ist das genau umgekehrt. Aber darüber erfahren Sie mehr, wenn ich diesen Diagrammtyp vorstelle. Nur, um Sie schon einmal zu warnen: Bei Excel wird dieses Format mit den umgedrehten Achsen *Balkendiagramm* genannt. Sie haben Diagramme wie das in Abbildung 3.1 bestimmt schon als Balkendiagramm bezeichnet gesehen. Ich auch. Bei Excel heißen diese Diagramme jedoch *Säulendiagramme*. So werde ich sie von nun an auch nennen.

Um auf die Begriffe »unabhängig« und »abhängig« zurückzukommen: Diese Begriffe lassen darauf schließen, dass Änderungen in der vertikalen Richtung von Änderungen in der horizontalen Richtung (zumindest teilweise) abhängen.

Ein wichtiges Prinzip beim Erstellen eines Diagramms: Schonen Sie die Augen des Betrachters! Wenn Sie zu viele Informationen in ein Diagramm packen oder zu viele Spezialeffekte verwenden, machen Sie die ganze Wirkung des Diagramms zunichte.

Für Abbildung 3.1 musste ich mir beispielsweise überlegen, wie ich die Säulen füllen wollte. Farblich gekennzeichnete Säulen wären hilfreich gewesen. Aber in diesem Buch sehen Sie nur Schwarz, Weiß und Grauschattierungen.

Vieles beim Erstellen von Diagrammen kommt mit der Erfahrung. Und letztlich müssen Sie sich einfach auf Ihr eigenes Urteilsvermögen verlassen. Bei diesem Beispiel nutzte ich mein Urteilsvermögen und fügte horizontale Gitternetzlinien ein. Bei den meisten Diagrammen verwende ich lieber keine. Hier verleihen sie dem Diagramm eine Struktur und der Betrachter kann die Beträge (die Werte) leichter den jeweiligen Spalten zuordnen. Aber wie gesagt, das ist nur meine Meinung.

Die Diagramm-Features von Excel

Wie ich bereits erwähnt habe, handelt es sich bei dem Diagramm in Abbildung 3.1 um ein Säulendiagramm. Das ist einer der zahlreichen Diagrammtypen, die Sie mit Excel erstellen können. Von den Diagramm-Features, die Excel bereitstellt, sind jedoch nur einige wenige für Statistiken brauchbar. Diese werden in diesem Kapitel beschrieben.

Ich werde Ihnen zeigen, wie ein Säulendiagramm, ein Kreisdiagramm, ein Balkendiagramm, ein Liniendiagramm und ein Punktdiagramm erstellt werden. Außerdem stelle ich ein spannendes neues Feature vor, das bei der Einführung Power View genannt wurde und das in der aktuellen Version den Namen 3D-Karte trägt.

Ein Diagramm einfügen

Wenn Sie ein Diagramm erstellen, fügen Sie es in eine Arbeitsmappe ein. Hieran können Sie sofort erkennen, dass sich die betreffenden Befehle auf der Registerkarte EINFÜGEN und dort in der Gruppe DIAGRAMME befinden (siehe Abbildung 3.2).

Abbildung 3.2: Die Gruppe DIAGRAMME auf der Registerkarte EINFÜGEN

Um ein Diagramm einzufügen, führen Sie die folgenden Schritte durch:

1. **Geben Sie die Daten in ein Tabellenblatt ein.**

2. **Markieren Sie die Daten, die im Diagramm dargestellt werden sollen.**

3. **Klicken Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME auf EMPFOHLENE DIAGRAMME.**

   Das Dialogfeld DIAGRAMM EINFÜGEN wird geöffnet. Auf der Registerkarte EMPFOHLENE DIAGRAMME sehen Sie, welchen Diagrammtyp Excel für die von Ihnen markierten Daten vorschlägt. Wählen Sie einen Diagrammtyp aus und Excel fügt das Diagramm in das Tabellenblatt ein.

4. **Bearbeiten Sie das Diagramm.**

   Klicken Sie das Diagramm an, damit Excel im Menüband die Registerkarten ENTWURF und FORMAT einblendet. Mit den Befehlen dieser Registerkarten können Sie alle möglichen Änderungen an dem Diagramm vornehmen. Sie können auch eines der Diagrammelemente anklicken (beispielsweise eine Achse oder einen Datenpunkt), damit Excel an der rechten Seite des Fensters einen Aufgabenbereich anzeigt. In diesem Aufgabenbereich können Sie das ausgewählte Element bearbeiten. Wenn Sie den Aufgabenbereich geöffnet lassen und

ein anderes Element anklicken, ändert sich der Inhalt des Aufgabenbereichs und Sie sehen dort sofort die Bearbeitungsmöglichkeiten, die Ihnen für das ausgewählte Element zur Verfügung stehen.

Ja, es ist wirklich so einfach. (Das Erstellen von Diagrammen scheint mit jeder neuer Excel-Version immer einfacher zu werden.) Der nächste Abschnitt beschreibt, was ich damit meine.

Hier noch eine wichtige Information zu den Excel-Diagrammen. In Excel ist ein Diagramm *dynamisch*. Nachdem Sie ein Diagramm erstellt haben, führt eine Änderung an den zugrunde liegenden Daten automatisch dazu, dass sich auch das Diagramm ändert.

In diesem und allen nachfolgenden Beispielen besteht der dritte Schritt immer darin, den Befehl EINFÜGEN | EMPFOHLENE DIAGRAMME auszuwählen.

Sie können auch direkt einen Diagrammtyp auswählen, ohne die Empfehlung von Excel zu berücksichtigen. Für jeden Diagrammtyp finden Sie in der Gruppe DIAGRAMME der Registerkarte EINFÜGEN eine eigene Schaltfläche. Wenn Sie im Dialogfeld DIAGRAMM EINFÜGEN die Registerkarte ALLE DIAGRAMME öffnen, stehen Ihnen ebenfalls alle Diagrammtypen zur Verfügung.

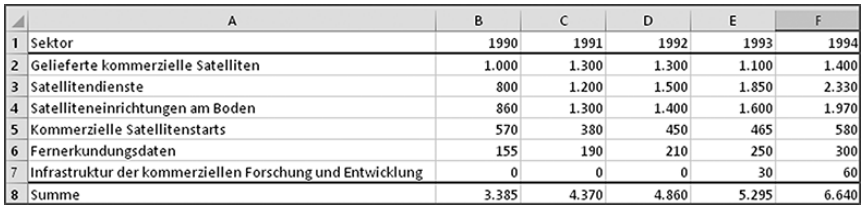
So werden Sie ein Säulenheiliger

In diesem Abschnitt erfahren Sie, wie Sie das großartige Diagramm in Abbildung 3.1 erstellen können. Führen Sie dazu folgende Schritte

durch:

1. **Geben Sie die Daten in ein Tabellenblatt ein.**

   Abbildung 3.3 zeigt die Daten aus Tabelle 3.1, nachdem diese in ein Tabellenblatt eingegeben wurden.



	A	B	C	D	E	F
1	Sektor	1990	1991	1992	1993	1994
2	Gelieferte kommerzielle Satelliten	1.000	1.300	1.300	1.100	1.400
3	Satellitendienste	800	1.200	1.500	1.850	2.330
4	Satelliteneinrichtungen am Boden	860	1.300	1.400	1.600	1.970
5	Kommerzielle Satellitenstarts	570	380	450	465	580
6	Fernerkundungsdaten	155	190	210	250	300
7	Infrastruktur der kommerziellen Forschung und Entwicklung	0	0	0	30	60
8	Summe	3.385	4.370	4.860	5.295	6.640

Abbildung 3.3: Die Daten aus Tabelle 3.1 wurden in ein Tabellenblatt eingegeben.

2. **Markieren Sie die Daten, die im Diagramm dargestellt werden sollen.**

   Ich habe den Zellbereich A1:F7 markiert. Diese Auswahl enthält die Beschriftungen für die Achsen, aber nicht Zeile 8, in der sich die Summen befinden.

3. **Klicken Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME auf EMPFOHLENE DIAGRAMME.**

   Wenn Sie den Befehl EINFÜGEN | DIAGRAMME | EMPFOHLENE DIAGRAMME wählen, wird das Dialogfeld DIAGRAMM EINFÜGEN geöffnet, das Sie in Abbildung 3.4 sehen. Ich habe in der Liste der empfohlenen Diagramme auf der linken Seite nach unten geblättert und dort die fünfte Empfehlung ausgewählt. (Anscheinend unterscheiden sich meine Vorlieben von denen von Excel. Möglicherweise entwickelt Excel in einer zukünftigen Version den gleichen Geschmack, den ich habe.) Dieser Diagrammtyp wird *Gruppierte Säulen* genannt.

4. **Bearbeiten Sie das Diagramm.**

   Abbildung 3.5 zeigt das erstellte Diagramm und die Registerkarten ENTWURF und FORMAT. Diese beiden

Registerkarten enthalten die Werkzeuge für Diagramme. Wie Sie sehen können, muss ich ein paar Anpassungen vornehmen. Excel hat nicht richtig vorhergesehen, wie ich das Diagramm gestalten möchte. Das Diagramm sieht ganz okay aus, jedoch wird es (zumindest für mich) besser, wenn ich die Legende verschiebe. (Die Legende ist der Teil unterhalb der x-Achse, der die Bedeutung der Farben erläutert.) Wie in Abbildung 3.1 zu sehen, bevorzuge ich es, wenn die Legende rechts neben dem Diagramm steht.

Um diese Änderung durchzuführen, doppelklicke ich auf die Legende. Excel zeigt daraufhin den Aufgabenbereich LEGENDE FORMATIEREN an (siehe Abbildung 3.5). Dort klicke ich auf OBEN RECHTS, um die Legende neu zu platzieren.

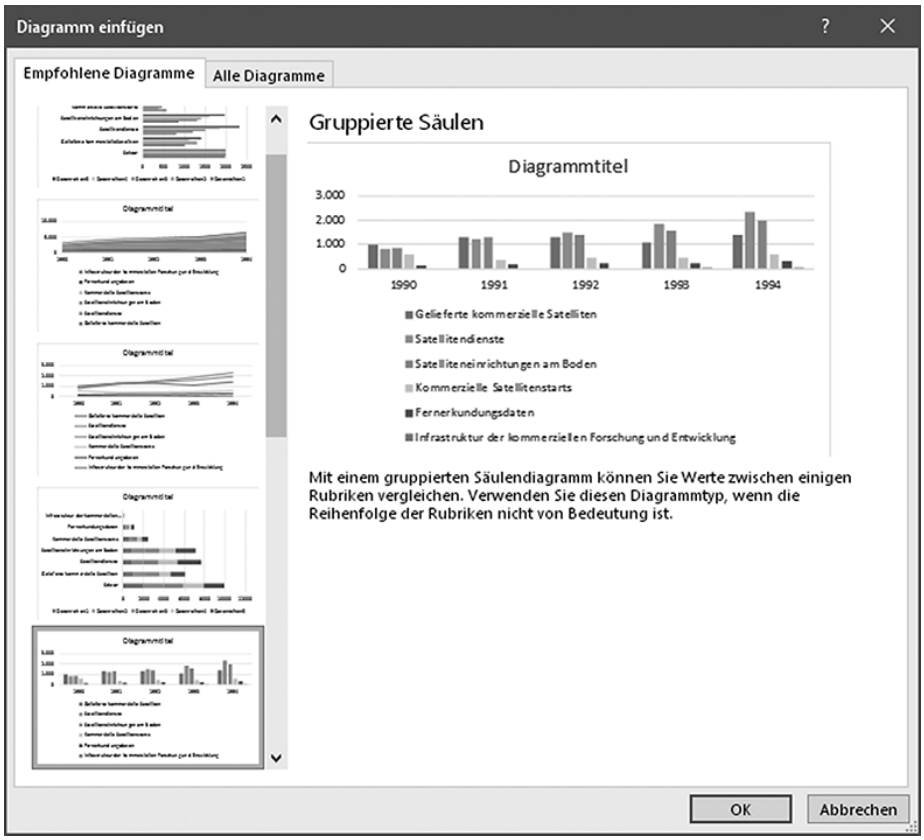


Abbildung 3.4: Das Dialogfeld DIAGRAMM EINFÜGEN

Es bleibt noch mehr zu tun. Aus irgendeinem Grund erstellt Excel für die y-Achse keine Linie und für die x-Achse wird eine hellgraue Linie

verwendet. Außerdem sind die Achsen noch nicht beschriftet und für den Diagrammtitel wurde lediglich ein Platzhalter eingefügt.

Ich beginne mit der Formatierung der Achsen. Wenn ich die y-Achse anklicke, ändert sich LEGENDE FORMATIEREN in ACHSE FORMATIEREN. Abbildung 3.6 zeigt diesen Aufgabenbereich, nachdem ich zuerst auf ACHSENOPTIONEN und dann auf LINIE geklickt habe. Ich habe die Schaltfläche FARBE verwendet, um die Farbe der y-Achse zu ändern. Ich kann dann anschließend die x-Achse markieren und deren Farbe wie bei der y-Achse ändern.

Anschließend ergänze ich die Achsenbeschriftungen und den Diagrammtitel. Hierzu zeige ich mit der Maus auf das Diagramm und klicke es an. Rechts neben dem Diagramm werden drei Schaltflächen angezeigt. Eine dieser Schaltflächen, mit einem Pluszeichen versehen, ist die Schaltfläche DIAGRAMMELEMENTE. Klicken Sie diese Schaltfläche an und schalten Sie dann im Menü der Schaltfläche das Kontrollkästchen ACHSENTITEL ein. Abbildung 3.7 zeigt die Schaltfläche DIAGRAMMELEMENTE mit geöffnetem Popupmenü und aktiviertem Kontrollkästchen ACHSENTITEL. Wenn Sie das Kontrollkästchen ACHSENTITEL einschalten, werden im Diagramm zwei Platzhalter für die Achsenbeschriftungen eingefügt.

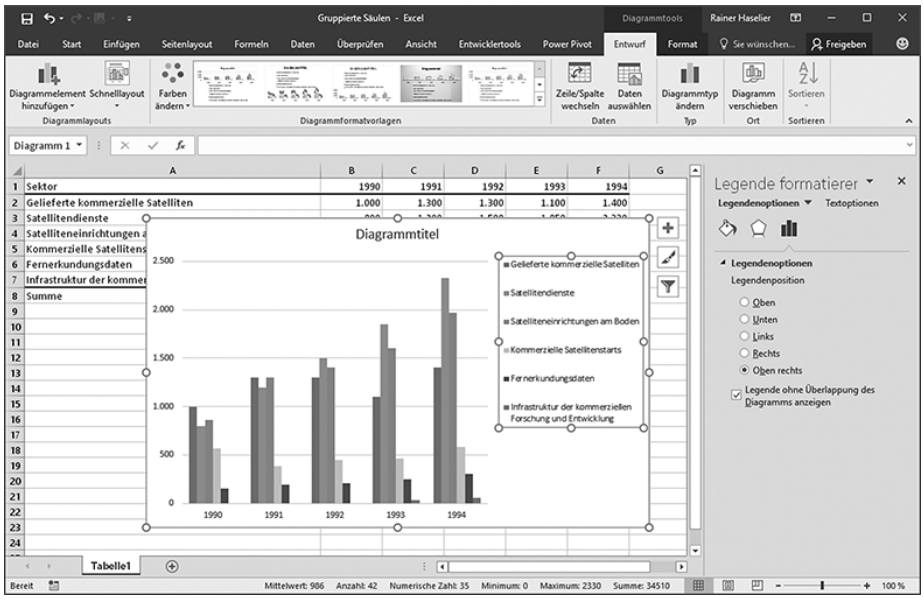


Abbildung 3.5: Der Aufgabenbereich LEGENDE FORMATIEREN

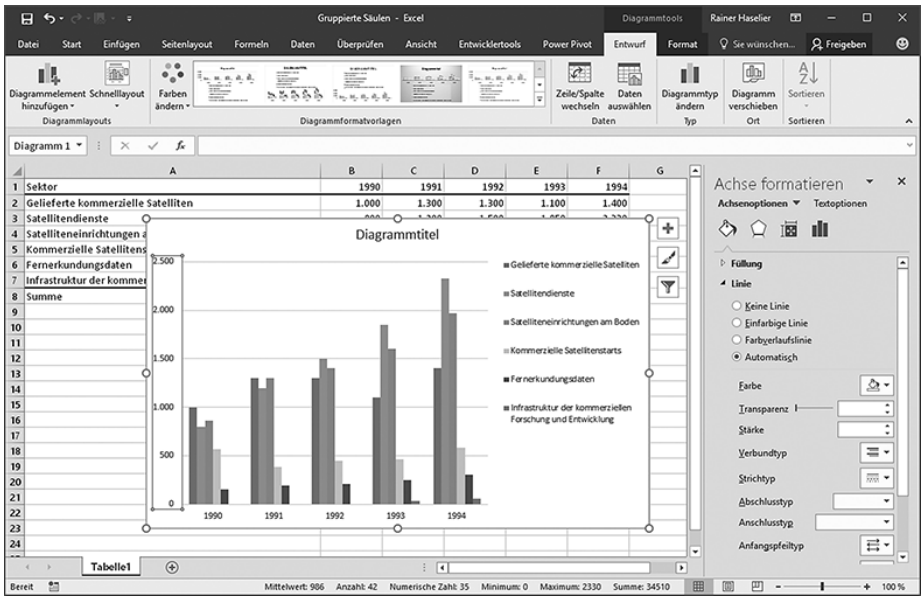


Abbildung 3.6: Der Aufgabenbereich ACHSE FORMATIEREN mit dem Abschnitt LINIE

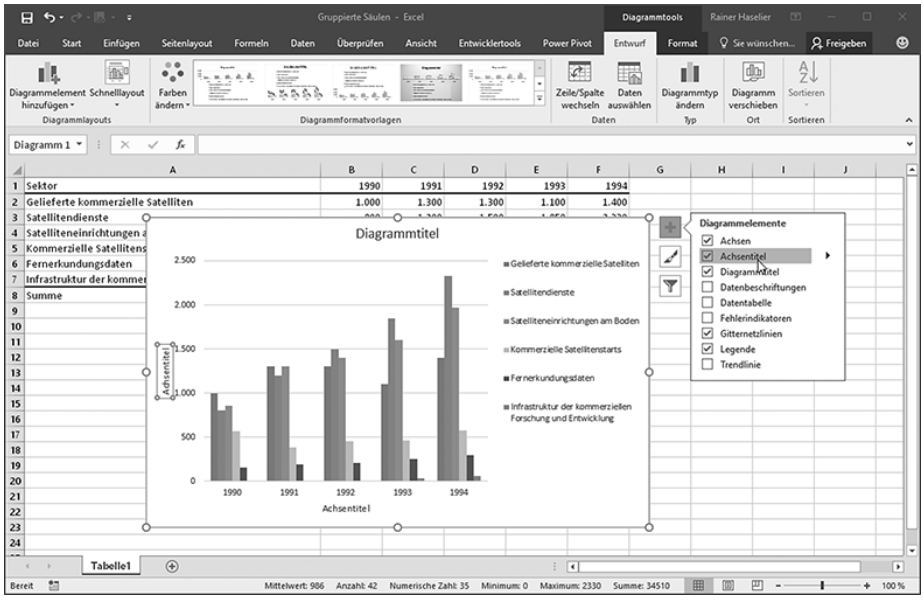


Abbildung 3.7: Das Menü der Schaltfläche DIAGRAMMELEMENTE mit aktiviertem Kontrollkästchen ACHSENTITEL

Verwenden Sie die Titeltextfelder, um die Titel zu ergänzen und um das Diagramm so fertig zu stellen, wie in Abbildung 3.1 gezeigt.

Um Diagrammelemente einzufügen, können Sie auch die Schaltfläche DIAGRAMMELEMENTE HINZUFÜGEN verwenden, die sich ganz links auf der Registerkarte DIAGRAMMTOOLS/ENTWURF befindet.

Ein schnelles Verfahren, um einen Diagrammtitel einzufügen, besteht darin, den Titel anzuklicken und dann einen neuen Titel einzugeben.

Wenn Sie einen Titel ergänzen (gleichgültig, ob für eine Achse oder das Diagramm), können Sie entweder das Textfeld markieren und dann den Text eingeben oder den gesamten Titeltext markieren und dann mit dem Tippen beginnen. Wenn Sie tippen, ohne zu markieren, erscheint der neue Titel in der Bearbeitungsleiste und wird erst dann in den Titelbereich eingefügt, nachdem Sie Enter gedrückt haben. Wenn Sie den bestehenden Text vorher markieren, erscheint der neue Text im Titelbereich, während Sie ihn eingeben.

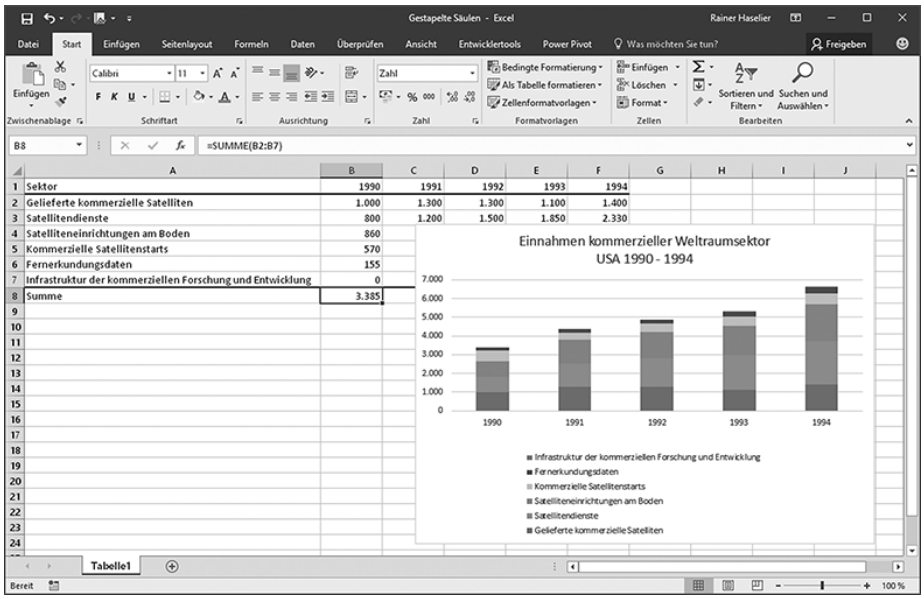
Sie haben verschiedene Möglichkeiten, um sich das Diagramm in einer Vorschau anzusehen. Wenn Sie im Dialogfeld DIAGRAMM EINFÜGEN eines der empfohlenen Diagramme anklicken, sehen Sie eine Vorschau des Diagramms mit Ihren Daten. Die Vorschau wird direkt im Dialogfeld angezeigt. Nachdem Sie das Diagramm erstellt und eingefügt haben, können Sie es markieren und dann auf der Registerkarte DIAGRAMMTOOLS/ENTWURF den Katalog mit dem DIAGRAMMFORMATVORLAGEN öffnen. Sie sehen dann weitere Miniaturansichten der Änderungen, die Sie hier vornehmen können.

Die drei Schaltflächen neben dem Diagramm enthalten viele weitere nützliche Tools. Die Schaltfläche mit dem Pinsel stellt

eine Vielzahl von Formatvorlagen und Farbschemata für Ihr Diagramm bereit. Die Schaltfläche mit dem Filtersymbol erlaubt es Ihnen, bestimmte Elemente aus dem Diagramm zu entfernen und stellt einen schnellen Weg bereit, um das Dialogfeld DATENQUELLE AUSWÄHLEN zu öffnen (das ich weiter hinten in diesem Kapitel im Abschnitt »Eine Linie ziehen« beschreibe).

Säulen stapeln

Hätte ich das siebte der von Excel empfohlenen Diagrammtypen ausgewählt, wären Säulen erstellt worden, die dieselben Informationen ein bisschen anders darstellen. Dieser Diagrammtyp wird *Gestapelte Säulen*genannt. Hierbei stellt jede Säule die Gesamtheit aller Datenreihen an einem Punkt auf der x-Achse dar. Jede Säule ist in Segmente unterteilt, wobei sich die Größe der einzelnen Segmente jeweils nach dem Anteil am Gesamten richtet. Abbildung 3.8 zeigt, was ich meine.



	A	B	C	D	E	F
1	Sektor	1990	1991	1992	1993	1994
2	Gelieferte kommerzielle Satelliten	1.000	1.300	1.300	1.100	1.400
3	Satellitendienste	800	1.200	1.500	1.850	2.330
4	Satelliteneinrichtungen am Boden	860				
5	Kommerzielle Satellitenstarts	570				
6	Fernerkundungsdaten	155				
7	Infrastruktur der kommerziellen Forschung und Entwicklung	0				
8	Summe	3.385				

Abbildung 3.8: Die Daten aus Tabelle 3.1 in einem Diagramm mit gestapelten Säulen

Bisher habe ich die Diagramme in das Tabellenblatt eingefügt. Excel erlaubt es, das Diagramm auf ein eigenes Blatt innerhalb der Arbeitsmappe zu verschieben. Wählen Sie dazu ENTWURF | DIAGRAMM

VERSCHIEBEN (die Schaltfläche befindet sich weit rechts auf der Registerkarte DIAGRAMMTOOLS/ENTWURF), um das Dialogfeld DIAGRAMM VERSCHIEBEN zu öffnen. Schalten Sie das Optionsfeld NEUES BLATT ein, um ein neues Blatt hinzuzufügen und das Diagramm dorthin zu verschieben. Abbildung 3.9 zeigt, wie das Diagramm auf einem eigenen Blatt aussieht. Wie Sie in Abbildung 3.9 sehen können, habe ich die Legende an eine andere Position gesetzt.

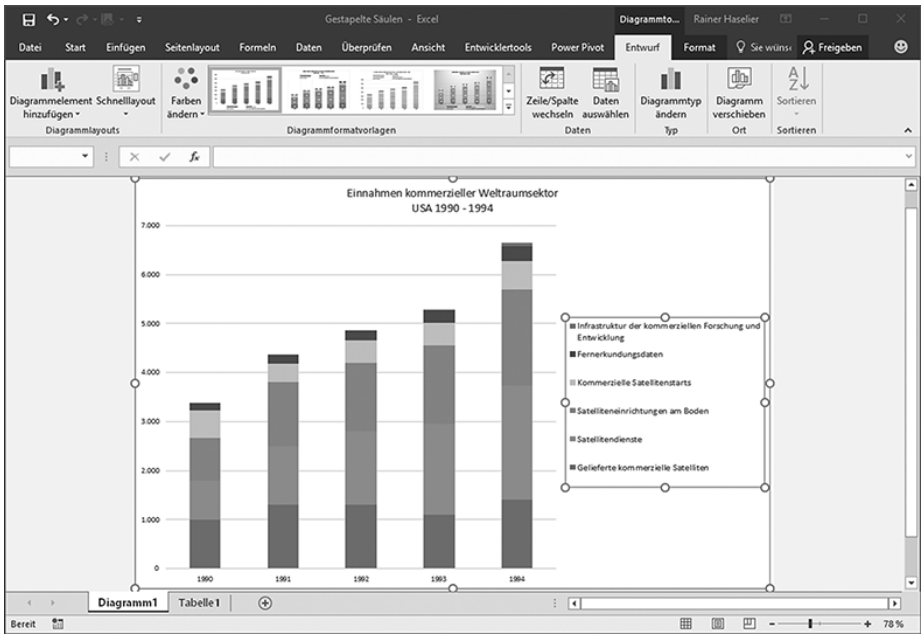


Abbildung 3.9: Das gespaltete Säulendiagramm auf einem eigenen Diagrammblatt

Auf diese Art können schön Änderungen in der Verteilung der Einzelwerte im Laufe der Zeit dargestellt werden. Wenn Sie nur die Verteilung in einem Jahr darstellen möchten, ist der folgende Diagrammtyp besser geeignet.

Segmente eines Kreisdiagramms

Schauen wir uns den nächsten Diagrammtyp an. Wenn Sie die Verteilung von Einzelwerten im Verhältnis zum Gesamtwert darstellen möchten, verwenden Sie am besten ein Kreisdiagramm.

Nehmen wir an, Sie möchten die Einnahmen der USA aus dem kommerziellen Weltraumsektor im Jahr 1994, also die letzte Spalte in Tabelle 3.1, darstellen. Sie werden die Aufmerksamkeit des Betrachters wecken, wenn Sie die Daten in Form eines Kreisdiagramms wie das in Abbildung 3.10 darstellen.

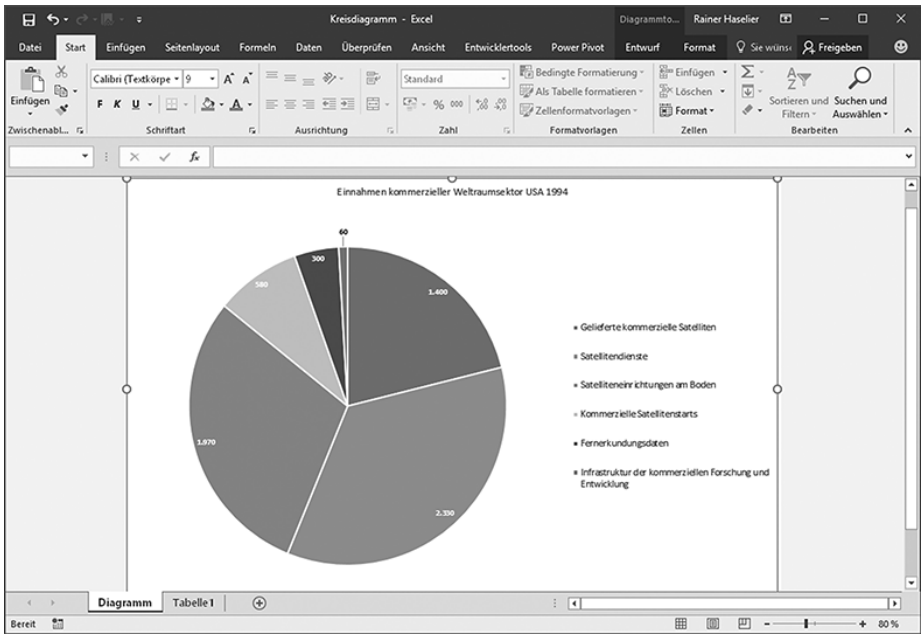


Abbildung 3.10: Die Daten aus Tabelle 3.1 in einem aus der letzten Spalte erstellten Kreisdiagramm

Und so erstellen Sie dieses Diagramm:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

   Dies ist ganz einfach. Ich habe dies bereits gemacht.

2. **Markieren Sie die Daten, die Sie im Diagramm darstellen wollen.**

Ich möchte die Namen aus Spalte A und die Daten aus Spalte F verwenden. Der Trick besteht darin, zuerst Spalte A (die Zellen A2 bis A7) auf die übliche Art zu markieren und dann die Taste Strg gedrückt zu halten. Während Sie diese Taste gedrückt halten, markieren Sie mit der Maus die Zellen F2 bis F7. Fertig! Sie haben nun Zellen markiert, die nicht direkt nebeneinander stehen. Excel nennt dies Mehrfachauswahl.

3. **Klicken Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME auf EMPFOHLENE DIAGRAMME und wählen Sie in der Liste auf der linken Seite des Dialogfelds das Kreisdiagramm aus.**

4. **Bearbeiten Sie das Diagramm.**

   Abbildung 3.11 zeigt das anfängliche Kreisdiagramm (nachdem ich den Titel ergänzt habe). Damit das Diagramm so aussieht wie in Abbildung 3.10, musste ich ein paar Änderungen vornehmen. Zuerst habe ich die Legende wie in Abbildung 3.10 gezeigt formatiert.

   Die Zahlen innerhalb der Kreissegmente werden Datenbeschriftungen genannt. Um diese hinzuzufügen, habe ich das Diagramm markiert (nicht nur ein Segment), die Schaltfläche DIAGRAMMELEMENTE angeklickt und dann im Popupmenü das Kontrollkästchen DATENBESCHRIFTUNGEN eingeschaltet.

   Um die Schriftfarbe der Beschriftungen zu ändern, klicken Sie eine der Datenbeschriftungen an und klicken dann im Aufgabenbereich DATENBESCHRIFTUNGEN FORMATIEREN auf TEXTOPTIONEN. Schalten Sie das Optionsfeld EINFARBIGE FÜLLUNG ein und ändern Sie die Farbe von Schwarz in Weiß ab. Drücken Sie Strg+F, um den Text fett zu formatieren. Die Datenbeschriftung, die sich außerhalb des Segments befindet (60), können Sie getrennt markieren und die Schriftfarbe wieder auf Schwarz setzen.

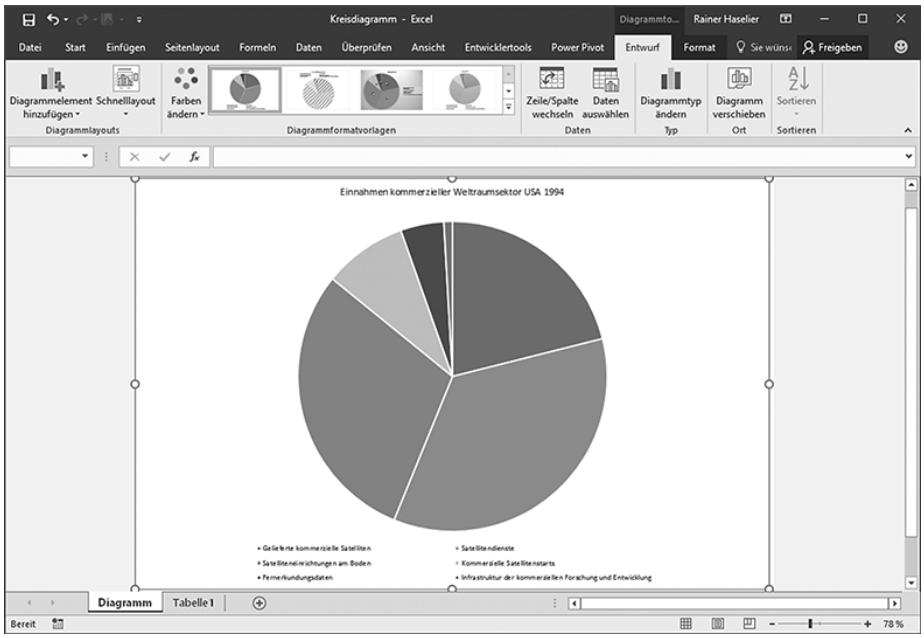


Abbildung 3.11: Das anfängliche Kreisdiagramm auf einem eigenen Diagrammblatt

Denken Sie beim Erstellen eines Kreisdiagramms immer an Folgendes:

Der Rat des Weisen

Der letzte große Gesellschaftskritiker, Anekdotenerzähler und ehemalige Baseballspieler Yogi Berra bestellte einmal in einer Pizzeria eine Pizza.

»In wie viele Stücke soll ich sie schneiden«, fragte die Bedienung, »in vier oder in acht?«

»Lieber in vier«, sagte Yogi. »Ich weiß nicht, ob ich acht schaffe.«

Yogis aufschlussreiche Analyse führt zu einem hilfreichen Lehrsatz für Kreisdiagramme. Sie lassen sich besser »verdauen«, wenn sie aus wenigen Segmenten bestehen. Wenn Sie ein Kreisdiagramm in zu schmale Segmente zergliedern, werden Sie Ihre Zuhörer mit Informationen überladen.

Wenn Sie ein Diagramm für eine Präsentation erstellen (beispielsweise für PowerPoint), sollten Sie immer Datenbeschriftungen verwenden. Sie verdeutlichen Ihren Zuhörern wichtige Punkte und Trends.

(Diese Yogi-Anekdote taucht auch in den vorherigen drei Ausgaben dieses Buches auf. Hat sich die Geschichte wirklich ereignet? Wir wissen es nicht sicher. Ein berühmtes Zitat von Mr. Berra lautet: »Die Hälfte der Lügen, die über mich erzählt werden, ist nicht wahr.«)

Eine Linie ziehen

Im vorhergehenden Beispiel habe ich mich auf eine Spalte mit Daten aus Tabelle 3.1 konzentriert. Bei diesem Beispiel hier verwende ich die Daten einer Zeile. Dabei geht es darum, die Entwicklung eines Weltraumsektors im Verlauf der Jahre 1990 bis 1994 zu veranschaulichen. In diesem Beispiel habe ich die Einnahmen aus den Satellitendiensten in einem Diagramm dargestellt. Das Endergebnis, auf einem eigenen Diagrammblatt dargestellt, sehen Sie in Abbildung 3.12.

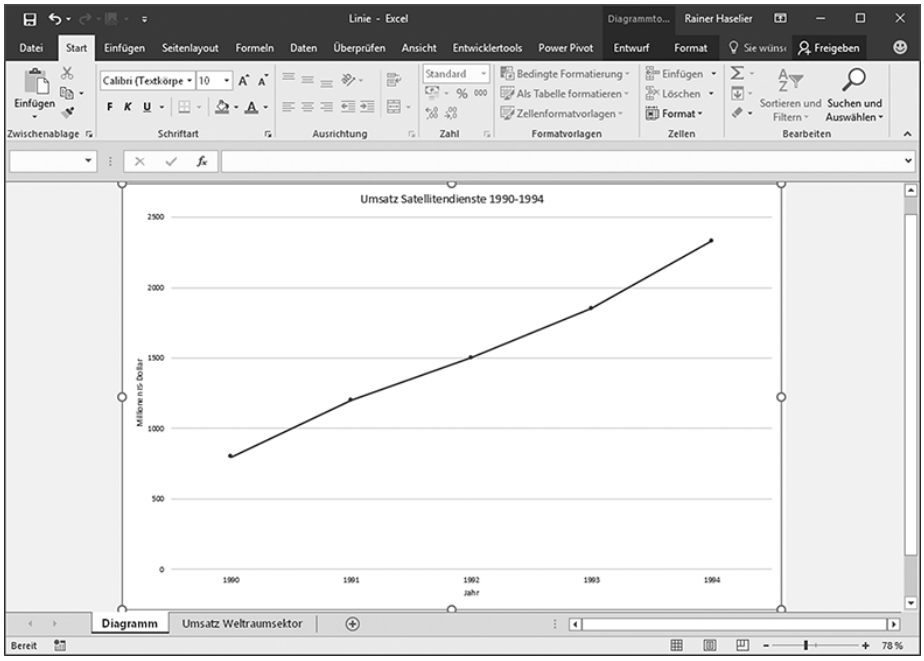


Abbildung 3.12: Die Daten aus der zweiten Zeile aus Tabelle 3.1 in einem Liniendiagramm mit Datenpunkten

Ein Liniendiagramm ist eine gute Möglichkeit, Veränderungen in einem Zeitraum darzustellen, solange nicht zu viele Datenreihen betrachtet werden. Wenn Sie alle sechs Sektoren in einem Liniendiagramm darstellen, sieht das Ganze eher wie ein Teller Spaghetti aus.

Und wie erstellen Sie nun ein Diagramm wie das in Abbildung 3.12? Ganz einfach:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

   Auch dies habe ich bereits erledigt.

2. **Markieren Sie die Daten, die das Diagramm darstellen soll.**

   In diesem Beispiel sind dies die Zellen A3 bis F3. Ja, die Beschriftung habe ich ebenfalls markiert.

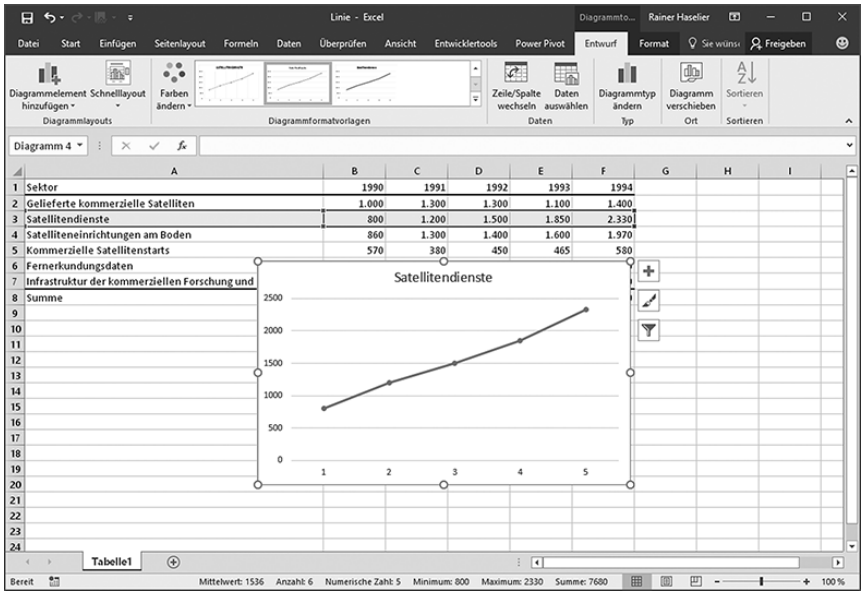
   Hmm. Habe ich nicht etwas vergessen? Was ist mit dem kleinen Trick, den ich Ihnen etwas weiter vorne verraten habe, bei dem Sie die Taste Strg gedrückt halten, um weitere Zellen auszuwählen? Hätte ich dies hier nicht ebenfalls tun können und die oberste Zeile mit den Jahren markieren sollen?

   Nein. Diesmal nicht. Würde ich dies tun, ginge Excel davon aus, dass es sich bei 1990, 1991, 1992, 1993 und 1994 um eine weitere Datenreihe handelt, die im Diagramm dargestellt werden soll. Ich zeige Ihnen gleich ein anderes Verfahren, mit dem Sie die Jahreszahlen in die x-Achse bekommen.

3. **Klicken Sie auf EINFÜGEN | EMPFOHLENE DIAGRAMME und wählen Sie einen Diagrammtyp aus.**

   Dieses Mal habe ich die Registerkarte ALLE DIAGRAMME geöffnet, in der Liste auf der linken Seite auf LINIE geklickt und dann den

Diagrammtyp LINIE MIT DATENPUNKTEN ausgewählt. Das Ergebnis sehen Sie in Abbildung 3.13.



	A	B	C	D	E	F
1	Sektor	1990	1991	1992	1993	1994
2	Gelieferte kommerzielle Satelliten	1.000	1.300	1.300	1.100	1.400
3	Satellitendienste	800	1.200	1.500	1.850	2.330
4	Satelliteneinrichtungen am Boden	860	1.300	1.400	1.600	1.970
5	Kommerzielle Satellitenstarts	570	380	450	465	580
6	Fernerkundungsdaten					
7	Infrastruktur der kommerziellen Forschung und					
8	Summe					

Abbildung 3.13: Das Ergebnis, nachdem auf der Registerkarte ALLE DIAGRAMME der Diagrammtyp LINIE MIT DATENPUNKTEN ausgewählt wurde.

4. **Bearbeiten Sie das Diagramm.**

   Möglicherweise ist die Linie im Diagramm nur schwer zu erkennen. Wenn Sie die Linie anklicken und dann ENTWURF | FARBEN ÄNDERN wählen, sehen Sie die Auswahlmöglichkeiten für die Linienfarbe. Ich habe Schwarz gewählt.

   Anschließend habe ich die Titel für die Achsen und das Diagramm ergänzt. Der einfachste Weg, um den Titel zu ändern (für den Excel die Beschriftung verwendet, die ich gemeinsam mit den Daten ausgewählt habe), ist, den Titel anzuklicken und die Änderung einzugeben.

   Um die Achsentitel zu ergänzen, habe ich die Schaltfläche DIAGRAMMELEMENTE angeklickt (die mit dem Pluszeichen), und dann im Popupmenü das Kontrollkästchen ACHSENTITEL eingeschaltet (siehe auch Abbildung 3.7). Anschließend habe ich die Achsentitel angeklickt, den Text ausgewählt und den neuen Titel eingegeben.

Jetzt fehlen auf der x-Achse noch die Jahreszahlen. Um diese zu ergänzen, klicke ich das Diagramm mit der rechten Maustaste an, um das Kontextmenü aus Abbildung 3.14 zu öffnen.

Abbildung 3.14: Wenn Sie das Diagramm mit der rechten Maustaste anklicken, wird dieses Kontextmenü geöffnet.

Klicken Sie in diesem Kontextmenü auf DATEN AUSWÄHLEN, um das Dialogfeld DATENQUELLE AUSWÄHLEN zu öffnen (siehe Abbildung 3.15). Klicken Sie im Bereich HORIZONTALE ACHSENBESCHRIFTUNGEN (RUBRIK) auf BEARBEITEN, um das Dialogfeld ACHSENBESCHRIFTUNGEN zu öffnen (siehe Abbildung 3.16). Die blinkende Einfügemarke im Feld ACHSENBESCHRIFTUNGSBEREICH zeigt an, dass Sie mit der Eingabe loslegen können. Die Auswahl der Zellen B1 bis F1 und ein Klick auf OK legen den Bereich fest und schließen das Dialogfeld. Ein Klick auf OK im Dialogfeld DATENQUELLE AUSWÄHLEN schließt auch dieses Dialogfeld und fügt die Jahreszahlen an der x-Achse ein.

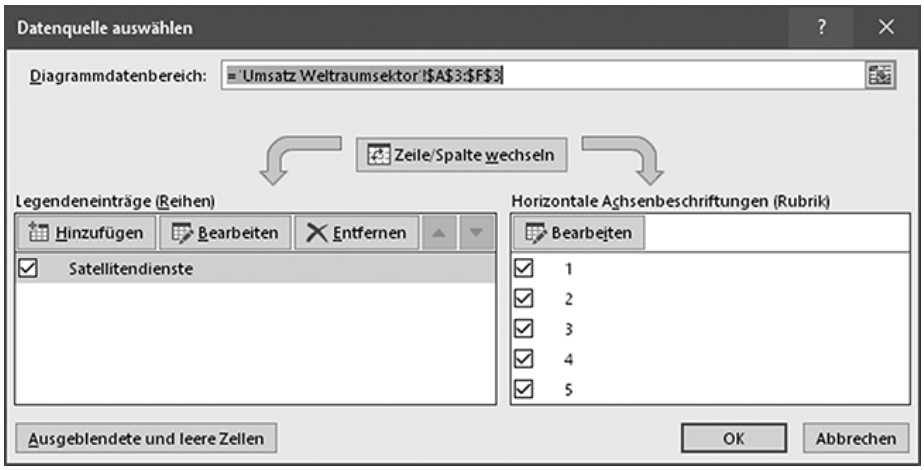


Abbildung 3.15: Das Dialogfeld DATENQUELLE AUSWÄHLEN

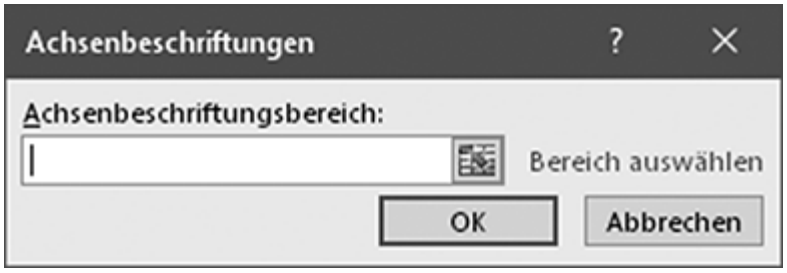


Abbildung 3.16: Das Dialogfeld ACHSENBESCHRIFTUNGEN

Minidiagramme mit Sparklines

Edward Tufte (der als »da Vinci der Daten« gilt), ist der geistige Vater der Sparklines. Dies sind kleine Diagramme, die Sie in Text oder eine Tabelle einfügen können, um schnell einen Trend zu illustrieren. Sie sind so groß, dass sie ungefähr den Platz eines Wortes einnehmen; daher werden Sparklines von Tufte auch »Wortgrafiken« genannt.

Es stehen drei Arten von Sparklines zur Verfügung: Ein Liniendiagramm, ein Säulendiagramm sowie ein dritter Typ, an dem vor allem Sportinteressierte ihre Freude haben. Er zeigt Gewinn und Verlust.

Um Ihnen zu zeigen, wie Sparklines aussehen, werde ich die ersten beiden Typen in die Daten aus Tabelle 3.1 einfügen. Zuerst habe ich zwischen den Spalten A und B zwei weitere Spalten eingefügt und

dann Zelle B2 in der neuen Spalte markiert. Mit dem Befehl EINFÜGEN | SPARKLINES | LINIE habe ich das Dialogfeld SPARKLINES ERSTELLEN geöffnet (siehe Abbildung 3.17).

Abbildung 3.17: Das Dialogfeld SPARKLINES ERSTELLEN

In das Feld DATENBEREICH habe ich D2:H2 eingegeben und dann auf OK geklickt. Die weiteren Zellen der Spalte wurden automatisch ausgefüllt. Für Spalte C habe ich diese Schritte wiederholt, jedoch im Menüband den Befehl EINFÜGEN | SPARKLINES | SÄULE gewählt. Das Ergebnis sehen Sie in Abbildung 3.18.

	A	B	C	D	E	F	G	H
1	Sektor			1990	1991	1992	1993	1994
2	Gelieferte kommerzielle Satelliten			1.000	1.300	1.300	1.100	1.400
3	Satellitendienste			800	1.200	1.500	1.850	2.330
4	Satelliteneinrichtungen am Boden			860	1.300	1.400	1.600	1.970
5	Kommerzielle Satellitenstarts			570	380	450	465	580
6	Fernerkundungsdaten			155	190	210	250	300
7	Infrastruktur der kommerziellen Forschung und Entwicklung			0	0	0	30	60
8	Summe			3.385	4.370	4.860	5.295	6.640
9								

Abbildung 3.18: Linien- und Säulen-Sparklines für die Daten aus Tabelle 3.1

Falls Sie in einer Präsentation unbedingt eine Tabelle zeigen wollen, dann sind Sparklines eine perfekte Ergänzung. Wenn ich diese Beispieldaten präsentieren müsste, würde ich den Sparklinetyp Säule verwenden.

Welche anderen Möglichkeiten gibt es, Sparklines zu verwenden? Abbildung 3.19 zeigt zwei Säulen-Sparklines, die in ein Word-Dokument eingefügt wurden. Es ist nicht ganz einfach, die Minidiagramme richtig zu kopieren und einzufügen; außerdem

müssen Sie die Sparklines als Grafik einfügen. Ich vermute, Sie stimmen mit mir überein, dass sich der Aufwand lohnt.

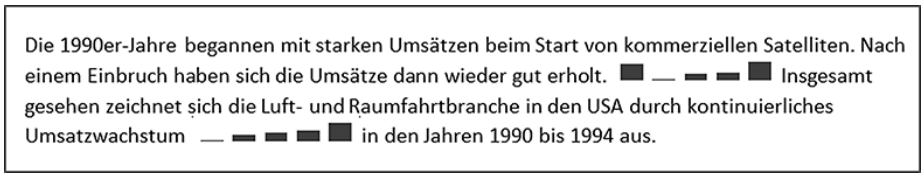


Abbildung 3.19: Sparklines in einem Word-Dokument

Mit dem Sparklinetyp *Gewinn/Verlust* lässt sich gut der Fortschritt einer Sportmannschaft innerhalb einer Spielsaison veranschaulichen. Die Sparklines in Abbildung 3.20 stellen die Monatswertung der Teams in der amerikanischen nationalen Baseball-Liga East Division dar.

	A	B	C	D	E	F	G	H	I
1	National League Eastern Division	Sparkline	April	Mai	Juni	Juli	August	September	
2	New York Mets		1	-1	-1	1	1	1	
3	Washington Nationals		-1	1	1	-1	-1	0	
4	Miami Marlins		-1	-1	-1	-1	-1	1	
5	Atlanta Braves		-1	1	-1	-1	-1	-1	
6	Philadelphia Phillies		-1	-1	-1	1	-1	-1	
7									

Abbildung 3.20: Gewinne und Verluste in der amerikanischen nationalen Baseball-Liga East Division

Eine 1 in den Daten gibt an, dass das Team mehr Spiele gewonnen als verloren hat, –1 gibt an, dass mehr Spiele verloren als gewonnen wurden und die 0 steht für eine gleiche Anzahl von gewonnenen und verlorenen Spielen. Ich habe den Oktober nicht angegeben, da jede Mannschaft in diesem Monat nur vier reguläre Spiele bestritten hat. (In den Monaten von April bis September stand jede Mannschaft für zwischen 22 und 28 Spiele auf dem Platz.) In den Sparklines wird ein gewonnener Monat als Kästchen oberhalb der Linie und ein verlorener Monat als Kästchen unterhalb der Linie angezeigt. Monate mit der gleichen Anzahl von gewonnenen und verlorenen Spielen bleiben leer.

Beachten Sie, dass die New York Mets, der Divisionssieger (und der Meister der National League), die einzige Mannschaft waren, die in drei aufeinanderfolgenden Monaten mehr Spiele gewonnen als

verloren haben. (Ja ich weiß, sie haben die World Series verloren. Bitte erinnern Sie mich nicht daran.)

Verwenden Sie nicht das Standardverfahren, um eine Sparkline zu löschen. Klicken Sie die Sparkline stattdessen mit der rechten Maustaste an und klicken Sie im Kontextmenü auf SPARKLINES. Sie sehen dann die Befehle, mit denen Sie Sparklines und Sparklinegruppen löschen können.

Balken über Balken

Das Balkendiagramm von Excel ist ein auf die Seite gelegtes Säulendiagramm. Hierbei wird die Horizontal-vertikal-Regel umgekehrt. Die vertikale Achse ist die x-Achse, an der die unabhängigen Variablen abgetragen werden. Die horizontale Achse ist die y-Achse, an der die abhängigen Variablen abgetragen werden.

Wann wird das Balkendiagramm verwendet? Dieser Diagrammtyp eignet sich gut, wenn Sie auf das Erreichen eines Ziels oder auf die Ungerechtigkeit beim Erreichen des Ziels hinweisen möchten.

In Tabelle 3.2 sind die Daten für die private Internetnutzung aufgeführt. Die Daten stammen aus dem statistischen Jahrbuch der Vereinigten Staaten (Statistical Abstract of the U.S.) vom US-amerikanischen Statistischen Bundesamt (Bureau of Census) und wurden für das Jahr 2013 erhoben. *Prozent* gibt den Anteil der Personen in jeder Einkommensgruppe an.

Haushaltseinkommen	Prozent
Unter $ 25.000	48,4
$ 25.000 bis $ 59.999	69
$ 50.000 bis $ 99.999	84,9

Haushaltseinkommen	Prozent
$ 100.000 bis $ 149.999	92,7
Über $ 150.000	94,9

Tabelle 3.2: Nutzung des Internets zu Hause (2013)

Die Zahlen in der Tabelle zeigen einen deutlichen Trend. In einem Balkendiagramm wird dieser Trend noch deutlicher, wie Sie in Abbildung 3.21 sehen können.

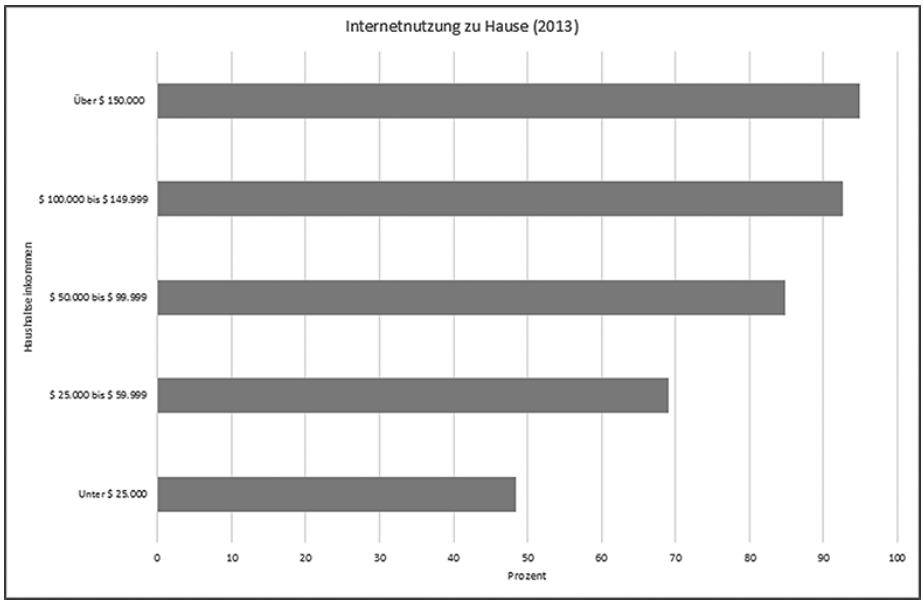


Abbildung 3.21: Die Daten aus Tabelle 3.2 in einem Balkendiagramm

Um dieses Diagramm zu erstellen, führen Sie diese Schritte durch:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

   Abbildung 3.22 zeigt, wie dieses Arbeitsblatt aussieht.

	A	B	C
1	Haushaltseinkommen	Prozent	
2	Unter $ 25.000	48,4	
3	$ 25.000 bis $ 59.999	69	
4	$ 50.000 bis $ 99.999	84,9	
5	$ 100.000 bis $ 149.999	92,7	
6	Über $ 150.000	94,9	
7			

Abbildung 3.22: Die Daten aus Tabelle 3.2 wurden in ein Arbeitsblatt eingegeben.

2. **Markieren Sie die Daten, die im Diagramm dargestellt werden sollen.**

   In diesem Beispiel befinden sich die Daten im Zellbereich A1:B8.

3. **Klicken Sie auf EINFÜGEN | EMPFOHLENE DIAGRAMME und wählen Sie in der Liste auf der linken Seite den Diagrammtyp aus, der Ihnen gefällt.**

   Ich habe die erste Option gewählt: GRUPPIERTE BALKEN. Abbildung 3.23 zeigt das Ergebnis.

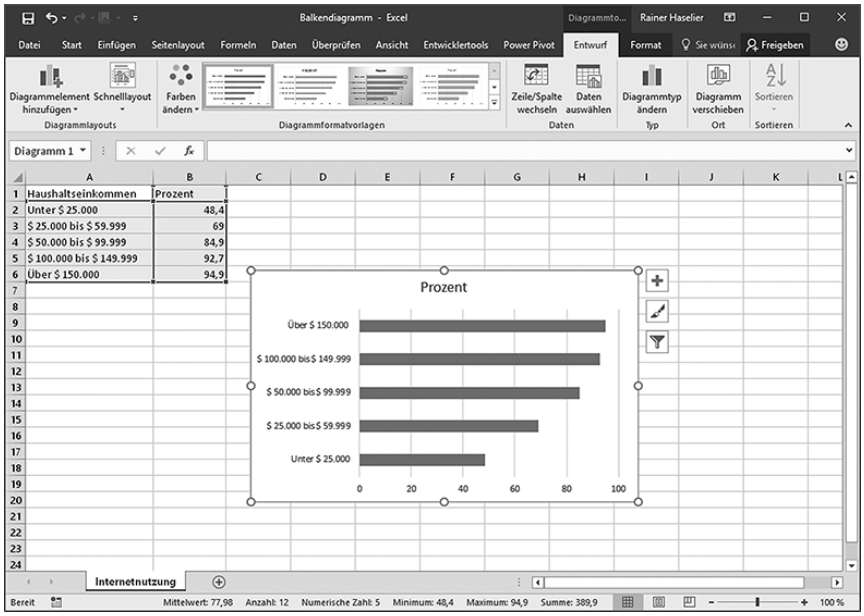


Abbildung 3.23: Das anfängliche Excel-Balkendiagramm

4. **Bearbeiten Sie das Diagramm.**

   Die erste Bearbeitung besteht darin, den Titel zu ändern. Sie können den vorhandenen Titel anklicken und dann den neuen Text eintippen. Anschließend habe ich die Achsenbeschriftungen ergänzt. Hierfür habe ich die Schaltfläche DIAGRAMMELEMENTE angeklickt (die mit dem Pluszeichen) und dann im Popupmenü das Kontrollkästchen ACHSENTITEL eingeschaltet. Excel fügt Platzhalter in das Diagramm ein, die ich geändert habe. Zum guten Schluss habe ich sowohl die Achsentitel als auch die Zahlen fett formatiert. Der einfachste Weg, um dies zu tun, ist, ein Element zu markieren und dann Strg + F zu drücken.

Punkt, Punkt, Komma, Strich …

Zum Ermitteln der Beziehung zwischen einer Variablen *x* und einer anderen Variablen *y* verwenden Sie eine wichtige Technik der Statistik, die als *lineare Regression* bezeichnet wird. Ausführlichere Informationen zur linearen Regression finden Sie in Kapitel 14.

Diese Technik basiert auf einem Diagramm, in dem für einzelne Elemente sowohl *x* als auch *y* gemessen und dargestellt wird. Im Diagramm ist jedes Element als ein Punkt dargestellt. Da dieses Diagramm aus vielen Punkten besteht, wird es in Excel als Punktdiagramm, sonst auch als *Streudiagramm* bezeichnet.

Nehmen wir einmal an, Sie möchten herausfinden, wie gut ein Verkaufstalenttest die Produktivität von Verkäufern vorhersagt. Sie führen den Test an einer Stichprobe von Verkäufern durch und erfassen, wie viel Geld diese durch Provisionen in einem Zeitraum von zwei Monaten verdienen. Jede Person positioniert sich durch ihre Wertepaare (Talentwert und Provision) im Punktdiagramm.

Um ein Punktdiagramm zu erstellen, gehen Sie wie folgt vor:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

   Abbildung 3.24 zeigt die eingegebenen Daten.

2. **Markieren Sie die Daten, die im Diagramm dargestellt werden sollen.**

   Im Hintergrund von Abbildung 3.25 sehen Sie den markierten Zellbereich B2 bis C21. (Wenn Sie B1 mit einschließen, erhalten Sie das gleiche Diagramm, dieses hat jedoch einen falschen Titel.) Die Zellen in Spalte A sind lediglich Platzhalter, die die Daten strukturieren.

3. **Klicken Sie auf EINFÜGEN | EMPFOHLENE DIAGRAMME und wählen Sie in der Liste auf der linken Seite den Diagrammtyp aus.**

   Ich habe hier die erste Option gewählt und erhalte so das Diagramm aus Abbildung 3.25.

	A	B	C	D
1	Vertreter	Talentpunkte	Provision (Tausend Euro)	
2	1	54	65	
3	2	34	45	
4	3	23	39	
5	4	46	36	
6	5	34	45	
7	6	47	50	
8	7	51	76	
9	8	54	62	
10	9	35	49	
11	10	45	50	
12	11	40	48	
13	12	39	48	
14	13	35	47	
15	14	56	61	
16	15	22	38	
17	16	58	64	
18	17	42	54	
19	18	33	50	
20	19	39	45	
21	20	50	62	
22				

Abbildung 3.24: Die Daten für das Punktdiagramm

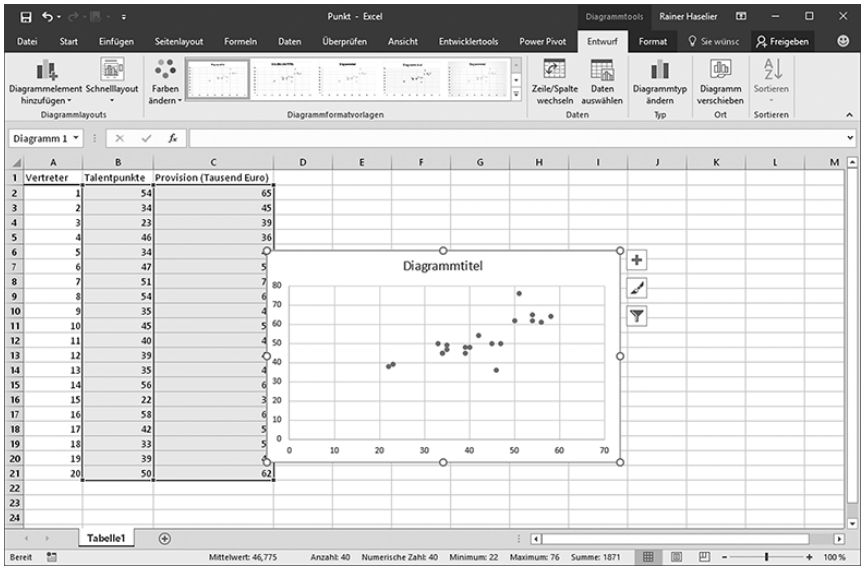
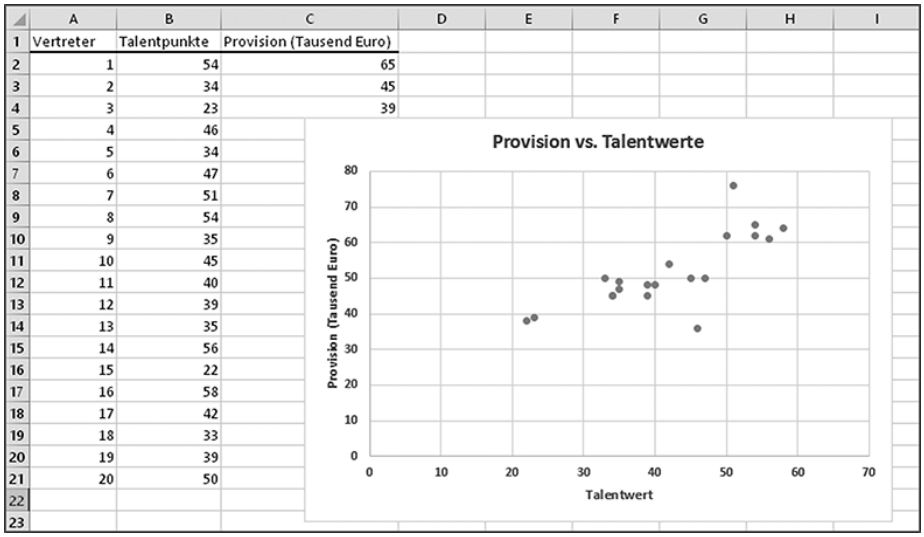


Abbildung 3.25: Die anfängliche Version des Punktdiagramms

4. Bearbeiten Sie das Diagramm.

Ich habe den allgemeinen Diagrammtitel angeklickt und einen neuen Titel eingegeben. Anschließend habe ich die Schaltfläche DIAGRAMMELEMENTE angeklickt (die mit dem Pluszeichen), im Popupmenü das Kontrollkästchen ACHSENTITEL eingeschaltet und den Text für die Achsentitel angepasst. Excel fügt Platzhalter in das Diagramm ein, die ich geändert habe. Zum guten Schluss habe ich die Achsentitel und den Diagrammtitel fett formatiert, indem ich Strg+F gedrückt habe. Das Ergebnis sehen Sie in Abbildung 3.26.



	A	B	C
1	Vertreter	Talentpunkte	Provision (Tausend Euro)
2	1	54	65
3	2	34	45
4	3	23	39
5	4	46	
6	5	34	
7	6	47	
8	7	51	
9	8	54	
10	9	35	
11	10	45	
12	11	40	
13	12	39	
14	13	35	
15	14	56	
16	15	22	
17	16	58	
18	17	42	
19	18	33	
20	19	39	
21	20	50	
22			
23			

Abbildung 3.26: Das fast fertige Punktdiagramm

Abbildung 3.27: Wenn Sie mit der rechten Maustaste auf einen Punkt im Punktdiagramm klicken, wird dieses Kontextmenü angezeigt.

Wenn Sie auf TRENDLINIE HINZUFÜGEN klicken, wird der Aufgabenbereich TRENDLINIE FORMATIEREN angezeigt (siehe Abbildung 3.28). Ich habe die Option LINEAR eingeschaltet (dies ist der Standardwert) und im unteren Bereich die beiden Kontrollkästchen aktiviert. (Eventuell müssen Sie ein wenig nach unten scrollen, um sie zu sehen.) Diese sind mit FORMEL IM DIAGRAMM ANZEIGEN und BESTIMMTHEITSMASS IM DIAGRAMM DARSTELLEN beschriftet.

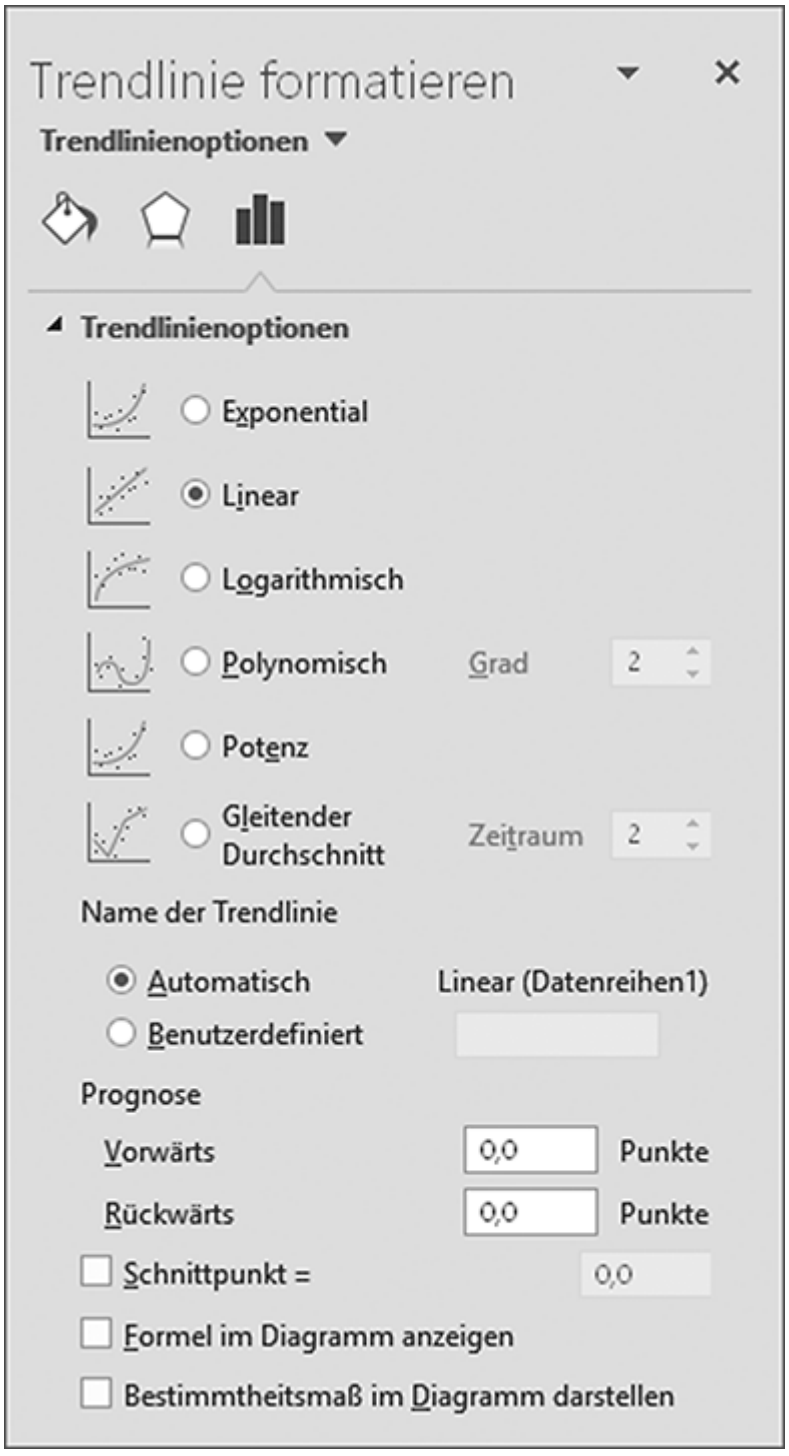
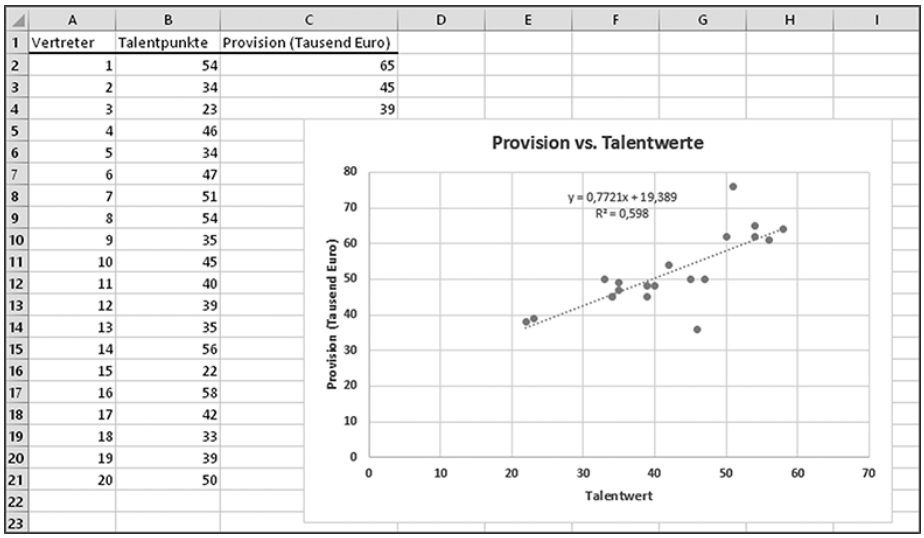


Abbildung 3.28: Der Aufgabenbereich TRENDLINIE FORMATIEREN

Klicken Sie auf SCHLIESSEN, um den Aufgabenbereich TRENDLINIE FORMATIEREN zu schließen. Nun befinden sich einige zusätzliche Elemente im Punktdiagramm (siehe Abbildung 3.29). Die Punkte

sind nun durch eine Linie miteinander verbunden. Bei Excel wird diese als *Trendlinie* bezeichnet, es handelt sich aber eigentlich um eine *Regressionslinie*. Außerdem werden Gleichungen angezeigt. (Ich habe diese übrigens aus der ursprünglichen Position verschoben.) Was hat das alles zu bedeuten? Was bedeuten nur all diese Zahlen?

Um das herauszufinden, müssen Sie Kapitel 14 lesen.



	A	B	C
1	Vertreter	Talentpunkte	Provision (Tausend Euro)
2	1	54	65
3	2	34	45
4	3	23	39
5	4	46	
6	5	34	
7	6	47	
8	7	51	
9	8	54	
10	9	35	
11	10	45	
12	11	40	
13	12	39	
14	13	35	
15	14	56	
16	15	22	
17	16	58	
18	17	42	
19	18	33	
20	19	39	
21	20	50	
22			
23			

Abbildung 3.29: Das Punktdiagramm mit zusätzlichen Informationen

Ein weiterer Einsatzzweck für das Punktdiagramm

Neben der Anwendung, die Sie im vorherigen Absatz kennengelernt haben, können Sie das Punktdiagramm auch verwenden, um so etwas wie ein Liniendiagramm zu erstellen. Das normale Liniendiagramm eignet sich gut, wenn die Werte auf der x-Achse gleichmäßig verteilt sind, wie es bei Tabelle 3.1 der Fall war.

Angenommen, Ihre Daten sehen so aus wie die Werte in Abbildung 3.30. Kenner älterer Excel-Versionen (2010 und früher) erinnern sich vielleicht daran, dass ein Punktdiagramm mit geraden Linien und Datenpunkten der beste Weg war, um diese Daten zu visualisieren.

	A	B	C
1	x	y	
2	10	12	
3	20	16	
4	50	22	
5	100	75	
6	200	112	
7			

Abbildung 3.30: Diese Daten weisen auf Liniendiagramm hin, jedoch sind die x-Werte nicht gleichmäßig verteilt.

Excel 2016 erkennt die Struktur der Daten und schlägt den richtigen Diagrammtyp vor. Wenn Sie die Daten markieren und dann EINFÜGEN | EMPFOHLENE DIAGRAMME wählen, schlägt Excel den Diagrammtyp *Punkte mit geraden Linien und Datenpunkten* vor, obwohl der Typ lediglich mit PUNKT (XY) beschriftet ist (siehe Abbildung 3.31). So wissen Sie schnell, welcher Diagrammtyp für diese Daten geeignet ist.

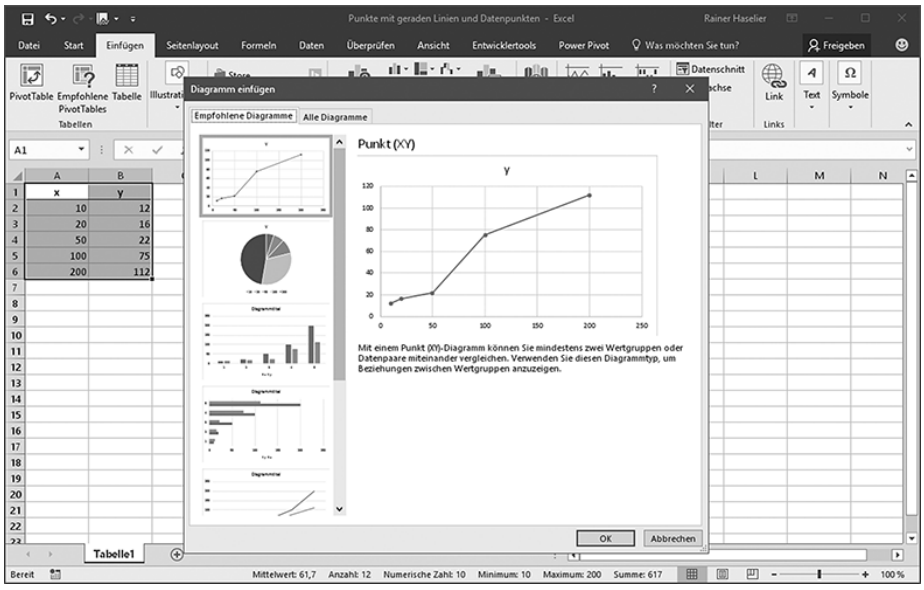


Abbildung 3.31: Der Befehl EMPFOHLENE DIAGRAMME schlägt für diese Daten den geeigneten Diagrammtyp vor.

In den Kapiteln 8, 10, 11 und 19 stelle ich einen weiteren Verwendungszweck des Punktdiagramms vor. Dort geht es dann um das Punktdiagramm mit geglätteten Linien.

Die Blasen schmecken

Ein Blasendiagramm dient dazu, in einem zweidimensionalen Diagramm drei Dimensionen darzustellen. Jeder Datenpunkt wird im Diagramm als Kreis beziehungsweise Blase dargestellt. Die Position der Blase in Bezug auf die x- und die y-Achse stellt zwei Dimensionen dar; die Größe der Blase repräsentiert die dritte Dimension.

Abbildung 3.32 zeigt die Daten, die ich in Kapitel 14 für das Thema multiple Regression verwende. Die Daten sind für die 20 Studenten in Spalte A. Spalte B zeigt den SAT-Index (sprachliche und mathematische Kenntnisse), Spalte C die High-School-Durchschnittsnoten und Spalte D den GPA-Wert (Grade Point Average auf einer Skala von 0,0 bis 4,0; entspricht dem Notendurchschnitt).

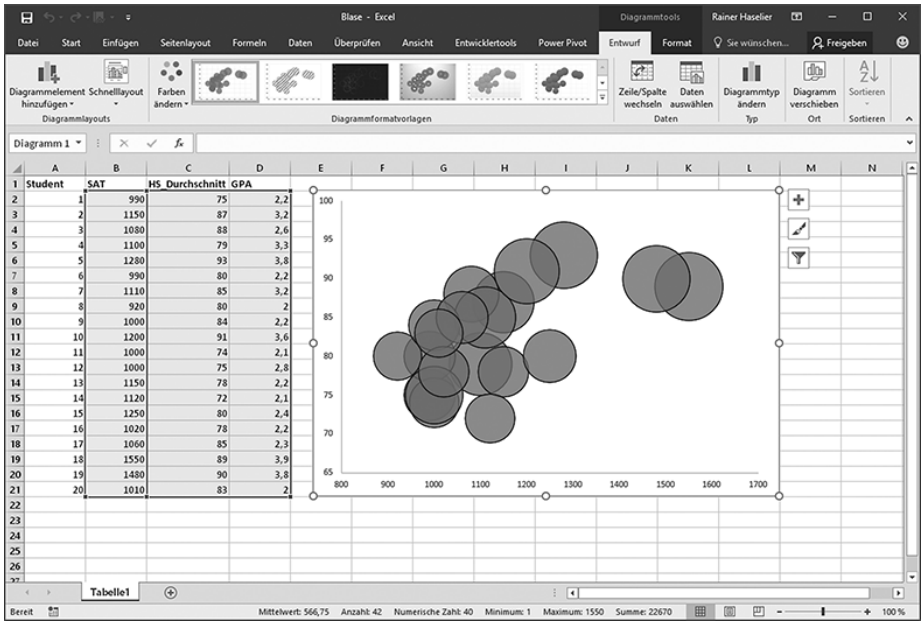


Abbildung 3.32: Das Blasendiagramm stellt drei Dimensionen in einem zweidimensionalen Diagramm dar.

In dem zugehörigen Blasendiagramm befindet sich der SAT-Index auf der x-Achse, die High-School-Durchschnittsnoten befinden sich auf der y-Achse und die Größe der Blase gibt den GPA-Wert des Colleges an.

Ich habe im Arbeitsblatt die Zellen B2 bis D21 markiert; das Blasendiagramm ist die sechste Auswahl der empfohlenen Diagramme. Ich habe die Achsen und die Blasen ein wenig formatiert, damit die dargestellten Informationen deutlicher werden.

Alles auf Kurs

Falls Sie die Kursentwicklung aller Unternehmen in Ihrem Aktienportfolio nachverfolgen, dann ist das Kursdiagramm für Sie genau das Richtige.

Abbildung 3.33 zeigt die Kurswerte der Google-Aktie für den Zeitraum 5. Januar bis 13. Januar 2016. (Die Daten sind unter www.google.com/finance/historical?q=NASDAQ:GOOG frei verfügbar.)

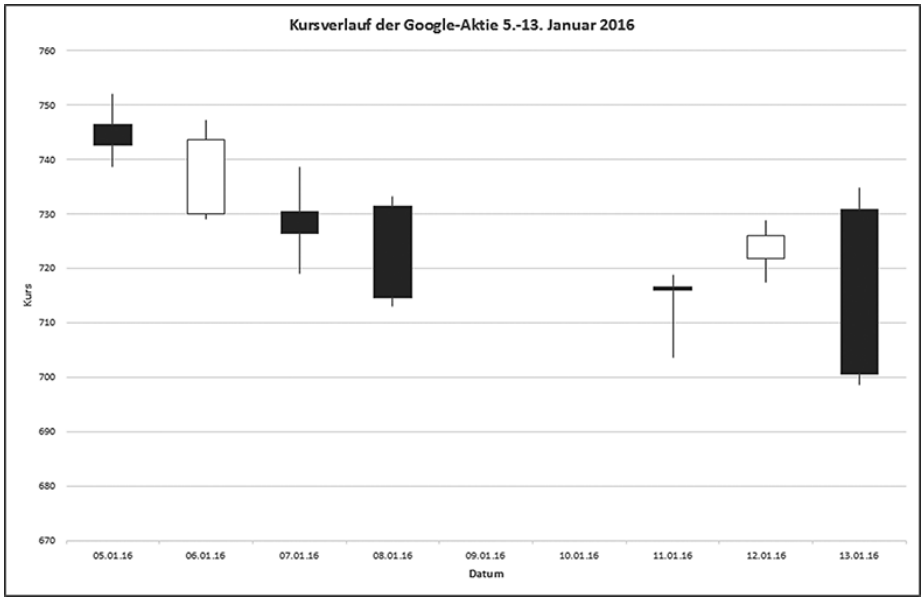


Abbildung 3.33: Ein Kursdiagramm, das den Kursverlauf der Google-Aktie für den Zeitraum 5. Bis 13. Januar 2016 darstellt.

Jeder Datenpunkt wird im Kursdiagramm mit einem Kästchen dargestellt, aus dem sowohl nach oben als auch nach unten eine Linie hinausragt. Die oberen und die untere Begrenzung des Kästchens geben den Eröffnungs- und den Schlusskurs für ein bestimmtes Datum an. Beachten Sie, dass ich *und* geschrieben habe und nicht *beziehungsweise*. Der Grund hierfür ist, dass der Eröffnungskurs manchmal höher ist als der Schlusskurs, und manchmal ist es genau umgekehrt.

Woran erkenne ich, was der Eröffnungs- und was der Schlusskurs ist? Wenn das Kästchen leer ist, dann ist der Eröffnungskurs die untere Begrenzung (und es gab an diesem Tag einen Kursgewinn). Wenn das Kästchen gefüllt ist, dann ist der Eröffnungskurs die obere Begrenzung (und es gab einen Kursverlust).

Der Endpunkt der nach oben zeigenden Linie stellt den Tageshöchstkurs dar und der Endpunkt der nach unten zeigen Linie den niedrigsten Kurs an diesem Tag.

Beachten Sie, dass Excel an der x-Achse auch Datumswerte ohne Kurse anzeigt (für die Tage, an denen die Börse geschlossen war).

Um dieses Diagramm zu erstellen, gehen Sie wie folgt vor:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

   Abbildung 3.34 zeigt die Daten, die ich für dieses Beispiel verwende.

	A	B	C	D	E	F
1	Datum	Eröffnung	Höchst	Niedrigst	Schluss	
2	13.01.16	730,85	734,74	698,61	700,56	
3	12.01.16	721,68	728,75	717,32	726,07	
4	11.01.16	716,61	718,86	703,54	716,03	
5	08.01.16	731,45	733,23	713	714,47	
6	07.01.16	730,31	738,5	719,06	726,309	
7	06.01.16	730	747,18	728,92	743,62	
8	05.01.16	746,46	752	738,64	742,58	
9						

Abbildung 3.34: Die Daten für das Google-Kursdiagramm

2. **Wählen Sie die Daten aus, die im Diagramm dargestellt werden sollen.**

   Für dieses Beispiel sind dies die Zellen A1 bis E8. Die Beschriftungen habe ich ebenfalls markiert.

3. **Wählen Sie EINFÜGEN | EMPFOHLENE DIAGRAMME und wählen Sie den Diagrammtyp aus.**

   Interessanterweise schlägt Excel nicht automatisch das Kursdiagramm vor. Darum habe ich die Registerkarte ALLE DIAGRAMME geöffnet, in der Liste auf der linken Seite KURS angeklickt und dann den Untertyp ERÖFFNUNGS-HÖCHST-TIEFST-SCHLUSSKURS ausgewählt. Das Ergebnis sehen Sie in Abbildung 3.35.

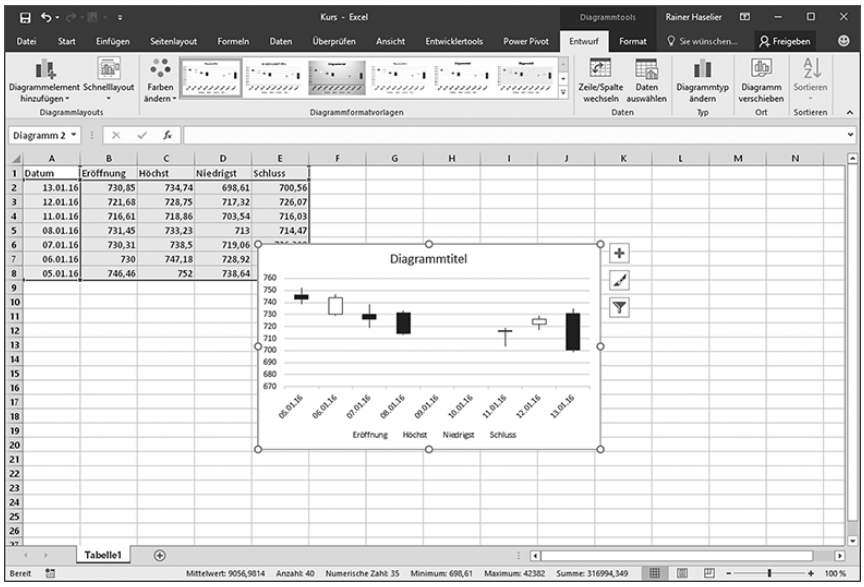


Abbildung 3.35: Ein Eröffnungs-Höchst-Tiefst-Schlusskurs-Diagramm erstellen

4. **Bearbeiten Sie das Diagramm.**

   Ich habe den allgemeinen Diagrammtitel angeklickt und einen neuen Titel eingegeben. Anschließend habe ich die Schaltfläche DIAGRAMMELEMENTE angeklickt (die mit dem Pluszeichen), im Popupmenü das Kontrollkästchen ACHSENTITEL eingeschaltet und den Text für die Achsentitel angepasst. Excel fügt Platzhalter in

das Diagramm ein, die ich geändert habe. Zum guten Schluss habe ich die Achsentitel und den Diagrammtitel fett formatiert, indem ich Strg+F gedrückt habe.

An der Oberfläche kratzen

Das Oberflächendiagramm ist ein dreidimensionales Diagramm, das die Ergebnisse aus der Kombination einer Variablen mit einer anderen darstellt. Das Arbeitsblatt in Abbildung 3.36 zeigt die Lernzeiten in Stunden in Kombination mit der Anzahl der Pausen. Die Zahl in jeder Zelle ist die in einer Prüfung erreichte Punktzahl.

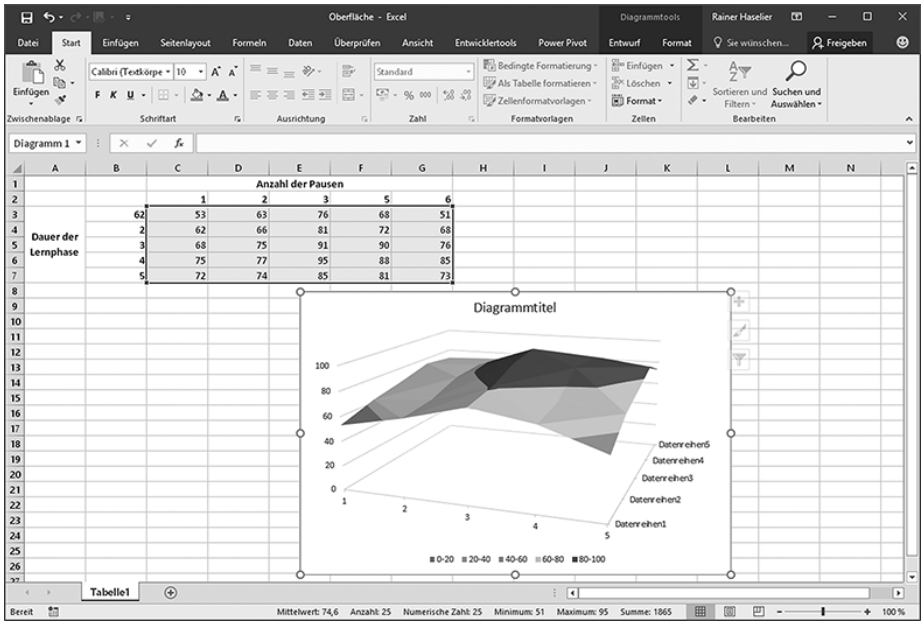


Abbildung 3.36: Ein Oberflächendiagramm erstellen

In dem zugehörigen Oberflächendiagramm werden die Lernzeiten auf der x-Achse dargestellt (das ist die Achse mit den Beschriftungen Datenreihen1 bis Datenreihen5), die Anzahl der Pausen befindet sich auf der y-Achse und die Performance (hier die erreichte Punktzahl) wird auf der z-Achse abgebildet. Farbige Bänder (die in der schwarzweißen Abbildung nicht gut zu unterscheiden sind), geben die Intervalle der Punktzahl an.

Ich habe die Daten in den Zellen C3 bis G7 markiert und dann auf der Registerkarte ALLE DIAGRAMME den Typ OBERFLÄCHE ausgewählt. Die erste Option führt zu dem in der Abbildung gezeigten Ergebnis.

Im Netz gefangen

Das Netzdiagramm stellt die Werte der einzelnen Kategorien in Form von konzentrischen Ringen dar. Die Werte erscheinen auf einer Achse, die in der Mitte beginnt und am äußeren Rand endet.

In Abbildung 3.37 sind die Kategorien *Madison, Wisconsin* und *Jacksonville, Florida*. Die Werte sind die durchschnittlichen Monatshöchsttemperaturen in °C für diese beiden Städte (von http://usclimatedata.com). Wie Sie vermutlich erwarten, sind die durchschnittlichen Höchsttemperaturen von Jacksonville in jedem Monat höher als für Madison.

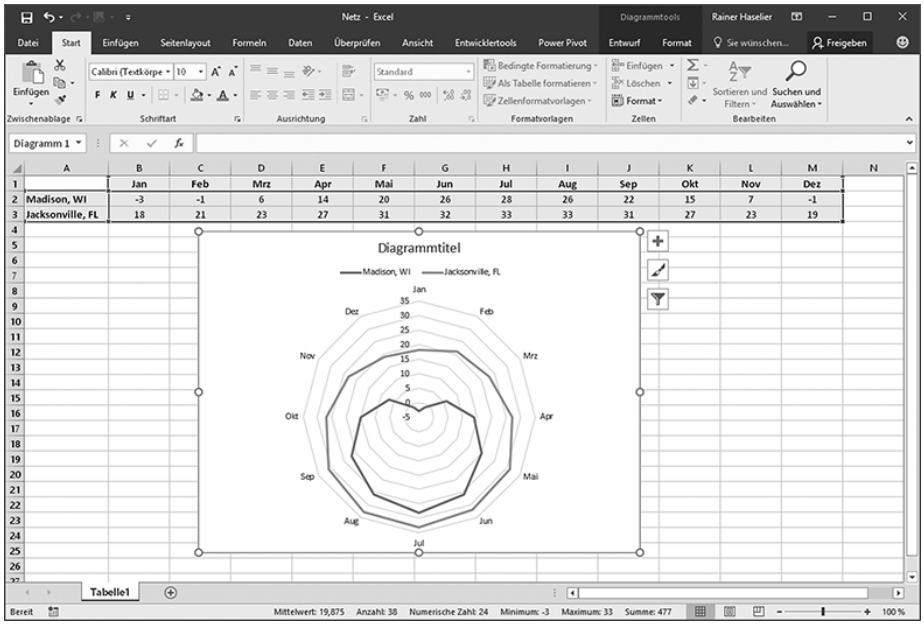


Abbildung 3.37: Ein Netzdiagramm erstellen

Ich habe die Zellen A1 bis M3 markiert und dann auf der Registerkarte ALLE DIAGRAMME die Option NETZ gewählt, um dieses Diagramm zu erstellen.

Ein dickes Sorry an alle Mac-User: Die im weiteren Verlauf dieses Kapitels vorgestellten Diagramme und die 3D-Karte stehen in Excel 2016 auf dem Mac nicht zur Verfügung.

Eine Treemap blühen und die Sonne hervorbrechen lassen

Eine Treemap ist ein hierarchisches Diagramm, das Muster in Daten darstellt. Rechtecke entsprechen dabei den Ästen und kleinere Rechtecke symbolisieren die Zweige.

Die Daten im Arbeitsblatt in Abbildung 3.40 zeigen die Ausgaben für Forschung und Entwicklung von drei US-Ministerien im Jahre 2010 (Quelle: www.census.gov/library/publications/2011/compendia/statab/131ed/science-technology.html). Die einzelnen Ministerien befinden sich auf der obersten Hierarchieebene (Ast) und die verschiedenen Bereiche stellen die nächstuntergeordnete Ebene da (Zweig). Um die in der Abbildung gezeigte Treemap zu erstellen, habe ich die Zellen A1 bis C8 markiert und in der Liste der empfohlenen Diagramme TREEMAP angeklickt.

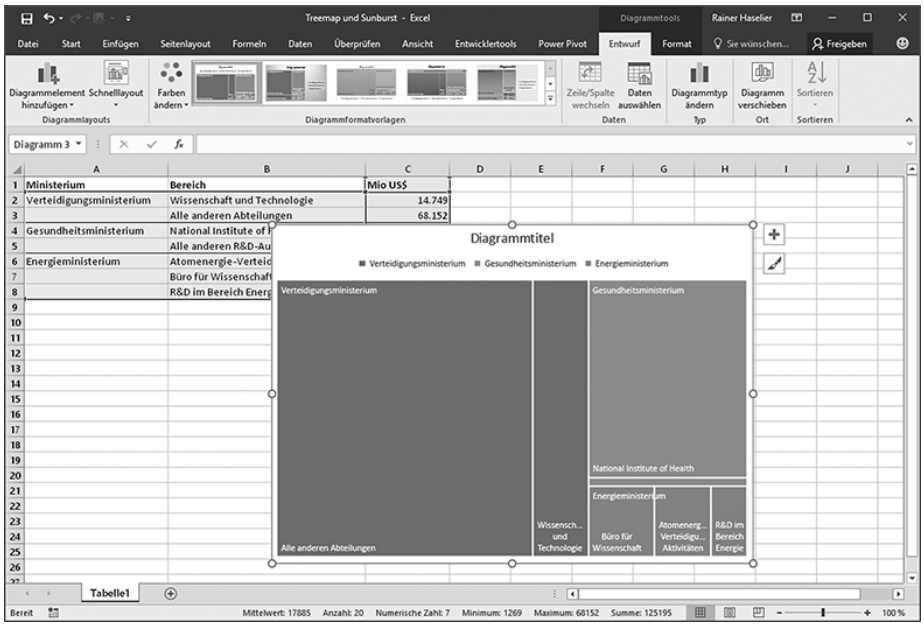


Abbildung 3.38: Eine Treemap erstellen

Beachten Sie, dass die Treemap keine Dollarbeträge anzeigt, sondern diese als anteilige Fläche darstellt.

Das Sunburst-Diagramm stellt die gleichen Arten von Daten dar wie die Treemap. Abbildung 3.39 zeigt das Sunburst-Diagramm für die Daten aus Abbildung 3.38. Der innere Ring stellt hier die oberste Hierarchieebene dar; die untergeordneten Ebenen befinden sich auf den äußeren Ringen. Wie auch bei der Treemap werden im Sunburst-Diagramm keine Dollarbeträge angezeigt; diese werden hier als anteilige Fläche mittels der Ringsegmente symbolisiert. Ich habe dieses Diagramm erstellt, indem ich die Zellen A1 bis C8 markiert und in der Liste der empfohlenen Diagramme SUNBURST ausgewählt habe.

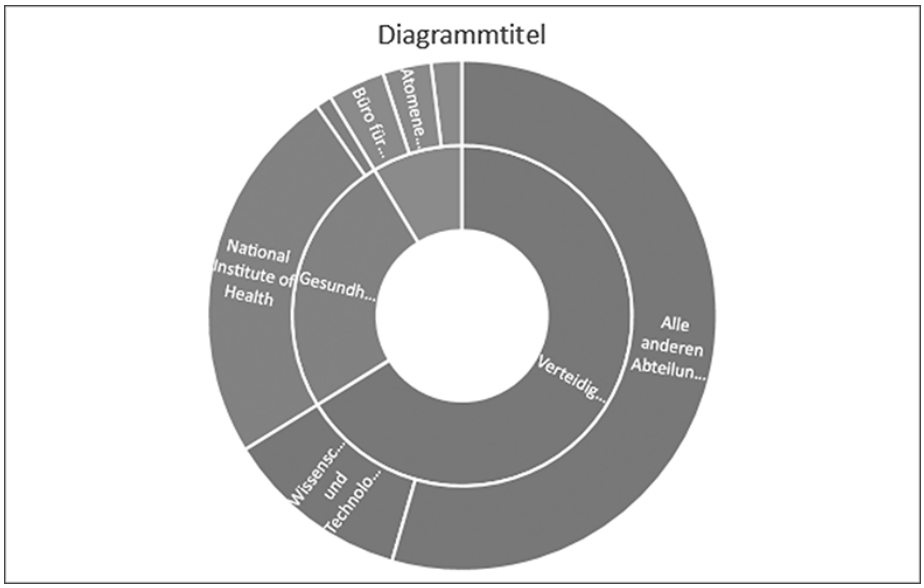


Abbildung 3.39: Ein Sunburst-Diagramm

Ein Histogramm erstellen

Histogramme gab es in Excel schon immer. Sie können entweder die Funktion HISTOGRAMM aus dem Add-In mit den Analyse-Funktionen verwenden oder die Funktion `HÄUFIGKEIT` einsetzen und die Daten dann als Säulendiagramm darstellen. In Excel 2016 können Sie Letzteres auch durchführen, indem Sie den Zellbereich mit den Daten markieren und dann das Diagramm Histogramm einfügen.

Die Daten in Abbildung 3.40 zeigen den SAT-Index der 20 Studenten aus Abbildung 3.32. Um das zugehörende Histogramm zu erstellen, habe ich die Zellen C2 bis C21 markiert und dann auf der Registerkarte ALLE DIAGRAMME die Option HISTOGRAMM ausgewählt. Wie Sie sehen können, versucht Excel, die Größe des Intervalls zu bestimmen (in diesem Beispiel 210). Um das Intervall zu ändern, doppelklicken Sie die x-Achse an und geben im Aufgabenbereich ACHSE FORMATIEREN im Feld CONTAINERBREITE einen anderen Wert ein.

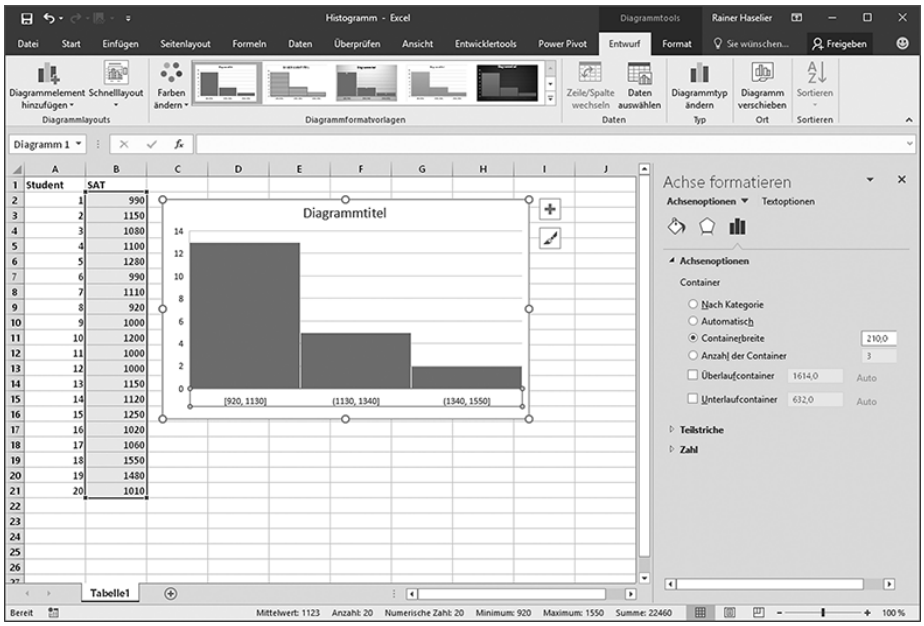


Abbildung 3.40: Ein Histogramm erstellen

Säulen sortieren: Pareto

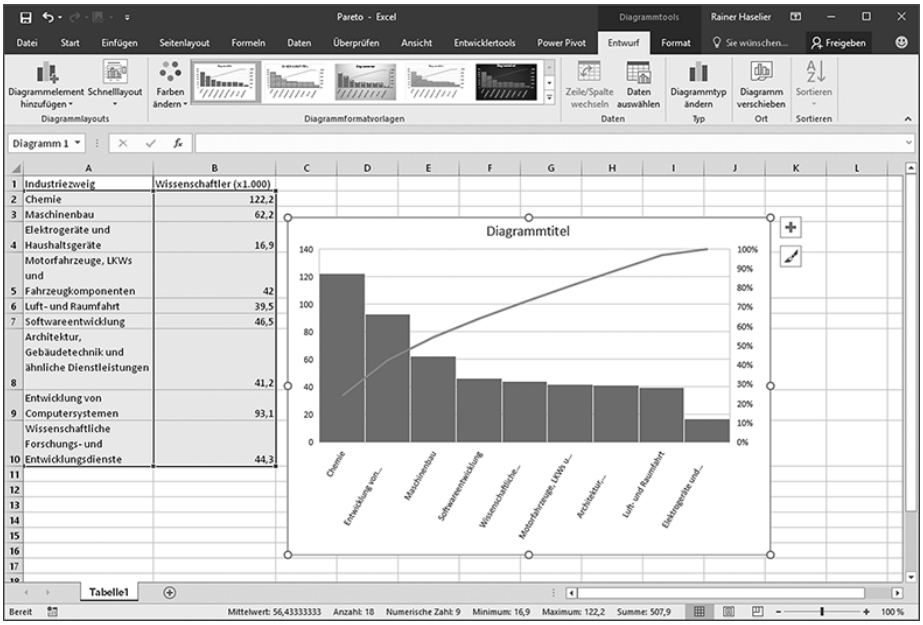
Das Pareto-Diagramm ist eine Kombination aus einem Säulen- und einem Liniendiagramm. Der Säulenteil ist besonders: Die Säulen werden absteigend in absteigender Reihenfolge der Häufigkeit angezeigt. Die Linie stellt den kumulativen Anteil an der Gesamtheit dar.

Abbildung 3.41 zeigt ein Beispiel. Bei den Daten handelt es sich um die Anzahl der Wissenschaftler (x1.000), die im Jahre 2006 in den USA in verschiedenen Industriezweigen arbeiteten (Quelle: www.census.gov/library/publications/2011/compendia/statab/131ed/science-technology.html).

Die Säulen des Diagramms repräsentieren die Anzahl der in den Industriezweigen arbeiteten Wissenschaftler in absteigender Reihenfolge. Die Linie stellt den kumulativen Prozentsatz des Industriezweigs dar; die Prozentwerte befinden sich auf der y-Achse an der rechten Seite des Diagramms.

Um dieses Diagramm zu erstellen, habe ich Zellen A2 bis B10 markiert und bei den empfohlenen Diagrammen die Option PARETO

ausgewählt.



	A	B
1	Industriezweig	Wissenschaftler (x1.000)
2	Chemie	122,2
3	Maschinenbau	62,2
4	Elektrogeräte und Haushaltsgeräte	16,9
5	Motorfahrzeuge, LKWs und Fahrzeugkomponenten	42
6	Luft- und Raumfahrt	39,5
7	Softwareentwicklung	46,5
8	Architektur, Gebäudetechnik und ähnliche Dienstleistungen	41,2
9	Entwicklung von Computersystemen	93,1
10	Wissenschaftliche Forschungs- und Entwicklungsdienste	44,3

Abbildung 3.41: Ein Pareto-Diagramm erstellen

Kastengrafik: Kästchen und Antennen

Das Kastengrafikdiagramm eignet sich gut für statistische Auswertungen. Auf den ersten Eindruck sieht es ein wenig so aus wie das Kursdiagramm, da jeder Datenpunkt als Kästchen mit vertikalen Linien dargestellt wird. Dann hören die Gemeinsamkeiten jedoch auf.

Im Arbeitsblatt in Abbildung 3.42 sind die Daten für die Anzahl der Internetverbindungen in acht Gemeinden eines fiktiven Landkreises für die Jahre 2014 bis 2016 dargestellt. Ich habe das Diagramm erstellt, indem ich die Zellen B2 bis D9 markiert und dann auf der Registerkarte ALLE DIAGRAMME den Typ KASTENGRAFIK ausgewählt habe.

Wie Sie sehen können, sehen die aus den Kästchen herausragenden Linien wie ein großes T aus. Dies sind die sogenannten Whisker (Antennen). Jedes Kästchen und dessen

Whisker fassen die Daten zusammen. Die Kästchen werden durch das obere Quartil (75 %-Quantil) und das untere Quartil (25 %-Quantil) begrenzt. (Weitere Informationen zu Quantilen finden Sie in Kapitel 6.) Eine horizontale Linie im Kästchen gibt den Median an (auch 50 %-Quantil genannt). Das X im Kästchen ist der Mittelwert. Der Endpunkt der nach oben zeigenden Linie stellt den größten und der Endpunkt der nach unten verlängerten Linie den kleinsten Wert des Datensatzes dar.

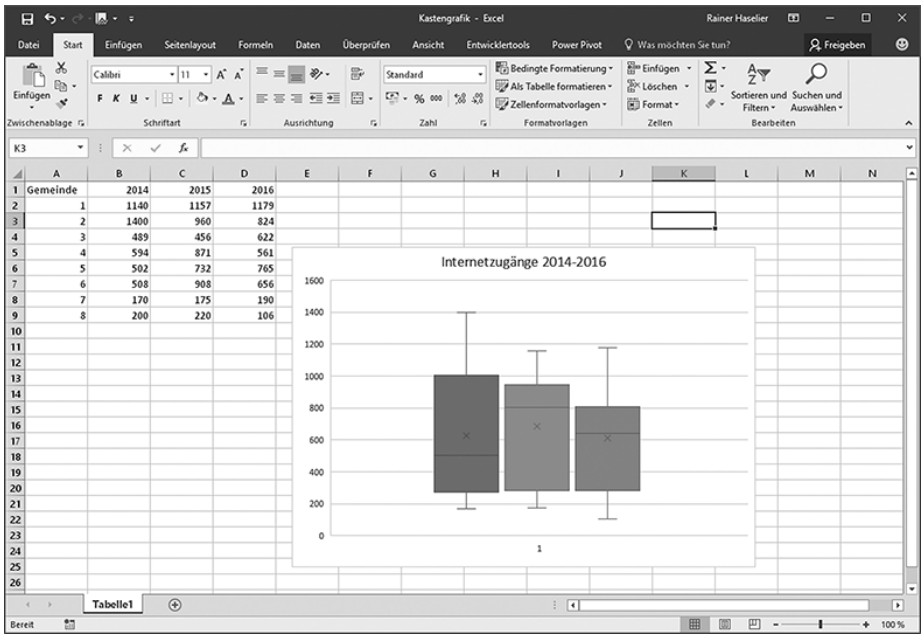


Abbildung 3.42: Ein Kastengrafikdiagramm erstellen

Wenn Sie die einzelnen Datenbereiche einzeln in das Kastengrafikdiagramm einfügen (und nicht wie hier gezeigt, alle auf einmal), zeigt jedes Kästchen alle Datenpunkte des Zellbereichs an. Um dies zu tun, wählen Sie den Diagrammtyp KASTENGRAFIK aus, bevor Sie irgendwelche Daten markieren. Um dann einen einzelnen Zellbereich hinzuzufügen, klicken Sie das Diagramm mit der rechten Maustaste an, klicken Sie im Kontextmenü auf DATEN AUSWÄHLEN und fügen Sie dann die Datenreihen hinzu.

Die letzten drei vorgestellten Diagrammtypen werden von Excel »statistische Diagramme« genannt.

3D-Karten

In Excel 2013 gab es dieses Feature unter dem Namen Power View; in Excel 2016 heißt es 3D-Karte. Wie es der Name andeutet, können mit 3D-Karten Daten auf einer dreidimensionalen Karte abgebildet werden. In diesem Abschnitt zeige ich, wie mit 3D-Karten ein Globus erstellt wird, auf dem Daten zu vielen Ländern dargestellt werden.

Abbildung 3.43 zeigt einen Ausschnitt eines Arbeitsblatts mit Daten zur Lebenserwartung bei Geburt, der voraussichtlichen und der durchschnittlichen Schulbesuchsdauer für 188 Länder (Quelle: http://hdr.undp.org/en/composite/HDI).

	A	B	C	D
1				
2	Land	Lebenserwartung bei Geburt	Voraussichtliche Schulbesuchsdauer	Durchschnittliche Schulbesuchsdauer
3	Norway	81,6	17,5	12,6
4	Australia	82,4	20,2	13,0
5	Switzerland	83,0	15,8	12,8
6	Denmark	80,2	18,7	12,7
7	Netherlands	81,6	17,9	11,9
8	Germany	80,9	16,5	13,1
9	Ireland	80,9	18,6	12,2
10	United States	79,1	16,5	12,9
11	Canada	82,0	15,9	13,0
12	New Zealand	81,8	19,2	12,5
13	Singapore	83,0	15,4	10,6
14	Hong Kong, China (SAR)	84,0	15,6	11,2
15	Liechtenstein	80,0	15,0	11,8
16	Sweden	82,2	15,8	12,1
17	United Kingdom	80,7	16,2	13,1
18	Iceland	82,6	19,0	10,6
19	Korea (Republic of)	81,9	16,9	11,9
20	Israel	82,4	16,0	12,5
21	Luxembourg	81,7	13,9	11,7
22	Japan	83,5	15,3	11,5
23	Belgium	80,8	16,3	11,3
24	France	82,2	16,0	11,1
25	Austria	81,4	15,7	10,8

Abbildung 3.43: Lebenserwartung bei Geburt sowie voraussichtliche und durchschnittliche Schulbesuchsdauer für 188 Länder

Ich habe 3D-Karten verwendet, um einen dreidimensionalen Globus zu erstellen, der diese Länder und die betreffenden Kenngrößen

darstellt. Die Weltkugel wird von der Microsoft-Suchmaschine Bing geliefert, daher ist eine aktive Internetverbindung erforderlich.

Um diese Weltkugel zu erstellen, gehen Sie wie folgt vor:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

2. **Markieren Sie die Daten, einschließlich der Spaltenüberschriften.**

3. **Klicken Sie auf EINFÜGEN | 3D-KARTE | 3D-KARTEN öffnen.**

   Nach ein paar Sekunden (haben Sie etwas Geduld) wird das Fenster geöffnet, das Abbildung 3.44 zeigt.

4. **Ziehen Sie Felder (die Namen der Spaltenüberschriften) aus der Feldliste in den Aufgabenbereich SCHICHT1.**

   Ich habe die Felder DURCHSCHNITTLICHE SCHULBESUCHSDAUER, LEBENSERWARTUNG BEI GEBURT und VORAUSSICHTLICHE SCHULBESUCHSDAUER in das Feld HÖHE gezogen. Das Ergebnis sehen Sie in Abbildung 3.45. Die Daten wurden als gestapelte Säulen in das Diagramm eingefügt und sind im Zentrum der einzelnen Länder zu sehen.

5. **Schauen Sie sich den gesamten Globus an.**

   Schließen Sie das kleine Fenster auf der linken und auf der rechten Seite und lassen Sie das Fenster SCHICHT 1 mit der Legende geöffnet (siehe Abbildung 3.46).

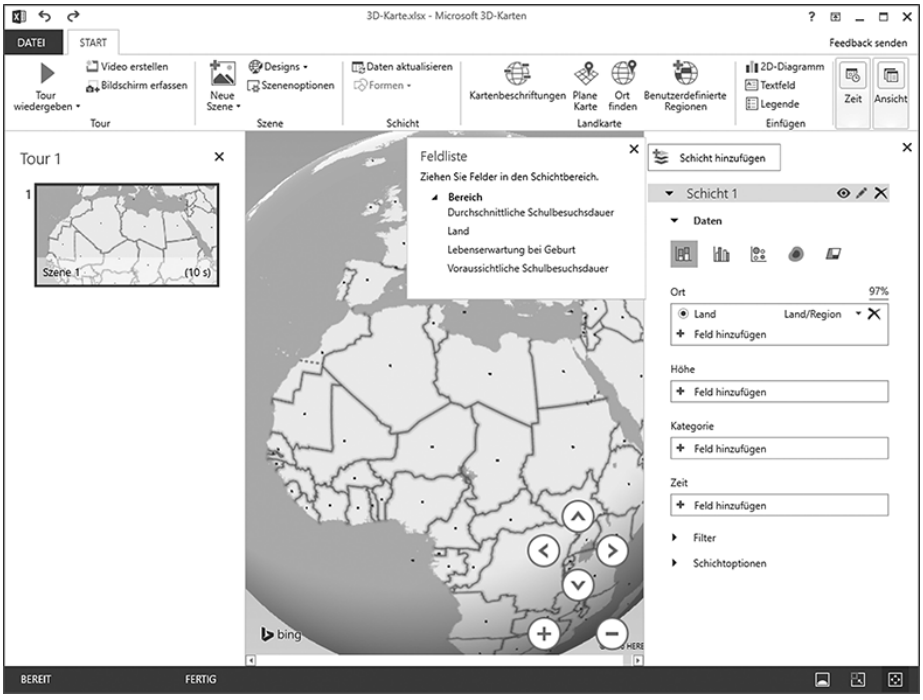


Abbildung 3.44: Wenn Sie den Befehl 3D-KARTEN ÖFFNEN wählen, wird dieses Fenster angezeigt.

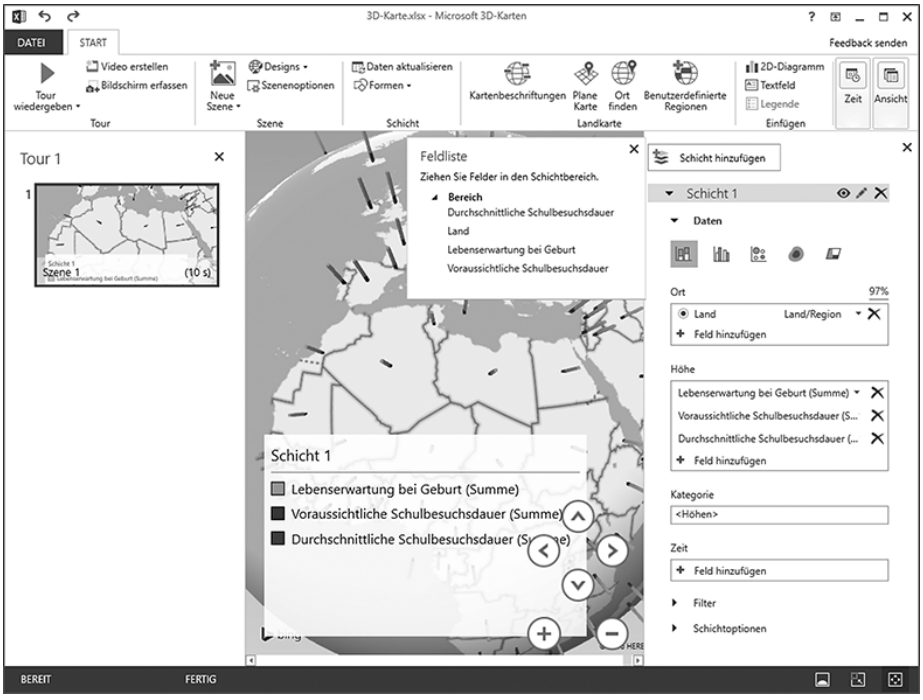


Abbildung 3.45: Die Daten wurden in den Globus als gestapelte Säulen eingefügt.

Abbildung 3.46: Der fertige Globus

Ein statischer Globus wäre nicht sehr hilfreich, da dann viele Länder nicht sichtbar wären. Mit den Schaltflächen unten rechts im Fenster können Sie den Globus drehen und hinein- beziehungsweise herauszoomen. Sie können außerdem die Farben der Säulen ändern und diese neu anordnen.

Das Feature 3D-Karten eignet sich nicht nur für die Darstellung von Daten auf Länderebene: Sie können ebenfalls Karten für einen viel kleineren Maßstab erstellen.

Kapitel 4

Suchen Sie Ihre Mitte

In diesem Kapitel

- Arbeiten Sie mit den Mittelwerten
- Bedingungen erfüllen
- Alles dreht sich um den Median
- Begeisterung für den Modalwert

Statistiker beschäftigen sich mit Zahlengruppen. Sie finden es häufig hilfreich, eine Zahlengruppe durch eine einzige Zahl zusammenzufassen. Aber woher kommt eine solche Zusammenfassungszahl?

Am besten sucht man nach einer Zahl irgendwo in der Mitte und verwendet diese Zahl als Repräsentant für die ganze Gruppe. Wenn Sie eine Zahlengruppe betrachten und versuchen, eine Zahl zu finden, die sich irgendwo in der Mitte befindet, befassen Sie sich mit der *zentralen Tendenz* (oder den *Lagemaßen*) dieser Zahlengruppe. Wie gute Eiscreme gibt es auch die zentrale Tendenz in verschiedenen Geschmacksrichtungen.

Mittelwert: Die Lehre vom Durchschnitt

Praktisch jeder verwendet Durchschnittswerte. Der statistische Ausdruck für einen Durchschnittwert ist *Mittelwert* oder *Mittel*. Irgendwann in Ihrem Leben haben Sie mit Sicherheit schon einmal einen Mittelwert berechnet. Mit dem Mittelwert können Sie Ihre

Noten, Ihr Vermögen oder vielleicht Ihre Leistung bei einer Aufgabe oder beim Sport über einen bestimmten Zeitraum hinweg schnell zusammenfassen.

Ein weiterer Grund für die Berechnung von Mittelwerten hat mit wissenschaftlicher Arbeit zu tun. Ein Wissenschaftler wendet ein bestimmtes Verfahren auf eine kleine Stichprobe von Menschen oder Dingen an und misst das Ergebnis auf irgendeine Art. Er verwendet das Ergebnis aus der Stichprobe, um die Auswirkungen des Verfahrens auf die Grundgesamtheit abzuschätzen, aus der die Stichprobe gezogen wurde. Dabei stellt der Mittelwert der Stichprobe den besten Schätzwert für den Mittelwert der Grundgesamtheit dar.

Den Mittelwert berechnen

Ich brauche Ihnen vermutlich nicht zu erklären, wie der Mittelwert berechnet wird. Aber ich werde es trotzdem tun. Dann werde ich Ihnen die statistische Formel zeigen. Mir geht es darum, dass Sie verstehen, wie statistische Formeln allgemein funktionieren, und ich möchte Ihnen erklären, wie Excel Mittelwerte berechnet.

Ein Mittelwert ist einfach die Summe aus einer Reihe von Zahlen, dividiert durch die Anzahl der addierten Zahlen. Ein Beispiel: Nehmen wir einmal an, Sie messen die Lesegeschwindigkeit von sechs Kindern in Wörtern pro Minute. Sie ermitteln folgende Lesegeschwindigkeiten:

56, 78, 45, 49, 55, 62

Die durchschnittliche Lesegeschwindigkeit dieser sechs Kinder beträgt

$$\frac{56+78+45+49+55+62}{6}=57,5$$

Das bedeutet, dass der Mittelwert bei diesem Beispiel 57,5 Wörter pro Minute beträgt.

Ein erster Versuch, wie die Formel aussehen könnte:

$$Mittelwert = \frac{Summe\ von\ Zahlen}{Anzahl\ addierter\ Zahlen}$$

Diese Formel ist in dieser Form sehr unhandlich. Daher verwenden Statistiker Abkürzungen. »Zahl« wird häufig mit *X* abgekürzt. Die »Anzahl addierter Zahlen« wird meist mit *N* abgekürzt. Mit diesen Abkürzungen sieht die Formel folgendermaßen aus:

$$Mittelwert = \frac{Summe\ von\ x}{n}$$

Eine weitere Abkürzung, die in der Statistik verwendet wird, wird für »Summe von« verwendet. Hierbei handelt es sich um den griechischen Großbuchstaben für *S*. Er heißt »Sigma« und sieht so aus: $\sum$. Und so sieht die Formel mit Sigma aus:

$$Mittelwert = \frac{\sum x}{n}$$

Und was ist mit dem »Mittelwert«? Statistiker haben auch dafür eine Abkürzung. M wäre eine gute Abkürzung, und manche Statistiker verwenden diese auch. Aber die meisten Statistiker verwenden $\bar{x}$ (x mit Querbalken) für den Mittelwert. Und so sieht die fertige Formel aus:

$$\bar{x} = \frac{\sum x}{n}$$

War's das? Noch nicht ganz. Lateinische Buchstaben wie $\bar{x}$ stellen Eigenschaften von Stichproben dar. Für Eigenschaften von Grundgesamtheiten werden griechische Buchstaben verwendet. Für den Mittelwert aus der Grundgesamtheit wird der griechische Buchstabe für *M* verwendet: μ (wie mü ausgesprochen). Die Formel für den Mittelwert aus der Grundgesamtheit lautet somit

$$\mu = \frac{\sum x}{n}$$

Und das war's dann.

MITTELWERT und MITTELWERTA

Mit der Tabellenblattfunktion `MITTELWERT` von Excel können Sie den Mittelwert einer Zahlengruppe berechnen. In Abbildung 4.1 sind die Daten und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `MITTELWERT` dargestellt.

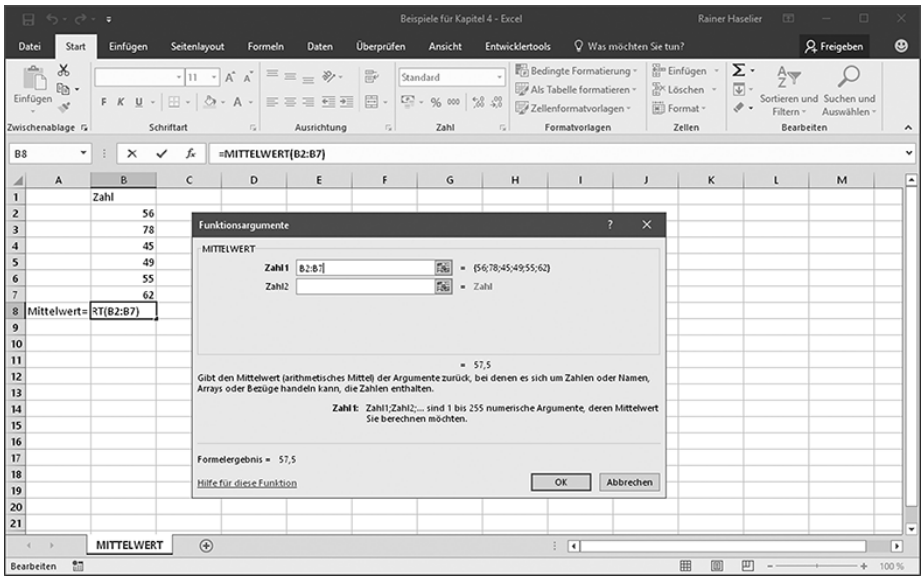


Abbildung 4.1: Die Funktion *MITTELWERT* verwenden

Und so müssen Sie vorgehen:

1. **Geben Sie im Arbeitsblatt die gewünschten Zahlen in einen Zellbereich ein, und wählen Sie die Zelle aus, in die die Funktion `MITTELWERT` das Ergebnis einfügen soll.**

   In diesem Beispiel habe ich in die Zellen B2 bis B7 die Zahlen 56, 78, 45, 49, 55 und 62 eingegeben und B8 zum Speichern des Ergebnisses angegeben.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MITTELWERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Wenn im Feld ZAHL1 nicht bereits der Zellbereich mit den Zahlen enthalten ist, geben Sie ihn dort ein. Der Mittelwert (in diesem Beispiel 57,5) wird in diesem Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen.**

   Damit wird der Mittelwert in die Zelle eingefügt, die Sie im Arbeitsblatt ausgewählt haben. In diesem Beispiel ist das die Zelle B8.

Wie Sie in der Bearbeitungsleiste in Abbildung 4.1 sehen können, wird folgende Formel verwendet:

```
=MITTELWERT(B2:B7)
```

Hätte ich dem Zellbereich B2:B7 den Namen `Zahlen` zugewiesen (siehe Kapitel 2), würde die Formel wie folgt aussehen:

```
=MITTELWERT(Zahlen)
```

`MITTELWERTA` bewirkt dasselbe wie `MITTELWERT`, jedoch mit einem wichtigen Unterschied. Wenn der Mittelwert mit `MITTELWERT` berechnet wird, werden Zellen mit Buchstaben oder mit Ausdrücken wie `WAHR` oder `FALSCH` nicht beachtet. Mit `MITTELWERTA` werden Buchstaben und Ausdrücke bei der Berechnung des Mittelwerts dagegen berücksichtigt. Wenn also eine Zelle Text oder den Ausdruck `FALSCH` enthält, hat sie den Wert 0. Wenn eine Zelle den logischen Wert `WAHR` enthält, hat sie den Wert 1. `MITTELWERTA` berücksichtigt diese Werte beim Berechnen des Mittelwerts.

Neben `MITTELWERTA` stellt Excel Tabellenblattfunktionen wie `VARIANZA` und `STABWA` bereit, die bei der Berechnung ebenfalls Zellen mit Text und mit logischen Ausdrücken berücksichtigen. Ich

weiß nicht, ob Sie diese Funktionen bei Ihrer täglichen Arbeit mit der Statistik jemals brauchen werden (ich habe sie bisher nicht gebraucht). Aber ich werde Ihnen vorsichtshalber mal erklären, wie sie funktionieren. So verwenden Sie die Funktion MITTELWERTA:

1. **Geben Sie die Zahlen ein und markieren Sie eine Zelle zum Speichern des Ergebnisses.**

   Geben Sie bei diesem Beispiel die Zahlen 56, 78, 45, 49, 55 und 62 in die Zellen B2 bis B7 ein und markieren Sie B9. Damit bleibt B8 leer.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MITTELWERTA, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Dieses Mal gebe ich in das Feld ZAHL1 den Zellbereich B2:B8 ein. Im Dialogfeld wird der Mittelwert (57,5) angezeigt. MITTELWERTA ignoriert also genau wie MITTELWERT leere Zellen.

4. **Klicken Sie auf OK, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen und den Mittelwert in die markierte Zelle einzufügen.**

Nun können Sie ein wenig experimentieren. Geben Sie in Zelle B8 **xxx** ein. Anstelle von 57,5 wird in B9 nun 49,28571 angezeigt. Geben Sie anschließend **WAHR** in B8 ein. Nun wird in B9 49,42857 angezeigt. Und geben Sie schließlich **FALSCH** in B8 ein. Nun wird wieder 49,28571 angezeigt.

Woher resultieren diese Änderungen? MITTELWERTA behandelt eine Zeichenfolge wie xxx als 0. Daher wird in diesem Fall der Mittelwert nicht aus 6, sondern aus 7 Zahlen berechnet; eine dieser Zahlen ist null. MITTELWERTA wertet WAHR als Wert 1 aus. Daher

wird auch hier der Mittelwert von 7 Zahlen berechnet; eine dieser Zahlen ist die 1. MITTELWERTA behandelt den logischen Wert FALSCH wie null und berechnet daher den gleichen Mittelwert wie für den Fall, das B8 den Text xxx enthält.

MITTELWERTWENN und MITTELWERTWENNS

Die beiden Funktionen MITTELWERTWENN und MITTELWERTWENNS berechnen den Mittelwert der Zellen, die eine Bedingung oder mehrere Bedingungen erfüllen. MITTELWERTWENN berücksichtigt den Wert der Zellen, falls eine Bedingung zutrifft. MITTELWERTWENNS berücksichtigt den Wert der Zellen, falls mehrere Bedingungen zutreffen.

Um zu zeigen, wie diese beiden Funktionen arbeiten, habe ich das Tabellenblatt erstellt, das Sie in Abbildung 4.2 sehen. Die dort eingegebenen Daten enthalten die Ergebnisse eines fiktionalen, psychologischen Experiments. In diesem Experiment sitzt eine Person vor einem Bildschirm, auf dem eine mit einer Farbe gefüllte Form angezeigt wird. Die Farbe ist entweder Rot oder Grün und die Form ist entweder ein Kreis oder ein Quadrat. Bei jedem Versuch wird eine zufällige Kombination aus Farbe und Form angezeigt; alle Kombinationen werden gleich oft verwendet. Im Fachjargon handelt es sich bei jeder angezeigten Form um einen Versuch. Daher zeigt das Tabellenblatt die Ergebnisse von 16 Versuchen an.

	A	B	C	D
1	Versuch	Farbe	Form	Reaktionszeit
2	1	Rot	Kreis	410
3	2	Rot	Quadrat	334
4	3	Grün	Quadrat	335
5	4	Grün	Kreis	336
6	5	Rot	Quadrat	398
7	6	Rot	Kreis	450
8	7	Rot	Kreis	440
9	8	Grün	Quadrat	467
10	9	Grün	Kreis	445
11	10	Rot	Quadrat	296
12	11	Grün	Quadrat	378
13	12	Rot	Kreis	496
14	13	Rot	Kreis	544
15	14	Grün	Quadrat	468
16	15	Rot	Quadrat	577
17	16	Grün	Kreis	448
18				
19			Mittelwert Gesamt =	426,375
20			Mittelwert Kreis=	446,125
21			Mittelwert grünes Quadrat =	412

Abbildung 4.2: Daten von 16 Versuchen eines fiktiven, psychologischen Experiments.

Die Person vor dem Computer drückt eine Taste, wenn sie eine Form sieht. Spalte D (mit *Reaktionszeit* beschriftet), gibt die Reaktionszeit der Versuchsperson in Millisekunden (Tausendstel einer Sekunde) an. Die Spalten B und C enthalten die Merkmale der in dem jeweiligen Versuch angezeigten Form. So können Sie beispielsweise aus Zeile 2 ablesen, dass im ersten Versuch ein roter Kreis angezeigt wurde und dass die Reaktionszeit 410 Millisekunden betrug.

Für alle Spalten habe ich den Text in der obersten Zeile verwendet, um einen Namen zu definieren, über den ich dann auf die Daten verweisen kann. (Wenn Sie nicht mehr genau wissen, wie das geht, werfen Sie einen Blick in Kapitel 2).

Ich habe dann drei verschiedene Mittelwerte berechnet. Der erste, Mittelwert Gesamt, der in Zelle D20 angezeigt wird, wird ganz einfach mit der folgenden Formel berechnet:

```
=MITTELWERT(Reaktionszeit)
```

Was ist mit den beiden anderen Mittelwerten? Zelle D20 enthält den Mittelwert der Versuche, bei denen ein Kreis angezeigt wurde. Dies ist es, was ich mit bedingtem Mittelwert meine: Hier ist es der Mittelwert der Versuche, bei denen die Bedingung `Form = Kreis` zutrifft.

Abbildung 4.3 zeigt das fertig ausgefüllte Dialogfeld FUNKTIONSARGUMENTE für `MITTELWERTWENN`. Nach dem Klick auf OK fügt Excel in die betreffende Zelle die folgende Formel ein:

```
=MITTELWERTWENN(Form;"Kreis";Reaktionszeit)
```

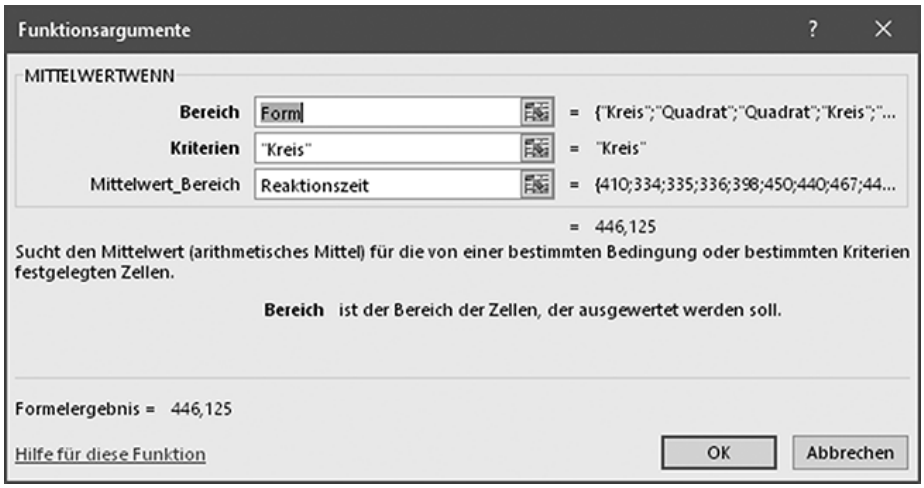


Abbildung 4.3: Das fertig ausgefüllte Dialogfeld für *MITTELWERTWENN*.

Die Formel und das Dialogfeld liefern Ihnen die folgenden Informationen: Excel berücksichtigt bei der Berechnung des Mittelwerts dann den Wert einer Zelle in Spalte `Reaktionszeit` (Spalte D), wenn die betreffende Zelle in Spalte FORM (Spalte B) den Wert `"Kreis"` enthält. Trifft diese Bedingung nicht zu, wird der Wert der Zelle nicht berücksichtigt.

Um diese Formel zu erstellen, führen Sie folgende Schritte durch:

1. **Geben Sie die Zahlen in das Tabellenblatt ein und markieren Sie die Zelle, in der das Ergebnis stehen soll.**

   Ich habe Zelle D20 ausgewählt.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MITTELWERTWENN, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Für dieses Beispiel ist der BEREICH `Form`, die KRITERIEN sind `"Kreis"` (Excel ergänzt die doppelten Anführungszeichen) und der Parameter für MITTELWERT_BEREICH ist `Reaktionszeit`.

4. **Klicken Sie auf OK, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen; das Ergebnis wird in der ausgewählten Zelle angezeigt.**

Hier weitere Beispiele für die Verwendung von `MITTELWERTWENN`. Um den Mittelwert der ersten acht Versuche zu berechnen, verwenden Sie die folgende Formel:

```
=MITTELWERTWENN(Versuch;<9;Reaktionszeit)
```

Um den Mittelwert der Versuche zu ermitteln, bei denen die Reaktionszeit schneller war als 400 Millisekunden, verwenden Sie die folgende Formel:

```
=MITTELWERTWENN(Reaktionszeit;<400;Reaktionszeit)
```

In den beiden letzten Beispielen steht der Vergleichsoperator »<« vor dem numerischen Wert. Falls Sie bei der Definition der Bedingung den Wert vor den Operator stellen, wird die Formel nicht funktionieren.

Was ist mit dem Mittelwert für grüne Quadrate, den Sie in Zelle D21 sehen? Abbildung 4.4 zeigt das fertig ausgefüllte Dialogfeld FUNKTIONSARGUMENTE für die Funktion `MITTELWERTWENNS`, bei der Sie mehr als eine Bedingung angeben können. Die Formel, um den

Mittelwert der Versuche zu berechnen, bei denen `Farbe = Grün` und `Form = Quadrat` ist, lautet:

```
=MITTELWERTWENNS(Reaktionszeit;Farbe;"Grün";Form;"Quadrat")
```

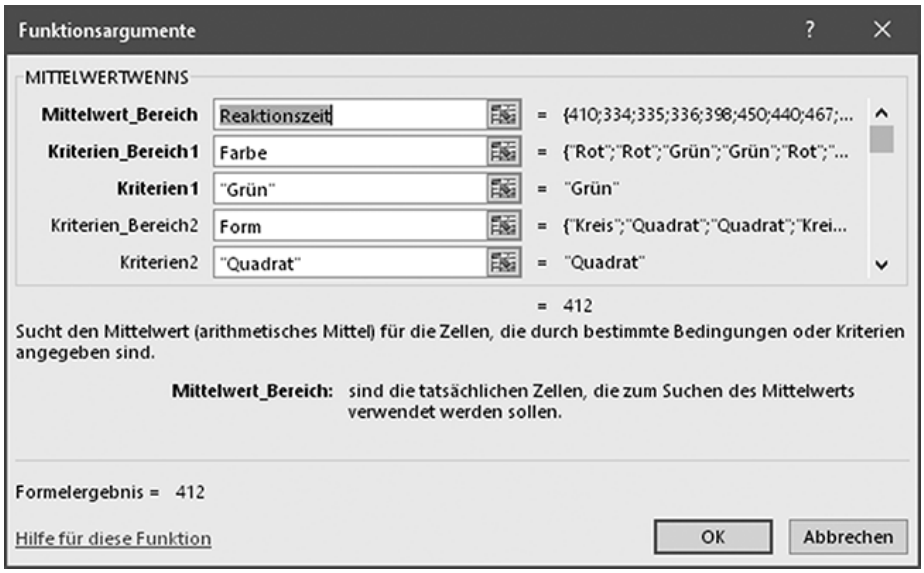


Abbildung 4.4: Das fertig ausgefüllte Dialogfeld für *MITTELWERTWENNS*

Beachten Sie, dass das Argument MITTELWERT_BEREICH bei `MITTELWERTWENNS` das erste und bei `MITTELWERTWENN` das letzte Argument ist.

Um diese Formel zu erstellen, führen Sie folgende Schritte durch:

1. **Geben Sie die Zahlen in das Tabellenblatt ein und markieren Sie die Zelle, in der das Ergebnis stehen soll.**

   Ich habe Zelle D21 ausgewählt.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MITTELWERTWENNS, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

Für dieses Beispiel ist der MITTELWERT_BEREICH die `Reaktionszeit`. Der KRITERIEN_BEREICH1 ist `Farbe` und das Argument für Kriterien1 ist `"Grün"`. Der KRITERIEN_BEREICH2 ist `Form` und das Argument für Kriterien2 ist `"Quadrat"`. (Excel ergänzt die doppelten Anführungszeichen da, wo es erforderlich ist.)

4. **Klicken Sie auf OK, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen; das Ergebnis wird in der ausgewählten Zelle angezeigt.**

Nach allem, was Sie jetzt gesehen haben, fragen Sie sich vielleicht, warum Excel sowohl `MITTELWERTWENN` als auch `MITTELWERTWENNS` zur Verfügung stellt. Die Formel

```
=MITTELWERTWENN(Form;"Kreis";Reaktionszeit)
```

liefert das gleiche Ergebnis wie

```
=MITTELWERTWENNS(Reaktionszeit;Form;"Kreis")
```

Warum also zwei Funktionen? Die kurze Antwort: Ich weiß es nicht. Die lange Antwort: Ich weiß es nicht.

GESTUTZTMITTEL

In einer Abwandlung eines bekannten Zitats zum Thema Statistik sagte jemand: »Es gibt drei Arten von Lügen: Lügen, verdammte Lügen und statistische Ausreißer.« Ein *Ausreißer* ist ein Extremwert innerhalb einer Wertegruppe, der so extrem ist, dass die Person, die die Werte zusammengetragen hat, glaubt, dass etwas nicht stimmt.

Ein Beispiel für Ausreißer findet sich in psychologischen Experimenten, bei denen die Zeit gemessen wird, die eine Person für eine Entscheidung benötigt. Gemessen wird in tausendstel Sekunden, wobei die Reaktionszeiten von der Komplexität der Entscheidung abhängen: Je komplexer die Entscheidung, desto länger die Reaktionszeit.

Eine Person, die an einem Experiment wie diesem teilnimmt, durchläuft in der Regel viele experimentelle Tests und trifft eine

Entscheidung pro Test. Ein Test mit einer außergewöhnlich schnellen Reaktionszeit (deutlich unter dem Durchschnitt) ist möglicherweise ein Hinweis darauf, dass die Person schnell geraten hat, ohne wirklich zu überlegen, was sie eigentlich tun sollte. Ein Test mit einer besonders langsamen Reaktionszeit (deutlich über dem Durchschnitt) ist möglicherweise ein Hinweis darauf, dass die Person zunächst nicht aufmerksam war, sich dann aber auf die Aufgabe konzentriert hat.

Beide Arten von Ausreißer können der Schlussfolgerung, die anhand des Mittelwerts der Daten gezogen wird, im Wege stehen. Daher sollten diese vor der Berechnung des Mittelwerts am besten ausgeschlossen werden. Statistiker bezeichnen diesen Vorgang als *Mitteln ohne Ausreißer*. Und diese Aufgabe übernimmt die Funktion `GESTUTZTMITTEL` von Excel.

Und so verwenden Sie die Funktion `GESTUTZTMITTEL`:

1. **Geben Sie die gewünschten Werte in einen Zellbereich ein und markieren Sie eine Zelle für das Ergebnis.**

   Bei diesem Beispiel habe ich die folgenden Zahlen in die Zellen B2 bis B11 eingegeben:

   2500, 280, 550, 540, 525, 595, 620, 1052, 591, 618

   Diese Werte können von einem psychologischen Experiment stammen, bei dem die Reaktionszeit in tausendstel Sekunden (Millisekunden) gemessen wird.

   Ich habe B12 für das Ergebnis gewählt.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf GESTUTZTMITTEL, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

Das Datenfeld wird im Feld MATRIX eingegeben. In diesem Beispiel geben Sie B2:B11 ein.

Anschließend geben Sie den Prozentwert der Werte ein, die bei der Berechnung nicht berücksichtigt werden sollen. Ich habe in das Feld PROZENT **,2** eingegeben. Dies weist die Funktion `GESTUTZTMITTEL` an, die äußeren 20 % der Werte bei der Berechnung des Mittelwerts auszuschließen. Die äußeren 20 % bedeutet 10 % der größten und 10 % kleinste Werte. In Abbildung 4.5 sind das Dialogfeld, der Wertebereich und die markierte Zelle dargestellt. Im Dialogfeld wird der Mittelwert ohne Ausreißer (567,375) angezeigt.

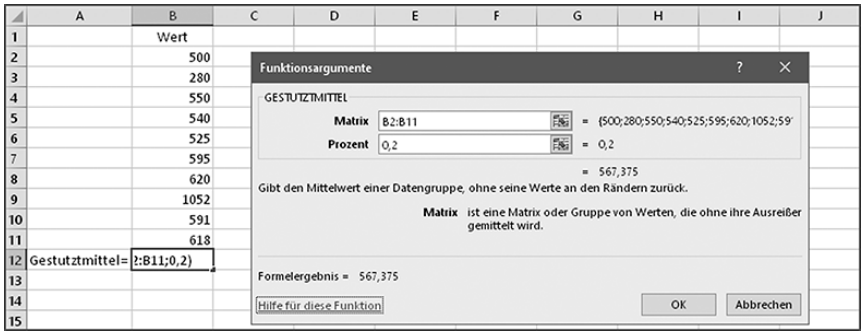


Abbildung 4.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *GESTUTZTMITTEL* mit dem Zellbereich und der markierten Zelle

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen; das Ergebnis wird in der ausgewählten Zelle angezeigt.**

Der Parametername *Prozent* ist hier etwas irreführend. Sie müssen den Prozentsatz als Dezimalwert angeben. Also geben Sie im Feld PROZENT anstelle von 20 den Wert ,2 ein, wenn Sie die äußeren 20 % ausschließen möchten. (Kurze Frage: Welchem Ergebnis entspricht es, wenn Sie im Feld PROZENT 0 eingeben? Antwort: `MITTELWERT(B2:B11)`.)

Welchen Prozentsatz der Werte sollten Sie ausschließen? Das bleibt Ihnen überlassen. Das hängt davon ab, was Sie messen, wie extrem die Werte sein können und wie genau Sie den Bereich kennen, den Sie untersuchen. Wenn Sie Werte ohne Extremwerte mitteln und den Mittelwert veröffentlichen, müssen Sie angeben, dass Sie Werte ausgeschlossen haben, und Sie müssen angeben, welchen Prozentsatz Sie ausgeschlossen haben.

Im Abschnitt über den Median zeige ich Ihnen eine andere Möglichkeit für den Umgang mit Extremwerten.

Und schließlich noch ein paar andere Mittel

In diesem Abschnitt werden zwei Arten von Durchschnittwerten beschrieben, die sich von dem unterscheiden, den Sie bereits kennen. Ich werde diese erläutern, weil sie Ihnen möglicherweise bei den statistischen Funktionen von Excel begegnen werden. (Wie viele unterschiedliche Arten Durchschnittswerte kann es überhaupt geben? Die Mathematiker im alten Griechenland haben ganze elf gezählt!)

Geometrisches Mittel

Nehmen wir einmal an, Sie haben eine Kapitalanlage mit einer Laufzeit von zwei Jahren, die im ersten Jahr 25 % und im zweiten Jahr 75 % abwirft. (Wenn Sie die tatsächlich haben, möchte ich mehr darüber erfahren!) Wie hoch ist die durchschnittliche jährliche Rendite?

Um diese Frage zu beantworten, werden Sie geneigt sein, den Mittelwert aus 25 und 75 zu berechnen (was übrigens 50 ist). Dabei vergessen Sie jedoch ein wichtiges Detail: Am Ende des ersten Jahres addieren Sie nicht 1,25 zu Ihrer Kapitalanlage, sondern Sie *multiplizieren* Ihre Kapitalanlage mit 1,25. Am Ende des zweiten Jahres multiplizieren Sie das Ergebnis des ersten Jahres mit 1,75.

Der normale Feld-Wald-und-Wiesen-Durchschnitt entspricht nicht der durchschnittlichen Rendite. Der Mittelwert wird vielmehr wie folgt berechnet:

$$\text{Durchschnittliche Rendite} = \sqrt{1,25 \times 1,75} = 1,479$$

Die durchschnittliche Rendite beträgt etwa 47,9 % und nicht 50 %. Diese Art von Durchschnitt wird als *geometrisches Mittel* bezeichnet.

Bei diesem Beispiel ist das geometrische Mittel die Quadratwurzel aus dem Produkt zweier Zahlen. Bei drei Zahlen ist das geometrische Mittel die Kubikwurzel aus dem Produkt der drei Zahlen. Bei vier Zahlen die vierte Wurzel aus deren Produkt usw. Allgemein ausgedrückt ist das geometrische Mittel von N Zahlen die Nte Wurzel aus deren Produkt.

Mit der Tabellenblattfunktion `GEOMITTEL` von Excel können Sie das geometrische Mittel einer Zahlengruppe berechnen. Gehen Sie dabei genauso vor wie bei der Funktion `MITTELWERT`, wählen Sie jedoch im Menü der statistischen Funktionen die Funktion GEOMITTEL aus.

Harmonisches Mittel

Noch ein anderer Mittelwert ist einer, der Ihnen begegnen wird, wenn Sie die Art von Aufgaben lösen müssen, wie sie in Algebra-Übungsbüchern zuhauf vorkommen.

Nehmen wir beispielsweise an, Sie fahren an einem Morgen gemütlich zur Arbeit und fahren von zu Hause bis zur Arbeitsstelle mit einer Geschwindigkeit von 40 km/h. Am gleichen Tag möchten Sie schnell nach Hause kommen. Auf dem Nachhauseweg (mit genau derselben Entfernung) fahren Sie daher von Ihrer Arbeitsstelle bis nach Hause mit 60 km/h. Wie hoch ist Ihre durchschnittliche Geschwindigkeit für die Gesamtzeit, die Sie auf der Straße verbringen?

Jedenfalls nicht 50 km/h, da Sie für die beiden Strecken unterschiedlich lange brauchen. Ohne zu weit ins Detail zu gehen,

kann ich Ihnen sagen, dass die Formel für diese Berechnung wie folgt lautet:

$$\frac{1}{\text{Durchschnitt}} = \frac{1}{2}\left[\frac{1}{40} + \frac{1}{60}\right] = \frac{1}{48}$$

Der Durchschnittwert beträgt bei diesem Beispiel 48. Diese Art von Durchschnitt wird als *harmonisches Mittel* bezeichnet. Ich habe Ihnen gezeigt, wie Sie diesen Mittelwert für zwei Zahlen berechnen. Sie können diesen Mittelwert jedoch für jede beliebige Anzahl von Zahlen berechnen. Setzen Sie einfach die einzelnen Zahlen jeweils in den Nenner eines Bruchs, wobei die 1 im Zähler steht. Mathematiker sprechen in diesem Zusammenhang vom *reziproken Wert* einer Zahl. 1/40 ist der reziproke Wert von 40. Addieren Sie alle reziproken Werte und bilden Sie deren Durchschnitt. Das Ergebnis ist der reziproke Wert des harmonischen Mittels.

Für den seltenen Fall, dass Sie tatsächlich einmal einen dieser Werte berechnen müssen, übernimmt Excel die dröge Rechenarbeit für Sie. Mit der Tabellenblattfunktion `HARMITTEL` können Sie das harmonische Mittel einer Zahlengruppe berechnen. Gehen Sie dabei genauso vor wie bei der Funktion `MITTELWERT`, wählen Sie jedoch im Menü der statistischen Funktionen die Funktion HARMITTEL aus.

Mediane: Auf halber Strecke erwischt

Die Ermittlung des Mittelwerts ist hilfreich, um eine Zahlengruppe zusammenzufassen. Dabei werden jedoch auch Extremwerte berücksichtigt: Wenn eine Zahl im Vergleich zu den anderen Zahlen aus dem Lot ist, ist schnell der ganze Mittelwert aus dem Lot. In diesem Fall ist der Mittelwert kein guter Repräsentant der Gruppe.

Wenn die folgenden Zahlen beispielsweise die Lesegeschwindigkeit (in Wörtern pro Minute) für eine Gruppe Kinder angeben

56, 78, 45, 49, 55, 62

beträgt der Mittelwert 57,5. Nehmen wir an, das Kind, das 78 Wörter pro Minute liest, verlässt die Gruppe und wird durch ein außergewöhnlich schnell lesendes Kind ersetzt. Dieses Kind liest mit einer Geschwindigkeit von 180 Wörtern pro Minute. Nun lauten die Lesegeschwindigkeiten der Gruppe wie folgt:

56, 180, 45, 49, 55, 62

Der neue Durchschnittwert beträgt nun 74,5. Dieser Wert ist irreführend, da außer dem neu hinzugekommenen Kind keins in der Gruppe auch nur annähernd so schnell liest. In einem Fall wie diesem sollte für die zentrale Tendenz ein anderes Maß verwendet werden: der *Median*.

Einfach ausgedrückt ist der Median der mittlere Wert in einer Gruppe von Zahlen. Wenn Sie die Zahlen sortieren, ist der Median der Wert, unter dem sich die eine Hälfte der Zahlen und über dem sich die andere Hälfte der Zahlen befinden.

Den Median ermitteln

In unserem Beispiel beträgt die Lesegeschwindigkeit der ersten Gruppe (in aufsteigender Folge)

45, 49, 55, 56, 62, 78

Der Median liegt direkt in der Mitte zwischen 55 und 56. Er beträgt somit 55,5.

Und wie sieht das mit der Gruppe mit dem neu hinzugekommenen Kind aus? Die Werte lauten

45, 49, 55, 56, 62, 180

Der Median beträgt nach wie vor 55,5. Der Median wird durch den Extremwert nicht verändert.

MEDIAN

Mit der Tabellenblattfunktion MEDIAN (Sie ahnen es schon) können Sie den Median einer Zahlengruppe berechnen. Und so müssen Sie vorgehen:

1. **Geben Sie die gewünschten Werte in einen Zellbereich ein und markieren Sie eine Zelle für das Ergebnis.**

   Ich habe in diesem Beispiel die Zahlen 45, 49, 55, 56, 62 und 78 in die Zellen B2 bis B7 eingegeben und B8 für den Median markiert. Ich habe die Zahlen in aufsteigender Folge angeordnet. Das brauchen Sie jedoch nicht zu tun, um die Funktion MEDIAN zu verwenden.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MEDIAN, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Das Dialogfeld FUNKTIONSARGUMENTE wird geöffnet und der Zellbereich wird im Feld ZAHL1 angezeigt. Der Median wird unten im Dialogfeld angezeigt. (Er beträgt in diesem Beispiel 55,5.) In Abbildung 4.6 sind das Dialogfeld, der Zellbereich und die markierte Zelle dargestellt.

Abbildung 4.6: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *MEDIAN* mit dem Zellbereich und der markierten Zelle für das Ergebnis.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen; der Median wird in der markierten Zelle angezeigt.**

Zur Übung ersetzen Sie die 78 in B7 durch 180. Sie werden sehen: Der Median bleibt derselbe.

Der Modalwert

Es gibt noch ein weiteres für die zentrale Tendenz wichtiges Maß. Hierbei handelt es sich um den Wert, der in einer Wertegruppe am häufigsten vorkommt. Dieser Wert wird als *Modalwert* bezeichnet.

Den Modalwert ermitteln

Das Ermitteln des Modalwerts ist nicht schwierig. Betrachten Sie die Werte, suchen Sie nach demjenigen, der am häufigsten vorkommt, und schon haben Sie den Modalwert ermittelt. Zwei Werte kommen dafür in Frage? In diesem Fall weist der Wertebereich zwei Modalwerte auf. (Der Fachausdruck hierfür lautet *bimodaler Wert*.)

Kann es mehr als zwei Modalwerte geben? Absolut.

Stellen Sie sich vor, jeder Wert kommt gleich oft vor. Dann gibt es keinen Modalwert.

Manchmal ist der Modalwert der repräsentativste Wert der zentralen Tendenz. Stellen Sie sich ein kleines Unternehmen mit 30 Beratern und zwei Geschäftsführern vor. Jeder Berater hat ein Jahresgehalt von 40.000 €. Jeder Geschäftsführer hat ein Jahresgehalt von 250.000 €. Das Durchschnittsgehalt in diesem Unternehmen beträgt 53.125 €.

Vermittelt der Durchschnittswert eine klare Vorstellung von der Gehaltsstruktur dieses Unternehmens? Würde der Durchschnittswert Ihre Erwartungshaltung beeinflussen, wenn Sie sich bei diesem Unternehmen bewerben würden? Mit dem Modalwert, der in diesem Fall 40.000 € beträgt, wären Sie wohl besser beraten.

MODUS.EINF und MODUS.VIELF

Mit der Excel-Funktion MODUS.EINF können Sie einen einzelnen Modalwert ermitteln:

1. **Geben Sie Ihre Daten in einen Zellbereich ein und markieren Sie eine Zelle für das Ergebnis.**

   Ich habe in diesem Beispiel die Zahlen 56, 23, 77, 75, 57, 75, 91, 59 und 75 benutzt. Die Zahlen habe ich in die Zellen B2 bis B10 eingegeben, und ich habe die Zelle B11 für den Modalwert markiert.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MODUS.EINF, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 4.7).**

Abbildung 4.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *MODUS.EINF* mit dem Zellbereich und der markierten Zelle für das Ergebnis

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Ich habe in das Feld ZAHL1 den Zellbereich B2:B10 eingegeben. Der Modalwert (in diesem Beispiel 75) wird im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen; das Ergebnis wird in der markierten Zelle angezeigt.**

Wenn eine Wertegruppe mehr als einen Modalwert besitzt (es handelt sich dann um eine multimodale Verteilung), verwenden Sie die Excel-Funktion `MODUS.VIELF`. Hierbei handelt es sich um eine Matrixfunktion: Sie gibt (möglicherweise) ein Array/eine Matrix mit Ergebnissen zurück und nicht nur ein Ergebnis. Sie markieren einen Zellbereich für die Ergebnissen und drücken, nachdem Sie die Parameter im Dialogfeld Funktionsargumente eingegeben haben, die Tastenkombination Strg+Shift+Enter, um das Array zu füllen.

Hier ist ein Beispiel für die Verwendung der Funktion `MODUS.VIELF`:

1. **Geben Sie die Daten in das Tabellenblatt ein und markieren Sie ein vertikales Array von Zellen für die Ergebnisse.**

   Für dieses Beispiel habe ich in die Zellen B2:B10 die Werte 57, 23, 77, 75, 57, 75, 91, 57 und 75 eingegeben. Als Ergebnisbereich habe ich die Zellen B11:B14 markiert. Beachten Sie, dass diese Wertegruppe zwei Modalwerte besitzt: 57 und 75.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf MODUS.VIELF, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 4.8).**

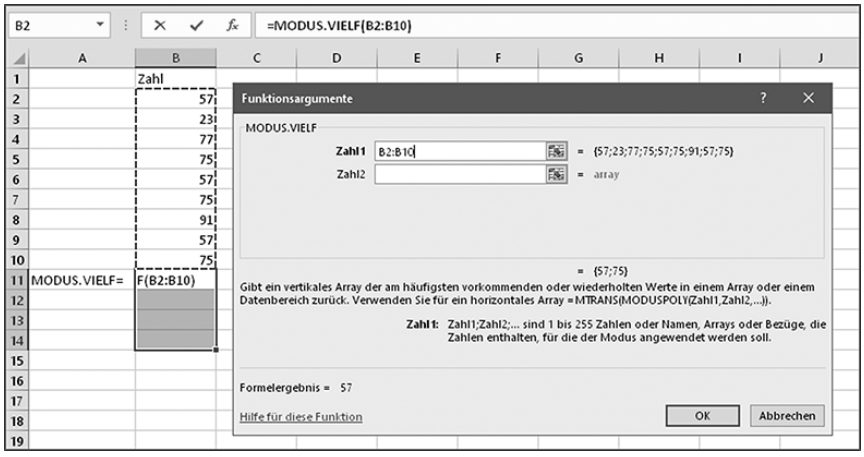


Abbildung 4.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *MODUS.VIELF*, mit dem Array der Zelldaten und dem Zellarray für die Ergebnisse.

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Wichtig: Klicken Sie nicht auf OK.

4. **Da es sich hierbei um eine Arrayformel handelt, drücken Sie Strg+Shift+Enter, um die Ergebnisse von `MODUS.VIELF` in das markierte Array einzufügen.**

   Im Dialogfeld finden Sie keinen Hinweis darauf, dass Sie dies gemacht haben.

   Abbildung 4.9 zeigt, was passiert, nachdem Sie Strg+Shift+Enter gedrückt haben. Da ich für das Ergebnis vier Zellen reserviert habe, es jedoch in der Datenreihe nur zwei Modalwerte gibt, werden in den beiden verbleibenden Zellen Fehlermeldungen angezeigt.

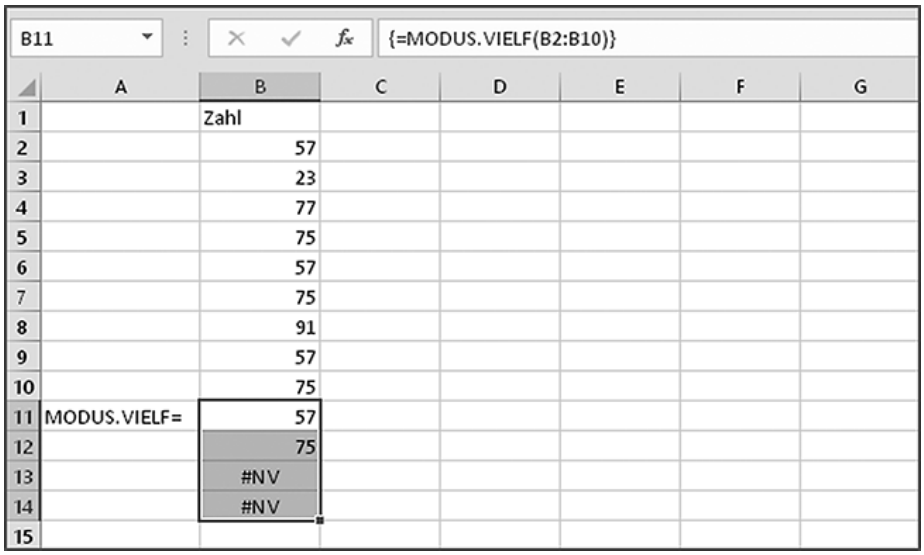


Abbildung 4.9: Die Ergebnisse von *MODUS.VIELF*. Die geschweiften Klammern in der Bearbeitungsleiste zeigen an, dass es sich um eine Arrayformel handelt.

Was passiert, wenn Sie `MODUS.EINF` bei einer multimodalen Datenreihe verwenden? Das Ergebnis ist der Modalwert, der zuerst in der Datenreihe auftaucht. Was passiert, wenn Sie `MODUS.VIELF`

bei einer Datenreihe verwenden, die nur einen Modalwert enthält? Meine erste Vermutung war, dass dieser eine Modalwerte in der ersten Zelle des Arrays angezeigt wird und dass die weiteren Zellen Fehlermeldungen enthalten. Stimmt aber nicht. `MODUS.VIELF` füllt das gesamte Array mit diesem Modalwert, und zwar auch dann, wenn die Anzahl der Zellen des Arrays größer ist als die Anzahl der Vorkommen des Modalwerts. Ja, das hat mich auch überrascht.

Falls Sie also eine lange Zahlenspalte habe und Sie den beziehungsweise die Modalwerte ermitteln wollen, verwenden Sie `MODUS.VIELF`. Das Schlimmste, was passieren kann, ist, dass Sie im Ergebnisarray ein paar Zellen mit Fehlermeldungen erhalten. Der Vorteil ist, dass Ihnen so keine Modalwerte entgehen.

Und das war der längste Text, den ich jemals über Modalwerte geschrieben habe,

Kapitel 5

Abweichungen vom Durchschnitt

In diesem Kapitel

- Das hat es mit der Streuung auf sich
- Varianz und Standardabweichung verwenden
- Tabellenblattfunktionen von Excel zum Berechnen der Streuung
- Hilfsroutinen für fehlender Tabellenblattfunktionen verwenden
- Weitere Tabellenblattfunktionen für die Streuung

Drei kluge Beschreibungen von Statistikern:

Beschreibung Nr. 1: »Ein Statistiker ist eine Person, die in einem Eimer mit eiskaltem Wasser steht, den Kopf in einen Backofen hält und sagt: ›Im Durchschnitt geht es mir gut.‹« (K. Dunning)

Beschreibung Nr. 2: »Ein Statistiker ist beim Überqueren eines Flusses mit einer durchschnittlichen Tiefe von 15 cm ertrunken.« (Anonym)

Beschreibung Nr. 3: »Drei Statistiker gehen mit Pfeil und Bogen auf die Jagd. Sie treffen auf einen kapitalen Bock und nehmen ihn ins Visier. Einer schießt und trifft drei Meter links daneben. Der zweite schießt und trifft drei Meter rechts daneben. Der dritte springt vor Freude in die Luft und ruft: ›Wir haben ihn! Wir haben ihn!‹« (Bill Butz, von Diana McLellan im *Washingtonian* zitiert)

Was haben diese Beschreibungen gemeinsam? Mit dem Berechnen des Mittelwerts lassen sich mehrere Zahlen gut zusammenfassen. Damit erhalten Sie jedoch nicht all die Informationen, die häufig erforderlich sind. Wenn Sie nur den Mittelwert heranziehen, übersehen Sie möglicherweise etwas Wichtiges.

Damit Sie nichts Wichtiges übersehen, brauchen Sie eine andere Art der statistischen Berechnung: eine, mit der die *Streuung* berechnet wird. Dabei geht es darum festzustellen, um wie viel die einzelnen Zahlen einer Zahlengruppe vom Gruppendurchschnitt abweichen. Zum Berechnen der Streuung gibt es mehrere Möglichkeiten. Alle beruhen auf demselben Prinzip: Je größer der berechnete Wert, umso mehr weichen die Zahlen vom Mittelwert ab. Je kleiner der Wert, umso weniger weichen sie ab.

Die Streuung berechnen

Nehmen wir einmal an, Sie messen, wie groß die Kinder einer Gruppe sind. Dabei ergeben sich folgende Größen (in cm):

122, 122, 122, 122, 122

Nun messen Sie, wie groß die Kinder einer anderen Gruppe sind. Dabei ergeben sich folgende Größen:

127, 120, 132, 117, 114

Wenn Sie den Mittelwert der beiden Gruppen berechnen, werden Sie feststellen, dass er bei beiden Gruppen gleich ist: 122 cm. Wenn Sie die Zahlen der beiden Gruppen betrachten, sehen Sie jedoch, dass sie unterschiedlich sind: Die Zahlen in der ersten Gruppe sind identisch, während die in der zweiten Gruppe etwas variieren.

Mittelwert von quadratischen Abweichungen: Varianz, und wie sie berechnet wird

Eine Möglichkeit, den Unterschied zwischen den beiden Gruppen darzustellen, besteht darin, die Abweichungen innerhalb der Gruppen zu untersuchen. Sie können sich eine *Abweichung* als die Differenz zwischen einem Wert und dem Mittelwert aller Werte in einer Gruppe vorstellen.

Und das meine ich damit: In Tabelle 5.1 sind die Werte der ersten Gruppe sowie deren Abweichungen dargestellt.

Größe	Größe – Mittelwert	Abweichung
122	122 – 122	0
122	122 – 122	0
122	122 – 122	0
122	122 – 122	0
122	122 – 122	0

Tabelle 5.1: Die Größen und Abweichungen der ersten Gruppe

Nun kann beispielsweise der Mittelwert der Abweichungen gebildet werden. Der Mittelwert der Zahlen in der Spalte Abweichung ist natürlich null.

In Tabelle 5.2 sind die Werte der zweiten Gruppe sowie deren Abweichungen dargestellt.

Größe	Größe – Mittelwert	Abweichung
127	127 – 122	5
120	120 – 122	–2
132	132 – 122	10
117	117 – 122	–5
114	114 – 122	–8

Tabelle 5.2: Die Größen und Abweichungen der zweiten Gruppe

Wie sieht es mit dem Mittelwert der Abweichungen in Tabelle 5.2 aus? Der beträgt … null!

Hmmm … Was nun?

Das Ermitteln der mittleren Abweichung trägt nicht dazu bei, einen Unterschied zwischen zwei Gruppen zu erkennen, da die mittlere Abweichung vom Mittelwert bei jeder Zahlengruppe *immer* null ist. Die alten Hasen unter den Statistikern werden Ihnen sogar sagen, dass es sich hierbei um eine zentrale Eigenschaft des Mittelwerts handelt.

Das Problem sind die negativen Zahlen. Wie gehen Statistiker damit um?

Vielleicht erinnern Sie sich noch an einen Trick aus der Algebra: Minus mal minus gibt plus. Kommt Ihnen das bekannt vor?

Bedeutet das also, dass jede Abweichung mit sich selbst multipliziert und anschließend aus den Ergebnissen der Mittelwert gebildet wird? Absolut. Das Multiplizieren einer Abweichung mit sich selbst wird als *Quadrierung einer Abweichung* bezeichnet. Der Mittelwert der quadratischen Abweichung ist so wichtig, dass er einen eigenen Namen bekommt: *Varianz*.

In Tabelle 5.3 sind die Werte der Gruppe aus Tabelle 5.2 mit den Abweichungen und den quadratischen Abweichungen dargestellt.

Größe	Größe – Mittelwert	Abweichung	Quadratische Abweichung
127	127 – 122	5	25
120	120 – 122	–2	4
132	132 – 122	10	100
117	117 – 122	–5	25
114	114 – 122	–8	64

Tabelle 5.3: Die Größen und quadratischen Abweichungen der zweiten Gruppe

Die Varianz, also der Mittelwert der quadratischen Abweichungen für diese Gruppe, beträgt (25 + 4 + 100 + 25 + 64)/5 = 218/5 = 43,6. Dieser Wert unterscheidet sich natürlich erheblich von dem der ersten Gruppe, bei der die Varianz null beträgt.

Um die Formel für die Varianz zu entwickeln und um Ihnen zu zeigen, wie sie funktioniert, werde ich verschiedene Symbole verwenden. *X* steht für die Größe in der ersten Spalte der Tabelle und $\bar{X}$ repräsentiert den Mittelwert. Da eine Abweichung das Ergebnis aus der Subtraktion des Mittelwerts von den einzelnen Zahlen ist, wird die Abweichung durch

$$(X - \bar{X})$$

ausgedrückt. Die Multiplikation einer Abweichung mit sich selbst? Das ist einfach

$$(X - \bar{X})^2$$

Zum Berechnen der Varianz quadrieren Sie die einzelnen Abweichungen, addieren die jeweiligen Ergebnisse und ermitteln den Durchschnitt der quadratischen Abweichungen. Wenn *n* für die Anzahl der quadratischen Abweichungen steht (in unserem Beispiel also fünf), lautet die Formel zum Berechnen der Varianz folgendermaßen:

$$\text{Varianz} = \frac{\Sigma(X - \bar{X})^2}{n}$$

Σ ist der griechische Großbuchstabe Sigma und steht für Summe.

Welches Symbol wird nun für die Varianz verwendet? Wie in Kapitel 1 bereits erwähnt, werden griechische Buchstaben für Parameter von Grundgesamtheiten und lateinische Buchstaben für Statistiken verwendet. Stellen Sie sich vor, unsere kleine Gruppe aus fünf

Zahlen stellt die gesamte Grundgesamtheit dar. Gibt es im griechischen Alphabet einen Buchstaben, der dem *V* entspricht, so wie *μ* (das Symbol für den Mittelwert der Grundgesamtheit) dem *M* entspricht?

Leider nicht. Stattdessen verwenden wir den dem Sigma entsprechenden Kleinbuchstaben! Dieser sieht so aus: σ. Da es hier jedoch um quadratische Werte geht, verwenden wir das Symbol σ^2.

Somit lautet die Formel zum Berechnen der Varianz (einer Grundgesamtheit) folgendermaßen

$$\sigma^2 = \frac{\Sigma(X - \bar{X})^2}{N}$$

Die Varianz ist groß, wenn die Zahlen in einer Gruppe erheblich vom Mittelwert abweichen. Bei geringen Abweichungen ist die Varianz klein.

Die Varianz, die wir uns gerade erarbeitet haben, stimmt, wenn die Gruppe aus fünf Werten eine Grundgesamtheit darstellt. Bedeutet das, dass die Varianz für eine Stichprobe anders aussieht? Ja, das tut sie. Und Sie werden gleich sehen, warum das so ist. Zuvor kehren wir jedoch zu Excel zurück.

VAR.P und VARIANZENA

Mit den beiden Tabellenblattfunktionen `VAR.P` und `VARIANZENA` von Excel können Sie die Varianz einer Grundgesamtheit berechnen. (Das P im Funktionsnamen steht für Population.)

Beginnen wir mit `VAR.P`. In Abbildung 5.1 sind die Daten und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `VAR.P` dargestellt.

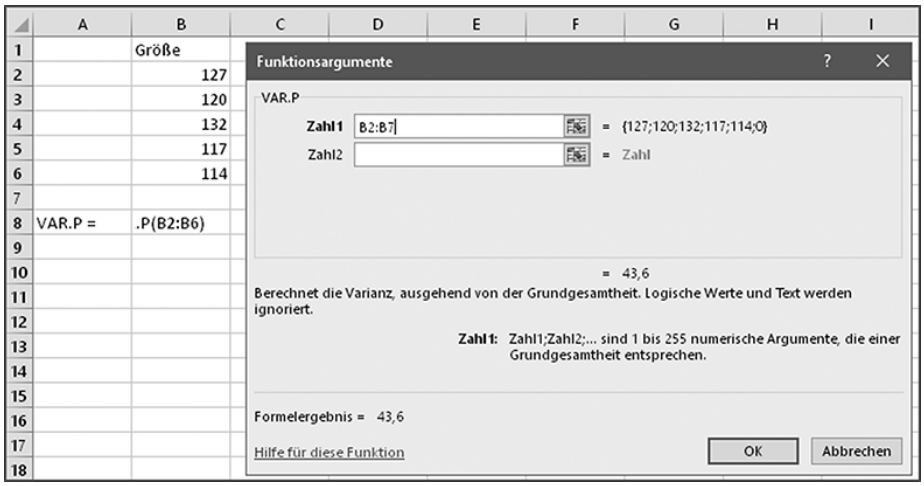


Abbildung 5.1: Die Funktion `VAR.P` verwenden

Führen Sie diese Schritte durch:

1. **Geben Sie die Daten in ein Arbeitsblatt ein und markieren Sie eine Zelle zum Anzeigen des Ergebnisses.**

   In Abbildung 5.1 sehen Sie, dass ich in diesem Beispiel die Zahlen 127, 120, 132, 117 und 114 in die Zellen B2 bis B6 eingegeben und die Zelle B8 für das Ergebnis ausgewählt habe.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf VAR.P, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld ZAHL1 habe ich den Zellbereich B2:B7 und nicht B2:B6 eingegeben. Dies habe ich gemacht, um zu zeigen, wie `VAR.P` mit Leerzellen umgeht. Die Varianz der Grundgesamtheit wird im Dialogfeld mit 43,6 angezeigt.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

Hätte ich dem Zellbereich B2:B7 den Namen Größe zugewiesen (siehe Kapitel 2), dann würde die folgende Formel in der Bearbeitungsleiste stehen:

```
=VAR.P(Größe)
```

Beim Berechnen der Varianz in einem Zellbereich mit VAR.P werden nur Zahlen berücksichtigt. Falls sich in einigen Zellen Text oder logische Werte befinden, werden diese ignoriert.

Nicht so VARIANZENA. VARIANZENA berücksichtigt bei der Berechnung der Varianz Text und logische Werte. Wie? Wenn eine Zelle Text enthält, wird diese so behandelt, als enthielte sie den Wert Null. Wenn eine Zelle den logischen Wert FALSCH enthält, zählt auch dieser bei VARIANZENA als Null. Der logische Wert WAHR zählt dagegen 1. Diese Nullen und Einsen werden hinzugezählt und wirken sich auf den Mittelwert und die Varianz aus.

Damit Sie sehen, was das bedeutet, habe ich die Zahlen in den Zellen B2 bis B6 unverändert übernommen und wieder Zelle B8 markiert. Dann bin ich genauso vorgegangen wie bei VAR.P, habe jedoch diesmal die Funktion VARIANZENA ausgewählt. Im Dialogfeld VARIANZENA gebe ich im Feld ZAHL1 **B2:B7** (B7, *nicht* B6) ein und klicke auf OK. In Zelle B8 wird dasselbe Ergebnis wie zuvor angezeigt, da VARIANZENA die leere Zelle B7 als »kein Eintrag« interpretiert.

Wenn ich nun in Zelle B7 **WAHR** eingebe, wird in B8 ein anderes Ergebnis angezeigt, da VARIANZENA B7 als 1 interpretiert (siehe Abbildung 5.2).

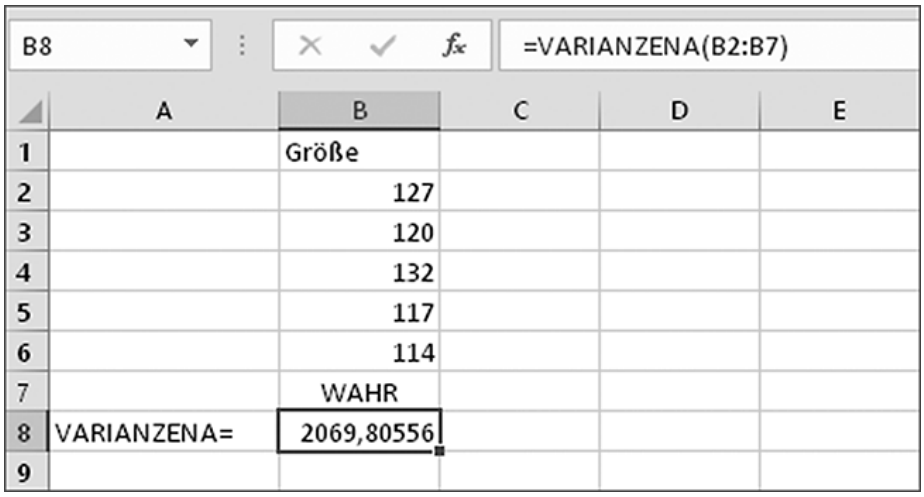


Abbildung 5.2: `VARIANZENA` interpretiert WAHR als 1. Daher wird ein anderes Ergebnis als das in Abbildung 5.1 angezeigt.

Geben Sie nun `FALSCH` (oder ein anderes Wort als `WAHR`) in B7 ein. Wieder wird ein anderer Wert in B8 angezeigt. Diesmal interpretiert `VARIANZENA` die Zelle B7 als null.

Varianz einer Stichprobe

Weiter oben habe ich erwähnt, dass diese Formel zum Berechnen der Varianz einer Grundgesamtheit verwendet wird:

$$\sigma^2 = \frac{\Sigma(X - \bar{X})^2}{n}$$

Ich habe außerdem darauf hingewiesen, dass die Varianz einer Stichprobe etwas anders berechnet wird. Und so unterscheiden sich die beiden Berechnungen: Wenn es sich bei den Zahlen um eine Stichprobe aus einer größeren Grundgesamtheit handelt, wollen Sie höchstwahrscheinlich mit der Varianz dieser Stichprobe die Varianz der Grundgesamtheit schätzen.

Mit der Formel, die Sie zum Berechnen der Varianz verwendet haben, können Sie aber nicht wirklich auf die Varianz der Grundgesamtheit schließen. Vom Mittelwert einer Stichprobe lässt sich zwar auf den Mittelwert der Grundgesamtheit schließen, bei der Varianz ist das so jedoch nicht möglich, und zwar aus Gründen, die den Rahmen dieses Buches bei weitem sprengen würden.

Wie wird ein guter Schätzwert für die Varianz der Grundgesamtheit berechnet? Das ist ganz einfach. Anstelle von *n* wird im Nenner einfach *n - 1* verwendet. (Die Gründe hierfür sollen uns im Rahmen dieses Buches nicht interessieren.)

Da wir es hier mit der Eigenschaft einer Stichprobe (nicht mit der einer Grundgesamtheit) zu tun haben, verwenden wir das lateinische Pendant des griechischen Buchstabens, also *s* statt σ. Somit lautet die Formel für die Berechnung der Stichprobenvarianz wie folgt:

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

Mit den quadratischen Abweichungen für unsere fünf Zahlen ergibt sich für den Wert von s^2

(25 + 4 + 100 + 25 + 64)/4 = 218/4 = 54,5

Wenn also die Zahlen

127, 120, 132, 117, 114

eine ganze Grundgesamtheit darstellen, beträgt die Varianz 43,6. Wenn es sich bei den Zahlen um eine Stichprobe aus einer größeren Grundgesamtheit handelt, beträgt unsere beste Schätzung für die Varianz dieser Grundgesamtheit 54,5.

VAR.S und VARIANZA

Mit den Tabellenblattfunktionen `VAR.S` und `VARIANZA` wird die Varianz einer Stichprobe berechnet.

In Abbildung 5.3 ist das Dialogfeld FUNKTIONSARGUMENTE für `VAR.S` dargestellt, wobei in die Zellen B2 bis B6 die Zahlen 127, 120, 132, 117 und 114 eingegeben wurden. Zelle B7 ist ebenfalls Teil des Zellbereichs; diese Zelle ist jedoch absichtlich leer gelassen.

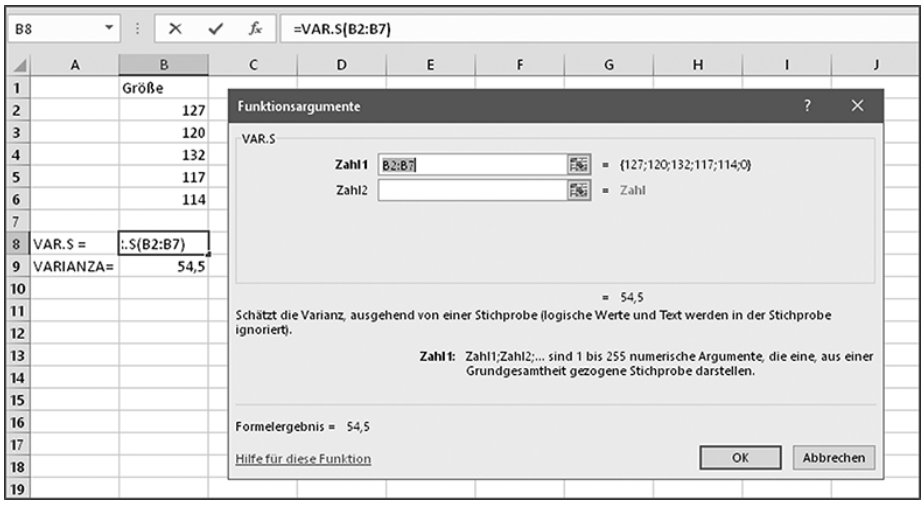


Abbildung 5.3: Die Funktion `VAR.S` im Einsatz

Die Beziehung zwischen `VAR.S` und `VARIANZA` ist dieselbe wie die zwischen `VAR.P` und `VARIANZENA`. `VAR.S` ignoriert Zellen mit logischen Werten (`WAHR` und `FALSCH`) und mit Text. Bei `VARIANZA` werden diese Zellen jedoch berücksichtigt. Dabei wird `WAHR` als 1 und `FALSCH` als 0 interpretiert. Text in einer Zelle wird von `VARIANZA` als 0 interpretiert.

Dies ist der Grund dafür, dass ich Zelle B7 leer gelassen haben. Wenn Sie ein wenig mit `VARIANZA` experimentieren und in Zelle B7 logische Werte oder Text eingeben, sehen Sie genau, was `VARIANZA` macht.

Zurück zu den Wurzeln: Standardabweichung

Beim Berechnen der Varianz einer Zahlenreihe ergibt sich ein Wert mit einer Einheit, die sich von der der ursprünglichen Werte unterscheidet. Wenn die ursprünglichen Werte beispielsweise die Einheit *cm* hatten, so hat deren Varianz die Einheit cm^2. Das kommt

daher, weil die Abweichungen vor dem Bilden des Mittelwerts quadriert werden.

In der Regel ist es anschaulicher, wenn das Ergebnis einer Variationsberechnung dieselbe Einheit wie die ursprünglichen Werte aufweist. Dies lässt sich ganz einfach erreichen. Sie brauchen lediglich die Quadratwurzel der Varianz zu ziehen.

Wie die Varianz, so ist auch das Ergebnis dieser Quadratwurzel so wichtig, dass es einen eigenen Namen bekommt: *Standardabweichung*.

Standardabweichung einer Grundgesamtheit

Die Standardabweichung einer Grundgesamtheit ist die Quadratwurzel der Varianz dieser Grundgesamtheit. Für die Standardabweichung einer Grundgesamtheit wird das Symbol σ (Sigma) verwendet. Die Formel für die Standardabweichung lautet:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X - \bar{X})^2}{n}}$$

Für die folgenden Werte (in cm)

127, 120, 132, 117, 114

beträgt die Varianz der Grundgesamtheit 43,6 cm^2 und die Standardabweichung der Grundgesamtheit 6,60 cm (gerundet).

STABW.N und STABWNA

Mit den Excel-Tabellenblattfunktionen `STABW.N` und `STABWNA` können Sie die Standardabweichung einer Grundgesamtheit berechnen. Gehen Sie dazu wie folgt vor:

1. **Geben Sie die gewünschten Werte in einen Zellbereich ein, und markieren Sie eine Zelle für das Ergebnis.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf STABW.N, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   Nachdem Sie den Zellbereich in das Feld ZAHL1 eingegeben haben, wird im Dialogfeld die berechnete Standardabweichung angezeigt. Dies können Sie in Abbildung 5.4 sehen.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

Wie `VARIANZENA` berücksichtigt auch `STABWNA` logische Werte und Text bei der Berechnung der Standardabweichung einer Grundgesamtheit. Dabei wird `WAHR` als 1 und `FALSCH` als 0 interpretiert. Text in einer Zelle wird als 0 interpretiert.

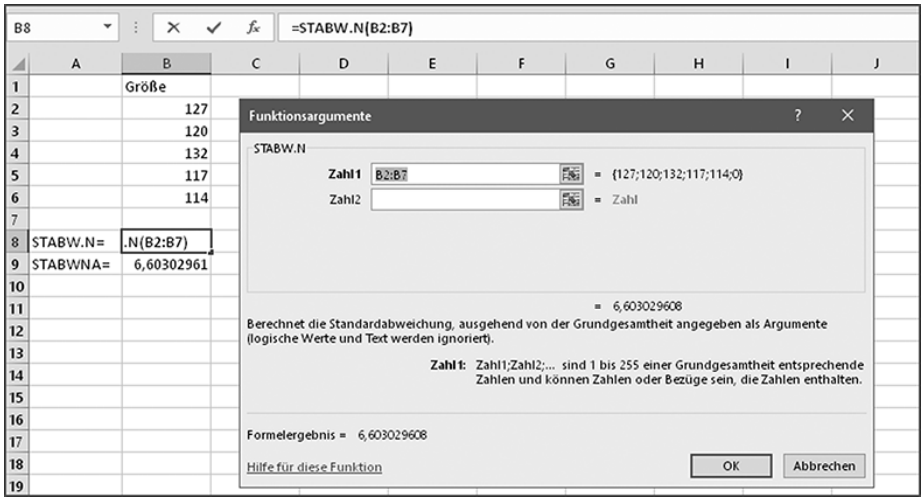


Abbildung 5.4: Das Dialogfeld FUNKTIONSARGUMENTE für `STABW.N` mit den Daten

Standardabweichung einer Stichprobe

Die Standardabweichung einer Stichprobe, also die Schätzung der Standardabweichung einer Grundgesamtheit, ist die Quadratwurzel

der Stichprobenvarianz. Dafür wird das Symbol *s* verwendet und die Formel lautet wie folgt:

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

Für die folgenden Werte (in cm)

127, 120, 132, 117 und 114

beträgt die Varianz der Stichprobe 54,5 cm^2 und die Standardabweichung der Stichprobe 7,38 cm (gerundet).

STABW.S und STABWA

Mit den Excel-Tabellenblattfunktionen `STABW.S` und `STABWA` von Excel können Sie die Standardabweichung einer Stichprobe berechnen. So verwenden Sie die Funktion `STABW.S`:

1. **Geben Sie die gewünschten Werte in einen Zellbereich ein, und markieren Sie eine Zelle für das Ergebnis.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf STABW.S, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   Nachdem Sie den Zellbereich in das Feld ZAHL1 eingegeben haben, wird im Dialogfeld die Standardabweichung der Stichprobe angezeigt. Dies können Sie in Abbildung 5.5 sehen.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

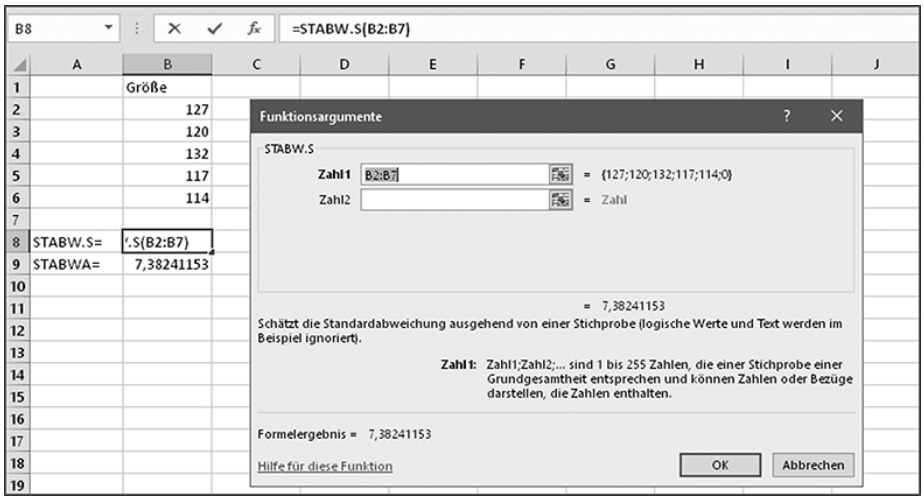


Abbildung 5.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion STABW.S

STABWA berücksichtigt Text und logische Werte bei der Berechnung. Zellen mit Text haben den Wert 0 und Zellen, deren Wert FALSCH ist, werden ebenfalls als 0 interpretiert. Zellen mit dem Wert WAHR werden als 1 interpretiert.

Die fehlenden Funktionen: STABWWENN und STABWWENNS

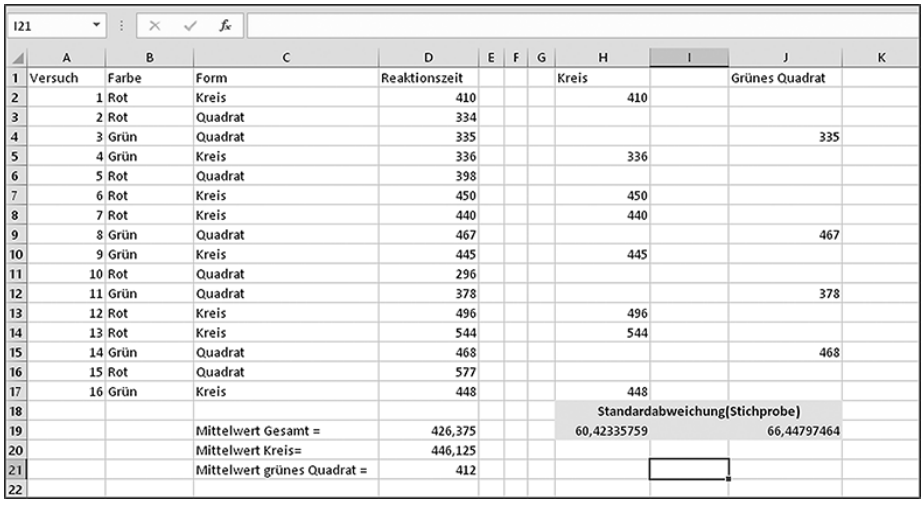
Hier eine Faustregel: Stellen Sie immer dann, wenn Sie einen Mittelwert berechnen, auch die Standardabweichung zur Verfügung. Verwenden Sie MITTELWERT und STABW.N im Tandem.

Wie bereits erwähnt, stellt Excel die beiden Funktionen MITTELWERTWENN und MITTELWERTWENNS zur Verfügung, bei denen die Berechnung des Mittelwertes von Bedingungen abhängig gemacht werden kann (siehe Kapitel 4). Daher wäre es hilfreich, wenn Excel auch die Funktionen STABWWENN und STABWWENNS zur Verfügung stellen würde, mit denen Sie, wenn Sie den Mittelwert in Abhängigkeit von Bedingungen berechnen, dies auch für die Standardabweichungen machen könnten.

Leider stellt Excel diese Funktionen jedoch nicht bereit. Daher stelle ich Ihnen nachfolgend Alternativlösungen vor, mit denen Sie die Standardabweichungen in Abhängigkeit von festgelegten Bedingungen berechnen können.

Die Alternativlösung filtert zuerst die Daten, die eine bestimmte Bedingung erfüllen, und berechnet dann die Standardabweichung für die gefilterten Daten. Abbildung 5.6 zeigt, was ich damit meine. Die Daten stammen aus dem fiktiven psychologischen Experiment, das ich auch in Kapitel 4 verwende.

In diesem Experiment sitzt eine Person vor einem Bildschirm, auf dem eine mit einer Farbe gefüllte Form angezeigt wird. Die Farbe ist entweder Rot oder Grün und die Form ist entweder ein Kreis oder ein Quadrat. Bei jedem Versuch wird eine zufällige Kombination aus Farbe und Form angezeigt; alle Kombinationen werden gleich oft verwendet. Im Fachjargon handelt es sich bei jeder angezeigten Form um einen Versuch. Daher zeigt das Tabellenblatt die Ergebnisse von 16 Versuchen an.



	A	B	C	D	E	F	G	H	I	J	K
1	Versuch	Farbe	Form	Reaktionszeit				Kreis		Grünes Quadrat	
2	1	Rot	Kreis	410				410			
3	2	Rot	Quadrat	334							
4	3	Grün	Quadrat	335						335	
5	4	Grün	Kreis	336				336			
6	5	Rot	Quadrat	398							
7	6	Rot	Kreis	450				450			
8	7	Rot	Kreis	440				440			
9	8	Grün	Quadrat	467						467	
10	9	Grün	Kreis	445				445			
11	10	Rot	Quadrat	296							
12	11	Grün	Quadrat	378						378	
13	12	Rot	Kreis	496				496			
14	13	Rot	Kreis	544				544			
15	14	Grün	Quadrat	468						468	
16	15	Rot	Quadrat	577							
17	16	Grün	Kreis	448				448			
18								Standardabweichung(Stichprobe)			
19			Mittelwert Gesamt =	426,375				60,42335759		66,44797464	
20			Mittelwert Kreis=	446,125							
21			Mittelwert grünes Quadrat =	412							
22											

Abbildung 5.6: Daten filtern, um die Standardabweichung in Abhängigkeit von Bedingungen zu berechnen

Die Person vor dem Computer drückt eine Taste, wenn sie eine Form sieht. Spalte D (mit Reaktionszeit beschriftet), gibt die Reaktionszeit der Versuchsperson in Millisekunden (Tausendstel einer Sekunde) an. Die Spalten B und C enthalten die Merkmale der in dem jeweiligen Versuch angezeigten Form. So können Sie beispielsweise aus Zeile 2 ablesen, dass im ersten Versuch ein roter Kreis angezeigt wurde und dass die Reaktionszeit 410 Millisekunden betrug.

Für alle Spalten habe ich den Text in der obersten Zeile verwendet, um einen Namen zu definieren, über den ich dann auf die Daten verweisen kann. (Wenn Sie nicht mehr genau wissen, wie das geht, werfen Sie einen Blick in Kapitel 2).

Zelle D20 zeigt den Mittelwert der Reaktionszeit an. Dieser wird ganz einfach mit dieser Formel berechnet:

```
=MITTELWERT(Reaktionszeit)
```

Zelle D20 enthält den Mittelwert der Versuche, bei denen ein Kreis angezeigt wurde. Dieser Wert wird mit dieser Formel berechnet, die die Funktion `MITTELLWERTWENN` verwendet:

```
=MITTELWERTWENN(Form;"Kreis";Reaktionszeit)
```

Zelle D21 wiederum zeigt den Mittelwert der Versuche, bei denen ein grünes Quadrat angezeigt wurde. Die Formel in dieser Zelle lautet:

```
=MITTELWERTWENNS(Reaktionszeit;Farbe;"Grün";Form;"
Quadrat")
```

In den Spalten H und K befinden sich die gefilterten Daten. Spalte H zeigt die Reaktionszeiten der Versuche an, bei denen ein Kreis angezeigt wurde. Zelle H19 zeigt die Standardabweichung (`STABW.S`) für diese Versuche an. Der Wert hier wird mit

```
=STABWWENN(Form;"Kreis";Reaktionszeit)
```

berechnet, wenn *diese Funktion* denn existieren würde.

Spalte K zeigt die Reaktionszeiten der Versuche an, bei denen ein grünes Quadrat angezeigt wurde. Zelle K19 zeigt die Standardabweichung (`STABW.S`) für diese Versuche an. Der Wert hier wird mit

```
=STABWWENNS(Reaktionszeit;Farbe;"Grün";Form;"Quadr
at")
```

berechnet, wenn *diese Funktion* denn existieren würde.

Wie habe ich die Daten gefiltert? Ich werde Ihnen dieses Geheimnis sofort verraten, nachdem ich kurz ein anderes Thema angeschnitten habe. Dieses lautet:

Ein wenig Logik

Damit Sie weitermachen können, müssen Sie zwei der logischen Funktionen von Excel kennen: WENN und UND. Sie verwenden diese Formeln, indem Sie auf FORMELN | LOGISCH klicken und dann im Menü der Schaltfläche LOGISCH auswählen.

Die Funktion WENN besitzt drei Argumente:

- eine logische Bedingung, die geprüft werden soll
- die Aktion, die unternommen werden soll, wenn die Bedingung erfüllt ist, das heißt, wenn der Wert des logischen Ausdrucks WAHR ist
- ein optionales Argument, das die Aktion für den Fall festlegt, dass die Bedingung nicht erfüllt ist, das heißt, wenn der Wert des logischen Ausdrucks FALSCH ist

Abbildung 5.7 zeigt das Dialogfeld FUNKTIONSARGUMENTE für die Funktion WENN.

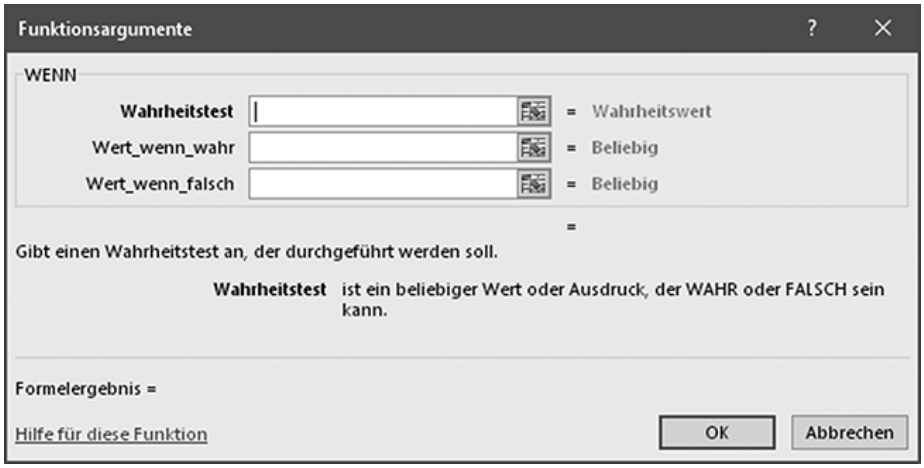


Abbildung 5.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion WENN

Die Funktion UND kann bis zu 255 Argumente entgegennehmen. UND prüft, ob alle Argumente die jeweils angegebene Bedingung erfüllen, ob also jede Bedingung WAHR ergibt. Wenn dies für alle

Bedingungen zutrifft, gibt UND den Wert WAHR zurück und anderenfalls FALSCH.

Abbildung 5.8 zeigt das Dialogfeld FUNKTIONSARGUMENTE für die Funktion UND.

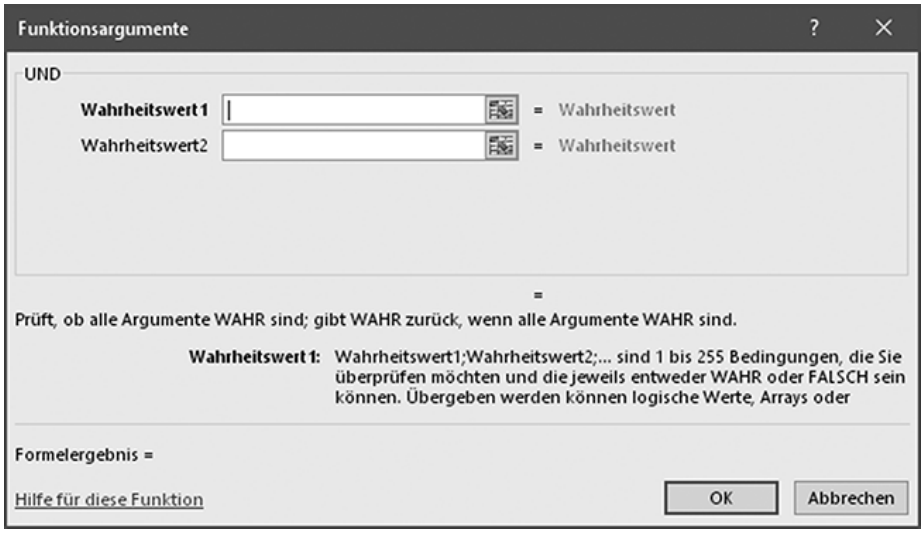


Abbildung 5.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion UND

Zurück zur eigentlichen Aufgabe

In diesem Beispiel verwende ich die Funktion WENN, um eine Zelle in Spalte H auf den betreffenden Wert der gleichen Zeile in Spalte D zu setzen, wenn der Wert in Spalte C der gleichen Zelle "Kreis" lautet. Die Formel in Zelle H2 lautet:

```
=WENN(C2="Kreis";D2;" ")
```

Würde man diese Formel in einem Satz ausdrücken, dann wäre es dieser: »Wenn der Wert in C2 'Kreis' lautet, dann füge in die aktuelle Zelle den Wert aus Zelle D2 ein. Falls nicht, lasse die aktuelle Zelle leer.« Wenn Sie die 15 darunter stehenden Zellen automatisch ausfüllen lassen, erhalten Sie die in Spalte H von Abbildung 5.6 gezeigten Daten. Die Standardabweichung in Zelle H19 ist der Wert, den STABWWENN liefern würde, würde diese Funktion denn existieren.

Ich könnte das dritte Argument (die beiden doppelten Anführungszeichen) auch einfach weglassen, ohne dass dies den Wert der Standardabweichung beeinflusst. Ohne das dritte Argument gibt Excel in die Zellen, die die Bedingungen nicht erfüllen, `FALSCH` ein, statt sie leer zu lassen.

Für die Zellen in Spalte K habe ich sowohl `UND` als auch `WENN` verwendet. In die Zelle dieser Spalte wird der betreffende Wert aus Spalte D eingefügt, wenn zwei Bedingungen wahr sind:

- Wenn der Wert in der betreffenden Zelle in Spalte B `"Grün"` ist.
- Wenn der Wert in der betreffenden Zelle in Spalte C `"Quadrat"` ist

Dies ist die Formel in Zelle K2:

```
=WENN(UND(B2="Grün";C2="Quadrat");D2;"")
```

Wenn diese Formel ein Satz wäre, würde er so lauten: »Wenn der Wert in B2 ›Grün‹ und der Wert in C2 ›Quadrat‹ lautet, füge in die aktuelle Zelle den Wert aus Zelle D2 ein. Falls nicht, lasse die aktuelle Zelle leer.« Wenn Sie die 15 darunter stehenden Zellen automatisch ausfüllen lassen, erhalten Sie die in Spalte K von Abbildung 5.6 gezeigten Daten. Die Standardabweichung in Zelle K19 ist der Wert, den `STABWWENNS` liefern würde, würde diese Funktion denn existieren.

Verwandte Funktionen

Bevor wir fortfahren, wollen wir rasch einige weitere Tabellenblattfunktionen betrachten, die ebenfalls mit Streuung zu tun haben.

SUMQUADABW

Mit SUMQUADABW können Sie die Summe der quadratischen Abweichungen vom Mittelwert (ohne Division durch *N* oder *N* – *1*) berechnen. Für die Zahlen

127, 120, 132, 117, 114

beträgt dieser Wert 218 (siehe Abbildung 5.9).

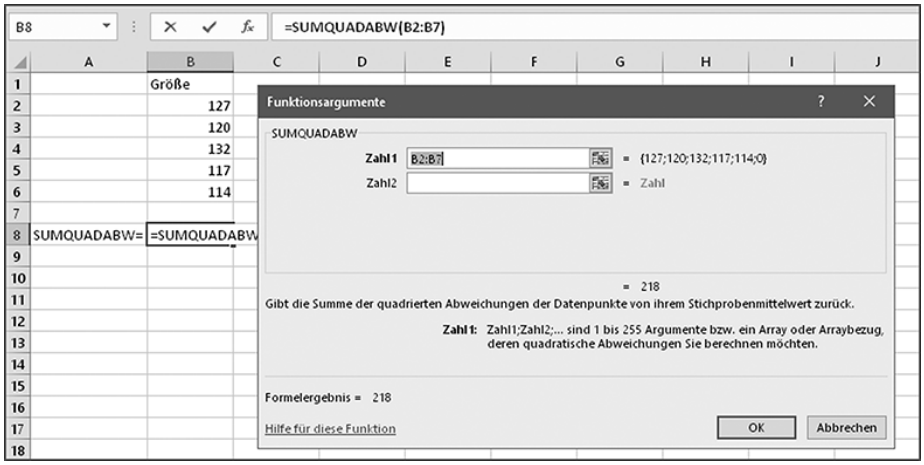


Abbildung 5.9: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion SUMQUADABW

Mittlere Abweichung

Eine weitere Excel-Funktion bezieht sich auf Abweichungen, ohne diese zu quadrieren.

Bei der Varianz und der Standardabweichung werden negative Abweichungswerte zunächst quadriert und anschließend wird der Mittelwert gebildet. Was geschieht aber, wenn wir das Minuszeichen einfach nicht beachten? Dann spricht man vom jeweiligen *Absolutwert* der Abweichung. (So drücken Mathematiker die Frage »Was geschieht aber, wenn wir das Minuszeichen einfach nicht beachten?« aus.)

Wenn wir also für die Werte

127, 120, 132, 117, 114

das Minuszeichen unberücksichtigt lassen, können wir die jeweiligen Absolutwerte für die Abweichung in eine Tabelle wie in Tabelle 5.4 eintragen.

Größe	Größe – Mittelwert	|Abweichung|
127	127 – 122	5
120	120 – 122	2
132	132 – 122	10
117	117 – 122	5
114	114 – 122	8

Tabelle 5.4: Eine Zahlengruppe und die entsprechenden absoluten Abweichungen

Beachten Sie in Tabelle 5.4 die vertikalen Linien vor und nach dem Wort *Abweichung* in der Überschrift der dritten Spalte. Vertikale Linien vor und nach einer Zahl symbolisieren den absoluten Wert einer Zahl. Das heißt, die vertikalen Linien sind das mathematische Symbol für »Was geschieht aber, wenn wir das Minuszeichen einfach nicht beachten?«.

Der Mittelwert der Zahlen in der dritten Spalte beträgt 6,0. Dieser Mittelwert wird als *mittlere absolute Abweichung* bezeichnet. Damit lässt sich die Streuung von Werten um deren Mittelwert rasch und einfach ausdrücken. Die Werte haben dieselbe Einheit wie die ursprünglichen Werte. Wenn die Werte also in *cm* angegeben sind, werden auch die mittleren absoluten Abweichungen in *cm* angegeben.

Wie bei der Varianz und der Standardabweichung bedeutet eine große mittlere absolute Abweichung eine große Streuung. Eine kleine mittlere absolute Abweichung bedeutet eine geringe Streuung.

Diese Berechnung ist im Vergleich zur Berechnung der Varianz oder der Standardabweichung einfacher, wird jedoch nur selten verwendet. Warum? Aus Gründen, die (wieder einmal) den Rahmen dieses Buches sprengen würden, können Statistiker das Ergebnis dieser Berechnung nicht als Grundlage für weitere statistische Berechnungen verwenden, die Varianz und Standardabweichung dagegen schon.

MITTELABW

Mit der Excel-Tabellenblattfunktion `MITTELABW` können Sie die absolute Abweichung vom Mittelwert einer Zahlengruppe berechnen. In Abbildung 5.10 ist das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `MITTELABW` dargestellt. Hier wird die absolute Abweichung vom Mittelwert der Zellen im angegebenen Bereich angezeigt.

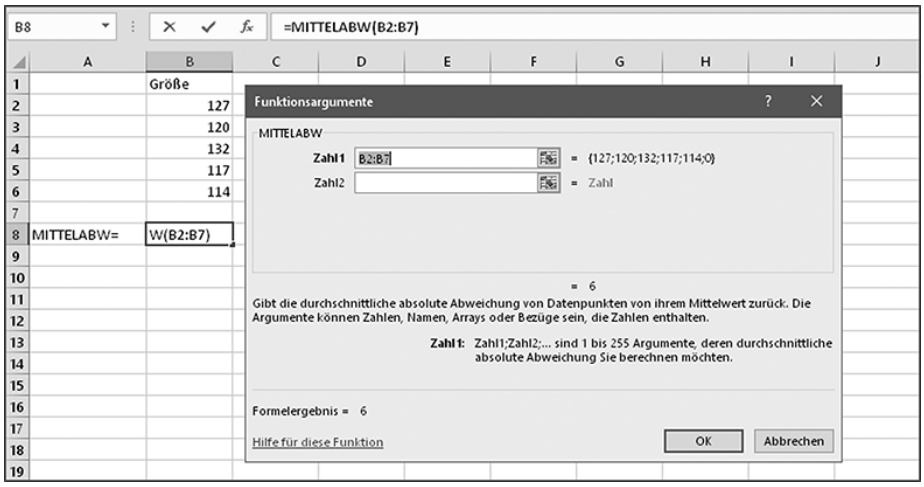


Abbildung 5.10: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `MITTELABW`

Kapitel 6

Standards und Wertungen kennenlernen

In diesem Kapitel

- So standardisieren Sie Werte
- Vergleiche anstellen
- Ränge in Dateien
- Quantile kommen ins Spiel

In meiner linken Hand halte ich 30 argentinische Pesos und in meiner rechten 100 chilenische Pesos. Was ist mehr wert? Beide Währungen heißen »Peso«, nicht wahr? Müsste also nicht 100 mehr wert sein als 30? Nicht unbedingt. »Peso« ist nur ein Wort – eine zufällige Übereinstimmung von Namen. Jede Währung kommt aus einem anderen Land und jedes Land hat seine eigene Wirtschaft.

Um die beiden Geldwerte miteinander vergleichen zu können, müssen Sie die Währungen in eine Standardeinheit umrechnen. Die anschaulichste Währung ist für uns immer die eigene Währung. Welchen Wert haben die einzelnen Beträge in Euro und/oder Cent? 15 argentinische Pesos haben einen Wert von circa 0,90 €. 100 chilenische Pesos sind etwa 14 Cent wert (Stand 14. November 2016).

In diesem Kapitel erfahren Sie, wie Sie mit statistischen Berechnungen Standardeinheiten ermitteln können. Anhand von Standardeinheiten können Sie feststellen, an welcher Position ein Wert in Bezug auf andere Werte in einer Gruppe steht. Außerdem

lernen Sie hier weitere Möglichkeiten zum Ermitteln der Position eines Werts innerhalb einer Gruppe kennen.

Z-Werte einfangen

Wie Sie im vorherigen Abschnitt gesehen haben, sagt eine Zahl allein nichts aus. Um die Bedeutung einer Zahl wirklich verstehen zu können, müssen Sie wissen, wie diese Zahl zustande gekommen ist. Um zwei Zahlen miteinander vergleichen zu können, müssen beide derselben Werteskala entsprechen.

In manchen Fällen, wie beispielsweise bei der Umrechnung von Währungen, ist es einfach, einen Standardwert zu ermitteln. In anderen Fälle, wie bei der Umrechnung von Temperaturwerten oder von englischen Maßen in Werte des metrischen Systems, ist eine Formel erforderlich.

Wenn Sie die Formel nicht kennen, können Sie mit dem Mittelwert und der Standardabweichung Werte aus unterschiedlichen Systemen standardisieren. Hierzu verwenden Sie den Mittelwert einer Reihe von Zahlen als Nullpunkt und die Standardabweichung als Einheitsmaß. Dann vergleichen Sie die Abweichung der einzelnen Zahlen vom Mittelwert mit der Standardabweichung. Sie fragen sich vielleicht: Wie groß ist eine bestimmte Abweichung relativ zu (etwas wie) einem Durchschnitt aller Abweichungen?

Um das herauszufinden, dividieren Sie die Abweichung des Werts durch die Standardabweichung. Damit transformieren Sie den Wert in eine andere Art Wert. Der transformierte Wert wird als *Standardwert* oder *z-Wert* bezeichnet.

Die Formel hierfür lautet wie folgt:

$$z = \frac{x - \bar{x}}{s}$$

wenn es sich um eine Stichprobe handelt, und

$$z = \frac{x - \mu}{\sigma}$$

wenn es sich um eine Grundgesamtheit handelt. In beiden Fällen steht *x* für den Wert, der in einen z-Wert transformiert wird.

Eigenschaften von z-Werten

Ein z-Wert kann positiv, negativ oder null sein. Ein negativer z-Wert ist ein Wert kleiner als der Mittelwert und ein positiver z-Wert ist ein Wert größer als der Mittelwert. Wenn ein Wert genauso groß ist wie der Mittelwert, beträgt der z-Wert null.

Wenn Sie den z-Wert für jeden Wert in einer Gruppe berechnen, ist der Mittelwert der z-Werte 0 und die Standardabweichung der z-Werte 1.

Wenn Sie diese Werte für eine Reihe von Werten berechnet haben, ist es legitim, einen Wert aus der einer Zahlengruppe mit einem Wert aus der anderen Zahlengruppe zu vergleichen. Haben zwei Zahlengruppen unterschiedliche Mittelwerte und Standardabweichungen, ist der Vergleich ohne Standardisierung wie der Vergleich von Äpfeln mit Birnen.

Anhand der folgenden Beispiele werden Sie sehen, wie mit z-Werten Vergleiche angestellt werden.

Bonds und Ruth

Eine wichtige Frage, die in ernsten metaphysischen Diskussionen immer wieder gestellt wird: Wer ist der größte Homerun-Schlagmann aller Zeiten? Barry Bonds oder Babe Ruth? Das ist keine leichte Frage. Aber wir können der Antwort näher kommen, wenn wir jeweils die beste Saison der beiden Spieler miteinander vergleichen. Bonds erzielte 2001 73 Homeruns und Ruth 1927 60. Auf den ersten Blick scheint Bonds erfolgreicher.

Das Jahr 1927 war jedoch ganz anders als das Jahr 2001. Im Baseball (und überall sonst auch) hat sich seit 1927 einiges verändert, was sich natürlich auf die Spielerstatistiken auswirkt. In den 20er-Jahren des letzten Jahrhunderts war es im Vergleich zum Jahr 2001 schwieriger, einen Homerun zu erzielen. Aber trotzdem, 73 zu 60? Hmmm …

Mit Standardwerten können wir ermitteln, wessen beste Saison besser war. Zum Standardisieren habe ich die 50 besten Homerun-Schlagmänner von 1927 und die besten 50 von 2001 herangezogen. Ich habe den Mittelwert und die Standardabweichung jeder Gruppe berechnet und anschließend die 60 von Ruth und die 73 von Bonds in z-Werte umgerechnet.

1927 haben die Schlagmänner im Durchschnitt 12,68 Homeruns mit einer Standardabweichung von 10,49 erzielt. 2001 haben die Schlagmänner im Durchschnitt 37,02 Homeruns mit einer Standardabweichung von 9,64 erzielt. Trotz der großen Differenz bei den Mittelwerten unterscheiden sich die Standardabweichungen nur wenig.

Und die z-Werte? Der z-Wert von Ruth beträgt

$$z = \frac{60 - 12{,}68}{10{,}49} = 4{,}51$$

Der z-Wert von Bonds beträgt

$$z = \frac{73 - 37{,}02}{9{,}64} = 3{,}73$$

Der eindeutige Gewinner im Wettbewerb um den besten z-Wert für Homeruns ist Babe Ruth. Punkt.

Nur damit Sie sehen, wie sich die Zeiten geändert haben: Lou Gehring schlug 1927 47 Homeruns (und war damit Zweiter nach Ruth) mit einem z-Wert von 3,27. 2001 ergeben 47 Homeruns einen z-Wert von 1,04.

Prüfungsergebnisse

Aber lassen wir den Sport und wenden wir uns einer praktischen Anwendung der z-Werte, nämlich der Zuweisung von Noten zu Prüfungsergebnissen zu. In den USA bewerten Lehrer die Leistung von Schülern auf der Grundlage eines prozentualen Bewertungssystems: 90 Punkte und mehr (von 100) entsprechen der Note A, 80 bis 89 Punkte entsprechen der Note B, 70 bis 79 Punkte entsprechen der Note C, 60 bis 69 Punkte entsprechen der Note D und weniger als 60 Punkte entsprechen der Note F. Sie bilden aus mehreren Prüfungsergebnissen den Mittelwert, um so zu einer Endnote zu gelangen.

Ist das fair? Ein Peso aus Argentinien ist mehr wert als ein Peso aus Chile und 1927 war es schwieriger, einen Homerun zu schlagen, als 2001. So stellt sich die Frage, ob ein Punkt in einer Prüfung tatsächlich genauso viel wert ist wie ein Punkt in einer anderen Prüfung. Ist das nicht wie beim Peso nur ein Wort?

Ja, das ist es. Ein Punkt für eine schwierige Prüfung ist per definitionem schwieriger zu erzielen als ein Punkt für eine leichte Prüfung. Da Punkte je nach Prüfung eine andere Bedeutung haben können, wäre es am fairsten, die Noten der einzelnen Prüfungen vor dem Berechnen des Mittelwerts in z-Werte umzurechnen. So werden Zahlen auf gleicher Ebene gemittelt.

In meinem Kurs gehe ich so vor. Dabei stelle ich häufig fest, dass eine schlechtere Note bei einer Prüfung oft einem besseren z-Wert entspricht als eine bessere Note bei einer anderen Prüfung. Beispiel: Bei einer Prüfung mit einem Mittelwert von 65 und der Standardabweichung 12 ergibt der Wert 71 einen z-Wert von 0,5. Bei einer anderen Prüfung mit einem Mittelwert von 69 und einer Standardabweichung von 14 entspricht der Wert 75 dem z-Wert 0,429. (Ja, das ist wie mit den 60 Homeruns von Ruth und den 73 Homeruns von Bonds.) Fazit: Zahlen allein haben keine Bedeutung. Sie müssen wissen, wie die Zahlen zustande gekommen sind.

STANDARDISIERUNG

Die Excel-Tabellenblattfunktion STANDARDISIERUNG berechnet z-Werte. In Abbildung 6.1 sind einige Prüfungsergebnisse, deren Mittelwert und die Standardabweichung dargestellt. Diese statistischen Werte habe ich mit den Funktionen MITTELWERT und STABW.N berechnet. Außerdem ist in der Abbildung das Dialogfeld FUNKTIONSARGUMENTE für die Funktion STANDARDISIERUNG zu sehen.

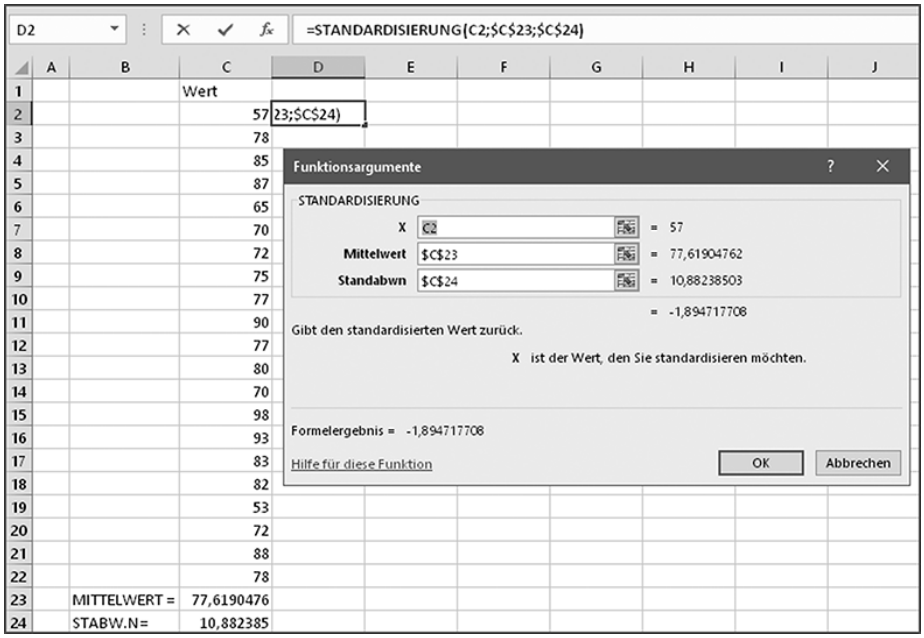


Abbildung 6.1: Prüfungsergebnisse und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion STANDARDISIERUNG

Und so müssen Sie vorgehen:

1. **Geben Sie die gewünschten Werte in einen Zellbereich ein und markieren Sie eine Zelle.**

   Die Daten befinden sich in C2:C22. Der z-Wert für C2 soll in Zelle D2 berechnet werden. Anschließend werde ich Spalte D automatisch ausfüllen, damit alle z-Werte neben den jeweiligen Prüfungsergebnissen stehen.

2. **Wählen Sie im Menü mit den statistischen Funktionen STANDARDISIERUNG, damit das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion geöffnet wird.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Zuerst gebe ich in das Feld X die Zelle ein, in der sich das erste Prüfungsergebnis befindet. In unserem Beispiel ist dies Zelle C2.

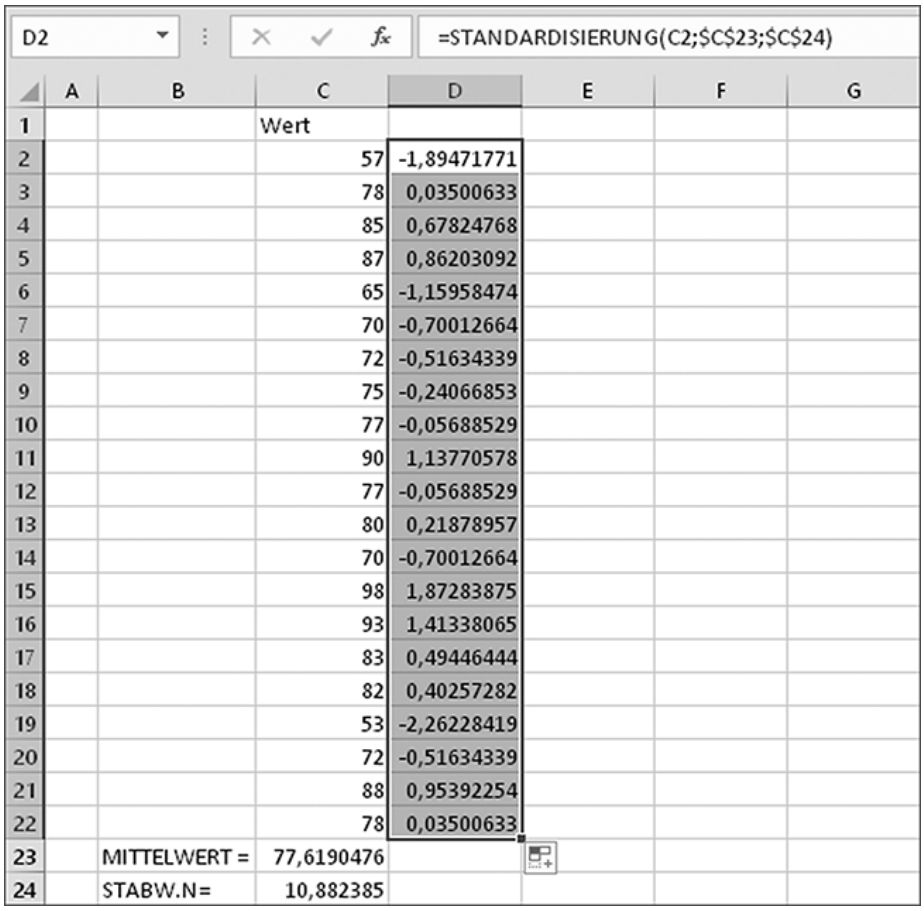
   In das Feld MITTELWERT gebe ich die Zelle ein, in der sich der Mittelwert befindet, in unserem Beispiel C23. Achten Sie darauf, dass Sie die Zelladresse als absoluten Zellbezug eingeben, Ihre Eingabe muss also **$C$23** sein. Sie können den Zellbezug entweder direkt eintippen oder die Zelle C23 auswählen, dann das Feld MITTELWERT markieren und die Taste F4 drücken.

   In das Feld STANDABWN gebe ich die Zelle ein, in der sich die Standardabweichung befindet. In diesem Beispiel ist dies Zelle C24. Auch hier müssen Sie einen absoluten Bezug eingeben. Dies sieht dann folgendermaßen aus: **$C$24**.

4. **Klicken Sie auf OK, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen und den z-Wert für das erste Prüfungsergebnis in die markierte Zelle einzufügen.**

Abschließend zeige ich mit dem Cursor auf das Ausfüllkästchen der markierten Zelle, halte die linke Maustaste gedrückt und ziehe die Maus, um die restlichen z-Werte automatisch auszufüllen.

In Abbildung 6.2 sind die automatisch ausgefüllten Zellen mit den z-Werten dargestellt.



D2 | =STANDARDISIERUNG(C2;$C$23;$C$24)

	A	B	C	D	E	F	G
1			Wert				
2			57	-1,89471771			
3			78	0,03500633			
4			85	0,67824768			
5			87	0,86203092			
6			65	-1,15958474			
7			70	-0,70012664			
8			72	-0,51634339			
9			75	-0,24066853			
10			77	-0,05688529			
11			90	1,13770578			
12			77	-0,05688529			
13			80	0,21878957			
14			70	-0,70012664			
15			98	1,87283875			
16			93	1,41338065			
17			83	0,49446444			
18			82	0,40257282			
19			53	-2,26228419			
20			72	-0,51634339			
21			88	0,95392254			
22			78	0,03500633			
23		MITTELWERT =	77,6190476				
24		STABW.N=	10,882385				

Abbildung 6.2: Das automatisch ausgefüllte Feld mit z-Werten

Z-Werte verstecken

Da negative z-Werte eine, nun, negative Konnotation haben können, gibt es Lehrer, die den z-Wert für die Bewertung der Schüler ändern. Genau genommen verstecken sie den z-Wert, aber das Konzept bleibt dasselbe: Standardisierung mit der Standardabweichung als Maßeinheit.

Eine bekannte Transformation wird als T-Wert bezeichnet. Beim T-Wert gibt es keine negativen Werte, da eine Gruppe T-Werte einen Mittelwert von 50 und eine Standardabweichung von 10 aufweist. Dabei geht es darum, eine Prüfung schreiben zu lassen, alle Tests zu benoten und den Mittelwert und die Standardabweichung zu berechnen. Anschließend wird jede Note in einen z-Wert umgerechnet. Dann wird die folgende Formel angewendet:

$$T = (z)\,(10) + 50$$

Meist wird der T-Wert auf die nächste ganze Zahl gerundet.

SAT-Wert ist eine weitere Transformation des z-Werts. (Manche nennen den SAT-Wert C-Wert.) Der SAT-Wert weist einen Mittelwert von 500 und eine Standardabweichung von 100 auf. Nach der Benotung der Prüfungen und der Berechnung von Mittelwert und Standardabweichung werden die einzelnen Prüfungsergebnisse ganz normal in z-Werte umgerechnet. Mit dieser Formel wird dann der z-Wert in einen SAT-Wert umgerechnet:

$$SAT = (z)\,(100) + 500$$

Auch hier wird der Wert auf die nächste ganze Zahl gerundet.

Der IQ-Wert ist noch eine Transformation des z-Werts. Der Mittelwert beträgt 100 und die Standardabweichung (bei der Stanford-Binet-Version) beträgt 16. Und wie wird der IQ berechnet? Sie können es sich bestimmt schon denken. Für eine Gruppe IQ-Werte werden der Mittelwert und die Standardabweichung und anschließend der z-Wert berechnet. Die Formel lautet dann wie folgt:

$$IQ = (z)\,(16) + 100$$

Wie die beiden anderen Werte wird auch beim IQ-Wert auf die nächste ganze Zahl gerundet.

Wo stehen Sie?

Anhand von Standardwerten wird erkennbar, an welcher Position sich ein Wert im Vergleich zu anderen Werten in derselben Gruppe befindet. Hierbei wird die Standardabweichung als Maßeinheit verwendet.

Wenn Sie die Standardabweichung nicht verwenden möchten, können Sie die relative Position eines Werts einfach ermitteln. Sie können den Rang des Werts innerhalb der Gruppe ermitteln. Der höchste Wert hat den Rang 1, der zweithöchste den Rang 2 usw.

RANG.GLEICH und RANG.MITTELW

Excel 2016 stellt zwei Funktionen zur Ermittlung des Rangs zur Verfügung. Sie unterscheiden sich darin, wie sie mit gleichen Zahlen umgehen. Wenn `RANG.GLEICH` auf gleiche Zahlen trifft, wird diesen Zahlen die gleiche, und zwar die größte Rangzahl zugewiesen.

Wenn also drei gleiche Werte auf dem zweiten Rang stehen, erhalten alle drei Werte die Rangzahl 2.

Wenn `RANG.MITTELW` auf gleiche Zahlen trifft, weist die Funktion all diesen Zahlen den Mittelwert der Ränge ein, die sie einnehmen. Diese Funktion weist daher drei gleichen Werten, die auf dem zweiten Rang stehen, die Rangzahl 3 zu (also den Mittelwert aus 2, 3 und 4).

Um zu verdeutlichen, wie diese beiden Funktionen arbeiten, zeigt Abbildung 6.3 die Ergebnisse, nachdem die Funktionen auf die Werte in Spalte B angewendet wurden.

	A	B	C	D	E
1		Wert	RANG.GLEICH=	RANG.MITTELW=	
2		45	10	10	
3		44	11	11	
4		34	12	12	
5		23	13	13,5	
6		22	15	15	
7		48	8	8,5	
8		48	8	8,5	
9		67	5	5	
10		65	6	6	
11		78	2	3	
12		78	2	3	
13		80	1	1	
14		78	2	3	
15		23	13	13,5	
16		54	7	7	
17					

Abbildung 6.3: Verwendung von `RANG.GLEICH` und `RANG.MITTELW`

So verwenden Sie die Funktion `RANG.GLEICH`:

1. **Geben Sie die Daten in einen Zellbereich ein und markieren Sie eine Zelle.**

   In diesem Beispiel habe ich die Werte in die Zellen B2 bis B16 eingegeben und die Zelle C2 markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen RANG.GLEICH, damit das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion geöffnet wird.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Feld ZAHL gebe ich die Zelle ein, in der sich der Wert befindet, dessen Rang in die markierte Zelle eingefügt werden soll. In diesem Beispiel ist das die Zelle B2.

   In das Feld BEZUG gebe ich den Zellbereich mit den Werten ein, also den Zellbereich B2:B16.

   Dieser Teil ist wichtig. Nachdem ich `RANG.GLEICH` in Zelle C2 eingegeben habe, will ich die Maus über Spalte C ziehen und so die Ränge der restlichen Werte automatisch ausfüllen lassen. Dazu muss ich Excel wissen lassen, dass die Zellen B2 bis B16 das Datenfeld für alle Werte darstellen, und nicht nur für einen einzigen.

   Das bedeutet, dass das Datenfeld im Feld BEZUG wie folgt angegeben werden muss: **$C$2:$C$16**. Sie können die Dollarzeichen ($) entweder manuell eingeben oder das Feld BEZUG markieren und die Taste F4 drücken.

   Im Feld REIHENFOLGE gebe ich die Reihenfolge für die Sortierung der Werte an. Um die Werte in absteigender Folge zu sortieren, kann ich das Feld REIHENFOLGE entweder leer lassen oder dort eine 0 (Null) eingeben. Um die Werte in aufsteigender Folge zu sortieren, muss ich einen anderen Wert als 0 eingeben. In diesem Beispiel habe ich das Feld REIHENFOLGE nicht ausgefüllt (siehe Abbildung 6.4).

4. **Klicken Sie auf OK, damit der Rang in der markierten Zelle angezeigt wird.**

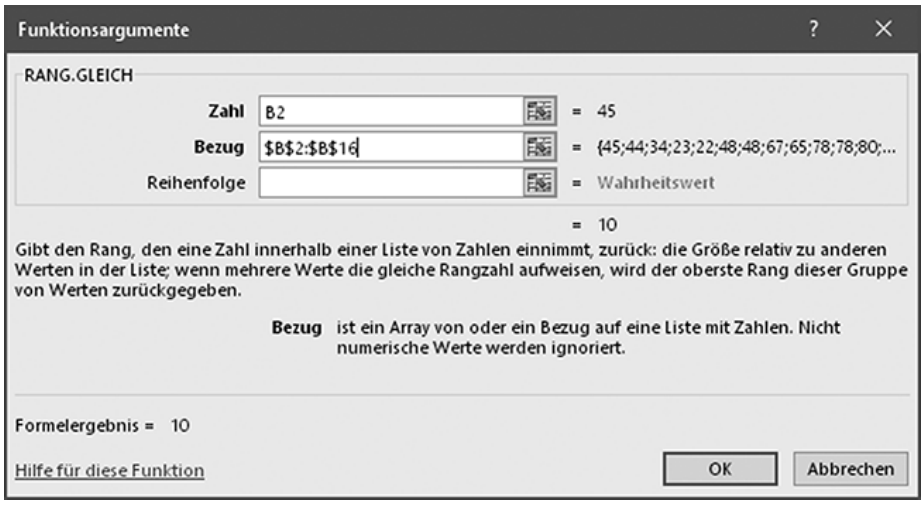


Abbildung 6.4: Das Dialogfeld FUNKTIONSARGUMENTE für `RANG.GLEICH`

Abschließend habe ich mit der Maus auf das Ausfüllkästchen der markierten Zelle gezeigt, die linke Maustaste gedrückt und die Maus gezogen, damit die Ränge der weiteren Werte automatisch ausgefüllt werden. So habe ich Spalte C in Abbildung 6.3 fertig gestellt).

KGRÖSSTE und KKLEINSTE

Sie können den Rangermittlungsprozess umkehren, indem Sie einen Rang angeben und fragen, welcher Wert diesem Rang zugeordnet ist. Mit den Tabellenblattfunktionen `KGRÖSSTE` und `KKLEINSTE` lässt sich diese Frage beantworten. Mit diesen Funktionen können Sie den fünftgrößten Wert, den drittkleinsten oder jeden beliebigen anderen Wert ermitteln.

Abbildung 6.5 zeigt das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `KGRÖSSTE` dargestellt. Im Feld MATRIX geben Sie die Zellen ein, in denen sich die Werte befinden. Im Feld K geben Sie den Rang ein, dessen Wert ermittelt werden soll. Um beispielsweise den siebtgrößten Wert zu ermitteln, geben Sie im Feld K **7** ein.

Abbildung 6.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion KGRÖSSTE

KKLEINSTE funktioniert genauso, nur dass damit der Werterang vom unteren Ende der Zahlengruppe ermittelt wird. Im Dialogfeld FUNKTIONSARGUMENTE für die Funktion KKLEINSTE befinden sich ebenfalls ein Feld mit der Bezeichnung MATRIX und ein Feld mit der Bezeichnung K. Wenn Sie im Feld K **7** eingeben, wird der siebtkleinste Wert der Zahlengruppe angezeigt.

QUANTIL.INKL und QUANTIL.EXKL

Eng verwandt mit dem Rang ist das *Quantil*, das die Position eines Werts in einer Gruppe als Prozentwert der darunterliegenden Werte angibt. Wenn Sie schon einmal standardisierte Tests wie den SAT durchgeführt haben, kennen Sie Quantile. Ein SAT-Wert im 80. Quantil ist größer als 80 % der anderen SAT-Werte.

Klingt einfach, oder? Jedoch ist die Definition von Quantil heutzutage ein wenig verschwommen. Einige definieren Quantil als »größer als« (wie im vorigen Absatz), andere definieren es als »größer oder gleich«. *Größer als* ist gleichbedeutend mit *exklusiv*. *Größer oder gleich* ist gleichbedeutend mit *inklusiv*.

Aus diesem Grund stellt Excel für den Umgang mit Quantilen zwei Tabellenblattfunktionen bereit. QUANTIL.INKL verwendet »größer oder gleich« und QUANTIL.EXKL verwendet »größer als«.

Um diese Funktionen zu verwenden, geben Sie einen Zellbereich mit Werten und ein Quantil ein. Bei diesen Formeln entspricht der Parameter K dem Quantil. QUANTIL.INKL findet den Wert, der

größer oder gleich dem angegebenen Quantil ist. QUANTIL.EXKL finden den kleinsten Wert, der größer ist als das Quantil.

Um zu zeigen, wie sich die Ergebnisse der beiden Funktionen unterscheiden, habe ich das Tabellenblatt erstellt, das Sie in Abbildung 6.6 sehen.

I7

	A	B	C	D	E	F	G
1			Wert		K	QUANTIL.INKL=	QUANTIL.EXKL=
2			80		0,95	78,6	#ZAHL!
3			78		0,90	78,0	78,8
4			78		0,85	78,0	78,0
5			78		0,80	78,0	78,0
6			67		**0,75**	**72,5**	**78,0**
7			65		0,70	66,6	69,2
8			54		0,65	65,2	65,8
9			48		0,60	58,4	60,6
10			48		0,55	52,2	52,8
11			45		**0,50**	**48,0**	**48,0**
12			44		0,45	48,0	48,0
13			34		0,40	46,8	46,2
14			23		0,35	44,9	44,6
15			23		0,30	44,2	42,0
16			22		**0,25**	**39,0**	**34,0**
17					0,20	31,8	25,2
18					0,15	24,1	23,0
19					0,10	23,0	22,6
20					0,05	22,7	#ZAHL!

Abbildung 6.6: QUANTIL.INKL **versus** QUANTIL.EXKL

Die Werte in Spalte C sind mit denen aus den vorherigen Beispielen dieses Kapitels identisch, jedoch sind sie in absteigender Reihenfolge sortiert. K gibt die Quantile an. Beachten Sie, dass sie als Dezimalzahlen angegeben sind; daher entspricht 0,95 dem 95 %-Quantil. Die Werte in den Spalten F und G wurden mit den in der Spaltenüberschrift angegebenen Funktionen berechnet. Der Wert in diesen Spalten bezieht sich auf das Quantil in der gleichen Zeile von Spalte C. Spalte F zeigt also die einschließlichen und

Spalte G die ausschließlichen Quantile an. Ich habe die 75 %-, 50 %- und 25 %-Quantile hervorgehoben; dies dient der Vorbereitung der Ausführungen im Kasten »Gezogen und quartiliert«, den Sie etwas weiter unten finden.

Einige Unterschiede fallen sofort ins Auge, insbesondere die Fehlermeldungen in Spalte G. `QUANTIL.EXKL` gibt eine Fehlermeldung aus, wenn das angeforderte Quantil größer ist als `n/(n+1)`, wobei `n` für die Anzahl der Werte steht. Dies ist der maximale Wert von K, mit dem die Formel arbeiten kann. In unserem Beispiel ist der Maximalwert 0,94.

`QUANTIL.EXKL` gibt ebenfalls eine Fehlermeldung aus, wenn das angeforderte Quantil kleiner ist als `n/(n+1)`. Dies ist der kleinste Wert, mit dem die Formel arbeiten kann. In unserem Beispiel ist dies 0,06.

Wie Sie in der Abbildung ebenfalls sehen können, gehen die beiden Funktionen beim Ermitteln der Quantile unterschiedlich vor. So liegt das inklusive 70 %-Quantil irgendwo zwischen 65 und 67. `QUANTIL.INKL` interpoliert den Wert zu 66,6. (Da der Wert 78 dreimal vorkommt, führt die Interpolation dazu, dass der Wert größer wird als 66). `QUANTIL.EXKL` extrapoliert das exklusive 70 %-Quantil zu 69,2; dies ist der kleinste Wert, der größer ist als die Nachbarschaft von 65-67.

Dies bedeutet jedoch nicht, dass das exklusive Quantil immer größer ist als das inklusive. Ein paar der Einträge in den Spalten F und G zeigen, dass manchmal genau das Umgekehrte passiert. Gleiche Werte und die Position, an denen sie sich befinden, spielen dabei eine entscheidende Rolle.

Ich glaube nicht, dass die Unterschiede in der Praxis viel bedeuten. Dies ist aber nur meine persönliche Meinung. Die Moral: Wenn Sie für Ihr Quantil »größer oder gleich« benötigen, verwenden Sie die INKL-Version. Falls Sie lediglich nach »größer als« suchen, dann verwenden Sie die EXKL-Variante der Funktion.

So setzen Sie die Funktion `QUANTIL.INKL` ein:

1. **Geben Sie die Daten in einen Zellbereich ein, und markieren Sie eine Zelle.**

2. **Wählen Sie im Menü mit den statistischen Funktionen QUANTIL.INKL, damit das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion geöffnet wird.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Parameter ein.**

   Abbildung 6.7 zeigt die Daten, die ausgewählte Zelle und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion QUANTIL.INKL. In das Feld ARRAY habe ich im Hinblick auf AutoAusfüllen als absoluten Bezug $C$2:$C$16 eingegeben; im Feld K steht der Bezug zur Zelle E2: Dort steht der Wert 0,95.

4. **Klicken Sie auf OK, um das Quantil in die ausgewählte Zelle einzutragen.**

Für die Verwendung der Formel QUANTIL.EXKL gehen Sie entsprechend vor.

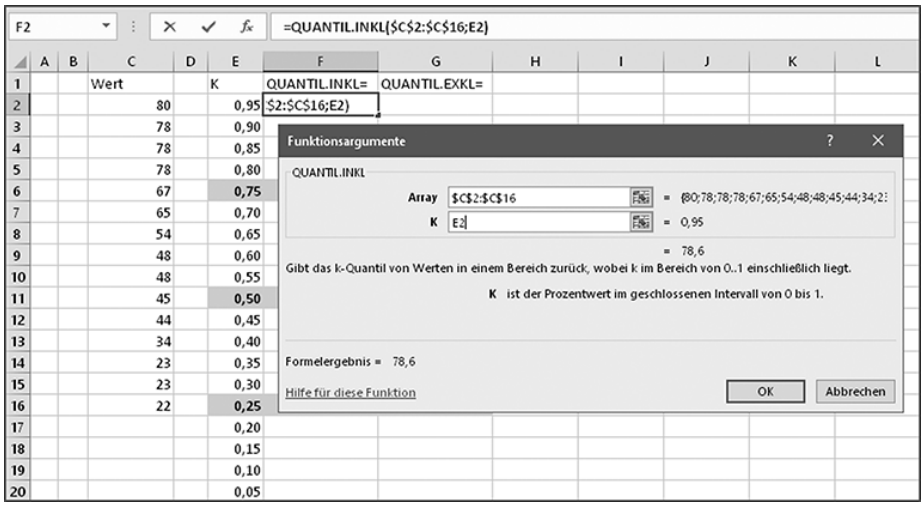


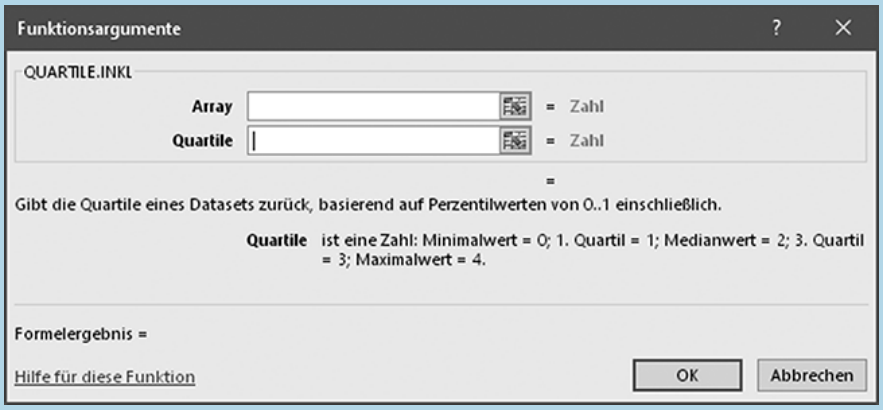
Abbildung 6.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion QUANTIL.INKL

Im Dialogfeld FUNKTIONSARGUMENTE können Sie für beide Funktionen das Quantil in das Feld K entweder als Dezimalzahl (0,75) oder als Prozentzahl eingeben (75 %). Falls Sie die zweite Variante wählen, müssen Sie das Prozentzeichen ebenfalls eingeben.

Gezogen und quartiliert

Einige bestimmte Quantile werden häufig verwendet, um Gruppen von Werten zusammenzufassen. Der Median, das 50 %- Quantil (weil größer als 50 % der Werte), ist ein Beispiel für ein solches häufiger verwendetes Quantil. Drei weitere Beispiele sind das 25 %-Quantil, das 75 %-Quantil und das 100 %-Quantil (der Maximalwert). Da diese Quantile eine Gruppe Werte in Viertel teilen, werden diese vier Quantile als *Quartile* (Viertelwerte) bezeichnet. Mit den Excel-Funktionen `QUARTILE.INKL` und `QUARTILE.EXKL` können Sie diese berechnen. Die INKL-Variante rechnet inklusive und die EXKL-Variante rechnet exklusive.

Wenn Sie im Dialogfeld FUNKTION EINFÜGEN auf `QUARTILE.INKL` klicken, wird das in der folgenden Abbildung dargestellte Dialogfeld FUNKTIONSARGUMENTE für die Funktion `QUARTILE.INKL` angezeigt. (Das Dialogfeld für `QUARTILE.EXKL` sieht genauso aus; beide Funktionen verwenden die gleichen Argumente.)



Wichtig dabei ist, dass im Feld QUARTILE die richtigen Zahlen eingegeben werden: 1 für das 25 %-Quantil (das 1. beziehungsweise untere Quartil), 2 für das 50 %-Quantil (das mittlere Quartil beziehungsweise der Medianwert), 3 für das 75 %-Quantil (das 3. beziehungsweise obere Quartil) und 4 für den Maximalwert. Wenn Sie im Feld QUARTILE eine 0 eingeben, wird der niedrigste Wert in der Gruppe angezeigt.

QUANTILSRANG.INKL und QUANTILSRANG.EXKL

Mit den beiden QUANTILSRANG-Tabellenblattfunktionen können Sie den prozentualen Rang eines angegebenen Werts ermitteln. Die INKL-Variante gibt die Quantile für größer oder gleich zurück. Die EXKL-Version gibt die Quantile für größer als zurück.

Auch hier möchte ich die Unterschiede zwischen den Funktionen mit einem Tabellenblatt verdeutlichen, das die Ergebnisse der Funktionen nebeneinander stellt. Abbildung 6.8 zeigt die Werte (in absteigender Reihenfolge) in Spalte C; die inklusiven Quantilsränge werden in Spalte D und die exklusiven Quantilsränge in Spalte E ausgegeben. Beachten Sie, dass der inklusive Quantilsrang für den höchsten Wert (80) 1 und dass der exklusive Quantilsrang 0,937 beträgt. Der inklusive Rang für den kleinsten Wert (20) beträgt 0; der exklusive Rang hingegen ist 0,062.

	A	B	C	D	E
1			Wert	QUANTILSRANG.INKL=	QUANTILSRANG.EXKL=
2			80	1	0,937
3			78	0,785	0,75
4			78	0,785	0,75
5			78	0,785	0,75
6			67	0,714	0,687
7			65	0,642	0,625
8			54	0,571	0,562
9			48	0,428	0,437
10			48	0,428	0,437
11			45	0,357	0,375
12			44	0,285	0,312
13			34	0,214	0,25
14			23	0,071	0,125
15			23	0,071	0,125
16			22	0	0,062

Abbildung 6.8: QUANTILSRANG.INKL **versus** QUANTILSRANG.EXKL

Führen Sie folgende Schritte durch, um die Funktion QUANTILSRANG.INKL zu verwenden:

1. **Geben Sie die Daten in einen Zellbereich ein, und markieren Sie eine Zelle.**

2. **Wählen Sie im Menü mit den statistischen Funktionen QUANTILSRANG.INKL, damit das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion geöffnet wird.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Parameter ein.**

   Abbildung 6.9 zeigt die Daten, die ausgewählte Zelle und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion QUANTILSRANG.INKL. In das Feld ARRAY habe ich den Zellbereich C2:C16 eingegeben; im Feld X steht der Bezug zur Zelle C2. Diese Zelle enthält den Wert, dessen Quantilsrang ich berechnen will. Mit dem Feld GENAUIGKEIT legen Sie die Anzahl der Nachkommastellen fest; wenn Sie das Feld leer lassen, werden drei Nachkommastellen verwendet.

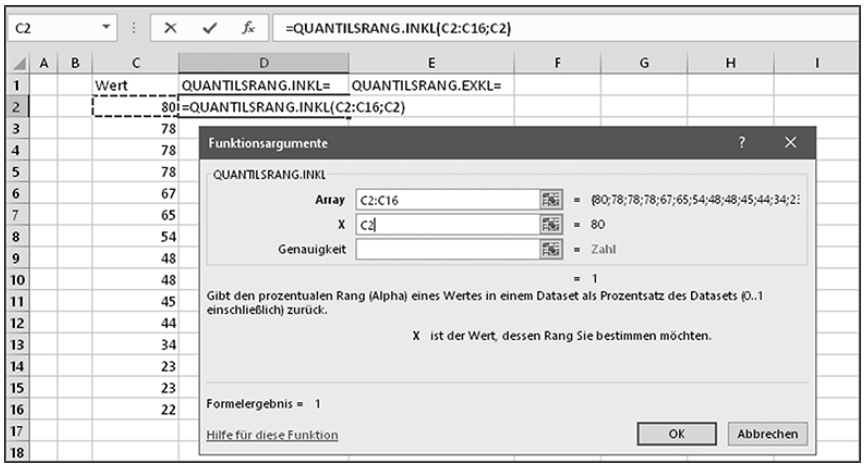


Abbildung 6.9: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion QUANTILSRANG.INKL

4. **Klicken Sie auf OK, um das Quantil in die ausgewählte Zelle einzutragen.**

Datenanalysetool: Rang und Quantil

Wie die Überschrift dieses Abschnitts bereits andeutet: Excel stellt ein Datenanalysetool bereit, mit dem Ränge und Quantile der einzelnen Werte in einer Gruppe berechnet werden können. Mit dem Tool *Rang und Quantil* werden beide Werte gleichzeitig berechnet. Somit sparen Sie gegenüber der Verwendung der beiden Tabellenblattfunktionen einige Schritte. (In Kapitel 2 wird beschrieben, wie die Datenanalysetools von Excel installiert werden.) In Abbildung 6.10 habe ich die Prüfungsergebnisse aus dem Beispiel für den z-Wert verwendet und das Dialogfeld RANG UND QUANTIL geöffnet. (*Mac-Anwender*: StatPlus stellt dieses Tool nicht zur Verfügung.)

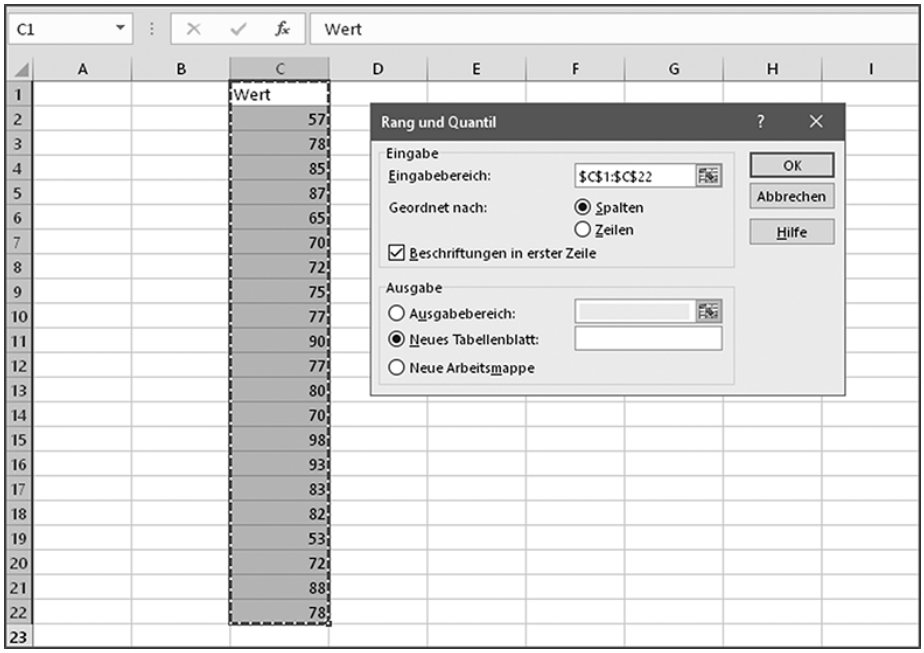


Abbildung 6.10: Das Datenanalysetool RANG UND QUANTIL

So verwenden Sie das Tool RANG UND QUANTIL:

1. **Geben Sie die Daten in einen Zellbereich ein.**

   In diesem Beispiel befinden sich die Daten in den Zellen C2 bis C22.

2. **Klicken Sie auf der Registerkarte DATEN in der Gruppe ANALYSE auf DATENANALYSE, um das Dialogfeld ANALYSE-**

FUNKTIONEN zu öffnen.

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option RANG UND QUANTIL aus.**

4. **Klicken Sie auf OK, um das Dialogfeld RANG UND QUANTIL zu öffnen.**

5. **Geben Sie im Dialogfeld RANG UND QUANTIL im Feld EINGABEBEREICH den Datenbereich ein. Achten Sie darauf, dass Sie den Zellbereich als absoluten Zellbezug eingeben.**

   In diesem Beispiel befindet sich in der ersten Zeile (in Zelle C1) eine Spaltenüberschrift. Ich möchte diese Überschrift in die Ausgabe übernehmen und gebe daher in das Feld EINGABEBEREICH **$C$22** ein und aktiviere das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE.

6. **Aktivieren Sie das Optionsfeld SPALTEN, um anzugeben, dass die Daten nach Spalten geordnet werden.**

7. **Aktivieren Sie das Optionsfeld NEUES TABELLENBLATT, um in der Arbeitsmappe ein neues Tabellenblatt zu erstellen und die Ergebnisse in diesem neu erstellten Tabellenblatt anzuzeigen.**

8. **Klicken Sie auf OK, um das Dialogfeld zu schließen. Öffnen Sie das neu erstellte Tabellenblatt, um die Ergebnisse anzuzeigen.**

In Abbildung 6.11 ist das neue Tabellenblatt mit den Ergebnissen dargestellt. In der Tabelle sind die Werte vom höchsten zum niedrigsten geordnet, wie anhand der Spalten WERT und RANG zu sehen ist. Die Spalte DATENPUNKT zeigt Ihnen die Position des Werts in der ursprünglichen Anordnung. Die 98 in Zelle B2 ist beispielsweise der 12. Wert in der ursprünglichen Datengruppe. In der Spalte PROZENT ist das Quantil der einzelnen Werte angegeben, so wie es auch `QUANTIL.INKL` macht.

	A	B	C	D	E
1	*Datenpunkt*	*Wert*	*Rang*	*Prozent*	
2	14	98	1	100,00%	
3	15	93	2	95,00%	
4	10	90	3	90,00%	
5	20	88	4	85,00%	
6	4	87	5	80,00%	
7	3	85	6	75,00%	
8	16	83	7	70,00%	
9	17	82	8	65,00%	
10	12	80	9	60,00%	
11	2	78	10	50,00%	
12	21	78	10	50,00%	
13	9	77	12	40,00%	
14	11	77	12	40,00%	
15	8	75	14	35,00%	
16	7	72	15	25,00%	
17	19	72	15	25,00%	
18	6	70	17	15,00%	
19	13	70	17	15,00%	
20	5	65	19	10,00%	
21	1	57	20	5,00%	
22	18	53	21	0,00%	
23					

Abbildung 6.11: Die Ausgabe des Datenanalysetools RANG UND QUANTIL

Kapitel 7

Alles zusammenfassen

In diesem Kapitel

- Mit großen und kleinen Dingen arbeiten
- Symmetrie, Spitzen und Plateaus verstehen
- Beschreibend werden
- Statistik auf dem silbernen Tablett servieren

Mit der zentralen Tendenz und Streuung lassen sich mehrere Werte prima zusammenfassen. Diese beiden Methoden sind jedoch nicht die einzigen Möglichkeiten. Die zentrale Tendenz und die Streuung sind nur ein Teil der Populationskenngrößen. Einige Populationskenngrößen wie Anzahl, Maximum und Minimum sind leicht verständlich. Andere, wie Schiefe und Wölbung, nicht.

In diesem Kapitel geht es um Populationskenngrößen und wie diese mit den Funktionen von Excel berechnet und angezeigt werden können.

Auszählen

Die wichtigste Populationskenngröße, die ich mir vorstellen kann, ist die Anzahl der Werte in einer Wertegruppe. Bei Excel gibt es fünf Möglichkeiten, diese Größe zu ermitteln. Ja, fünf Möglichkeiten. Zählen Sie nach.

ANZAHL, ANZAHL2, ANZAHLLEEREZELLEN,

ZÄHLENWENN und ZÄHLENWENNS

Wenn Sie einen Zellbereich haben, können Sie mit ANZAHL die Anzahl der Zellen ermitteln, die numerische Daten enthalten. In Abbildung 7.1 habe ich mehrere Werte eingegeben, eine Zelle für das Ergebnis der Funktion ANZAHL markiert und das Dialogfeld FUNKTIONSARGUMENTE für ANZAHL geöffnet.

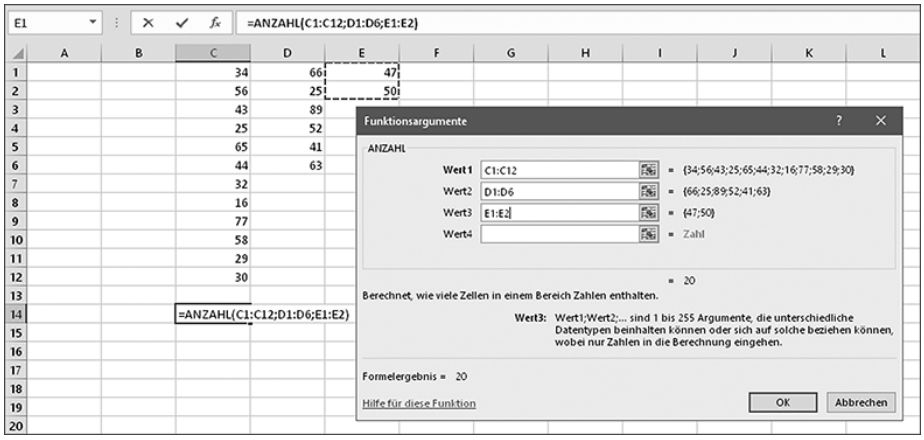


Abbildung 7.1: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion ANZAHL mit mehreren Argumenten

Und so müssen Sie vorgehen:

1. **Geben Sie die gewünschten Werte in das Arbeitsblatt ein, und markieren Sie eine Zelle für das Ergebnis.**

   Ich habe Zellbereiche aus den Spalten C, D und E eingegeben, um zu zeigen, dass ANZAHL mit mehreren Argumenten umgehen kann. Für das Ergebnis habe ich die Zelle C14 markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion ANZAHL aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE Werte für die Argumente ein.**

- In das Feld WERT1 habe ich eine der Datenspalten dieses Beispiels eingegeben, und zwar **C1:C12**.
- Ich habe in das Feld WERT2 geklickt und eine weitere Datenspalte eingegeben; in diesem Beispiel ist es **D1:D6**.
- Ich habe in das Feld WERT3 geklickt und dort **E1:E2**, den letzten Zellbereich, eingegeben.

4. **Klicken Sie auf OK, damit die Anzahl in der markierten Zelle angezeigt wird.**

ANZAHL2 funktioniert wie ANZAHL, außer dass die Anzahl hierbei auch Zellen mit Text und logischen Werten beinhaltet.

Mit ANZAHLLEEREZELLEN wird die Anzahl der leeren Zellen in einem Zellbereich gezählt. In Abbildung 7.2 ist zu sehen, dass ich die Zahlen aus dem vorhergehenden Beispiel verwendet, jedoch den Bereich um die Zellen D7 bis D12 und E3 bis E12 erweitert habe. Im Feld BEREICH ist C1:E12 eingegeben. Im Dialogfeld FUNKTIONSARGUMENTE für ANZAHLLEEREZELLEN wird die Anzahl leerer Zellen (16 in diesem Beispiel) angezeigt.

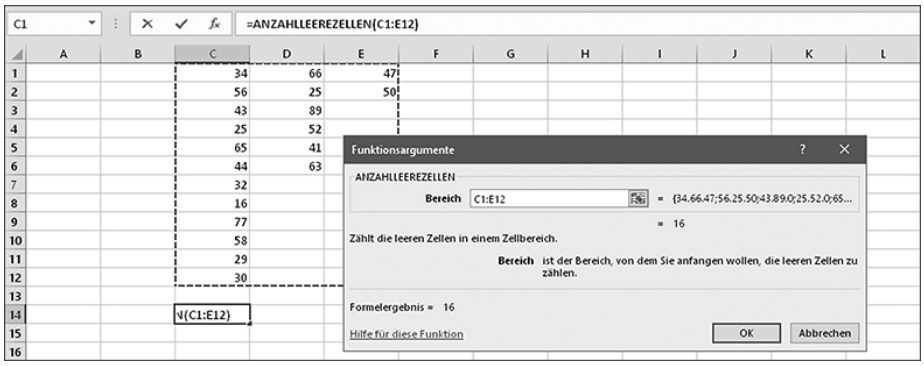


Abbildung 7.2: Mit ANZAHLLEEREZELLEN werden die leeren Zellen in einem angegebenen Bereich gezählt.

ZÄHLENWENN zeigt die Anzahl der Zellen an, deren Wert ein bestimmtes Kriterium erfüllt. In Abbildung 7.3 habe ich die Werte noch einmal verwendet, um das Dialogfeld FUNKTIONSARGUMENTE für

ZÄHLENWENN aufzurufen. Der Bereich umfasst die Zellen C1:E12. ZÄHLENWENN berücksichtigt leere Zellen jedoch nicht.

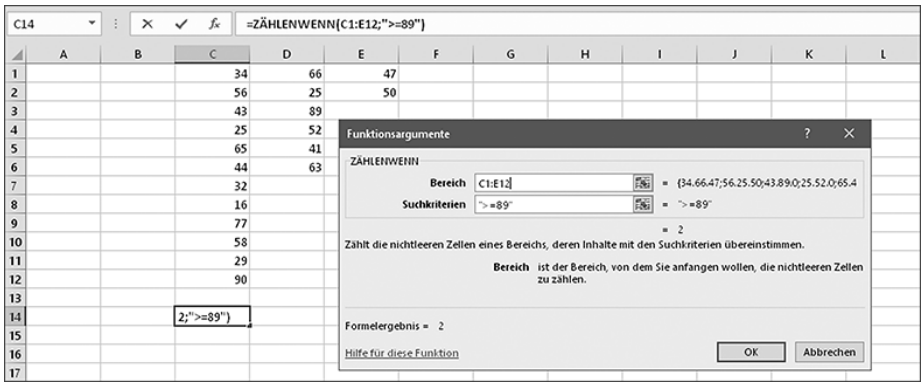


Abbildung 7.3: ZÄHLENWENN zeigt die Anzahl der Zellen an, deren Wert ein bestimmtes Kriterium erfüllt.

Mit dem hier verwendeten Kriterium (>=89) wird ZÄHLENWENN angewiesen, nur die Zellen zu zählen, deren Wert größer als oder gleich 89 ist. In diesem Beispiel ist die Anzahl 2.

Bei der Verwendung von ZÄHLENWENN macht es nicht so viel aus, jedoch können Sie an diesem Beispiel eine kleine Macke von Excel erkennen. Wenn Sie die Kriterien in doppelte Anführungszeichen einschließen, zeigt Excel das Ergebnis an, bevor Sie auf OK klicken. Falls Sie dies nicht machen, wird das Formelergebnis nicht angezeigt. Wenn Sie auf OK klicken oder in ein anderes Eingabefeld wechseln, werden die doppelten Anführungszeichen von Excel automatisch ergänzt.

ZÄHLENWENNS unterstützt beim Zählen von Zellen mehrere Bedingungen. Wenn die Kriterien aus zwei Zellbereichen stammen, müssen diese die gleiche Anzahl von Zellen aufweisen. Dies liegt daran, dass ZÄHLENWENNS Zellpaare zählt. Ein bestimmtes Zellpaar wird nur dann beim Zählen berücksichtigt, wenn eine der Zellen das eine und eine der anderen Zellen das andere Kriterium erfüllt. Werfen Sie einen Blick auf Abbildung 7.4.

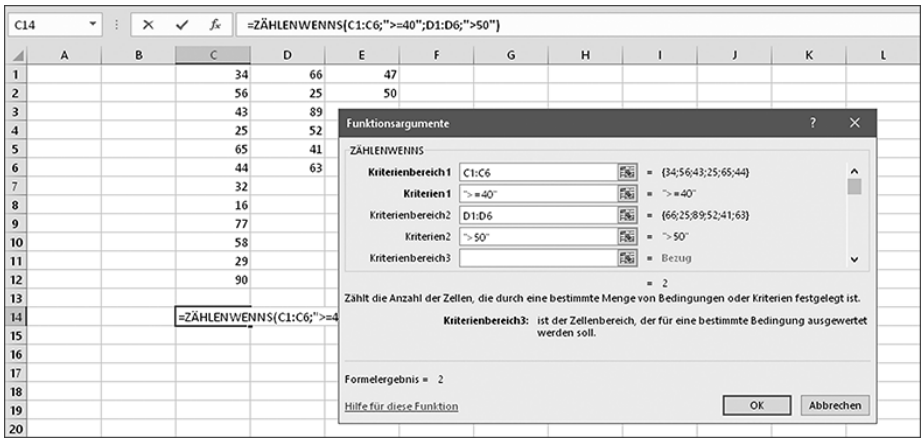


Abbildung 7.4: Die Funktion ZÄHLENWENNS verwenden

In diesem Beispiel verwendet `ZÄHLENWENNS` die Kriterienbereiche C1:C6 und D1:D6. Das Kriterium für die Zellen in Spalte C ist > 40. Das Kriterium für die Zellen in Spalte D ist > 50. Dies bedeutet, dass `ZÄHLENWENNS` die Zellpaare zählt, bei denen die Zelle in Spalte C größer ist als 40 und die betreffende Zelle in Spalte D größer ist als 50. Diese Bedingung trifft lediglich auf zwei Zellpaare zu, wie Sie im Dialogfeld sehen können.

Sie können den gleichen Zellbereich auch mehr als einmal bei `ZÄHLENWENNS` angeben. Ein Beispiel:

```
=ZÄHLENWENNS(C1:C6;">30";C1:C6;"<60")
```

Diese Formel gibt die Anzahl der Zellen zurück, deren Wert zwischen 30 und 60 liegt (30 und 60 nicht eingeschlossen).

Groß und klein

Es gibt zwei weitere Populationskenngrößen, die wohl ebenfalls nicht ausführlich erklärt werden müssen: das Minimum und das Maximum. Hierbei handelt es sich natürlich um den größten und den kleinsten Wert in einer Wertegruppe.

MAX, MAXA, MIN und MINA

Excel enthält Tabellenblattfunktionen, mit denen Sie den größten und kleinsten Wert in einer Gruppe ermitteln können. Ich werde die

Funktion MAX ausführlich erläutern. Die anderen Funktionen arbeiten ähnlich.

Abbildung 7.5 verwendet die Werte aus den vorherigen Beispielen.

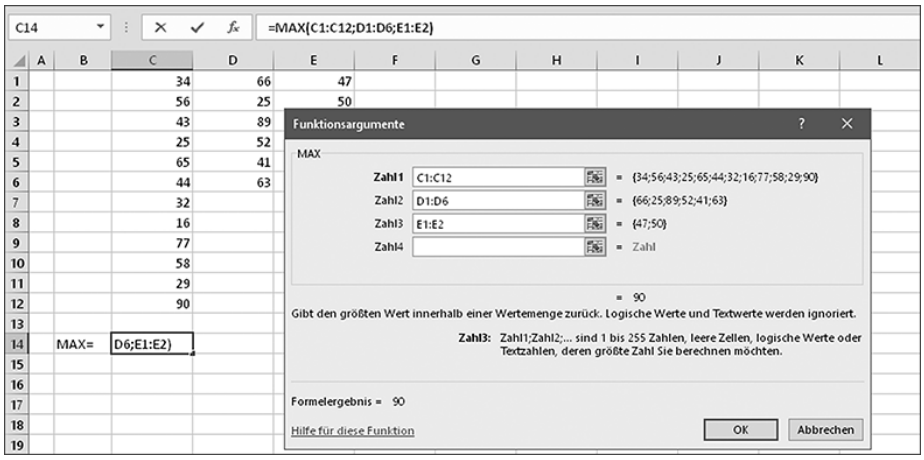


Abbildung 7.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion MAX

Ich habe eine Zelle für den größten Wert markiert und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion MAX geöffnet. Und so müssen Sie vorgehen:

1. **Geben Sie die gewünschten Werte in das Arbeitsblatt ein und markieren Sie eine Zelle für den größten Wert.**

   Ich habe Werte in die Spalten C, D und E eingegeben, um zu zeigen, dass MAX mit mehreren Argumenten umgehen kann. In diesem Beispiel habe ich die Zelle C14 als Ergebniszelle markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion MAX aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE Werte für die Argumente ein.**

   - In das Feld ZAHL1 habe ich eine der Datenspalten eingegeben: **C1:C12**.

- Das Anklicken von Feld ZAHL2 erstellt ein weiteres Feld, ZAHL3. In das Feld ZAHL2 habe ich einen weiteren Zellbereich eingegeben: **D1:D6**.

- Ich habe in das Feld ZAHL3 geklickt und dort **E1:E2**, den dritten und letzten Zellbereich dieses Beispiels, eingegeben.

4. **Klicken Sie auf OK, damit der größte Wert in C14 angezeigt wird.**

MAX berücksichtigt Text und logische Werte nicht. MAXA berücksichtigt dagegen Text und logische Werte beim Ermitteln des größten Werts. MAXA interpretiert den logischen Wert WAHR als 1. Den logischen Wert FALSCH oder einen anderen Text als WAHR interpretiert die Funktion als 0.

MIN und MINA funktionieren so wie MAX und MAXA, ermitteln jedoch anstelle des größten Werts den kleinsten. Verwenden Sie die Funktion MINA mit Vorsicht, da sich die Interpretation von logischen Werten und Text als 0 und 1 auf das Ergebnis auswirkt. Bei den Zahlen aus dem vorherigen Beispiel ergibt sich für den kleinsten Wert 22. Wenn Sie in einem der Bereiche in eine Zelle FALSCH oder einen anderen Text eingeben, ermittelt MINA 0 als den kleinsten Wert. Wenn Sie WAHR eingeben, zeigt MINA 1 als kleinsten Wert an.

Jetzt wird es esoterisch

In diesem Abschnitt werden seltener verwendete Kenngrößen beschrieben, die mit dem Mittelwert und der Varianz verwandt sind. Meist reichen Mittelwert und Varianz zum Beschreiben einer Datenmenge aus. Diese anderen Kenngrößen, *Schiefe* und *Wölbung*, gehen nur ein kleines Stückchen weiter. Möglicherweise werden Sie diese Kenngrößen eines Tages einmal brauchen, um eine riesige Menge Daten ausführlich zu beschreiben.

Stellen Sie sich den Mittelwert so vor, dass Sie eine Gruppe von Werten dadurch *finden*, dass angezeigt wird, wo sich deren Zentrum

befindet. Dies ist der Ausgangspunkt für die anderen Kenngrößen. In Bezug auf den Mittelpunkt

- beschreibt die Varianz, wie die Werte *gestreut* sind
- gibt die Schiefe an, wie *symmetrisch* die Werte verteilt sind
- zeigt die Wölbung, ob die Verteilung der Werte um den Mittelpunkt eine *Spitze* aufweist.

Die Schiefe und die Wölbung sind mit dem Mittelwert und der Varianz auf ziemlich mathematische Weise verbunden. Die Varianz ist die Summe der quadratischen Abweichungen der Werte um den Mittelwert. Die Schiefe ist die Summe der dritten Potenz der Abweichungen um den Mittelwert. Bei der Wölbung geht es um eine noch höhere Potenz, nämlich die vierte Potenz, um genau zu sein. Ich werde das alles in den nachfolgenden Abschnitten etwas genauer erklären.

SCHIEFE und SCHIEFE.P

In Abbildung 7.6 sehen Sie drei verschiedene Histogramme. Das erste ist symmetrisch, die anderen beiden nicht. Die Symmetrie und die Asymmetrie werden mit der Kenngröße Schiefe beschrieben.

Beim symmetrischen Histogramm ist die Schiefe null. Beim zweiten Histogramm, das nach rechts hin abflacht, ist der Wert der Schiefe positiv. Man spricht hier auch von einer rechtsschiefen oder linkssteilen Verteilung. Beim dritten Histogramm, das nach links hin abflacht, ist der Wert der Schiefe negativ. Man spricht hier auch von einer linksschiefen oder rechtssteilen Verteilung.

Wie kommt es zu einer positiven, einer negativen und keiner Schiefe? Von dieser Formel:

$$Schiefe = \frac{\Sigma(x-\bar{x})^3}{(n-1)s^3}$$

In der Formel steht $\bar{x}$ für den Mittelwert der Werte, *n* für die Anzahl der Werte und *s* für die Standardabweichung. Dies ist die Formel für die Schiefe einer Stichprobe, die Formel für die Schiefe einer Grundgesamtheit verwendet anstelle von *n-1* lediglich *n*.

Ich nenne diese Formel nur der Vollständigkeit halber. Wenn Sie jemals mit Schiefe einer Stichprobe zu tun haben, werden Sie diese Formel sowieso nicht verwenden, da Sie den Wert mit der Excel-Funktion `SCHIEFE` berechnen können.

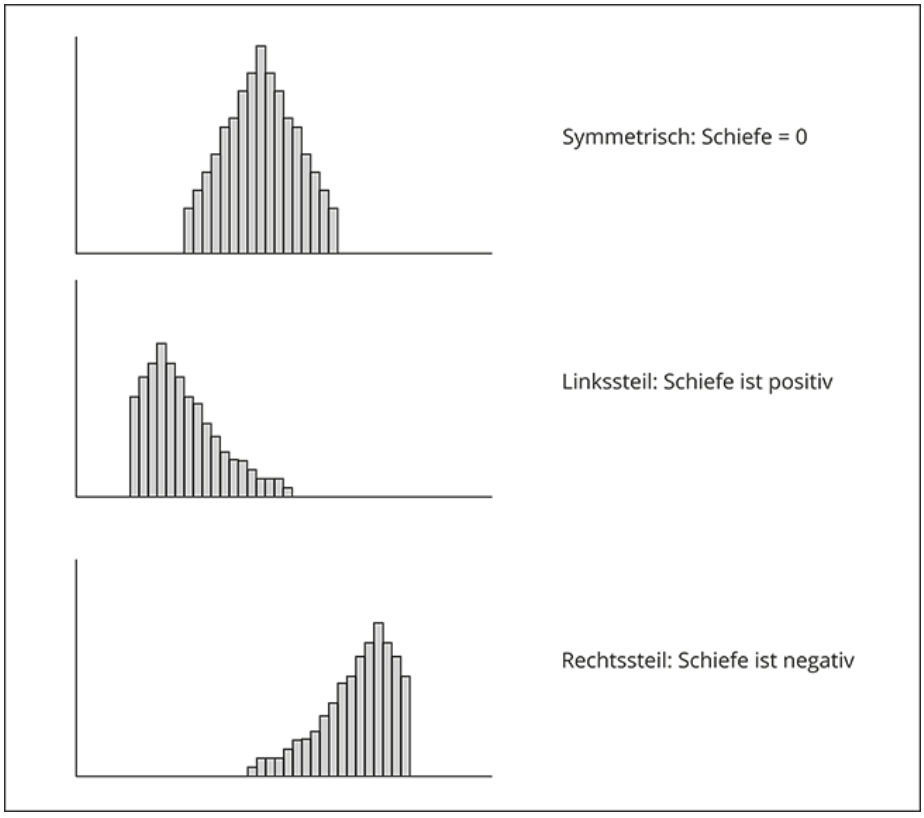


Abbildung 7.6: Drei Diagramme, die die unterschiedlichen Schiefen zeigen

So verwenden Sie die Funktion `SCHIEFE`:

1. **Geben Sie die gewünschten Zahlen in ein Arbeitsblatt ein, und markieren Sie eine Zelle für das Ergebnis.**

   In diesem Beispiel habe ich die Werte in die ersten zehn Zeilen der Spalten B, C, D und E eingegeben (siehe Abbildung 7.7). Ich habe die Zelle H2 für das Ergebnis gewählt.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion SCHIEFE aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE Werte für die Argumente ein.**

   In das Feld ZAHL1 habe ich den Zellbereich mit den Daten eingegeben. In diesem Beispiel ist dies der Bereich B1:E10. Nachdem der Zellbereich festgelegt wurde, wird im Dialogfeld FUNKTIONSARGUMENTE die Schiefe angezeigt, die in diesem Beispiel negativ ist.

4. **Klicken Sie auf OK, damit der Wert der Schiefe in die markierte Zelle eingefügt wird.**

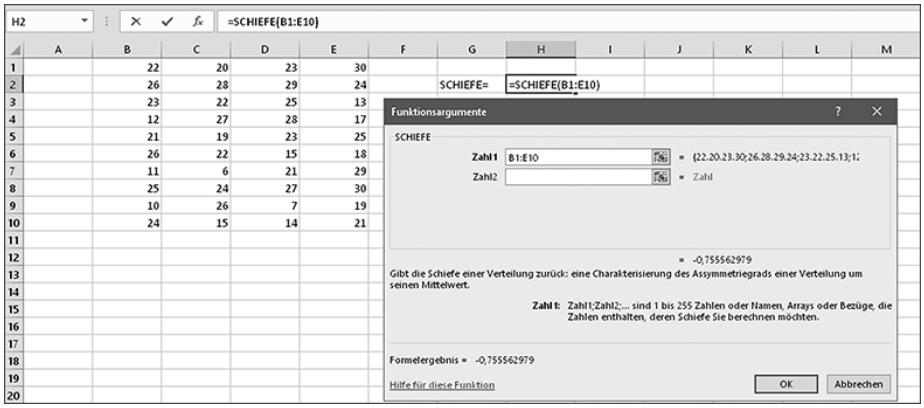


Abbildung 7.7: Mit der Funktion *SCHIEFE* die Schiefe berechnen

Das Dialogfeld FUNKTIONSARGUMENTE für `SCHIEFE.P` (Schiefe einer Grundgesamtheit) sieht genauso aus. Bei der Berechnung der Schiefe einer Population wird jedoch *N* und nicht wie bei einer Stichprobe *n-1* verwendet.

KURT

In Abbildung 7.8 sehen Sie zwei Histogramme. Beim ersten befindet sich eine Spitze in der Mitte, das zweite ist flach. Das erste wird als *leptokurtisch* oder *schmalgipflig* bezeichnet. Die Kurtosis ist positiv. Das zweite Histogramm ist *platykurtisch* oder *breitgipflig*. Die Kurtosis ist negativ.

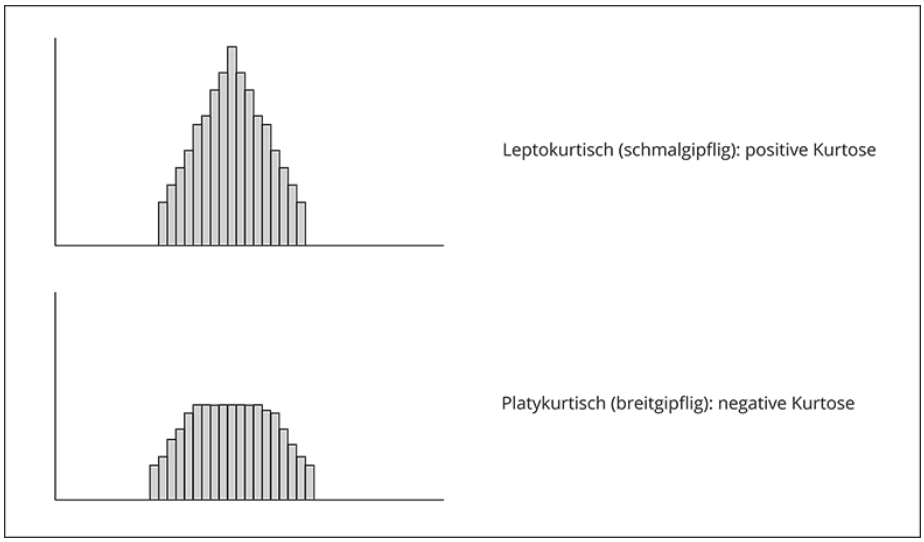


Abbildung 7.8: Zwei Histogramme mit unterschiedlicher Kurtosis

Negativ? Aber einen Moment mal. Wie kann das sein? Ich habe weiter oben erwähnt, dass sich die Wölbung aus der Summe der vierten Potenz der Abweichungen vom Mittelwert berechnet. Da 4 eine gerade Zahl ist, ist die vierte Potenz auch bei einer negativen Abweichung positiv. Wie kann sich also beim Addieren aller positiven Zahlen ein negativer Wölbungswert ergeben?

Ganz einfach. Die Formel für die Wölbung lautet

$$Kurtosis = \frac{\Sigma(x-\bar{x})^2}{(n-1)s^4} - 3$$

wobei $\bar{x}$ für den Mittelwert der Werte, *n* für die Anzahl der Werte und *s* für die Standardabweichung steht.

Hm … warum 3? Die 3 kommt ins Spiel, da dieser Wert das Wölbungsmaß der so genannten *Standardnormalverteilung* angibt. (Die Normalverteilungen werden in Kapitel 8 beschrieben.) Statistiker bezeichnen diese Formel als *Kurtosis* oder *Exzess*. Die Kurtosis ist ein Maß für die Wölbung einer Verteilung von Werten im Vergleich zur Standardnormalverteilung. Fragen Sie lieber nicht, warum die Wölbung der Standardnormalverteilung 3 beträgt.

Dies ist eine weitere Formel, die Sie wahrscheinlich nie verwenden werden, weil Sie die Wölbung mit der Excel-Funktion `KURT` berechnen können. In Abbildung 7.9 sind die Werte aus dem vorherigen Beispiel, eine markierte Zelle und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `KURT` dargestellt.

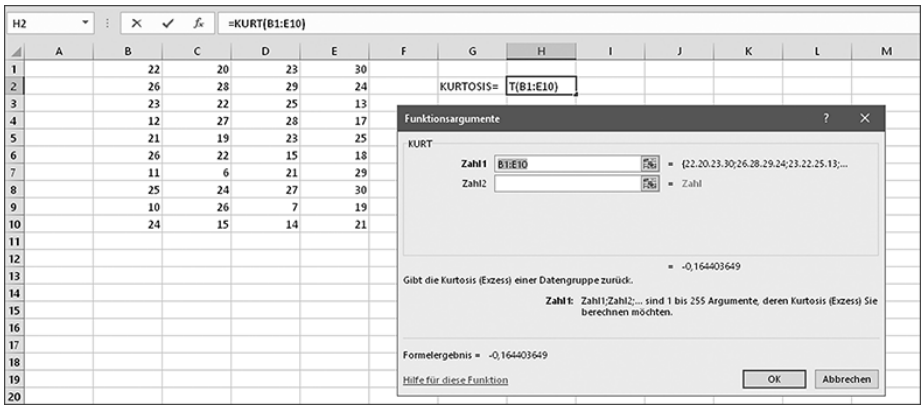


Abbildung 7.9: Mit `KURT` die Wölbung berechnen

So verwenden Sie die Funktion `KURT`:

1. **Geben Sie die gewünschten Zahlen in ein Arbeitsblatt ein und markieren Sie eine Zelle für das Ergebnis.**

   In diesem Beispiel habe ich die Werte in die ersten zehn Zeilen der Spalten B, C, D und E eingegeben. Ich habe die Zelle H2 für das Ergebnis gewählt.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion KURT aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE Werte für die Argumente ein.**

   In das Feld ZAHL1 habe ich den Zellbereich mit den Daten eingegeben. In diesem Beispiel umfasst der Zellbereich die Zellen B1:E10. Nachdem der Zellbereich eingegeben wurde, zeigt das Dialogfeld FUNKTIONSARGUMENTE die Wölbung (Kurtosis) an, die in diesem Beispiel negativ ist.

4. **Klicken Sie auf OK, damit das Ergebnis in der markierten Zelle angezeigt wird.**

Nun kommt die Häufigkeit ins Spiel

Die Berechnung von Schiefe und Wölbung ist gut und schön. Damit wird aber nichts darüber ausgesagt, wie die Werte verteilt sind. Dazu müssen Sie eine *Häufigkeitsverteilung* erstellen. Das ist eine Tabelle, mit der die möglichen Werte in Intervalle unterteilt und die Anzahl der Werte dargestellt wird, die sich in den einzelnen Intervallen befinden.

Mit Excel haben Sie zwei Möglichkeiten, eine Häufigkeitsverteilung zu erstellen. Sie können entweder eine Tabellenblattfunktion oder ein Datenanalysetool verwenden.

HÄUFIGKEIT

Die Tabellenblattfunktion `HÄUFIGKEIT` habe ich bereits in Kapitel 2 im Zusammenhang mit den Arrayformeln erläutert. Hier betrachte ich sie von einer anderen Warte. Im folgenden Beispiel verwende ich die Daten aus dem Beispiel zur Schiefe und zur Wölbung, damit Sie sehen, wie diese Werte verteilt sind.

In Abbildung 7.10 sind die Daten noch einmal mit einem markierten Zellbereich mit der Bezeichnung *Häufigkeit* dargestellt. Außerdem wurde eine Spalte mit der Überschrift *Intervall* versehen. In dieser Spalte befinden sich die Intervallgrenzen. Die Zahlen in dieser Spalte stellen jeweils die obere Grenze eines Intervalls dar. In der Abbildung ist außerdem das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `HÄUFIGKEIT` dargestellt.

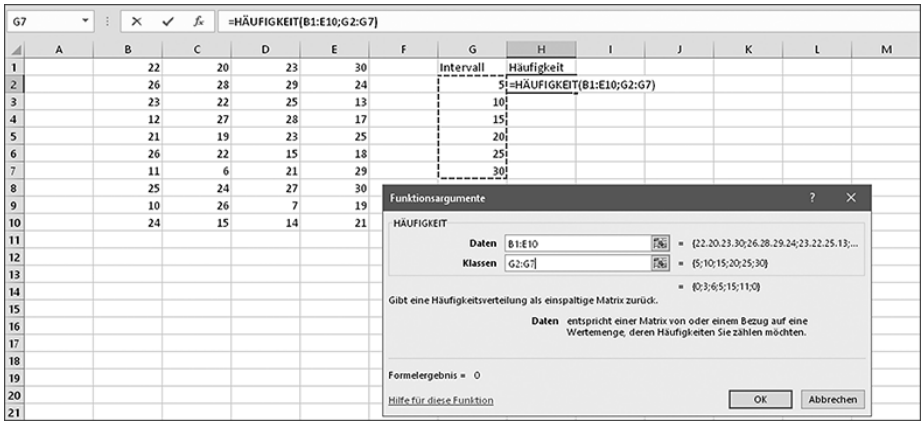


Abbildung 7.10: Die Häufigkeiten in einem Zellbereich suchen

Da es sich hierbei um eine Arrayfunktion handelt, sind die Schritte zur Verwendung von `HÄUFIGKEIT` ein wenig anders als bei den bisherigen Funktionen in diesem Kapitel.

1. **Geben Sie die Werte in einen Zellbereich ein.**

   Die Daten befinden sich hier wie beim vorherigen Beispiel im Zellbereich B1:E10.

2. **Geben Sie die Intervalle in einen Zellbereich ein.**

   Ich habe in den Zellbereich G2:G7 die Werte 5, 10, 15, 20, 25 und 30 eingegeben.

3. **Wählen Sie einen Zellbereich für die Häufigkeiten aus.**

   Ich habe Spalte H mit *Häufigkeit* beschriftet und die Zellen H2 bis H7 für die sich ergebenden Häufigkeiten markiert.

4. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion HÄUFIGKEIT aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

5. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE Werte für die Argumente ein.**

   In das Feld DATEN habe ich den Zellbereich eingegeben, in dem sich die Werte befinden. In diesem Beispiel ist das B1:E10.

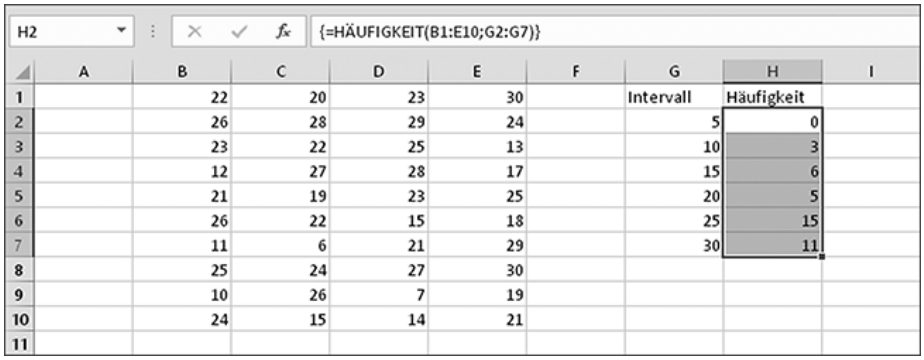
   Bei der Funktion HÄUFIGKEIT werden Intervalle als Klassen bezeichnet und die Intervalle werden im Feld KLASSEN eingegeben. In diesem Beispiel steht im Feld KLASSEN G2:G7

   Nachdem beide Zellbereiche angegeben wurden, werden im Dialogfeld die Häufigkeiten in geschwungenen Klammern angezeigt. Wenn Sie sich Abbildung 7.10 genau ansehen, stellen Sie fest, dass Excel am Ende der Liste mit den Häufigkeiten (dritte Textzeile auf der rechten Seite) die Häufigkeit 0 eingefügt hat.

6. **Drücken Sie die Tastenkombination Strg+Shift+Enter, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen.**

   Sie müssen diese Tastenkombination verwenden, da es sich bei HÄUFIGKEIT um eine Arrayfunktion handelt.

Nach dem Schließen des Dialogfelds FUNKTIONSARGUMENTE werden die Häufigkeiten in den gewünschten Zellen angezeigt (siehe Abbildung 7.11).



H2 | {=HÄUFIGKEIT(B1:E10;G2:G7)}

	A	B	C	D	E	F	G	H	I
1		22	20	23	30		Intervall	Häufigkeit	
2		26	28	29	24		5	0	
3		23	22	25	13		10	3	
4		12	27	28	17		15	6	
5		21	19	23	25		20	5	
6		26	22	15	18		25	15	
7		11	6	21	29		30	11	
8		25	24	27	30				
9		10	26	7	19				
10		24	15	14	21				
11									

Abbildung 7.11: `HÄUFIGKEIT`s Häufigkeiten

Hätte ich dem Zellbereich mit den Werten den Namen *Daten* und dem Zellbereich mit den Intervallen den Namen *Intervall* zugewiesen und dann im Dialogfeld FUNKTIONSARGUMENTE diese Namen verwendet, würde die Formel wie folgt aussehen:

```
=HÄUFIGKEIT(Daten;Intervall)
```

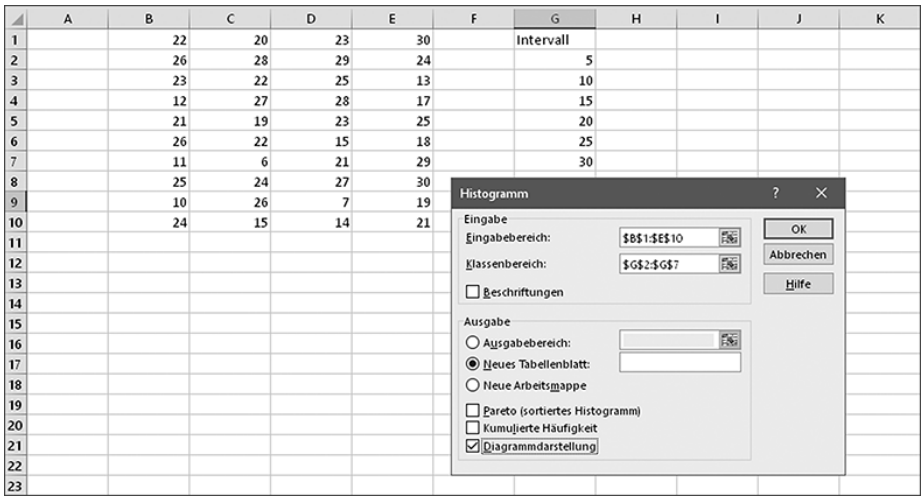
Diese Formel ist leichter zu verstehen als diese:

```
=HÄUFIGKEIT(B1:E10;G2:G7)
```

(Wissen Sie nicht mehr, wie Sie einen Zellbereich mit einem Namen versehen? Werden Sie dann einen Blick in Kapitel 2.)

Datenanalysetool: Histogramm

Eine weitere Möglichkeit, eine Häufigkeitsverteilung zu erstellen, ist das Datenanalysetool *Histogramm*. Damit Sie sehen, dass die beiden Methoden gleichwertig sind, verwende ich die Daten aus dem Beispiel für die Funktion `HÄUFIGKEIT`. In Abbildung 7.12 sind die Daten und das Dialogfeld HISTOGRAMM dargestellt.



	A	B	C	D	E	F	G	H	I	J	K
1		22	20	23	30		Intervall				
2		26	28	29	24		5				
3		23	22	25	13		10				
4		12	27	28	17		15				
5		21	19	23	25		20				
6		26	22	15	18		25				
7		11	6	21	29		30				
8		25	24	27	30						
9		10	26	7	19						
10		24	15	14	21						
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											

Abbildung 7.12: Das Datenanalysetool HISTOGRAMM

1. **Geben Sie die Werte in einen und die Intervalle in einen anderen Zellbereich ein.**

2. **Wählen Sie den Befehl DATEN | ANALYSE | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option HISTOGRAMM, um das Dialogfeld HISTOGRAMM zu öffnen.**

4. **Geben Sie im Dialogfeld HISTOGRAMM die Werte ein, die verwendet werden sollen.**

   In diesem Beispiel beinhaltet der Datenbereich die Zellen E1 bis E10. Am einfachsten geben Sie diesen Bereich ein, indem Sie auf B1 klicken, die Shift-Taste gedrückt halten und dann auf E10 klicken. Damit wird der Eingabebereich mit absoluten Zellbezügen ($E$1:$E$10) in das Feld EINGABEBEREICH übernommen.

   In das Feld KLASSENBEREICH gebe ich die Zellen ein, in denen sich die Intervalle befinden. In diesem Beispiel sind das die Zellen G2 bis G7. Ich klicke Zelle G2 an, halte dann die Shift-Taste gedrückt und klicke abschließend auf Zelle G7. Der

Bereich wird im Feld KLASSENBEREICH mit absoluten Zellbezügen ($G$2:$G$7) angezeigt.

5. **Aktivieren Sie das Optionsfeld NEUES TABELLENBLATT, um ein neues Tabellenblatt zu erstellen und die Ergebnisse in diesem neu erstellten Tabellenblatt anzuzeigen.**

6. **Aktivieren Sie das Kontrollkästchen DIAGRAMMDARSTELLUNG, um ein Histogramm zu erstellen und das Ergebnis anzuzeigen.**

7. **Klicken Sie auf OK, um das Dialogfeld zu schließen.**

In Abbildung 7.13 ist das Ergebnis des Datenanalysetools Histogramm dargestellt. In der Tabelle werden dieselben Ergebnisse angezeigt wie die, die sich aus der Funktion `HÄUFIGKEIT` ergeben haben. Beachten Sie, dass im Histogramm bei den Klassen die Spalte *und größer* ergänzt wird. Das Histogramm ist zunächst etwas kleiner. Ich habe die Größe mit der Maus verändert, bis es so groß war wie in der Abbildung dargestellt. Das Histogramm zeigt, dass die Verteilung nach links hin abflacht (entspricht der negativen Schiefe) und keine ausgeprägte Spitze aufweist (entspricht der negativen Kurtosis). Beachten Sie ebenfalls, dass rechts neben dem Diagramm die drei Schaltflächen DIAGRAMMELEMENTE, DIAGRAMMFORMATVORLAGEN und DIAGRAMMFILTER angezeigt werden, mit denen Sie das Histogramm verändern können (mehr hierzu finden Sie in Kapitel 3).

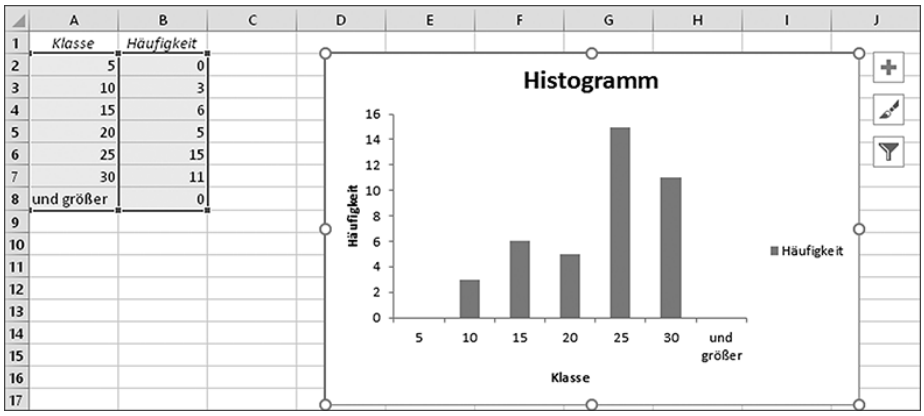


Abbildung 7.13: Das Ergebnis des Tools HISTOGRAMM (nachdem die Größe des Diagramms verändert wurde)

Im Dialogfeld HISTOGRAMM gibt es übrigens noch zwei weitere Kontrollkästchen: PARETO (SORTIERTES HISTOGRAMM) und KUMULIERTE HÄUFIGKEIT. Beim Pareto-Diagramm werden die Intervalle in absteigender Folge vom größten zum kleinsten Häufigkeitswert sortiert. Bei der kumulierten Häufigkeit wird die Häufigkeit von Werten in einem Intervall zusammen mit der Häufigkeit in allen vorherigen Intervallen dargestellt. Wenn Sie dieses Kontrollkästchen aktivieren, wird außerdem eine Linie für die kumulierte Häufigkeit in das Histogramm eingefügt. (Um das Pareto-Diagramm zu erstellen, das ich in Kapitel 3 beschreibe, müssen Sie sowohl das Kontrollkästchen PARETO (SORTIERTES HISTOGRAMM) als auch KUMULIERTE HÄUFIGKEIT einschalten.)

Können Sie mir eine Beschreibung geben?

Wenn Sie es mit einzelnen Populationskenngrößen zu tun haben, reichen die bisher beschriebenen Tabellenblattfunktionen aus. Wenn Sie jedoch einen Übersichtsbericht möchten, in dem alle Populationskenngrößen enthalten sind, müssen Sie das Datenanalysetool verwenden, das im nächsten Abschnitt beschrieben wird.

Datenanalysetool: Populationskenngrößen

In Kapitel 2 habe ich das Tool *Populationskenngrößen* als Einführung in die Datenanalysetools von Excel beschrieben. Im Folgenden geht es um ein etwas komplexeres Beispiel. In Abbildung 7.14 sind drei Spalten mit Werten und das Dialogfeld POPULATIONSKENNGRÖSSEN dargestellt. Die Spalten haben die Überschriften *Erste*, *Zweite* und *Dritte*, damit Sie sehen können, wie dieses Tool Beschriftungen verwendet.

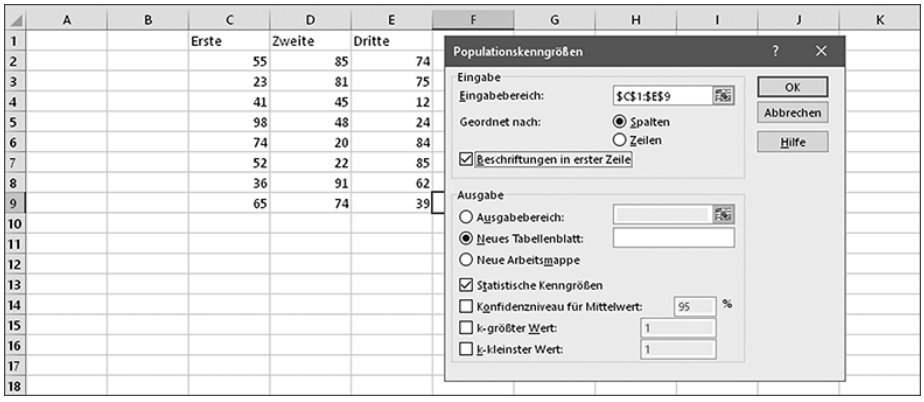


Abbildung 7.14: Das Tool POPULATIONSKENNGRÖSSEN

Hier folgen die Schritte, um dieses Tool zu verwenden:

1. **Geben Sie die Daten in einen Zellbereich ein.**

2. **Wählen Sie den Befehl DATEN | ANALYSE | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option POPULATIONSKENNGRÖSSEN, um das Dialogfeld POPULATIONSKENNGRÖSSEN zu öffnen.**

4. **Geben Sie im Dialogfeld POPULATIONSKENNGRÖSSEN die Werte ein, die verwendet werden sollen.**

   Im Feld EINGABEBEREICH habe ich die Daten eingegeben. Der einfachste Weg, um dies zu tun, ist, Zelle C1 anzuklicken, dann die Shift-Taste gedrückt zu halten und abschließend die untere rechte Ecke des Zellbereichs (Zelle E9) anzuklicken. Damit wird $C$1:$E$9 in das Feld EINGABEBEREICH eingefügt.

5. **Aktivieren Sie das Optionsfeld SPALTEN, um anzugeben, dass die Daten nach Spalten geordnet werden.**

6. **Aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE, da der Eingabebereich auch die Spaltenüberschriften enthält.**

7. **Aktivieren Sie das Optionsfeld NEUES TABELLENBLATT, um im aktuellen Arbeitsblatt ein neues Tabellenblatt zu erstellen und die Ergebnisse in diesem neu erstellten Tabellenblatt anzuzeigen.**

8. **Aktivieren Sie das Kontrollkästchen STATISTISCHE KENNGRÖSSEN. Die anderen Kontrollkästchen lassen Sie deaktiviert.**

9. **Klicken Sie auf OK, um das Dialogfeld zu schließen.**

   Das neue Tabellenblatt wird geöffnet und zeigt die Kenngrößen der Daten an.

   In Abbildung 7.15 ist für jede Spalte eine eigene Übersicht dargestellt. Wenn Sie dieses Tabellenblatt zum ersten Mal öffnen, sind die Spalten mit den Bezeichnungen für die Kenngrößen zu schmal. Daher habe ich die Spalten so verbreitert, wie es die Abbildung zeigt.

	A	B	C	D	E	F	G
1	*Erste*		*Zweite*		*Dritte*		
2							
3	Mittelwert	55,5	Mittelwert	58,25	Mittelwert	56,875	
4	Standardfehler	8,343089527	Standardfehler	10,01026259	Standardfehler	9,99006203	
5	Median	53,5	Median	61	Median	68	
6	Modus	#NV	Modus	#NV	Modus	#NV	
7	Standardabweichung	23,59782072	Standardabweichung	28,31329824	Standardabweichung	28,2561624	
8	Stichprobenvarianz	556,8571429	Stichprobenvarianz	801,6428571	Stichprobenvarianz	798,410714	
9	Kurtosis	0,288278448	Kurtosis	-1,75169467	Kurtosis	-1,30336482	
10	Schiefe	0,567053259	Schiefe	-0,310932636	Schiefe	-0,66103577	
11	Wertebereich	75	Wertebereich	71	Wertebereich	73	
12	Minimum	23	Minimum	20	Minimum	12	
13	Maximum	98	Maximum	91	Maximum	85	
14	Summe	444	Summe	466	Summe	455	
15	Anzahl	8	Anzahl	8	Anzahl	8	
16							

Abbildung 7.15: Das Ergebnis des Tools POPULATIONSKENNGRÖSSEN

Das Tool Populationskenngrößen zeigt Werte für die folgenden Kenngrößen an: Mittelwert, Standardfehler, Median, Modus, Standardabweichung, Stichprobenvarianz, Kurtosis, Schiefe, Wertebereich, Minimum, Maximum, Summe und Anzahl. Inzwischen kennen Sie schon alle mit Ausnahme von Standardfehler und Wertebereich.

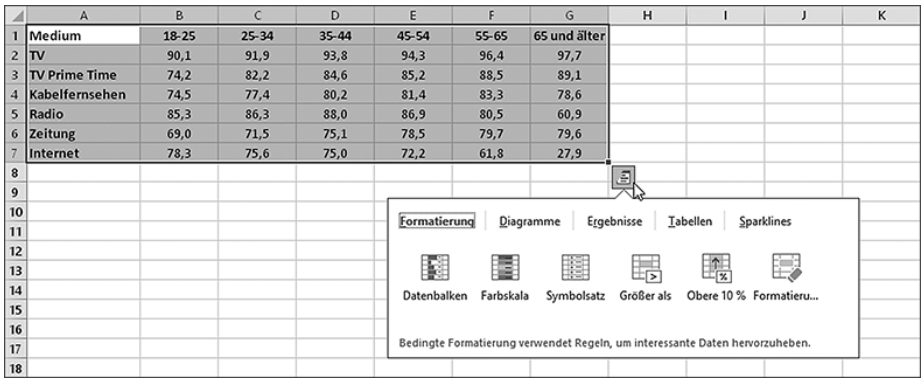
Der Wertebereich, auch als Spannweite bezeichnet, gibt lediglich die Differenz zwischen dem größten und dem kleinsten Wert an. Die Kenngröße Standardfehler ist etwas komplizierter und wird in Kapitel 9 erläutert. Hier nur so viel dazu, dass sich der Standardfehler aus der Standardabweichung dividiert durch die Quadratwurzel der Stichprobengröße berechnet. Das soll für den Moment genügen.

Eines der Kontrollkästchen, das im Beispiel in Schritt 6 nicht aktiviert wurde, gibt übrigens das so genannte Konfidenzniveau für den Mittelwert an. Auch diese Größe wird in Kapitel 9 beschrieben. Die beiden anderen Kontrollkästchen K-GRÖSSTER WERT und K-KLEINSTER WERT haben dieselbe Funktion wie die Funktionen `KGRÖSSTE` und `KKLEINSTE`.

Schnelle Analysen

Das Werkzeug Schnellanalyse ist eine fantastische Neuerung, die es seit Excel 2013 gibt, die jedoch leider auf dem Mac noch nicht zur Verfügung steht. Sie markieren einen Zellbereich mit Daten; danach wird unten rechts neben Ihrer Auswahl die Schaltfläche SCHNELLANALYSE angezeigt. Wenn Sie diese Schaltfläche anklicken (oder wenn Sie Strg+Q drücken), wird ein Menü mit zahlreichen Optionen zur Visualisierung und zur Zusammenfassung der Daten angezeigt. Wenn Sie mit der Maus auf die verschiedenen Optionen zeigen, sehen Sie im Tabellenblatt eine Vorschau dessen, was Sie durch Anklicken der jeweiligen Option erhalten.

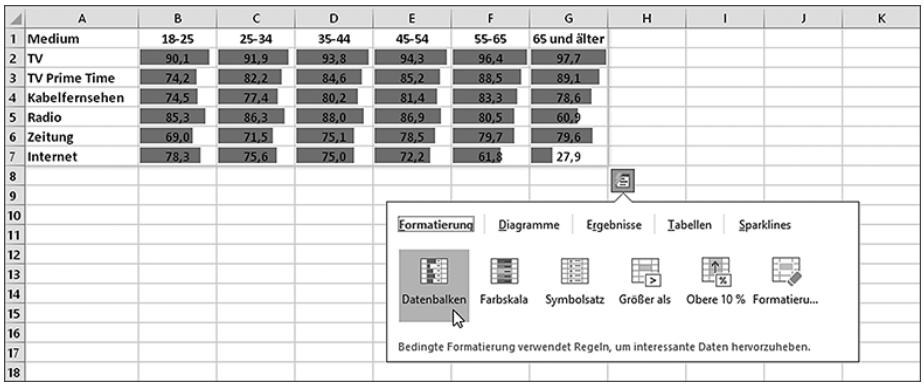
Das Tabellenblatt in Abbildung 7.16 zeigt für verschiedene Altersgruppen die Nutzung verschiedener Medien in Prozenten. (Die Daten stammen aus dem U.S. Statistical Abstracts.) Ich habe die Daten markiert, wodurch die Schaltfläche SCHNELLANALYSE eingeblendet wird. Anschließend habe ich diese Schaltfläche angeklickt, um das Menü mit den Optionen zu öffnen.



	A	B	C	D	E	F	G
1	Medium	18-25	25-34	35-44	45-54	55-65	65 und älter
2	TV	90,1	91,9	93,8	94,3	96,4	97,7
3	TV Prime Time	74,2	82,2	84,6	85,2	88,5	89,1
4	Kabelfernsehen	74,5	77,4	80,2	81,4	83,3	78,6
5	Radio	85,3	86,3	88,0	86,9	80,5	60,9
6	Zeitung	69,0	71,5	75,1	78,5	79,7	79,6
7	Internet	78,3	75,6	75,0	72,2	61,8	27,9

Abbildung 7.16: Markieren Sie die Daten, klicken Sie auf SCHNELLANALYSE und wählen Sie im Menü die gewünschte Option zur Visualisierung und Zusammenfassung der Daten aus.

Abbildung 7.17 zeigt, was passiert, wenn ich mit der Maus auf FORMATIERUNG | DATENBALKEN zeige.



	A	B	C	D	E	F	G
1	Medium	18-25	25-34	35-44	45-54	55-65	65 und älter
2	TV	90,1	91,9	93,8	94,3	96,4	97,7
3	TV Prime Time	74,2	82,2	84,6	85,2	88,5	89,1
4	Kabelfernsehen	74,5	77,4	80,2	81,4	83,3	78,6
5	Radio	85,3	86,3	88,0	86,9	80,5	60,9
6	Zeitung	69,0	71,5	75,1	78,5	79,7	79,6
7	Internet	78,3	75,6	75,0	72,2	61,8	27,9

Abbildung 7.17: Mit der Maus auf FORMATIERUNG | DATENBALKEN zeigen

Möchten Sie die Daten in einem Säulendiagramm wie dem folgenden darstellen? Zeigen Sie mit der Maus auf DIAGRAMME | GRUPPIERTE SÄULEN (siehe Abbildung 7.18).

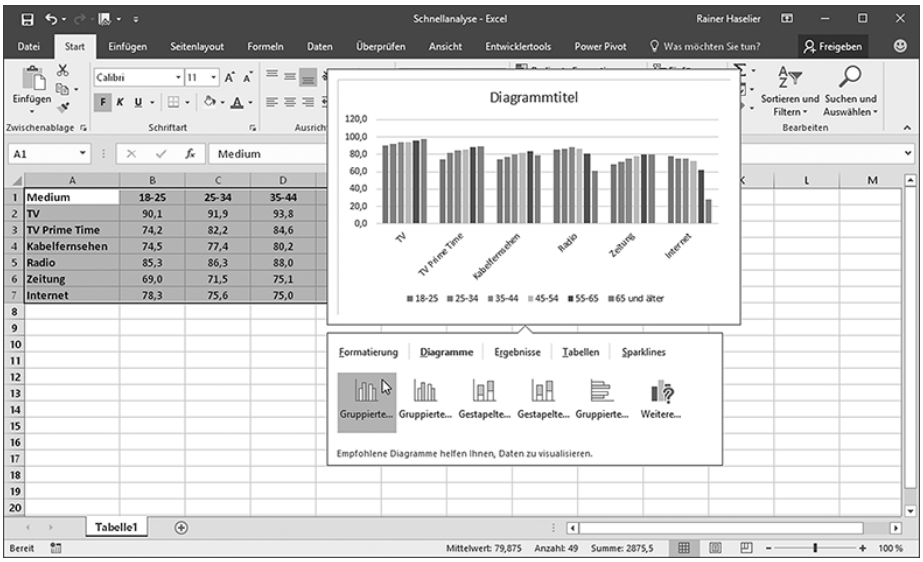
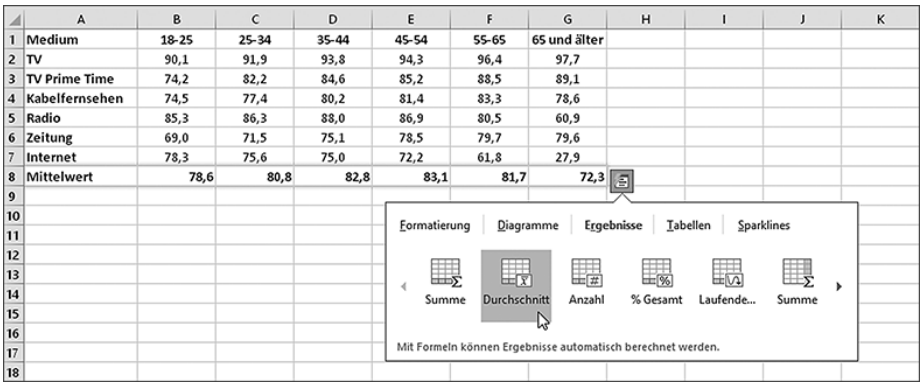


Abbildung 7.18: Mit der Maus auf DIAGRAMME | GRUPPIERTE SÄULEN zeigen

Wie fügen Sie die Mittelwerte ein? Klicken Sie auf ERGEBNISSE | DURCHSCHNITT, wie in Abbildung 7.19 gezeigt.



	A	B	C	D	E	F	G
1	Medium	18-25	25-34	35-44	45-54	55-65	65 und älter
2	TV	90,1	91,9	93,8	94,3	96,4	97,7
3	TV Prime Time	74,2	82,2	84,6	85,2	88,5	89,1
4	Kabelfernsehen	74,5	77,4	80,2	81,4	83,3	78,6
5	Radio	85,3	86,3	88,0	86,9	80,5	60,9
6	Zeitung	69,0	71,5	75,1	78,5	79,7	79,6
7	Internet	78,3	75,6	75,0	72,2	61,8	27,9
8	Mittelwert	78,6	80,8	82,8	83,1	81,7	72,3

Abbildung 7.19: Mit der Maus auf ERGEBNISSE | DURCHSCHNITT zeigen

Ich könnte den ganzen Tag so weitermachen, jedoch zeige ich Ihnen nur ein paar wenige, weitere Beispiele. Falls Sie den ausgewählten Zellbereich etwas professioneller aussehen lassen wollen, versuchen Sie mal TABELLEN | TABELLE (siehe Abbildung 7.20).

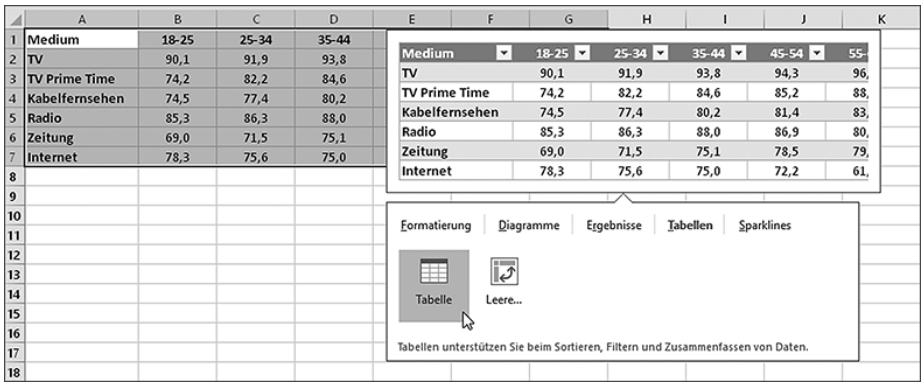


Abbildung 7.20: Mit der Maus auf TABELLEN | TABELLE zeigen

Ich kann dieses Thema nicht abschließen, ohne vorher noch einen Blick auf die Sparklines in der Schnellanalyse zu werfen. Wenn ich mit der Maus auf SPARKLINES | SÄULE zeige, erhalte ich die Ausgabe, wie in Abbildung 7.21 gezeigt. Diese Säulen machen wichtige, altersabhängige Trends deutlich: Vergleichen Sie die Sparklines für die Internetnutzung mit denen für TV und Zeitung.

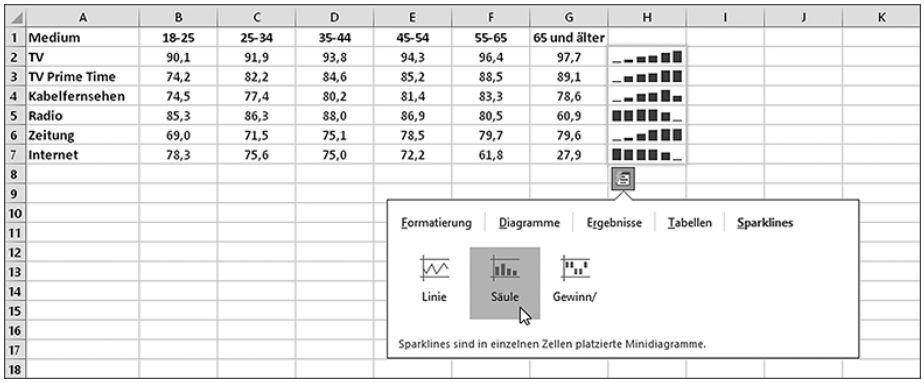


Abbildung 7.21: Mit der Maus auf SPARKLINES | SÄULE zeigen

Schnelle Kenngrößen

Angenommen, Sie arbeiten mit einem Zellbereich mit vielen Daten. Vielleicht wollen Sie schnell den Durchschnitt oder vielleicht andere statistische Kenngrößen ermitteln, ohne hierfür mehrere statistische Funktionen verwenden zu müssen. Sie können die Statusleiste unterhalb des Tabellenblatts konfigurieren, damit dort die gewünschten Kenngrößen angezeigt werden, nachdem Sie einen Zellbereich markiert haben. Um die Statusleiste zu konfigurieren,

klicken Sie diese mit der rechten Maustaste an, um das Menü STATUSLEISTE ANPASSEN zu öffnen (siehe Abbildung 7.22). Wenn Sie im dritten Bereich von unten alle Optionen einschalten, werden die betreffenden statistischen Kenngrößen (und die Anzahl der numerischen und nicht numerischen Zellen) in der Statusleiste angezeigt.

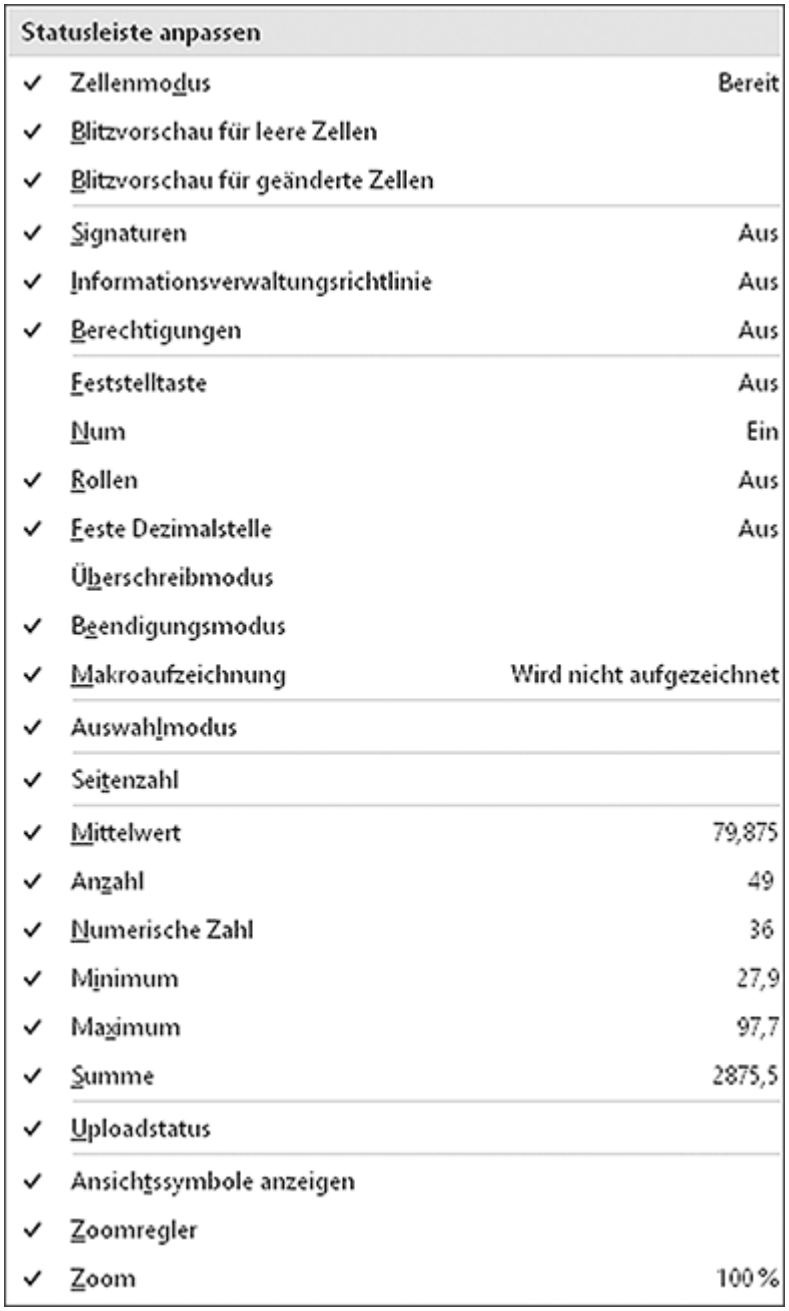


Abbildung 7.22: Das Menü STATUSLEISTE ANPASSEN

Abbildung 7.23 zeigt, wie die Kenngrößen der von mir markieren Zellen in der Statusleiste angezeigt werden.

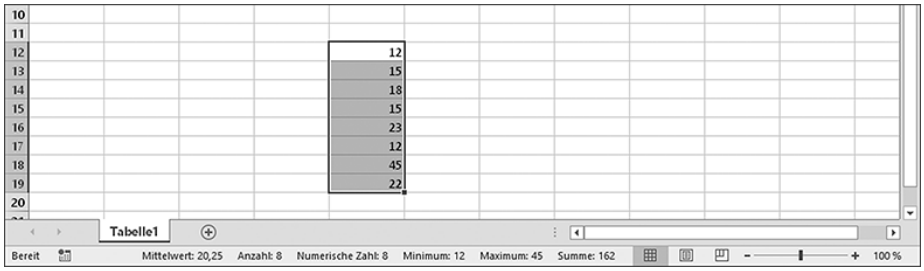


Abbildung 7.23: Kenngrößen in der Statusleiste

Kapitel 8

Was ist normal?

In diesem Kapitel

- Lernen Sie die Familie der Normalverteilung kennen
- Mit Standardabweichungen und Normalverteilungen arbeiten
- Excel-Funktionen für die Normalverteilung verstehen

Eine wichtige Aufgabe der Statistiker ist das Schätzen von Parametern von Grundgesamtheiten. Die Aufgabe wird etwas einfacher, wenn für die untersuchten Grundgesamtheiten einige Voraussetzungen gelten.

Eine Voraussetzung lässt sich immer wieder anwenden. Ein bestimmtes Merkmal, eine bestimmte Eigenschaft oder Fähigkeit tritt in einer Gesellschaft immer wieder auf. Die meisten Menschen entsprechen dabei etwa dem Mittelwert für dieses Attribut, wobei die Anzahl der Menschen mit immer extremeren Werten immer mehr abnimmt. In diesem Kapitel geht es um diese Voraussetzung und deren Bedeutung für die Statistik. Hier finden Sie außerdem Informationen zu den entsprechenden Excel-Funktionen.

So kratzen Sie die Kurve

Wenn Sie etwas wie die Länge oder das Gewicht messen, haben Sie es mit Objekten zu tun, die Sie sehen und berühren können. Statistiker, Sozialwissenschaftler, Marktforscher und Kaufleute messen dagegen häufig etwas, das sie weder sehen noch berühren können. Eigenschaften wie Intelligenz, Musikalität oder das Verlangen, ein neues Produkt zu kaufen, gehören in diese Kategorie.

Diese Art von Eigenschaften findet sich in der ganzen Gesellschaft. Die meisten Menschen entsprechen dabei etwa dem Durchschnitt, wobei die Anzahl der Menschen mit Extremwerten stetig abnimmt. Da dies so häufig vorkommt, wurde dies zu einer Annahme über die Verteilung der meisten Eigenschaften.

Diese »Die meisten Menschen entsprechen dem Durchschnitt«-Annahme lässt sich grafisch darstellen. In Abbildung 8.1 ist die bekannte *Glockenkurve* dargestellt, die die Verteilung einer Reihe von Merkmalen aufzeigt. Die Fläche unter der Kurve stellt die Grundgesamtheit dar. Die horizontale Achse repräsentiert die Werte für die jeweils untersuchte Fähigkeit. Die gedachte vertikale Linie in der Mitte der Kurve entspricht dem Durchschnitt der Größen.

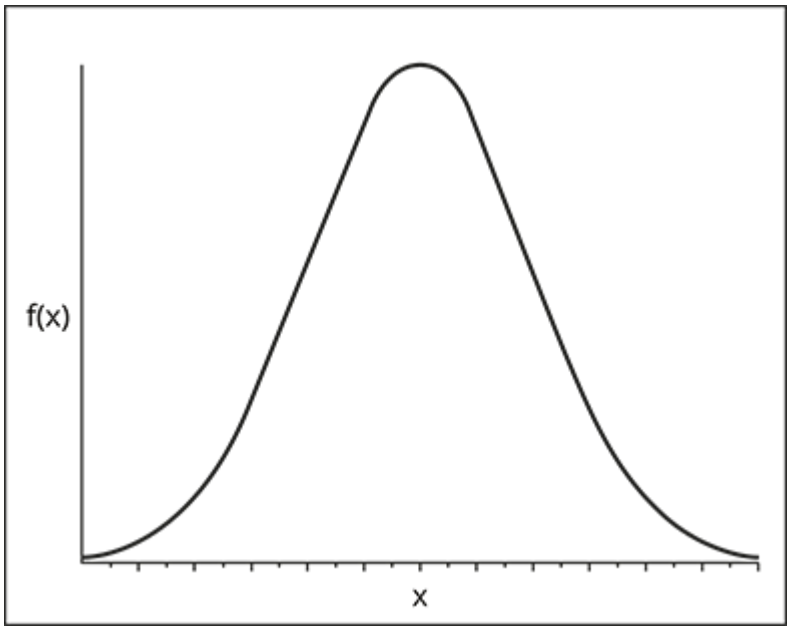


Abbildung 8.1: Die Glockenkurve

Wenn wir also davon ausgehen, dass es möglich ist, eine Eigenschaft wie Intelligenz zu messen, und wenn wir außerdem davon ausgehen, dass diese Kurve die Verteilung der Intelligenz in der Gesellschaft bzw. Grundgesamtheit darstellt, können wir Folgendes sagen: Die Glockenkurve zeigt, dass die meisten Menschen über eine durchschnittliche Intelligenz verfügen, dass nur wenige nicht so intelligent und ebenfalls nur wenige Genies sind.

Das passt doch perfekt zu unserer Vorstellung von Intelligenz, nicht wahr?

Tiefer graben

Die horizontale Achse in Abbildung 8.1 ist mit *x* und die vertikale Achse mit *f(x)* gekennzeichnet. Was bedeutet das? Die horizontale Achse repräsentiert, wie bereits erwähnt, die Werte, so dass Sie sich jeden Wert als ein *x* vorstellen können.

Was *f(x)* bedeutet, ist etwas schwieriger zu erklären. Eine mathematische Beziehung zwischen *x* und *f(x*) ergibt die Glockenkurve und ermöglicht uns deren Darstellung. Diese Beziehung ist ziemlich kompliziert und ich möchte Sie damit nicht belasten. Es ist völlig ausreichend, wenn Sie wissen, dass *f(x)* die Höhe der Kurve für einen bestimmten Wert von *x* darstellt. Sie geben einen Wert für *x* (und einige andere Dinge) an und aufgrund dieser komplizierten Beziehung ergibt sich der Wert von *f(x)*.

Hier noch einige Erläuterungen: Die Glockenkurve wird offiziell als *Normalverteilung* bezeichnet. Der Ausdruck *f(x)* wird als *Wahrscheinlichkeitsdichte* bezeichnet. Die Normalverteilung ist somit ein Beispiel für eine *Wahrscheinlichkeitsdichtefunktion*. Ich möchte Ihnen keine technische Definition der Wahrscheinlichkeitsdichte geben. Stellen Sie sich stattdessen die Wahrscheinlichkeitsdichte einfach als etwas vor, das aus der Fläche unter der Kurve eine Wahrscheinlichkeit macht. Wahrscheinlichkeit von … ja, wovon eigentlich? Das erfahren Sie im nächsten Abschnitt.

Parameter einer Normalverteilung

Häufig wird von *der* Normalverteilung gesprochen. Das ist eigentlich nicht richtig. Eigentlich handelt es sich um mehrere Verteilungen. Die einzelnen Verteilungen unterscheiden sich durch zwei Parameter voneinander. Ja, *Parameter*, weil es ja um Grundgesamtheiten geht. Diese beiden Parameter sind der Mittelwert (μ) und die Standardabweichung (σ). Der Mittelwert gibt das Zentrum der Verteilung an und die Standardabweichung gibt an, wie sich die Werte um den Mittelwert verteilen. Der Mittelwert ist die Mitte der

Verteilung. Jede Normalverteilung ist symmetrisch, die linke Seite der Verteilung ist das Spiegelbild der rechten.

Die Merkmale der Normalverteilung sind den Statistikern wohlbekannt. Was noch wichtiger ist: Sie können diese Merkmale bei Ihrer Arbeit verwenden.

Wie? Hier kommen wir zurück zur Wahrscheinlichkeit. Sie können einige hilfreiche Wahrscheinlichkeiten ermitteln, wenn Sie vier Dinge tun können:

- wenn Sie eine Linie ziehen können, die die Skala des gemessenen Merkmals repräsentiert
- wenn Sie auf der Linie angeben können, wo sich der Mittelwert der Werte befindet
- wenn Sie die Standardabweichung kennen
- wenn Sie wissen (oder davon ausgehen können), dass das Merkmal in der Grundgesamtheit normalverteilt ist

Nehmen wir IQ-Werte, damit Sie sehen, was ich meine. Die Werte im IQ-Test nach Stanford-Binet sind normalverteilt. Der Mittelwert der Verteilung dieser Werte beträgt 100 und die Standardabweichung 16. In Abbildung 8.2 ist diese Verteilung dargestellt.

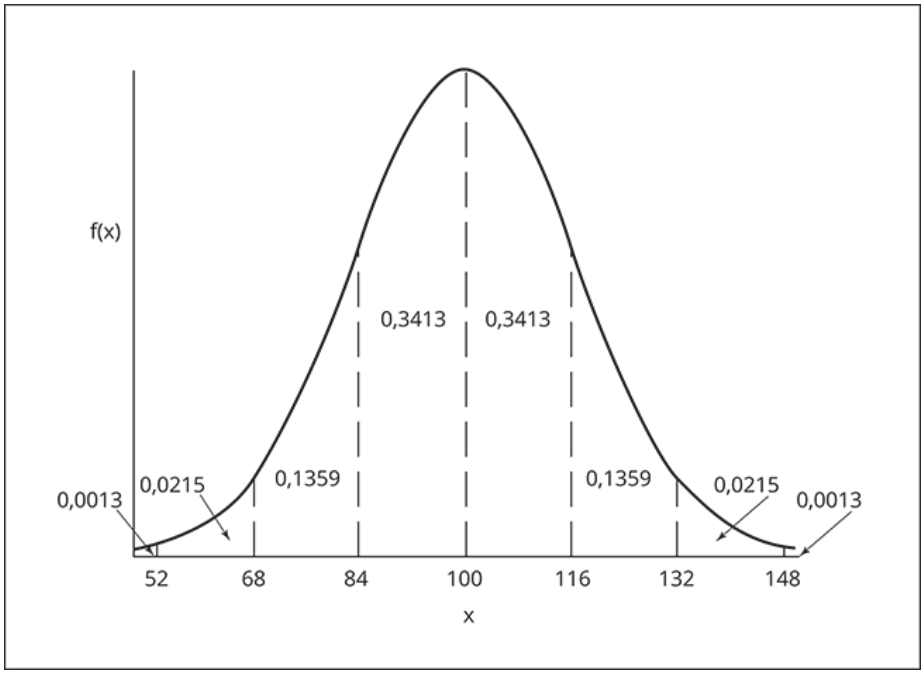


Abbildung 8.2: Die Normalverteilung des Intelligenzquotienten, unterteilt in Standardabweichungen

Wie Sie in der Abbildung sehen können, habe ich eine Linie für die IQ-Skala gezogen. Jeder Punkt auf der Linie repräsentiert einen IQ-Wert. Der Mittelwert 100 ist der Bezugspunkt. Davon ausgehend habe ich jeweils nach 16 Punkten (Standardabweichung) Markierungen eingefügt. Ich habe eine gestrichelte Linie vom Mittelpunkt aus bis *f*(100) (die Höhe der Normalverteilung ist *x* = 100) und eine gestrichelte Linie von jedem Standardabweichungspunkt aus gezogen.

Anhand der Abbildung ist außerdem der Flächenanteil zu erkennen, der von der Kurve und der x-Achse sowie von den aufeinanderfolgenden Standardabweichungen begrenzt wird. Darüber hinaus sind auch die Anteile außerhalb der dreifachen Standardabweichung auf beiden Seiten zu sehen (52 und 148). Beachten Sie, dass die Kurve niemals die horizontale Achse berührt. Sie nähert sich der Achse an, berührt sie jedoch nicht. (Mathematiker sagen dazu, dass die Kurve *asymptotisch* zur horizontalen Achse verläuft.)

Zwischen dem Mittelwert und einer Standardabweichung, zwischen 100 und 116, liegen somit 0,3413 (bzw. 34,13 %) der Werte in der

Grundgesamtheit. Oder anders ausgedrückt: Die Wahrscheinlichkeit, dass ein IQ-Wert zwischen 100 und 116 liegt, beträgt 0,3413. An den Enden der Verteilung liegen jeweils 0,0013 (0,13 %) der Werte auf jeder Seite.

Die in Abbildung 8.2 dargestellten Flächenanteile unter einer Normalverteilung gelten nicht nur für die IQ-Werte nach Stanford-Binet, sondern für alle Normalverteilungen. In Kapitel 6 habe ich beispielsweise die SAT-Werte erwähnt, die einen Mittelwert von 500 und eine Standardabweichung von 100 aufweisen. Diese Werte sind ebenfalls normalverteilt. Das bedeutet, dass 34,13 % der SAT-Werte zwischen 500 und 600 liegen, 34,13 % zwischen 400 und 500 und … Nun, Sie können die anderen Anteile anhand von Abbildung 8.2 ablesen.

NORM.VERT

In Abbildung 8.2 sind nur Flächen dargestellt, die nach Werten der Standardabweichungen aufgeteilt sind. Aber was ist mit dem Anteil der IQ-Werte zwischen 100 und 125? Oder zwischen 75 und 91? Oder größer als 118? Wenn Sie schon einmal an einem Statistikkurs teilgenommen haben, erinnern Sie sich vielleicht noch daran, dass es gar nicht so einfach war, die Flächenanteile unter der Normalverteilung zu ermitteln. Möglicherweise erinnern Sie sich auch noch daran, dass Sie die Anteile mit Tabellen der Normalverteilung ermittelt haben.

Mit der Excel-Tabellenblattfunktion `NORM.VERT` können Sie die Flächen unter der Normalverteilung ohne Tabellen ermitteln. `NORM.VERT` ermittelt eine *kumulierte Fläche*. Sie geben einen Wert, einen Mittelwert und eine Standardabweichung für eine Normalverteilung an und `NORM.VERT` ermittelt den Anteil der Fläche links des angegebenen Werts (auch als *kumulierter Anteil* oder *kumulierte Wahrscheinlichkeit* bezeichnet). Anhand von Abbildung

8.2 ist beispielsweise zu sehen, dass sich bei der IQ-Verteilung 0,8413 Anteile der Fläche links von 116 befinden.

Wie komme ich auf diesen Anteil? Die Summe aller Anteile links von 100 ergibt 0,5000. (Die Summe aller Anteile rechts von 100 ergibt ebenfalls 0,5000.) Wenn Sie diese 0,5000 mit den 0,3413 zwischen 100 und 116 addieren, ergibt sich 0,8413.

Dies lässt sich folgendermaßen umformulieren: Die Wahrscheinlichkeit für einen IQ-Wert kleiner als oder gleich 116 beträgt 0,8413.

In Abbildung 8.3 wurde dieser Anteil mit NORM.VERT ermittelt. Und so müssen Sie vorgehen:

1. **Markieren Sie eine Zelle für das Ergebnis von NORM.VERT.**

   In diesem Beispiel habe ich die Zelle C2 markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen NORM.VERT aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Parameter für die Funktion ein.**

   In das Feld X habe ich den Wert eingegeben, für den die kumulierte Fläche ermittelt werden soll. In diesem Beispiel ist das der Wert **116**.

   In das Feld MITTELWERT habe ich den Mittelwert der Verteilung und in das Feld STANDABWN die Standardabweichung eingegeben. Hier ist der Mittelwert **100** und die Standardabweichung **16**.

   In das Feld KUMULIERT habe ich **Wahr** eingegeben. Damit wird NORM.VERT mitgeteilt, dass die kumulierte Fläche ermittelt werden soll. Im Dialogfeld wird das Ergebnis angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

In Abbildung 8.3 ist zu sehen, dass die kumulierte Fläche 0,84134476 beträgt. Wenn Sie im Feld Kumuliert **Falsch** eingeben, ermittelt `NORM.VERT` die Höhe der Normalverteilung bei 116.

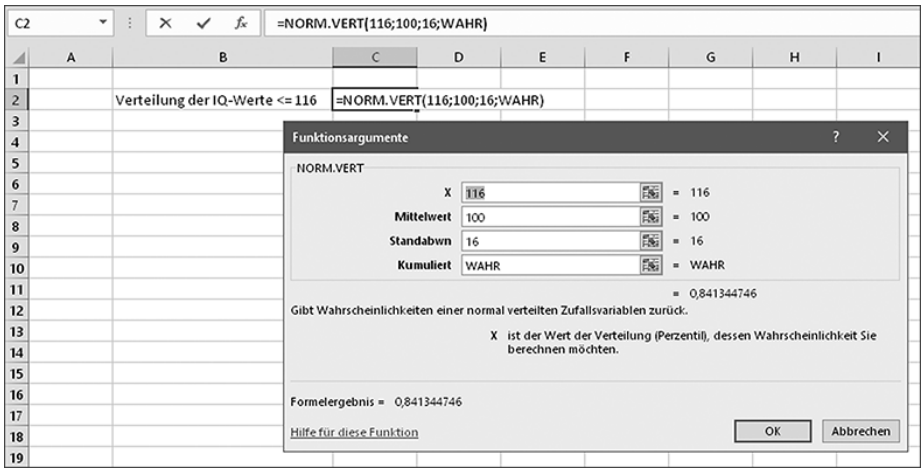


Abbildung 8.3: Die Funktion `NORM.VERT` verwenden

Um den Anteil der IQ-Werte größer als 116 zu ermitteln, brauchen Sie nur das Ergebnis von 1,0 zu subtrahieren. (Nur der Vollständigkeit halber: Das ist 0,15865524.)

Aber was ist mit dem Anteil der IQ-Werte zwischen 116 und 125? Wenden Sie die Funktion `NORM.VERT` auf alle Werte an, und subtrahieren Sie die Ergebnisse. Für dieses Beispiel lautet die Formel folgendermaßen:

```
=NORM.VERT(125;100;16;WAHR)-
NORM.VERT(116;100;16;WAHR)
```

Das Ergebnis lautet übrigens 0,09957.

NORM.INV

`NORM.INV` ist die Umkehrfunktion von `NORM.VERT`. Sie geben eine kumulierte Wahrscheinlichkeit, einen Mittelwert und eine Standardabweichung an und `NORM.INV` ermittelt den Wert, der die kumulierte Wahrscheinlichkeit umkehrt. Wenn Sie beispielsweise

0,5000 zusammen mit einem Mittelwert und einer Standardabweichung eingeben, gibt NORM.INV den Mittelwert zurück.

Diese Funktion ist hilfreich, wenn Sie den Wert für ein bestimmtes Quantil in einer Normalverteilung berechnen müssen. Wiederholen Sie die Schritte aus dem Abschnitt über NORM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE für NORM.INV zu öffnen (siehe Abbildung 8.4). Geben Sie **0,75** als kumulierte Wahrscheinlichkeit, **500** als Mittelwert und **100** als Standardabweichung ein. Da die SAT-Werte eine Normalverteilung mit dem Mittelwert 500 und der Standardabweichung 100 aufweisen, entspricht das Ergebnis für den SAT-Wert dem Wert beim 75 %-Quantil. (Ausführlichere Informationen zu Quantilen finden Sie in Kapitel 6.)

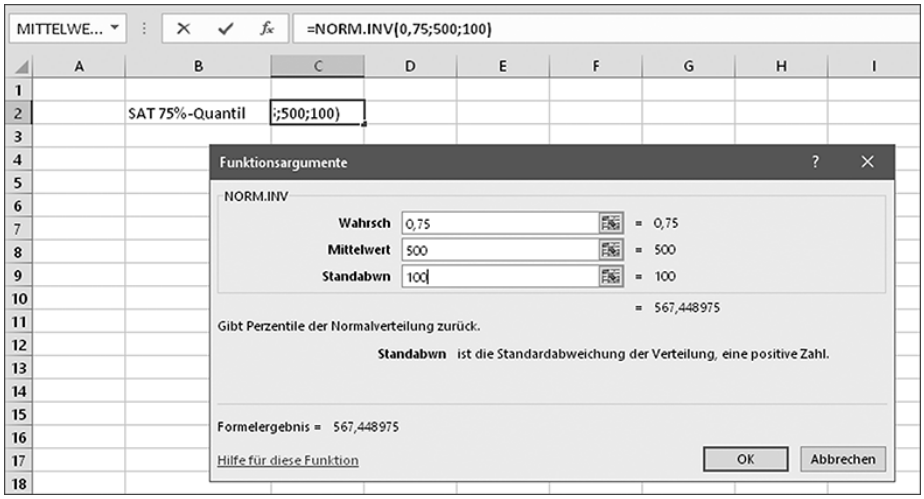


Abbildung 8.4: Die Funktion NORM.INV verwenden

Eine ganz besondere Verteilung

Um eine Gruppe von Werten zu standardisieren, damit sie mit einer anderen Gruppe von Werten verglichen werden kann, wird jeder Wert in einen *z-Wert* umgerechnet (siehe Kapitel 6). Die Formel zum Umrechnen eines Werts in einen z-Wert (auch als Standardwert bezeichnet) lautet:

$$z = \frac{x - \mu}{\sigma}$$

Dabei geht es darum, dass die Standardabweichung als Einheitsmaß verwendet wird. Bei der Stanford-Binet-Version des IQ-Tests beispielsweise beträgt der Mittelwert 100 und die Standardabweichung 16. Bei der Wechsler-Version beträgt der Mittelwert dagegen 100 und die Standardabweichung 15. Was bedeutet der Stanford-Binet-Wert 110 im Vergleich zum Wechsler-Wert 110?

Diese Frage lässt sich leicht beantworten: Bringen Sie die beiden Versionen auf die gleiche Ebene und standardisieren Sie beide Werte. Für den Stanford-Binet-Wert ergibt sich:

$$z = \frac{110 - 100}{16} = 0{,}625$$

Für den Wechsler-Wert ergibt sich:

$$z = \frac{110 - 100}{15} = 0{,}667$$

Der Wechsler-Wert 110 ist also ein etwas höherer Wert als der Stanford-Binet-Wert 110.

Wenn Sie nun alle Werte in einer Normalverteilung (wie zum Beispiel in eine Version der IQ-Werte) umrechnen, ergibt sich eine Normalverteilung von z-Werten. Jede Menge z-Werte (ob normalverteilt oder nicht) weist den Mittelwert 0 und die Standardabweichung 1 auf. Wenn eine Normalverteilung diese Parameter aufweist, handelt es sich um eine *Standardnormalverteilung*, also um eine Normalverteilung von Standardwerten.

Von dieser Art der Normalverteilung haben die meisten schon einmal gehört. Es ist die, die ihnen aus dem Statistikunterricht am besten in Erinnerung geblieben ist und dies ist die Verteilung, die fast jeder meint, wenn er (fälschlicherweise) von *der* Normalverteilung spricht. An diese Verteilung denken die meisten auch, wenn sie von *z-Werten* sprechen. Diese Verteilung verleitet viele zu der falschen Vorstellung, dass Werte durch Umrechnen in z-Werte eine Normalverteilung ergeben.

In Abbildung 8.5 ist die Standardnormalverteilung dargestellt. Sie sieht fast wie Abbildung 8.2 aus, nur dass der Mittelwert 0 beträgt und dass sich die Kennzeichnungen für die Einheiten der Standardabweichung an den entsprechenden Stellen befinden.

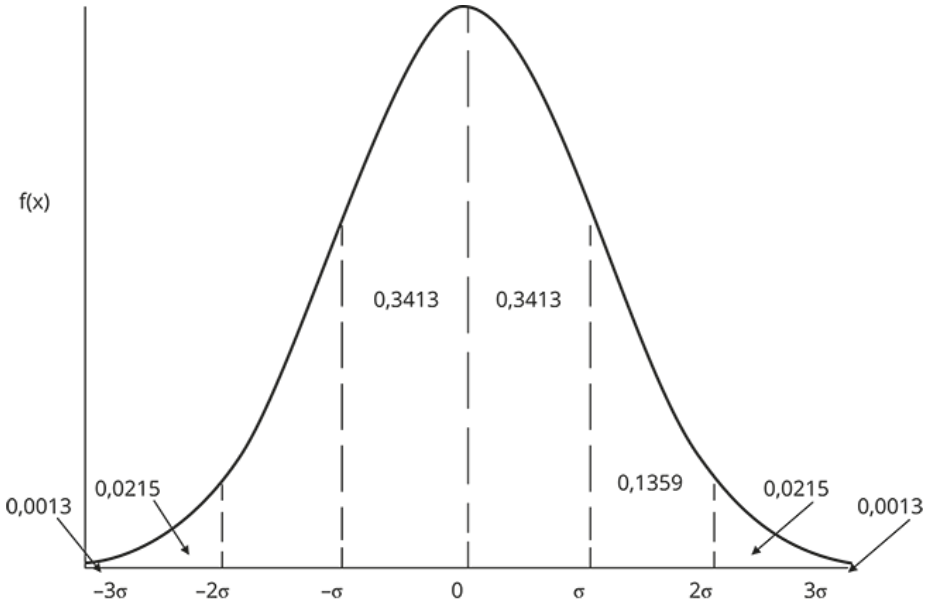


Abbildung 8.5: Die Standardnormalverteilung, aufgeteilt in Standardabweichungen

In den nächsten beiden Abschnitten werden die Excel-Funktionen für die Arbeit mit der Standardabweichung beschrieben.

NORM.S.VERT

Die Funktion NORM.S.VERT funktioniert ähnlich wie ihr Pendant NORM.VERT. Die beiden Funktionen unterscheiden sich lediglich dadurch, dass NORM.S.VERT für eine Normalverteilung mit dem Mittelwert 0 und der Standardabweichung 1 gedacht ist. Wiederholen Sie die Schritte aus dem Abschnitt über NORM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE für NORM.S.VERT zu öffnen. Wenn Sie einen z-Wert eingeben, wird die Fläche links vom z-Wert, also die Wahrscheinlichkeit, dass ein z-Wert kleiner als oder gleich dem angegebenen Wert ist, angezeigt.

In Abbildung 8.6 wurde im Dialogfeld FUNKTIONSARGUMENTE für den z-Wert 1 angegeben. Im Dialogfeld wird das Ergebnis 0,841344746 angezeigt. Dies ist die Wahrscheinlichkeit, dass ein z-Wert in einer Standardnormalverteilung kleiner als oder gleich 1,00 ist. Klicken Sie auf OK, damit das Ergebnis in der markierten Zelle angezeigt wird.

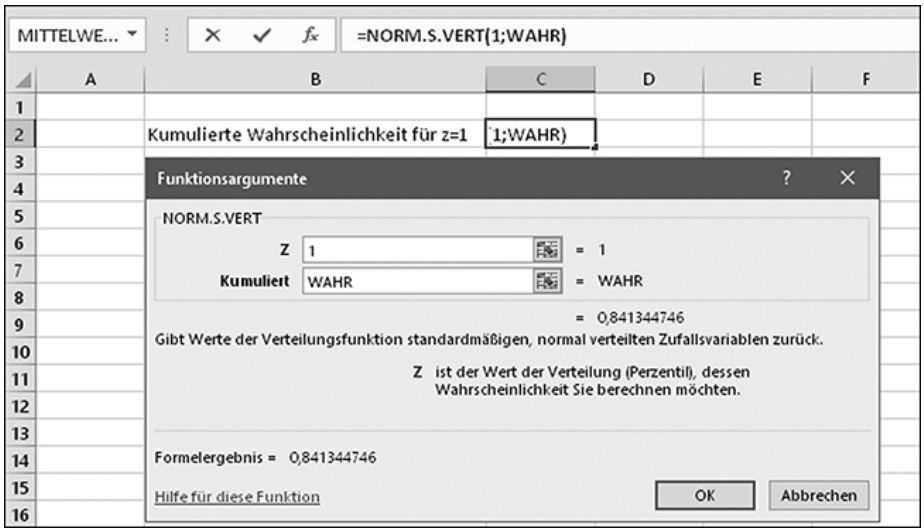


Abbildung 8.6: Die Funktion NORM.S.VERT verwenden

NORM.S.INV

NORM.S.INV ist die Umkehrfunktion von NORM.S.VERT. Wiederholen Sie die Schritte aus dem Abschnitt über NORM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE für NORM.S.INV zu öffnen. Sie geben eine kumulierte Wahrscheinlichkeit an und NORM.S.INV ermittelt den z-Wert, der die kumulierte Wahrscheinlichkeit umkehrt.

Wenn Sie beispielsweise 0,5000 angeben, zeigt NORM.S.INV 0 an, also den Mittelwert der Standardnormalverteilung.

In Abbildung 8.7 wurde im Dialogfeld FUNKTIONSARGUMENTE für NORM.S.INV für die kumulierte Wahrscheinlichkeit 0,75 angegeben. Im Dialogfeld wird das Ergebnis 0,67448975, der z-Wert des 75 %-Quantils der Standardnormalverteilung, angezeigt.

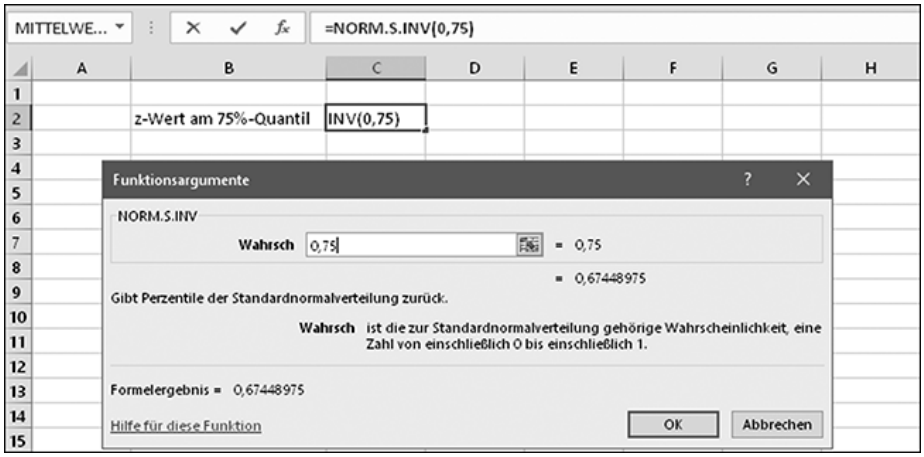


Abbildung 8.7: Die Funktion NORM.S.INV verwenden

PHI und GAUSS

Diese beiden Tabellenblattfunktionen arbeiten mit der Standardnormalverteilung. Beiden Funktionen wird jeweils ein Argument entgegengestellt.

```
=PHI(x)
```

gibt die Höhe (das ist der Wert der Dichte) der Standardnormalverteilung an *x* zurück.

```
=GAUSS(x)
```

gibt einen Wert zurück, der um 0,5 kleiner ist als die kumulierte Wahrscheinlichkeit von *x* (in der Standardnormalverteilung). Diese Funktion ist oft nützlich, wenn Sie schnell die kumulierte Wahrscheinlichkeit an nur einer Seite der Verteilung ermitteln, Sie hierfür nicht NORM.S.VERT verwenden, dann alle Argumente und vom Ergebnis 0,5 subtrahieren wollen.

Eine Standardnormalverteilung als Diagramm darstellen

Meiner Erfahrung nach hilft es, eine Verteilung in einem Diagramm darzustellen, um die Verteilung zu verstehen. Daher stelle ich in diesem Abschnitt die Schritte vor, mit denen Sie eine Standardnormalverteilung visualisieren können.

Die Beziehung zwischen *x* und *f(x)* für die Normalverteilung ist, wie ich bereits erwähnt habe, ziemlich kompliziert. Die Gleichung dafür sieht so aus:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Wenn Sie Werte für μ (Mittelwert), σ (Standardabweichung) und x (ein Wert) eingeben, wird mit der Gleichung ein Wert für *f(x)*, die Höhe der Normalverteilung bei *x*, ermittelt. π und e sind wichtige Konstanten in der Mathematik. π beträgt etwa 3,1416 (Verhältnis von Kreisumfang und Durchmesser) und e etwa 2,71828. Das hat mit dem so genannten *natürlichen Logarithmus* und einer Reihe anderer mathematischer Konzepte zu tun. (Über Logarithmen und e erfahren Sie mehr in Kapitel 22, und zwar im Abschnitt »Logarithmische Esoteriken«.)

Bei einer Standardnormalverteilung ist μ = 0 und σ = 1, so dass die Formel wie folgt lautet:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Ich habe aus dem *x* ein *z* gemacht, weil wir es bei dieser Normalverteilung mit z-Werten zu tun haben.

Ich habe dann in Schritten von 0,2 in Zellen von Spalte B z-Werte im Bereich von –4 und 4 eingetragen.

Anschließend habe ich in Zelle C2 die Gleichung für die Standardnormalverteilung als Excel-Formel eingegeben:

```
=((1/WURZEL(2*PI())))*EXP(-(B2^2)/2)
```

`PI` ist eine Excel-Funktion, die den Wert π angibt. Die Funktion `EXP` potenziert die Zahl e mit dem Wert, der in der nachfolgenden Klammer enthalten ist. (Auch hierzu finden Sie weitere Informationen in Kapitel 22.)

Dies ist der harte, der mathematische Weg, um die Aufgabe zu lösen. Die Tabellenblattfunktion `PHI`, die ich weiter vorne erwähnt habe, ist ein einfacherer Weg, um die *f(z)*-Werte zu erhalten. Wenn Sie in Zelle C2 diese Formel eingeben:

```
=PHI(B2)
```

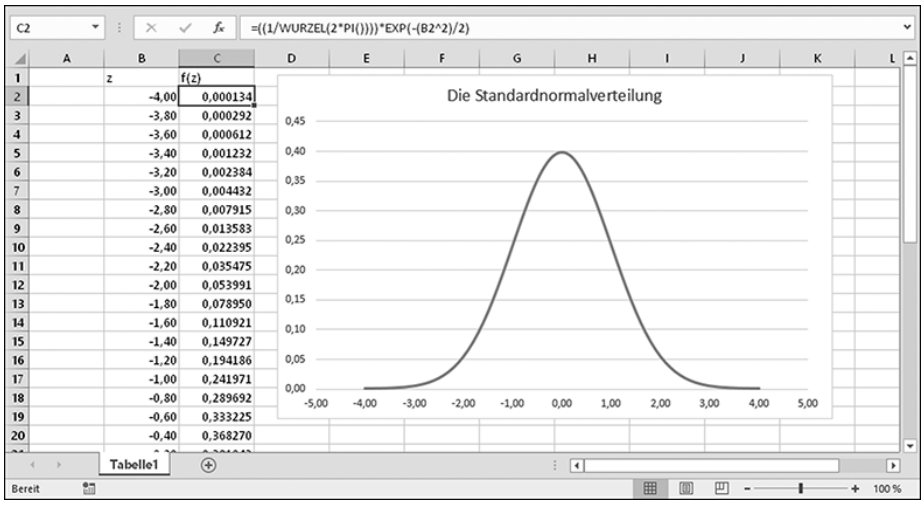
erhalten Sie dasselbe Ergebnis wie mit der komplizierten Formel.

Der nächste Schritt besteht darin, Spalte C automatisch auszufüllen.

Wählen Sie dann den Befehl EINFÜGEN | EMPFOHLENE DIAGRAMME | ALLE DIAGRAMME und wählen Sie den Diagrammtyp PUNKTE MIT INTERPOLIERTEN LINIEN. Sie müssen ein paar Anpassungen vornehmen, damit das Diagramm so aussieht wie das in Abbildung 8.8. Die Abbildung stellt die Standardnormalverteilung für die Werte `z=-4` bis `z=4` gut dar. Abbildung 8.8 zeigt ebenfalls die automatisch ausgefüllten Werte.

Beachten Sie, dass in der Bearbeitungsleiste von Excel die Gleichung für die Standardnormalverteilung angezeigt wird `[=((1/WURZEL(2*PI())))*EXP(-(B2^2)/2)]`, die ich weiter oben bereits erwähnt habe.

Ich beschreibe dies alles so ausführlich, damit Sie sich daran gewöhnen, wichtige Verteilungen grafisch darzustellen. Ich werde dies im weiteren Verlauf des Buches ebenfalls tun.



	A	B	C
1		z	f(z)
2		-4,00	0,000134
3		-3,80	0,000292
4		-3,60	0,000612
5		-3,40	0,001232
6		-3,20	0,002384
7		-3,00	0,004432
8		-2,80	0,007915
9		-2,60	0,013583
10		-2,40	0,022395
11		-2,20	0,035475
12		-2,00	0,053991
13		-1,80	0,078950
14		-1,60	0,110921
15		-1,40	0,149727
16		-1,20	0,194186
17		-1,00	0,241971
18		-0,80	0,289692
19		-0,60	0,333225
20		-0,40	0,368270

Abbildung 8.8: Das fertige Diagramm für die Standardnormalverteilung

Teil III

Rückschlüsse aus Daten ziehen

In diesem Teil …

Stichprobenverteilungen erstellen

Vertrauensgrenzen ermitteln

Mit t-Tests arbeiten

Die Varianzanalyse verwenden

Visualisieren von t-, Chi-Quadrat- und F-Verteilungen

Korrelation und Regression verstehen

Mit Zeitreihen arbeiten

Parameterfreie Statistiken verstehen

Kapitel 9

Die Sache mit dem Vertrauen: Schätzung

In diesem Kapitel

- Einführung in die Stichprobenverteilung
- Den Standardfehler verstehen
- Die Stichprobenverteilung des Mittelwerts simulieren
- Schätzungen mit Vertrauensgrenzen verknüpfen

Grundgesamtheiten und Stichproben sind recht einfache Konzepte. Eine Grundgesamtheit ist eine große Menge Einzelelemente, aus denen eine Stichprobe gezogen wird. Bewerten Sie die Objekte der Stichprobe nach Eigenschaft oder Merkmal, berechnen Sie statistische Werte, mit denen Stichproben zusammengefasst werden, und schon sind Sie mitten drin.

Mit den statistischen Werten können Sie jedoch nicht nur die Werte in der Stichprobe zusammenfassen, sondern auch die Parameter der Grundgesamtheit schätzen. Das ist keine Kleinigkeit. Aus einem kleinen Anteil an Objekten aus der Grundgesamtheit können Sie auf die Grundgesamtheit schließen.

Dabei taucht jedoch eine Frage auf: Wie sehr können Sie dieser Schätzung trauen? Um diese Frage zu beantworten, müssen Sie die Schätzung in einen Kontext bringen. Wie wahrscheinlich ist sie? Wie wahrscheinlich ist es, dass sich der wahre Wert eines Parameters zwischen einer unteren und einer oberen Grenze befindet?

In diesem Kapitel lernen Sie den Kontext für Schätzungen kennen. Sie werden erfahren, wie dies in das Vertrauen in diese Schätzungen hineinspielt. Außerdem werden Sie eine Excel-Funktion kennenlernen, mit der Sie das Konfidenzniveau berechnen können.

Stichprobenverteilungen verstehen

Stellen Sie sich eine Grundgesamtheit vor, aus der Sie eine Stichprobe ziehen. Sie messen ein bestimmtes Merkmal an den einzelnen Stichprobenobjekten und berechnen den Mittelwert der Stichprobe. Sie geben die Stichprobenobjekte an die Grundgesamtheit zurück. Sie ziehen eine weitere Stichprobe, bewerten die Objekte der neuen Stichprobe und berechnen *deren* Mittelwert. Sie wiederholen diesen Vorgang immer wieder und verwenden dabei immer genau so viele Objekte wie bei der ersten Stichprobe. Wenn Sie diesen Vorgang unendlich oft wiederholen könnten (jeweils mit einer Stichprobe derselben Größe), könnten Sie eine unendliche Menge Stichprobenmittelwerte ermitteln. Diese Stichprobenmittelwerte bilden eine eigene Verteilung. Diese Verteilung wird als *Stichprobenverteilung des Mittelwerts* bezeichnet.

Für einen Stichprobenmittelwert ist dies der Kontext, den ich zu Beginn dieses Kapitels erwähnt habe. Wie jede andere Zahl auch hat ein statistischer Wert für sich genommen keine Bedeutung. Um die Bedeutung eines Werts zu verstehen, müssen Sie wissen, wie der Wert zustande kommt. Ein statistischer Wert kommt natürlich durch eine mit Stichprobendaten durchgeführte Berechnung zustande. Oder anders ausgedrückt: Ein statistischer Wert ist Teil einer Stichprobenverteilung.

Allgemein lässt sich sagen: *Eine Stichprobenverteilung ist die Verteilung aller möglichen Werte einer Statistik für einen*

gegebenen Stichprobenumfang.

Ich habe die Definition hervorgehoben, weil sie ganz besonders wichtig ist. Nach vielen Jahren Erfahrung im Unterrichten von Statistik kann ich Ihnen sagen, dass sich an diesem Konzept festmachen lässt, wer Statistik versteht und wer nicht.

Wenn Sie also verstehen, was eine Stichprobenverteilung ist, verstehen Sie auch, worum es in der Statistik geht. Wenn nicht, dann eben nicht. So einfach ist das.

Wenn Sie nicht verstehen, was eine Stichprobenverteilung ist, wird die Statistik für Sie wie ein Kochbuch funktionieren. Sie werden, wann immer Sie mit Statistik zu tun haben, Zahlen in Formeln einsetzen und hoffen, dass alles gut geht. Wenn Sie jedoch verstehen, was eine Stichprobenverteilung ist, werden Sie auch verstehen, worum es bei der Inferenzstatistik geht.

Anhand von Abbildung 9.1 wird das Konzept einer Stichprobenverteilung deutlicher. Hier sind die Schritte zum Erstellen einer Stichprobenverteilung des Mittelwerts zusammengefasst.

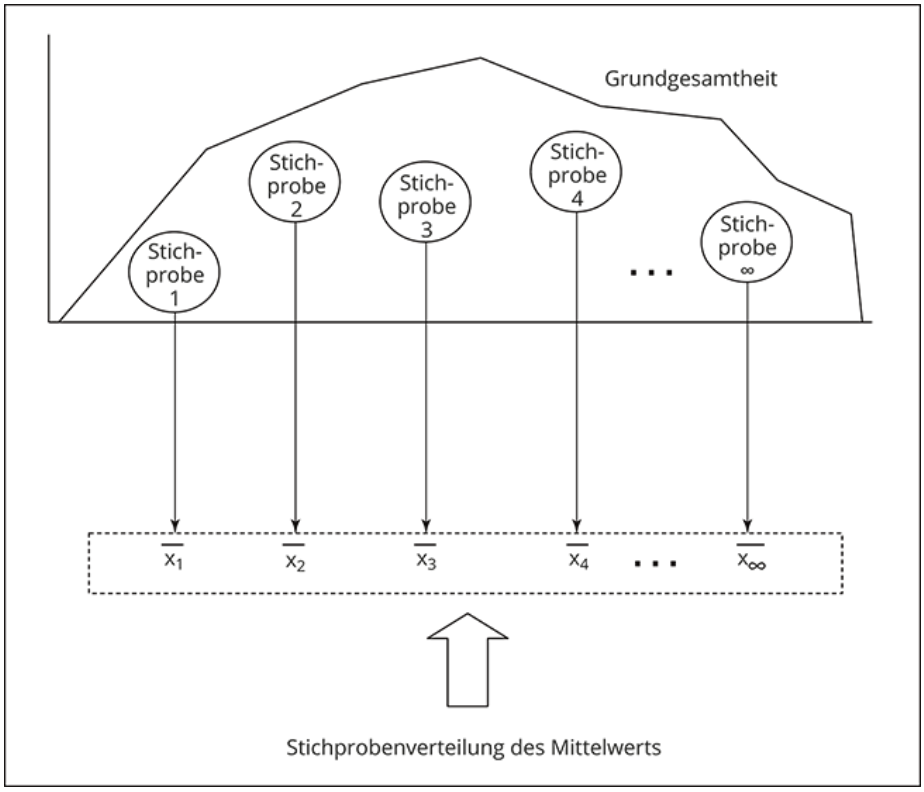


Abbildung 9.1: Erstellen einer Stichprobenverteilung des Mittelwerts

Eine Stichprobenverteilung hat wie jede andere Gruppe von Werten einen Mittelwert und eine Standardabweichung. Das Symbol für den Mittelwert der Stichprobenverteilung des Mittelwerts (hey, ist das ein Bandwurm!) ist $\mu_{\bar{x}}$.

Die Standardabweichung einer Stichprobenverteilung ist ein ziemlich heißes Eisen. Sie hat einen eigenen Namen: *Standardfehler*. Bei der Stichprobenverteilung des Mittelwerts wird die Standardabweichung als *Standardfehler des Mittelwerts* bezeichnet. Als Symbol wird $\sigma_{\bar{x}}$ verwendet.

Ein BESONDERS wichtiges Konzept: der zentrale Grenzwertsatz

Ich habe Sie gebeten, sich eine Situation vorzustellen, die es im echten Leben nicht gibt. Sie werden niemals die Mittelwerte aus unendlich vielen Stichproben berechnen und Sie werden niemals eine Stichprobenverteilung des Mittelwerts erstellen. Normalerweise ziehen Sie eine Stichprobe und berechnen die statistischen Werte.

Wenn Sie also nur eine Stichprobe haben, wie können Sie dann jemals etwas über eine Stichprobenverteilung, also über eine theoretische Verteilung, wissen, die unendlich viele Stichproben umfasst? Ist das alles ein fruchtloses Unterfangen?

Nein, das ist es nicht. Wegen eines besonderen Geschenks der Mathematik an die Statistik können Sie über eine Stichprobenverteilung eine ganze Menge herausfinden. Dieses Geschenk wird als *zentraler Grenzwertsatz* bezeichnet.

Nach dem zentralen Grenzwertsatz

- ist die Stichprobenverteilung des Mittelwerts in etwa normalverteilt, wenn der Stichprobenumfang groß genug ist.

  Groß genug bedeutet, etwa 30 oder mehr Stichproben.

- entspricht der Mittelwert der Stichprobenverteilung des Mittelwerts dem Mittelwert der Grundgesamtheit.

  Als Gleichung sieht das folgendermaßen aus:

$$\mu_{\bar{x}} = \mu$$

- entspricht die Standardabweichung der Stichprobenverteilung des Mittelwerts (auch als Standardfehler des Mittelwerts bezeichnet) der Standardabweichung der Grundgesamtheit dividiert durch die Quadratwurzel des Stichprobenumfangs.

  Die Gleichung hierfür lautet wie folgt:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Der zentrale Grenzwertsatz sagt übrigens nichts über die Grundgesamtheit aus. Er besagt lediglich, dass bei einem ausreichend großen Stichprobenumfang die Stichprobenverteilung des Mittelwerts eine Normalverteilung mit den angegebenen Parametern ist. Die Grundgesamtheit, die die Stichproben liefert, muss laut zentralem Grenzwertsatz nicht normalverteilt sein.

Aber was gilt, wenn die Grundgesamtheit normalverteilt ist? In diesem Fall ist die Stichprobenverteilung des Mittelwerts unabhängig vom Stichprobenumfang eine Normalverteilung.

In Abbildung 9.2 ist die Stichprobenverteilung des Mittelwerts unterteilt in Standardfehlereinheiten allgemein dargestellt.

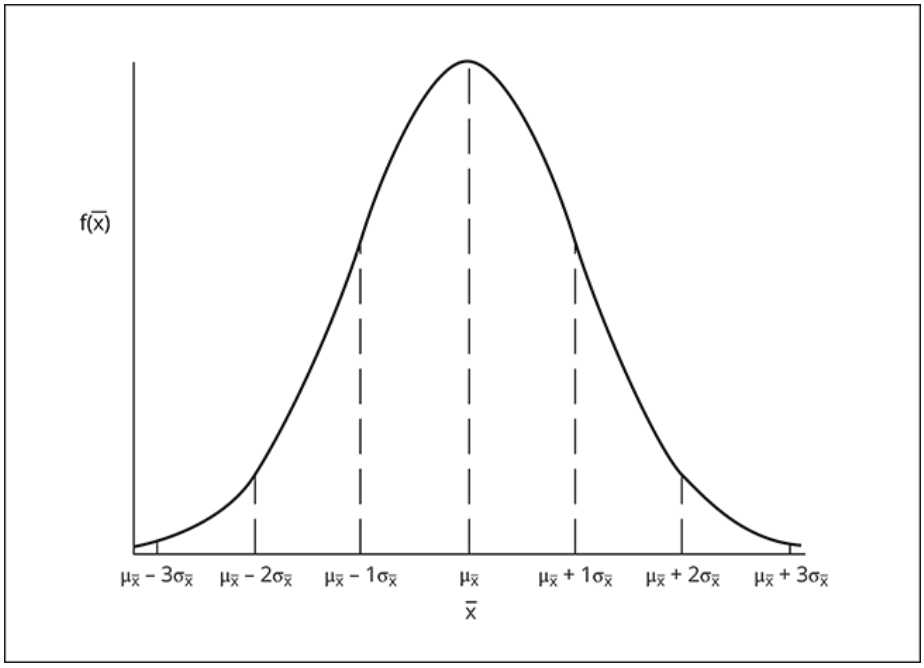


Abbildung 9.2: Die Stichprobenverteilung des Mittelwerts, in Standardfehlereinheiten unterteilt

(Näherungsweise) den zentralen Grenzwertsatz simulieren

Das ist kaum zu glauben: Wie kann sich aus einer Grundgesamtheit, die nicht normalverteilt ist, eine normalverteilte Stichprobenverteilung ergeben?

Damit Sie sich besser vorstellen können, wie der zentrale Grenzwertsatz funktioniert, habe ich eine Simulation erstellt. Mit dieser Simulation wird so etwas wie eine Stichprobenverteilung des Mittelwerts für eine sehr kleine Stichprobe auf der Grundlage einer nicht normalverteilten Grundgesamtheit erstellt. Wie Sie sehen, erinnert die Stichprobenverteilung des Mittelwerts an eine Normalverteilung, obwohl die Grundgesamtheit nicht normalverteilt ist und obwohl die Stichprobe klein ist.

Stellen Sie sich eine riesige Grundgesamtheit vor, die nur aus den drei Werten 1, 2 und 3 besteht, und dass jeder dieser Werte gleich

häufig in einer Stichprobe auftreten kann. (Diese Art der Grundgesamtheit ist definitiv *nicht* normalverteilt.) Stellen Sie sich außerdem vor, dass Sie nach dem Zufallsprinzip eine Stichprobe mit drei Werten aus dieser Grundgesamtheit ziehen können. In Tabelle 9.1 sind alle möglichen Stichproben und die zugehörigen Mittelwerte aufgeführt.

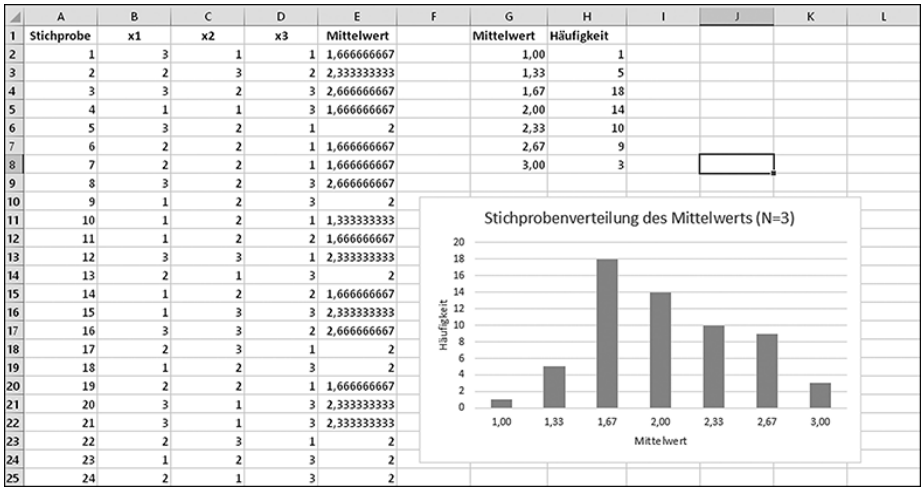
Stichprobe	Mittelwert	Stichprobe	Mittelwert	Stichprobe	Mittelwert
1,1,1	1,00	2,1,1	1,33	3,1,1	1,67
1,1,2	1,33	2,1,2	1,67	3,1,2	2,00
1,1,3	1,67	2,1,3	2,00	3,1,3	2,33
1,2,1	1,33	2,2,1	1,67	3,2,1	2,00
1,2,2	1,67	2,2,2	2,00	3,2,2	2,33
1,2,3	2,00	2,2,3	2,33	3,2,3	2,67
1,3,1	1,67	2,3,1	2,00	3,3,1	2,33
1,3,2	2,00	2,3,2	2,33	3,3,2	2,67
1,3,3	2,33	2,3,3	2,67	3,3,3	3,00

Tabelle 9.1: Alle mit drei Werten möglichen Stichproben (und die zugehörigen Mittelwerte) aus einer Grundgesamtheit, die aus den drei Werten 1, 2 und 3 besteht

Wenn Sie sich die Tabelle genau ansehen, können Sie fast schon erkennen, was in der Simulation geschehen wird. Der Stichprobenmittelwert, der am häufigsten vorkommt, ist der Wert 2,00. Die Stichprobenmittelwerte, die am wenigsten oft vorkommen, sind die Werte 1,00 und 3,00. Hmmm …

In der Simulation wähle ich nach dem Zufallsprinzip zunächst einen Wert aus der Grundgesamtheit und dann noch zwei weitere. Diese Gruppe der drei Werte ist eine Stichprobe. Anschließend berechne ich den Mittelwert dieser Stichprobe. Ich wiederhole diesen Vorgang für insgesamt 60 Stichproben und berechne somit 60 Stichprobenmittelwerte. Schließlich stelle ich die Verteilung der Stichprobenmittelwerte in einem Diagramm dar.

Wie sieht die simulierte Stichprobenverteilung des Mittelwerts aus? In Abbildung 9.3 sehen Sie ein Arbeitsblatt, das diese Frage beantwortet.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Stichprobe	x1	x2	x3	Mittelwert		Mittelwert	Häufigkeit				
2	1	3	1	1	1,666666667		1,00	1				
3	2	2	3	2	2,333333333		1,33	5				
4	3	3	2	3	2,666666667		1,67	18				
5	4	1	1	3	1,666666667		2,00	14				
6	5	3	2	1	2		2,33	10				
7	6	2	2	1	1,666666667		2,67	9				
8	7	2	2	1	1,666666667		3,00	3				
9	8	3	2	3	2,666666667							
10	9	1	2	3	2							
11	10	1	2	1	1,333333333							
12	11	1	2	2	1,666666667							
13	12	3	3	1	2,333333333							
14	13	2	1	3	2							
15	14	1	2	2	1,666666667							
16	15	1	3	3	2,333333333							
17	16	3	3	2	2,666666667							
18	17	2	3	1	2							
19	18	1	2	3	2							
20	19	2	2	1	1,666666667							
21	20	3	1	3	2,333333333							
22	21	3	1	3	2,333333333							
23	22	2	3	1	2							
24	23	1	2	3	2							
25	24	2	1	3	2							

Abbildung 9.3: Simulieren der Stichprobenverteilung des Mittelwerts (n=3) aus einer Grundgesamtheit, die aus den Werten 1, 2 und 3 besteht. Die Simulation setzt sich aus 60 Stichproben zusammen.

Im Arbeitsblatt stellt jede Zeile eine Stichprobe dar. In den Spalten mit den Überschriften *x1*, *x2* und *x3* befinden sich die drei Werte für die Stichproben. In Spalte G ist jeweils der Mittelwert für die Stichprobe der Zeile enthalten. In Spalte I befinden sich die möglichen Werte für den Stichprobenmittelwert und in Spalte J ist angegeben, wie häufig die einzelnen Mittelwerte in den 60 Stichproben auftreten. Anhand der Spalten I und J und der Grafik wird deutlich, dass am häufigsten der Stichprobenmittelwert 2,00 auftritt. Die Häufigkeit nimmt ab, je weiter die Stichprobenmittelwerte vom Wert 2,00 entfernt sind.

Bei all dem geht es darum, dass die Grundgesamtheit bei weitem nicht normalverteilt und der Stichprobenumfang sehr klein ist. Aber sogar mit diesen Einschränkungen gleicht die Stichprobenverteilung des Mittelwerts bei 60 Stichproben einer Normalverteilung.

Wie sieht es mit den Parametern aus, die der zentrale Grenzwertsatz für die Stichprobenverteilung prognostiziert? Beginnen wir mit der Grundgesamtheit. Der Mittelwert der Grundgesamtheit beträgt 2,00 und die Standardabweichung der

Grundgesamtheit 0,67. (Zum Ermitteln der Parameter für diese Art von Grundgesamtheit sind etwas kompliziertere mathematische Berechnungen erforderlich. Da diese Berechnungen hier ein wenig zu weit führen würden, möchte ich es hierbei belassen.)

Fahren wir mit der Stichprobenverteilung fort. Der Mittelwert der 60 Stichproben beträgt 1,97 und die Standardabweichung (eine Schätzung des Standardfehlers des Mittelwerts) 0,51. Diese Zahlen entsprechen fast den durch den zentralen Grenzwertsatz prognostizierten Parametern für die Stichprobenverteilung des Mittelwerts, 2,00 (gleich dem Mittelwert der Grundgesamtheit), und 0,47 (die Standardabweichung, 0,67, dividiert durch die Quadratwurzel von 3, dem Stichprobenumfang).

Für den Fall, dass Sie die in Abbildung 9.3 dargestellte Simulation nachstellen möchten, gehen Sie wie folgt vor:

1. **Wählen Sie eine Zelle für die erste zufällig ausgewählte Zahl aus.**

   Ich habe die Zelle B2 ausgewählt.

2. **Wählen Sie mit der Tabellenblattfunktion** `ZUFALLSBEREICH` **1, 2 oder 3 aus.**

   Damit wird die Auswahl einer Zahl aus einer Grundgesamtheit bestehend aus den Zahlen 1, 2 und 3 simuliert, bei der die Wahrscheinlichkeit für die Auswahl für alle Zahlen gleich groß ist. Sie können entweder den Befehl FORMELN | MATHEMATIK UND TRIGONOMETRIE | ZUFALLSBEREICH und dann das Dialogfeld FUNKTIONSARGUMENTE verwenden oder einfach

   ```
   =ZUFALLSBEREICH(1;3)
   ```

   in B2 eingeben und die Enter-Taste drücken. Das erste Argument ist die kleinste Zahl, die `ZUFALLSBEREICH` zurückgibt, und das zweite Argument ist die größte Zahl.

3. **Markieren Sie die Zelle rechts neben der ersten Zelle und geben Sie eine weitere zufällig gewählte Zahl zwischen 1 und 3 ein. Und geben Sie schließlich eine dritte zufällig gewählte Zahl in die Zelle rechts neben der zweiten ein.**

   Am einfachsten füllen Sie die beiden Zellen rechts neben der ersten Zelle automatisch aus. In meinem Arbeitsblatt sind das die Zellen C2 und D2.

4. **Betrachten Sie diese drei Zellen als Stichprobe und berechnen Sie den jeweiligen Mittelwert in der Zelle rechts neben der dritten Zelle.**

   Geben Sie hierzu in Zelle E2 einfach

   ```
   =MITTELWERT(B2:D2)
   ```

   ein und drücken Sie die Enter-Taste.

5. **Wiederholen Sie diesen Vorgang für alle Stichproben, die in der Simulation berücksichtigt werden sollen. Dabei entspricht jede Zeile einer Stichprobe.**

Ich habe 60 Stichproben verwendet. Am schnellsten und einfachsten geht das Ganze, wenn Sie die erste Zeile mit drei zufällig ausgewählten Zahlen und den dazugehörigen Mittelwert markieren und dann die restlichen Zeilen automatisch ausfüllen. Die Stichprobenmittelwerte in Spalte E stellen die simulierte Stichprobenverteilung des Mittelwerts dar. Ermitteln Sie mit den Funktionen `MITTELWERT` und `STABW.N` den jeweiligen Mittelwert und die Standardabweichung.

Damit Sie sehen, wie diese simulierte Stichprobenverteilung aussieht, wenden Sie die Arrayfunktion `HÄUFIGKEIT` auf die Stichprobenmittelwerte in Spalte E an. Gehen Sie dazu wie folgt vor:

1. **Geben Sie die möglichen Werte des Stichprobenmittelwerts in einen Bereich ein.**

Ich habe hierfür Spalte G verwendet. Die möglichen Werte des Stichprobenmittels habe ich als Brüche (3/3, 4/3, 5/3, 6/3, 7/3, 8/3 und 9/3) in die Zellen G2 bis G8 eingegeben. Excel wandelt die Brüche in Dezimalzahlen um.

2. **Markieren Sie einen Bereich für die Häufigkeiten der möglichen Werte des Stichprobenmittelwerts.**

   Ich habe für die Häufigkeiten die Spalte H ausgewählt, indem ich die Zellen H2 bis H8 markiert habe.

3. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion HÄUFIGKEIT aus, um das Dialogfeld FUNKTIONSARGUMENTE für die Funktion HÄUFIGKEIT zu öffnen.**

4. **Geben Sie den im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion HÄUFIGKEIT ein.**

   Geben Sie im Feld DATEN die Zellen an, in denen sich die Stichprobenmittelwerte befinden. In diesem Beispiel ist das E2:E61.

5. **Geben Sie den Zellbereich an, in dem sich die möglichen Werte des Stichprobenmittelwerts befinden.**

   Bei der Funktion HÄUFIGKEIT wird dieser Bereich im Feld KLASSEN angegeben. In meinem Arbeitsblatt steht im Feld KLASSEN G2:G8. Wenn Sie beide Felder angegeben haben, werden im Dialogfeld FUNKTIONSARGUMENTE die Häufigkeiten in geschweiften Klammern angezeigt (siehe Abbildung 9.4).

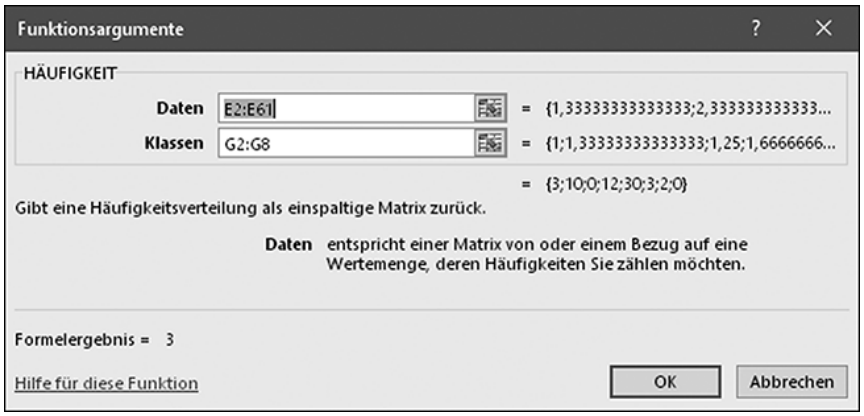


Abbildung 9.4: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion HÄUFIGKEIT im Arbeitsblatt »Simulierte Stichprobenverteilung«

6. **Drücken Sie die Tastenkombination Strg+Shift+Enter, um das Dialogfeld FUNKTIONSARGUMENTE zu schließen und die Häufigkeiten anzuzeigen.**

   Verwenden Sie diese Tastenkombination, da es sich bei HÄUFIGKEIT um eine Arrayfunktion handelt. (Ausführlichere Informationen zur Funktion HÄUFIGKEIT finden Sie in Kapitel 7.)

Lassen Sie den Zellbereich H2:H8 ausgewählt, verwenden Sie den Befehl EINFÜGEN | EMPFOHLENE DIAGRAMME und wählen Sie dann den Diagrammtyp GRUPPIERTE SÄUMEN aus, um das Diagramm mit den Häufigkeiten zu erstellen. Ich habe das Diagramm überarbeitet, damit es so aussieht wie in der Abbildung gezeigt. (Weitere Informationen zu Diagrammen finden Sie in Kapitel 3.) Vermutlich sieht Ihr Diagramm anders aus als meines, da Sie mit großer Wahrscheinlichkeit mit anderen Zufallszahlen arbeiten als ich.

Übrigens wiederholt Excel die Zufallsauswahl immer dann, wenn Sie etwas tun, das Excel veranlasst, das Arbeitsblatt neu zu berechnen. So ändern sich die Zahlen ständig, während Sie arbeiten. (Hierdurch wird die Simulation erneut ausgeführt.) Beispielsweise ändern sich die Zahlen, wenn Sie eine der Zeilen erneut automatisch ausfüllen lassen, und folglich ändert sich auch das Diagramm.

Die Vertrauensgrenzen

Ich habe Ihnen von Stichprobenverteilungen erzählt, weil sie Ihnen bei der Beantwortung der Frage helfen, die ich zu Beginn dieses Kapitels gestellt habe. Wie gut können Sie Ihrer eigenen Schätzung trauen?

Es geht darum, einen statistischen Wert zu berechnen und mit ihm obere und untere Grenzen für die Parameter der Grundgesamtheit mit, sagen wir, einer statistischen Sicherheit von 95 % zu erstellen. Das können Sie nur tun, wenn Sie die Stichprobenverteilung des statistischen Werts und den Standardfehler kennen. Im nächsten Abschnitt erfahren Sie, wie Sie dabei beim Mittelwert vorgehen.

So suchen Sie die Vertrauensgrenzen für einen Mittelwert

Das Unternehmen FarBlonJet ist ein Hersteller von Navigationssystemen. (Das Firmenmotto lautet: »Wenn Sie reisen, brauchen Sie FarBlonJet.«) Die Firma hat für sein tragbares Modell eine neue Batterie entwickelt. Um das System besser zu verkaufen, möchte FarBlonJet wissen, wie lange eine solche Batterie im Durchschnitt hält.

Das Unternehmen möchte für den Durchschnittswert eine Schätzung mit einer Sicherheit von 95 %. Dazu wird eine Stichprobe mit 100 Batterien getestet und festgestellt, dass der Stichprobenmittelwert bei einer Standardabweichung von 20 Stunden 60 Stunden beträgt. Der zentrale Grenzwertsatz besagt, dass sich bei einer ausreichend großen Stichprobe (mindestens 30) die Stichprobenverteilung des Mittelwerts einer Normalverteilung nähert. Der Standardfehler des Mittelwerts (die Standardabweichung der Stichprobenverteilung des Mittelwerts) berechnet sich aus

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Der Stichprobenumfang *n* beträgt 100. Und was ist mit σ? Dieser Wert ist nicht bekannt, er muss also geschätzt werden. Wenn σ bekannt wäre, wäre auch μ bekannt, und es müssten keine Vertrauensgrenzen festgelegt werden.

Die beste Schätzung für σ ist die Standardabweichung. In diesem Beispiel ist das der Wert 20. Damit ergibt sich für die Schätzung des Standardfehlers des Mittelwerts Folgendes:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

Die beste Schätzung des Mittelwerts der Grundgesamtheit ist der Stichprobenmittelwert 60. Mit diesen Informationen (geschätzter Mittelwert, geschätzter Standardfehler des Mittelwerts, Normalverteilung) können Sie die Stichprobenverteilung des Mittelwerts ermitteln, was ich in Abbildung 9.5 getan habe. In Übereinstimmung mit Abbildung 9.2 ist jede Standardabweichung ein Standardfehler des Mittelwerts.

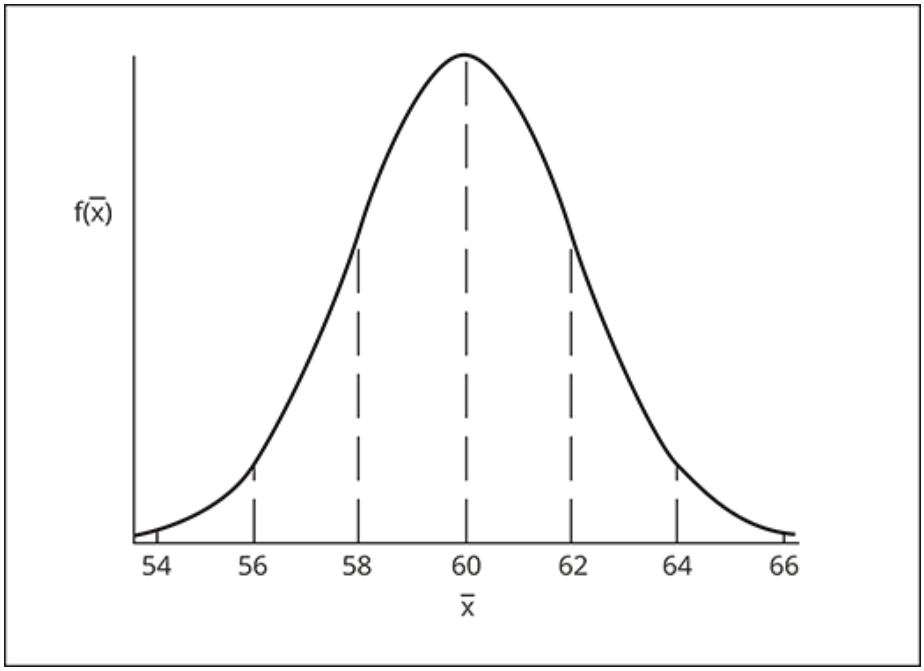


Abbildung 9.5: Die Stichprobenverteilung des Mittelwerts für die FarBlonJet-Batterie

Wenn Sie die Stichprobenverteilung ermittelt haben, können Sie die 95%igen Vertrauensgrenzen für den Mittelwert festlegen. Wie weit müssen Sie also von der Mitte der Verteilung auf beiden Seiten nach außen gehen, bis Sie 95 % der Fläche unter der Kurve erfasst haben? (Weitere Informationen zur Fläche unter der Normalverteilungskurve und deren Bedeutung finden Sie in Kapitel 8.)

Diese Frage lässt sich mit der Standardnormalverteilung beantworten, indem der z-Wert ermittelt wird, der 47,5 % auf der rechten und 47,5 % auf der linken Seite abtrennt. (Ja, wieder Kapitel 8.) Beim Wert auf der rechten Seite handelt es sich um einen positiven und bei dem auf der linken um einen negativen z-Wert. Multiplizieren Sie anschließend die z-Werte mit dem Standardfehler. Addieren Sie die Ergebnisse zum Stichprobenmittelwert, um so die obere und die untere Vertrauensgrenze zu ermitteln.

Dabei stellt sich heraus, dass der z-Wert für die Grenze auf der rechten Seite der Standardnormalverteilung 1,96 und für die Grenze auf der linken Seite −1,96 beträgt. Sie können diese Werte berechnen (schwierig), sie einer Tabelle der Standardnormalverteilung entnehmen, die sich in jedem Statistikbuch befindet (einfacher), oder die Tabellenfunktion von Excel verwenden, die ich im nächsten Abschnitt beschreibe und die die ganze Rechenarbeit übernimmt (viel einfacher). Es geht darum, dass die obere Grenze in der Stichprobenverteilung bei 63,92 (= 60 + 1,96 $s_{\bar{X}}$) und die untere Grenze bei 56,08 (= 60 – 1,96 $s_{\bar{X}}$) liegt. In Abbildung 9.6 sind diese Grenzen auf der Stichprobenverteilungskurve dargestellt.

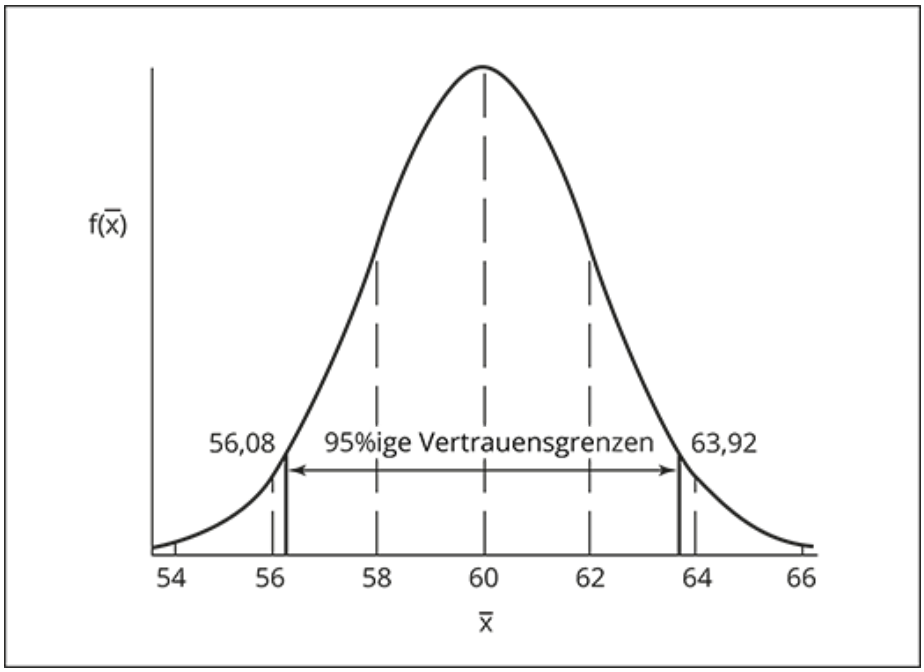


Abbildung 9.6: Die 95%igen Vertrauensgrenzen auf der FarBlonJet-Stichprobenverteilungskurve

Das bedeutet, dass Sie mit einer Sicherheit von 95 % sagen können, dass die FarBlonJet-Batterie im Durchschnitt eine Lebensdauer zwischen 56,08 Stunden und 63,92 Stunden aufweist. Sie möchten einen engeren Bereich? Dann können Sie entweder das Konfidenzniveau senken (auf beispielsweise 90 %) oder eine größere Stichprobe Batterien testen.

KONFIDENZ.NORM

Die Tabellenblattfunktion `KONFIDENZ.NORM` nimmt Ihnen den Großteil der Arbeit beim Festlegen von Konfidenzintervallen ab. Sie geben das Konfidenzniveau, die Standardabweichung und den Stichprobenumfang an und `KONFIDENZ.NORM` gibt das Ergebnis aus der Multiplikation des entsprechenden z-Werts mit dem Standardfehler des Mittelwerts zurück. Um die obere Grenze des Vertrauensbereichs festzulegen, addieren Sie das Ergebnis zum Stichprobenmittelwert. Um die untere Grenze zu ermitteln, subtrahieren Sie das Ergebnis vom Stichprobenmittelwert.

Betrachten wir noch einmal das Beispiel mit den FarBlonJet-Batterien, damit Sie sehen, wie das Ganze funktioniert. Und so

müssen Sie vorgehen:

1. **Wählen Sie eine Zelle aus.**

2. **Wählen Sie im Menü mit den statistischen Funktionen KONFIDENZ.NORM aus, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 9.7).**

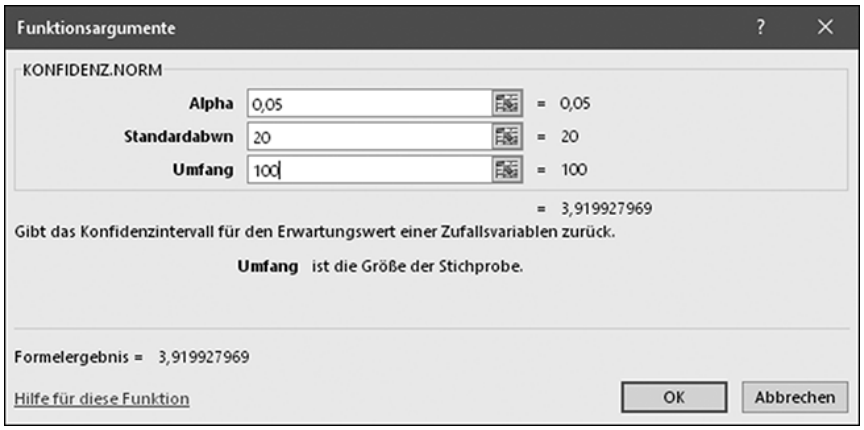


Abbildung 9.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `KONFIDENZ.NORM`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld ALPHA wird das Ergebnis aus der Subtraktion des Konfidenzniveaus von 1,00 eingegeben.

   Ja, das ist etwas verwirrend. Statt 0,95 für die 95%ige Vertrauensgrenze muss ich 0,05 eingeben. Stellen Sie sich einfach vor, Sie geben nicht den Anteil der Fläche *innerhalb* der Vertrauensgrenzen, sondern den Anteil der Fläche *außerhalb* der Vertrauensgrenzen ein. Und warum heißt dieses Feld *Alpha*? Das werde ich in Kapitel 10 erläutern.

   In das Feld STANDABWN gebe ich die Standardabweichung der Stichprobe ein. In diesem Beispiel beträgt die Standardabweichung 20.

Das Feld UMFANG ist für die Anzahl der Elemente in der Stichprobe vorgesehen. In diesem Beispiel wurden 100 Batterien getestet. Nachdem ich diese Zahl eingegeben habe, wird das Ergebnis (3,919928) im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, damit das Ergebnis in die markierte Zelle eingefügt wird.**

Zum guten Schluss addiere ich das Ergebnis zum Stichprobenmittelwert (60), um die obere Vertrauensgrenze (63,92) zu ermitteln, und subtrahiere das Ergebnis vom Mittelwert, um die untere Vertrauensgrenze (56,08) zu ermitteln.

Passend für ein t

Der zentrale Grenzwertsatz beschreibt (näherungsweise) eine Normalverteilung für große Stichproben. Häufig können Sie sich jedoch den Luxus großer Stichproben nicht leisten und die Normalverteilung passt nicht. Was also tun?

Bei kleinen Stichproben ist die Stichprobenverteilung des Mittelwerts eine so genannte *t-Verteilung*. Der Parameter, der dafür sorgt, dass sich diese Verteilung von den anderen Verteilungen unterscheidet, wird als *Freiheitsgrade* bezeichnet.

Sie können sich Freiheitsgrade als Bezeichnung für die Varianzschätzung vorstellen. Wenn eine Stichprobe beispielsweise aus 25 Objekten besteht, berechnet sich die Stichprobenvarianz, mit der die Varianz der Grundgesamtheit geschätzt wird, wie folgt:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})^2}{25-1} = \frac{\sum(x-\bar{x})^2}{24}$$

Die Zahl im Nenner ist 24, und das ist der Wert der Freiheitsgrade. Allgemein sind die Freiheitsgrade (df) = *n* – 1

(*n* ist der Stichprobenumfang), wenn die *t*-Verteilung verwendet wird, wie ich das tun werde.

In Abbildung 9.8 sind zwei t-Verteilungen (df = 3 und df = 10) sowie die Normalverteilung zum Vergleich dargestellt. Wie die Abbildung zeigt, nähert sich *t* umso mehr einer Normalverteilung, je größer df ist.

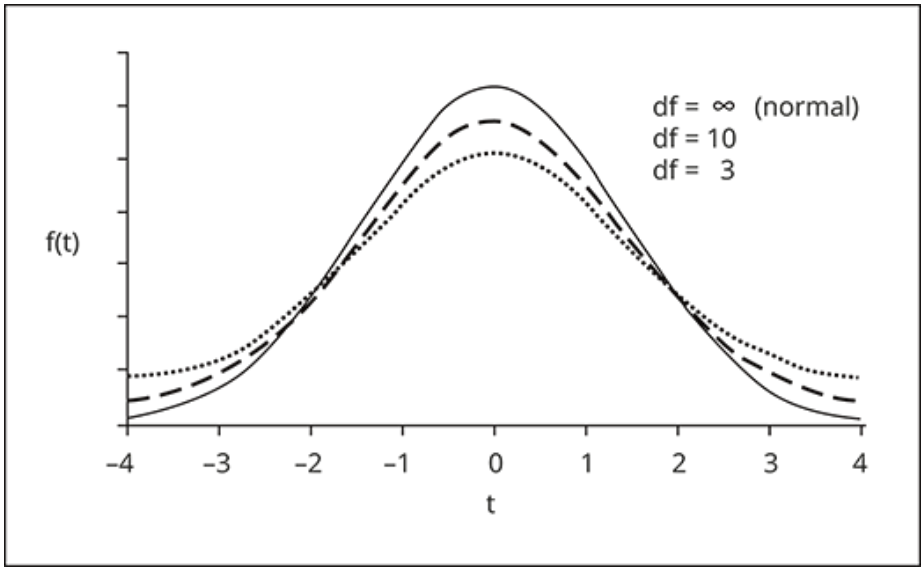


Abbildung 9.8: Einige t-Verteilungskurven

Um also das 95%ige Konfidenzniveau bei einer kleinen Stichprobe zu ermitteln, verwenden Sie die *t*-Verteilung mit geeignetem df-Wert. Suchen Sie den Wert, der 2,5 % der Fläche von der rechten Seite und 2,5 % der Fläche von der linken Seite der Verteilung abtrennt. Beim Wert auf der rechten Seite handelt es sich um einen positiven Wert und bei dem auf der linken um einen negativen. Multiplizieren Sie anschließend die Werte mit dem Standardfehler. Addieren Sie die Ergebnisse zum Mittelwert, um so die obere und die untere Vertrauensgrenze zu ermitteln.

Nehmen wir an, die Stichprobe beim Beispiel mit den FarBlonJet-Batterien besteht aus 25 Batterien mit einem Mittelwert von 60 und einer Standardabweichung von 20. Der Schätzwert für den Standardfehler des Mittelwerts beträgt

$$s_{\bar{X}} = \frac{s}{\sqrt{N}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

df = *n* − 1 = 24. Der Wert, der 2,5 % der Fläche von rechts von dieser Verteilung abschneidet, beträgt 2,064 und der, der 2,5 % der Fläche von links von dieser Verteilung abschneidet, beträgt −2,064. Wie bereits erwähnt, können Sie diese Werte berechnen (schwierig), in einer Tabelle in einem Statistikübungsbuch nachschlagen (einfacher) oder die Excel-Funktion verwenden, die im nächsten Abschnitt beschrieben wird (viel einfacher).

Es geht darum, dass die obere Vertrauensgrenze 68,256 (= 60 + 2,064 $s_{\bar{X}}$) und die untere Vertrauensgrenze 51,744 (= 60 − 2,064 $s_{\bar{X}}$) beträgt. Bei einer Stichprobe mit 25 Batterien können Sie mit einer 95%igen Sicherheit sagen, dass die durchschnittliche Lebensdauer einer FarBlonJet-Batterie zwischen 51,744 und 68,256 Stunden beträgt. Beachten Sie, dass bei einer kleineren Stichprobe der Bereich für das im vorhergehenden Beispiel verwendete Konfidenzniveau größer ist.

KONFIDENZ.T

Die Excel-Tabellenblattfunktion `KONFIDENZ.T` arbeitet so wie die Funktion `KONFIDENZ`, nur dass sie die *t*-Verteilung verwendet und nicht die Standardnormalverteilung. Verwenden Sie diese Funktion, wenn Ihre Daten die Anforderungen für die Standardnormalverteilung nicht erfüllen. Das Dialogfeld FUNKTIONSARGUMENTE von `KONFIDENZ.T` ist genauso aufgebaut wie das von `KONFIDENZ` und Sie verwenden es, indem Sie die gleichen Schritte durchführen.

Für das zweite FarBlonJet-Beispiel (Mittelwert = 60, Standardabweichung = 20 und Stichprobengröße = 25) gibt `KONFIDENZ.T` den Wert 8,256 zurück. Ich addiere dieses Ergebnis zum Stichprobenmittelwert (60), um die obere Vertrauensgrenze (68,256) zu ermitteln, und subtrahiere das Ergebnis von 60, um die untere Vertrauensgrenze (51,744) zu ermitteln.

Kapitel 10

Ein-Stichproben-Hypothesentest

In diesem Kapitel

- Einführung in Hypothesentests
- Hypothesen zu Mittelwerten testen
- Hypothesen zu Varianzen testen
- Verteilungen visualisieren

Unabhängig von Ihrer Beschäftigung müssen Sie häufig einschätzen, ob etwas außer der Reihe passiert ist. Manchmal beginnen Sie mit einer Stichprobe aus einer Grundgesamtheit, über deren Parameter Sie einiges wissen. Sie müssen entscheiden, ob diese Stichprobe dem Rest der Grundgesamtheit entspricht oder nicht.

Messen Sie diese Stichprobe und berechnen Sie die zugehörigen statistischen Werte. Vergleichen Sie schließlich diese Werte mit den Parametern der Grundgesamtheit. Sind sie gleich? Sind sie anders? Repräsentiert die Stichprobe etwas, das außerhalb der Reihe ist? Die richtige Anwendung der Statistik hilft Ihnen bei der Entscheidung.

Manchmal kennen Sie die Parameter der betreffenden Grundgesamtheit nicht. Was dann? In diesem Kapitel lernen Sie Techniken der Statistik sowie Tabellenblattfunktionen für den Umgang mit beiden Beispielen kennen.

Hypothesen, Tests und Fehler

Eine *Hypothese* ist eine Annahme darüber, wie die Welt funktioniert. Sie ist eine vorläufige Erklärung eines natürlichen oder künstlichen Vorgangs. Bevor ein Forscher die Elemente in einer Stichprobe untersucht und misst, formuliert er Hypothesen, die vorhersagen, wie die Daten aussehen sollen.

Eine Hypothese sagt meist voraus, dass die Daten nichts Neues oder Interessantes zeigen. Die so genannte *Nullhypothese* (abgekürzt mit H_0) besagt, dass es reiner Zufall ist, wenn Daten in irgendeiner Weise von der Norm abweichen. Die so genannte *Alternativhypothese* (abgekürzt mit H_1) beschreibt die Dinge anders. Nach der Alternativhypothese zeigen die Daten etwas Wichtiges.

Nach dem Zusammentragen der Daten muss der Forscher eine Entscheidung treffen. Aufgrund der Funktionsweise der Logik konzentriert sich die Entscheidung auf die Nullhypothese. Der Forscher muss sich entscheiden, ob er die Nullhypothese verwirft. Beim *Testen von Hypothesen* werden Hypothesen formuliert, Daten gesammelt und es wird entschieden, ob die Nullhypothese verworfen wird.

In der Logik ist es nicht möglich, eine Hypothese *anzunehmen*. Die Logik sieht auch keine Entscheidung über die Alternativhypothese vor. Es geht nur darum, die Nullhypothese zu verwerfen oder nicht zu verwerfen.

Unabhängig von der Entscheidung, eine Hypothese zu verwerfen oder nicht zu verwerfen, sind Fehler möglich. Eine Art Fehler tritt auf, wenn Sie glauben, dass die Daten etwas Wichtiges zeigen, und Sie die Nullhypothese verwerfen, die Daten in Wirklichkeit aber auf einen Zufall zurückzuführen sind. Hierbei spricht man von einem Fehler 1. Art. Zu Beginn einer Untersuchung legen Sie die Kriterien für das Verwerfen der Nullhypothese fest. Dabei legen Sie die

Wahrscheinlichkeit für einen Fehler 1. Art fest. Diese Wahrscheinlichkeit wird als *Alpha* (α) bezeichnet.

Die andere Art Fehler tritt auf, wenn Sie die Nullhypothese nicht verwerfen, die Daten in Wirklichkeit aber auf etwas Außergewöhnliches zurückzuführen sind. Aus irgendeinem Grund haben Sie das übersehen. Hierbei spricht man von einem Fehler 2. Art. Die zugehörige Wahrscheinlichkeit wird als *Beta* (β) bezeichnet. In Tabelle 10.1 sind die möglichen Entscheidungen und Fehler zusammengefasst.

		»Tatsächlicher Zustand« der Welt	
		ist wahr	ist wahr
	verwerfen	Fehler 1. Art	Richtige Entscheidung
Entscheidung			
	nicht verwerfen	Richtige Entscheidung	Fehler 2. Art

Tabelle 10.1: Entscheidungen und Fehler beim Testen von Hypothesen

Sie wissen nie, wie die Realität wirklich aussieht. Sie können immer nur die Elemente in einer Stichprobe messen, die statistischen Werte berechnen und eine Entscheidung über H_0 treffen.

Hypothesentests und Stichprobenverteilungen

In Kapitel 9 ging es um Stichprobenverteilungen. Erinnern Sie sich noch? Eine Stichprobenverteilung ist die Menge aller möglichen Werte einer Statistik für einen gegebenen Stichprobenumfang.

Ebenfalls in Kapitel 9 wurde der zentrale Grenzwertsatz beschrieben. Dieser Lehrsatz besagt, dass sich die Stichprobenverteilung des Mittelwerts einer Normalverteilung nähert, wenn der Stichprobenumfang groß ist (aus praktischen Gründen mindestens 30). Dies gilt auch dann, wenn die Grundgesamtheit

nicht normalverteilt ist. Wenn die Grundgesamtheit normalverteilt ist, ist die Stichprobenverteilung für jeden Stichprobenumfang normal. Zwei weitere Punkte aus dem zentralen Grenzwertsatz:

- Der Mittelwert der Stichprobenverteilung des Mittelwerts ist gleich dem Mittelwert der Grundgesamtheit.

  Die Gleichung hierfür lautet wie folgt:

$$\mu_{\bar{X}} = \mu$$

- Der Standardfehler des Mittelwerts (auch als Standardabweichung der Stichprobenverteilung bezeichnet) ist gleich der Standardabweichung der Grundgesamtheit dividiert durch die Quadratwurzel des Stichprobenumfangs.

  Diese Gleichung lautet folgendermaßen:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Die Stichprobenverteilung des Mittelwerts spielt bei der Art von Hypothesentest, der in diesem Kapitel beschrieben wird, eine bedeutende Rolle. Theoretisch entspricht beim Testen einer Nullhypothese gegen eine Alternativhypothese jede Hypothese einer eigenen Stichprobenverteilung.

Anhand von Abbildung 10.1 wird deutlich, was ich meine. In der Abbildung sind zwei Normalverteilungen dargestellt. Ich habe sie willkürlich platziert. Jede Normalverteilung repräsentiert eine Stichprobenverteilung des Mittelwerts. Es wird die Verteilung von möglichen Stichprobenmittelwerten dargestellt: links, wenn die Nullhypothese der Realität entspricht, und rechts, wenn die Alternativhypothese stimmt.

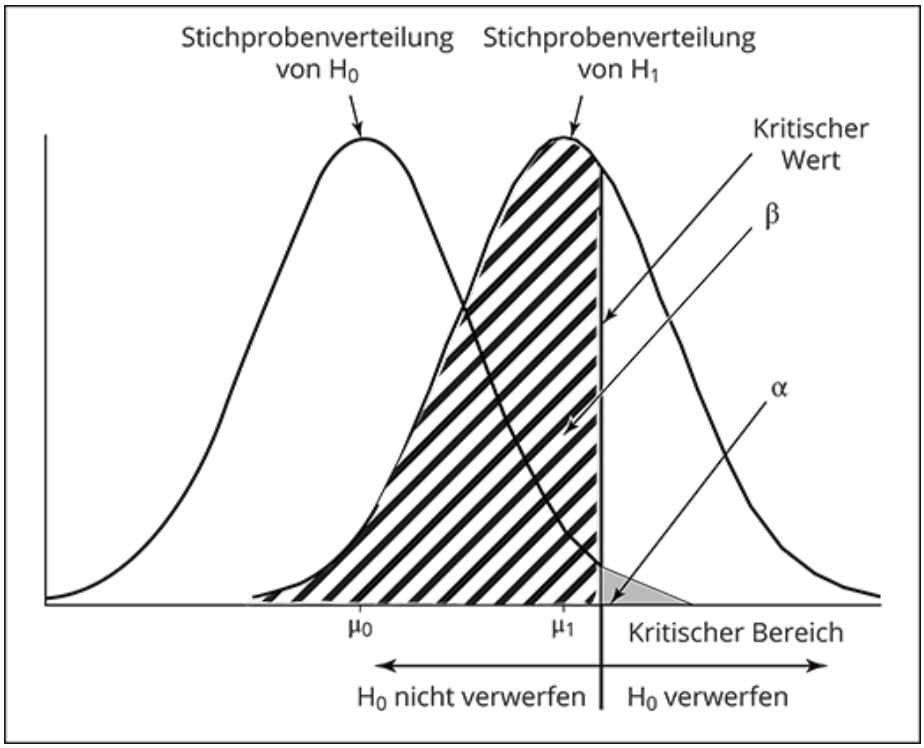


Abbildung 10.1: H_0 und H_1 entsprechen jeweils einer Stichprobenverteilung.

Wenn Sie eine Hypothese testen, wissen Sie natürlich nie, welche Verteilung zu den Beobachtungen führt. Sie arbeiten mit einem Stichprobenmittelwert, einem Punkt auf der x-Achse. Sie müssen entscheiden, welcher Verteilung der Stichprobenmittelwert angehört. Sie legen einen *kritischen Wert*, ein Entscheidungskriterium, fest. Wenn sich der Stichprobenmittelwert auf einer Seite des kritischen Werts befindet, verwerfen Sie H_0. Wenn nicht, dann nicht.

In der Abbildung sind auch α und β dargestellt. Hierbei handelt es sich, wie bereits erwähnt, um die Wahrscheinlichkeiten von Entscheidungsfehlern. Der Bereich, der α entspricht, befindet sich in der H_0-Verteilung. Dieser Bereich ist dunkelgrau schattiert. Er repräsentiert die Wahrscheinlichkeit, dass ein Stichprobenmittelwert von der H_0-Verteilung stammt, jedoch so extrem ist, dass H_0 verworfen wird.

Der kritische Wert wird durch α bestimmt. Bei den meisten Hypothesentests wird für α 0,05 festgelegt. Das bedeutet, dass Sie bereit sind, einen Fehler der 1. Art (irrtümliches Verwerfen von H_0) in 5 % der Fälle zu tolerieren. Grafisch schneidet der kritische Wert 5 % der Fläche der Stichprobenverteilung ab. Wenn übrigens von den 5 % der Fläche auf der rechten Seite der Verteilung (wie in Abbildung 10.1) die Rede ist, sind damit die oberen 5 % gemeint. Wenn es um die 5 % auf der linken Seite geht, sind es die unteren 5 %.

Der Bereich, der β entspricht, befindet sich in der H_1-Verteilung. Dieser Bereich ist hellgrau schattiert. Dieser Bereich repräsentiert die Wahrscheinlichkeit, dass ein Stichprobenmittelwert aus der H_1-Verteilung stammt, sich jedoch nah genug am Zentrum der H_0-Verteilung befindet, dass H_0 nicht verworfen wird. β wird nicht von Ihnen festgelegt. Die Größe dieser Fläche hängt vom Abstand der Mittelwerte der beiden Verteilungen ab und dieser wird von der Welt, in der wir leben, bestimmt, und nicht von Ihnen.

Diese Stichprobenverteilungen sind geeignet, wenn Ihre Arbeit den Bedingungen des zentralen Grenzwertsatzes entspricht: wenn Sie wissen, dass die Grundgesamtheit, mit der Sie arbeiten, normalverteilt ist, oder wenn die Stichprobe groß ist.

Noch einmal z-Werte

Hier sehen Sie ein Beispiel für einen Hypothesentest mit einer Stichprobe aus einer normalverteilten Grundgesamtheit. Weil die Grundgesamtheit normalverteilt ist, stellt jeder Stichprobenumfang eine normalverteilte Stichprobenverteilung dar. Weil es sich um eine Normalverteilung handelt, werden für den Hypothesentest *z-Werte* verwendet:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Und hier noch ein Grund: Weil für den Hypothesentest der z-Wert verwendet wird, spricht man in diesem Zusammenhang von der *Testgröße* oder auch *Teststatistik*.

Nehmen wir an, Sie sind der Meinung, dass Menschen, die in einem bestimmten Postleitzahlenbereich leben, einen überdurchschnittlichen IQ haben. Sie ziehen eine Stichprobe von 16 Personen aus diesem Postleitzahlenbereich, unterziehen sie einem IQ-Test, erfassen die Ergebnisse und berechnen die statistischen Werte. Es ist bekannt, dass für die Grundgesamtheit von IQ-Werten $\mu = 100$ und $\sigma = 16$ beträgt (zumindest für die Version von Standford-Binet).

Die Hypothesen lauten wie folgt:

H_0: $\mu_{PLZ} \leq 100$

H_1: $\mu_{PLZ} > 100$

Nehmen wir an, dass $\alpha = 0{,}05$ ist. Das ist der schattierte Bereich auf der rechten Seite der H_0-Verteilung in Abbildung 10.1.

Wozu das ≤ in H_0? Dieses Symbol wird verwendet, weil H_0 nur dann verworfen wird, wenn der Stichprobenmittelwert größer als der hypothetische Wert ist. Alles andere spricht dafür, H_0 nicht zu verwerfen.

Nehmen wir an, der Stichprobenmittelwert beträgt 107,75. Kann H_0 verworfen werden?

Für den Test muss 107,75 in einen Standardwert in der Stichprobenverteilung des Mittelwerts umgerechnet werden:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{107{,}75 - 100}{\frac{16}{\sqrt{16}}} = \frac{7{,}75}{\frac{16}{4}} = \frac{7{,}75}{4} = 1{,}94$$

Ist der Wert der Testgröße groß genug, damit Sie H_0 mit $\alpha = 0{,}05$ verwerfen können? Ja, ist er. Der kritische Wert, also der Wert von z, der 5 % der Fläche in einer Standardnormalverteilung abtrennt, beträgt 1,64. (Nach jahrelanger Erfahrung mit der Standardnormalverteilung weiß ich das einfach. In Kapitel 8 lernen Sie die Excel-Funktion `NORM.S.INV` kennen, mit der Sie Informationen wie diese ebenfalls jederzeit griffbereit haben.) Der berechnete Wert 1,94 ist größer als 1,645 und befindet sich somit im kritischen Bereich. Also wird H_0 verworfen.

Das bedeutet, dass, wenn H_0 wahr ist, die Wahrscheinlichkeit für eine Testgröße, die mindestens so groß ist, weniger als 0,05 beträgt. Das ist Beweis genug, um H_0 zu verwerfen. Ein Statistiker bezeichnet das Ergebnis beim Verwerfen von H_0 als »statistisch signifikant«.

Diese Art Hypothesentest wird als *einseitig* bezeichnet, weil sich der kritische Bereich auf einer Seite der Stichprobenverteilung befindet.

Ein Hypothesentest kann auch einseitig in die andere Richtung sein. Stellen Sie sich vor, Sie haben Grund zu der Annahme, dass die Menschen in dem Postleitzahlenbereich einen unterdurchschnittlichen IQ haben. In dem Fall lauten die Hypothesen wie folgt:

$H_0: \mu_{PLZ} \geq 100$

$H_1: \mu_{PLZ} < 100$

Bei diesem Hypothesentest beträgt der kritische Wert der Testgröße –1,645, wenn $\alpha = 0{,}05$ ist.

Ein Hypothesentest kann zweiseitig sein, das heißt, der kritische Bereich befindet sich auf beiden Seiten der H_0-Stichprobenverteilung. Das ist dann der Fall, wenn die Hypothesen folgendermaßen lauten:

$H_0: \mu_{PLZ} = 100$

$H_1: \mu_{PLZ} \neq 100$

In diesem Fall gibt die Alternativhypothese lediglich an, dass sich der Mittelwert vom Wert der Nullhypothese unterscheidet. Sie sagt aber nichts darüber aus, ob er größer oder kleiner ist. In Abbildung 10.2 ist der zweiseitige kritische Bereich für α = 0,05 dargestellt. Die 5 % sind gleichmäßig auf die linke Seite (auch als untere Seite bezeichnet) und rechte Seite (auch als obere Seite bezeichnet) verteilt.

Bei einer Standardnormalverteilung beträgt der z-Wert, der 2,5 % auf der rechten Seite abschneidet, zufällig 1,96. Der z-Wert, der 2,5 % auf der linken Seite abschneidet, beträgt –1,96. (Wie erwähnt, nach jahrelanger Erfahrung mit der Standardnormalverteilung kenne ich diese Werte einfach.) Der z-Wert im vorhergehenden Beispiel, 1,94, ist nicht größer als 1,96. H_0 wird also beim zweiseitigen Test *nicht* verworfen.

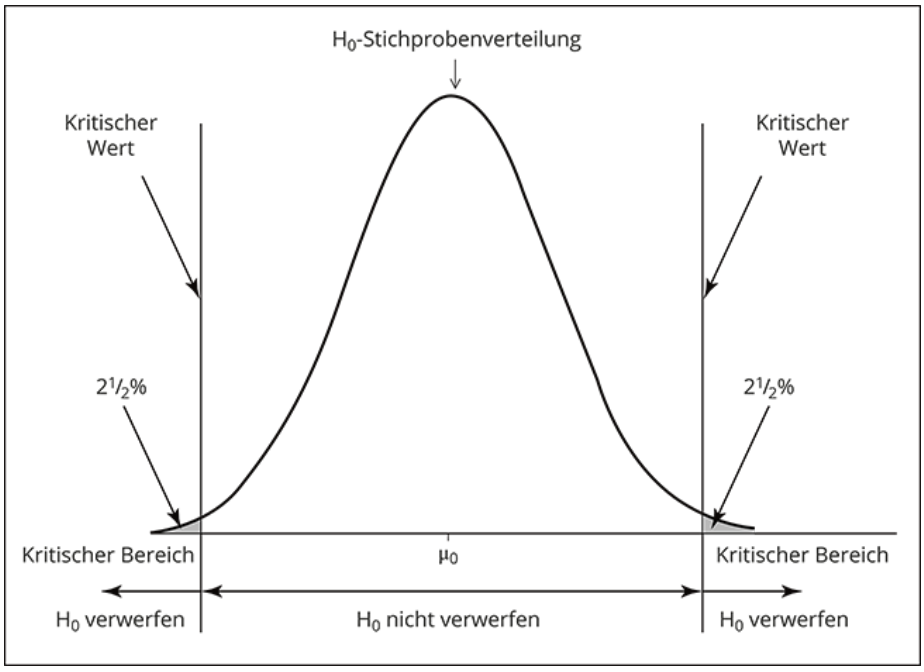


Abbildung 10.2: Der zweiseitige kritische Bereich für α = 0,05.

Hier kommen wir zu einem wichtigen Punkt. Mit einem einseitigen Hypothesentest kann H_0 verworfen werden, während das bei einem zweiseitigen nicht möglich ist. Ein

zweiseitiger Test dient dazu, nach der Differenz zwischen dem Stichprobenmittelwert und dem Mittelwert der Nullhypothese zu suchen, wobei jedoch nicht bekannt ist, in welcher Richtung gesucht werden soll. Ein einseitiger Test zeigt, dass Sie eine relativ gute Vorstellung davon haben, was sich als Differenz ergibt. Praktisch bedeutet das, dass Sie versuchen sollten, so viele Informationen zusammenzutragen, dass Sie einen einseitigen Test durchführen können.

G.TEST

Die Excel-Tabellenblattfunktion `G.TEST` übernimmt die Berechnung für Hypothesentests mit z-Werten in einer Standardnormalverteilung. Sie geben Stichprobendaten, einen Nullhypothesenwert und die Standardabweichung einer Grundgesamtheit an. `G.TEST` gibt die Wahrscheinlichkeit am einen Ende der H_0-Stichprobenverteilung zurück.

Das ist etwas anders als bei den Formeln, die ich Ihnen eben gezeigt habe. Mit der Formel wird ein z-Wert berechnet. Dabei bleibt es Ihnen überlassen festzustellen, wo sich der Wert in einer Standardnormalverteilung in Bezug auf die Wahrscheinlichkeit befindet. `G.TEST` geht ohne Umweg (die Notwendigkeit, den z-Wert zu berechnen) direkt auf die Wahrscheinlichkeit zu.

In Abbildung 10.3 sind die Daten und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `G.TEST` dargestellt. Bei den Daten handelt es sich um IQ-Werte für 16 Personen aus dem Beispiel mit dem Postleitzahlenbereich im vorhergehenden Abschnitt. Bei diesem Beispiel wurde die Hypothese getestet, dass Menschen in einem bestimmten Postleitzahlenbereich einen überdurchschnittlichen IQ haben.

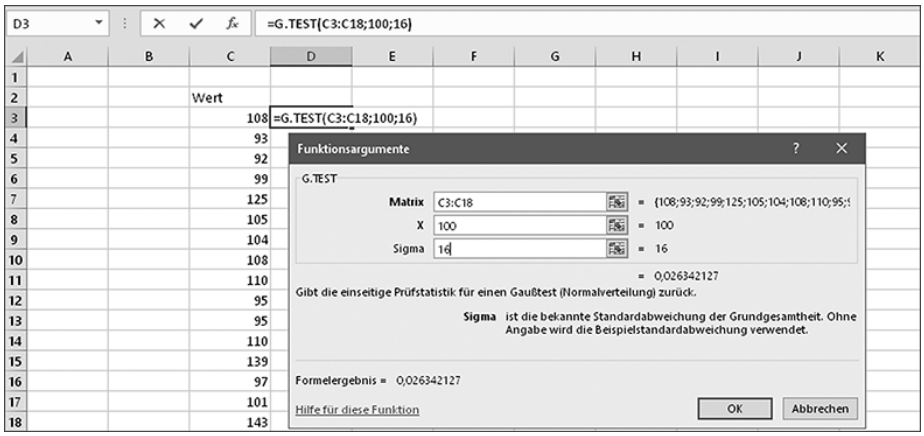


Abbildung 10.3: Daten und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `G.TEST`

Und so müssen Sie vorgehen:

1. **Geben Sie die Daten in einen Zellbereich ein und wählen Sie eine Zelle für das Ergebnis aus.**

   In diesem Beispiel befinden sich die Daten in den Zellen C3 bis C18. Das Ergebnis der Funktion soll in Zelle D3 ausgegeben werden.

2. **Wählen Sie im Menü mit den statistischen Funktionen G.TEST aus, um das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `G.TEST` zu öffnen (siehe Abbildung 10.3).**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente ein, die Sie an die Funktion übergeben wollen.**

   In diesem Beispiel ist die MATRIX der Zellbereich C3:C18. In das Feld X gebe ich den Mittelwert ein. In diesem Beispiel beträgt der Mittelwert 100. Das ist der Mittelwert der IQ-Werte in der Grundgesamtheit. In das Feld SIGMA gebe ich **16** ein, das ist die Standardabweichung der Grundgesamtheit für den IQ. Sobald ich diese Zahl eingegeben habe, wird das Ergebnis (0,026342) im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, damit das Ergebnis in der markierten Zelle angezeigt wird.**

Bei α = 0,05 und einem einseitigen Test (H_1: μ > 100) wird H_0 verworfen, da das Ergebnis (0,026) kleiner als 0,05 ist. Bei einem zweiseitigen Test (H_1: μ ≠ 100) wird H_0 nicht verworfen. Das kommt daher, weil 2 × 0,026 größer als 0,05 ist. Nur wenig größer (0,052). Aber wenn Sie die Linie bei 0,05 ziehen, können Sie H_0 nicht verwerfen.

t-Test für eine Stichprobe

Im vorhergehenden Beispiel habe ich mit IQ-Werten gearbeitet. Die Grundgesamtheit von IQ-Werten weist eine Normalverteilung mit bekanntem Mittelwert und bekannter Standardabweichung auf. Daher konnte ich mit dem zentralen Grenzwertsatz arbeiten und die Stichprobenverteilung des Mittelwerts als eine Normalverteilung beschreiben. Daraufhin konnte ich den z-Wert als Testgröße verwenden.

In der Realität haben Sie jedoch selten den Luxus, mit derart gut definierten Grundgesamtheiten zu arbeiten. Meist stehen Ihnen kleine Stichproben zur Verfügung und Sie messen in der Regel etwas, das nicht so bekannt ist wie der IQ. Sie kennen also häufig die Parameter der Grundgesamtheit nicht und wissen meist auch nicht, ob die Grundgesamtheit normalverteilt ist.

In diesem Fall schätzen Sie mit den Stichprobendaten die Standardabweichung der Grundgesamtheit und Sie behandeln die Stichprobenverteilung des Mittelwerts als eine so genannte *t*-Verteilung. Dabei dient *t* als Testgröße. In Kapitel 9 haben Sie diese Verteilung kennengelernt und erfahren, dass sich Verteilungen dieser Art durch einen Parameter unterscheiden, der als *Freiheitsgrade* (df) bezeichnet wird.

Die Formel für die Testgröße lautet wie folgt:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Stellen Sie sich df als Nenner des Schätzers der Varianz der Grundgesamtheit vor. Bei den Hypothesentests in diesem Abschnitt ist das *n* – 1, wobei *n* für die Anzahl der Werte in der Stichprobe steht. Je höher der df-Wert, umso mehr gleicht die *t*-Verteilung der Normalverteilung.

Ein Beispiel: FarKlempt Robotics vertreibt Mikroroboter. Das Unternehmen behauptet, dass sein Produkt pro Einheit durchschnittlich vier Fehler aufweist. Eine Verbrauchergruppe ist der Ansicht, dass der Durchschnittswert höher liegt. Die Verbrauchergruppe ermittelt anhand einer Stichprobe von neun FarKlempt-Mikrorobotern einen Durchschnitt von sieben Fehlern bei einer Standardabweichung von 3,16. Der Hypothesentest lautet wie folgt:

H_0: $\mu \leq 4$

H_1: $\mu > 4$

$\alpha = 0{,}05$

Die Gleichung hierfür lautet:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{7-4}{\frac{3{,}16}{\sqrt{9}}} = \frac{3}{\frac{3{,}16}{3}} = 2{,}85$$

Kann H_0 verworfen werden? Diese Frage beantworten Ihnen die Excel-Funktionen im nächsten Abschnitt.

T.VERT, T.VERT.RE und T.VERT.2S

Mit der Familie der `T.VERT`-Tabellenblattfunktionen können Sie entscheiden, ob Ihr berechneter *t*-Wert im kritischen Bereich liegt. Bei der Funktion `T.VERT` geben Sie einen Wert für *t*, einen Wert für df und das Argument KUMULIERT an. `T.VERT` ermittelt die

Wahrscheinlichkeit, einen *t*-Wert zu erhalten, der mindestens so groß ist wie Ihrer, wenn H_0 wahr ist. Wenn diese Wahrscheinlichkeit geringer ist als Ihr α, wird H_0 verworfen.

Gehen Sie wie folgt vor:

1. **Wählen Sie eine Zelle aus, in die das Ergebnis eingefügt werden soll.**

2. **Wählen Sie im Menü mit den statistischen Funktionen T.VERT aus, um das Dialogfeld FUNKTIONSARGUMENTE für `T.VERT` zu öffnen (siehe Abbildung 10.4).**

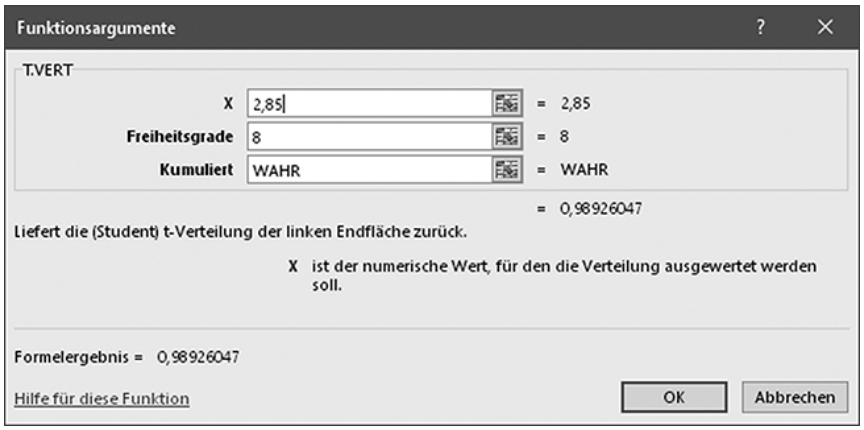


Abbildung 10.4: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `T.VERT`

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die Argumente für `T.VERT` ein.**

   Der berechnete t-Wert wird in das Feld x eingegeben. In diesem Beispiel ist der berechnete t-Wert 2,85.

   Geben Sie in das Feld FREIHEITSGRADE die Freiheitsgrade ein. Die Freiheitsgrade betragen bei diesem Beispiel 8 (9 Werte − 1).

   In das Feld KUMULIERT geben Sie entweder `WAHR` oder `FALSCH` ein. Wenn ich hier `WAHR` eingebe, gibt `T.VERT` die kumulierte Verteilungsfunktion zurück, und ich erhalte so die Wahrscheinlichkeit, dass ich in der t-Verteilung den Wert x oder kleiner erhalte, und zwar für die angegebene Anzahl der

Freiheitsgrade. Excel verwendet hierfür den Begriff linksseitige Student-t-Verteilung. Wenn ich hier FALSCH eingebe, erhalte ich die Höhe der t-Verteilung an x (die Wahrscheinlichkeitsdichtefunktion). Ich verwende diese Option etwas weiter hinten in diesem Kapitel bei der Erstellung des Diagramms einer t-Verteilung. Ansonsten fällt mir kein anderer Fall ein, für den Sie FALSCH in dieses Feld eingeben müssten.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

Der Wert im Dialogfeld in Abbildung 10.4 ist kleiner als 0,95. Somit wird H_0 verworfen.

Möglicherweise halten Sie T.VERT.RE für besser geeignet, zumindest für dieses Beispiel. Das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion sieht so aus wie das in Abbildung 10.4, jedoch fehlt das Feld KUMULIERT. Diese Funktion ermittelt die Wahrscheinlichkeit, in der t-Verteilung mindestens den Wert von X zu erhalten. Das RE im Funktionsnamen steht für rechtsseitig. Für dieses Beispiel gibt die Funktion den Wert 0,01073953 zurück. Da dieser Wert kleiner ist als 0,05, lautet die Entscheidung, H_0 zu verwerfen.

T.VERT.2S gibt die zweiseitige Wahrscheinlichkeit, also die t-Verteilung für zwei Endflächen, zurück. Das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion sieht so aus wie das für T.VERT.RE. Sie gibt die Wahrscheinlichkeit rechts von X in der t-Verteilung und die Wahrscheinlichkeit links von –X in der Verteilung zurück.

T.INV und T.INV.2S

Die beiden T.INV-Tabellenblattfunktionen sind die Umkehrungen der entsprechenden T.VERT-Funktionen. Übergeben Sie T.INV eine Wahrscheinlichkeit und einen Wert für die Anzahl der Freiheitsgrade und Sie erhalten den Wert von t, der die Wahrscheinlichkeit an der linken Seite der t-Verteilung abschneidet. Gehen Sie so vor, um T.INV zu verwenden:

1. **Wählen Sie die Zelle aus, in die Sie das Ergebnis einfügen wollen.**

2. **Wählen Sie im Menü mit den statistischen Funktionen T.INV aus, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen (siehe Abbildung 10.5).**

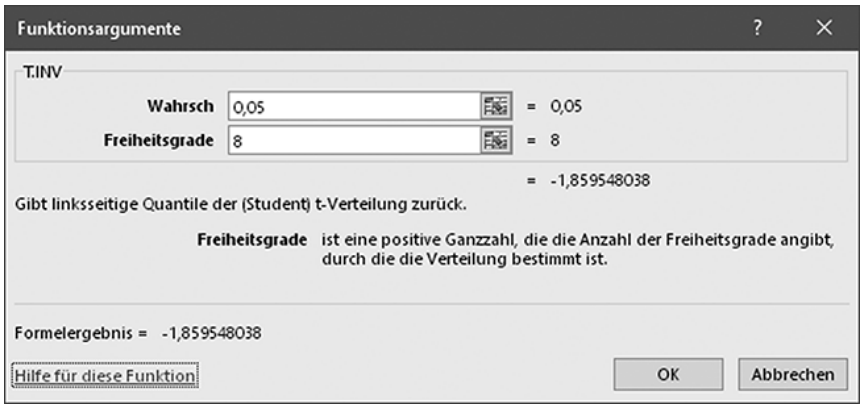


Abbildung 10.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `T.INV`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   In das Feld WAHRSCH habe ich **0,05** und in das Feld FREIHEITSGRADE **8** eingegeben. Das Ergebnis der Funktion (−1,859548038) wird im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und um das Ergebnis in die ausgewählte Zelle einzufügen.**

Das Dialogfeld FUNKTIONSARGUMENTE für `T.INV.2S` sieht genauso aus wie das für `T.INV`. Sie übergeben eine Wahrscheinlichkeit und die Freiheitsgrade und die Funktion schneidet die Wahrscheinlichkeit in zwei Teile. Die Funktion gibt den Wert von t auf der rechten Seite zurück, der die Wahrscheinlichkeit halbiert. Was ist mit der anderen Hälfte? Dies ist der gleiche numerische Wert, jedoch mit –1 multipliziert. Dieser negative Wert von *t* schneidet die andere Hälfte der Wahrscheinlichkeit ab, und zwar in der linken Seite der Verteilung.

Eine t-Verteilung visualisieren

Wie ich bereits in Kapitel 8 erwähnt habe, hilft die Visualisierung einer Verteilung oft dabei, sie zu verstehen. Die Visualisierung ist einfach und lehrreich. Abbildung 10.6 zeigt das Diagramm einer *t*-Verteilung. Hier werden die Funktion `VERT.T` und für das Argument KUMULIERT der Wert `FALSCH` verwendet.

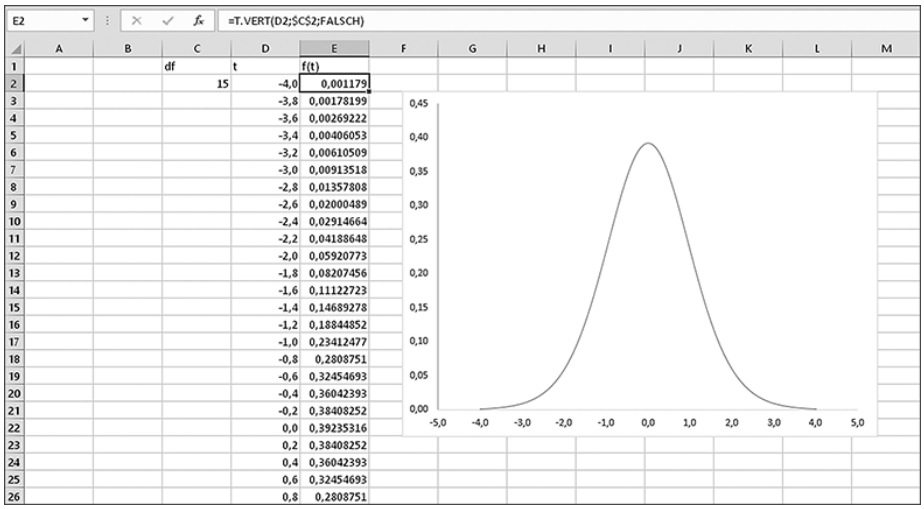


Abbildung 10.6: Visualisierung einer t-Verteilung

Hier die einzelnen Schritte:

1. **Geben Sie die Anzahl der Freiheitsgrade in eine Zelle ein.**

   Ich habe in Zelle C2 **15** eingegeben.

2. **Erstellen Sie eine Spalte mit Werten für die Statistik.**

   In den Zellbereich D2 bis D42 hab ich die Werte zwischen **–4** und **4** in Schritten von 0,2 eingegeben.

3. **Geben Sie in die erste Zelle der benachbarten Spalte den Wert der Wahrscheinlichkeitsdichte für den ersten Datenwert ein.**

Da ich eine t-Verteilung visualisieren will, verwende ich in Zelle E2 die Funktion VERT.T. Für den Wert von X klicke ich Zelle D2 an. Das Argument FREIHEITSGRADE steht in Zelle C2. Nachdem ich diese Zelle angeklickt habe drücke ich F4, um einen absoluten Zellbezug zu erstellen. In das Feld KUMULIERT gebe ich FALSCH ein, um die Höhe der Verteilung für diesen Wert von t zu erhalten. Abschließend klicke ich auf OK.

4. **Füllen Sie Spalte E automatisch aus.**

5. **Erstellen Sie das Diagramm.**

   Markieren Sie dazu beide Spalten. Klicken Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME auf PUNKTE MIT INTERPOLIERTEN LINIEN.

6. **Bearbeiten Sie das Diagramm.**

   Anfänglich befindet sich die y-Achse in der Mitte des Diagramms. Klicken Sie die x-Achse an, um den Aufgabenbereich ACHSE FORMATIEREN zu öffnen. Klicken Sie im Abschnitt ACHSENOPTIONEN unterhalb von VERTIKALE ACHSE SCHNEIDET auf ACHSENWERT, und geben Sie in das betreffende Feld **–5** ein. Anschließend können Sie das Diagramm anklicken, damit die Schaltfläche DIAGRAMMELEMENTE angezeigt wird und dann die Achsentitel (t und f(t)) ergänzen. Ich habe den Diagrammtitel und die Gitternetzlinien entfernt, aber das ist eine Frage des persönlichen Geschmacks. Außerdem habe ich das Diagramm vergrößert.

7. **Verändern Sie die Diagrammdaten.**

   Um ein besseres Gefühl für die t-Verteilung zu bekommen, können Sie den Wert *df* (die Anzahl der Freiheitsgrade) ändern und so sehen, wie sich dies auf das Diagramm auswirkt.

Testen einer Varianz

Bisher haben Sie den Ein-Stichproben-Hypothesentest für Mittelwerte kennengelernt. Sie können aber auch Hypothesen zu Varianzen testen.

Dieses Thema taucht manchmal im Zusammenhang mit Herstellungsprozessen auf. Das Unternehmen FarKlempt Robotics stellt beispielsweise ein Bauteil her, das eine bestimmte Länge haben muss und nur eine sehr geringe Streuung aufweisen darf. Sie können eine Stichprobe von Bauteilen ziehen, diese messen, die Streuung der Stichprobe ermitteln und einen Hypothesentest zur gewünschten Streuung durchführen.

Die Familie der Verteilungen für diesen Test wird *Chi-Quadrat* genannt. Als Symbol wird χ^2 verwendet. Keine Angst: Ich werde Ihnen nun nicht den ganzen mathematischen Hintergrund erläutern. Nur so viel: df ist auch hier der Parameter, der den Unterschied zu den anderen Verteilung ausmacht. In Abbildung 10.7 sind zwei Chi-Quadrat-Verteilungen dargestellt.

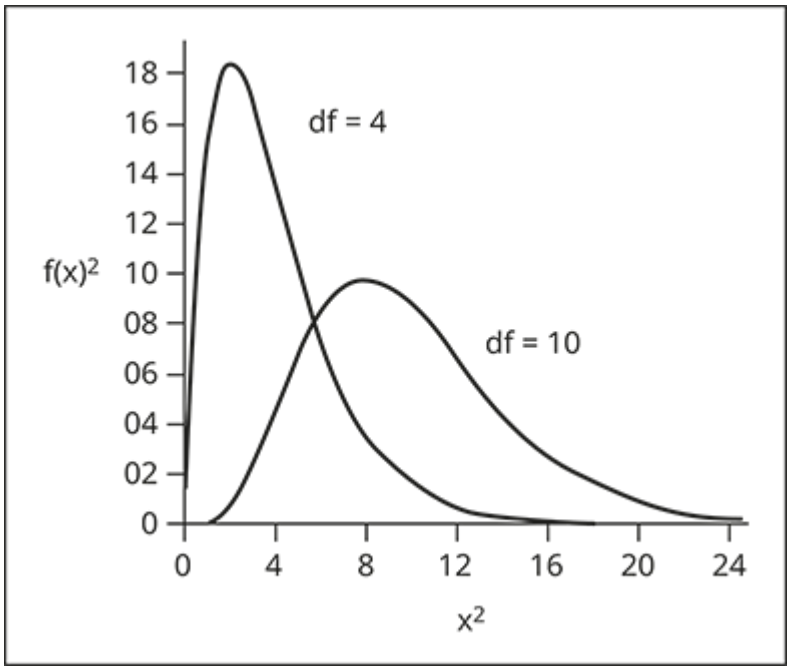


Abbildung 10.7: Zwei Chi-Quadrat-Verteilungskurven

Die Formel für diese Testgröße lautet wie folgt:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

n ist die Anzahl der Werte in der Stichprobe, s^2 ist die Stichprobenvarianz und σ^2 ist die Varianz der Grundgesamtheit unter H_0.

Bei diesem Test müssen Sie für das, was Sie messen, eine Normalverteilung voraussetzen.

Nehmen wir an, dass es beim Herstellungsprozess bei FarKlempt nur zu einer Standardabweichung von maximal 38,1 mm in der Länge kommen darf. (Ich habe übrigens *Standardabweichung* verwendet. So ist es möglich, von mm zu sprechen. Wenn ich *Varianz* geschrieben hätte, wäre die fragliche Einheit mm^2.) Nach dem Messen einer Stichprobe von 26 Bauteilen ergibt sich eine Standardabweichung von 45,72 mm.

Die Hypothesen lauten wie folgt:

H_0: $\sigma^2 \leq 1451{,}61$ (Denken Sie daran, die maximale Standardabweichung von 38,1 mm zu quadrieren.)

In die Formel eingesetzt:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(26-1)(45{,}72)^2}{(38{,}1)^2} = \frac{(25)(2090{,}3184)}{(1451{,}61)} = 36$$

Kann H_0 verworfen werden? Lesen Sie weiter.

CHIQU.VERT und CHIQU.VERT.RE

Nachdem Sie einen Wert für Ihre Chi-Quadrat-Testgröße berechnet haben, bewerten Sie sie mit der Tabellenblattfunktion `CHIQU.VERT`. Sie geben den Chi-Quadrat-Wert und den df-Wert an. Wie bei `T.VERT` geben Sie für den Parameter KUMULIERT entweder `FALSCH` oder `WAHR` an. Wenn Sie `WAHR` angeben, ermittelt `CHIQU.VERT` die Wahrscheinlichkeit, höchstens einen so großen Wert zu erhalten, wenn H_0 wahr ist. (Dies ist eine linksseitige Wahrscheinlichkeit).

Wenn diese Wahrscheinlichkeit größer ist als 1 – α, verwerfen Sie H_0.

Damit Sie sehen, wie Sie mit dieser Funktion arbeiten können, verwende ich die Informationen aus dem Beispiel im vorherigen Abschnitt. Gehen Sie dazu wie folgt vor:

1. **Wählen Sie die Zelle aus, in der Sie das Ergebnis einfügen wollen.**

2. **Wählen Sie im Menü mit den statistischen Funktionen CHIQU.VERT aus, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen (siehe Abbildung 10.8).**

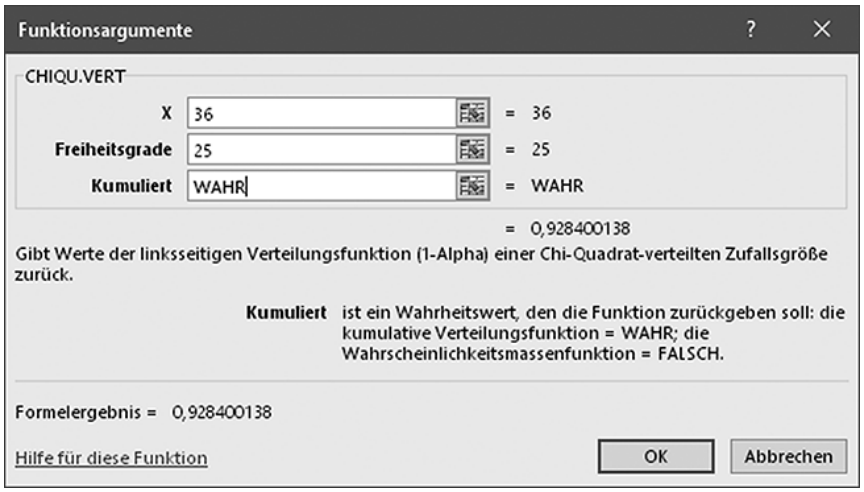


Abbildung 10.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion CHIQU.VERT

3. **Geben Sie in die verschiedenen Felder des Dialogfeldes FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   In das Feld X habe ich den berechneten Chi-Quadrat-Wert eingegeben. In diesem Beispiel ist das der Wert **36**.

   In das Feld FREIHEITSGRADE habe ich die Anzahl der Freiheitsgrade eingegeben. In diesem Beispiel betragen die Freiheitsgrade **25** (26–1).

In das Feld KUMULIERT habe ich WAHR eingegeben. So erhalte ich den Wert der Verteilungsfunktion, also die Wahrscheinlichkeit, dass ich höchstens den in das Feld x eingegebenen Wert erhalte. Wenn ich hier FALSCH verwende, wird der Wert der Dichtefunktion berechnet, also die Höhe der Chi-Quadrat-Verteilung an x. Dies ist lediglich in den Fällen nützlich, in denen Sie die Chi-Quadrat-Verteilung in einem Diagramm darstellen wollen, was ich etwas weiter hinten machen werde.

Nachdem Sie WAHR eingegeben haben, zeigt das Dialogfeld die Wahrscheinlichkeit an, einen höchstens so großen Wert für Chi zu erhalten, wenn H_0 wahr ist.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und die Lösung in der markierten Zelle zu speichern.**

Der Wert im Dialogfeld in Abbildung 10.8 ist größer als 1–0,05. Somit wird H_0 nicht verworfen. (Können wir daraus schließen, dass sich die Maße bei der Herstellung innerhalb akzeptabler Streuungsgrenzen befinden? Diese Frage wird im Abschnitt *Etwas zum Nachdenken* beantwortet.)

CHIQU.VERT.RE arbeitet so wie CHIQU.VERT, jedoch finden Sie im Dialogfeld kein Feld KUMULIERT. Geben Sie einen Wert für Chi-Quadrat in das Feld x und die Anzahl der Freiheitsgrade ein. Sie erhalten dann die rechtsseitige Wahrscheinlichkeit, also die Wahrscheinlichkeit, einen Wert für Chi-Quadrat zu erhalten, der mindestens so groß ist wie der in das Feld x eingegebene Wert.

CHIQU.INV und CHIQU.INV.RE

CHIQU.INV ist die Umkehrfunktion von CHIQU.VERT und CHIQU.INV.RE die Umkehrfunktion von CHIQU.RE. Sie geben eine Wahrscheinlichkeit und einen df-Wert an, und CHIQU.INV ermittelt den entsprechenden Chi-Quadrat-Wert, der die Wahrscheinlichkeit an der linken Seite der Chi-Quadrat-Verteilung abschneidet. Führen Sie diese Schritte durch:

1. **Wählen Sie die Zelle aus, in die Sie das Ergebnis einfügen wollen.**

2. **Wählen Sie im Menü mit den statistischen Funktionen CHIQU.INV aus, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen (siehe Abbildung 10.9).**

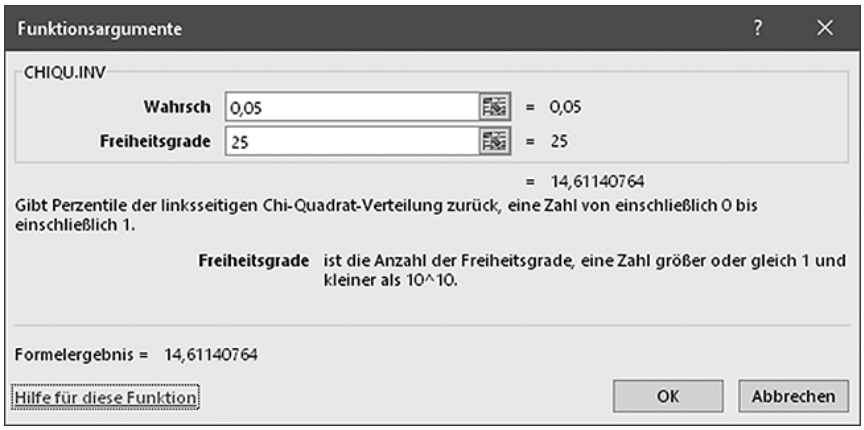


Abbildung 10.9: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion CHIQU.INV

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   In das Feld WAHRSCH habe ich **0,05** eingegeben. Dies ist die Wahrscheinlichkeit, an der ich in diesem Beispiel interessiert bin.

   In das Feld FREIHEITSGRADE habe ich die die Freiheitsgrade ein. Der Wert für Freiheitsgrade beträgt bei diesem Beispiel 25 (26 – 1). Nachdem ich den df-Wert eingegeben haben, wird im Dialogfeld der Wert (14,61140764) angezeigt, der die unteren 5 % der Fläche in dieser Chi-Quadrat-Verteilung abtrennt.

4. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion CHIQU.INV.RE ist mit dem für die Funktion CHIQU.INV identisch. Die RE-Version der Funktion gibt den Chi-Quadrat-Wert zurück, der die oberen 5 %

der Fläche der Wahrscheinlichkeit abtrennt. Dies ist nützlich, wenn Sie den Wert ermitteln wollen, den Sie überschreiten müssen, um H_0 zu verwerfen. Für dieses Beispiel habe ich als Argumente die Werte 0,05 und 25 eingegeben. Das Ergebnis der Funktion ist 37,65248. Der berechnete Wert, 36, verfehlt den Abschneidewert nur knapp. Aber knapp daneben ist auch vorbei und so können Sie H_0 nicht verwerfen.

Etwas zum Nachdenken

Betrachten Sie noch einmal das obige Beispiel. FarKlempt Robotics möchte nachweisen, dass der Herstellungsprozess des Unternehmens innerhalb akzeptabler Streuungsgrenzen liegt. Die Nullhypothese besagt tatsächlich, dass der Prozess akzeptabel ist. Die Daten liefern keinen Anhaltspunkt zum Verwerfen von H_0. Der Wert der Testgröße verfehlt den kritischen Wert nur knapp. Bedeutet das, dass die Werte für den Herstellungsprozess innerhalb akzeptabler Grenzen liegen?

Statistiken sind ein Hilfsmittel für den gesunden Menschenverstand, sie ersetzen ihn aber nicht. Wenn die Daten nur gerade mal in einem akzeptablen Bereich liegen, sollte das die Alarmglocken läuten lassen.

Normalerweise sollten Sie versuchen, H_0 zu verwerfen. Dies ist ein seltenes Beispiel, bei dem das Verwerfen von H_0 nicht erwünscht ist, weil das Nichtverwerfen etwas Positives impliziert: Der Herstellungsprozess verläuft ordnungsgemäß. Können in einer Situation wie dieser dennoch die Techniken der Hypothesentests verwendet werden?

Ja, aber mit einer wichtigen Änderung. Anstelle eines kleinen Werts für α (wie 0,05) wählen Sie einen großen Wert, zum Beispiel 0,20. Damit stehen die Chancen gut, dass die Nullhypothese tatsächlich zutrifft, wenn sie nicht verworfen wird. Denn bei einem großen Alpha können bereits kleine Testgrößenwerte zu einer Ablehnung von H_0 führen. Wenn α in diesem Beispiel 0,20 wäre, würde der kritische Wert 30,6752 betragen. (Sie können das mit `CHIQU.INV.RE` überprüfen.) Da der berechnete Wert 36 größer als der kritische Wert ist, muss H_0 bei diesem α verworfen werden.

Ein großer Wert für α wird nur selten verwendet. Dies ist jedoch sehr zu empfehlen, wenn H_0 *nicht* verworfen werden soll.

Eine Chi-Quadrat-Verteilung visualisieren

Damit Sie die Chi-Quadrat-Verteilungen besser verstehen können zeige ich Ihnen nun, wie Sie für diese Verteilung ein Diagramm erstellen können. Sie verwenden hierfür die Funktion `CHIQU.VERT` und für das Argument KUMULIERT den Wert `FALSCH`. Abbildung 10.10 zeigt die Zahlen und das fertige Diagramm.

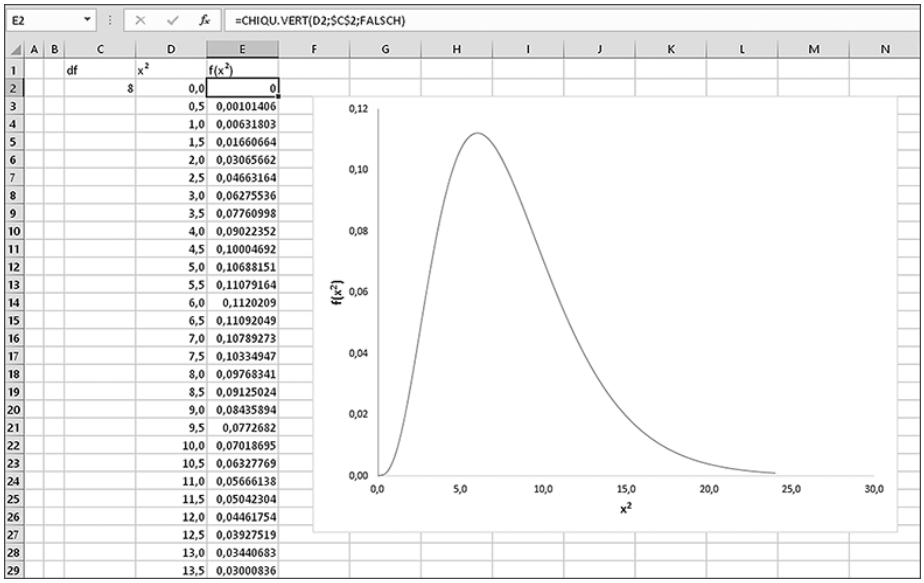


Abbildung 10.10: Visualisierung einer Chi-Quadrat-Verteilung

Hier die einzelnen Schritte:

1. **Geben Sie die Anzahl der Freiheitsgrade in eine Zelle ein.**

   Ich habe in Zelle C2 **8** eingegeben.

2. **Erstellen Sie eine Spalte mit Werten für die Statistik.**

   In den Zellbereich D2 bis D50 habe ich die Werte zwischen **0** und **24** in Schritten von 0,5 eingegeben.

3. **Geben Sie in die erste Zelle der benachbarten Spalte den Wert der Wahrscheinlichkeitsdichte für den ersten Datenwert ein.**

   Da ich eine Chi-Quadrat-Verteilung visualisieren will, verwende ich in Zelle E2 die Funktion `CHIQU.VERT`. Für den Wert von X klicke ich Zelle D2 an. Das Argument FREIHEITSGRADE steht in Zelle C2. Nachdem ich diese Zelle angeklickt habe, drücke ich F4, um einen absoluten Zellbezug zu erstellen. In das Feld KUMULIERT gebe ich `FALSCH` ein, um die Höhe der Verteilung für diesen Wert von χ^2 zu erhalten. Abschließend klicke ich auf OK.

4. **Füllen Sie Spalte E automatisch aus.**

5. **Erstellen Sie das Diagramm.**

   Markieren Sie dazu beide Spalten. Klicken Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME auf PUNKTE MIT INTERPOLIERTEN LINIEN.

6. **Bearbeiten Sie das Diagramm.**

   Ich klicke das das Diagramm an, damit die Schaltfläche DIAGRAMMELEMENTE angezeigt wird und dann die Achsentitel (χ^2 und $f(\chi^2)$) ergänzt. Ich entferne den Diagrammtitel und die Gitternetzlinien, aber das ist eine Frage des persönlichen Geschmacks. Außerdem habe ich das Diagramm vergrößert.

7. **Verändern Sie die Diagrammdaten.**

   Um ein besseres Gefühl für die Verteilung zu bekommen, können Sie den Wert *df* (die Anzahl der Freiheitsgrade) ändern und so sehen, wie sich dies auf das Diagramm auswirkt.

Kapitel 11

Zwei-Stichproben-Hypothesentest

In diesem Kapitel

- Differenzen zwischen Mittelwerten zweier Stichproben testen
- Mittelwerte abhängiger Stichproben testen
- Hypothesen zu Varianzen testen

Im Geschäftsleben, in der Ausbildung und in der wissenschaftlichen Forschung müssen häufig zwei Stichproben miteinander verglichen werden. Manchmal sind die Stichproben unabhängig voneinander und manchmal hängen sie irgendwie zusammen. Jede Stichprobe entstammt einer anderen Grundgesamtheit. Ziel ist es zu entscheiden, ob sich die Grundgesamtheiten, aus denen die Stichproben gezogen wurden, voneinander unterscheiden.

Dazu müssen in der Regel Hypothesentests zu den Mittelwerten der Grundgesamtheiten durchgeführt werden. Sie können auch Hypothesen zu Varianzen der Grundgesamtheiten testen. In diesem Kapitel zeige ich Ihnen, wie diese Tests durchgeführt werden. Ich werde außerdem nützliche Tabellenblattfunktionen und Datenanalysetools vorstellen, mit denen Sie diese Tests durchführen können.

Hypothesen für zwei

Wie in dem Fall mit nur einer Stichprobe (siehe Kapitel 10) beginnt der Hypothesentest mit zwei Stichproben ebenfalls mit einer

Nullhypothese (H_0) und einer Alternativhypothese (H_1). Die Nullhypothese besagt, dass der Unterschied zwischen zwei Stichproben ausschließlich dem Zufall zuzuschreiben ist. Die Alternativhypothese besagt, dass ein Unterschied echt und kein Zufallsprodukt ist.

Beim einseitigen Test gibt die Alternativhypothese die Richtung der Differenz zwischen den beiden Mittelwerten an, beim zweiseitigen Test tut sie das nicht.

Bei einem einseitigen Test lauten die Hypothesen folgendermaßen:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 > 0$

oder:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 < 0$

Bei einem zweiseitigen Test lauten die Hypothesen folgendermaßen:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Für diese Hypothesen wird üblicherweise die Null verwendet. Sie können jedoch auch jeden anderen Wert verwenden: Ersetzen Sie dazu einfach die Null durch den entsprechenden Wert.

Um den Test durchzuführen, legen Sie zunächst α, die Wahrscheinlichkeit eines Fehlers 1. Art, fest, mit dem Sie zu leben bereit sind (siehe Kapitel 9). Anschließend berechnen Sie den Mittelwert und die Standardabweichung der Stichproben, subtrahieren einen Mittelwert vom anderen und rechnen das Ergebnis mit einer Formel in eine Testgröße um. Anschließend vergleichen Sie diese Testgröße mit einer Stichprobenverteilung von Testgrößen. Wenn die Testgröße sich in dem durch α definierten kritischen Bereich befindet, verwerfen Sie H_0. Wenn nicht, dann verwerfen Sie H_0 nicht.

Noch einmal Stichprobenverteilungen

In Kapitel 9 wurde das Konzept der Stichprobenverteilung, einer Verteilung aller möglichen Werte einer Statistik, für einen bestimmten Stichprobenumfang und die Stichprobenverteilung des Mittelwerts beschrieben. In Kapitel 10 erfahren Sie, wie diese mit dem Ein-Stichproben-Hypothesentest zusammenhängt.

Für Zwei-Stichproben-Hypothesentests ist eine weitere Stichprobenverteilung erforderlich. Dabei handelt es sich um die *Stichprobenverteilung der Differenz zwischen Mittelwerten*.

Die Stichprobenverteilung der Differenz zwischen Mittelwerten ist die Verteilung aller möglichen Differenzen zwischen Stichprobenmittelwertepaaren, wobei der Stichprobenumfang von Paar zu Paar konstant gehalten wird. (Ja, da staunen Sie, was?) *Von Paar zu Paar konstant gehalten* bedeutet, dass aus der einen Grundgesamtheit immer dieselbe Anzahl Stichproben gezogen wird und aus der zweiten ebenfalls. Die beiden Stichprobenumfänge sind jedoch nicht unbedingt gleich groß.

Innerhalb der Paare entstammt jede Stichprobe aus einer anderen Grundgesamtheit. Alle Stichproben sind unabhängig voneinander, so dass sich die Auswahl von Objekten für die eine Stichprobe nicht auf die für die andere Stichprobe auswirkt.

Abbildung 11.1 zeigt, wie diese Stichprobenverteilung erstellt wird. Das werden Sie in der Realität niemals tun. Das ist reine Theorie. Wie Sie in der Abbildung sehen können, geht es darum, aus der einen sowie aus der anderen Grundgesamtheit je eine Stichprobe zu ziehen, die jeweiligen Mittelwerte zu berechnen und einen Mittelwert vom anderen zu subtrahieren. Geben Sie die Stichproben wieder in die Grundgesamtheiten zurück und wiederholen Sie den Vorgang

immer und immer wieder. Daraus ergibt sich eine Reihe von Differenzen der Mittelwerte. Diese Reihe von Differenzen stellt die Stichprobenverteilung dar.

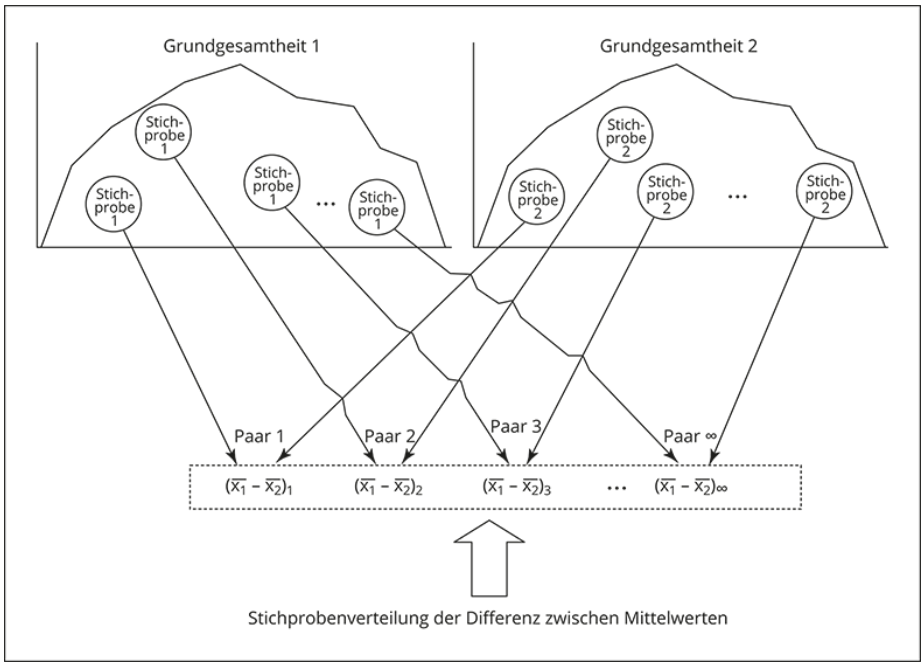


Abbildung 11.1: Erstellen einer Stichprobenverteilung der Differenz zwischen Mittelwerten

Den zentralen Grenzwertsatz anwenden

Wie jede andere Zahlengruppe hat auch diese Stichprobenverteilung einen Mittelwert und eine Standardabweichung. Und wie bei der Stichprobenverteilung des Mittelwerts (Kapitel 9 und 10) gilt auch hier der zentrale Grenzwertsatz.

Nach dem zentralen Grenzwertsatz gilt: Wenn die Stichproben groß sind, nähert sich die Stichprobenverteilung der Differenzen zwischen Mittelwerten einer Normalverteilung. Wenn die Grundgesamtheiten normalverteilt sind, stellt die Stichprobenverteilung eine Normalverteilung dar, auch wenn die Stichproben klein sind.

Der zentrale Grenzwertsatz sagt auch etwas über den Mittelwert und die Standardabweichung dieser Stichprobenverteilung aus. Nehmen

wir an, die Parameter für die erste Grundgesamtheit sind μ_1 und σ_1, und die Parameter für die zweite Grundgesamtheit sind μ_2 und σ_2. Der Mittelwert der Stichprobenverteilung lautet

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Die Standardabweichung der Stichprobenverteilung lautet

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

N_1 steht für die Anzahl der Objekte in der Stichprobe aus der ersten Grundgesamtheit. N_2 steht für die Anzahl der Objekte in der Stichprobe aus der zweiten Grundgesamtheit.

Diese Standardabweichung wird als *Standardfehler der Differenz zwischen Mittelwerten* bezeichnet.

In Abbildung 11.2 ist die Stichprobenverteilung zusammen mit den zugehörigen Parametern gemäß dem zentralen Grenzwertsatz dargestellt.

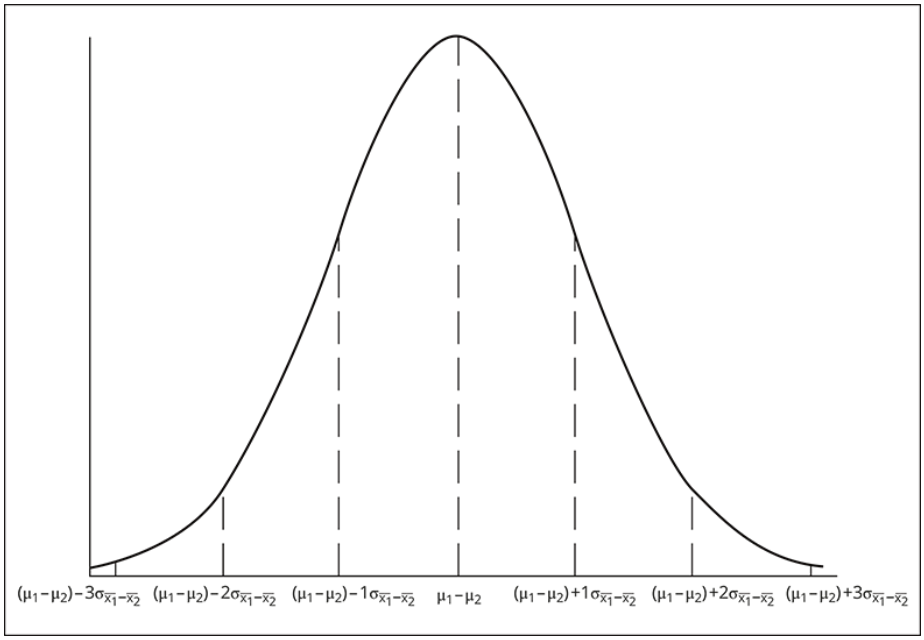


Abbildung 11.2: Die Stichprobenverteilung der Differenz zwischen Mittelwerten nach dem zentralen Grenzwertsatz

Noch einmal z-Werte

Da der zentrale Grenzwertsatz besagt, dass die Stichprobenverteilung bei großen Stichproben (oder bei kleinen Stichproben aus normalverteilten Grundgesamtheiten) nahezu normalverteilt ist, verwenden Sie die z-Werte als Testgröße. Statt »die z-Werte als Testgröße verwenden« kann man auch »einen z-Test durchführen« sagen. Die Formel hierfür lautet wie folgt:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

Der Ausdruck ($\mu_1 - \mu_2$) steht für die Differenz zwischen den Mittelwerten in H_0.

Mit dieser Formel wird die Differenz zwischen Stichprobenmittelwerten in einen Standardwert umgerechnet. Vergleichen Sie den Standardwert mit einer Standardnormalverteilung, das heißt einer Normalverteilung, für die gilt: $\mu = 0$ und $\sigma = 1$. Wenn sich der Wert in dem durch α definierten

kritischen Bereich befindet, wird H_0 verworfen. Wenn nicht, dann verwerfen Sie H_0 nicht.

Diese Formel wird verwendet, wenn die Werte von σ_1^2 und σ_2^2 bekannt sind.

Ein Beispiel: Stellen Sie sich eine neue Trainingsmethode vor, mit der sich der IQ verbessern lässt. Wenden Sie diese neue Methode bei einer Stichprobe von 25 Personen an. Eine andere Stichprobe von 25 Personen erhält kein spezielles Training. Nehmen wir an, es ergibt sich, dass der Stichprobenmittelwert für die Stichprobe mit der neuen Methode 107 und für die Stichprobe ohne Training 101,2 beträgt. Der Hypothesentest lautet wie folgt:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 > 0$

Für α lege ich 0,05 fest.

Für den IQ-Wert ist bekannt, dass die Standardabweichung 16 beträgt. Ich gehe davon aus, dass die Standardabweichung für die Grundgesamtheit der Personen, die nach der neuen Methode trainiert haben, dieselbe ist. Diese Grundgesamtheit gibt es natürlich nicht. Wir gehen nur davon aus; wenn es sie gäbe, müsste sie für die Standardabweichung denselben Wert aufweisen wie die normale Grundgesamtheit von IQ-Werten. Hat der Mittelwert dieser (theoretischen) Grundgesamtheit denselben Wert wie die normale Grundgesamtheit? H_0 sagt ja. H_1 sagt, er ist größer.

Die Testgröße beträgt:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(107 - 101{,}2)}{\sqrt{\frac{16^2}{25} + \frac{16^2}{25}}} = \frac{5{,}8}{4{,}53} = 1{,}28$$

Wenn α = 0,05 ist, beträgt der kritische Wert von *z*, also der Wert, der die oberen 5 % der Fläche unter der Standardnormalverteilung

abschneidet, 1,64. (Sie können das mit der Tabellenblattfunktion `NORM.S.INV` aus Kapitel 10 überprüfen.) Der berechnete Wert der Testgröße ist kleiner als der kritische Wert. H_0 wird also nicht verworfen. In Abbildung 11.3 ist das alles zusammengefasst.

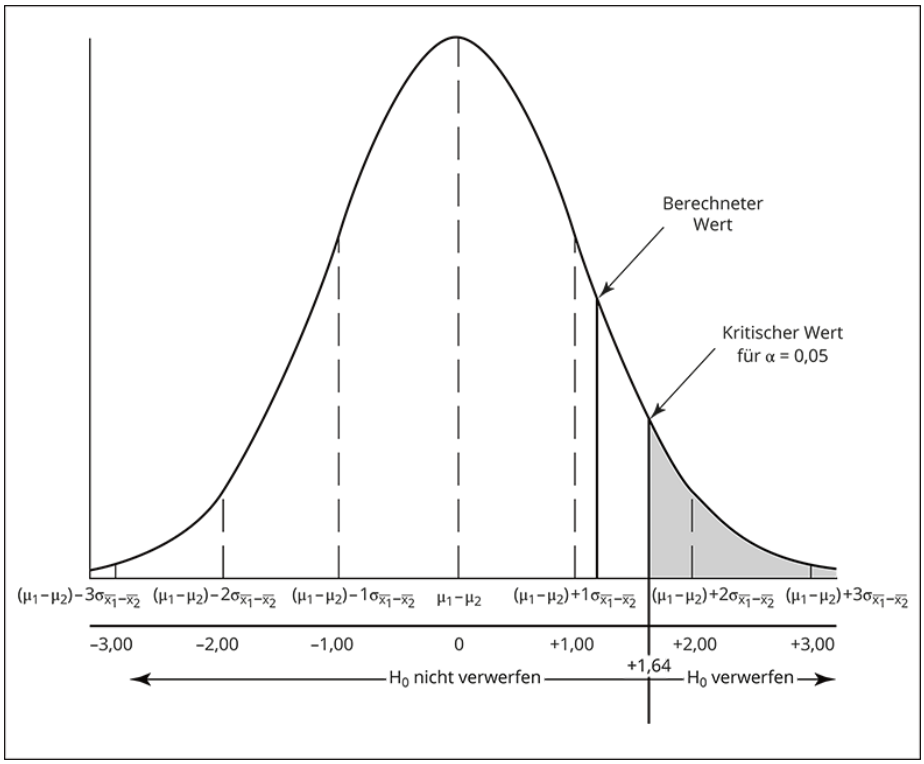
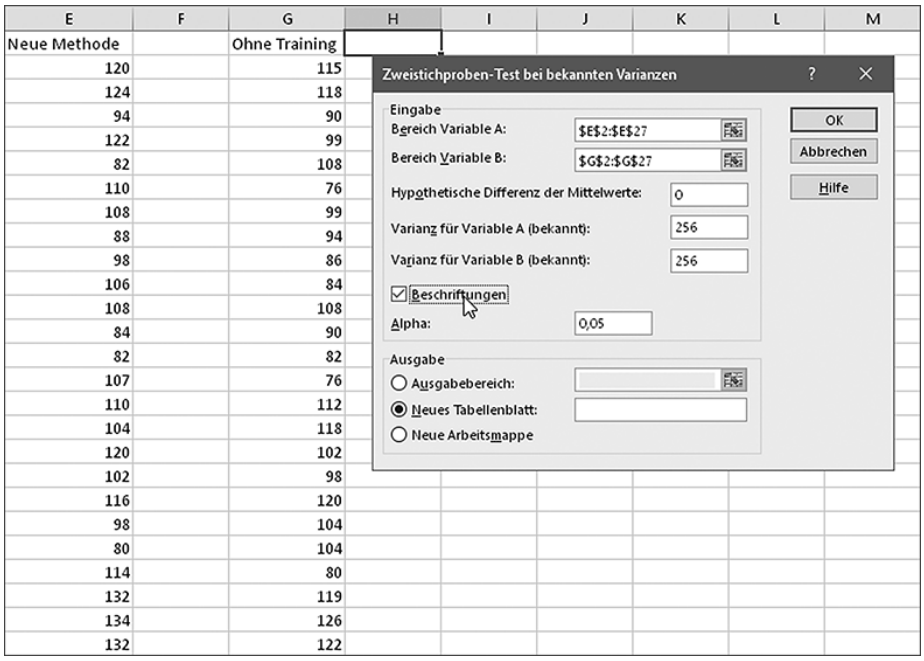


Abbildung 11.3: Die Stichprobenverteilung der Differenz zwischen Mittelwerten zusammen mit dem kritischen Wert für α = 0,05 und dem berechneten Wert der Testgröße im Beispiel mit dem IQ

Datenanalysetool: Zweistichproben-Test bei bekannten Varianzen

Excel enthält ein Datenanalysetool, das die Ausführung von Tests wie den im Beispiel mit dem IQ erleichtert. Dieses Tool trägt den Namen *Zweistichproben-Test bei bekannten Varianzen*. In Abbildung 11.4 ist das Dialogfeld für dieses Tool zusammen mit Stichprobendaten aus dem Beispiel mit dem IQ dargestellt.



E	F	G	H	I	J	K	L	M
Neue Methode		Ohne Training						
120		115						
124		118						
94		90						
122		99						
82		108						
110		76						
108		99						
88		94						
98		86						
106		84						
108		108						
84		90						
82		82						
107		76						
110		112						
104		118						
120		102						
102		98						
116		120						
98		104						
80		104						
114		80						
132		119						
134		126						
132		122						

Abbildung 11.4: Das Datenanalysetool für den Zweistichproben-Test bei bekannten Varianzen und Daten aus zwei Stichproben

Gehen Sie wie folgt vor, um dieses Tool zu verwenden:

1. **Geben Sie die Daten für die beiden Stichproben jeweils in einen eigenen Zellbereich ein.**

   Bei diesem Beispiel befinden sich die Daten aus der Stichprobe *Neue Methode* in Spalte E und die Daten für die Stichprobe *Ohne Training* in Spalte G.

2. **Wählen Sie den Befehl DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ZWEISTICHPROBEN-TEST BEI BEKANNTEN VARIANZEN aus.**

4. **Klicken Sie auf OK und öffnen Sie so das Dialogfeld ZWEISTICHPROBEN T-TEST BEI BEKANNTEN VARIANZEN (siehe Abbildung 11.4).**

5. **Geben Sie im Feld BEREICH VARIABLE A den Zellbereich ein, in dem sich die Daten für eine der Stichproben befinden.**

Die Daten für die Stichprobe *Neue Methode* befinden sich in unserem Beispiel in $E$2:$E$27. (Beachten Sie die Dollarzeichen [$] für den absoluten Bezug.)

6. **Geben Sie im Feld BEREICH VARIABLE B den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden.**

   Die Daten für die Stichprobe *Ohne Training* befinden sich in $G$2:$G$27.

7. **Geben Sie im Feld HYPOTHETISCHE DIFFERENZ DER MITTELWERTE die durch H_0 definierte Differenz zwischen μ_1 und μ_2 ein.**

   In diesem Beispiel beträgt diese Differenz 0.

8. **Geben Sie im Feld VARIANZ FÜR VARIABLE A (BEKANNT) die Varianz der ersten Stichprobe ein.**

   Die Standardabweichung der Grundgesamtheit der IQ-Werte beträgt 16. Somit ergibt sich für die Varianz der Wert 256.

9. **Geben Sie im Feld VARIANZ FÜR VARIABLE B (BEKANNT) die Varianz der zweiten Stichprobe ein.**

   In diesem Beispiel beträgt diese Varianz ebenfalls 256.

10. **Wenn der Zellbereich Spaltenüberschriften enthält, schalten Sie das Kontrollkästchen BESCHRIFTUNGEN ein.**

    Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich auch diese Option aktiviert.

11. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05.**

    In diesem Beispiel habe ich für α immer den Standardwert verwendet.

12. **Wählen Sie im Bereich AUSGABE ein Optionsfeld aus und legen Sie so fest, wo das Ergebnis ausgegeben werden**

soll.

Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

13. **Klicken Sie auf OK.**

    Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird ein neues Arbeitsblatt mit dem Ergebnis erstellt.

In Abbildung 11.5 habe ich das von diesem Tool berechnete Ergebnis mit erweiterten Spalten dargestellt. In den Zeilen 4, 5 und 7 befinden sich die Werte, die Sie im Dialogfeld eingegeben haben. In Zeile 6 steht die Anzahl der Werte in der jeweiligen Stichprobe.

	A	B	C
1	Zweistichproben-Test bei bekannten Varianzen, Gausstest		
2			
3		*Neue Methode*	*Ohne Training*
4	Mittelwert	107	101,2
5	Bekannte Varianz	256	256
6	Beobachtungen	25	25
7	Hypothetische Differenz der Mittelwerte	0	
8	z	1,281631041	
9	P(Z<=z) einseitig	0,099986053	
10	Kritischer z-Wert bei einseitigem Test	1,644853627	
11	P(Z<=z) zweiseitig	0,199972106	
12	Kritischer z-Wert bei zweiseitigem Test	1,959963985	
13			

Abbildung 11.5: Ergebnis des Datenanalysetools für den Zweistichproben-Test, Gausstest

Der Wert der Testgröße befindet sich in Zelle B8. Der kritische Wert für einen einseitigen Test befindet sich in Zelle B10, der kritische Wert für einen zweiseitigen Test in Zelle B12.

In Zelle B9 ist der Anteil der Fläche angegeben, den die Testgröße auf der einen Seite der Standardnormalverteilung abschneidet. In Zelle B11 wird das Doppelte dieses Werts angezeigt. Hierbei handelt es sich um den Anteil der Fläche, der vom positiven Wert der Testgröße (auf der rechten Seite der Verteilung) abgeschnitten wird,

plus dem Anteil, der vom negativen Wert der Testgröße (auf der linken Seite der Verteilung) abgeschnitten wird.

t-Test für zwei Stichproben

Das Beispiel im vorherigen Abschnitt beschreibt eine Situation, wie sie in der Realität praktisch nicht vorkommt: Beide Varianzen sind bekannt. Wenn die Varianz einer Grundgesamtheit bekannt ist, ist im Prinzip auch der Mittelwert der Grundgesamtheit bekannt. Wenn der Mittelwert bekannt ist, müssen dazu eigentlich keine Hypothesentests ausgeführt werden.

Wenn die Varianzen nicht bekannt sind, hilft der zentrale Grenzwertsatz nicht weiter. Das bedeutet, dass die Normalverteilung nicht als Näherung der Stichprobenverteilung der Differenz zwischen Mittelwerten verwendet werden kann. Stattdessen wird die *t*-Verteilung verwendet, die in Kapitel 9 vorgestellt und in Kapitel 10 auf einen Ein-Stichproben-Hypothesentest angewendet wurde. Die einzelnen *t*-Verteilungen unterscheiden sich im Hinblick auf den Parameter Freiheitsgrade (df) voneinander. Den df-Wert können Sie sich als den Nenner der Schätzung der Varianz vorstellen, die Sie zum Berechnen eines Werts von *t* als Testgröße verwenden. Statt »Berechnen eines Werts von *t* als Testgröße« kann man auch »Durchführen eines *t*-Tests« sagen.

Wenn die Varianzen der Grundgesamtheit nicht bekannt sind, gibt es zwei Möglichkeiten für Hypothesentests. Eine Möglichkeit besteht darin, dass Sie trotzdem Grund zu der Annahme haben, dass diese gleich sind. Die andere Möglichkeit ist, dass Sie eben nicht annehmen können, dass die Varianzen gleich sind. In den nachfolgenden Abschnitten beschreibe ich diese Möglichkeiten.

Wie ein Ei dem anderen: gleiche Varianzen

Wenn Sie die Varianz einer Grundgesamtheit nicht kennen, schätzen Sie diese mit der Stichprobenvarianz. Wenn Sie zwei Stichproben

vorliegen haben, bilden Sie sozusagen den Mittelwert der beiden Stichprobenvarianzen, um zu einer Schätzung zu gelangen.

Wenn Sie Stichprobenvarianzen zum Schätzen der Varianz einer Grundgesamtheit zusammenführen, wird dies als *Pooling* bezeichnet. Bei zwei Stichprobenvarianzen gehen Sie folgendermaßen vor:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

In dieser Formel steht s_p^2 für die gepoolte Schätzung. Der Nenner dieser Schätzung lautet (*n1* – 1) + (*n2* – 1). Ist das der df-Wert? Jawohl!

Die Formel zur Berechnung von *t* lautet wie folgt:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Ein Beispiel: FarKlempt Robotics muss sich zwischen zwei Maschinen zum Herstellen eines Bauteils für den neuen Mikroroboter entscheiden. Da Geschwindigkeit besonders wichtig ist, stellt jede Maschine zehn Bauteile her, und FarKlempt misst die Zeit für jeden Produktionslauf. Die Hypothesen lauten wie folgt:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2$

FarKlempt legt für α den Wert 0,05 fest. Hierbei handelt es sich um einen zweiseitigen Test, da das Unternehmen im Voraus nicht weiß, welche Maschine schneller ist.

In Tabelle 11.1 sind die Daten für die Produktionszeiten in Minuten aufgeführt.

	Maschine 1	Maschine 2
Mittlere Produktionszeit	23,00	20,00
Standardabweichung	2,71	2,79
Stichprobenumfang	10	10

Tabelle 11.1: Stichprobenwerte aus der FarKlempt-Maschinen-Studie

Der gepoolte Schätzwert von σ^2 lautet somit

$$\sigma_p^2 = \frac{(n_1-1)s_1^2+(n_s-1)s_2^2}{(n-1)+(n_2-1)} = \frac{(10-1)(2{,}71)^2+(10-1)(2{,}79)^2}{(10-1)+(10-1)}$$
$$= \frac{(9)(7{,}3441)+(9)(7{,}7841)}{9+9} = \frac{66{,}0969+70{,}0569}{18} = 7{,}56$$

Der Schätzwert von σ beträgt 2,75, die Quadratwurzel aus 7,56.

Die Testgröße beträgt

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(23-20)}{2{,}75\sqrt{\frac{1}{10}+\frac{1}{10}}} = \frac{3}{1{,}23} = 2{,}44$$

Für diese Testgröße ist df = 18 der Nenner des Varianzschätzwerts. Bei einer *t*-Verteilung mit 18 Freiheitsgraden beträgt der kritische Wert 2,10 für die rechte (obere) Seite und –2.10 für die linke (untere) Seite. Wenn Sie es nicht glauben, dann berechnen Sie die Werte mit `T.INV.2S` (siehe Kapitel 10). Der berechnete Wert der Testgröße ist größer als 2,10. H_0 wird also verworfen. Die Daten zeigen, dass Maschine 2 signifikant schneller ist als Maschine 1. (Sie können den Ausdruck *signifikant* immer dann verwenden, wenn Sie H_0 verwerfen.)

Wie Äpfel und Birnen: unterschiedliche Varianzen

Der Fall der unterschiedlichen Varianzen stellt eine Herausforderung dar. Wenn die Varianzen nicht gleich sind, ist die *t*-Verteilung mit $(n_1 - 1) + (n_2 - 1)$ keine so gute Annäherung an die Stichprobenverteilung, wie sie die Statistiker gerne hätten.

Statistiker begegnen dieser Herausforderung durch eine Reduzierung der Freiheitsgrade. Dazu verwenden sie eine ziemlich komplizierte Formel, die von den Standardabweichungen der Stichproben und dem jeweiligen Stichprobenumfang abhängt.

Da die Varianzen unterschiedlich sind, ist eine gepoolte Schätzung nicht anwendbar. Daher wird der *t*-Test anders berechnet:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Dabei wird die Testgröße mit einer *t*-Verteilung mit den geringeren Freiheitsgraden verglichen.

T.TEST

Dank der Tabellenblattfunktion `T.TEST` brauchen Sie sich nicht durch die ganzen Formeln für den *t*-Test durchzuarbeiten.

In Abbildung 11.6 sind die Daten für das Beispiel mit den FarKlempt-Maschinen von weiter oben dargestellt. In der Abbildung ist außerdem das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `T.TEST` dargestellt.

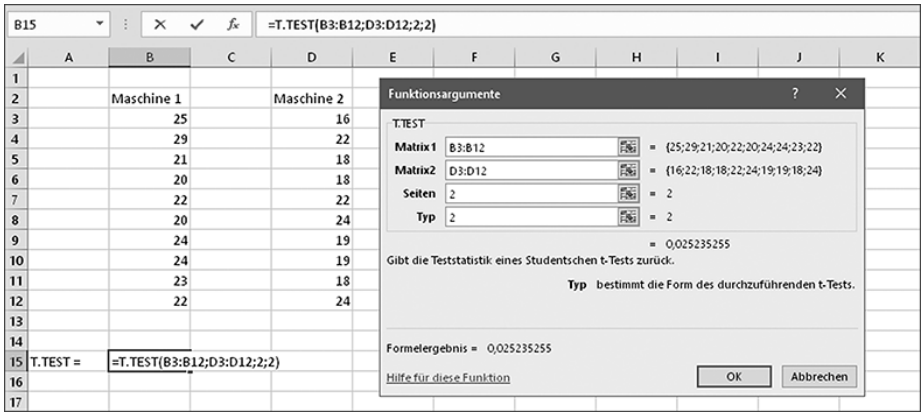


Abbildung 11.6: Die Funktion `T.TEST` verwenden

Gehen Sie dazu wie folgt vor:

1. **Geben Sie die Daten für die beiden Stichproben jeweils in einen Zellenbereich ein und markieren Sie die Zelle, die das Ergebnis erhalten soll.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe der *Maschine 1* in Spalte B und die Daten für die Stichprobe der *Maschine 2* in Spalte D.

2. **Wählen Sie im Menü mit den statistischen Funktionen T.TEST aus, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente ein.**

   Geben Sie im Feld MATRIX1 den Zellbereich ein, in dem sich die Daten für eine der Stichproben befinden. In diesem Beispiel befinden sich die Daten der Maschine 1 in den Zellen B3 bis B12.

   Geben Sie im Feld MATRIX2 den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden. Die Daten der Maschine 2 befinden sich in den Zellen D3 bis D12.

Im Feld SEITEN geben Sie an, ob es sich um einen einseitigen oder um einen zweiseitigen Test handelt. Da es sich in diesem Beispiel um einen zweiseitigen Test handelt, gebe ich eine **2** ein.

In das Feld TYP wird eine Zahl eingegeben, die den Typ des *t*-Tests angibt. Zur Auswahl stehen 1 für einen Test für abhängige Stichproben (den Sie im folgenden Abschnitt *n* kennenlernen werden), 2 für zwei Stichproben unter der Annahme gleicher Varianzen und 3 für zwei Stichproben unter der Annahme ungleicher Varianzen. Ich habe **2** eingegeben.

Sobald Sie in dieses Feld eine Zahl eingeben, wird im Dialogfeld die Wahrscheinlichkeit zum *t*-Wert für die Daten angezeigt. Es wird nicht, wie das Dialogfeld von Excel vermuten lässt, der Wert von *t* angezeigt.

4. **Klicken Sie auf OK, damit das Ergebnis in der markierten Zelle angezeigt wird.**

Der Wert im Dialogfeld in Abbildung 11.6 ist kleiner als 0,05. Somit wird H_0 verworfen.

Wenn Sie übrigens im Feld TYP **3** (für ungleiche Varianzen) eingeben, unterscheidet sich die Wahrscheinlichkeit geringfügig von der Wahrscheinlichkeit, die sich für den Test mit gleichen Varianzen ergibt. Der Unterschied ist nur geringfügig, weil die Stichprobenvarianzen nahezu gleich sind und der Stichprobenumfang jeweils identisch ist.

Datenanalysetool: Zweistichproben t-Test: Gleicher Varianzen

Excel stellt Datenanalysetools für *t*-Tests bereit. Ein Tool wird für Verteilungen mit gleichen Varianzen, das andere für Verteilungen mit ungleichen Varianzen verwendet. Wie Sie gleich sehen werden, erhalten Sie mit diesen Tools mehr Informationen als mit der Funktion `T.TEST`.

Hier ein Beispiel, in dem das Tool für den *t*-Test bei gleichen Varianzen auf die Daten aus dem Beispiel mit den FarKlempt-Maschinen angewendet wird. In Abbildung 11.7 sind die Daten mit dem Dialogfeld für den *t*-Test für zwei Stichproben bei gleichen Varianzen dargestellt.

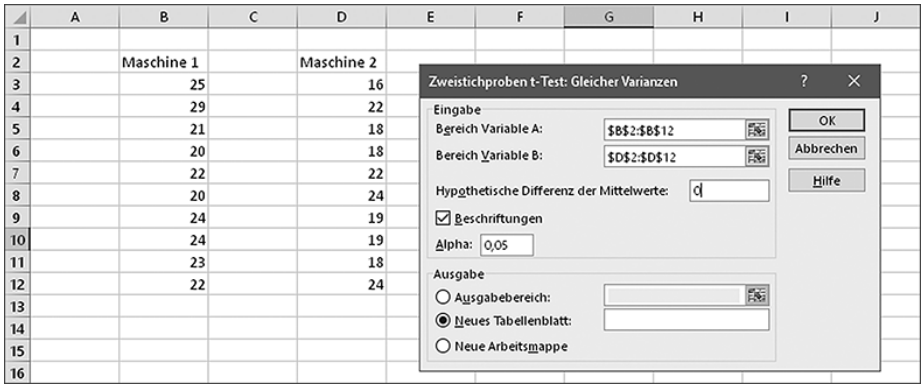


Abbildung 11.7: Das Datenanalysetool für den *t*-Test bei gleichen Varianzen und Daten aus zwei Stichproben

Gehen Sie wie folgt vor, um dieses Tool zu verwenden:

1. **Geben Sie die Daten für die beiden Stichproben jeweils in einen eigenen Zellbereich ein.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe der *Maschine* 1 in Spalte B und die Daten für die Stichprobe der *Maschine 2* in Spalte D.

2. **Wählen Sie den Befehl DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ZWEISTICHPROBEN-TEST: GLEICHER VARIANZEN aus.**

4. **Klicken Sie auf OK und öffnen Sie so das Dialogfeld ZWEISTICHPROBEN-TEST: GLEICHER VARIANZEN.**

   Das ist das Dialogfeld in Abbildung 11.7.

5. **Geben Sie im Feld BEREICH VARIABLE A den Zellbereich ein, in dem sich die Daten für eine der Stichproben befinden.**

   Die Daten für die Stichprobe *Maschine 1* befinden sich in unserem Beispiel in $B$2:$B$12. (Beachten Sie die Dollarzeichen [$] für den absoluten Bezug.)

6. **Geben Sie im Feld BEREICH VARIABLE B den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden.**

   Die Daten für die Stichprobe *Maschine 2* befinden sich in $D$2:$D$12.

7. **Geben Sie im Feld HYPOTHETISCHE DIFFERENZ DER MITTELWERTE die durch H_0 definierte Differenz zwischen μ_1 und μ_2 ein.**

   In diesem Beispiel beträgt diese Differenz 0.

8. **Wenn der Zellbereich Spaltenüberschriften enthält, schalten Sie das Kontrollkästchen BESCHRIFTUNGEN ein.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich auch diese Option aktiviert.

9. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05.**

   In diesem Beispiel habe ich für α immer den Standardwert verwendet.

10. **Wählen Sie im Bereich AUSGABE ein Optionsfeld aus und legen Sie so fest, wo das Ergebnis ausgegeben werden soll.**

    Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

11. **Klicken Sie auf OK.**

Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird ein neues Arbeitsblatt mit dem Ergebnis erstellt.

In Abbildung 11.8 ist das von diesem Tool berechnete Ergebnis mit erweiterten Spalten dargestellt. In den Zeilen 4 bis 7 befinden sich die Stichprobenwerte. Zelle B8 enthält die von H_0 definierte Differenz zwischen den Mittelwerten der Grundgesamtheit und B9 die Freiheitsgrade.

	A	B	C	D
1	Zweistichproben t-Test unter der Annahme gleicher Varianzen			
2				
3		*Maschine 1*	*Maschine 2*	
4	Mittelwert	23	20	
5	Varianz	7,333333333	7,777777778	
6	Beobachtungen	10	10	
7	Gepoolte Varianz	7,555555556		
8	Hypothetische Differenz der Mittelwerte	0		
9	Freiheitsgrade (df)	18		
10	t-Statistik	2,44046765		
11	P(T<=t) einseitig	0,012617628		
12	Kritischer t-Wert bei einseitigem t-Test	1,734063607		
13	P(T<=t) zweiseitig	0,025235255		
14	Kritischer t-Wert bei zweiseitigem t-Test	2,10092204		
15				

Abbildung 11.8: Ergebnis des Datenanalysetools für den *t*-Test unter der Annahme gleicher Varianzen

In den restlichen Zeilen befinden sich Informationen zum *t*-Wert. Der berechnete Wert der Testgröße befindet sich in Zelle B10. Zelle B11 nennt den Anteil der Fläche, den der positive Wert der Testgröße an der oberen Seite der *t*-Verteilung mit dem angegebenen df-Wert abtrennt. Zelle B12 gibt den kritischen Wert für einen einseitigen Test an: Das ist der Wert, der den Anteil der Fläche an der oberen Seite abschneidet, der α entspricht.

In Zelle B13 wird das Doppelte des Anteils in B11 angezeigt. In dieser Zelle befindet sich der Anteil der Fläche von B11, der zum Anteil der Fläche addiert wird, den der negative Wert der Testgröße an der unteren Seite abschneidet. Zelle B14 gibt den kritischen Wert für einen zweiseitigen Test an: Hierbei handelt es sich um den positiven Wert, der α/2 an der oberen Seite abschneidet. Der entsprechende negative Wert (hier nicht dargestellt) schneidet α/2 an der unteren Seite ab.

Die Stichproben in meinen Beispielen haben dieselbe Anzahl an Werten und in etwa gleiche Varianzen. Wenn wir auf diese Daten das Datenanalysetool für den *t*-Test bei ungleichen Varianzen anwenden würden, würden sich die Ergebnisse kaum voneinander unterscheiden.

So habe ich ein neues Beispiel erstellt, das in Tabelle 11.2 zusammengefasst ist. Die Stichproben in diesem Beispiel haben jeweils einen anderen Umfang und deutlich unterschiedliche Varianzen.

	Stichprobe 1	Stichprobe 2
Mittelwert	52,50	41,33
Standardabweichung	499,71	41,87
Stichprobenumfang	8	6

Tabelle 11.2: Stichprobenwerte für einen *t*-Test bei ungleichen Varianzen

Damit Sie sehen, wie sich das Datenanalysetool für den *t*-Test bei gleichen Varianzen von dem bei unterschiedlichen Varianzen unterscheidet, habe ich beide Tools auf diese Daten angewendet und die Ergebnisse nebeneinandergestellt. In Abbildung 11.9 sind die beiden Ergebnisse zu sehen. Um das Datenanalysetool für unterschiedliche Varianzen auszuführen, gehen Sie genauso vor wie beim Tool für gleiche Varianzen, jedoch mit einer Ausnahme: Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ZWEISTICHPROBEN T-TEST: UNTERSCHIEDLICHER VARIANZEN aus.

	A	B	C	D	E	F	G
1	Zweistichproben t-Test unter der Annahme unterschiedlicher Varianzen				Zweistichproben t-Test unter der Annahme gleicher Varianzen		
2							
3		*Maschine 1*	*Maschine 2*			*Maschine 1*	*Maschine 2*
4	Mittelwert	23	20		Mittelwert	23	20
5	Varianz	7,333333333	7,777777778		Varianz	7,333333333	7,777777778
6	Beobachtungen	10	10		Beobachtungen	10	10
7	Hypothetische Differenz der Mittelwerte	0			Gepoolte Varianz	7,555555556	
8	Freiheitsgrade (df)	18			Hypothetische Differenz der Mittelwerte	0	
9	t-Statistik	2,44046765			Freiheitsgrade (df)	18	
10	P(T<=t) einseitig	0,012617628			t-Statistik	2,44046765	
11	Kritischer t-Wert bei einseitigem t-Test	1,734063607			P(T<=t) einseitig	0,012617628	
12	P(T<=t) zweiseitig	0,025235255			Kritischer t-Wert bei einseitigem t-Test	1,734063607	
13	Kritischer t-Wert bei zweiseitigem t-Test	2,10092204			P(T<=t) zweiseitig	0,025235255	
14					Kritischer t-Wert bei zweiseitigem t-Test	2,10092204	

Abbildung 11.9: Ergebnisse aus der Berechnung des Datenanalysetools für den *t*-Test bei gleichen Varianzen und für den *t*-Test bei unterschiedlichen Varianzen für die in Tabelle 11.2 zusammengefassten Daten

In Abbildung 11.9 ist ein offensichtlicher Unterschied zwischen den beiden Datenanalysetools zu erkennen: Das Datenanalysetool für unterschiedliche Varianzen ermittelt keinen gepoolten Schätzwert für σ^2, weil dieser für den *t*-Test in diesem Fall nicht verwendet wird. Ein weiterer Unterschied stellt der df-Wert dar. Wie bereits erwähnt, verwenden wir bei ungleichen Varianzen entsprechend der Stichprobenvarianzen und dem jeweiligen Stichprobenumfang kleinere Freiheitsgrade. Bei dem Beispiel mit den gleichen Varianzen beträgt der df-Wert 12, bei dem Beispiel mit den ungleichen Varianzen beträgt er 10.

Die Auswirkungen dieser Unterschiede zeigen sich in den restlichen Werten. Die *t*-Werte, die kritischen Werte und die Wahrscheinlichkeiten sind jeweils unterschiedlich.

Ein passendes Paar: Hypothesentest für abhängige Stichproben

Bei den bisher beschriebenen Hypothesentests sind die Stichproben unabhängig voneinander. Die Auswahl eines Objekts für die eine Stichprobe hat keine Auswirkungen auf die Auswahl eines Objekts für die andere Stichprobe.

Manchmal hängen Stichproben jedoch zusammen. Ein sehr anschauliches Beispiel hierfür ist der Fall, bei dem dasselbe Objekt unter zwei Bedingungen je einen Wert liefert, wie das bei Vorher-Nachher-Untersuchungen der Fall ist. Stellen Sie sich beispielsweise Personen vor, die an einem Diätprogramm teilnehmen. Diese werden vor Beginn des Programms und zum Beispiel nach einem Monat gewogen. Die wichtigen Daten sind die Werte der Vorher-Nachher-Differenz. In Tabelle 11.3 sind diese Daten in Pfund aufgeführt.

Person	Gewicht vor Diätbeginn	Gewicht nach einen Monat	Differenz
1	198	194	4
2	201	203	–2
3	210	200	10
4	185	183	2
5	204	200	4
6	156	153	3
7	167	166	1
8	197	197	0
9	220	215	5
10	186	184	2
Mittelwert			2,9
Standardabweichung			3,25

Tabelle 11.3: Daten für das Diät-Beispiel

Hierbei geht es darum, sich diese Differenzen als eine Stichprobe von Werten vorzustellen und wie in einem *t*-Test mit einer Stichprobe (siehe Kapitel 10) zu behandeln.

Sie testen die folgenden Hypothesen:

H_0: $\mu_d \leq 0$

H_1: $\mu_d > 0$

Das tiefgestellte d steht für *D*ifferenz. Legen Sie α = 0,05 fest.

Die Formel für diesen *t*-Test lautet wie folgt:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

Bei dieser Formel ist $\bar{d}$ der Mittelwert der Differenzen. Um $s_{\bar{d}}$ zu ermitteln, berechnen Sie die Standardabweichung der Differenzen

und dividieren durch die Quadratwurzel der Anzahl der Paare:

$$s_d = \frac{s}{\sqrt{n}}$$

Der df-Wert beträgt *N* – 1.

Aus Tabelle 11.3 ergibt sich:

$$t = \frac{\bar{d} - \mu_d}{s_d} = \frac{2{,}9}{\frac{3{,}25}{\sqrt{10}}} = 2{,}82$$

Für df = 9 (Anzahl der Paare − 1) und α = 0,05 beträgt der kritische Wert 2,26. (Sie können das mit `T.INV` überprüfen.) Der berechnete Wert ist größer als dieser Wert. Somit wird H_0 verworfen.

Wenn Sie sich diesen Test ansehen und denken, »Hmm – das sieht doch so aus wie ein Einstichproben-*t*-Test, nur dass diese eine Stichprobe die Differenz von Wertepaarten enthält«, dann sind Sie genau auf dem richtigen Weg.

T.TEST für abhängige Stichproben

Weiter vorn habe ich die Tabellenblattfunktion `T.TEST` beschrieben und Ihnen gezeigt, wie Sie sie auf unabhängige Stichproben anwenden können. Hier werde ich diese Funktion für die abhängigen Stichproben aus dem Diätbeispiel verwenden. In Abbildung 11.10 sind die Daten aus dem Diätbeispiel und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `T.TEST` dargestellt.

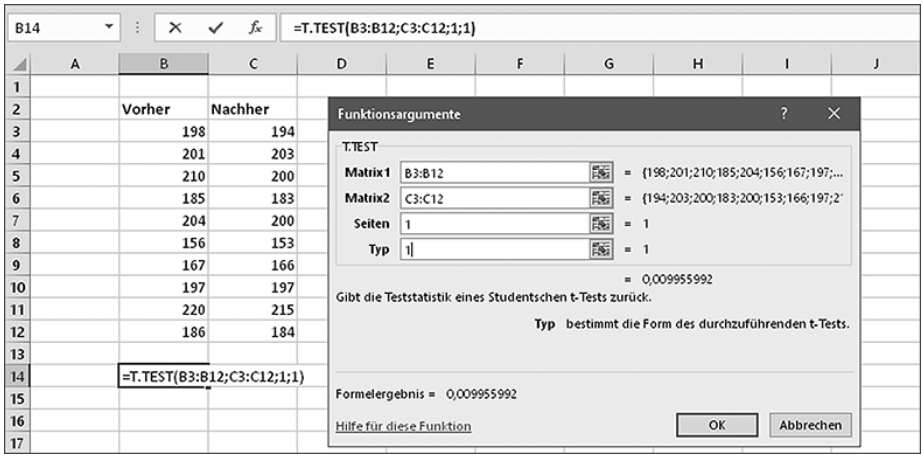


Abbildung 11.10: Das Dialogfeld FUNKTIONSARGUMENTE für `T.TEST` mit abhängigen Daten

Und so müssen Sie vorgehen:

1. **Geben Sie die Daten für die beiden Stichproben in jeweils einen eigenen Zellbereich ein und markieren Sie die Zelle, in die das Ergebnis soll.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe *Vorher* in Spalte B und die Daten für die Stichprobe *Nachher* in Spalte C.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion T.TEST aus, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte der Argumente ein.**

   Geben Sie im Feld MATRIX1 den Zellbereich ein, in dem sich die Daten für eine der Stichproben befinden. In diesem Beispiel befinden sich die Daten für *Vorher* in den Zellen B3 bis B12.

   Geben Sie im Feld MATRIX2 den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden. Die Daten für *Nachher* befinden sich in den Zellen C3 bis C12.

Geben Sie im Feld SEITEN für einen einseitigen Test eine 1 oder für einen zweiseitigen Test eine 2 ein. In diesem Beispiel handelt es sich um einen einseitigen Test, daher habe ich **1** eingegeben.

Im Feld TYP geben Sie eine Zahl ein, die den *t*-Testtyp angibt. Zur Auswahl stehen 1 für einen Test für abhängige Stichproben, 2 für zwei Stichproben bei gleichen Varianzen und 3 für zwei Stichproben mit ungleichen Varianzen. Ich habe **1** eingegeben.

Sobald Sie in dieses Feld eine Zahl eingeben, wird im Dialogfeld die Wahrscheinlichkeit zum *t*-Wert für die Daten angezeigt. Es wird nicht, wie das Dialogfeld von Excel vermuten lässt, der Wert von *t* angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die markierte Zelle einzufügen.**

   Der Wert im Dialogfeld in Abbildung 11.10 ist kleiner als 0,05. Somit wird H_0 verworfen.

Hätte ich die Spaltenüberschriften als Namen für die entsprechenden Zellbereiche definiert, würde die Formel in der Bearbeitungsleiste so aussehen:

```
=T.TEST(Vorher;Nachher;1;1)
```

Dieses Format ist einfacher zu verstehen, wenn Sie das Arbeitsblatt jemandem zeigen wollen. (Falls Sie nicht mehr genau wissen, wie Sie einen Namen für einen Zellbereich festlegen, werfen Sie einen Blick in Kapitel 2.)

Datenanalysetool: Zweistichproben t-Test bei abhängigen Stichproben

Excel stellt ein Datenanalysetool bereit, das sich um abhängige Stichproben kümmert. Dieses Tool trägt den Namen *Zweistichproben t-Test bei abhängigen Stichproben*. In diesem Abschnitt wird es für die Diätdaten verwendet.

In Abbildung 11.11 sind die Daten mit dem Dialogfeld für den *t*-Test bei zwei abhängigen Stichproben dargestellt.

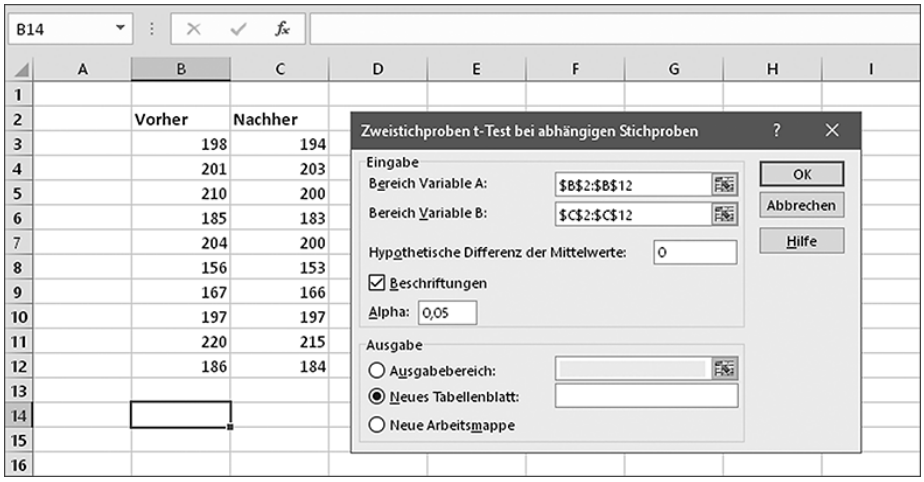


Abbildung 11.11: Das Datenanalysetool ZWEISTICHPROBEN T-TEST BEI ABHÄNGIGEN STICHPROBEN und Daten aus Stichprobenpaaren

Und so müssen Sie vorgehen:

1. **Geben Sie die Daten für die beiden Stichproben in jeweils einen eigenen Zellbereich ein.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe *Vorher* in Spalte B und die Daten für die Stichprobe *Nachher* in Spalte C.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ZWEISTICHPROBEN T-TEST BEI ABHÄNGIGEN STICHPROBEN aus.**

4. **Klicken Sie auf OK, um das Dialogfeld für dieses Tool zu öffnen.**

   In Abbildung 11.11 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld BEREICH VARIABLE A den Zellbereich ein, in dem sich die Daten für eine der Stichproben befinden.**

In diesem Beispiel befinden sich die Daten für *Vorher* in den Zellen $B$2:$B$12. (Beachten Sie die Dollarzeichen $ für den absoluten Bezug.)

6. **Geben Sie im Feld BEREICH VARIABLE B den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden.**

   Die Daten für *Nachher* befinden sich in den Zellen $C$2:$C$12.

7. **Geben Sie im Feld HYPOTHETISCHE DIFFERENZ DER MITTELWERTE die durch H_0 definierte Differenz zwischen μ_1 und μ_2 ein.**

   In diesem Beispiel beträgt diese Differenz 0.

8. **Wenn die Zellbereiche Spaltenüberschriften enthalten, schalten Sie das Kontrollkästchen BESCHRIFTUNGEN ein.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich dieses Kontrollkästchen aktiviert.

9. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie ihn, falls Sie für Alpha einen anderen Wert verwenden wollen.**

10. **Legen Sie im Bereich AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

    Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Blatt in der aktuellen Arbeitsmappe angezeigt wird.

11. **Klicken Sie auf OK.**

    Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird ein neues Blatt mit dem Ergebnis erstellt.

In Abbildung 11.12 habe ich das von diesem Tool berechnete Ergebnis mit erweiterten Spalten dargestellt. In den Zeilen 4 bis 7

befinden sich die Stichprobenwerte. Das einzig neue Element ist die Zahl in Zelle B7, der Korrelationskoeffizient nach Pearson. Hierbei handelt es sich um eine Zahl zwischen −1 und +1, die angibt, wie stark die Daten der ersten Stichprobe mit den Daten der zweiten zusammenhängen. Wenn diese Zahl (wie in diesem Beispiel) nahe 1 ist, werden hohe Werte in der einen Stichprobe mit hohen Werten in der anderen Stichprobe und niedrige Werte in der einen Stichprobe mit niedrigen Werten in der anderen verknüpft. Wenn die Zahl nahe −1 ist, werden hohe Werte in der einen Stichprobe mit niedrigen Werten in der anderen Stichprobe und niedrige Werte in der einen Stichprobe mit hohen Werten in der anderen verknüpft. Wenn die Zahl nahe null ist, werden die Werte in der ersten Stichprobe nicht mit denen in der anderen Stichprobe verknüpft. Da die beiden Stichproben aus Werten derselben Personen bestehen, wird ein hoher Wert erwartet. (Dieses Thema wird in Kapitel 15 ausführlicher beschrieben.)

Zelle B8 enthält die von H_0 definierte Differenz zwischen den Mittelwerten der Grundgesamtheit und B9 die Freiheitsgrade.

In den restlichen Zeilen befinden sich Informationen zum *t*-Wert. Der berechnete Wert der Testgröße befindet sich in Zelle B10. Zelle B11 zeigt den Anteil der Fläche, den der positive Wert der Testgröße an der oberen Seite der *t*-Verteilung mit dem angegebenen df-Wert abschneidet. Zelle B12 gibt den kritischen Wert für einen einseitigen Test an: Das ist der Wert, der den Anteil der Fläche an der oberen Seite abschneidet, der α entspricht.

In Zelle B13 wird das Doppelte des Anteils in B11 angezeigt. In dieser Zelle befindet sich der Flächenanteil von B11, zu dem der Flächenanteil addiert wird, den der negative Wert der Testgröße an der unteren Seite abschneidet. Zelle B13 gibt den kritischen Wert für einen zweiseitigen Test an: Hierbei handelt es sich um den positiven Wert, der α/2 an der oberen Seite abschneidet. Der entsprechende negative Wert (hier nicht dargestellt) schneidet α/2 an der unteren Seite ab.

	A	B	C	D
1	Zweistichproben t-Test bei abhängigen Stichproben (Paarvergleichstest)			
2				
3		*Vorher*	*Nachher*	
4	Mittelwert	192,4	189,5	
5	Varianz	377,6	342,9444444	
6	Beobachtungen	10	10	
7	Pearson Korrelation	0,986507688		
8	Hypothetische Differenz der Mittelwerte	0		
9	Freiheitsgrade (df)	9		
10	t-Statistik	2,824139508		
11	P(T<=t) einseitig	0,009955992		
12	Kritischer t-Wert bei einseitigem t-Test	1,833112933		
13	P(T<=t) zweiseitig	0,019911984		
14	Kritischer t-Wert bei zweiseitigem t-Test	2,262157163		
15				

Abbildung 11.12: Ergebnis des Datenanalysetools für den *t*-Test bei zwei abhängigen Stichproben

Zwei Varianzen testen

Die bisher beschriebenen Zwei-Stichproben-Hypothesentests bezogen sich auf Mittelwerte. Es ist jedoch auch möglich, Hypothesen zu Varianzen zu testen.

In diesem Abschnitt erweitere ich das Beispiel mit dem Herstellungsprozess aus Kapitel 10 um eine Varianz. Das Unternehmen FarKlempt Robotics stellt ein Bauteil her, das eine bestimmte Länge haben muss und nur eine sehr geringe Streuung aufweisen darf. Im Unternehmen wird über zwei Maschinen nachgedacht, die dieses Bauteil herstellen und von denen diejenige mit der geringsten Streuung ausgewählt werden soll. FarKlempt zieht eine Stichprobe von Bauteilen von beiden Maschinen, misst diese, ermittelt die Varianz für die Stichproben und führt einen Hypothesentest durch, um festzustellen, ob sich die Varianz der einen Maschine signifikant von der Varianz der anderen unterscheidet.

Die Hypothesen lauten wie folgt:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Wie immer ist α ein Muss. Und wie gewöhnlich lege ich den Wert auf 0,05 fest.

Beim Testen zweier Varianzen wird nicht die eine von der anderen subtrahiert. Stattdessen wird zum Ermitteln der Testgröße die eine Varianz durch die andere dividiert. Sir Ronald Fisher ist ein berühmter Statistiker, der die mathematischen Grundlagen und die Verteilungen für diese Art von Arbeit mit Varianzen entwickelt hat. Die Testgröße wurde ihm zu Ehren benannt. Sie wird als *F*-Quotient und der Test als *F*-Test bezeichnet. Die Verteilung für diesen Test wird als *F*-Verteilung bezeichnet.

Um nicht zu tief in die Mathematik vorzudringen, sei hier nur so viel gesagt: df ist auch hier der Parameter, der für den Unterschied bei dieser Verteilung zuständig ist. Das Besondere an diesen Verteilungen ist, dass es hier um zwei Varianzschätzer geht, das heißt, jede Verteilung ist mit zwei df-Werten verknüpft (und nicht nur mit einem wie beim *t*-Test). Die *F*-Verteilung unterscheidet sich außerdem dadurch von den anderen, die Sie bisher kennengelernt haben, dass *F* keinen negativen Wert annehmen kann. In Abbildung 11.13 sind zwei *F*-Verteilungen dargestellt.

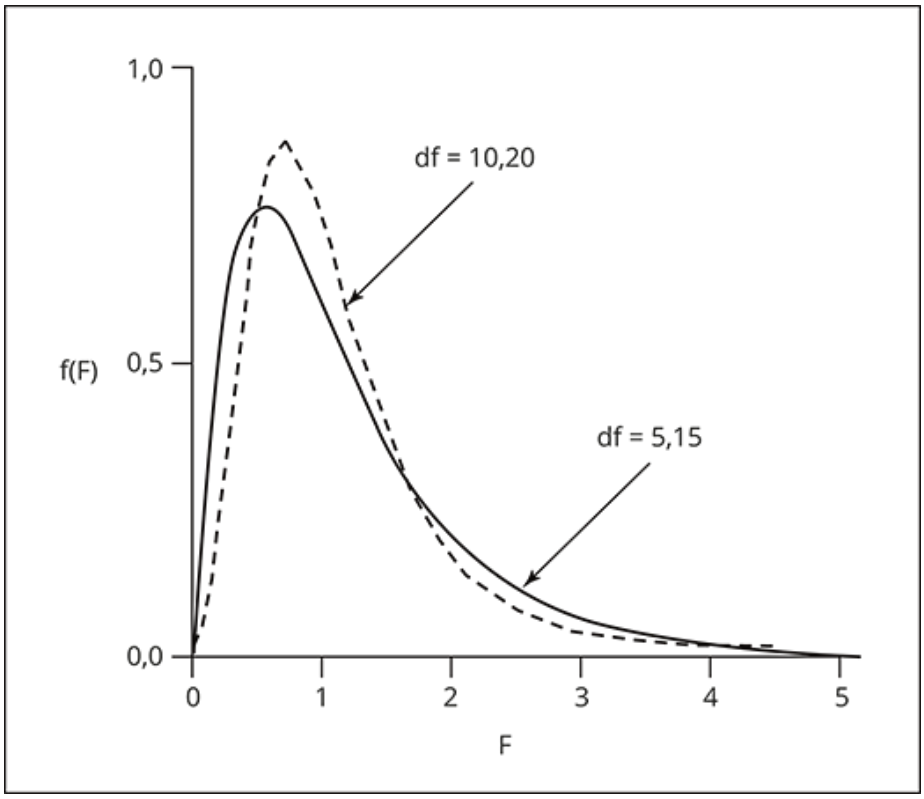


Abbildung 11.13: Zwei *F*-Verteilungen

Die Testgröße lautet:

$$F = \frac{\textit{größere } s^2}{\textit{kleinere } s^2}$$

Nehmen wir an, FarKlempt Robotics stellt 10 Bauteile mit Maschine 1 her und ermittelt eine Stichprobenvarianz von 0,60 cm^2. Mit Maschine 2 werden 15 Bauteile hergestellt und eine Stichprobenvarianz von 0,44 cm^2 ermittelt. Kann das Unternehmen H_0 verwerfen?

Berechnen wir die Testgröße:

$$F = \frac{0,60}{0,44} = 1,36$$

Die df-Werte betragen 9 und 14. Der Schätzwert der Varianz im Zähler des *F*-Quotienten beruht auf 10 Fällen und der Schätzwert

der Varianz im Nenner auf 15 Fällen.

Wenn die df-Werte 9 und 14 betragen und es sich um einen zweiseitigen Test bei α = 0,05 handelt, beträgt der kritische Wert von *F* 3,21. (Ich werde Ihnen gleich eine Excel-Funktion zeigen, mit der Sie diesen Wert ermitteln können.) Der berechnete Wert ist kleiner als der kritische Wert. H_0 wird also nicht verworfen.

Es ist ein Unterschied, welcher df-Wert im Zähler und welcher im Nenner steht. Die *F*-Verteilung für df = 9 und df = 14 ist eine andere als die *F*-Verteilung für df = 14 und df = 9. Der kritische Wert im letzten Fall beträgt beispielsweise 3,98 und nicht 3,21.

F zusammen mit t verwenden

Die *F*-Verteilung kann beispielsweise zusammen mit dem *t*-Test für unabhängige Stichproben verwendet werden. Vor dem Durchführen des *t*-Tests entscheiden Sie mit der *F*-Verteilung, ob in den Stichproben gleiche oder unterschiedliche Varianzen angenommen werden können.

Mit dem Beispiel zum *t*-Test mit gleichen Varianzen habe ich Ihnen bereits gezeigt, dass die Standardabweichungen 2,71 und 2,79 betragen. Die Varianzen betragen 7,34 und 7,78. Der *F*-Quotient dieser Varianzen berechnet sich wie folgt:

$$F = \frac{7,78}{7,34} = 1,06$$

Die beiden Stichproben beruhen auf je 10 Beobachtungen. Damit ergibt sich für die Stichprobenvarianzen jeweils df = 9. Ein *F*-Quotient von 1,06 schneidet die oberen 47 % der *F*-Verteilung ab, deren df-Werte jeweils 9 betragen. Somit kann für diese Daten die Version des *t*-Tests mit gleichen Varianzen verwendet werden.

Im Abschnitt *Etwas zum Nachdenken* am Ende von Kapitel 10 habe ich erwähnt, dass es in seltenen Fällen angebracht ist, einen hohen Wert für α zu verwenden. Wenn H_0 ein wünschenswertes Ergebnis ist und möglichst nicht verworfen werden soll, dann sichern Sie sich ab, indem Sie Alpha auf ein so hohes Niveau setzen, dass schon geringe Differenzen zu einem Verwerfen von H_0 führen. Wenn Sie H_0 dann dennoch nicht verwerfen müssen, können Sie sich recht sicher sein, dass H_0 tatsächlich zutrifft.

Hier haben wir einer dieser seltenen Fälle. Es ist wünschenswerter, den *t*-Test bei gleichen Varianzen durchzuführen, was in der Regel einen größeren Wert für die Freiheitsgrade ergibt als der *t*-Test bei unterschiedlichen Varianzen. Wenn Sie beim *F*-Test für α einen hohen Wert (0,20 ist ein guter Wert) festlegen, können Sie ruhigen Gewissens gleiche Varianzen annehmen.

F.TEST

Die Tabellenblattfunktion `F.TEST` berechnet anhand der Daten aus zwei Stichproben einen *F*-Quotienten. Sie gibt jedoch nicht den *F*-Quotienten an. Vielmehr gibt diese Funktion die zweiseitige Wahrscheinlichkeit des berechneten *F*-Quotienten unter H_0 an. Das bedeutet, dass das Ergebnis den Anteil der Fläche rechts des *F*-Quotienten und links des reziproken *F*-Quotienten (1 dividiert durch den *F*-Quotienten) angibt.

In Abbildung 11.14 sind die eben genannten Daten für das Beispiel mit den FarKlempt-Maschinen dargestellt. In der Abbildung ist außerdem das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `F.TEST` zu sehen.

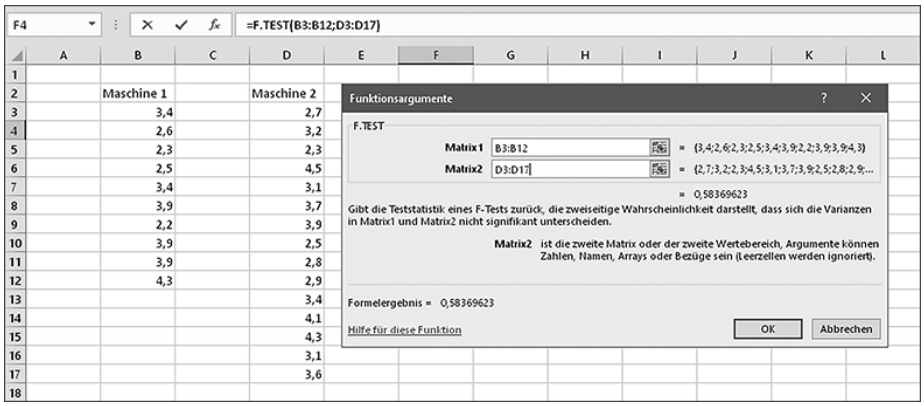


Abbildung 11.14: Die Funktion `F.TEST` verwenden

Gehen Sie dazu wie folgt vor:

1. **Geben Sie die Daten für die beiden Stichproben jeweils in einen eigenen Zellbereich ein und markieren Sie die Zelle für das Ergebnis.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe der *Maschine 1* in Spalte B und die Daten für die Stichprobe der MASCHINE 2 in Spalte D.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf F.TEST, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion `F.TEST` ein.**

   Geben Sie im Feld MATRIX1 den Zellbereich ein, in dem sich die Daten für die Stichprobe mit der größeren Varianz befinden. In diesem Beispiel befinden sich die Daten der Maschine 1 in den Zellen B3 bis B12.

   Geben Sie im Feld MATRIX2 den Zellbereich ein, in dem sich die Daten für die andere Stichprobe befinden. Die Daten der Maschine 2 befinden sich in den Zellen D3 bis D17.

Wenn Sie die Daten in diesem Feld eingegeben haben, wird das Ergebnis im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die markierte Zelle einzufügen.**

Der Wert im Dialogfeld (siehe Abbildung 11.14) ist größer als 0,05. Somit wird H_0 nicht verworfen. In Abbildung 11.15 ist die durch das Ergebnis angegebene Fläche dargestellt.

Hätte ich Namen für die beiden Zellbereichen definiert, würde die Formel in der Bearbeitungsleiste so aussehen:

```
=F.TEST(Maschine_1;Maschine_2)
```

Wenn Sie nicht wissen, wie Sie für einen Zellbereich einen Namen definieren, schlagen Sie in Kapitel 2 nach. Dort wird auch erläutert, warum ich in die Namen jeweils einen Unterstrich eingefügt habe.

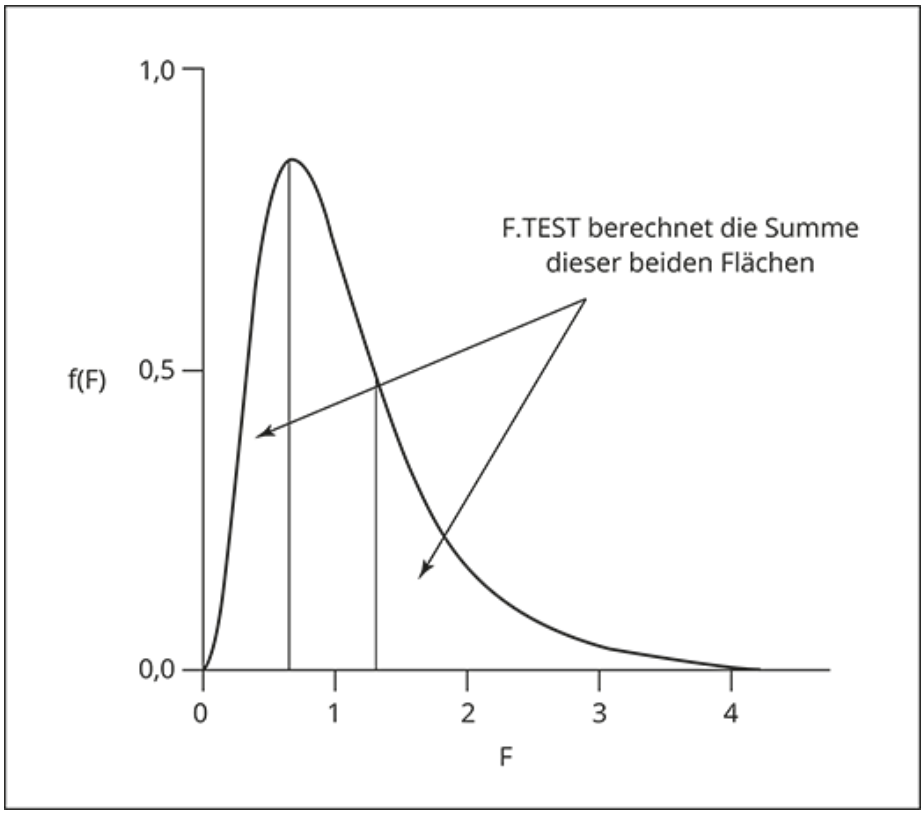


Abbildung 11.15: Die Ergebnisse von `F.TEST`

F.VERT und F.VERT.RE

Sie verwendendie Tabellenblattfunktionen F.VERT und F.VERT.RE, um zu entscheiden, ob Ihr berechneter *F*-Quotient im kritischen Bereich liegt. Bei der Funktion F.VERT geben Sie einen Wert für *F*, einen Wert für df und für das Argument mit dem Namen KUMULIERT einen logischen Wert (WAHR oder FALSCH) an. Wenn der Wert für KUMULIERT WAHR ist, ermittelt die Wahrscheinlichkeit, einen *F*-Quotienten, der höchstens so groß ist wie Ihrer, wenn H_0 wahr ist. (Excel nennt dies linksseitige Wahrscheinlichkeit.) Wenn die Wahrscheinlichkeit größer ist als 1– α, wird H_0 verworfen. Wenn der Wert für KUMULIERT FALSCH ist, gibt F.VERT die Höhe der *F*-Verteilung an Ihrem Wert von *F* zurück. Ich verwende diese Option weiter hinten in diesem Kapitel, um für die *F*-Verteilung ein Diagramm zu erstellen.

F.VERT.RE gibt die Wahrscheinlichkeit zurück, einen *F*-Quotienten zu erhalten, der mindestens so groß ist wie Ihrer, wenn H_0 wahr ist. (Excel nennt dies rechtsseitige Wahrscheinlichkeit.) Wenn dieser Wert für die Wahrscheinlichkeit kleiner ist als Ihr Wert für α, wird H_0 verworfen. Daher ist die Funktion F.VERT.RE etwas zielgerichteter.

Im Folgenden wende ich F.VERT.RE auf das obige Beispiel an. Der *F*-Quotient beträgt 1,36, die df-Werte sind 9 und 14.

Gehen Sie wie folgt vor:

1. **Wählen Sie eine Zelle für das Ergebnis aus.**

2. **Klicken im Menü mit den statistischen Funktionen auf F.VERT.RE, um das Dialogfeld FUNKTIONSARGUMENTE für F.VERT.RE zu öffnen (siehe Abbildung 11.16).**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente ein.**

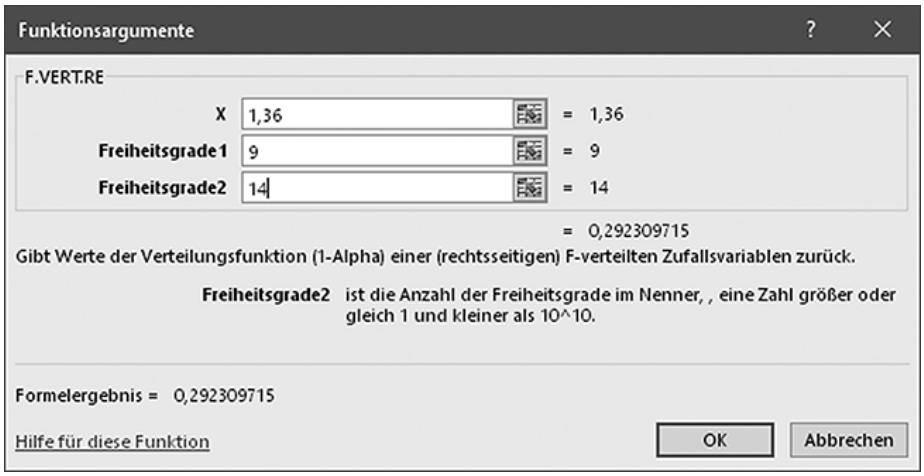


Abbildung 11.16: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion F.VERT.RE

Geben Sie im Feld X den berechneten *F*-Wert ein. In diesem Beispiel beträgt der berechnete *F*-Wert 1,36.

Geben Sie im Feld FREIHEITSGRADE1 die Freiheitsgrade für die Varianzschätzung im Zähler des *F*-Quotienten ein. Die Freiheitsgrade betragen für den Zähler in diesem Beispiel 9 (10 Werte − 1).

Geben Sie im Feld FREIHEITSGRADE2 die Freiheitsgrade für die Varianzschätzung im Nenner des *F*-Quotienten ein. Die Freiheitsgrade betragen für den Nenner in diesem Beispiel 14 (15 Werte − 1).

Wenn Sie diese Zahl eingegeben haben, wird das Ergebnis im Dialogfeld angezeigt.

1. **Klicken Sie auf OK, um das Dialogfeld zu schließen und das Ergebnis in die markierte Zelle einzufügen.**

   Der Wert im Dialogfeld in Abbildung 11.16 ist größer als 0,05. Somit wird H_0 nicht verworfen.

F.INV und F.INV.RE

Die Tabellenblattfunktionen F.INV sind die Umkehrungen der entsprechenden F.VERT-Funktionen. F.INV ermittelt den Wert in

der *F*-Verteilung, der einen bestimmten Anteil der Fläche der unteren (linken) Seite abschneidet. `F.INV.RE` ermittelt den Wert in der *F*-Verteilung, der einen bestimmten Anteil der Fläche der oberen (rechten) Seite abschneidet. Mit `F.INV.RE` können Sie den kritischen Wert von *F* ermitteln.

Hier verwende ich `F.INV.RE`, um den kritischen Wert für den zweiseitigen Test am Beispiel mit den FarKlempt-Maschinen zu ermitteln.

1. **Wählen Sie eine Zelle für das Ergebnis aus.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf F.INV.RE, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die entsprechenden Werte für die Argumente ein.**

   Geben Sie im Feld WAHRSCH den Anteil der Fläche an der oberen Seite ein. In diesem Beispiel ist das der Wert 0,025, da es sich um einen zweiseitigen Test mit α = 0,05 handelt.

   Geben Sie im Feld FREIHEITSGRADE1 die Freiheitsgrade für den Zähler ein. In diesem Beispiel beträgt df für den Zähler = 9.

   Geben Sie im Feld FREIHEITSGRADE2 die Freiheitsgrade für den Nenner ein. In diesem Beispiel beträgt df für den Nenner = 14.

   Wenn Sie diesen Wert eingegeben haben, wird das Ergebnis, nämlich 3,209 (und noch ein paar Dezimalstellen mehr), im Dialogfeld angezeigt (siehe Abbildung 11.17).

4. **Klicken Sie auf OK, um das Ergebnis in die markierte Zelle einzufügen.**

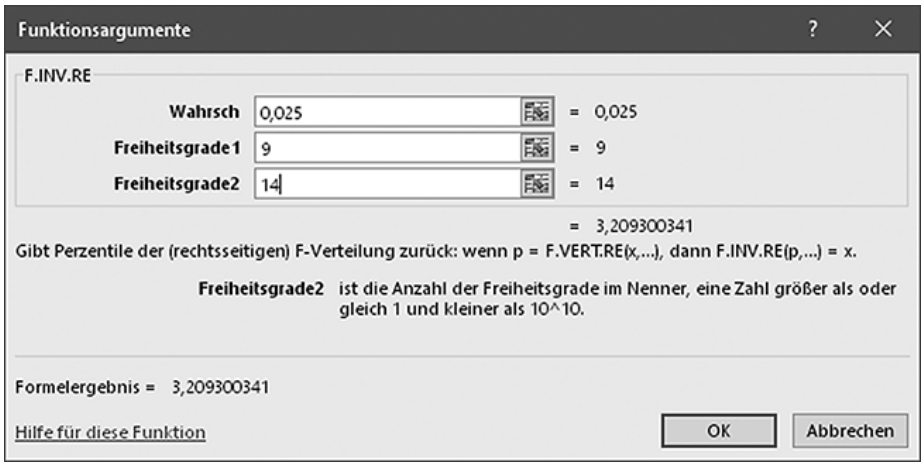


Abbildung 11.17: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `F.INV.RE`

Datenanalysetool: Zwei-Stichproben F-Test

Excel stellt ein Datenanalysetool für einen *F*-Test an zwei Stichprobenvarianzen bereit. Ich verwende dieses Tool hier für das obige Beispiel für Stichprobenvarianzen. In Abbildung 11.18 sind die Daten mit dem Dialogfeld ZWEI-STICHPROBEN F-TEST dargestellt.

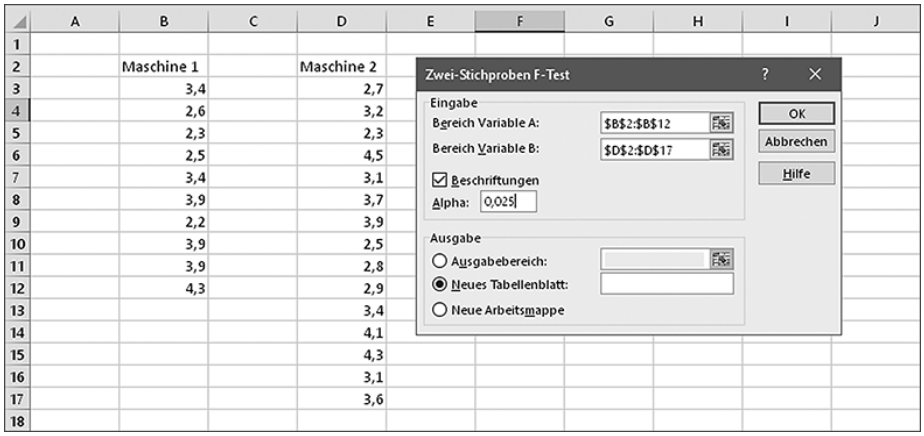


Abbildung 11.18: Das Datenanalysetool für den *F*-Test und Daten aus zwei Stichproben

Gehen Sie wie folgt vor, um dieses Tool zu verwenden:

1. **Geben Sie die Daten für die beiden Stichproben jeweils in einen eigenen Zellbereich ein.**

   Bei diesem Beispiel befinden sich die Daten für die Stichprobe der *Maschine 1* in Spalte B und die Daten für die Stichprobe der *Maschine 2* in Spalte D.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ZWEI-STICHPROBEN F-TEST aus.**

4. **Klicken Sie auf OK, um das Dialogfeld für dieses Tool zu öffnen.**

   In Abbildung 11.18 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld BEREICH VARIABLE A den Zellbereich ein, in dem sich die Daten für die erste Stichprobe befinden.**

   Die Daten für *Maschine 1* befinden sich in unserem Beispiel in $B$2:$B$12. (Beachten Sie die Dollarzeichen $ für den absoluten Bezug.)

6. **Geben Sie im Feld BEREICH VARIABLE B den Zellbereich ein, in dem sich die Daten für die zweite Stichprobe befinden.**

   Die Daten für *Maschine 2* befinden sich in den Zellen $D$2:$D$17.

7. **Wenn die Zellbereiche Spaltenüberschriften enthalten, schalten Sie das Kontrollkästchen BESCHRIFTUNGEN ein.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich auch dieses Kontrollkästchen aktiviert.

8. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie den Wert für ein anderes α.**

   Im Feld ALPHA ist der Alphawert für einen einseitigen Test voreingestellt. Ich möchte jedoch einen zweiseitigen Test durchführen. Also habe ich diesen Wert in 0,025 geändert.

9. **Legen Sie im Bereich AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

   Ich habe das Optionsfeld NEUES TABELLENBLATT ausgewählt, damit das Ergebnis in der aktuellen Arbeitsmappe auf einem neuen Blatt angezeigt wird.

10. **Klicken Sie auf OK.**

    Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird das Blatt mit den Ergebnissen angezeigt.

In Abbildung 11.19 ist das von diesem Tool berechnete Ergebnis mit erweiterten Spalten dargestellt. In den Zeilen 4 bis 6 befinden sich die Stichprobenwerte. In Zeile 7 werden die Freiheitsgrade angezeigt.

	A	B	C	D
1	Zwei-Stichproben F-Test			
2				
3		*Maschine 1*	*Maschine 2*	
4	Mittelwert	3,24	3,34	
5	Varianz	0,600444444	0,441142857	
6	Beobachtungen	10	15	
7	Freiheitsgrade (df)	9	14	
8	Prüfgröße (F)	1,361111111		
9	P(F<=f) einseitig	0,291848115		
10	Kritischer F-Wert bei einseitigem Test	3,209300341		
11				

Abbildung 11.19: Ergebnis des Datenanalysetools für den Zwei-Stichproben-*F*-Test

In den restlichen Zeilen befinden sich Informationen zum *F*-Quotienten. Der berechnete Wert für *F* befindet sich in Zelle B8. In

Zelle B9 ist der Anteil der Fläche angegeben, den der berechnete *F*-Quotient an der oberen Seite der *F*-Verteilung abschneidet. In Abbildung 11.15 ist dies die Fläche auf der rechten Seite. Zelle B10 gibt den kritischen Wert für einen einseitigen Test an: Das ist der Wert, der den Anteil der Fläche an der oberen Seite abschneidet, der dem Wert im Feld ALPHA entspricht.

Die F-Verteilung visualisieren

Wie Sie im nächsten Kapitel noch sehen werden, ist die *F*-Verteilung in der Statistik sehr wichtig. Damit Sie die *F*-Verteilung besser verstehen, zeige ich Ihnen, wie Sie diese in einem Diagramm darstellen können. Abbildung 11.20 zeigt die Zahlen und das fertige Ergebnis.

Gehen Sie folgendermaßen vor, um das Diagramm zu erstellen.

1. **Geben Sie die Freiheitsgrade in Zellen ein.**

   In Zelle B1 habe ich 10 und in Zelle B2 15 eingegeben.

2. **Erzeugen Sie eine Spalte mit Werten für die Statistik.**

   In die Zellen D2 mit D42 habe ich die Werte 0 bis 8 in Schritten von 0,2 eingegeben.

3. **Berechnen Sie in der ersten Zelle der benachbarten Spalte den Wert der Dichtefunktion für den ersten Wert der Statistik.**

   Da ich eine *F*-Verteilung darstellen will, verwende ich in Zelle E2 die Funktion `F.VERT`. Klicken Sie für den Wert von X Zelle D2 an. Für FREIHEITSGRADE1 klicke ich Zelle B1 an und drücke F4, um hieraus einen absoluten Zellbezug zu machen. Für FREIHEITSGRADE2 klicke ich B2 an und drücke ebenfalls F4. Im Feld KUMULIERT gebe ich `FALSCH` ein, um die Höhe der

Verteilung für diesen Wert von x berechnen zu lassen. Abschließend klicke ich auf OK.

4. **Verwenden Sie automatisches Ausfüllen, um diese Spalte mit den weiteren Werten zu füllen.**

5. **Erstellen Sie das Diagramm.**

   Markieren Sie beide Spalten. Wählen Sie in der Gruppe DIAGRAMME der Registerkarte EINFÜGEN den Typ PUNKTE MIT INTERPOLIERTEN LINIEN aus.

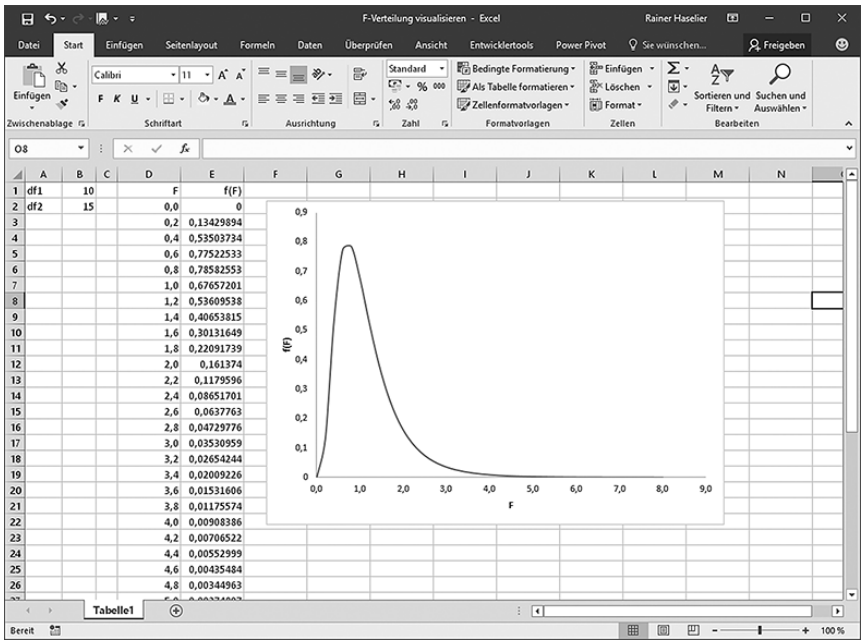


Abbildung 11.20: Die *F*-Verteilung visualisieren

6. **Formatieren Sie das Diagramm.**

   Ich habe das Diagramm angeklickt, damit die Schaltfläche DIAGRAMMELEMENTE (mit den Pluszeichen) sichtbar wird. Im Menü dieser Schaltfläche habe ich die Achsentitel eingeschaltet und den Diagrammtitel sowie die Gitternetzlinien entfernt, aber das ist eine Frage des persönlichen Geschmacks. Außerdem habe ich das Diagramm ein wenig vergrößert.

7. **Manipulieren Sie das Diagramm.**

Um ein besseres Gefühl für die Wirkung der Freiheitsgrade zu bekommen, können Sie die Werte für df ändern und dann sehen, wie sich diese Änderung auf das Diagramm auswirkt.

Kapitel 12

Mehr als zwei Stichproben testen

In diesem Kapitel

- Warum mehrfache *t*-Tests nicht funktionieren
- Die Varianzanalyse verwenden
- So geht es nach der Varianzanalyse weiter
- Mit wiederholten Messungen arbeiten
- Eine Trendanalyse durchführen

Die Statistik wäre langweilig, wenn nur Rückschlüsse aufgrund von einer oder zwei Stichproben gezogen werden könnten. In diesem Kapitel erfahren Sie, wie Sie Hypothesen bei drei und mehr Stichproben testen können. Ich zeige Ihnen, wie Sie vorgehen müssen, wenn Stichproben unabhängig voneinander sind, und wie Sie vorgehen müssen, wenn sie es nicht sind. Für beide Fälle erläutere ich, wie es nach dem Testen der Hypothesen weitergeht.

Ich stelle Ihnen außerdem die Datenanalysetools von Excel vor, die Ihnen die ganze Arbeit abnehmen. Auch wenn diese Tools nicht so gut sind wie die in einer speziellen Statistik-Software, so können Sie doch mit diesen Tools und den Standardfunktionen von Excel recht passable Analysen durchführen.

Mehr als zwei Stichproben testen

Stellen Sie sich folgende Situation vor: Sie werden von Ihrem Unternehmen gebeten, drei unterschiedliche Methoden für die Weiterbildung von Mitarbeitern für eine bestimmte Tätigkeit zu bewerten. Sie wählen nach dem Zufallsprinzip je 30 Mitarbeiter für die drei Methoden aus. Sie lassen diese an der Weiterbildung teilnehmen, testen sie, erfassen die Ergebnisse in einer Tabelle und ziehen Ihre Schlüsse. Bevor Sie die Untersuchung abschließen können, verlassen drei Mitarbeiter das Unternehmen: ein Mitarbeiter aus der Gruppe *Methode 1* und zwei Mitarbeiter aus der Gruppe *Methode 3*.

In Tabelle 12.1 sind diese Daten aufgeführt.

	Methode 1	Methode 2	Methode 3
	95	83	68
	91	89	75
	89	85	79
	90	89	74
	99	81	75
	88	89	81
	96	90	73
	98	82	77
	95	84	
		80	
Mittelwert	93,44	85,20	75,25
Varianz	16,28	14,18	15,64
Standardabweichung	4,03	3,77	3,96

Tabelle 12.1: Daten aus drei Weiterbildungsmethoden

Ergeben sich bei den drei Methoden unterschiedliche Ergebnisse oder sind diese so gleich, dass sich kaum ein Unterschied ausmachen lässt? Um dies zu entscheiden, müssen Sie einen Hypothesentest durchführen:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Nicht H_0

wobei $\alpha = 0{,}05$.

Eine harte Nuss

Das hört sich ganz einfach an, vor allem, wenn Sie Kapitel 11 gelesen haben. Nehmen Sie den Mittelwert der Werte aus Methode 1, den Mittelwert der Werte aus Methode 2 und führen Sie einen *t*-Test durch, um festzustellen, ob diese sich unterscheiden. Gehen Sie für Methode 1 und Methode 3 und für Methode 2 und Methode 3 genauso vor. Wenn mindestens einer dieser *t*-Tests eine signifikante Differenz aufweist, verwerfen Sie H_0. Kein Problem also, oder?

Doch. Wenn α für jeden *t*-Test 0,05 beträgt, begehen Sie einen Fehler 1. Art, wobei die Wahrscheinlichkeit höher als beabsichtigt ist. Die Wahrscheinlichkeit, dass mindestens einer der drei *t*-Tests eine signifikante Differenz aufweist, liegt weit über 0,05. Genau gesehen, liegt der Wert bei 0,14 und damit weit außerhalb des akzeptablen Bereichs. (Die mathematische Berechnung dieser Zahl ist etwas kompliziert. Daher werde ich nicht weiter darauf eingehen.)

Bei mehr als drei Stichproben wird die Situation noch dramatischer. Für vier Gruppen sind sechs *t*-Tests erforderlich und die Wahrscheinlichkeit, dass mindestens einer davon signifikant ist, liegt bei 0,26. In Tabelle 12.2 ist dargestellt, was bei zunehmender Anzahl von Stichproben geschieht.

Anzahl der Stichproben	Anzahl der -Tests	(mindestens ein signifikanter -Wert)
3	3	0,14
4	6	0,26

Anzahl der Stichproben	Anzahl der -Tests	(mindestens ein signifikanter -Wert)
5	10	0,40
6	15	0,54
7	21	0,66
8	28	0,76
9	36	0,84
10	45	0,90

Tabelle 12.2: Der unglaublich zunehmende Alphawert

Mehrere *t*-Tests sind also eindeutig keine Lösung. Was also tun?

Eine Lösung

Wir müssen uns wohl etwas anderes einfallen lassen. Eine Lösung bietet sich in der Verwendung von Varianzen anstelle von Mittelwerten.

Ich möchte, dass Sie sich Varianzen etwas anders vorstellen. Wenn Sie sich noch erinnern: Die Formel für die Schätzung der Varianz der Grundgesamtheit lautet wie folgt:

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

Da die Varianz nahezu ein Mittelwert der quadratischen Abweichungen vom Mittelwert ist, wird sie von Statistikern auch als *mittlere Quadratsumme* (engl. Mean Square, MS) bezeichnet. Irgendwie ist der Name nicht so ganz passend, weil hier die Abweichung vom Mittelwert fehlt. Aber so ist das nun mal.

Der Zähler der Varianz, Verzeihung, der mittleren Quadratsumme, ist die Summe der quadratischen Abweichungen vom Mittelwert. Da haben wir schon die nächste Bezeichnung: *Quadratsummen* (engl. Sum of Squares, SS). Im Nenner stehen, wie in Kapitel 10 bereits

erwähnt, die *Freiheitsgrade* (df). Die etwas andere Betrachtungsweise von Varianz sieht somit folgendermaßen aus:

$$Mittlere\ Quadratsumme = \frac{Quadratsummen}{df}$$

Das lässt sich wie folgt abkürzen:

$$MS = \frac{SS}{df}$$

Nun wollen wir aber die harte Nuss knacken. Ein wichtiger Schritt besteht darin, die mittleren Quadratsummen zu finden, die sich in den Daten verbergen. Ein weiterer Schritt besteht darin zu wissen, dass mit diesen mittleren Quadratsummen die Varianzen der Grundgesamtheiten geschätzt werden, aus denen diese Stichproben gezogen wurden. Für dieses Beispiel nehmen wir an, diese Varianzen seien gleich, so dass praktisch nur eine Varianz geschätzt wird. In einem letzten Schritt werden mit diesen Schätzwerten die zu Beginn des Kapitels aufgestellten Hypothesen getestet.

In den Daten in Tabelle 12.1 verbergen sich drei verschiedene mittlere Quadratsummen. Beginnen wir mit allen 27 Werten und lassen für den Moment außen vor, dass diese in drei Gruppen unterteilt sind. Nehmen wir an, wir möchten mit diesen 27 Werten einen Schätzwert der Varianz der Grundgesamtheit berechnen. (Sehr gewagt, aber lassen Sie mich mal.) Der Mittelwert dieser 27 Werte beträgt 85. Ich nenne diesen Mittelwert den *Gesamtmittelwert*, weil er der Mittelwert von allen Werten ist.

Somit ergibt sich für die mittlere Quadratsumme Folgendes:

$$\frac{(95-85)^2+(91-85)^2+\ldots+(73-85)^2+(77-85)^2}{(27-1)} = 68{,}08$$

Im Nenner stehen 26 (27 − 1) Freiheitsgrade. Diese Varianz bezeichne ich als *Gesamtvarianz* oder unter Berücksichtigung der

neuen Betrachtungsweise MS_{Gesamt}. Häufig wird sie auch als MS_G abgekürzt.

Ein weiteres Beispiel für eine Varianz, die es zu berücksichtigen gilt: In Kapitel 11 wurde der *t*-Test für zwei Stichproben bei gleichen Varianzen beschrieben. Für diesen Test werden zwei Stichprobenvarianzen zu einem *gepoolten* Schätzwert der Varianz der Grundgesamtheit zusammengefasst. Die Daten in Tabelle 12.1 enthalten drei Stichprobenvarianzen für einen gepoolten Schätzwert: 16,28, 14,18 und 15,64. Angenommen, diese Zahlen stellen gleiche Varianzen der Grundgesamtheit dar, dann folgt für den gepoolten Schätzwert:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + (n_3 - 1)\, s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$
$$= \frac{(9-1)\,(16{,}28) + (10-1)\,(14{,}18) + (8-1)\,(15{,}64)}{(9-1) + (10-1) + (8-1)} = 15{,}31$$

Da dieser gepoolte Schätzwert von der Varianz *innerhalb* der Gruppen stammt, wird er als $MS_{Innerhalb}$ oder MS_I bezeichnet.

Noch eine weitere mittlere Quadratsumme: die Varianz der Stichprobenmittelwerte um den Gesamtmittelwert. In diesem Beispiel bedeutet das die Varianz in diesen Zahlen: 93,44, 85,20 und 75,25 – oder so. Ich sage »oder so«, weil es sich um Mittelwerte, nicht um Werte handelt. Wenn Sie sich mit Mittelwerten befassen, müssen Sie die Anzahl der Werte berücksichtigen, aus denen sich die Mittelwerte zusammensetzen. Dazu multiplizieren Sie die quadratische Abweichung mit der Anzahl der Werte in der Stichprobe.

Somit ergibt sich für diese Varianz Folgendes:

$$\frac{(9)\,(93{,}44 - 85)^2 + (10)\,(85{,}20 - 85)^2 + (8)\,(75{,}25 - 85)^2}{3 - 1} = 701{,}34$$

Der df-Wert für diese Varianz beträgt 2 (die Anzahl der Stichproben – 1).

Statistiker bezeichnen dies als die Varianz *zwischen* Stichprobenmittelwerten. Diese Varianz wird als $MS_{Zwischen}$ oder MS_Z bezeichnet.

Sie kennen nun also drei Schätzwerte der Varianz der Grundgesamtheit: MS_G, MS_I und MS_Z. Und was fangen Sie damit an?

Erinnern Sie sich, dass das ursprüngliche Ziel war, eine Hypothese über drei Mittelwerte zu testen? Die Nullhypothese besagt, dass Unterschiede zwischen den drei Stichprobenmittelwerten ausschließlich dem Zufall zuzuschreiben sind. Daraus folgt, dass die Varianz zwischen diesen Mittelwerten dieselbe ist wie die Varianz von drei beliebigen Zahlen, die zufällig aus der Grundgesamtheit ausgewählt werden.

Wenn wir die Varianzen zwischen den Mittelwerten (das ist MS_Z, erinnern Sie sich?) irgendwie mit der Varianz der Grundgesamtheit vergleichen könnten, wüssten wir, ob das stimmt. Wenn wir nur einen Schätzwert der Varianz der Grundgesamtheit hätten, der unabhängig von den Differenzen unter den Gruppen wäre, dann wäre das alles kein Problem.

Aber wir haben diesen Schätzwert doch. Wir haben MS_I, einen Schätzwert, der auf dem Pooling der Varianzen innerhalb der Stichproben beruht. Wenn wir annehmen, dass diese Varianzen gleiche Varianzen der Grundgesamtheit darstellen, ist das eine recht solide Schätzung. In diesem Beispiel beruht sie auf 24 Freiheitsgraden.

Nun können wir wie folgt argumentieren: Wenn MS_Z etwa gleich groß ist wie MS_I, stimmt der Beweis mit H_0 überein. Wenn MS_Z signifikant größer ist als MS_I, stimmt der Beweis mit H_0 nicht überein. Sie formen diese Hypothesen:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Nicht H_0

in die folgenden Hypothesen um:

H_0: $\sigma_Z^2 \leq \sigma_I^2$

H_1: $\sigma_Z^2 > \sigma_I^2$

Statt mehrerer *t*-Tests zwischen Stichprobenmittelwerten testen Sie die Differenz zwischen zwei Varianzen.

Mit welchem Test tun Sie das? In Kapitel 11 haben Sie den Test für Hypothesen bei zwei Varianzen kennengelernt. Dieser Test wird als *F*-Test bezeichnet. Um diesen Test durchzuführen, dividieren Sie eine Varianz durch die andere. Sie schätzen das Ergebnis anhand der so genannten *F*-Verteilungen. Da zwei Varianzen beteiligt sind, wird jede Verteilung durch zwei Werte für die Freiheitsgrade definiert.

Bei diesem Beispiel weist die *F*-Verteilung die Werte df = 2 (für MS_Z) und df = 24 (für MS_I) auf. In Abbildung 12.1 ist diese *F*-Verteilung dargestellt. Für unsere Zwecke ist dies die Verteilung von möglichen *F*-Werten, wenn H_0 wahr ist. (Siehe auch den Abschnitt »Die F-Verteilung visualisieren« in Kapitel 11.)

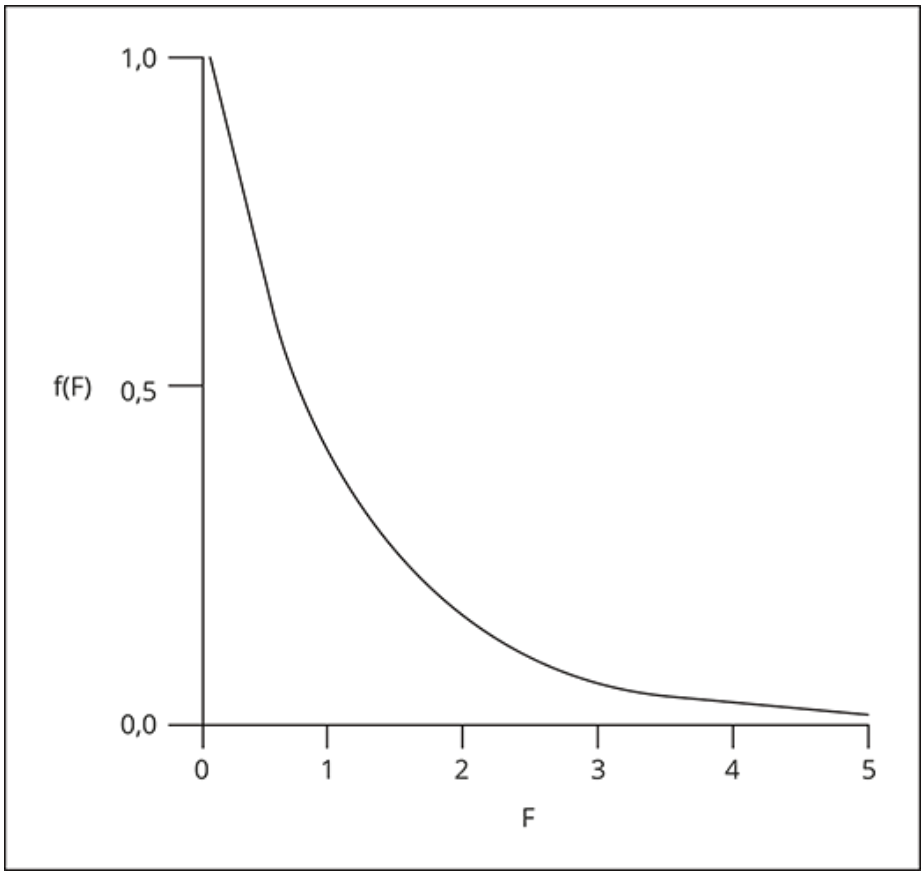


Abbildung 12.1: Die *F*-Verteilung mit den Freiheitsgraden 2 und 24

Die Testgröße für dieses Beispiel lautet wie folgt:

$$F = \frac{701,34}{15,31} = 45,82$$

Welchen Anteil der Fläche schneidet dieser Wert an der oberen Seite der *F*-Verteilung ab? In Abbildung 12.1 können Sie erkennen, dass dieser Anteil mikroskopisch klein ist, da die Werte an der x-Achse nur bis 5 reichen. (Und der Anteil der Fläche nach 5 ist sehr klein.) Er ist deutlich kleiner als 0,05.

Das bedeutet, es ist sehr unwahrscheinlich, dass Differenzen zwischen den Mittelwerten auf einen Zufall zurückzuführen sind. Das bedeutet, dass H_0 verworfen wird.

Dieses ganze Verfahren, bei dem mehr als zwei Stichproben getestet werden, ist als *Varianzanalyse* oder auch ANOVA (engl. ANalysis Of VAriance) bekannt. Im Zusammenhang mit der Varianzanalyse wird der Nenner eines *F*-Quotienten als *Fehlerterm* oder *Zufallsfehler* bezeichnet. Die unabhängige Variable wird auch als *Faktor* bezeichnet. Somit handelt es sich hierbei um eine einfaktorielle Varianzanalyse.

In diesem Beispiel ist der Faktor die Unterrichtsmethode. Die möglichen Werte der unabhängigen Variablen werden als *Ausprägung* bezeichnet. Die unabhängige Variable in diesem Beispiel hat drei Ausprägungen.

Komplexere Studien haben mehrere Faktoren und jeder Faktor hat mehrere Ausprägungen.

Wichtige Zusammenhänge

Betrachten wir noch einmal die mittlere Quadratsumme in diesem Beispiel jeweils mit den zugehörigen Quadratsummen und Freiheitsgraden. Als wir weiter oben die mittleren Quadratsummen berechnet haben, habe ich Ihnen nicht jede einzelne Quadratsumme gezeigt. Das werde ich hier nachholen:

$$MS_Z = \frac{SS_Z}{df_Z} = \frac{1402,68}{2} = 701,34$$

$$MS_I = \frac{SS_I}{df_I} = \frac{367,32}{24} = 15,31$$

$$MS_G = \frac{SS_G}{df_G} = \frac{1770}{26} = 68,08$$

Beginnen wir mit den Freiheitsgraden: $df_Z = 2$, $df_I = 24$ und $df_G = 26$. Ist es Zufall, dass sie eine Summe bilden? Wohl kaum. Es ist immer so, dass

$$df_Z + df_I = df_G$$

ergibt. Und wie verhält es sich mit den Quadratsummen?

$$1402{,}68 + 367{,}32 = 1770$$

Auch das ist kein Zufall. Bei der Varianzanalyse trifft Folgendes immer zu:

$$SS_Z + SS_I = SS_G$$

Statistiker, die mit der Varianzanalyse arbeiten, sprechen von der Quadratsummenzerlegung (will heißen »Zerlegen in einander nicht überschneidende Teile«) von SS_G in einen Teil für SS_Z und einen zweiten Teil für SS_I sowie von der Zerlegung von df_G in eine Menge für df_Z und eine andere für df_I.

Nach dem F-Test

Mit dem *F*-Test können Sie entscheiden, ob H_0 verworfen werden soll oder nicht. Wenn Sie entschieden haben, H_0 zu verwerfen, was dann? Sie können nur sagen, dass sich irgendwo innerhalb der Gruppe der Mittelwerte irgendetwas von etwas anderem unterscheidet. Der *F*-Test gibt keine Auskunft, was »irgendetwas« und »etwas anderes« ist.

Geplante Vergleiche

Wenn Sie Näheres wissen möchten, müssen Sie weitere Tests durchführen. Und nicht nur das. Sie müssen diese Tests bereits vor der Varianzanalyse planen.

Welche Tests sind das? Angesichts dessen, was ich Ihnen vorher gesagt habe, werden Sie erstaunt sein: *t*-Tests. Auf den ersten Blick scheint das nicht mit dem zunehmenden Alphawert bei mehreren *t*-Tests zusammenzupassen. Tut es aber. Wenn Sie mit einer Varianzanalyse H_0 verwerfen können, können Sie sich mit *t*-Tests auf die Daten konzentrieren und herausfinden, wo sich die Differenzen befinden. Und wo ich gerade dabei bin: Dieser *t*-Test unterscheidet

sich geringfügig von dem *t*-Test, den Sie in Kapitel 11 kennengelernt haben.

Diese nach der Varianzanalyse durchgeführten *t*-Tests werden als *geplante Vergleiche* bezeichnet. Manche Statistiker bezeichnen sie auch als *A-priori-Tests*. Ich werde das Ganze an einem Beispiel erläutern. Nehmen wir an, Sie hatten vor dem Sammeln der Daten Grund zur Annahme, dass Methode 1 bessere Werte ergibt als Methode 2 und dass Methode 2 bessere Werte ergibt als Methode 3. In diesem Fall planen Sie im Voraus, die Mittelwerte dieser Stichproben miteinander zu vergleichen, wenn sich aufgrund der Varianzanalyse ergibt, dass H_0 verworfen werden soll.

Die Formel für diesen *t*-Test lautet wie folgt:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_I \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Dabei werden folgende Hypothesen getestet:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

MS_I nimmt die Stelle des gepoolten Schätzwerts s_p^2 ein, den Sie in Kapitel 11 kennengelernt haben. Als ich Ihnen MS_I vorgestellt habe, habe ich Ihnen gezeigt, dass es sich hierbei lediglich um einen gepoolten Schätzwert handelt, der Varianzen von mehr als zwei Stichproben beinhalten kann. Der df-Wert für diesen *t*-Test ist df_I anstelle von $(n_1 - 1) + (n_2 - 1)$.

Bei diesem Beispiel lautet der Vergleich von Methode 1 mit Methode 2 wie folgt:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_I \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{93,44 - 85,2}{\sqrt{15,31\left[\frac{1}{9} + \frac{1}{10}\right]}} = 4,59$$

Bei df = 24 schneidet dieser Wert von *t* einen minimalen Anteil der Fläche an der oberen Seite der *t*-Verteilung ab. Also wird H_0 verworfen.

Die Formel für den geplanten Vergleich mit *t*-Test, die ich Ihnen gezeigt habe, passt zum *t*-Test für zwei Stichproben. Sie können die Formel für den geplanten Vergleich mit *t*-Test so schreiben, dass sich zusätzliche Möglichkeiten eröffnen. Beginnen wir mit dem Zähler:

$$\bar{X}_1 - \bar{X}_2$$

etwas anders geschrieben:

$$(+1)\bar{X}_1 + (-1)\bar{X}_2$$

+1 und –1 sind *Vergleichskoeffizienten*, die auch *Kontraste* genannt werden. Ich werde sie als c_1 und c_2 bezeichnen. c_3 und $\bar{X}_3$ können in den Vergleich mit aufgenommen werden, auch wenn nur $\bar{X}_1$ mit $\bar{X}_2$ verglichen wird:

$$(+1)(\bar{X}_1 + (-1)\bar{X}_2 + (0)\bar{X}_3$$

Wichtig ist, dass die Summe der Koeffizienten null ergibt.

Und so passen die Vergleichskoeffizienten in die Formel für den geplanten Vergleich mittels *t*-Test für eine Studie mit drei Stichproben:

$$t = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3}{\sqrt{MS_I\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}\right]}}$$

Und so wird mit dieser Formel Methode 2 mit Methode 3 verglichen:

$$t = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3}{\sqrt{MS_I\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}\right]}} = \frac{(0)(93,44) + (+1)(85,2) + (-1)(75,25)}{\sqrt{MS_I\left[\frac{0^2}{9} + \frac{1^2}{10} + \frac{1^2}{8}\right]}} = 5,36$$

Der Wert für *t* zeigt, dass die Ergebnisse aus Methode 2 signifikant höher sind als die aus Methode 3.

Sie können auch einen komplexeren Vergleich planen, zum Beispiel Methode 1 mit dem Mittelwert von Methode 2 und Methode 3. Beginnen wir mit dem Zähler:

$$\bar{x}_1 - \frac{(\bar{x}_2 + \bar{x}_3)}{2}$$

Mit Vergleichskoeffizienten können wir diese Gleichung wie folgt schreiben:

$$(+1)\bar{x}_1 + \left(-\frac{1}{2}\right)\bar{x}_2 + \left(-\frac{1}{2}\right)\bar{x}_3$$

Wenn Ihnen ganze Zahlen lieber sind, können Sie auch Folgendes schreiben:

$$(+2)\bar{x}_1 + (-1)\bar{x}_2 + (-1)\bar{x}_3$$

Wenn Sie diese Zahlen in die Formel einsetzen, ergibt sich Folgendes:

$$t = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3}{\sqrt{MS_I\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}\right]}} = \frac{(2)(93,44) + (-1)(85,2) + (-1)(75,25)}{\sqrt{15,31\left[\frac{2^2}{9} + \frac{1^2}{10} + \frac{1^2}{8}\right]}} = 8,26$$

Auch hier spricht alles dafür, H_0 zu verwerfen.

Nicht geplante Vergleiche

Das alles wäre langweilig, wenn mit diesen nach der Varianzanalyse durchgeführten Tests nur im Voraus geplante Vergleiche angestellt werden könnten. Manchmal möchten Sie sich auch einfach ein bisschen in den Daten umschauen und sehen, ob sich möglicherweise etwas Interessantes ergibt. Manchmal entdecken Sie etwas, was Sie nicht erwartet hätten.

In diesem Fall können Sie Vergleiche anstellen, die Sie nicht geplant haben. Diese Art von Vergleich wird als *A-posteriori-Test*, *Post-hoc-Test* oder einfach als *nicht geplanter Vergleich* bezeichnet. Statistiker haben eine ganze Palette dieser Tests entwickelt, von denen einige recht exotische Namen tragen und von denen viele von speziellen Stichprobenverteilungen abhängen.

Bei diesen Tests bezahlen Sie einen Preis dafür, dass Sie den Vergleich nicht im Voraus geplant haben. Dieser Preis hat damit zu tun, dass verhindert wird, dass H_0 für den jeweiligen Vergleich verworfen wird.

Von all den nicht geplanten Tests gefällt mir der des berühmten Statistikers Henry Scheffé am besten. Dieser besteht nicht aus esoterischen Formeln und Verteilungen, sondern aus dem Test, den ich Ihnen bereits gezeigt habe, und noch ein paar ganz einfachen Extras.

Das erste Extra besteht darin, dass Sie den Zusammenhang zwischen *t* und *F* kennen müssen. Den *F*-Test für drei Stichproben haben Sie bereits kennengelernt. Sie können einen *F*-Test aber auch für zwei Stichproben durchführen. Für diesen *F*-Test gilt: df = 1 und df = $(n_1–1) + (n_2–1)$. Der df-Wert für den *t*-Test lautet übrigens auch $(n_1−1) + (n_2−1)$. Hmmm … es scheint so, als würde hier ein Zusammenhang bestehen.

Und tatsächlich. Zwischen dem Zwei-Stichproben-*t*-Wert und dem Zwei-Stichproben-*F*-Wert besteht folgender Zusammenhang:

$$t^2 = F$$

Nun kann ich Ihnen die einzelnen Schritte des Scheffé-Tests verraten:

1. **Berechnen Sie den geplanten Vergleich mittels *t*-Test.**

2. **Quadrieren Sie das Ergebnis, um zu *F* zu gelangen.**

3. **Ermitteln Sie den kritischen Wert von *F* für df_Z und df_I bei α = 0,05 (oder welchen Wert für Alpha Sie auch immer gewählt haben).**

4. **Multiplizieren Sie diesen kritischen Wert von *F* mit der Anzahl der Stichproben −1.**

   Das Ergebnis ist der kritische Wert von *F* für den nicht geplanten Vergleich. Ich nenne diesen Wert *F′*.

5. **Vergleichen Sie den berechneten *F*-Wert mit *F′*.**

   Wenn der berechnete *F*-Wert größer ist, verwerfen Sie H_0 für diesen Test. Wenn er nicht größer ist, verwerfen Sie H_0 für diesen Test nicht.

Stellen Sie sich vor, Sie hätten nicht im Voraus geplant, den Mittelwert von Methode 1 mit dem Mittelwert von Methode 3 zu vergleichen. (In einer Studie mit drei Stichproben ist das kaum vorstellbar, das kann ich Ihnen sagen.) Die Formel für den *t*-Test lautet wie folgt:

$$t = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3}{\sqrt{MS_I\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3}\right]}} = \frac{(+1)(93,44) + (0)(85,2) + (-1)(75,25)}{\sqrt{15,31\left[\frac{1^2}{9} + \frac{0^2}{10} + \frac{-1^2}{8}\right]}} = 9,57$$

Wenn Sie dieses Ergebnis quadrieren, ergibt sich Folgendes:

$$F = t^2 = (9,57)^2 = 91,61$$

Für *F* beträgt der kritische Wert 3,403 bei df = 2 und 24 und α = 0,05. (Sie können das in einer Tabelle in einem Statistikbuch

nachschlagen oder mit der Tabellenblattfunktion `F.INV.RE` berechnen.) Somit gilt:

$$F' = (3 - 1)F = (2)(3{,}403) = 6{,}806$$

Da der berechnete Wert *F* mit 91,61 größer ist als *F′*, wird H_0 verworfen. Das zeigt, dass die Ergebnisse aus Methode 1 anders sind als die Ergebnisse aus Methode 3.

Datenanalysetool: Einfaktorielle Varianzanalyse

Die Berechnungen für die Varianzanalyse können recht umfangreich sein. Excel stellt ein Datenanalysetool bereit, das Ihnen einen Großteil der Arbeit abnimmt. Dieses Tool heißt *Einfaktorielle Varianzanalyse*. In Abbildung 12.2 ist dieses Tool zusammen mit den Daten aus dem vorherigen Beispiel dargestellt.

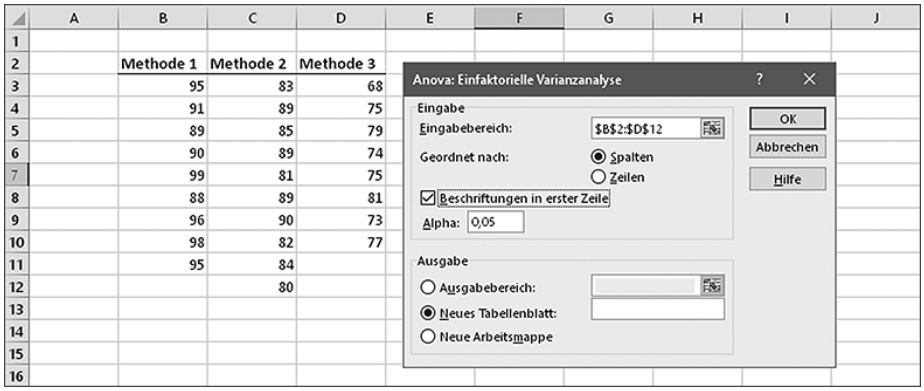


Abbildung 12.2: Das Datenanalysetool für die einfaktorielle Varianzanalyse

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie die Daten für die Stichproben in separate Zellbereiche ein.**

   Bei diesem Beispiel befinden sich die Daten aus der Stichprobe zu Methode 1 in Spalte B, die Daten aus der Stichprobe zu Methode 2 in Spalte C und die Daten aus der Stichprobe zu Methode 3 in Spalte D.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ANOVA: EINFAKTORIELLE VARIANZANALYSE aus.**

4. **Klicken Sie auf OK, um das Dialogfeld ANOVA: EINFAKTORIELLE VARIANZANALYSE zu öffnen.**

   In Abbildung 12.2 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld EINGABEBEREICH die Zellen an, in denen sich die Werte befinden.**

   In diesem Beispiel befinden sich die Daten in den Zellen $B$2:$D$12. (Beachten Sie die Dollarzeichen [$] für den absoluten Bezug.)

6. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich dieses Kontrollkästchen aktiviert.

7. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie diesen Wert, falls Sie möchten.**

8. **Legen Sie im Abschnitt AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, so dass das Ergebnis auf einem neuen Arbeitsblatt angezeigt wird.

9. **Klicken Sie auf OK.**

   Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird ein neues Arbeitsblatt mit dem Ergebnis erstellt.

In Abbildung 12.3 habe ich das von diesem Tool berechnete Ergebnis mit angepassten Spaltenbreiten dargestellt. Das Ergebnis besteht aus zwei Tabellen: *ZUSAMMENFASSUNG* und *ANOVA*. Die Tabelle *ZUSAMMENFASSUNG* enthält zusammenfassende Informationen zu den Stichproben: die Anzahl der Werte in den einzelnen Gruppen, die Gruppensummen, die Mittelwerte und die Varianzen. In der Tabelle *ANOVA* sind die Quadratsummen, die Freiheitsgrade, die mittlere Quadratsumme, die Prüfgröße *F*, der P-Wert und der kritische *F*-Wert für die angegebenen Freiheitsgrade aufgeführt. Der P-Wert ist der Anteil der Fläche, den der *F*-Quotient an der oberen Seite der *F*-Verteilung abtrennt. Wenn dieser Wert kleiner als 0,05 ist, wird H_0 verworfen.

	A	B	C	D	E	F	G	H
1	Anova: Einfaktorielle Varianzanalyse							
2								
3	ZUSAMMENFASSUNG							
4	*Gruppen*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*			
5	Methode 1	9	841	93,44444444	16,27777778			
6	Methode 2	10	852	85,2	14,17777778			
7	Methode 3	8	602	75,25	15,64285714			
8								
9								
10	ANOVA							
11	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*	
12	Unterschiede zwischen den Gruppen	1402,677778	2	701,3388889	45,82389062	6,38098E-09	3,402826105	
13	Innerhalb der Gruppen	367,3222222	24	15,30509259				
14								
15	Gesamt	1770	26					
16								

Abbildung 12.3: Ergebnis des Analysetools ANOVA: EINFAKTORIELLE VARIANZANALYSE

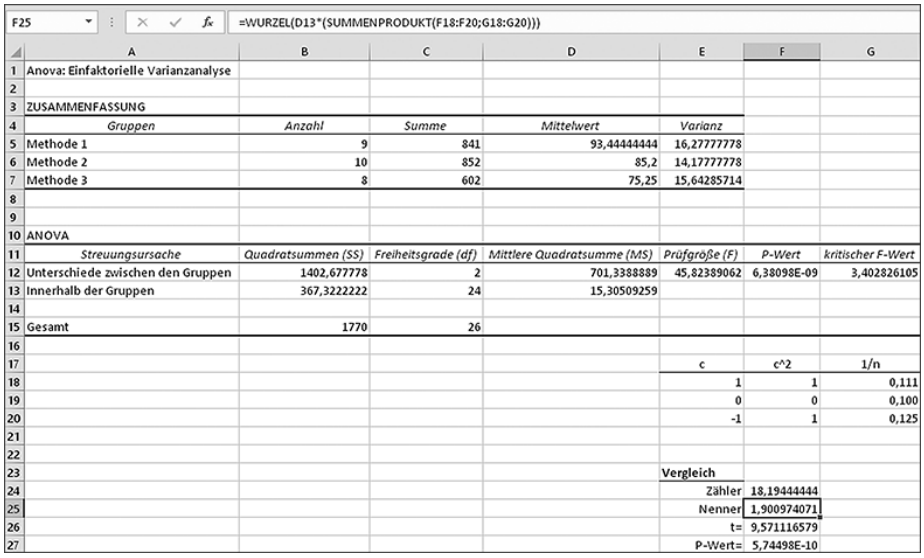
Mittelwerte vergleichen

Das Datenanalysetool Einfaktorielle Varianzanalyse enthält keine Funktion zum geplanten (oder nicht geplanten) Vergleichen der Mittelwerte. Mit ein bisschen Erfindungsgeist können Sie diese Vergleiche jedoch mit der Excel-Tabellenblattfunktion `SUMMENPRODUKT` durchführen.

Das Arbeitsblatt mit dem Ergebnis der Varianzanalyse ist der Ausgangspunkt für die geplanten Vergleiche. In diesem Abschnitt führen wir einen geplanten Vergleich durch: Mittelwert von Methode 1 im Vergleich zum Mittelwert von Methode 2.

Zuerst erstellen wir Spalten mit den für die Vergleiche wichtigen Informationen. In Abbildung 12.4 wird deutlich, was ich meine. Ich habe die Vergleichskoeffizienten in Spalte E, die Quadrate dieser

Koeffizienten in Spalte F und die reziproken Werte der einzelnen Stichprobenumfänge (1/*n*) in Spalte G ab Zeile 18 eingegeben.



F25 =WURZEL(D13*(SUMMENPRODUKT(F18:F20;G18:G20)))

	A	B	C	D	E	F	G
1	Anova: Einfaktorielle Varianzanalyse						
2							
3	ZUSAMMENFASSUNG						
4	*Gruppen*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
5	Methode 1	9	841	93,44444444	16,27777778		
6	Methode 2	10	852	85,2	14,17777778		
7	Methode 3	8	602	75,25	15,64285714		
8							
9							
10	ANOVA						
11	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
12	Unterschiede zwischen den Gruppen	1402,677778	2	701,3388889	45,82389062	6,38098E-09	3,402826105
13	Innerhalb der Gruppen	367,3222222	24	15,30509259			
14							
15	Gesamt	1770	26				
16							
17					c	c^2	1/n
18					1	1	0,111
19					0	0	0,100
20					-1	1	0,125
21							
22							
23					Vergleich		
24					Zähler	18,19444444	
25					Nenner	1,900974071	
26					t=	9,571116579	
27					P-Wert=	5,74498E-10	

Abbildung 12.4: Einen geplanten Vergleich durchführen

Ein paar Zeilen darunter habe ich die Informationen zum *t*-Test eingegeben: den Zähler, den Nenner und den Wert von *t*. Ich habe den Zähler und den Nenner in getrennte Zellen eingegeben, um die Formeln zu vereinfachen. Sie können sie aber auch zusammen in eine große Formel eingeben und nur eine Zelle für *t* einrichten. Dann ist es aber schwierig, den Überblick zu behalten.

Mit `SUMMENPRODUKT` werden Zahlen in Zellbereichen multipliziert und die Produkte addiert. (Diese Funktion befindet sich nicht im Menü mit den statistischen Funktionen, sondern im Menü der Schaltfläche MATHEMATIK UND & TRIGONOMETRIE.) Ich habe mit `SUMMENPRODUKT` die einzelnen Koeffizienten mit den einzelnen Mittelwerten multipliziert und anschließend die Produkte addiert. Das Ergebnis habe ich in F24 gespeichert. Es ist der Zähler für den geplanten Vergleich. Die Formel für F24 lautet wie folgt:

```
=SUMMENPRODUKT(E18:E20;D5:D7)
```

In den Zellen E18:E20 befinden sich die Vergleichskoeffizienten und in den Zellen D5:D7 die Stichprobenmittelwerte.

In F25 befindet sich der Nenner. Die Formel hierfür (die Sie auch in der Bearbeitungsleiste in Abbildung 12.4 sehen) lautet wie folgt:

```
=WURZEL(D13*(SUMMENPRODUKT(F18:F20;G18:G20)))
```

In D13 befindet sich die mittlere Quadratsumme innerhalb der Gruppen (MS_I). `SUMMENPRODUKT` multipliziert die quadrierten Koeffizienten in F18:F20 mit den reziproken Werten der Stichprobenumfänge in G18:G20. `WURZEL` zieht die Wurzel aus dem Ganzen.

In F26 befindet sich der Wert für *t*. Dieser Wert berechnet sich einfach aus F24 dividiert durch F25.

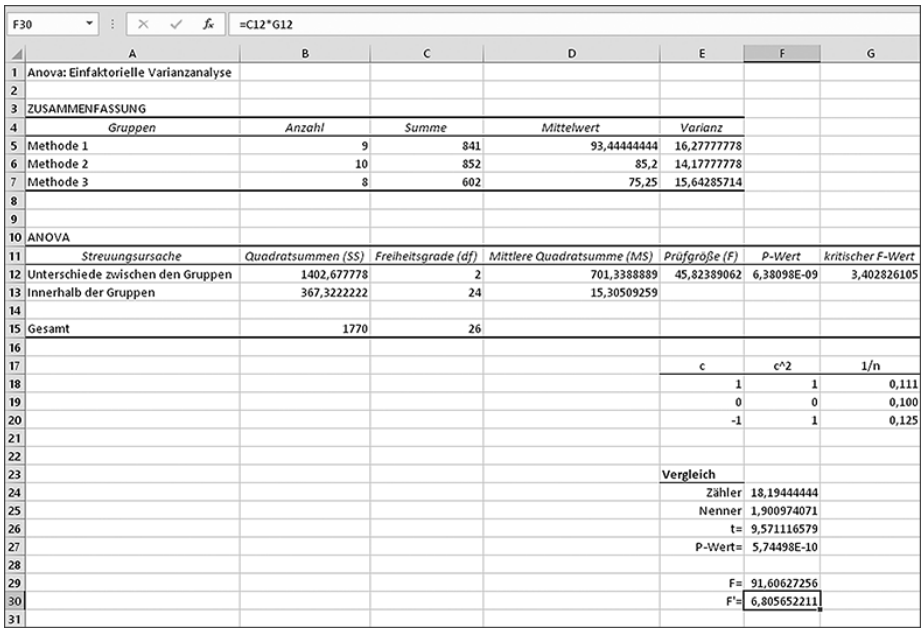
In F27 befindet sich der P-Wert für *t*: der Anteil der Fläche, den *t* bei df = 24 an der oberen Seite der *t*-Verteilung abschneidet. Die Formel für diese Zelle lautet wie folgt:

```
=T.VERT.RE(F26;C13)
```

Die Argumente sind der berechnete Wert für *t* (in F26), die Freiheitsgrade für MS_I (in C13).

Wenn Sie die Koeffizienten in E18:E20 ändern, stellen Sie sofort einen anderen Vergleich an.

Ich werde das jetzt tun und Ihnen den Post-hoc-Vergleich von Scheffé zeigen. Bei dem Vergleich in diesem Beispiel wird der Mittelwert von Methode 1 mit dem Mittelwert von Methode 3 verglichen. In Abbildung 12.5 sind die zusätzlichen Informationen für diesen Test ein paar Zeilen unter dem *t*-Test dargestellt.



F30 =C12*G12

	A	B	C	D	E	F	G
1	Anova: Einfaktorielle Varianzanalyse						
2							
3	ZUSAMMENFASSUNG						
4	*Gruppen*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
5	Methode 1	9	841	93,44444444	16,27777778		
6	Methode 2	10	852	85,2	14,17777778		
7	Methode 3	8	602	75,25	15,64285714		
8							
9							
10	ANOVA						
11	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
12	Unterschiede zwischen den Gruppen	1402,677778	2	701,3388889	45,82389062	6,38098E-09	3,402826105
13	Innerhalb der Gruppen	367,3222222	24	15,30509259			
14							
15	Gesamt	1770	26				
16							
17					c	c^2	1/n
18					1	1	0,111
19					0	0	0,100
20					-1	1	0,125
21							
22							
23					Vergleich		
24					Zähler	18,19444444	
25					Nenner	1,900974071	
26					t=	9,571116579	
27					P-Wert=	5,74498E-10	
28							
29					F=	91,60627256	
30					F'=	6,805652211	
31							

Abbildung 12.5: Einen Post-hoc-Vergleich durchführen

In der Zelle F29 befindet sich *F*, das Quadrat des *t*-Werts in F26. F30 enthält *F'*, das Produkt aus C12 (df_Z, die Anzahl der Stichproben − 1) und G12 (der kritische Wert von *F* für die Freiheitsgrade 2 und 24 und α = 0,05). F29 ist größer als F30, also wird H_0 für diesen Vergleich verworfen.

Eine andere Art Hypothese, eine andere Art Test

Die vorherige Varianzanalyse funktioniert mit unabhängigen Stichproben. Wie in Kapitel 11 erwähnt, kann es auch vorkommen, dass Sie mit zusammenhängenden Stichproben zu tun haben. So kann es beispielsweise sein, dass Daten über eine Person unter unterschiedlichen Bedingungen geliefert werden. In diesem Abschnitt lernen Sie die Varianzanalyse kennen, die bei mehr als zwei zusammenhängenden Stichproben verwendet wird.

Bei dieser Art von Varianzanalyse spricht man von *wiederholten Messungen*. In diesem Zusammenhang werden Ihnen auch Begriffe

wie *Blöcke mit zufälliger Zuteilung* oder *Innersubjekt-Design* begegnen.

Mit wiederholten Messungen bei der Varianzanalyse arbeiten

Damit Sie sehen, wie das funktioniert, erweitere ich das Beispiel aus Kapitel 11. In diesem Beispiel nehmen zehn Personen an einem Diätprogramm teil. In Tabelle 12.3 sind die Daten dieser Personen für einen Zeitraum von drei Monaten in Pfund aufgeführt.

Person	Vorher	Ein Monat	Zwei Monate	Drei Monate	Mittelwert
1	198	194	191	188	192,75
2	201	203	200	196	200,00
3	210	200	192	188	197,50
4	185	183	180	178	181,50
5	204	200	195	191	197,50
6	156	153	150	145	151,00
7	167	166	167	166	166,50
8	197	197	195	192	195,25
9	220	215	209	205	212,25
10	186	184	179	175	181,00
Mittelwert	192,4	189,5	185,8	182,4	187,525

Tabelle 12.3: Daten für das Beispiel mit dem Diätprogramm

Ist die Diät effektiv? Diese Frage verlangt nach einem Hypothesentest:

H_0: $\mu_{Vorher} = \mu_1 = \mu_2 = \mu_3$

H_1: Nicht H_0

Wie immer lege ich für α den Wert 0,05 fest.

Wie bei der vorherigen Varianzanalyse beginnen wir mit den Varianzen in den Daten. MS_G ist die Varianz in allen 40 Werten aus dem großen Mittelwert, der 187,525 beträgt:

$$MS_G = \frac{(198-187{,}525)^2 + (201-187{,}525)^2 + \ldots + (175-187{,}525)^2}{40-1} = 318{,}20$$

Die Personen, die an dem Diätprogramm teilnehmen, liefern ebenfalls eine Varianz. Der Gesamtmittelwert jeder Person (der Durchschnitt aus den vier Messungen) weicht vom großen Mittelwert ab. Da sich diese Daten in den Zeilen befinden, nenne ich sie MS_{Zeilen}:

$$MS_{Zeilen} = \frac{(4)(192{,}75-187{,}525)^2 + (4)(200-187{,}252)^2 + \ldots + (4)(181-187{,}525)^2}{10-1} = 1292{,}41$$

Dabei wird berücksichtigt, dass dieser Mittelwert der einzelnen Personen bereits aus vier Beobachtungen zusammengefasst wurde.

Die Mittelwerte aus den Spalten weichen ebenfalls vom großen Mittelwert ab:

$$MS_{Spalten} = \frac{\left(\begin{array}{l} (10)(192{,}4-187{,}525)^2 + (10)(189{,}5-187{,}525)^2 \\ + (10)(185{,}8-187{,}525)^2 + (10)(182{,}4-187{,}525)^2 \end{array} \right)}{4-1} = 189{,}69$$

Auch hier wird berücksichtigt, dass sich der Mittelwert der Spalten aus zehn Werten zusammensetzt.

Noch eine weitere Quelle für Varianzen befindet sich in den Daten. Diese können Sie sich als die Varianz vorstellen, die nach dem Abziehen der Varianz in den Zeilen und der Varianz in den Spalten von der Gesamtvarianz übrig bleibt. Es ist richtiger zu sagen: Diese Varianz basiert auf den Quadratsummen, die nach der Subtraktion von SS_{Zeilen} und $SS_{Spalten}$ von SS_{Gesamt} übrig bleiben.

Diese Varianz wird als MS_{Fehler} bezeichnet. Wie bereits erwähnt, wird in der Varianzanalyse der Nenner eines *F*-Quotienten als *Fehlerterm* bezeichnet. Das Wort »Fehler« ist ein Hinweis darauf, dass diese mittlere Quadratsumme (MS) der Nenner für einen *F*-Quotienten ist.

Zum Berechnen von MS_{Fehler} wird der Zusammenhang zwischen den Quadratsummen und zwischen den Freiheitsgraden genutzt:

$$MS_{Fehler} = \frac{SS_{Fehler}}{df_{Fehler}} = \frac{SS_{Gesamt} - SS_{Zeilen} - SS_{Spalten}}{df_{Gesamt} - df_{Zeilen} - df_{Spalten}} = \frac{209{,}175}{27} = 7{,}75$$

Und so können Sie df_{Fehler} auch berechnen:

$$df_{Fehler} = (\text{Anzahl der Zeilen} - 1)\,(\text{Anzahl der Spalten} - 1)$$

Zum Durchführen des Hypothesentests wird *F* berechnet:

$$F = \frac{MS_{Spalte}}{MS_{Fehler}} = \frac{189{,}69}{7{,}75} = 24{,}49$$

Für *F* beträgt der kritische Wert 2,96 bei df = 3 und 27 und α = 0,05. (Schlagen Sie den Wert nach oder verwenden Sie die Excel-Tabellenblattfunktion `F.INV.RE`.) Der berechnete *F*-Quotient ist größer als der kritische Wert für *F*. H_0 wird also verworfen.

Und wie verhält es sich mit einem *F*-Quotienten mit MS_{Zeilen}? Das geht bei diesem Beispiel nicht mit H_0 zusammen. Wenn Sie einen signifikanten *F*-Quotienten ermitteln, zeigt das nur, dass sich die Personen im Hinblick auf das Gewicht voneinander unterscheiden, und das sagt nicht viel aus.

Wie bei der Varianzanalyse, die ich Ihnen vorher gezeigt habe, planen Sie Vergleiche, um die Unterschiede herauszuarbeiten. Sie können dieselbe Formel verwenden, müssen jedoch anstelle von MS_I MS_{Fehler} verwenden:

$$t = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3 + c_4\bar{x}_4}{\sqrt{MS_{Fehler}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3} + \frac{c_4^2}{n_4}\right]}} = \frac{c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3 + c_4\bar{x}_4}{\sqrt{MS_{Fehler}\left[\frac{c_1^2 + c_2^2 + c_3^2 + c_4^2}{n}\right]}}$$

Die Formel wird in den Ausdruck auf der rechten Seite aufgelöst, da bei wiederholten Messungen *n* immer den gleichen Wert besitzt.

Der df-Wert für diesen Test ist der Wert df_{Fehler}.

Beim Post-hoc-Test von Scheffé gehen Sie genauso vor wie eben, ersetzen jedoch MS_I durch MS_{Fehler}. Außerdem ersetzen Sie beim Ermitteln von *F'* df_Z durch $df_{Spalten}$ und df_I durch df_{Fehler}.

Jetzt wird es trendy

In Situationen wie in dem Beispiel mit dem Diätprogramm haben Sie eine unabhängige quantitative Variable, das heißt, die Ausprägung dieser Variablen sind Zahlen (0 Monate, 1 Monat, 2 Monate, 3 Monate). Nicht nur das. In diesem Beispiel sind zudem die Intervalle gleich.

Bei dieser Art einer unabhängigen Variablen ist es häufig angebracht, in den Daten nach Trends zu suchen als nur Vergleiche unter Mittelwerten zu planen. Wenn wir die Mittelwerte im Beispiel mit dem Diätprogramm in einem Diagramm abtragen, scheinen sich diese einer Linie zu nähern (siehe Abbildung 12.6). Die *Trendanalyse* ist das statistische Verfahren, mit dem dieses Muster untersucht wird. Ziel dabei ist es herauszufinden, ob das Muster die signifikanten Differenzen unter den Mittelwerten unterstützt.

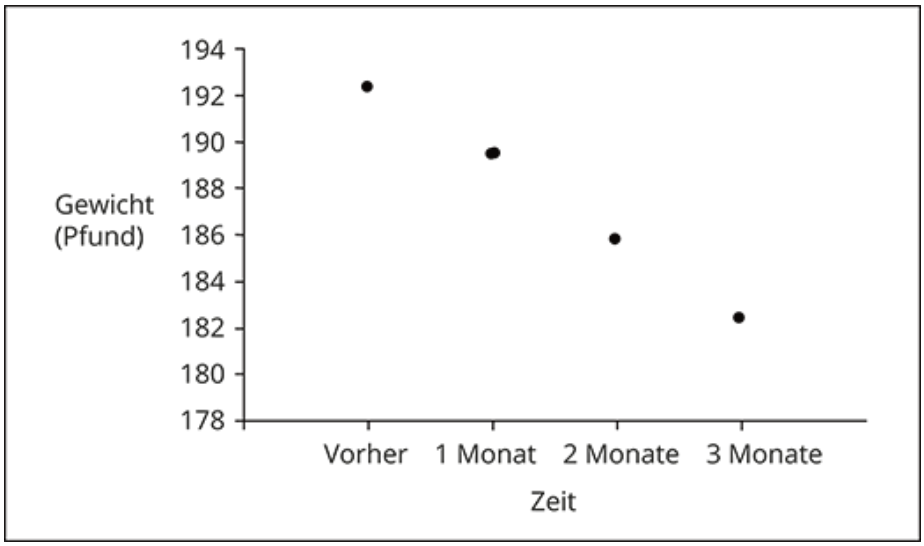


Abbildung 12.6: Die Mittelwerte für das Beispiel mit dem Diätprogramm

Ein Trend kann linear (wie er es in diesem Beispiel offensichtlich ist) oder nicht linear sein (dann bilden die Mittelwerte eine Kurve). In diesem Beispiel geht es nur um einen linearen Trend.

Zum Analysieren eines Trends verwenden Sie Vergleichskoeffizienten, die Zahlen, die in geplanten Vergleichen verwendet werden. Diese werden allerdings ein wenig anders eingesetzt.

Hier werden Vergleichskoeffizienten zum Ermitteln einer Quadratsumme für einen linearen Trend verwendet. Ich kürze das mit SS_{Linear} ab. Das ist ein Teil von $SS_{Spalten}$. Genau genommen:

$$SS_{Linear} + SS_{Nichtlinear} = SS_{Spalten}$$

Und:

$$df_{Linear} + df_{Nichtlinear} = df_{Spalten}$$

Wenn Sie SS_{Linear} berechnet haben, dividieren Sie den Wert durch df_{Linear}, um MS_{Linear} zu berechnen. Das ist kinderleicht, weil $df_{Linear} = 1$ ist. Dividieren Sie MS_{Linear} durch MS_{Fehler}, und schon haben Sie einen *F*-Quotienten. Wenn dieser *F*-Quotient größer ist als der kritische

Wert von *F* bei df = 1 und df_{Fehler} auf α-Ebene, dann nimmt das Gewicht während der Zeitdauer des Diätprogramms linear ab.

Die Vergleichskoeffizienten sind je nach Anzahl der Stichproben unterschiedlich. Bei vier Stichproben betragen die Koeffizienten beispielsweise −3, −1, 1 und 3. Die Formel zum Bilden von SS_{Linear} lautet wie folgt:

$$SS_{Linear} = \frac{n\left(\sum c\bar{x}\right)^2}{\sum c^2}$$

In dieser Formel steht *n* für die Anzahl der Personen und *c* für die Koeffizienten. Setzen wir die Zahlen aus diesem Beispiel in die Formel ein:

$$SS_{Linear} = \frac{n\left(\sum c\bar{x}\right)^2}{\sum c^2} = \frac{10\left[(-3)(192,4)+(-1)(189,5)+(1)(185,8)+(3)(182,4)\right]^2}{(-3)^2+(-1)^2+(3)^2+(1)^2}$$
$$= 567,845$$

Das ist ein so großer Teil von $SS_{Spalten}$, dass $SS_{Nichtlinear}$ einen wirklich kleinen Wert ergibt:

$$SS_{Nichtlinear} = SS_{Spalten} - SS_{Linear} = 569,075 - 567,845 = 1,23$$

Wie bereits erwähnt, ist df = 1, so dass $MS_{Linear} = SS_{Linear}$.

Und schließlich:

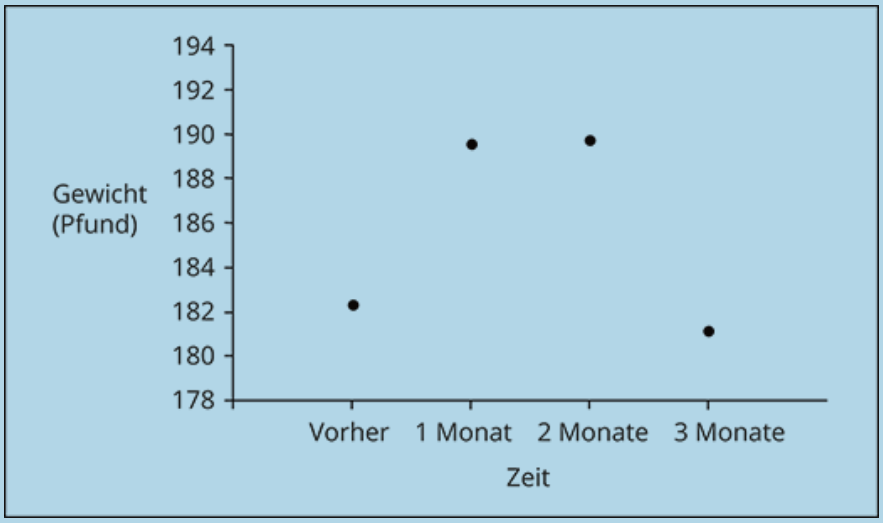
$$F = \frac{MS_{Linear}}{MS_{Fehler}} = \frac{567,85}{7,75} = 73,30$$

Für *F* beträgt der kritische Wert 4,21 bei df = 1 und 27 und α = 0,05 (den ich mit der Funktion `F.INV.RE` berechnet habe). Da der berechnete Wert größer ist als der kritische Wert, sagen Statistiker, die Daten zeigen eine *signifikant lineare Komponente*. Das bestätigt natürlich das, was in Abbildung 12.6 zu sehen ist.

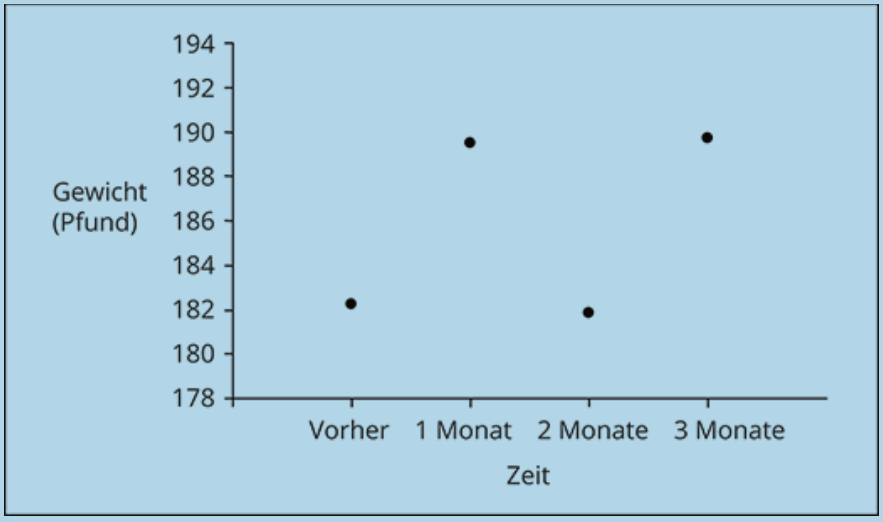
Noch ein bisschen Trend

Die Koeffizienten, die ich Ihnen gezeigt habe, stehen für eine mögliche Komponente dessen, was den Differenzen unter den vier Mittelwerten im Beispiel zugrunde liegt: die lineare Komponente. Bei vier Mittelwerten kann es auch noch andere Komponenten geben. Ich habe diese anderen Komponenten zu einer Kategorie zusammengefasst und als nicht linear bezeichnet. Diese werde ich nun ausführlich erläutern.

Eine Möglichkeit besteht darin, dass sich vier Mittelwerte voneinander unterscheiden und einen Trend bilden, der wie in der nachfolgenden Abbildung wie eine Kurve aussieht.



Vier Mittelwerte können aber auch noch einen anderen Trend bilden, wie die nachfolgende Abbildung zeigt:



Die erste Art, bei der der Trend die Richtung nur einmal ändert, wird als *quadratische Komponente* bezeichnet. In der ersten Abbildung nimmt sie erst

zu und dann ab. Die zweite Art, bei der der Trend die Richtung zweimal ändert, wird als *kubische Komponente* bezeichnet. In der zweiten Abbildung nimmt sie zu, ab und dann wieder zu. Der Trend in Abbildung 12.6 ist linear und ändert die Richtung nicht (er nimmt nur ab).

Quadratische und kubische Komponenten haben ebenfalls Koeffizienten:

Quadratisch: 1, −1, −1, 1

Kubisch: −1, 3, −3, 1

Diese Komponenten werden genauso wie die lineare Komponente getestet. Ein Trend kann sich aus mehreren Komponenten zusammensetzen: Wenn Sie einen signifikanten *F*-Quotienten vorliegen haben, ist mindestens eine dieser Trendkomponenten signifikant.

Bei vier Mittelwerten gibt es lineare, quadratische und kubische Komponenten. Mehr nicht. Bei fünf Mittelwerten gibt es zudem eine *biquadratische Komponente* (drei Richtungswechsel) und bei sechs Mittelwerten noch eine *quintische Komponente* (vier Richtungswechsel). Und wie sehen die Koeffizienten aus?

Bei fünf Mittelwerten:

Linear: −2, −1, 0 , 1, 2

Quadratisch: 2, −1, −2, −1, 2

Kubisch: −1, 2, 0, −2, 1

Biquadratisch: 1, −4, 6, −4, 1

Und bei sechs Mittelwerten:

Linear: −5, −3, −1 , 1, 3, 5

Quadratisch: 5, −1, −4, −4. −1, 5

Kubisch: −5, 7, 4, −4, −7, 5

Quartisch: 1, −3, 2, 2, −3, 1

Quintisch: −1, 5, −10, 10, −5, 1

Ich könnte mit noch mehr Mittelwerte, Koeffizienten und exotischen Komponentennamen (hexisch? septisch?) weitermachen, aber das reicht jetzt. Das müsste für eine Weile genug sein.

Datenanalysetool: Zweifaktorielle Varianzanalyse ohne

Messwiederholung

Wie bitte? Ist das ein Druckfehler? *Zweifaktoriell???* Ohne Messwiederholung?? Was ist das denn?

Das kann ich Ihnen sagen: Wenn Sie in den Datenanalysetools nach so etwas wie *Einfaktorielle Varianzanalyse mit wiederholten Messungen* suchen, werden Sie nicht fündig. Das Tool, das Sie suchen, gibt es, es verbirgt sich aber hinter einem anderen Namen.

In Abbildung 12.7 ist das Dialogfeld dieses Tools zusammen mit den Daten aus dem eben beschriebenen Beispiel mit dem Diätprogramm dargestellt.

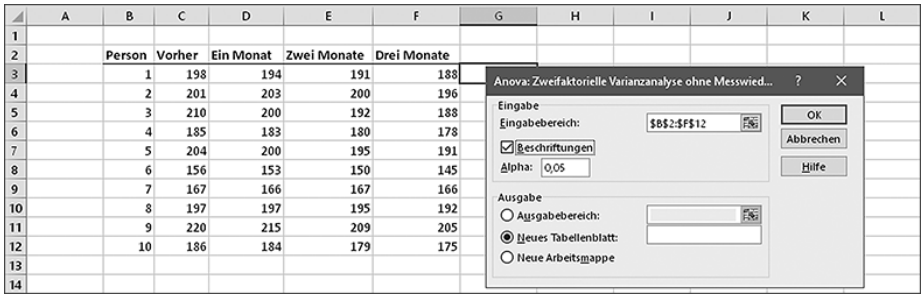


Abbildung 12.7: Das Datenanalysetool für die zweifaktorielle Varianzanalyse ohne Messwiederholung

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie die Daten für jede Stichprobe jeweils in einen eigenen Zellbereich ein. Geben Sie die Beschriftung für jede Person in eine Zelle ein.**

   In diesem Beispiel befinden sich die Beschriftungen für *Person* in Spalte B. Die Daten für die Stichprobe *Vorher* befinden sich in Spalte C, die Daten für die Stichprobe *1 Monat* in Spalte D, die Daten für die Stichprobe *2 Monate* in Spalte E und die Daten für die Stichprobe *3 Monate* in Spalte F.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG aus.**

4. **Klicken Sie auf OK, um das Dialogfeld ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG zu öffnen.**

   In Abbildung 12.7 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld EINGABEBEREICH den Zellbereich an, in dem sich die Werte befinden.**

   In diesem Beispiel befinden sich die Daten in den Zellen $B$2:$F$12. Beachten Sie die Dollarzeichen ($) für den absoluten Bezug. Beachten Sie außerdem – und das ist wichtig –, dass die Spalte *Person* Teil der Daten ist.

6. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich dieses Kontrollkästchen aktiviert.

7. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie den Wert, wenn Sie ein anderes α möchten.**

8. **Legen Sie im Bereich AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, so dass das Ergebnis auf einer neuen Seite im Arbeitsblatt angezeigt wird.

9. **Klicken Sie auf OK.**

Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird eine neue Seite mit dem Ergebnis erstellt.

In Abbildung 12.8 habe ich das von diesem Tool berechnete Ergebnis dargestellt, nachdem ich die Spaltenbreite angepasst habe. Das Ergebnis besteht aus zwei Tabellen: *ZUSAMMENFASSUNG* und *ANOVA*.

	A	B	C	D	E	F	G
1	Anova: Zweifaktorielle Varianzanalyse ohne Messwiederholung						
2							
3	*ZUSAMMENFASSUNG*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
4	1	4	771	192,75	18,25		
5	2	4	800	200	8,666666667		
6	3	4	790	197,5	94,33333333		
7	4	4	726	181,5	9,666666667		
8	5	4	790	197,5	32,33333333		
9	6	4	604	151	22		
10	7	4	666	166,5	0,333333333		
11	8	4	781	195,25	5,583333333		
12	9	4	849	212,25	43,58333333		
13	10	4	724	181	24,66666667		
14							
15	Vorher	10	1924	192,4	377,6		
16	Ein Monat	10	1895	189,5	342,9444444		
17	Zwei Monate	10	1858	185,8	298,8444444		
18	Drei Monate	10	1824	182,4	296,2666667		
19							
20							
21	ANOVA						
22	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
23	Zeilen	11631,725	9	1292,413889	166,8228756	2,70981E-21	2,250131477
24	Spalten	569,075	3	189,6916667	24,48512011	7,3047E-08	2,960351318
25	Zufallsfehler	209,175	27	7,747222222			
26							
27	Gesamt	12409,975	39				

Abbildung 12.8: Ergebnis aus dem Datenanalysetool für die zweifaktorielle Varianzanalyse ohne Messwiederholung

Die Tabelle *ZUSAMMENFASSUNG* besteht aus zwei Teilen. Der erste Teil enthält die zusammenfassenden Informationen für die Zeilen und der zweite Teil die für die Spalten. Zu diesen Informationen zählen die Anzahl der Werte in den Zeilen und Spalten sowie die Summen, Mittelwerte und Varianzen.

In der Tabelle *ANOVA* sind die Quadratsummen, die Freiheitsgrade, die mittlere Quadratsumme, die Prüfgröße *F*, die P-Werte und die kritischen *F*-Quotienten für die angegebenen Freiheitsgrade aufgeführt. In der Tabelle sind zwei Werte für *F* angegeben: ein *F*-Quotient für die Zeilen und ein weiterer für die Spalten. Der P-Wert ist der Anteil der Fläche, den der *F*-Quotient an der oberen Seite der *F*-Verteilung abschneidet. Wenn dieser Wert kleiner als 0,05 ist, wird H_0 verworfen.

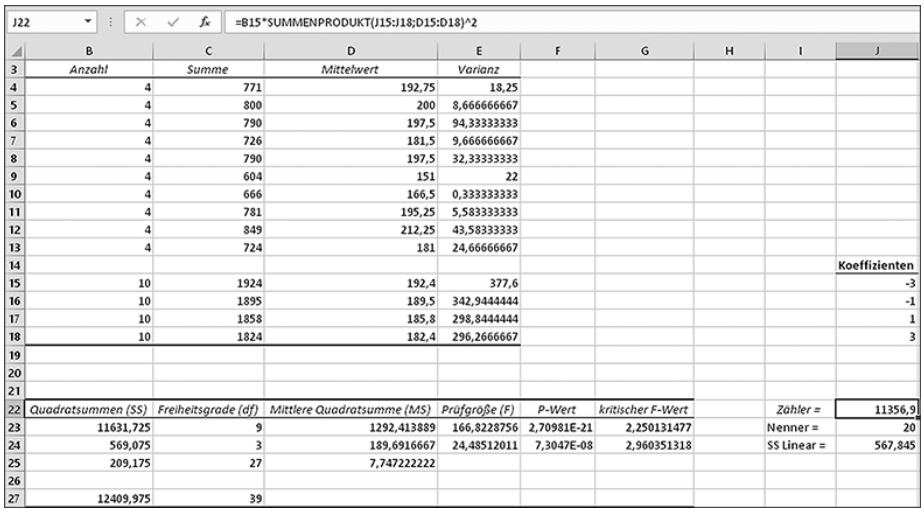
Auch wenn in der Tabelle *ANOVA* ein *F*-Quotient für die Zeilen angegeben ist, braucht Sie das in diesem Fall nicht zu kümmern, da es bei H_0 nur um die Spalten in den Daten geht. Jede Zeile enthält

die Daten einer Person. Ein großer *F*-Wert ist lediglich ein Hinweis darauf, dass die Personen verschieden sind, und das ist nichts Neues. Den Trend analysieren

Mit dem Excel-Tool für die zweifaktorielle Varianzanalyse ohne Messwiederholung kann keine Trendanalyse durchgeführt werden. Aber wie bei den geplanten Vergleichen kommen Sie auch hier mit ein bisschen Einfallsreichtum ein ganzes Stück weiter. Die Tabellenblattfunktionen `SUMMENPRODUKT` und `QUADRATESUMME` von Excel helfen bei der Berechnung.

Auf dem Tabellenblatt mit dem Ergebnis der Varianzanalyse finden Sie die Informationen, die Sie für den Anfang benötigen. In diesem Abschnitt werde ich die einzelnen Schritte der Analyse eines linearen Trends erläutern.

Zuerst gebe ich die Vergleichskoeffizienten für einen linearen Trend in die Zellen J15 bis J18 ein (siehe Abbildung 12.9).



J22: =B15*SUMMENPRODUKT(J15:J18;D15:D18)^2

	B	C	D	E	F	G	H	I	J
3	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*					
4	4	771	192,75	18,25					
5	4	800	200	8,666666667					
6	4	790	197,5	94,33333333					
7	4	726	181,5	9,666666667					
8	4	790	197,5	32,33333333					
9	4	604	151	22					
10	4	666	166,5	0,333333333					
11	4	781	195,25	5,583333333					
12	4	849	212,25	43,58333333					
13	4	724	181	24,66666667					
14									Koeffizienten
15	10	1924	192,4	377,6					-3
16	10	1895	189,5	342,9444444					-1
17	10	1858	185,8	298,8444444					1
18	10	1824	182,4	296,2666667					3
19									
20									
21									
22	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*		*Zähler =*	11356,9
23	11631,725	9	1292,413889	166,8228756	2,70981E-21	2,250131477		Nenner =	20
24	569,075	3	189,6916667	24,48512011	7,3047E-08	2,960351318		SS Linear =	567,845
25	209,175	27	7,747222222						
26									
27	12409,975	39							

Abbildung 12.9: Eine Trendanalyse durchführen

In den Zellen J22 bis J24 habe ich Informationen zu SS_{Linear} eingegeben: den Zähler, den Nenner und den Wert der Quadratsumme. Ich habe den Zähler und den Nenner in getrennte Zellen eingegeben, um die Formeln zu vereinfachen.

Wie bereits erwähnt, werden mit `SUMMENPRODUKT` Zahlen in Zellbereichen multipliziert und die Produkte addiert. (Diese Funktion

befindet sich im Menü der Schaltfläche MATHEMATIK UND TRIGONOMETRIE und nicht in der Kategorie STATISTIK.) Ich habe mit `SUMMENPRODUKT` die einzelnen Koeffizienten mit den einzelnen Mittelwerten multipliziert und anschließend die Produkte addiert. Das Ergebnis habe ich in J22 gespeichert. Das ist der Zähler für SS_{Linear}. Ich habe J22 markiert, damit Sie die Formel in der Bearbeitungsleiste sehen können:

```
=B15*SUMMENPRODUKT(J15:J18;D15:D18)^2
```

Der Wert in B15 gibt die Anzahl in jeder Spalte an. In den Zellen J15:J18 befinden sich die Vergleichskoeffizienten und in den Zellen D15:D18 die Spaltenmittelwerte.

In J23 steht der Nenner. Die Formel lautet:

```
=QUADRATESUMME(J15:J18)
```

Mit `QUADRATESUMME` werden die Koeffizienten in J15:J18 quadriert und die Quadrate anschließend addiert.

In J24 befindet sich der Wert für SS_{Linear}. Dieser entsteht, wenn J22 durch J23 dividiert wird.

In Abbildung 12.9 ist zu sehen, dass ich in die Tabelle ANOVA über der Zeile für den Zufallsfehler zwei Zeilen eingefügt habe. In diesen Zeilen befinden sich die Quadratsummen, die Freiheitsgrade, die mittleren Quadratsummen, die *F*-Quotienten, die P-Werte und die kritischen *F*-Werte für den linearen und für den nicht linearen Trend. $SS_{Nichtlinear}$ in B26 berechnet sich aus B24 minus B25.

Der *F*-Quotient für den linearen Trend berechnet sich aus D25 dividiert durch D27. Die Formel für den P-Wert in F25 lautet:

```
=F.VERT.RE(E25;C25;C27)
```

Das erste Argument, E25, ist der *F*-Quotient. Das zweite und dritte Argument sind die df-Werte.

Die Formel für den kritischen *F*-Wert in G25 lautet:

```
=F.INV.RE(0,05;C25;C27)
```

Das erste Argument ist α und das zweite und dritte sind die df-Werte.

Kapitel 13

Etwas komplexere Tests

In diesem Kapitel

- Mit zwei Variablen arbeiten
- Mit Messwiederholungen arbeiten
- Grundlegendes zu Wechselwirkungen
- Variablentypen mischen

In Kapitel 11 haben Sie erfahren, wie Sie Hypothesen mit zwei Stichproben testen können, und in Kapitel 12, wie Sie Hypothesen mit mehr als zwei Stichproben testen. Aus diesen beiden Kapiteln wissen Sie bereits, dass jeweils eine unabhängige Variable (auch als *Faktor* bezeichnet) beteiligt ist.

Nun kommt es häufig vor, dass die Auswirkungen mehrerer Faktoren getestet werden müssen. In diesem Kapitel erfahren Sie, wie zwei Faktoren innerhalb derselben Daten analysiert werden. Es gibt mehrere mögliche Situationen und ich beschreibe die Excel-Datenanalysetools für den Umgang mit all diesen Situationen.

Die Kombinationen knacken

FarKlempt Robotics stellt batteriebetriebene Mikroroboter her. Das Unternehmen möchte drei wiederaufladbare Batterien für diese Roboter an mehreren Aufgaben (Klettern, Gehen und Zusammenbauen) testen. Welche Kombination aus Batterie und Aufgabenstellung gewährleistet die längste Batterielebensdauer?

Es wird eine Stichprobe von neun Robotern getestet. Jedem Roboter werden nach Zufallsprinzip eine Batterie und eine Aufgabe zugewiesen. FarKlempt ermittelt die Anzahl der Tage, die jeder Roboter arbeitet, bevor die Batterie aufgeladen werden muss. Die Daten sind in Tabelle 13.1 aufgeführt.

Aufgabe	Batterie 1	Batterie 2	Batterie 3	Mittelwert
Klettern	12	15	20	15,67
Gehen	14	16	19	16,33
Zusammenbauen	11	14	18	14,33
Mittelwert	12,33	15,00	19,00	15,44

Tabelle 13.1: FarKlempt-Roboter: Anzahl der Tage bis zum Aufladen bei drei Aufgaben und drei Batterien

Das verlangt nach zwei Hypothesentests:

H_0: $\mu_{Batterie1} = \mu_{Batterie2} = \mu_{Batterie3}$
H_1: Nicht H_0

und

H_0: $\mu_{Klettern} = \mu_{Gehen} = \mu_{Zusammenbauen}$
H_1: Nicht H_0

Legen Sie bei beiden Tests für $\alpha = 0{,}05$ fest.

Die Varianzen analysieren

Die für diese Tests geeignete Analyse ist eine Varianzanalyse (ANOVA). Dabei werden die Variablen *Batterien* und *Aufgaben* auch als *Faktor* bezeichnet. Somit handelt es sich hierbei um eine *zweifaktorielle Varianzanalyse*.

Um diese Varianzanalyse zu verstehen, betrachten wir die Varianzen innerhalb der Daten. Zunächst konzentrieren wir uns auf die Varianz in allen neun Zahlen: MS_G. (Dabei steht »G« für »Gesamt«.) Der

Mittelwert dieser Zahlen beträgt 15,44. Da es sich um den Mittelwert aller Zahlen handelt, wird er als *Gesamtmittelwert* bezeichnet.

Diese Varianz lautet folgendermaßen:

$$MS_G = \frac{(12-15{,}44)^2 + (15-15{,}44)^2 + \ldots + (18-15{,}44)^2}{9-1} = \frac{76{,}22}{8} = 9{,}53$$

Die Mittelwerte der drei Batterien (die Spaltenmittelwerte) streuen ebenfalls um 15,44. Diese Varianz lautet folgendermaßen:

$$MS_{Batterien} = \frac{(3)\,(12{,}33-15{,}44)^2 + (3)\,(15{,}00-15{,}44)^2 + (3)\,(19{,}00-15{,}44)^2}{3-1}$$
$$= \frac{67{,}62}{2} = 33{,}81$$

Warum ist die 3 als Multiplikator für die quadratischen Abweichungen angegeben? Weil bei Mittelwerten die Anzahl der Werte, aus denen sich die Mittelwerte zusammensetzen, berücksichtigt werden muss.

Und auch die Mittelwerte der Aufgaben (die Zeilenmittelwerte) streuen um 15,44:

$$MS_{Aufgaben} = \frac{(3)\,(15{,}67-15{,}44)^2 + (3)\,(16{,}33-15{,}44)^2 + (3)\,(14{,}33-15{,}44)^2}{3-1}$$
$$= \frac{6{,}23}{2} = 3{,}12$$

Bleibt noch eine Varianz. Diese wird als MS_{Fehler} bezeichnet. Sie ergibt sich, wenn $SS_{Batterien}$ und $SS_{Aufgaben}$ von SS_{Gesamt} subtrahiert werden und das Ergebnis durch die Freiheitsgrade dividiert wird, die sich bei der Subtraktion von $df_{Batterien}$ und $df_{Aufgaben}$ von df_{Gesamt} ergeben.

$$MS_{Fehler} = \frac{SS_G - SS_{Batterien} - SS_{Aufgaben}}{df_{Gesamt} - df_{Batterien} - df_{Aufgaben}} = \frac{2{,}44}{4} = 0{,}61$$

Um die Hypothesen zu testen, berechnen Sie einen *F*-Quotienten für die Auswirkungen der Batterien und einen für die Auswirkungen der Aufgaben. Bei beiden Berechnungen ist der Nenner (der so genannte *Fehlerterm*) MS_{Fehler}:

$$F = \frac{MS_{Batterien}}{MS_{Fehler}} = \frac{33,81}{0,61} = 55,42$$
$$F = \frac{MS_{Aufgaben}}{MS_{Fehler}} = \frac{3,12}{0,61} = 5,11$$

Für beide *F*-Quotienten gelten die Freiheitsgrade 2 und 4. Bei α = 0,05 beträgt der kritische *F*-Wert jeweils 6,94. Im Hinblick auf die Batterien (sie unterscheiden sich stärker voneinander, als dass dies Zufall sein könnte) wird H_0 verworfen, im Hinblick auf die Aufgaben jedoch nicht.

Um die Differenzen bei den Batterien näher zu untersuchen, führen Sie geplante Vergleiche unter den Spaltenmittelwerten durch. (Ausführlichere Informationen hierzu finden Sie in Kapitel 12.)

Datenanalysetool: Zweifaktorielle Varianzanalyse ohne Messwiederholung

Mit dem Excel-ANOVA-Tool für die zweifaktorielle Varianzanalyse ohne Messwiederholung können Sie die eben beschriebene Analyse durchführen. (Dieses Tool habe ich in Kapitel 12 bereits für eine andere Analyse verwendet.) *Ohne Messwiederholung* bedeutet, dass jeder Batterie-Aufgaben-Kombination nur ein Roboter zugewiesen wird. Wenn jeder Kombination mehrere Roboter zugewiesen würden, spräche man von *Messwiederholung*.

In Abbildung 13.1 ist das Dialogfeld für dieses Tool zusammen mit den Daten für das Batterien-Aufgaben-Beispiel dargestellt.

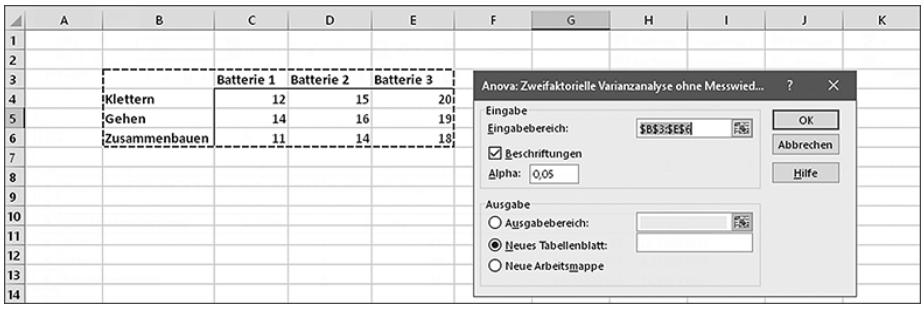


Abbildung 13.1: Das Datenanalysetool für die zweifaktorielle Varianzanalyse ohne Messwiederholung mit den Batterien-Aufgaben-Daten

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie die Daten sowie Beschriftungen für die Zeilen und Spalten in das Arbeitsblatt ein.**

   Bei diesem Beispiel befinden sich die Beschriftungen für die Aufgaben in den Zellen B4, B5 und B6. Die Beschriftungen für die Batterien befinden sich in den Zellen C3, D3 und E3. Die Daten befinden sich in den Zellen C4 bis E6.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG aus.**

4. **Klicken Sie auf OK, um das Dialogfeld ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG zu öffnen.**

   Dies ist das Dialogfeld in Abbildung 13.1.

5. **Geben Sie im Feld EINGABEBEREICH den Zellbereich an, in dem sich die Werte befinden.**

   In diesem Beispiel befinden sich die Daten in den Zellen $B$3:$E$6. Beachten Sie die Dollarzeichen ($) für den absoluten Bezug. Beachten Sie außerdem – und das ist wichtig –, dass sowohl die Zeilen- als auch die Spaltenbeschriftungen

Teil der Daten sind. Die erste Zelle im Eingabebereich B3 ist leer. Dies stellt jedoch kein Problem dar.

6. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich auch diese Option aktiviert.

7. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie den Wert, wenn Sie ein anderes α möchten.**

8. **Legen Sie im Bereich AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

9. **Klicken Sie auf OK.**

   Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird eine neue Seite mit dem Ergebnis erstellt.

In Abbildung 13.2 ist das von diesem Tool berechnete Ergebnis mit angepassten Spaltenbreiten dargestellt. Das Ergebnis besteht aus zwei Tabellen: *ZUSAMMENFASSUNG* und *ANOVA*.

Die Tabelle *ZUSAMMENFASSUNG* besteht aus zwei Teilen. Der erste Teil enthält die zusammenfassenden Informationen für die Zeilen und der zweite Teil die für die Spalten. Zu diesen Informationen zählen die Anzahl der Werte in den Zeilen und Spalten sowie die Summen, Mittelwerte und Varianzen.

In der Tabelle *ANOVA* sind die Quadratsummen, die Freiheitsgrade, die mittlere Quadratsumme, die Prüfgröße *F*, der P-Wert und der kritische *F*-Wert für den angegebenen df-Wert aufgeführt. In der Tabelle sind zwei Werte für *F* angegeben: ein *F*-Quotient für die Zeilen und einer für die Spalten. Der P-Wert ist der Anteil der Fläche,

den der *F*-Quotient an der oberen Seite der *F*-Verteilung abtrennt. Wenn dieser Wert kleiner als 0,05 ist, wird H_0 verworfen.

In diesem Beispiel wird H_0 für die Batterien (die Spalten) verworfen, für die Aufgaben (die Zeilen) nicht.

	A	B	C	D	E	F	G
1	Anova: Zweifaktorielle Varianzanalyse ohne Messwiederholung						
2							
3	*ZUSAMMENFASSUNG*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
4	Klettern	3	47	15,66666667	16,33333333		
5	Gehen	3	49	16,33333333	6,333333333		
6	Zusammenbauen	3	43	14,33333333	12,33333333		
7							
8	Batterie 1	3	37	12,33333333	2,333333333		
9	Batterie 2	3	45	15	1		
10	Batterie 3	3	57	19	1		
11							
12							
13	ANOVA						
14	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
15	Zeilen	6,222222222	2	3,111111111	5,090909091	0,079552926	6,94427191
16	Spalten	67,55555556	2	33,77777778	55,27272727	0,001219451	6,94427191
17	Zufallsfehler	2,444444444	4	0,611111111			
18							
19	Gesamt	76,22222222	8				

Abbildung 13.2: Ergebnis aus dem Datenanalysetool für die zweifaktorielle Varianzanalyse ohne Messwiederholung

Noch einmal Kombinationen knacken

Bei der vorherigen Analyse wird für jede Kombination aus den beiden Faktoren ein Wert berücksichtigt. Jeder Kombination ein Objekt zuzuweisen, ist bei Robotern und anderen Gegenständen in Ordnung, die mehr oder weniger identisch sind.

Wenn Menschen beteiligt sind, sieht die Sache etwas anders aus. Die individuellen Eigenheiten bei Menschen dürfen nicht einfach ignoriert werden. Daher muss einer Kombination aus Faktoren eine Stichprobe von Menschen zugewiesen werden und nicht nur eine Person.

Zeilen und Spalten

Ich werde das anhand eines Beispiels erklären. Stellen Sie sich vor, ein Unternehmen wendet zwei Methoden zum Präsentieren von Schulungsunterlagen an. Bei einer Methode hält eine Person einen

Vortrag, bei der anderen bestehen die Schulungsunterlagen aus Text. Stellen Sie sich außerdem vor, dass die in den Unterlagen enthaltenen Informationen entweder humorvoll oder technisch präsentiert werden. Den ersten Faktor bezeichne ich als Präsentationsmethode und den zweiten als Präsentationsstil.

Die Kombination aus den jeweils zwei Ausprägungen der Präsentationsmethode und der Präsentationsstile ergibt vier Möglichkeiten. Das Unternehmen weist jeder Kombinationsmöglichkeit vier Personen zu. Das ergibt insgesamt 16 Personen. Nach der Schulung werden die 16 Personen daraufhin getestet, was sie von der Schulung verstanden haben.

In Abbildung 13.3 sind die Kombinationen, die vier Werte für den Lernerfolg innerhalb der einzelnen Kombinationen und eine Zusammenfassung der Werte für die Kombinationen, Zeilen und Spalten dargestellt.

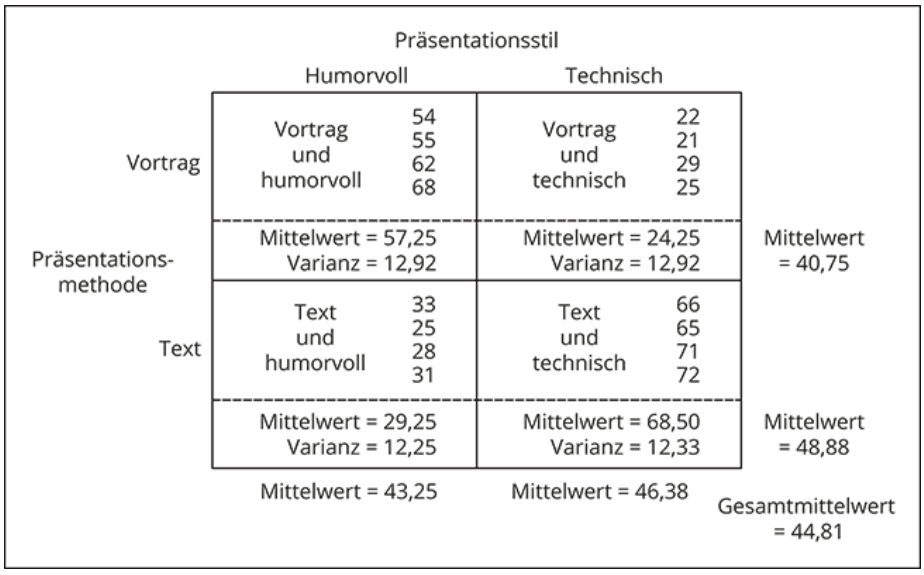


Abbildung 13.3: Kombination aus den Ausprägungen der Präsentationsmethode mit den Ausprägungen des Präsentationsstils

Die Hypothesen lauten wie folgt:

H_0: $\mu_{Vortrag} = \mu_{Text}$

H_1: Nicht H_0

und

H_0: $\mu_{Humorvoll} = \mu_{Technisch}$

H_1: Nicht H_0

Da sich die beiden Präsentationsmethoden (Vortrag und Text) in den Zeilen befinden, bezeichne ich den Präsentationstyp als *Zeilenfaktor*. Die beiden Präsentationsstile (Humorvoll und Technisch) befinden sich in den Spalten, so dass der Präsentationsstil *Spaltenfaktor*genannt wird.

Wechselwirkungen

Wenn Sie Daten in Zeilen und Spalten vorliegen haben und Hypothesen zum Zeilenfaktor und zum Spaltenfaktor testen, müssen Sie noch etwas berücksichtigen. Und zwar müssen Sie sich um die Zeilen-Spalten-Kombinationen Gedanken machen. Haben die Kombinationen bestimmte Auswirkungen?

Bei dem Beispiel von oben ist es möglich, dass die Kombination aus Vortrag/Text und Humorvoll/Technisch ein unerwartetes Ergebnis hervorbringt. Anhand der Daten in Abbildung 13.3 können Sie es sehen: Beim Vortrag führt der humorvolle Stil zu einem höheren Durchschnittswert als der technische Stil. Beim Text führt der humorvolle Stil zu einem geringeren Durchschnittswert als der technische Stil.

Eine Situation wie diese wird als *Wechselwirkung* oder auch *Interaktion* bezeichnet. Formal ausgedrückt entsteht eine Wechselwirkung, wenn sich die Ausprägungen eines Faktors auf die Ausprägungen des anderen Faktors anders auswirken. Die Wechselwirkung wird mit Zeilenfaktor × Spaltenfaktor bezeichnet, in diesem Beispiel also mit Methode × Stil.

Die Hypothesen hierfür lauten wie folgt:

H_0: Zwischen Präsentationsmethode und Präsentationsstil gibt es keine Wechselwirkung.

H_1: Nicht H_0

Die Analyse

Die statistische Analyse ist wieder einmal eine Varianzanalyse (ANOVA). Wie bei den anderen Varianzanalysen, die Sie kennengelernt haben, geht es um die Varianzen in den Daten.

Die erste Varianz ist die Gesamtvarianz MS_G. Dies ist die Varianz aller 16 Werte um deren Mittelwert (*Gesamtmittelwert*) 44,81:

$$MS_G = \frac{(54-44{,}81)^2 + (55-44{,}81)^2 + \ldots + (72-44{,}81)^2}{16-1}$$
$$= \frac{5674{,}44}{15} = 378{,}30$$

Der Nenner besagt, dass df für MS_G 15 beträgt.

Die nächste Varianz stammt vom Zeilenfaktor. Das ist $MS_{Methode}$ und ist die Varianz der Zeilenmittelwerte um den Gesamtmittelwert:

$$MS_{Methode} = \frac{(8)(40{,}75-44{,}81)^2 + (8)(48{,}44-44{,}81)^2}{2-1} = \frac{264{,}06}{1} = 264{,}06$$

Jede quadratische Abweichung wird mit 8 multipliziert, da die Anzahl der Werte berücksichtigt werden muss, aus denen sich die Zeilenmittelwerte zusammensetzen. Der df-Wert für $MS_{Methode}$ entspricht der Anzahl an Zeilen –1, also 1.

Entsprechend berechnet sich die Varianz für den Spaltenfaktor wie folgt:

$$MS_{Stil} = \frac{(8)(43{,}25-44{,}81)^2 + (8)(46{,}38-44{,}81)^2}{2-1} = \frac{39{,}06}{1} = 39{,}06$$

Der df-Wert für MS_{Stil} beträgt 1 (die Anzahl der Spalten – 1).

Eine weitere Varianz ist die gepoolte Schätzung basierend auf den Varianzen innerhalb der vier Zeile-Spalte-Kombinationen. Diese Varianz wird als $MS_{Innerhalb}$ oder MS_I bezeichnet. (Ausführlichere Informationen zu MS_I und gepoolten Schätzwerten finden Sie in Kapitel 12.) Für dieses Beispiel gilt:

$$MS_I = \frac{(4-1)(12{,}92)+(4-1)(12{,}92)+(4-1)(12{,}25)+(4-1)(12{,}33)}{(4-1)+(4-1)+(4-1)+(4-1)}$$
$$= \frac{151{,}25}{12} = 12{,}60$$

Dies ist der Fehlerterm (Nenner) für die einzelnen berechneten *F*-Quotienten. Der Nenner besagt, dass df für MS_I 12 beträgt.

Die letzte Varianz stammt von der Wechselwirkung zwischen dem Zeilenfaktor und dem Spaltenfaktor. In diesem Beispiel wird sie mit $MS_{Methode \times Stil}$ bezeichnet. Es gibt mehrere Möglichkeiten, diese Varianz zu berechnen. Die einfachste Möglichkeit besteht darin, folgende allgemeine Beziehung zu nutzen:

$$SS_{Zeile \times Spalte} = SS_G - SS_{Zeilenfaktor} - SS_{Spaltenfaktor} - SS_I$$

Oder diese hier:

$$df_{Zeile \times Spalte} = df_G - df_{Zeilenfaktor} - df_{Spaltenfaktor} - df_I$$

Eine weitere Berechnungsmöglichkeit:

$$df_{Fehler \times Spalte} = (\text{Anzahl der Zeilen} - 1)\,(\text{Anzahl der Spalten} - 1)$$

Die mittlere Quadratsumme lautet

$$MS_{Zeile \times Spalte} = \frac{SS_{Zeile \times Spalte}}{df_{Zeile \times Spalte}}$$

Für dieses Beispiel gilt:

$$MS_{Methode \times Stil} = \frac{SS_{Methode \times Stil}}{df_{Methode \times Stil}} = \frac{5764,44 - 264,06 - 39,06 - 151,25}{15 - 12 - 1 - 1}$$
$$= \frac{5220,06}{1} = 5220,06$$

Zum Testen der Hypothesen werden *F*-Quotienten berechnet:

$$F = \frac{MS_{Stil}}{MS_I} = \frac{39,06}{12,60} = 3,10$$

$$F = \frac{MS_{Methode}}{MS_I} = \frac{264,06}{12,60} = 20,95$$

$$F = \frac{MS_{Methode \times Stil}}{MS_I} = \frac{5220,06}{12,60} = 414,15$$

Für df = 1 und 12 beträgt der kritische *F*-Wert 4,75 bei α = 0,05. (Sie können das mit der Excel-Funktion `F.INV.RE` überprüfen.) H_0 wird also für die Präsentationsmethode und für die Wechselwirkung zwischen Methode und Stil verworfen, für den Präsentationsstil jedoch nicht.

Datenanalysetool: Zweifaktorielle Varianzanalyse mit Messwiederholung

Excelenthält ein Datenanalysetool, das sich um alles kümmert. Dieses Tool trägt den Namen *Zweifaktorielle Varianzanalyse mit Messwiederholung*. »Messwiederholung« bedeutet, dass sich in jeder Zeile-Spalte-Kombination mehrere Werte befinden.

In Abbildung 13.4 ist das Dialogfeld für dieses Tool zusammen mit den Daten für das Stil-Methode-Beispiel dargestellt.

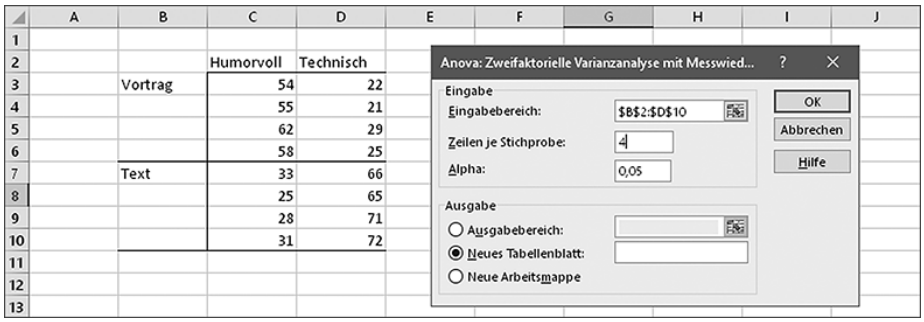


Abbildung 13.4: Das Datenanalysetool für die zweifaktorielle Varianzanalyse mit Messwiederholung mit den Stil-Methode-Daten

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie die Daten sowie die Beschriftungen für die Zeilen und Spalten in das Arbeitsblatt ein.**

   Bei diesem Beispiel befinden sich die Beschriftungen für die Präsentationsmethoden in den Zellen B3 und B7. Die Beschriftungen für den Präsentationsstil befinden sich in den Zellen C2 und D2. Die Daten befinden sich in den Zellen C3 bis D10.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE MIT MESSWIEDERHOLUNG aus.**

4. **Klicken Sie auf OK, um das Dialogfeld ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE MIT MESSWIEDERHOLUNG zu öffnen.**

   Dieses Dialogfeld zeigt Abbildung 13.4.

5. **Geben Sie im Feld EINGABEBEREICH den Zellbereich an, in dem sich die Werte befinden.**

   In diesem Beispiel befinden sich die Daten in den Zellen $B$2:$D$10. Beachten Sie die Dollarzeichen ($) für den

absoluten Bezug. Beachten Sie außerdem – und das ist wichtig –, dass die Beschriftungen für den Zeilenfaktor (Präsentationsmethode) Teil der Daten sind. Dies trifft auch für den Spaltenfaktor zu. Die erste Zelle im Zellbereich B2 ist leer. Dies stellt kein Problem dar.

6. **Im Feld ZEILEN PRO STICHPROBE geben Sie an, wie viele Beobachtungen Sie für jede Zeile-Spalte-Kombination haben.**

   In diesem Fall sind das **4**.

7. **Im Feld ALPHA befindet sich standardmäßig der Wert 0,05. Ändern Sie den Wert, wenn Sie ein anderes α möchten.**

8. **Legen Sie im Bereich AUSGABE fest, wo das Ergebnis ausgegeben werden soll.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

9. **Klicken Sie auf OK.**

   Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird das neue Arbeitsblatt mit dem Ergebnis erstellt und geöffnet.

In Abbildung 13.5 ist das von diesem Tool berechnete Ergebnis mit erweiterten Spalten dargestellt. Das Ergebnis besteht aus zwei Tabellen: ZUSAMMENFASSUNG und ANOVA.

	A	B	C	D	E	F	G
1	Anova: Zweifaktorielle Varianzanalyse mit Messwiederholung						
2							
3	ZUSAMMENFASSUNG	Humorvoll	Technisch	Gesamt			
4	*Vortrag*						
5	Anzahl	4	4	8			
6	Summe	229	97	326			
7	Mittelwert	57,25	24,25	40,75			
8	Varianz	12,91666667	12,91666667	322,2142857			
9							
10	*Text*						
11	Anzahl	4	4	8			
12	Summe	117	274	391			
13	Mittelwert	29,25	68,5	48,875			
14	Varianz	12,25	12,33333333	450,6964286			
15							
16	*Gesamt*						
17	Anzahl	8	8				
18	Summe	346	371				
19	Mittelwert	43,25	46,375				
20	Varianz	234,7857143	570,2678571				
21							
22							
23	ANOVA						
24	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
25	Stichprobe	264,0625	1	264,0625	20,95041322	0,000635547	4,747225347
26	Spalten	39,0625	1	39,0625	3,099173554	0,10377106	4,747225347
27	Wechselwirkung	5220,0625	1	5220,0625	414,153719	1,13844E-10	4,747225347
28	Fehler	151,25	12	12,60416667			
29							
30	Gesamt	5674,4375	15				

Abbildung 13.5: Ergebnis aus dem Datenanalysetool für die zweifaktorielle Varianzanalyse mit Messwiederholung

Die Tabelle *ZUSAMMENFASSUNG* besteht aus zwei Teilen. Der erste Teil enthält die zusammenfassenden Informationen für die Faktorkombinationen und für den Zeilenfaktor und der zweite die für den Spaltenfaktor. Zu diesen Informationen zählen die Anzahl der Werte in jeder Zeile-Spalte-Kombination, in jeder Zeile und in jeder Spalte sowie die Anzahl, Summen, Mittelwerte und Varianzen.

In der Tabelle *ANOVA* sind die Quadratsummen, die Freiheitsgrade, die mittlere Quadratsumme, die Prüfgröße *F*, der P-Wert und der kritische *F*-Wert für den angegebenen df-Wert aufgeführt. In der Tabelle sind drei Werte für *F* angegeben: ein *F*-Quotient für den Zeilenfaktor, ein *F*-Quotient für den Spaltenfaktor und ein *F*-Quotient für die Wechselwirkung. In der Tabelle wird der Zeilenfaktor als Stichprobe bezeichnet. Der P-Wert ist der Anteil der Fläche, den der *F*-Quotient an der oberen Seite der *F*-Verteilung abschneidet. Wenn dieser Wert kleiner als 0,05 ist, wird H_0 verworfen.

Bei diesem Beispiel wird H_0 für die Präsentationsmethode (der in der Tabelle als Stichprobe bezeichnete Zeilenfaktor) und für die Wechselwirkung verworfen, für den Präsentationsstil (Spaltenfaktor) nicht.

Zwei Arten von Variablen, und zwar gleichzeitig

Was passiert, wenn Sie sowohl eine Zwischengruppenvariable als auch eine Innergruppenvariable vorliegen haben, und zwar zur gleichen Zeit? (Dies wird gemischtes Design genannt.) Wie kann dies passieren?

Ganz einfach. Hier ein Beispiel. Angenommen, Sie wollen den Einfluss von Präsentationsmedien auf die Lesegeschwindigkeit von Viertklässlern untersuchen. Sie weisen Ihre Viertklässler (ich nenne sie Subjekte), zufällig einer der Gruppen E-Reader beziehungsweise Buch zu. Dies ist die Zwischengruppenvariable.

Nehmen wir weiterhin an, dass Sie auch die Auswirkungen der verwendeten Schriftart untersuchen wollen. Daher sollen alle Subjekte den Text in diesen drei Schriftarten lesen: Haettenschweiler, Arial und Calibri. (Ich habe noch nie ein Dokument gesehen, das komplett in Haettenschweiler formatiert war. Dennoch ist dies meine bevorzugte Schriftart, da sich der Name recht lustig anhört.) Da jedes Subjekt alle drei Schriftarten liest, ist dies die Innergruppenvariable. Aus Gründen der Vollständigkeit müssen Sie die Reihenfolge, in der die Subjekte den Text in einer der Schriftarten lesen, zufällig anordnen.

Wie würde eine entsprechende ANOVA-Tabelle aussehen? Tabelle 13.2 zeigt einen allgemeinen Aufbau. Die Tabelle ist in einen Satz von Quellen unterteilt, die die Variabilität zwischen den Gruppen ausmacht und einen weiteren Satz, die die Variabilität innerhalb der Gruppen beschreibt.

In der Kategorie *Zwischen* ist A der Name der Zwischengruppenvariablen. »S/A« können Sie als »S innerhalb A« lesen. Dies sagt aus, dass Personen in einer Stufe von A sich von denen in anderen Stufen unterscheiden.

In der Kategorie *Innerhalb* ist B der Name der Innergruppenvariablen. A × B beschreibt die Interaktion zwischen

den beiden Variablen. B × S/A ist so etwas wie die B-Variable, die mit den Subjekten in A in einer Wechselwirkung steht. Wie Sie sehen können, fällt alles, das mit B verknüpft wird, in die Kategorie »innerhalb der Gruppen«.

Quelle
Zwischen
Innerhalb
Gesamt

Tabelle 13.2: Die ANOVA-Tabelle für gemischtes ANOVA (eine Zwischengruppenvariable und eine Innergruppenvariable)

Das Erste, was hier auffällt, sind die drei Prüfgrößen, die F-Quotienten. Der erste testet die Unterschiede in den Stufen von A, der zweite testet die Unterschiede in den Stufen von B und der dritte testet die Wechselwirkung zwischen diesen beiden. Beachten Sie außerdem, dass sich der Nenner des ersten F-Quotienten von dem der anderen beiden Quotienten unterscheidet. Dies tritt immer öfter auf, je komplexer die ANOVAs werden.

Außerdem ist es wichtig, sich einige der Beziehungen bewusst zu machen. Auf der obersten Ebene sind dies:

$$SS_{Zwischen} + SS_{Innerhalb} + = SS_{Gesamt}$$
$$df_{Zwischen} + df_{Innerhalb} = df_{Gesamt}$$

Die Komponente *Zwischen den Gruppen* ist folgendermaßen untergliedert:

$$SS_{A} + SS_{S/A} = SS_{Zwischen}$$
$$df_{A} + df_{S/A} = df_{Zwischen}$$

Auch die Komponente *Innerhalb der Gruppen* ist untergliedert:

$$SS_B + SS_{A \times B} + SS_{B \times S/A} = SS_{Innerhalb}$$
$$df_B + df_{A \times B} \; df_{B \times S/A} = df_{Innerhalb}$$

Mit dem Wissen um diese Beziehungen können Sie ANOVA-Tabelle vervollständigen, nachdem Excel seine Arbeit erledigt hat.

Excel mit einem gemischten Design verwenden

Irgendwann einmal enthält die Werkzeugsammlung mit den Analyse-Funktionen vielleicht eine Option mit dem Namen *ANOVA: Gemischtes Design*. Leider ist dies heute noch nicht der Fall. Stattdessen verwende ich zwei der Werkzeuge aus den Analyse-Funktionen und das Wissen um diese Art von Design, um die Analyse durchzuführen.

Ich beginne mit den Daten der Untersuchung, die ich weiter vorne beschrieben habe. Abbildung 13.6 zeigt das Arbeitsblatt mit den Daten. Die Stufen der Zwischengruppenvariablen, Medium (die Variable A), befinden sich in der linken Spalte. Die Stufen der Innergruppenvariablen, Schriftart (die Variable B), befinden sich in der obersten Zeile. Die gemessenen Werte in den Zellen geben die Lesegeschwindigkeit in Wörtern je Minute an.

	A	B	C	D	E	F	G
1			Subjekt	Haettenschweiler	Arial	Calibri	
2	Buch		1	48	40	38	
3			2	55	43	45	
4			3	46	45	44	
5			4	61	53	53	
6	E-Reader		5	43	45	47	
7			6	50	52	54	
8			7	56	57	57	
9			8	53	53	55	
10							

Abbildung 13.6: Daten einer Studie mit einer Zwischengruppenvariablen und einer Innergruppenvariablen

Abbildung 13.7 zeigt die fertige ANOVA-Tabelle.

Wie habe ich diese erstellt? Überraschenderweise ist dies recht einfach, obwohl hierfür zahlreiche Schritte erforderlich sind. Alles, was Sie machen müssen, ist, zwei ANOVAs für die gleichen Daten auszuführen und anschließend die beiden ANOVA-Tabellen miteinander zu kombinieren. Die Beziehungen, die ich im vorherigen Abschnitt beschrieben habe, sind die fehlenden Teile in diesem Puzzle.

	A	B	C	D	E	F	G
1							
2	ANOVA						
3	*Streuungsursache*	*SS*	*df*	*MS*	*F*	*P-Wert*	*kritischer F-Wert*
4	Zwischen den Gruppen	638,2916667	*7*	91,18452381			
5	Medium	108,375	1	108,375	1,22707973	0,310395735	5,987378
6	Subjekt/Medium	529,9166667	6	88,31944444			
7							
8	Innerhalb der Gruppen	202,6666667	*16*				
9	Schriftart	40,08333333	2	20,04166667	5,681102	0,01836557	3,885294
10	Medium × Schriftart	120,25	2	60,125	17,043307	0,00031163	3,885294
11	Schriftart × Subjekt/Medium	42,33333333	12	3,527777778			
12							
13	Gesamt	840,9583333	23				
14							

Abbildung 13.7: Die fertige ANOVA-Tabelle für die Analyse der Daten aus Abbildung 13.6

Führen Sie diese Schritte durch;

1. **Nachdem Sie die Daten in ein Arbeitsblatt eingegeben haben, wählen Sie DATEN | DATENANALYSE.**

   Hierdurch wird das Dialogfeld ANALYSE-FUNKTIONEN geöffnet.

2. **WählenSie in Dialogfeld ANALYSE-FUNKTIONEN die Option ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG aus.**

   Hierdurch wird das Dialogfeld ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE OHNE MESSWIEDERHOLUNG geöffnet.

3. **Geben Sie im Feld EINGABEBEREICH den Zellbereich ein, in dem sich die Daten befinden.**

   In diesem Beispiel ist dies der Zellbereich C1:F9. Da dieser Bereich auch Überschriften enthält, schalten Sie das

Kontrollkästchen BESCHRIFTUNGEN ein.

4. **Achten Sie darauf, dass im Bereich AUSGABE das Optionsfeld NEUES TABELLENBLATT ausgewählt ist, und klicken Sie auf OK.**

   Das Ergebnis dieser Schritte ist die ANOVA-Tabelle, die Sie in Abbildung 13.8 sehen.

18	ANOVA						
19	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
20	Zeilen	638,2916667	7	91,18452381	7,851870835	0,000590448	2,764199257
21	Spalten	40,08333333	2	20,04166667	1,72578165	0,213823725	3,738891832
22	Zufallsfehler	162,5833333	14	11,61309524			
23							
24	Gesamt	840,9583333	23				

Abbildung 13.8: Die ANOVA-Tabelle der ersten von zwei ANOVAs

5. **Bearbeiten Sie die ANOVA-Tabelle. Fügen Sie Zeilen für die Bedingungen aus der zweiten ANOVA-Analyse ein, ändern Sie die Namen der Streuungsursachen und löschen Sie nicht benötigte Werte.**

   Fügen Sie zuerst zwischen *Zeilen* und *Spalten* vier Spalten ein (also zwischen Zeile 20 und der ursprünglichen 21. Zeile).

   Ändern Sie die Beschriftung *Zeilen* in **Zwischen den Gruppen** und die Beschriftung *Spalten* in **Schriftart** (der Name der B-Variablen) ab.

   Löschen Sie anschließend alle Informationen aus der Zeile, in der in der Spalte Streuungsursache *Zufallsfehler* steht.

   Abschließen löschen Sie die F-Quotienten, die P-Werte und die kritischen F-Werte. Die ANOVA-Tabelle sieht nun so aus, wie es Abbildung 13.9 zeigt.

18	ANOVA						
19	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
20	Zwischen den Gruppen	638,2916667	7	91,18452381			
21							
22							
23							
24							
25	Schriftart	40,08333333	2	20,04166667			
26							
27							
28	Gesamt	840,9583333	23				

Abbildung 13.9: Die erste ANOVA-Tabelle nach der Bearbeitung

6. **Wählen Sie ein weiteres Mal den Befehl DATEN | DATENANALYSE.**

7. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN diesmal die Option ANOVA: ZWEIFAKTORIELLE VARIANZANALYSE MIT MESSWIEDERHOLUNG aus.**

8. **Geben Sie im Feld EINGABEBEREICH den Zellbereich ein, in dem sich die Daten einschließlich der Spaltenüberschriften befinden.**

   Wählen Sie hierfür im Arbeitsblatt mit den Daten den Zellbereich C1:F9 aus.

9. **Geben Sie in das Feld ZEILEN JE STICHPROBE die Anzahl der Subjekte innerhalb jeder Stufe der Zwischengruppenvariablen ein.**

   In unserem Beispiel sind dies **4**.

10. **Achten Sie darauf, dass im Bereich AUSGABE das Optionsfeld NEUES TABELLENBLATT ausgewählt ist, und klicken Sie auf OK.**

11. **Kopieren Sie die erstellte ANOVA-Tabelle und fügen Sie diese in das Arbeitsblatt mit den Ergebnissen der ersten Varianzanalyse ein, und zwar direkt unterhalb der ersten ANOVA-Tabelle.**

    Das Arbeitsblatt sollte nun so aussehen, wie es Abbildung 13.10 zeigt.

18	ANOVA						
19	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
20	Zwischen den Gruppen	638,2916667	7	91,18452381			
21							
22							
23							
24							
25	Schriftart	40,08333333	2	20,04166667			
26							
27							
28	Gesamt	840,9583333	23				
29							
30	ANOVA						
31	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
32	Stichprobe	108,375	1	108,375	3,408912189	0,081355668	4,413873419
33	Spalten	40,08333333	2	20,04166667	0,630406291	0,543727248	3,554557146
34	Wechselwirkung	120,25	2	60,125	1,891218873	0,179677714	3,554557146
35	Fehler	572,25	18	31,79166667			
36							
37	Gesamt	840,9583333	23				

Abbildung 13.10: Die bearbeitete erste ANOVA-Tabelle und die eingefügte zweite ANOVA-Tabelle

12. **Fügen Sie in die erste ANOVA-Tabelle, und zwar vier Zeilen unterhalb von** *Zwischen Gruppen,* **den Text** *Innerhalb der Gruppen* **ein und berechnen Sie die Werte für SS und df.**

    Geben Sie in Zelle A24 den Text **Innerhalb der Gruppen** ein. Die Quadratsumme für innerhalb wird folgendermaßen berechnet: SS Gesamt – SS Zwischen den Gruppen (B28–B20). Die Freiheitsgrade für Innerhalb der Gruppen werden so berechnet: df Gesamt – df Zwischen den Gruppen (C28–C20).

13. **Kopieren Sie aus der zweiten ANOVA-Tabelle die Zeile** *Stichprobe* **und fügen Sie diese in die erste ANOVA-Tabelle direkt unterhalb von Zwischen Gruppen ein.**

    Kopieren Sie den Namen der Streuungsursache (*Stichprobe*) sowie den SS- und den df-wert.

14. **Ändern Sie** *Stichprobe* **in den Namen der Zwischengruppenvariablen (der A-Variablen) ab.**

    Ändern Sie *Stichprobe* in **Medium** ab.

15. **Geben Sie in die nächste Zeile den Namen der Quelle für S/A ein und berechnen Sie deren SS-, df- und MS-Wert.**

    Geben Sie **Subjekt/Medium** ein. Der SS-Wert wird so berechnet: SS Zwischen den Gruppen – SS Medium (B20–B21). Der Freiheitsgrad df wird so berechnet: df Zwischen den Gruppen – df Medium (C20–C21). Der MS-Wert ist die Quadratsumme SS geteilt durch die Freiheitsgrade (df): B22/C22.

16. **Berechnen Sie in der entsprechenden Zelle den F-Quotienten für die A-Variable.**

Das Ergebnis kommt in Zelle E21. Der F-Quotient wird folgendermaßen berechnet: MS Medium geteilt durch MS Subjekt/Medium (D21/D22). Die ANOVA-Tabelle sieht nun so aus, wie es Abbildung 13.11 zeigt.

18	ANOVA						
19	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
20	Zwischen den Gruppen	638,2916667	7	91,18452381			
21	Medium	108,375	1	108,375			
22	Subjekt/Medium	529,9166667	6	88,31944444			
23							
24	Innerhalb der Gruppen	202,6666667	16				
25	Schriftart	40,08333333	2	20,04166667			
26							
27							
28	Gesamt	840,9583333	23				

Abbildung 13.11: Die ANOVA-Tabelle mit den Informationen für die Zwischengruppenvariable und einem Teil der Informationen für die Innergruppenvariable

17. **Kopieren Sie aus der zweiten ANOVA-Tabelle die Wechselwirkung, deren SS-, df- und MS-Wert und fügen Sie diese in die erste ANOVA-Tabelle direkt unter dem Namen der B-Variablen ein. Ändern Sie den Text** *Wechselwirkung* **in den Namen der Wechselwirkung zwischen der A-Variablen und der B-Variablen ab.**

    Kopieren Sie die Informationen aus Zeile 34 in Zeile 26, direkt unterhalb von *Schriftart*. Ändern Sie den Text *Wechselwirkung* in **Medium × Schriftart** ab.

18. **Geben Sie in die nächste Zeile den Namen der Quelle von B × S/A ein und berechnen Sie deren SS-, df- und MS-Wert.**

    Geben Sie in Zelle A27 **Schriftart × Subjekt/Medium** ein. Der Wert für SS wird folgendermaßen berechnet: SS Innerhalb der Gruppen – SS Schriftart – SS Medium × Schriftart (B24–B25–B26). Der Wert für df wird folgendermaßen berechnet: df Innerhalb der Gruppen – df Schriftart – df Medium × Schriftart (C24–C25–C26). Die mittlere Quadratsumme MS wird folgendermaßen berechnet: SS geteilt durch df (B27/C27).

19. **Berechnen Sie in den entsprechenden Zellen die verbleibenden F-Quotienten.**

In Zelle E25 teilen Sie D25 durch D27. In Zelle E26 teilen Sie D26 durch D27. Fügen Sie aus Gründen der Übersichtlichkeit oberhalb von *Gesamt* eine leere Zeile ein. Die Tabelle sieht nun so aus, wie es Abbildung 13.12 zeigt.

18	ANOVA						
19	*Streuungsursache*	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
20	Zwischen den Gruppen	638,2916667	7	91,18452381			
21	Medium	108,375	1	108,375	1,22707973		
22	Subjekt/Medium	529,9166667	6	88,31944444			
23							
24	Innerhalb der Gruppen	202,6666667	16				
25	Schriftart	40,08333333	2	20,04166667	0,742283951		
26	Medium × Schriftart	120,25	2	60,125	17,04330709		
27	Schriftart × Subjekt/Medium	42,33333333	12	3,527777778			
28							
29	Gesamt	840,9583333	23				

Abbildung 13.12: Die ANOVA-Tabelle mit allen Quadratsummen, Freiheitsgraden, mittleren Quadratsummen und Prüfgrößen

Damit die Tabelle so aussieht wie die in Abbildung 13.7, verwenden Sie `F.VERT.RE`, um die *P*-Werte zu berechnen und `F.INV.RE`, um die kritischen *F*-Werte zu ermitteln. (Diese beiden Funktionen sind in Kapitel 11 beschrieben.) Sie können außerdem die Werte von $MS_{\text{Zwischen den Gruppen}}$ löschen, da diese keinen Zweck erfüllen.

Um die Tabelle optisch ansprechender zu gestalten, können Sie die Quellen unterhalb der Hauptkategorien (*Zwischen den Gruppen* und *Innerhalb der Gruppen*) einrücken und die Freiheitsgrade (df) der Hauptkategorien zentrieren.

Und was ist nun mit der Analyse? Die fertige ANOVA-Tabelle lässt für das Medium keinen Effekt erkennen, einen signifikanten Effekt bei der Schriftart und eine Wechselwirkung zwischen Medium und Schriftart.

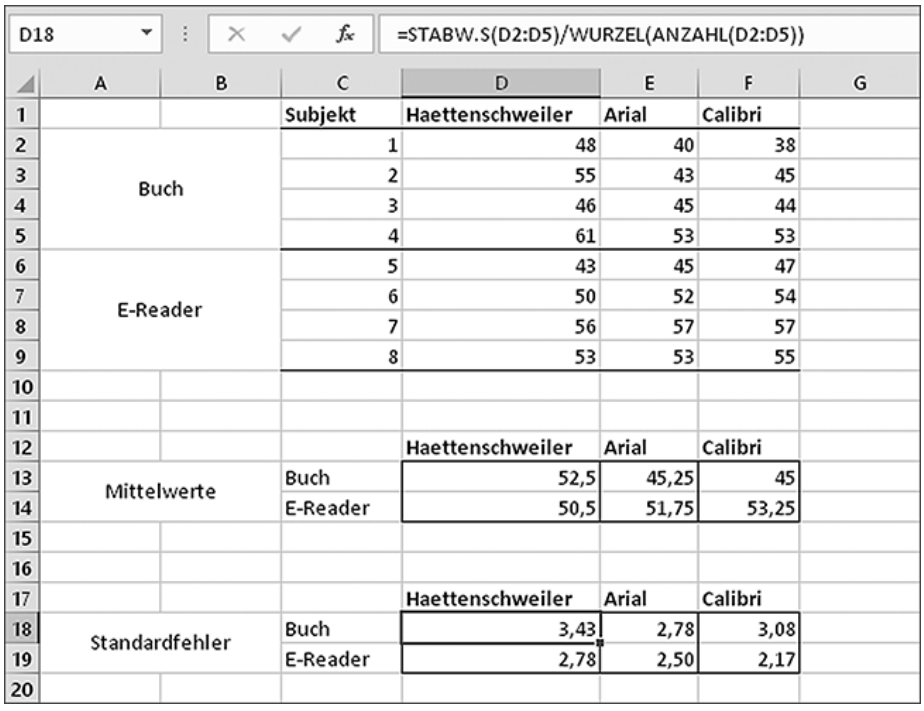
Die oben beschriebene Vorgehensweise verwendet Anova: zweifaktoriell mit Messwiederholung. Dieses Werkzeug setzt eine gleiche Anzahl von Replikationen (Zeilen) für jede Kombination der Faktoren voraus. Damit diese Vorgehensweise funktioniert, benötigen Sie also in jeder Stufe der Zwischengruppenvariablen die gleiche Anzahl von Personen.

Die Ergebnisse als Diagramm darstellen

Da es sich bei der im vorigen Abschnitt beschriebenen Vorgehensweise nicht um ein integriertes Werkzeug handelt, stellt Excel für jede Kombination der Bedingungen keine deskriptiven Statistiken zur Verfügung. Diese Statistiken (Mittelwert und Standardfehler) benötigen Sie jedoch, um das Ergebnis in einem Diagramm darstellen zu können.

Dies ist aber nicht schwer zu erreichen. Abbildung 13.13 zeigt die Tabelle der Mittelwerte sowie die Tabelle der Standardfehler zusammen mit den Ausgangsdaten. Jeder Eintrag in der Tabelle mit den Mittelwerten wird einfach mit der Funktion `MITTELWERT` berechnet, der der betreffende Zellbereich übergeben wird. So lautet die Formel in Zelle D13 beispielsweise folgendermaßen:

```
=MITTELWERT(D2:D5)
```



D18 | =STABW.S(D2:D5)/WURZEL(ANZAHL(D2:D5))

	A	B	C	D	E	F	G
1			Subjekt	Haettenschweiler	Arial	Calibri	
2	Buch		1	48	40	38	
3			2	55	43	45	
4			3	46	45	44	
5			4	61	53	53	
6	E-Reader		5	43	45	47	
7			6	50	52	54	
8			7	56	57	57	
9			8	53	53	55	
10							
11							
12				Haettenschweiler	Arial	Calibri	
13	Mittelwerte		Buch	52,5	45,25	45	
14			E-Reader	50,5	51,75	53,25	
15							
16							
17				Haettenschweiler	Arial	Calibri	
18	Standardfehler		Buch	3,43	2,78	3,08	
19			E-Reader	2,78	2,50	2,17	
20							

Abbildung 13.13: Die Daten sowie die Tabelle mit den Mittelwerten und die Tabelle mit den Standardfehlern

Jeder Wert in der Tabelle mit den Standardfehlern ist das Ergebnis einer Formel, in der drei verschiedene Excel-Tabellenblattfunktionen für den jeweiligen Zellbereich angewendet werden. Um Ihnen ein Beispiel zu zeigen, habe ich in Abbildung 13.13 Zelle D18 markiert, damit in der Bearbeitungsleiste diese Formel sichtbar ist:

```
=STABW.S(D2:D5)/WURZEL(ANZAHL(D2:D5))
```

Abbildung 13.14 zeigt das Säulendiagramm für die Ergebnisse. Um das Diagramm zu erstellen, habe ich die Tabelle mit den Mittelwerten (C12:F14) markiert, EINFÜGEN | EMPFOHLENE DIAGRAMME angeklickt und dann die Option GRUPPIERTE SÄULEN ausgewählt. Anschließend habe ich die Fehlerindikatoren für die Standardfehler ergänzt (siehe Kapitel 22) und das Diagramm optimiert (siehe Kapitel 3).

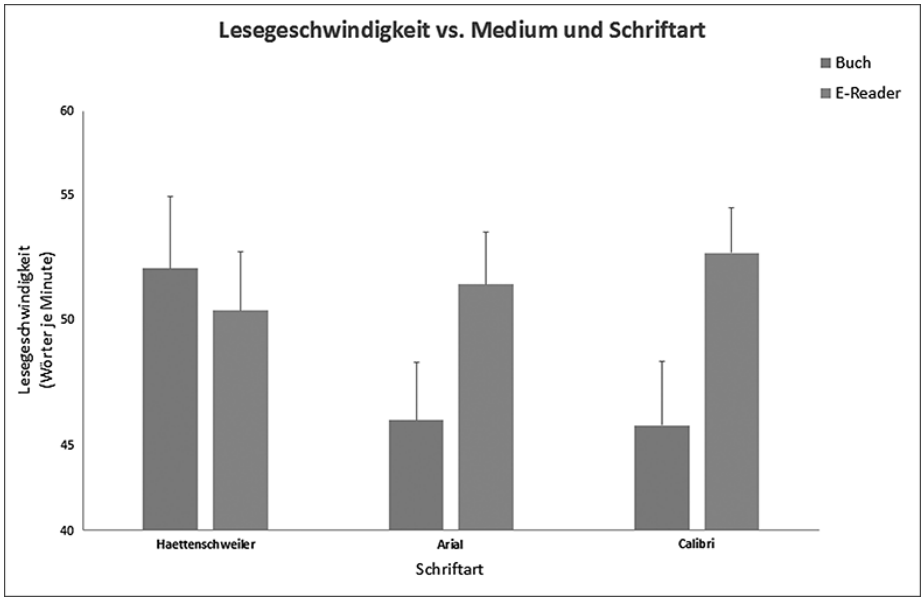


Abbildung 13.14: Die Ergebnisse als Diagramm darstellen

Das Diagramm verdeutlicht sehr gut die Wechselwirkung Medium × Schriftart. Bei der Schrift Haettenschweiler ist die Lesegeschwindigkeit beim Medium Buch besser. Bei allen anderen Schriftarten ist die Lesegeschwindigkeit beim E-Reader besser.

Nach der ANOVA

Wie ich bereits in Kapitel 12 erwähnte, weist Sie ein signifikantes Ergebnis in einer ANOVA-Analyse darauf hin, dass sich irgendwo in den Daten ein Effekt verbirgt. Tests, die Sie nach der Analyse durchführen, sagen Ihnen, wo. Hierbei sind zwei verschiedene Tests möglich, entweder geplant oder nicht geplant. Details hierzu finden Sie ebenfalls in Kapitel 12.

In unserem Beispiel besitzt die Zwischengruppenvariable lediglich zwei Stufen. Daher wäre hier, wenn sich ein signifikanter Effekt ergibt, kein weiterer Test erforderlich. Die Innergruppenvariable, Schriftart, ist signifikant. Normalerweise würde der Rest wie in Kapitel 12 beschrieben fortgeführt. In diesem Fall macht es die Wechselwirkung zwischen Medium und Schriftart erforderlich, einen anderen Weg einzuschlagen.

Wegen dieser Wechselwirkung können die postanalytischen Tests entweder einen (oder auch beide) der folgenden Wege einschlagen. Sie können die Effekte untersuchen, die jede Stufe der A-Variablen auf die Stufen der B-Variablen hat, oder sie können untersuchen, welche Effekte jede Stufe der B-Variablen auf die Stufen der A-Variablen hat. Statistiker nennen dies *einfache Haupteffekte*.

Bei unserem Beispiel würden beim ersten Weg die Mittelwerte der drei Schriftarten im Buch und die Mittelwerte der drei Schriftarten im E-Reader untersucht. Beim zweiten Ansatz würden Sie die Mittelwerte für das Buch im Vergleich zu den Mittelwerten für den E-Reader untersuchen, und zwar mit der Schriftart Haettenschweiler, der Schriftart Arial und der Schriftart Calibri.

Sie finden in statistischen Fachbüchern komplizierte Formeln, um diese Analysen zu berechnen. Excel macht es Ihnen einfach. Um die drei Schriftarten im Buch zu analysieren, führen Sie für Subjekt 1 bis 4 eine ANOVA-Analyse mit Messwiederholungen durch. Um die drei Schriftarten im E-Reader zu analysieren, führen Sie ebenfalls eine ANOVA-Analyse mit Messwiederholungen durch, diesmal aber für Subjekt 5 bis 8.

Für die Analyse der Schriftart Haettenschweiler im Buch beziehungsweise E-Reader verwenden Sie für die Haettenschweiler-Daten eine einfaktorielle ANOVA-Analyse. Um dies zu tun, müssen

Sie die Daten jedoch in zwei Spalten anordnen. Für die anderen Schriftarten würden Sie selbstverständlich eine ähnliche Vorgehensweise wählen.

Kapitel 14

Lineare und multiple Regression

In diesem Kapitel

- Einen Zusammenhang zusammenfassen
- Mit der Regression arbeiten
- Hypothesentests und Regression
- Viele Zusammenhänge austarieren

Eines der wichtigen Dinge, die Sie bei der Arbeit mit statistischen Werten tun: Sie treffen Vorhersagen. Dabei schätzen Sie den Wert einer Variablen mit den Daten einer oder mehrerer anderer Variablen. Um dies tun zu können, müssen Sie wissen, wie Zusammenhänge zwischen Variablen zusammengefasst und wie Hypothesen über diese Zusammenhänge getestet werden.

In diesem Kapitel lernen Sie die *Regression* kennen. Hierbei handelt es sich um ein statistisches Verfahren, mit dem Sie genau das tun können. Zudem können Sie dank der Regression ausführliche Informationen zu Zusammenhängen für Schätzungen verwenden. Zunächst zeige ich Ihnen, wie Sie den Zusammenhang zwischen zwei Variablen analysieren, und dann, wie Sie das zwischen einer Variablen und zwei anderen tun. Für diese Analysen sind einige Berechnungen erforderlich, die Excel gerne für Sie übernimmt.

Das Streudiagramm

Die Universität Sahutsket ist eine spannende, dynamische Institution, aber leider für diese Beispiele erfunden. Jedes Jahr bewerben sich an dieser Einrichtung Tausende von Studenten. Dabei sieht sich die Zulassungsstelle vor folgende Herausforderung gestellt: Die Bewerber möchten wissen, wie hoch ihr GPA (Grade Point Average = gewichteter Notendurchschnitt auf einer Skala von 0,0 bis 4,0) sein wird, wenn sie Sahutsket besuchen.

Welche Schätzung ist die beste? Ohne den Bewerber zu kennen und nur mit den GPAs der eigenen Studenten, ist die Antwort klar: Der GPA ist der durchschnittliche GPA an der Universität Sahutsket. Ohne den Bewerber zu kennen, ist das alles, was die Zulassungsstelle ohne weitere Informationen sagen kann.

Eine genauere Schätzung wird erst mit mehr Informationen zu den Studenten und zu den Bewerbern möglich. Wenn Sahutsket beispielsweise die Gesamt-SAT-Werte der eigenen Studenten speichert, kann die Zulassungsstelle zum GPA der Studenten den passenden SAT-Wert finden und prüfen, ob die beiden Informationen irgendwie zusammenhängen. (Der SAT-Wert ist das Ergebnis eines Tests, den Studienplatzbewerber an amerikanischen Universitäten durchführen müssen.)

Wenn ein Zusammenhang besteht, kann ein Bewerber seinen SAT-Wert angeben, und die Zulassungsstelle kann mit diesem Wert eine Vorhersage treffen.

In Abbildung 14.1 sind die GPA-SAT-Zusammenhänge grafisch dargestellt. Da diese Darstellung Punkte enthält, die im Diagramm verstreut sind, wird dieses Diagramm in Excel als *Punktdiagramm*, sonst auch als *Streudiagramm*bezeichnet. Es ist üblich, an der *y-Achse* (vertikale Achse) die Werte abzutragen, die geschätzt werden sollen. Diese Werte werden als *abhängige Variablen* oder auch als *y-Variablen* bezeichnet. In diesem Beispiel ist das der GPA. Es ist außerdem üblich, an der *x-Achse* (horizontale Achse) die Werte abzutragen, die für die Schätzung verwendet werden. Diese Werte werden als *unabhängige Variablen* oder auch als *x-Variablen* bezeichnet. In diesem Beispiel sind das die SAT-Werte.

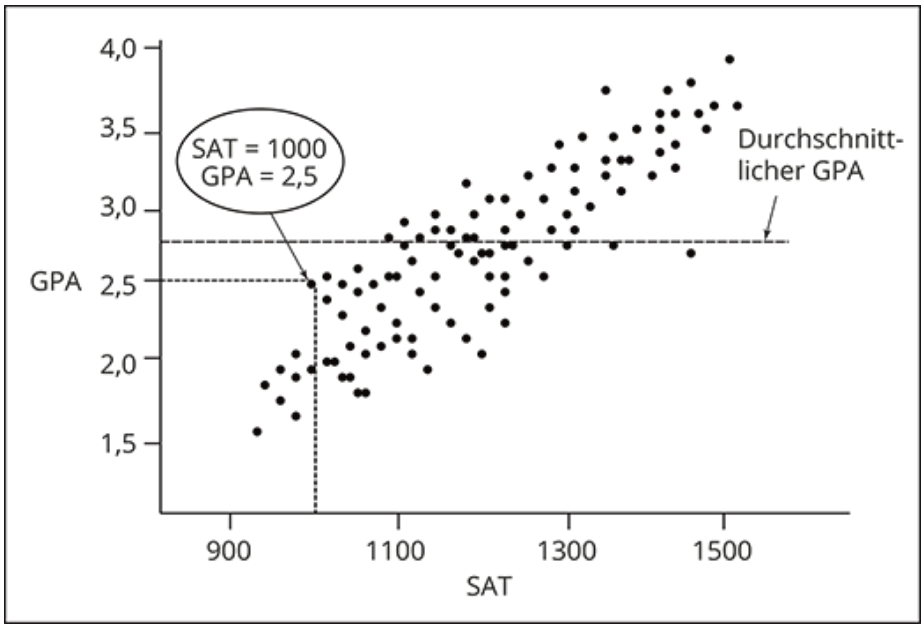


Abbildung 14.1: GPAs und SAT-Werte für die Studenten der Universität Sahutsket

Jeder Punkt im Diagramm stellt den GPA und den SAT-Wert eines einzelnen Studenten dar. Im Streudiagramm der Studenten einer echten Universität wären natürlich wesentlich mehr Punkte enthalten als in diesem Diagramm. Die allgemeine Tendenz der Punkte scheint so zu sein, dass hohe SAT-Werte mit hohen GPAs und niedrige SAT-Werte mit niedrigen GPAs verknüpft sind.

Ich habe einen Punkt aus den anderen herausgenommen. Er steht für einen Sahutsket-Studenten mit einem SAT-Wert von 1.000 und einem GPA von 2,5. Ich habe außerdem den durchschnittlichen GPA angegeben, damit Sie sehen, dass es von Vorteil ist, nicht nur den Mittelwert, sondern auch den GPA-SAT-Zusammenhang zu kennen.

Und wie können Sie diesen Vorteil nutzen? Zunächst fassen Sie den Zusammenhang zwischen SAT und GPA zusammen. Die Zusammenfassung ergibt eine Gerade durch die Punkte. Wie und wo zeichnen Sie die Gerade?

Das erfahren Sie gleich. Zunächst muss ich Ihnen noch ein bisschen etwas über Geraden im Allgemeinen erzählen.

Geraden zeichnen

In der Welt der Mathematik ist eine Gerade eine Methode zum Darstellen eines Zusammenhangs zwischen einer unabhängigen Variablen (x) und einer abhängigen Variablen (y). In der Beziehung y = 4 + 2x gilt: Wenn ich einen Wert für *x* angebe, lässt sich der entsprechende Wert für *y* berechnen. Die Gleichung besagt, dass der *x*-Wert mit 2 multipliziert und zum Ergebnis 4 addiert werden soll.

Beispiel: Wenn *x* = 1, ist *y* = 6. Wenn *x* = 2, ist *y* = 8. In Tabelle 14.1 sind einige *x*-*y*-Paare in diesem Zusammenhang aufgeführt, darunter auch das Paar für *x* = 0.

0	4
1	6
2	8
3	10
4	12
5	14
6	16

Tabelle 14.1: x-y-Paare in y = 4 + 2x

In Abbildung 14.2 sind diese Paare als Punkte an der *x*- und *y*-Achse abgetragen und die Punkte mit einer Geraden miteinander verbunden. Immer, wenn ein *x*-*y*-Paar in Klammern angegeben ist, ist der erste Wert der *x*-Wert.

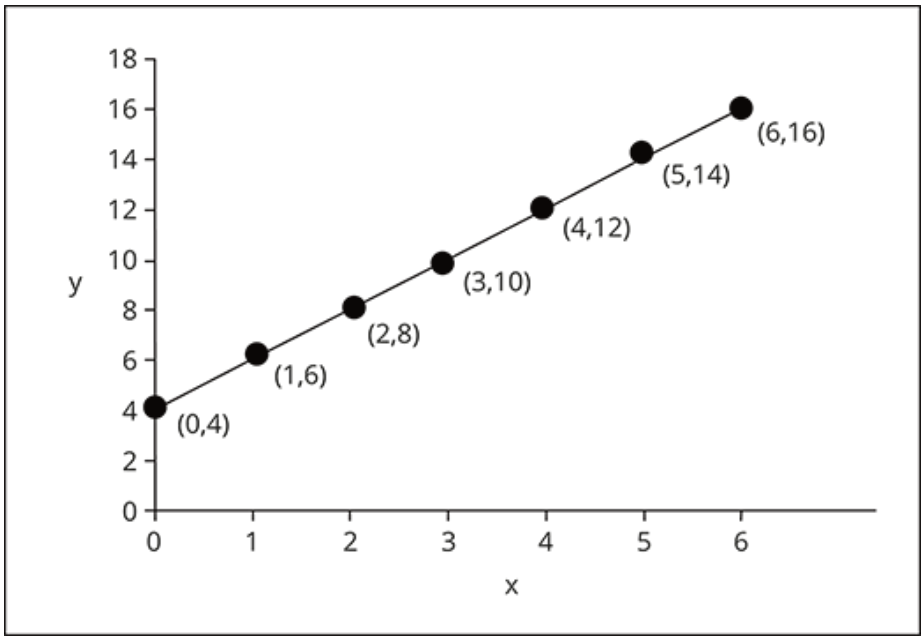


Abbildung 14.2: Das Diagramm für y = 4 + 2x

Wie anhand der Abbildung zu sehen ist, liegen die Punkte alle wunderschön auf einer Geraden. Die Gerade ist eine *grafische Darstellung* der Gleichung $y = 4 + 2x$. Wann immer Sie eine Gleichung wie diese vorliegen haben, bei der x nicht quadriert oder mit einer anderen Zahl als 1 potenziert wird, ergibt sich eine Gleichung, die die Mathematiker als *Geradengleichung* bezeichnen. (Wenn x mit einer höheren Zahl als 1 potenziert wird, ergeben die Punkte keine Gerade, sondern eine Kurve.)

Hier einige Dinge, die Sie sich zu einer Geraden merken sollten: Sie können eine Gerade im Hinblick darauf beschreiben, wie schief sie ist und an welcher Stelle sie die *y*-Achse schneidet.

Der »wie schief sie ist«-Teil wird als *Steigung* bezeichnet. Die Steigung beschreibt, um wie viel sich der *y*-Wert ändert, wenn sich der *x*-Wert um eine Einheit ändert. Bei der Geraden in Abbildung 14.2 ändert sich der *y*-Wert um zwei (von 12 auf 14), wenn sich der *x*-Wert um eins (beispielsweise von 4 auf 5) ändert.

Der »an welcher Stelle sie die *y*-Achse schneidet«-Teil wird als *y-Achsenabschnitt* (oder auch nur als *Achsenabschnitt*) bezeichnet. Das ist der Wert von *y*, bei dem $x = 0$ ist. In Abbildung 14.2 liegt der Achsenabschnitt bei 4.

Diese Zahlen finden Sie in der Gleichung wieder. Die Steigung ist die Zahl, die mit *x* multipliziert wird, und der Achsenabschnitt ist die Zahl, die zu *x* addiert wird. Allgemein gilt:

$$y = a + bx$$

wobei *a* für den Achsenabschnitt und *b* für die Steigung steht.

Die Steigung kann positiv, negativ oder null sein. Die Steigung in Abbildung 14.2 ist positiv. Wenn die Steigung negativ ist, ist die Gerade in die entgegengesetzte Richtung von der Geraden in Abbildung 14.2 geneigt. Eine negative Steigung bedeutet, dass der *y*-Wert mit zunehmendem *x*-Wert abnimmt. Wenn die Steigung null ist, verläuft die Gerade parallel zur *x*-Achse, da sich der *y*-Wert bei sich änderndem *x*-Wert nicht ändert.

Dasselbe gilt für den Achsenabschnitt: Er kann positiv, negativ oder null sein. Wenn der Achsenabschnitt positiv ist, schneidet die Gerade die *y*-Achse *oberhalb* der *x*-Achse. Wenn der Achsenabschnitt negativ ist, schneidet die Gerade die *y*-Achse *unterhalb* der *x*-Achse. Wenn der Achsenabschnitt null ist, schneidet die Gerade die *y*-Achse und die *x*-Achse im so genannten *Koordinatenursprung* oder *Ursprung*.

Und nun zurück zum eigentlichen Thema.

Regression: Was für eine Gerade!

Ich habe bereits erwähnt, dass eine Gerade die beste Möglichkeit ist, die Zusammenhänge im Streudiagramm in Abbildung 14.1 zusammenzufassen. Es ist möglich, unendlich viele Geraden durch

das Streudiagramm zu ziehen. Welche aber fasst die Zusammenhänge am besten zusammen?

Intuitiv wäre die »am besten angepasste« Gerade diejenige, die die meisten Punkte schneidet und die von den anderen Punkten nicht allzu weit entfernt ist. Für Statistiker hat diese Line eine bestimmte Eigenschaft: Wenn Sie im Streudiagramm diese Gerade und dann die Abstände (in vertikaler Richtung) zwischen den Punkten und der Geraden einzeichnen und die Abstände quadrieren und die Quadrate anschließend addieren, ist die am besten angepasste Gerade diejenige, bei der die Summe der quadratischen Abstände am kleinsten ist.

Statistiker nennen diese Gerade *Regressionsgerade* und geben sie wie folgt an:

$$y' = a + bx$$

Jeder *y'*-Wert ist ein Punkt auf der Geraden. Sie stellt die beste Schätzung von *y* für einen bestimmten *x*-Wert dar.

Um herauszufinden, wo genau sich diese Gerade befindet, können Sie die Steigung und den Achsenabschnitt berechnen. Bei einer Regressionsgeraden werden die Steigung und der Achsenabschnitt als *Regressionskoeffizienten* bezeichnet.

Die Formeln für die Regressionskoeffizienten sind unkompliziert. Die Formel für die Steigung lautet:

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

Die Formel für den Achsenabschnitt lautet:

$$a = \bar{y} - b\bar{x}$$

Ich werde das anhand eines Beispiels erklären. Damit die Zahlen berechenbar und überschaubar bleiben, verwende ich anstelle der mehreren tausend Studenten einer Universität nur eine kleine

Stichprobe. In Tabelle 14.2 sind die Stichprobendaten von 20 Studenten der Universität Sahutsket aufgeführt.

Student	SAT	GPA
1	990	2,2
2	1150	3,2
3	1080	2,6
4	1100	3,3
5	1280	3,8
6	990	2,2
7	1110	3,2
8	920	2,0
9	1000	2,2
10	1200	3,6
11	1000	2,1
12	1150	2,8
13	1070	2,2
14	1120	2,1
15	1250	2,4
16	1020	2,2
17	1060	2,3
18	1550	3,9
19	1480	3,8
20	1010	2,0
Mittelwert	1126,5	2,705
Varianz	26171,32	0,46
Standardabweichung	161,78	0,82

Tabelle 14.2: SAT-Werte und GPAs für 20 Studenten der Universität Sahutsket

Für diese Daten berechnet sich die Steigung der Regressionsgeraden wie folgt:

$$b = \frac{(990-1126{,}5)(2{,}2-2{,}705)+(1150-1126{,}5)(3{,}2-2{,}705)+\ldots+(1010-1126{,}5)(2{,}0-2{,}705)}{(2{,}2-2{,}705)^2+(3{,}2-2{,}705)^2+\ldots+(2{,}0-2{,}705)^2} = 0{,}0034$$

Der Achsenabschnitt berechnet sich so:

$$a = \bar{y} - b\bar{x} = 2{,}705 - 0{,}0034(1126{,}5) = -1{,}1538$$

Somit lautet die Gleichung für die am besten angepasste Gerade durch diese 20 Punkte:

$y' = -1{,}1538 + 0{,}0034x$

Oder mit GPAs und SAT-Werten ausgedrückt:

Schätzung von GPA = –1,1538 + 0,0034(*SAT*)

Die Regression für Schätzungen verwenden

Anhand dieser Stichprobe und dieser Regressionsgeraden können Sie mit dem SAT-Wert eines Bewerbers seinen GPA schätzen. Setzen wir für den SAT-Wert beispielsweise 1230 ein:

Schätzung von GPA = –1,1538 + 0,0034(1230) = 3,028

Ohne diese Regel könnte nur der mittlere GPA (2,705) geschätzt werden.

Streuung um die Regressionsgerade

In Kapitel 5 habe ich geschrieben, dass der Mittelwert nicht alles über eine Datengruppe aussagt. Sie müssen zusätzlich angeben,

wie die Werte um den Mittelwert streuen. Dazu habe ich die Varianz und die Standardabweichung eingeführt.

Eine ähnliche Situation liegt hier vor. Um ein vollständiges Bild von den Zusammenhängen in einem Streudiagramm zu erhalten, müssen Sie angeben, wie die Werte um die Regressionsgerade streuen. Hier führe ich nun die *Restvarianz* und den *Standardfehler eines Schätzwerts* ein, die der Varianz und der Standardabweichung entsprechen.

Die Restvarianz ist so etwas wie ein Durchschnittswert der quadratischen Abweichungen der beobachteten *y*-Werte um die geschätzten *y*-Werte. Jede Abweichung eines Datenpunkts von einem geschätzten Punkt ($y - y'$) wird als *Residue* bezeichnet, daher auch der Name. Die Gleichung hierfür lautet:

$$s_{yx}^2 = \frac{\Sigma(y-y')^2}{N-2}$$

Ich sagte »so etwas wie«, weil der Nenner nicht *N*, sondern $N - 2$ lautet. Den Grund hierfür zu erläutern, würde zu weit führen. Wie bereits erwähnt, ist der Nenner eines Varianzschätzwerts df (*Freiheitsgrade*) und Sie werden gleich sehen, wie praktisch das ist.

Der Standardfehler für den Schätzwert lautet:

$$s_{yx} = \sqrt{s_{yx}^2} = \sqrt{\frac{\Sigma(y-y')^2}{(N-2)}}$$

In Tabelle 14.3 können Sie sehen, wie die Restvarianz und der Standardfehler des Schätzwerts die Daten im Beispiel nutzen. Bei dieser Tabelle wurde um den geschätzten GPA für die einzelnen SAT-Werte erweitert.

Student	SAT	GPA	Schätzung von GPA
1	990	2,2	2,24
2	1150	3,2	2,79

Student	SAT	GPA	Schätzung von GPA
3	1080	2,6	2,55
4	1100	3,3	2,61
5	1280	3,8	3,23
6	990	2,2	2,24
7	1110	3,2	2,65
8	920	2,0	2,00
9	1000	2,2	2,27
10	1200	3,6	2,96
11	1000	2,1	2,27
12	1150	2,8	2,79
13	1070	2,2	2,51
14	1120	2,1	2,68
15	1250	2,4	3,13
16	1020	2,2	2,34
17	1060	2,3	2,48
18	1550	3,9	4,16
19	1480	3,8	3,92
20	1010	2,0	2,31
Mittelwert	1126,5	2,705	
Varianz	26171,32	0,46	
Standardabweichung	161,78	0,82	

Tabelle 14.3: SAT-Werte, GPAs und geschätzte GPAs für 20 Studenten der Universität Sahutsket

Wie Sie in der Tabelle sehen können, ist der geschätzte GPA manchmal recht nah, manchmal nicht. Ein Schätzwert (4,16) liegt total daneben, da der maximale Wert des GPA 4,00 beträgt.

Für diese Daten lautet die Restvarianz wie folgt:

$$s_{yx}^2 = \frac{\Sigma(y-y')^2}{N-2} = \frac{(2{,}2-2{,}24)^2+(3{,}2-2{,}79)^2+\ldots+(2{,}0-2{,}31)^2}{20-2}$$
$$= \frac{2{,}91}{18} = 0{,}16$$

Der Standardfehler für den Schätzwert lautet:

$$s_{yx} = \sqrt{s_{yx}^2} = \sqrt{0{,}16} = 0{,}40$$

Wenn die Restvarianz und der Standardfehler des Schätzwerts klein sind, ist die Regressionsgerade eine gute Anpassung an die Daten im Streudiagramm. Wenn die Restvarianz und der Standardfehler des Schätzwerts groß sind, ist die Regressionsgerade eine schlechte Anpassung.

Was ist »klein«? Was ist »groß«? Was ist eine »gute« Anpassung?

Lesen Sie weiter.

Hypothesen über die Regression testen

Mit der Regressionsgleichung, mit der Sie gearbeitet haben,

$$y' = a + bx$$

wird ein Zusammenhang in einem Streudiagramm einer Stichprobe zusammengefasst. Die Regressionskoeffizienten *a* und *b* sind Stichprobenwerte. Mit diesen Werten können Sie Hypothesen über die Parameter einer Grundgesamtheit testen, und genau das werden Sie in diesem Abschnitt tun.

Die Regressionsgerade durch die Grundgesamtheit, aus der die Stichprobe stammt (wie zum Beispiel alle Studenten der Universität Sahutsket, früher und heute), ist die Darstellung einer Gleichung, die nicht aus Statistiken, sondern aus Parametern besteht. Wenn Sie sich noch erinnern: Es ist üblich, griechische Buchstaben für Parameter zu verwenden. Somit lautet die Regressionsgleichung für die Grundgesamtheit wie folgt:

$$y' = a + \beta x + \varepsilon$$

Die ersten beiden griechischen Buchstaben auf der rechten Seite sind α (Alpha) und β (Beta), die Pendants zu *a* und *b*. Und was ist mit dem letzten Zeichen? Das sieht ein bisschen wie das griechische Pendant zum Buchstaben *e* aus. Aber was tut das hier?

Dieser letzte Ausdruck ist der griechische Buchstabe *Epsilon*. Er steht für »Fehler« (engl. Error) in der Grundgesamtheit. Fehler ist ein etwas ungeschickter Ausdruck. Er fasst alles zusammen, was wir nicht kennen oder über das wir keine Kontrolle haben. Fehler spiegeln sich in den Residuen, den Abweichungen von den Schätzwerten, wider. Je mehr Sie das, was Sie messen, verstehen, umso geringer ist die Fehlerquote.

Den Fehler im Zusammenhang zwischen SAT und GPA können wir nicht messen, aber er ist da. So ist es beispielsweise durchaus möglich, dass eine Person einen niedrigen SAT-Wert erzielt und dennoch die Universität mit einem höheren GPA als geschätzt abschließt. In einem Streudiagramm sieht der SAT-GPA-Punkt dieser Person wie ein Schätzfehler aus. Wenn Sie mehr über diese Person erfahren, finden Sie möglicherweise heraus, dass sie am Tag des SAT-Tests krank war, und das erklärt den »Fehler«.

Sie können Hypothesen über Alpha, Beta und Epsilon testen, und das werden wir in den nächsten Abschnitten tun.

Die Anpassung testen

Sie beginnen mit einem Test dazu, wie gut die Regressionsgerade an das Streudiagramm angepasst ist. Bei diesem Test wird ε, der Fehler im Zusammenhang, getestet.

Ziel ist es zu entscheiden, ob die Gerade wirklich einen Zusammenhang zwischen den Variablen darstellt. Es kann sein, dass das, was wie ein Zusammenhang aussieht, reiner Zufall ist und dass die Gleichung der Regressionsgeraden nichts bedeutet (da die Fehlergröße extrem ist) – oder es kann sein, dass zwischen den Variablen ein enger Zusammenhang besteht.

Diese Möglichkeiten können getestet werden und Sie formulieren Hypothesen, um sie zu testen:

H_0: Kein wirklicher Zusammenhang

H_1: Nicht H_0

Obwohl sich diese Hypothesen gut anhören, sind dies keine statistischen Tests. Für einen Test müssen die Varianzen betrachtet werden. Um Varianzen betrachten zu können, beginnen wir mit den Abweichungen. In Abbildung 14.3 ist ein Punkt im Streudiagramm und die zugehörige Abweichung von der Regressionsgeraden (die Residue) und vom Mittelwert der *y*-Variablen dargestellt. In dieser Abbildung ist außerdem die Abweichung der Regressionsgeraden vom Mittelwert dargestellt.

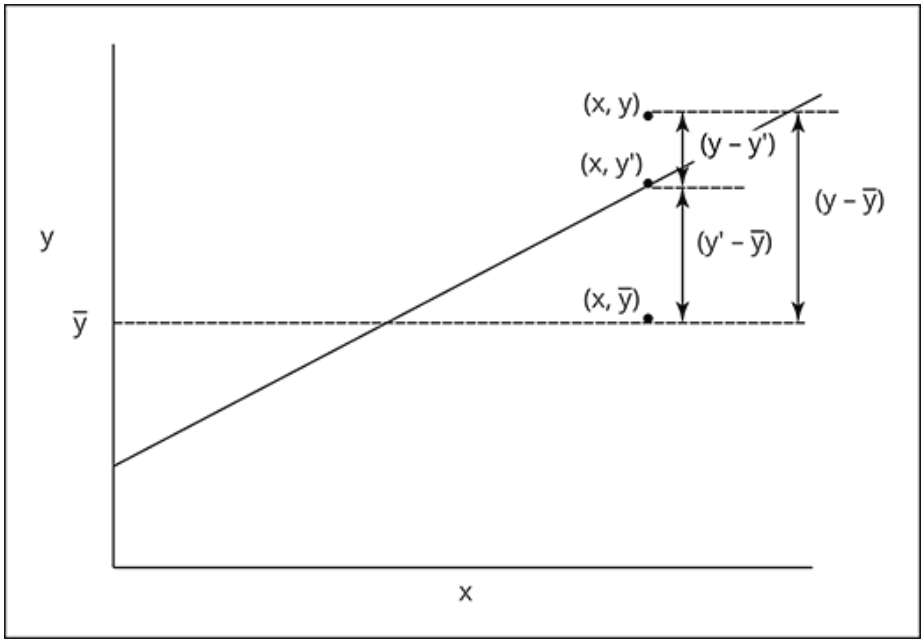


Abbildung 14.3: Die Abweichungen in einem Streudiagramm

Wie in der Abbildung zu sehen ist, ergibt die Summe aus dem Abstand zwischen dem Punkt und der Regressionsgeraden und dem Abstand zwischen der Regressionsgeraden und dem Mittelwert den Abstand zwischen dem Punkt und dem Mittelwert:

$$(y-y')+(y'-\bar{y})=(y-\bar{y})$$

Dies ist die Grundlage für einige andere wichtige Zusammenhänge.

Beginnen wir mit dem Quadrieren der einzelnen Abweichungen. Das ergibt $(y-y')^2$, $(y'-\bar{y})^2$ und $(y-\bar{y})^2$. Wenn Sie all diese quadratischen Abweichungen addieren, ergibt sich Folgendes:

$$\Sigma(y-y')^2$$

Diese Summe kennen Sie bereits. Sie ist der Zähler für die Restvarianz. Sie stellt die Streuung um die Regressionsgerade, also den weiter oben beschriebenen »Fehler«, dar. Mit der Terminologie aus Kapitel 12 wird der Zähler einer Varianz als Quadratsumme oder SS (Sum of Squares) bezeichnet. Somit ist dies $SS_{Residue}$.

$$\Sigma(y'-\bar{y})^2$$

Diese Summe ist neu. Die Abweichung $(y'-\bar{y})^2$ stellt die Verbesserung der Schätzung dar, weil die Regressionsgerade statt des Mittelwerts verwendet wurde. Die Summe spiegelt diese Verbesserung wider und wird als $SS_{Regression}$ bezeichnet.

$$\Sigma(y-\bar{y})^2$$

Diese Summe haben Sie in Kapitel 5 kennengelernt. Dort habe ich jedoch anstelle von *y* ein *x* verwendet. Hierbei handelt es sich um den Zähler der Varianz von *y*. Mit der Terminologie aus Kapitel 12 ausgedrückt, ist diese Summe der Zähler der *Gesamtvarianz*. Sie wird als SS_{Gesamt} bezeichnet.

Für den Zusammenhang zwischen diesen drei Summen gilt Folgendes:

$$SS_{Residue} + SS_{Regression} = SS_{Gesamt}$$

Jeder Summe ist ein Wert für Freiheitsgrade zugeordnet: der Nenner eines Varianzschätzers. Wie bereits im vorherigen Abschnitt

erwähnt, beträgt der Nenner für $SS_{Residue}$ $N - 2$. Der df-Wert für SS_{Gesamt} beträgt $N - 1$ (siehe Kapitel 5 und 12). Ebenso wie die Quadratsummen bilden auch die Freiheitsgrade eine Summe:

$$df_{Regression} + df_{Restwert} = df_{Gesamt}$$

Damit bleibt ein Freiheitsgrad für die Regression.

Wozu das alles und was hat das mit Hypothesentests zu tun? Nun, da Sie fragen: Varianzschätzwerte berechnen sich aus der Division von SS durch df. Jeder Varianzschätzwert wird als mittlere Quadratsumme (MS für engl. Mean Square) bezeichnet (siehe Kapitel 12):

$$MS_{Regression} = \frac{SS_{Regression}}{df_{Regression}}$$
$$MS_{Residue} = \frac{SS_{Residue}}{df_{Residue}}$$
$$MS_{Gesamt} = \frac{SS_{Gesamt}}{df_{Gesamt}}$$

Nun kommen wir zum Hypothesenteil. Wenn H_0 wahr ist, und wenn das, was wie ein Zusammenhang zwischen *x* und *y* aussieht, nicht wirklich etwas zu bedeuten hat, darf der Teil, der die Verbesserung der Schätzung aufgrund der Regressionsgeraden ($MS_{Regression}$) darstellt, nicht größer sein als die Streuung um die Regressionsgerade ($MS_{Residue}$). Wenn H_0 nicht wahr ist und die Verbesserung der Schätzung enorm ist, muss $MS_{Regression}$ deutlich größer sein als $MS_{Residue}$.

Somit lauten die Hypothesen nun wie folgt:

$$H_0 : \sigma^2_{Regression} \leq \sigma^2_{Residue}$$
$$H_1 : \sigma^2_{Regression} > \sigma^2_{Residue}$$

Das sind die Hypothesen, die Sie testen können. Wie? Zum Testen einer Hypothese über zwei Varianzen verwenden Sie einen *F*-Test

(siehe Kapitel 11). Die Testgröße hier beträgt

$$F = \frac{MS_{Regression}}{MS_{Residue}}$$

Damit Sie sehen, wie das alles funktioniert, wende ich die Formeln auf das Sahutsket-Beispiel an. $MS_{Residue}$ ist dasselbe wie s_{yx2} aus dem vorhergehenden Abschnitt und dieser Wert beträgt 0,16. $MS_{Regression}$ beträgt:

$$MS_{Regression} = \frac{(2,24-2,705)^2 + (2,79-2,705)^2 + \ldots + (2,31-2,705)^2}{1} = 5,83$$

Damit ergibt sich für den *F*-Quotienten Folgendes:

$$F = \frac{MS_{Regression}}{MS_{Residue}} = \frac{5,83}{0,16} = 36,03$$

Der kritische Wert für *F* beträgt 4,41, wobei df = 1 und 18 und α = 0,05 ist. (Sie können das mit der Tabellenfunktion FINV überprüfen.) Der berechnete *F*-Quotient ist größer als der kritische Wert für *F*. H_0 wird also verworfen. Das bedeutet, dass die Regressionsgerade eine gute Anpassung an die Daten in der Stichprobe darstellt.

Die Steigung testen

Eine weitere Frage, die sich im Zusammenhang mit der linearen Regression stellt, ist die, ob sich die Steigung der Regressionsgeraden signifikant von null unterscheidet. Wenn nicht, ist der Mittelwert für die Schätzung ebenso gut geeignet wie die Regressionsgerade.

Die Hypothesen für diesen Test lauten wie folgt:

H_0: $\beta \leq 0$
H_1: $\beta > 0$

Hier kommt der *t*-Test zur Anwendung, der in den Kapiteln 9, 10 und 11 im Zusammenhang mit Mittelwerten beschrieben wurde. Der *t*-Test für die Steigung lautet wie folgt:

$$t = \frac{b - \beta}{s_b}$$

wobei df = *N* – 2. Der Nenner stellt den Schätzwert für den Standardfehler der Steigung dar. Das hört sich komplizierter an, als es ist. Die Gleichung hierfür lautet:

$$s_b = \frac{s_{yx}}{s_x \sqrt{(N-1)}}$$

wobei s_x für die Standardabweichung der *x*-Variablen steht. Für die Daten in diesem Beispiel gilt:

$$s_b = \frac{s_{yx}}{s_x \sqrt{(N-1)}} = \frac{0,402}{(161,776)\sqrt{(20-1)}} = 0,00057$$

Die Formel für den *t*-Test lautet wie folgt:

$$t = \frac{b - \beta}{s_b} = \frac{0,0034 - 0}{0,00057} = 5,96$$

Der tatsächliche Wert beträgt 6,00. Wenn s_{yx} und s_b nicht auf Zahlen mit praktischeren Dezimalstellen gerundet werden, ergibt die Berechnung den Wert 5,96. Wie dem auch sei, dieser Wert ist größer als der kritische Wert von *t* bei einem df-Wert von 18 und α = 0,05 (2,10), so dass H_0 verworfen wird. Dieses Beispiel zeigt im Übrigen, warum es wichtig ist, Hypothesen zu testen. Die Steigung 0,0034 sieht wie eine sehr kleine Zahl aus (möglicherweise, weil es eine sehr kleine Zahl ist). Dennoch ist sie groß genug, um H_0 in diesem Fall zu verwerfen.

Den Achsenabschnitt testen

Der Vollständigkeit halber führe ich auch den Hypothesentest für den Achsenabschnitt an. Ich bezweifle jedoch, dass Sie ihn jemals verwenden werden. Er tritt jedoch im Ergebnis einiger Excel-Funktionen auf, die mit der Regression zu tun haben. Und schließlich möchte ich, dass Sie alle Aspekte dieses Ergebnisses (von dem ich Ihnen gleich erzählen werde) verstehen.

Die Hypothesen lauten wie folgt:

H_0: $\alpha = 0$

H_1: $\alpha \neq 0$

Auch hierbei wird der *t*-Test verwendet. Die Gleichung hierfür lautet:

$$t = \frac{a - \alpha}{S_a}$$

Der Nenner ist der Schätzwert des Standardfehlers des Achsenabschnitts. Ohne weiter ins Detail zu gehen, lautet die Formel für s_a:

$$S_a = S_{YX} \sqrt{\left[\frac{1}{N} + \frac{\bar{x}^2}{(N-1)s_x^2} \right]}$$

wobei $\bar{x}^2$ der quadratische Mittelwert der *x*-Variablen und s_x^2 die Varianz der *x*-Variablen sind. Setzen wir die Zahlen aus diesem Beispiel in die Formel ein:

$$S_a = S_{YX} \sqrt{\left[\frac{1}{N} + \frac{\bar{x}^2}{(N-1)s_x^2} \right]} = 0{,}649$$

Die Formel für den *t*-Test lautet wie folgt:

$$t = \frac{a - \alpha}{S_a} = \frac{-1,15}{0,649} = -1,78$$

Bei 18 Freiheitsgraden und einer Wahrscheinlichkeit für einen Fehler 1. Art von 0,05 liegt der kritische *t*-Wert bei einem zweiseitigen Test bei 2,45. Es handelt sich um einen zweiseitigen Test, weil H_1 besagt, dass der Achsenabschnitt ungleich null ist. Die Hypothese sagt nichts darüber aus, ob der Achsenabschnitt größer oder kleiner als null ist. Da der berechnete Wert nicht negativer ist als der negative kritische Wert, wird H_0 nicht verworfen.

Tabellenblattfunktionen für die Regression

Excel ist eine große Hilfe für rechenintensive Arbeit wie zum Beispiel die lineare Regression. Eine große Auswahl an Funktionen und Datenanalysetools erleichtern das Leben hierbei erheblich. In diesem Abschnitt werden Sie die entsprechenden Tabellenfunktionen und zwei Feldfunktionen kennenlernen.

In Abbildung 14.4 sind die Daten dargestellt, die ich zum Veranschaulichen der einzelnen Funktionen verwendet habe. Bei den Daten handelt es sich um GPAs und SAT-Werte für die 20 Studenten aus dem Beispiel weiter oben. Wie Sie der Abbildung entnehmen können, befinden sich die SAT-Werte in den Zellen C3:C22 und die GPAs in den Zellen D3:D22. Der SAT-Wert stellt die *x*-Variable dar und der GPA die *y*-Variable.

Um zu verdeutlichen, was die Funktionen machen, definiere ich Namen für die Zellbereiche. Dem Zellbereich C3:C22 habe ich den Namen SAT und dem Zellbereich D3:D22 den Namen GPA zugewiesen. Wenn Sie nicht wissen, wie Sie für einen Zellbereich einen Namen definieren, werfen Sie einen Blick in Kapitel 2.

	A	B	C	D	E
1					
2		Student	SAT	GPA	
3		1	990	2,2	
4		2	1150	3,2	
5		3	1080	2,6	
6		4	1100	3,3	
7		5	1280	3,8	
8		6	990	2,2	
9		7	1110	3,2	
10		8	920	2,0	
11		9	1000	2,2	
12		10	1200	3,6	
13		11	1000	2,1	
14		12	1150	2,8	
15		13	1070	2,2	
16		14	1120	2,1	
17		15	1250	2,4	
18		16	1020	2,2	
19		17	1060	2,3	
20		18	1550	3,9	
21		19	1480	3,8	
22		20	1010	2,0	
23					

Abbildung 14.4: Daten für Tabellenblattfunktionen zur Regression

STEIGUNG, ACHSENABSCHNITT, STFEHLERYX

Diese drei Funktionen sind mehr oder weniger gleich, so dass ich ihre Funktionsweise allgemein beschreiben und bei Bedarf die jeweiligen Eigenheiten erläutern werde.

1. **Markieren Sie nach der Eingabe der Daten eine Zelle.**

2. **Klicken Sie im Menü mit den statistischen Funktionen eine Regressionsfunktion an, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

- Um die Steigung einer Regressionsgeraden durch die Daten zu berechnen, wählen Sie die Funktion STEIGUNG aus.
- Um den Achsenabschnitt zu berechnen, wählen Sie die Funktion ACHSENABSCHNITT aus.
- Um den Standardfehler des Schätzwerts zu berechnen, wählen Sie die Funktion STFEHLERYX aus.

In den Abbildungen 14.5, 14.6 und 14.7 ist das Dialogfeld FUNKTIONSARGUMENTE für diese drei Funktionen abgebildet.

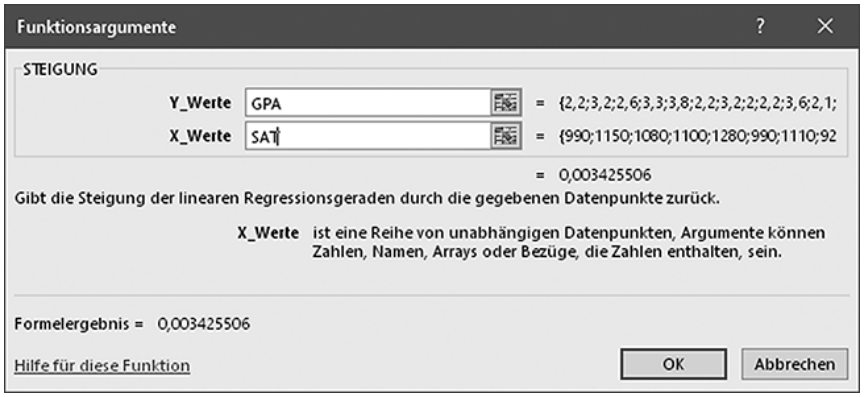


Abbildung 14.5: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion STEIGUNG

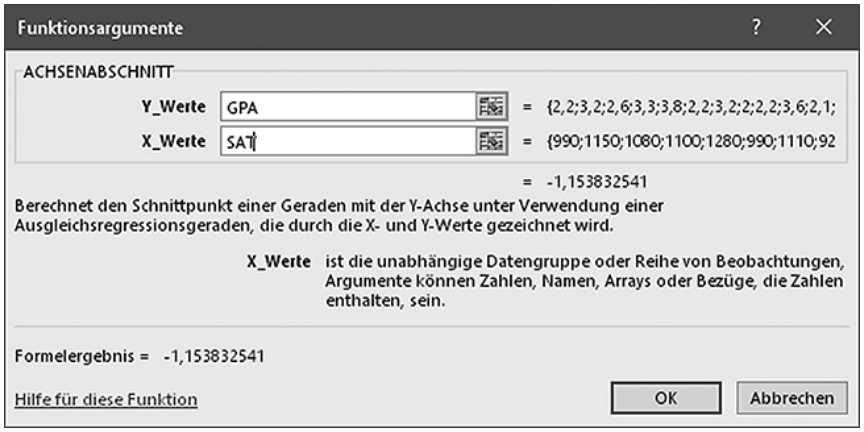


Abbildung 14.6: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion ACHSENABSCHNITT

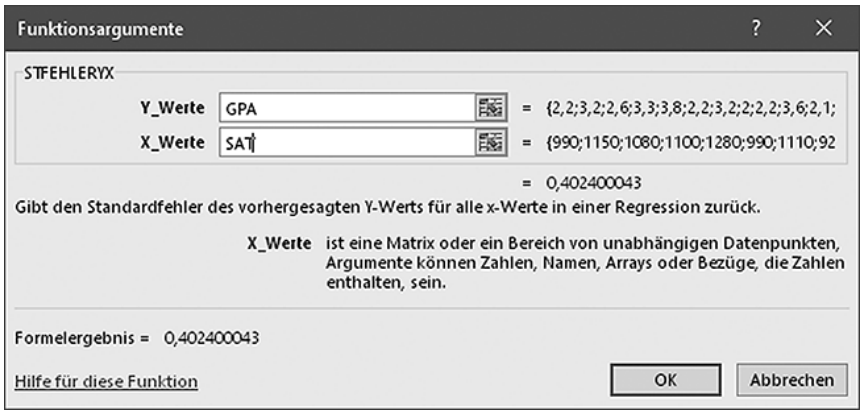


Abbildung 14.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion STFEHLERYX

3. **Geben Sie Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Feld Y_WERTE habe ich den Namen des Zellbereichs eingegeben, in dem sich die Werte für die y-Variable befinden. In diesem Beispiel ist dies GPA (der für den Zellbereich C3:C22 definierte Name).

   In das Feld X_WERTE habe ich den Namen des Zellbereichs eingegeben, in dem sich die Werte für die x-Variable befinden. In diesem Beispiel ist dies SAT (der für den Zellbereich D3:D22 definierte Name). Nachdem ich die Namen eingegeben habe, wird das Ergebnis im Dialogfeld angezeigt.

   - Das Ergebnis der Funktion `STEIGUNG` beträgt 0,00342556 (siehe Abbildung 14.5).
   - Das Ergebnis der Funktion `ACHSENABSCHNITT` beträgt −1,153832541 (siehe Abbildung 14.6).
   - Das Ergebnis der Funktion `STFEHLERYX` beträgt 0,402400043 (siehe Abbildung 14.7).

4. **Klicken Sie auf OK, um das Ergebnis in die markierte Zelle einzufügen.**

PROGNOSE.LINEAR

Diese Funktion unterscheidet sich etwas von den vorherigen drei. Neben den Spalten für die *x*- und *y*-Variablen geben Sie bei der Funktion `PROGNOSE.LINEAR` (die in früheren Windows-Versionen `SCHÄTZER` hieß und die in Excel für Mac *noch immer* so heißt) zudem einen Wert für *x* an. Das Ergebnis ist ein Schätzwert, der auf dem Zusammenhang der linearen Regression zwischen der *x*-Variablen und der *y*-Variablen beruht.

In Abbildung 14.8 ist das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `PROGNOSE.LINEAR` dargestellt. Im Feld X habe ich 1290 eingegeben. Für diesen SAT-Wert beträgt der geschätzte GPA 3,265070236.

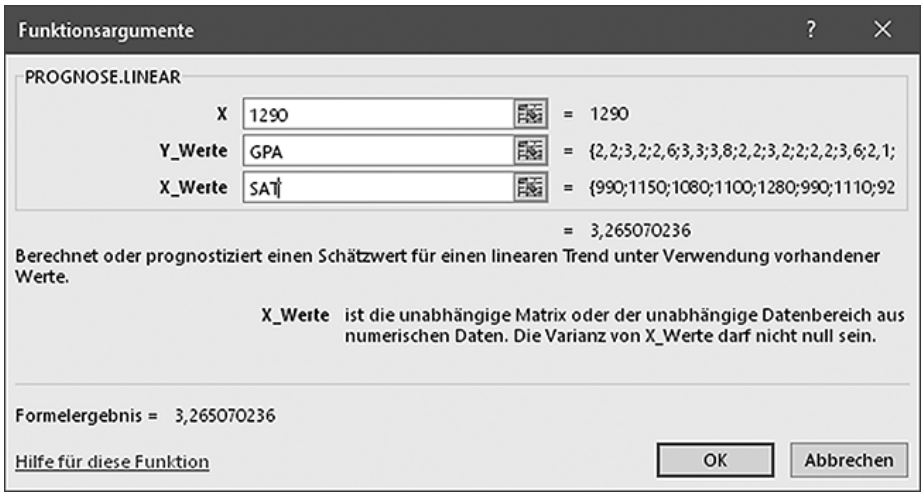


Abbildung 14.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `PROGNOSE.LINEAR`

Arrayfunktion: TREND

`TREND` ist eine vielseitige Funktion. Mit `TREND` können Sie für die *x*-Werte in der Stichprobe geschätzte *y*-Werte ermitteln.

Sie können auch neue *x*-Werte angeben und basierend auf dem linearen Zusammenhang in der Stichprobe neue *y*-Werte schätzen. Das ist so, als würden Sie die Funktion `PROGNOSE.LINEAR` mehrere Male hintereinander anwenden.

In diesem Abschnitt werden beide Anwendungen beschrieben.

y-Werte für die *x*-Werte in der Stichprobe schätzen

Zunächst schätzen wir mit der Funktion TREND die GPAs für die 20 Studenten in der Stichprobe. In Abbildung 14.9 ist das Dialogfeld für die Funktion TREND entsprechend ausgefüllt. In der Abbildung ist auch die Bearbeitungsleiste enthalten, damit Sie sehen, wie die Formel für diese Verwendung der Funktion TREND aussieht.

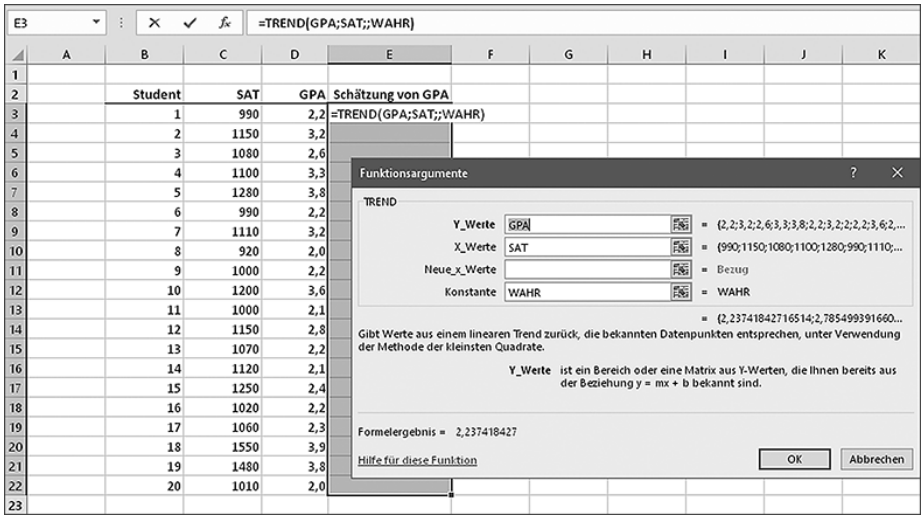


Abbildung 14.9: Das Dialogfeld FUNKTIONSARGUMENTE für TREND mit Daten. Das Dialogfeld ist zum Schätzen der GPAs für die SAT-Werte in der Stichprobe ausgefüllt.

1. **Markieren Sie nach der Eingabe der Daten eine Spalte für das Ergebnis aus der Berechnung der Funktion TREND.**

   Ich habe die Zellen E3:E22 markiert. Damit werden die geschätzten GPAs rechts neben den GPAs aus der Stichprobe eingefügt.

1. **Klicken Sie im Menü mit den statistischen Funktionen auf TREND, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

2. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

In das Feld Y_WERTE habe ich den Namen des Zellbereichs eingegeben, in dem sich die Werte für die y-Variable befinden. In diesem Beispiel ist dies `GPA` (der für den Zellbereich D3:D22 definierte Name).

In das Feld X_WERTE habe ich den Namen des Zellbereichs eingegeben, in dem sich die Werte für die x-Variable befinden. In diesem Beispiel ist dies `SAT` (der für den Zellbereich C3:C22 definierte Name).

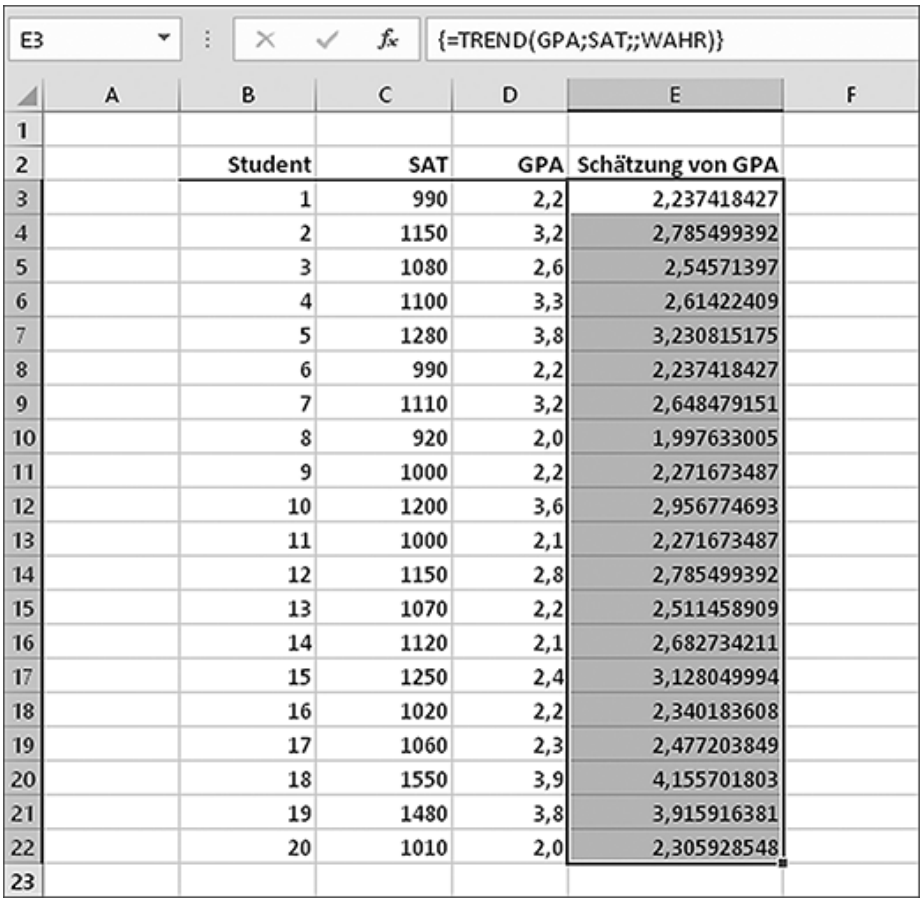
Das Feld NEUE_X_WERTE habe ich leer gelassen.

In das Feld KONSTANTE habe ich `WAHR` eingegeben (ich hätte es auch leer lassen können), um den *y*-Achsenabschnitt zu berechnen. Geben Sie `FALSCH` ein, um den *y*-Achsenabschnitt auf null zu setzen.

(Ich wüsste nicht, warum jemand `FALSCH` eingeben sollte.) Ein wichtiger Hinweis: Im Dialogfeld bezieht sich die Anweisung für das Feld KONSTANTE auf *b*. Das ist der *y*-Achsenabschnitt. Weiter vorn in diesem Kapitel habe ich *a* zur Darstellung des *y*-Achsenabschnitts und *b* zur Darstellung der Steigung verwendet. Eine einheitliche Verwendung ist hierfür nicht festgelegt.

3. **WICHTIG: Klicken Sie nicht auf OK. Da es sich hierbei um eine Arrayfunktion handelt, müssen Sie die Tastenkombination Strg+Shift+Enter drücken, damit das Ergebnis der Berechnung der Funktion `TREND` in die markierte Spalte eingefügt wird.**

In Abbildung 14.10 ist das Ergebnis in den Zellen E3:E22 dargestellt. Die Abbildung zeigt ebenfalls die Bearbeitungsleiste, damit Sie die geschweiften Klammern sehen können, mit denen Excel eine Arrayfunktion einschließt.



E3 | {=TREND(GPA;SAT;;WAHR)}

	A	B	C	D	E	F
1						
2		Student	SAT	GPA	Schätzung von GPA	
3		1	990	2,2	2,237418427	
4		2	1150	3,2	2,785499392	
5		3	1080	2,6	2,54571397	
6		4	1100	3,3	2,61422409	
7		5	1280	3,8	3,230815175	
8		6	990	2,2	2,237418427	
9		7	1110	3,2	2,648479151	
10		8	920	2,0	1,997633005	
11		9	1000	2,2	2,271673487	
12		10	1200	3,6	2,956774693	
13		11	1000	2,1	2,271673487	
14		12	1150	2,8	2,785499392	
15		13	1070	2,2	2,511458909	
16		14	1120	2,1	2,682734211	
17		15	1250	2,4	3,128049994	
18		16	1020	2,2	2,340183608	
19		17	1060	2,3	2,477203849	
20		18	1550	3,9	4,155701803	
21		19	1480	3,8	3,915916381	
22		20	1010	2,0	2,305928548	
23						

Abbildung 14.10: Das Ergebnis der Funktion `TREND`: Geschätzte GPAs für die SAT-Werte in der Stichprobe

Neue y-Werte für neue x-Werte schätzen

In diesem Beispiel verwende ich die Funktion `TREND` zum Schätzen der GPAs für neue SAT-Werte. In Abbildung 14.11 ist das Dialogfeld für die Funktion `TREND` entsprechend ausgefüllt. In der Abbildung ist außerdem die für das Ergebnis markierte Spalte zu sehen. Auch hier habe ich die Bearbeitungsleiste in die Abbildung mit aufgenommen, damit Sie die Formel für diese Verwendung der Funktion sehen.

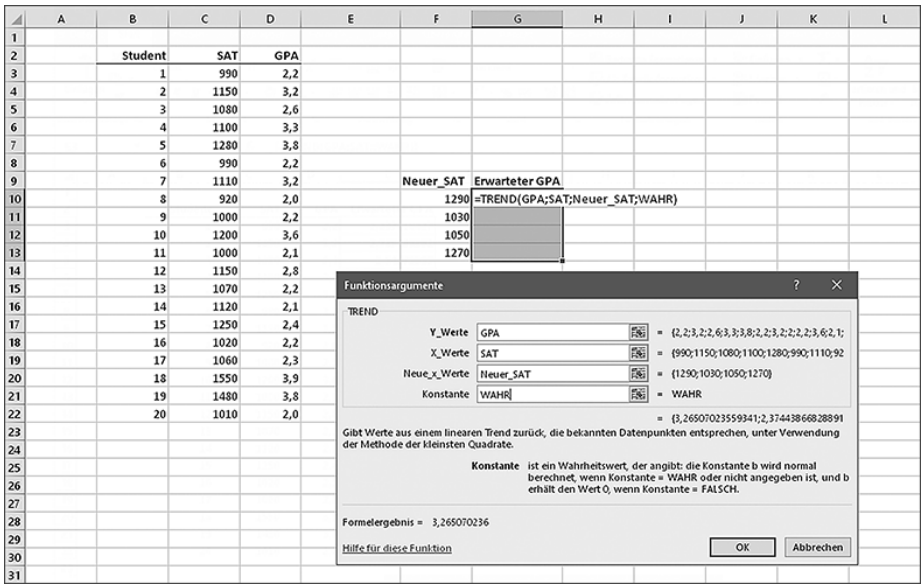


Abbildung 14.11: Das Dialogfeld FUNKTIONSARGUMENTE für `TREND` mit Daten. Die Funktion `TREND` wird zum Schätzen der GPAs für neue SAT-Werte verwendet.

1. **Markieren Sie nach der Eingabe der Daten eine Spalte für das Ergebnis aus der Berechnung der Funktion `TREND`.**

   Ich habe die Zellen G10:G13 markiert.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf TREND, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   Geben Sie in das Feld `Y_WERTE` den Namen des Zellbereichs ein, in dem sich die Werte für die *y*-Variable befinden. In diesem Beispiel ist dies `GPA` (der für den Zellbereich D3:D22 definierte Name).

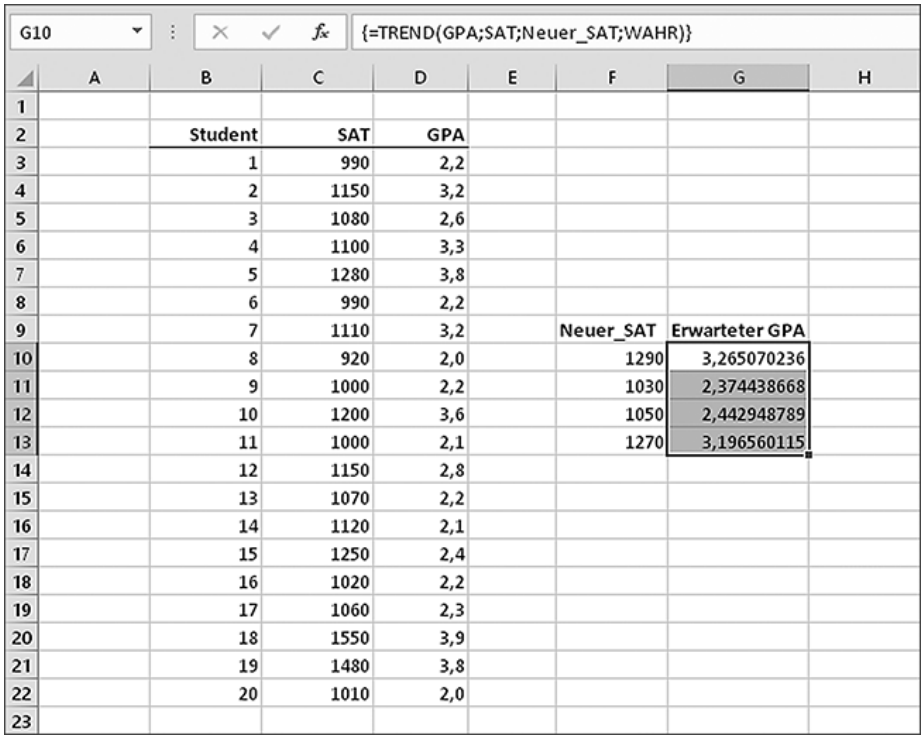
   Geben Sie in das Feld `X_WERTE` den Namen des Zellbereichs ein, in dem sich die Werte für die *x*-Variable befinden. In diesem Beispiel ist dies `SAT` (der für den Zellbereich C3:C22 definierte Name).

Geben Sie in das Feld NEUE_X_WERTE den Namen des Zellbereichs ein, in den die neuen Werte der *x*-Variablen eingefügt werden sollen. In diesem Beispiel ist dies `Neuer_SAT`.

In das Feld KONSTANTE können Sie `WAHR` eingegeben (oder es leer lassen), um den *y*-Achsenabschnitt zu berechnen. Geben Sie `FALSCH` ein, um den *y*-Achsenabschnitt auf null zu setzen. (Wie gesagt: Ich wüsste nicht, warum jemand `FALSCH` eingeben sollte.)

4. **WICHTIG: Klicken Sie nicht auf OK. Da es sich hierbei um eine Feldfunktion handelt, müssen Sie die Tastenkombination Strg+Shift+Enter drücken, damit das Ergebnis der Berechnung der Funktion `TREND` in die markierte Spalte eingefügt wird.**

In Abbildung 14.12 ist das Ergebnis in den Zellen G10:G13 dargestellt. Auch hier können Sie in der Bearbeitungsleiste sehen, dass Excel die Arrayformel in geschweifte Klammern einfasst.



	A	B	C	D	E	F	G	H
1								
2		Student	SAT	GPA				
3		1	990	2,2				
4		2	1150	3,2				
5		3	1080	2,6				
6		4	1100	3,3				
7		5	1280	3,8				
8		6	990	2,2				
9		7	1110	3,2		Neuer_SAT	Erwarteter GPA	
10		8	920	2,0		1290	3,265070236	
11		9	1000	2,2		1030	2,374438668	
12		10	1200	3,6		1050	2,442948789	
13		11	1000	2,1		1270	3,196560115	
14		12	1150	2,8				
15		13	1070	2,2				
16		14	1120	2,1				
17		15	1250	2,4				
18		16	1020	2,2				
19		17	1060	2,3				
20		18	1550	3,9				
21		19	1480	3,8				
22		20	1010	2,0				
23								

Abbildung 14.12: Das Ergebnis der Funktion TREND: Geschätzte GPAs für neue SAT-Werte

Arrayfunktion: RGP

RGP ist eine Kombination aus STEIGUNG, ACHSENABSCHNITT und STFEHLERYX mit einigen zusätzlichen Extras. In Abbildung 14.13 habe ich das Dialogfeld FUNKTIONSARGUMENTE für die Funktion RGP zusammen mit den Daten und dem markierten Zellbereich für das Ergebnis dargestellt. Der Zellbereich besteht aus fünf Zeilen und zwei Spalten. So muss der markierte Zellbereich für die lineare Regression aussehen. Woher würden Sie wissen, aus wie vielen Zeilen und Spalten der Bereich besteht, wenn ich es Ihnen nicht gesagt hätte? Nun, … Sie könnten es nicht wissen.

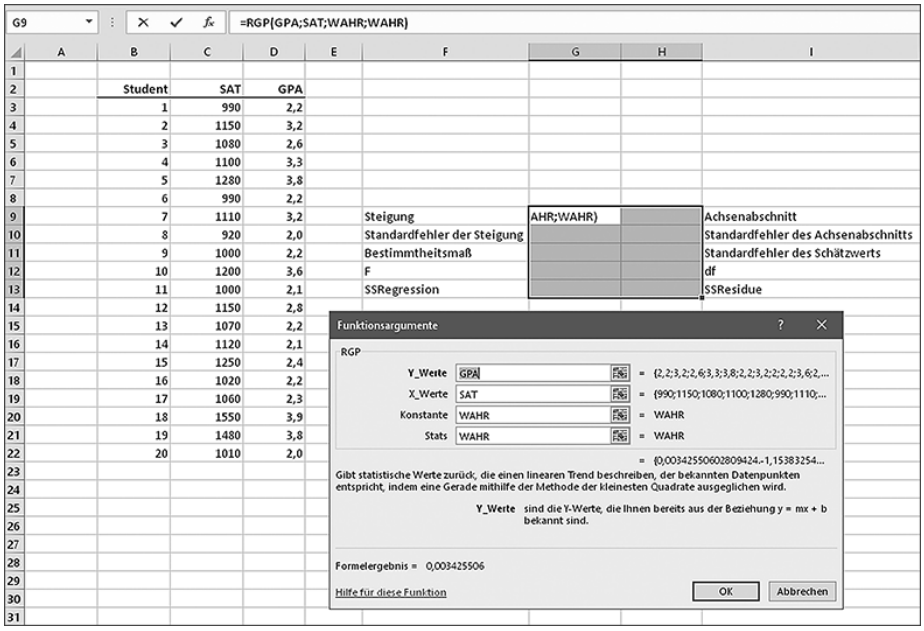


Abbildung 14.13: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion RGP zusammen mit den Daten und dem markierten Zellbereich für das Ergebnis

So verwenden Sie die Funktion RGP:

1. **Markieren Sie nach der Eingabe der Daten einen Zellbereich mit fünf Zeilen und zwei Spalten für das Ergebnis der Funktion RGP.**

Ich habe die Zellen G9:H13 markiert.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf RGP, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Geben Sie in das Feld Y_WERTE den Namen des Zellbereichs ein, in dem sich die Werte für die *y*-Variable befinden. In diesem Beispiel ist dies GPA (der für den Zellbereich D3:D22 definierte Name).

   Geben Sie in das Feld X_WERTE den Namen des Zellbereichs ein, in dem sich die Werte für die *x*-Variable befinden. In diesem Beispiel ist dies SAT (der für den Zellbereich C3:C22 definierte Name).

   In das Feld KONSTANTE können Sie WAHR eingegeben (oder es leer lassen), um den *y*-Achsenabschnitt zu berechnen. Geben Sie FALSCH ein, um den *y*-Achsenabschnitt auf null zu setzen. Ich habe WAHR eingegeben.

   Geben Sie im Feld STATS WAHR ein, um neben der Steigung und dem Achsenabschnitt auch Regressionswerte zu ermitteln. Geben Sie FALSCH ein (oder lassen Sie das Feld leer), um nur die Steigung und den Achsenabschnitt zu ermitteln. Auch hier habe ich WAHR eingegeben.

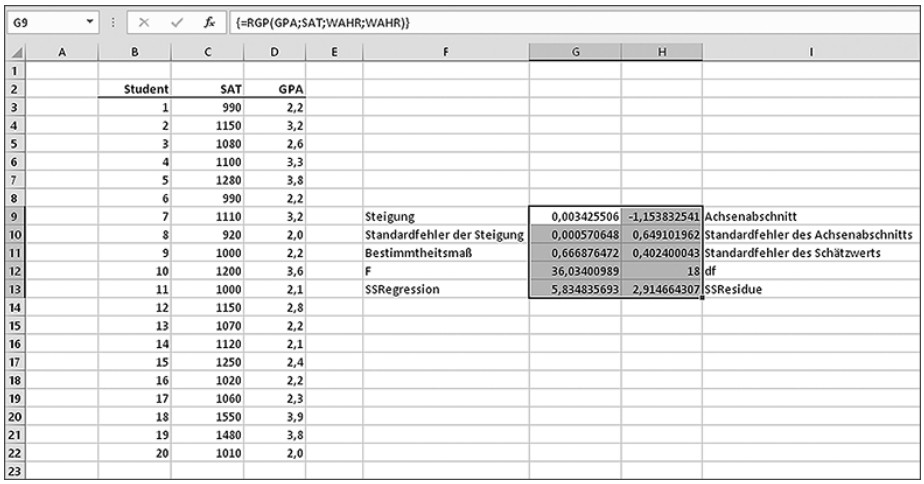
   Im Dialogfeld bezieht sich *b* auf den Achsenabschnitt und der *m-Koeffizient* auf die Steigung. Wie bereits erwähnt, gibt es hierfür keine einheitlichen Symbole.

4. **WICHTIG: Klicken Sie nicht auf OK. Da es sich hierbei um eine Feldfunktion handelt, müssen Sie die Tastenkombination Strg+Shift+Enter drücken, damit das**

Ergebnis der Berechnung der Funktion RGP in den markierten Zellbereich eingefügt wird.

In Abbildung 14.14 ist das Ergebnis der Funktion `RGP` dargestellt. Das Ergebnis ist nicht beschriftet. Daher habe ich die Beschriftungen im Arbeitsblatt hinzugefügt. In der linken Spalte stehen die Steigung, der Standardfehler der Steigung, etwas, das *Bestimmtheitsmaß* heißt, *F* und $SS_{Regression}$. Was ist ein Bestimmtheitsmaß? Das ist ein weiteres Maß dafür, wie eng SAT und GPA in der Stichprobe zusammenhängen. Ausführlichere Informationen hierzu finden Sie in Kapitel 15.

In der rechten Spalte stehen der Achsenabschnitt, der Standardfehler des Achsenabschnitts, der Standardfehler des Schätzwerts, Freiheitsgrade und $SS_{Residue}$.



	A	B	C	D	E	F	G	H	I
1									
2		Student	SAT	GPA					
3		1	990	2,2					
4		2	1150	3,2					
5		3	1080	2,6					
6		4	1100	3,3					
7		5	1280	3,8					
8		6	990	2,2					
9		7	1110	3,2		Steigung	0,003425506	-1,153832541	Achsenabschnitt
10		8	920	2,0		Standardfehler der Steigung	0,000570648	0,649101962	Standardfehler des Achsenabschnitts
11		9	1000	2,2		Bestimmtheitsmaß	0,666876472	0,402400043	Standardfehler des Schätzwerts
12		10	1200	3,6		F	36,03400989	18	df
13		11	1000	2,1		SSRegression	5,834835693	2,914664307	SSResidue
14		12	1150	2,8					
15		13	1070	2,2					
16		14	1120	2,1					
17		15	1250	2,4					
18		16	1020	2,2					
19		17	1060	2,3					
20		18	1550	3,9					
21		19	1480	3,8					
22		20	1010	2,0					
23									

Abbildung 14.14: Das Ergebnis der Funktion RGP im markierten Zellbereich

Datenanalysetool: Regression

Das Excel-Datenanalysetool Regression hat dieselbe Funktion wie `RGP`, übernimmt noch ein paar weitere Aufgaben und versieht das Ergebnis außerdem mit Beschriftungen. In Abbildung 14.15 ist das Dialogfeld für das Tool Regression zusammen mit den Daten für das SAT-GPA-Beispiel dargestellt.

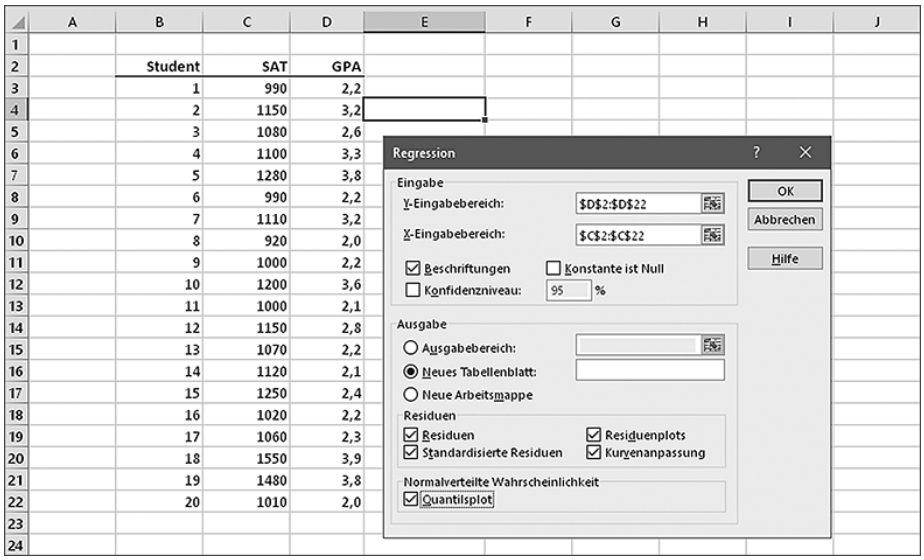


Abbildung 14.15: Das Datenanalysetool REGRESSION zusammen mit den SAT-GPA-Daten

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie die Daten in das Tabellenblatt sowie Beschriftungen für die Spalten ein.**

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option REGRESSION aus.**

4. **Klicken Sie auf OK, um das Dialogfeld REGRESSION zu öffnen**

   In Abbildung 14.15 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld Y-EINGABEBEREICH den Zellbereich ein, in dem sich die Daten für die *y*-Variable befinden.**

   In diesem Beispiel befinden sich die GPAs (mit Beschriftung) in den Zellen $D$2:$D$22. Beachten Sie die Dollarzeichen ($) für den absoluten Bezug.

6. **Geben Sie im Feld X-EINGABEBEREICH den Zellbereich ein, in dem sich die Daten für die *x*-Variable befinden.**

   Die SAT-Werte (mit Beschriftung) befinden sich in den Zellen $C$2:$C$22.

7. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich auch dieses Kontrollkästchen aktiviert.

8. **Im Feld KONFIDENZNIVEAU befindet sich standardmäßig der Wert 95 %, was einem Niveau von Alpha von 0,05 entspricht. Ändern Sie den Wert, wenn Sie ein anderes Alpha möchten.**

9. **Wählen Sie im Bereich AUSGABE eine Option für die Ausgabe des Ergebnisses.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, so dass das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

10. **Der Bereich RESIDUEN enthält vier Kontrollkästchen zum Anzeigen der Abweichungen zwischen den Datenpunkten und den geschätzten Punkten. Aktivieren Sie so viele Kontrollkästchen, wie Sie mögen, oder gar keins.**

    Ich habe alle vier Kontrollkästchenaktiviert. Ich werde diese Kontrollkästchen bei der Beschreibung des Ergebnisses erläutern.

11. **Aktivieren Sie das Kontrollkästchen QUANTILSPLOT, wenn Sie ein Diagramm mit den Quantilen der *y*-Variablen erstellen möchten.**

Ich habe diese Option aktiviert, damit ich Ihnen das Ergebnis zeigen kann.

12. **Klicken Sie auf** OK.

    Da ich die Option NEUES TABELLENBLATT ausgewählt habe, wird ein neues Arbeitsblatt mit dem Ergebnis erstellt.

Ergebnis in Tabellen

In Abbildung 14.16 ist die obere Hälfte der von diesem Tool erstellten Ergebnistabelle mit erweiterten Spalten dargestellt. Diese Tabelle trägt die Bezeichnung *AUSGABE: ZUSAMMENFASSUNG*. Dieser Ergebnisteil enthält eine Tabelle für die Regressionsstatistik, eine weitere für *ANOVA* und eine für die Regressionskoeffizienten.

	A	B	C	D	E	F	G	H	I
1	AUSGABE: ZUSAMMENFASSUNG								
2									
3	*Regressions-Statistik*								
4	Multipler Korrelationskoeffizient	0,81662505							
5	Bestimmtheitsmaß	0,66687647							
6	Adjustiertes Bestimmtheitsmaß	0,64836961							
7	Standardfehler	0,40240004							
8	Beobachtungen	20							
9									
10	ANOVA								
11		*Freiheitsgrade (*	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*			
12	Regression	1	5,834835693	5,834835693	36,03400989	1,12048E-05			
13	Residue	18	2,914664307	0,161925795					
14	Gesamt	19	8,7495						
15									
16		*Koeffizienten*	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*Untere 95,0%*	*Obere 95,0%*
17	Schnittpunkt	-1,15383254	0,649101962	-1,777582888	0,09237211	-2,517545159	0,209880078	-2,517545159	0,209880078
18	SAT	0,00342551	0,000570648	6,002833489	1,12048E-05	0,002226619	0,004624393	0,002226619	0,004624393
19									

Abbildung 14.16: Die obere Hälfte der Ergebnistabelle des Datenanalysetools Regression

Die ersten drei Zeilen der Tabelle *Regressions-Statistik* enthalten Informationen zum Bestimmtheitsmaß, ein Maß dafür, wie eng SAT und GPA in der Stichprobe zusammenhängen (siehe Kapitel 15). In der vierten Zeile ist der Standardfehler des Schätzwerts und in der fünften Zeile die Anzahl der Objekte in der Stichprobe angegeben.

Die Tabelle ANOVA enthält das Ergebnis aus folgendem Test:

H_0: $\sigma_{Regression} \leq \sigma_{Residue}$

H_1: $\sigma_{Regression} > \sigma_{Residue}$

Wenn der Wert in der Spalte *F krit* kleiner als 0,05 ist (oder welchen Alpha-Wert auch immer Sie verwenden), wird H_0 verworfen. In diesem Beispiel ist der Wert kleiner als 0,05.

Direkt unter der Tabelle ANOVA befindet sich eine Tabelle mit Informationen zu den Regressionskoeffizienten. Excel hat für diese Tabelle keinen Namen, aber ich nenne sie Koeffiziententabelle. Die Spalte *Koeffizienten* enthält die Werte für den Achsenabschnitt und die Steigung. Die Steigung ist mit dem Namen der *x*-Variablen gekennzeichnet. Die Spalte *Standardfehler* enthält den Standardfehler des Achsenabschnitts und den der Steigung.

Die weiteren Spalten enthalten die Ergebnisse für die *t*-Tests des Achsenabschnitts und der Steigung. Anhand der Informationen in der Spalte *P-Wert* können Sie erkennen, ob H_0 für den jeweiligen Test verworfen wird oder nicht. Wenn der Wert kleiner als Ihr Alpha-Wert ist, wird H_0 verworfen. In diesem Beispiel wird H_0 für die Steigung verworfen, für den Achsenabschnitt nicht.

In Abbildung 14.17 ist die untere Hälfte der Ergebnistabelle des Tools Regression dargestellt.

21							
22	AUSGABE: RESIDUENPLOT					AUSGABE: QUANTILSPLOT	
23							
24	*Beobachtung*	*ätzung für G*	*Residuen*	*Standardisierte Residuen*		*Quantil*	*GPA*
25	1	2,23741843	-0,037418427	-0,095536221		2,5	2
26	2	2,78549939	0,414500608	1,058297332		7,5	2
27	3	2,54571397	0,05428603	0,138602356		12,5	2,1
28	4	2,61422409	0,68577591	1,750913753		17,5	2,1
29	5	3,23081518	0,569184825	1,453234976		22,5	2,2
30	6	2,23741843	-0,037418427	-0,095536221		27,5	2,2
31	7	2,64847915	0,551520849	1,408135554		32,5	2,2
32	8	1,99763301	0,002366995	0,006043379		37,5	2,2
33	9	2,27167349	-0,071673487	-0,182995776		42,5	2,2
34	10	2,95677469	0,643225307	1,642274131		47,5	2,3
35	11	2,27167349	-0,171673487	-0,43831442		52,5	2,4
36	12	2,78549939	0,014500608	0,037022757		57,5	2,6
37	13	2,51145891	-0,311458909	-0,795212664		62,5	2,8
38	14	2,68273421	-0,582734211	-1,487829085		67,5	3,2
39	15	3,12804999	-0,728049994	-1,858847373		72,5	3,2
40	16	2,34018361	-0,140183608	-0,357914887		77,5	3,3
41	17	2,47720385	-0,177203849	-0,452434465		82,5	3,6
42	18	4,1557018	-0,255701803	-0,652854376		87,5	3,8
43	19	3,91591638	-0,115916381	-0,295956132		92,5	3,8
44	20	2,30592855	-0,305928548	-0,78109262		97,5	3,9
45							

Abbildung 14.17: Die untere Hälfte der Ergebnistabelle des Datenanalysetools Regression

Hier sehen Sie die Tabellen *AUSGABE: RESIDUENPLOT* und *AUSGABE: QUANTILSPLOT*. In der Tabelle *AUSGABE: RESIDUENPLOT* sind der Schätzwert und die Residue (y - y') für die

einzelnen Objekte in der Stichprobe aufgeführt. Außerdem enthält diese Tabelle die *standardisierte Residue* für die einzelnen Beobachtungen, die sich wie folgt berechnet:

$$\text{Standardisierte Residue} = \frac{\text{Residue} - \text{Mittelwert der Residuen}}{s_{yx}}$$

Die Daten zu Residue und standardisierter Residue in der Tabelle sind für die Analyse der Streuung um die Regressionsgerade hilfreich. So können Sie in diesen Daten beispielsweise nach Ausreißern suchen oder prüfen, ob Ausreißer mit bestimmten Werten der *x*-Variablen verknüpft sind. (Wenn dies der Fall ist, ist das möglicherweise ein Hinweis darauf, dass sich in Ihrer Stichprobe eigenartige Dinge tun.)

In der Tabelle *AUSGABE: QUANTILSPLOT* sind die Quantile in den Daten der *y*-Variablen in der Stichprobe aufgeführt.

Die grafische Ausgabe

In den Abbildungen 14.18, 14.19 und 14.20 sind die grafischen Ausgaben des Tools Regression dargestellt. Das Quantilsplot in Abbildung 14.18 ist eine grafische Darstellung der Tabelle *AUSGABE: QUANTILSPLOT*. Im SAT-Residuenplot in Abbildung 14.19 sind die Residuen gegen die *x*-Variablen abgetragen. Für jeden SAT-Wert in der Stichprobe gibt es in diesem Diagramm eine entsprechende Residue. In Abbildung 14.20 ist die SAT-Kurvenanpassung dargestellt: eine Darstellung des Streudiagramms und der geschätzten *y*-Werte.

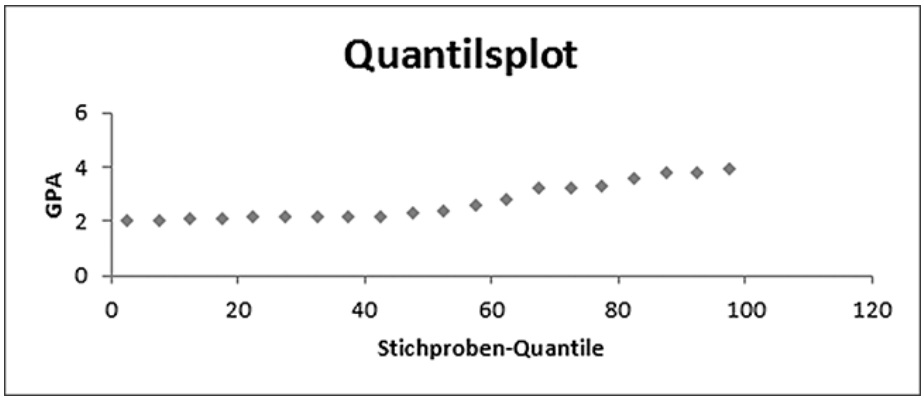


Abbildung 14.18: Im Quantilsplot sind die Quantile für die y-Variable dargestellt.

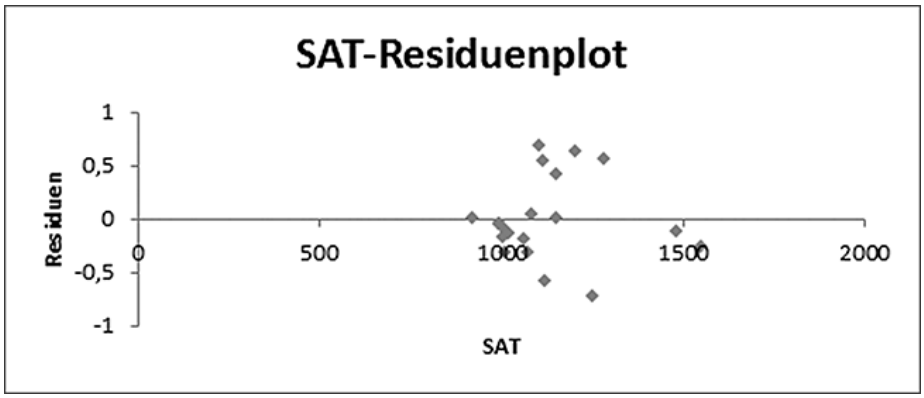


Abbildung 14.19: Im SAT-Residuenplot sind die Residuen gegen die x-Variable abgetragen.

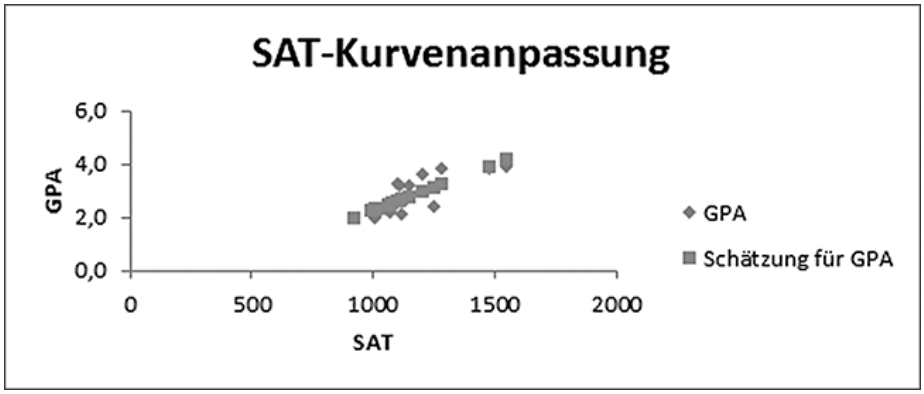


Abbildung 14.20: SAT-Kurvenanpassung ist ein Streudiagramm mit den geschätzten Punkten.

Falls Sie das Datenanalysetool Regression in Excel 2007 verwendet haben, fällt Ihnen auf, dass die Diagramme Quantilsplot und SAT-

Kurvenanpassung anders aussehen. In Excel 2007 wurden hier Säulendiagramme verwendet. Sie können die Diagrammfeatures von Excel verwenden, um die Diagramme so abzuändern, wie Sie möchten.

Irrsinnig viele Zusammenhänge auf einmal: multiple Regression

Die lineare Regression eignet sich hervorragend für Schätzungen. Wenn die Steigung und der Achsenabschnitt der Geraden, die zwei Variablen miteinander verbindet, bekannt sind, können Sie anhand eines neuen *x*-Werts einen neuen *y*-Wert schätzen. In dem Beispiel, das wir eben durchgearbeitet haben, haben wir anhand eines SAT-Werts einen GPA für einen Studenten an der Universität Sahutsket geschätzt.

Was wäre, wenn nicht nur der SAT-Wert für die einzelnen Studenten bekannt wäre? Was wäre, wenn für jeden Studenten auch der Notendurchschnitt von der High School (auf einer Skala von 100) bekannt wäre und diese Informationen ebenfalls verwendet werden könnten? Wenn Sie die SAT-Werte mit dem HS-Durchschnitt kombinieren könnten, ergäbe das möglicherweise einen genaueren Schätzwert als nur mit den SAT-Werten.

Wenn Sie mit mehr als einer unabhängigen Variablen arbeiten, bewegen Sie sich im Reich der *multiplen Regression*. Wie bei der linearen Regression können Sie Regressionskoeffizienten für die am besten angepasste Gerade durch ein Streudiagramm ermitteln. Wie bereits erwähnt, bedeutet »am besten angepasst«, dass die Summe der quadrierten Abstände von den Datenpunkten zur Geraden am kleinsten ist.

Bei zwei unabhängigen Variablen kann ein Streudiagramm jedoch nicht zweidimensional dargestellt werden. Eine dritte Dimension wird erforderlich, und das wird schwierig zu zeichnen. Stattdessen werde ich Ihnen die Gleichung der Regressionsgeraden zeigen:

$$y' = a + b_1x_1 + b_2x_2$$

Für das SAT-GPA-Beispiel ergibt sich somit Folgendes:

$$\textit{Schätzwert von GPA} = a + b_1(\text{SAT}) + b_2\ (\text{High-School-Durchschnittsnote})$$

Sie können Hypothesen über die gesamte Anpassung sowie über alle drei Regressionskoeffizienten testen.

Ich werde nicht alle Formeln zum Ermitteln der Koeffizienten beschreiben, weil das *richtig* schwierig wird. Stattdessen werde ich mich gleich den Funktionen von Excel zuwenden.

Hier noch einige Dinge, die Sie sich merken sollten, bevor ich fortfahre:

- **Sie können beliebig viele *x*-Variablen verwenden.** Im folgenden Beispiel verwende ich nur zwei.
- **Der Koeffizient für SAT wird sich von linearer in multiple Regression ändern.** Auch der Achsenabschnitt wird sich ändern.
- **Der Standardfehler des Schätzwerts nimmt von der linearen zur multiplen Regression ab.** Da für die multiple Regression mehr Informationen verwendet werden als für die lineare Regression, verringert sich üblicherweise die Fehlerquote.

Excel-Tools für multiple Regression

Das Gute an den Excel-Tools für multiple Regression ist, dass es dieselben sind, die Sie im Zusammenhang mit der linearen Regression kennengelernt haben. Sie verwenden die Tools nur etwas anders.

Das Schlechte an den Tools ist … nun … hm … mir fällt nichts Schlechtes ein!

Noch einmal TREND

Ich beginne mit der Funktion TREND. Ich habe Ihnen bereits gezeigt, wie Sie mit dieser Funktion Werte basierend auf einer *x*-Variablen schätzen können. Ändern Sie Ihre Eingaben im Dialogfeld, und schon können Sie Werte basierend auf mehreren Variablen schätzen.

In Abbildung 14.21 sind das Dialogfeld FUNKTIONSARGUMENTE für die Funktion TREND sowie die Daten für 20 Studenten dargestellt. Bei den Daten habe ich eine Spalte für den Notendurchschnitt an der High School der einzelnen Studenten hinzugefügt. In der Abbildung ist außerdem die für die Schätzung der Funktion TREND markierte Spalte zu sehen. Auch die Bearbeitungsleiste ist sichtbar, damit Sie sehen, wie die Formel aussieht.

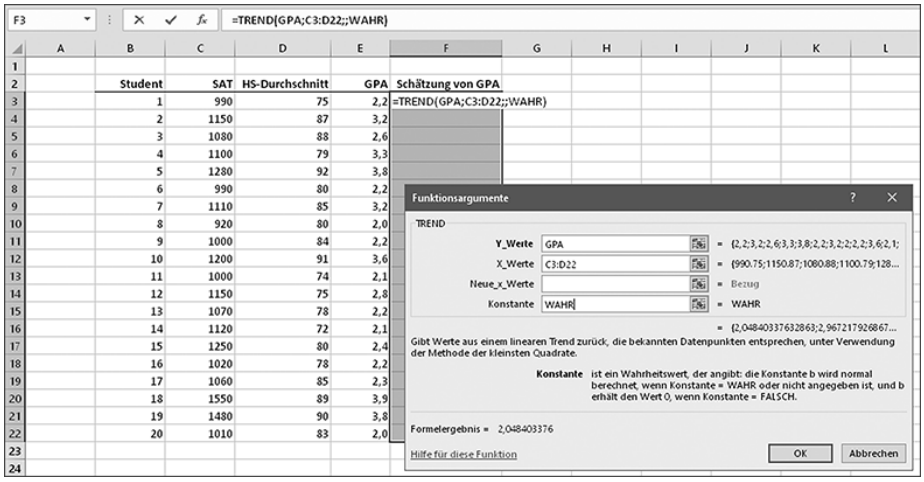


Abbildung 14.21: Das Dialogfeld FUNKTIONSARGUMENTE für TREND mit Daten. Das Dialogfeld ist zum Schätzen der GPAs für die SAT-Werte und die High-School-Durchschnittsnoten ausgefüllt.

Führen Sie diese Schritte durch:

1. **Markieren Sie nach der Eingabe der Daten eine Spalte für das Ergebnis aus der Berechnung der Funktion TREND.**

Ich habe die Zellen F3:F22 markiert. Damit werden die geschätzten GPAs rechts neben den GPAs aus der Stichprobe eingefügt.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf TREND, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   In das Feld Y_WERTE habe ich den Namen des Zellbereichs eingegeben, in dem sich die Werte für die *y*-Variable befinden. In diesem Beispiel ist dies GPA (der für den Zellbereich E3:E22 definierte Name).

   In das Feld X_WERTE habe ich den Zellbereich eingegeben, in dem sich die Werte für die *x*-Variable befinden. In diesem Beispiel ist dies der Zellbereich C3:D22.

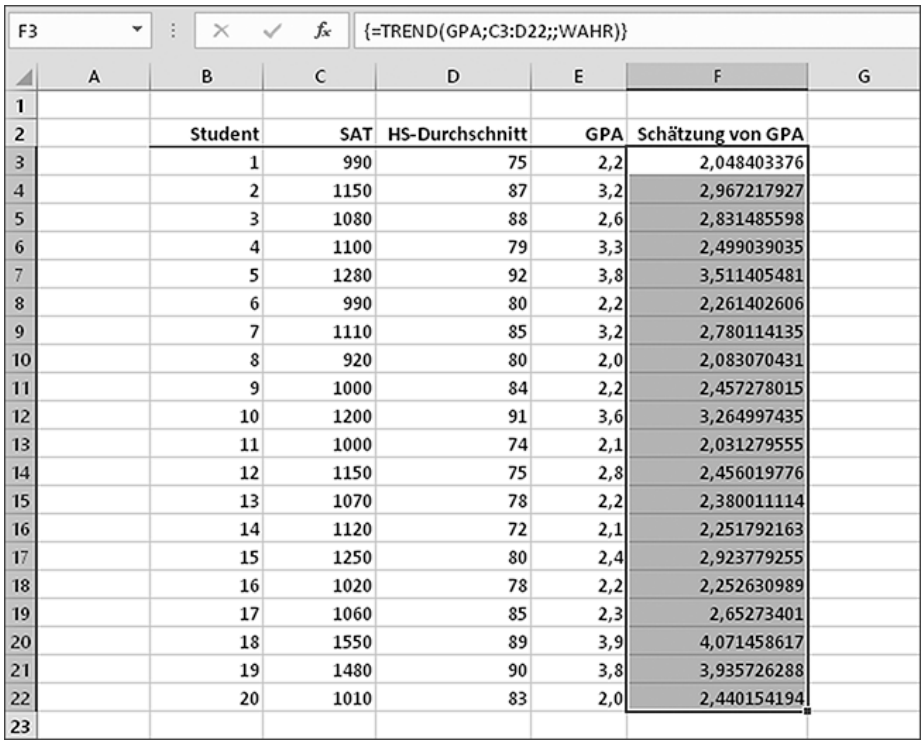
   Das Feld NEUE_X_WERTE habe ich leer gelassen.

   In das Feld KONSTANTE kann WAHR eingegeben werden (es kann auch leer bleiben), um den *y*-Achsenabschnitt zu berechnen. Mit FALSCH wird der *y*-Achsenabschnitt auf null gesetzt. Ich habe WAHR eingegeben. (Ich wüsste nicht, warum jemand FALSCH eingeben sollte.) Ein wichtiger Hinweis: Im Dialogfeld bezieht sich die Anweisung für das Feld KONSTANTE auf *b*. Das ist der *y*-Achsenabschnitt. Weiter vorn in diesem Kapitel habe ich *a* zur Darstellung des *y*-Achsenabschnitts und *b* zur Darstellung der Steigung verwendet. Hierfür gibt es keine einheitliche Regel. Außerdem liest sich der Text im Dialogfeld so, als könnte diese Funktion nur für lineare Regression angewendet werden. Wie Sie gleich sehen werden, lässt sie sich aber auch für multiple Regression verwenden.

4. **WICHTIG: Klicken Sie nicht auf OK. Da es sich hierbei um eine Feldfunktion handelt, müssen Sie die**

Tastenkombination Strg+Shift+Enter drücken, damit das Ergebnis der Funktion TREND in die markierte Spalte eingefügt wird.

In Abbildung 14.22 ist das Ergebnis in den Zellen F3:F22 dargestellt. Beachten Sie den Unterschied in den Bearbeitungsleisten in Abbildung 14.21: Das Dialogfeld FUNKTIONSARGUMENTE für *TREND* mit Daten. Das Dialogfeld ist zum Schätzen der GPAs für die SAT-Werte und die High-School-Durchschnittsnoten ausgefüllt. Abbildung 14.21 und Abbildung 14.22. Nachdem die Funktion Ihre Arbeit erledigt hat, ergänzt Excel die geschweiften Klammern, um anzuzeigen, dass es sich um eine Arrayformel handelt.



F3 {=TREND(GPA;C3:D22;;WAHR)}

	A	B	C	D	E	F	G
1							
2		Student	SAT	HS-Durchschnitt	GPA	Schätzung von GPA	
3		1	990	75	2,2	2,048403376	
4		2	1150	87	3,2	2,967217927	
5		3	1080	88	2,6	2,831485598	
6		4	1100	79	3,3	2,499039035	
7		5	1280	92	3,8	3,511405481	
8		6	990	80	2,2	2,261402606	
9		7	1110	85	3,2	2,780114135	
10		8	920	80	2,0	2,083070431	
11		9	1000	84	2,2	2,457278015	
12		10	1200	91	3,6	3,264997435	
13		11	1000	74	2,1	2,031279555	
14		12	1150	75	2,8	2,456019776	
15		13	1070	78	2,2	2,380011114	
16		14	1120	72	2,1	2,251792163	
17		15	1250	80	2,4	2,923779255	
18		16	1020	78	2,2	2,252630989	
19		17	1060	85	2,3	2,65273401	
20		18	1550	89	3,9	4,071458617	
21		19	1480	90	3,8	3,935726288	
22		20	1010	83	2,0	2,440154194	
23							

Abbildung 14.22: Das Ergebnis von *TREND*: Geschätzte GPAs für die SAT-Werte und die High-School-Durchschnittsnoten in der Stichprobe

Mit der Funktion TREND können Sie also die Werte schätzen und ich habe Ihnen noch nicht einmal gezeigt, wie Sie die Koeffizienten ermitteln!

Noch einmal RGP

Um die Koeffizienten der multiplen Regression zu ermitteln, verwende ich wieder RGP.

In Abbildung 14.23 ist das Dialogfeld FUNKTIONSARGUMENTE für die Funktion RGP zusammen mit den Daten und dem markierten Zellbereich für das Ergebnis dargestellt. Der markierte Zellbereich besteht aus fünf Zeilen und drei Spalten. Er besteht immer aus fünf Zeilen. Die Anzahl der Spalten entspricht dagegen der Anzahl der Regressionskoeffizienten. Bei der linearen Regression sind es zwei: die Steigung und der Achsenabschnitt. Bei diesem Beispiel der multiplen Regression sind es drei.

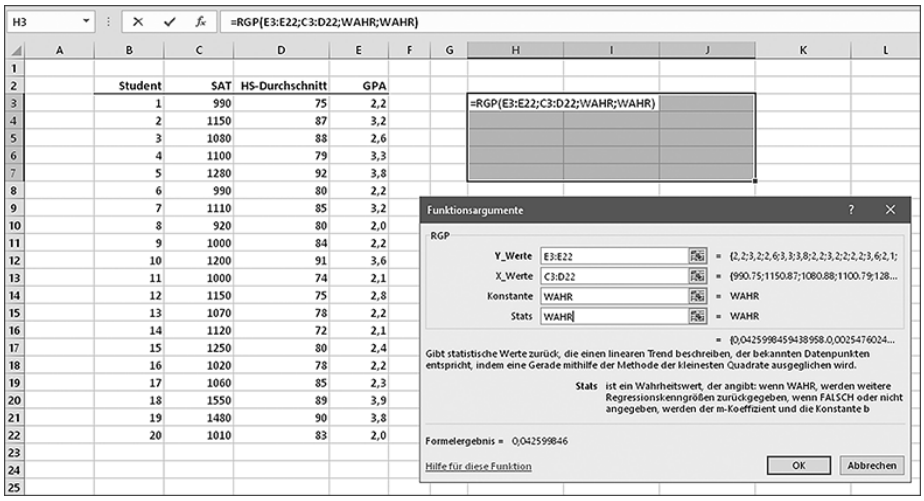


Abbildung 14.23: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *RGP* zusammen mit den Daten und dem markierten Zellbereich für das Ergebnis einer multiplen Regression

Und so verwenden Sie die Funktion RGP für eine multiple Regression mit drei Koeffizienten:

1. **Markieren Sie nach der Eingabe der Daten einen Zellbereich mit fünf Zeilen und drei Spalten für das Ergebnis der Funktion RGP.**

   Ich habe die Zellen H3:J7 markiert.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf RGP, um das Dialogfeld FUNKTIONSARGUMENTE für diese**

Funktion zu öffnen.

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Geben Sie in das Feld Y_WERTE den Zellbereich ein, in dem sich die Werte für die *y*-Variable befinden. In diesem Beispiel ist dies der Bereich E3:E22 mit den GPAs.

   Geben Sie in das Feld X_WERTE den Zellbereich ein, in dem sich die Werte für die *x*-Variable befinden. In diesem Beispiel ist dies der Zellbereich C3:D22 mit den SAT-Werten und den High-School-Durchschnittsnoten.

   Geben Sie im Feld KONSTANTE `WAHR` ein (oder lassen Sie es leer), um den *y*-Achsenabschnitt zu berechnen. Geben Sie `FALSCH` ein, um den *y*-Achsenabschnitt auf null zu setzen. Ich habe WAHR eingegeben.

   Geben Sie im Feld STATS `WAHR` ein, um neben der Steigung und dem Achsenabschnitt auch Regressionswerte zu ermitteln. Geben Sie `FALSCH` ein (oder lassen Sie das Feld leer), um nur die Steigung und den Achsenabschnitt zu ermitteln. Ich habe `WAHR` eingegeben. Im Dialogfeld wird der Achsenabschnitt als *b* und die anderen Koeffizienten als *m-Koeffizienten* bezeichnet. Ich verwende *a* für die Steigung und *b* für die anderen Koeffizienten. Hierfür gibt es keinen einheitlichen Standard.

4. **WICHTIG: Klicken Sie nicht auf OK. Da es sich hierbei um eine Feldfunktion handelt, müssen Sie die Tastenkombination Strg+Shift+Enter drücken, damit das Ergebnis der Berechnung der Funktion `RGP` in den markierten Zellbereich eingefügt wird.**

In Abbildung 14.24 ist das Ergebnis der Funktion `RGP` dargestellt. Es ist nicht beschriftet. Daher habe ich die Beschriftungen im

Arbeitsblatt hinzugefügt. Ich habe auch einen Teil des Ergebnisses umrahmt, damit deutlicher wird, was wozu gehört.

Die Einträge, die auffallen, sind die hässlichen #NV-Zeichen in den letzten drei Zeilen in der rechten Spalte. Diese zeigen an, dass `RGP` in diese Zellen nichts einfügt.

In den oberen beiden Zeilen des Zellbereichs sind die Werte und Standardfehler für die Koeffizienten enthalten. Ich habe diese Zeilen umrandet, um sie von den anderen drei Zeilen zu trennen, in denen Informationen anders angezeigt werden. Bevor ich zu diesen Zeilen komme, möchte ich Ihnen nur kurz sagen, dass die oberste Zeile die Informationen enthält, die Sie zum Aufstellen der Regressionsgleichung benötigen.

$$Y' = -3{,}67 + 0{,}0025x_1 + 0{,}043x_2$$

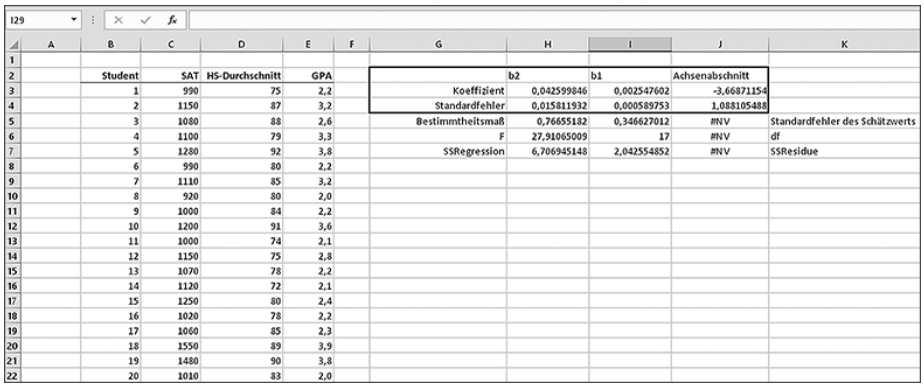
Unter Berücksichtigung von SAT-Wert, GPA und Durchschnittsnote an der High School ergibt sich Folgendes:

Schätzwert von GPA = –3,67 + 0,0025(SAT) + *0,043* (High-School-Durchschnittsnote)

In der dritten Zeile sind das Bestimmtheitsmaß (ein Maß dafür, wie eng GPA und die anderen beiden Variablen zusammenhängen; mehr dazu in Kapitel 15) und der Standardfehler des Schätzwerts angegeben. Wenn Sie den Standardfehler des Schätzwerts für die multiple Regression mit dem Standardfehler für die lineare Regression vergleichen, werden Sie feststellen, dass derjenige für die multiple Regression kleiner ist. (Aber nur keine Panik. Ich übernehme das für Sie. Der Wert beträgt 0,04 für die lineare Regression und 0,35 für die multiple Regression.)

In der vierten Zeile werden der *F*-Quotient (mit dem die Hypothese getestet wird, ob die Gerade gut an das Streudiagramm angepasst ist) und der df-Wert für den Nenner des *F*-Quotienten angezeigt. Der df-Wert für den Zähler (nicht angezeigt) setzt sich aus der Anzahl der Koeffizienten − 1 zusammen. Mit der Funktion FINV können Sie prüfen, ob dieser *F*-Quotient bei df = 2 und 17 signifikant ist.

In der letzten Zeile werden $SS_{Regression}$ und $SS_{Residue}$ angezeigt.



	A	B	C	D	E	F	G	H	I	J	K
1											
2		Student	SAT	HS-Durchschnitt	GPA			b2	b1	Achsenabschnitt	
3		1	990	75	2,2		Koeffizient	0,042599846	0,002547602	-3,66871154	
4		2	1150	87	3,2		Standardfehler	0,015811932	0,000589753	1,088105488	
5		3	1080	88	2,6		Bestimmtheitsmaß	0,76655182	0,346627012	#NV	Standardfehler des Schätzwerts
6		4	1100	79	3,3		F	27,91065009	17	#NV	df
7		5	1280	92	3,8		SSRegression	6,706945148	2,042554852	#NV	SSResidue
8		6	990	80	2,2						
9		7	1110	85	3,2						
10		8	920	80	2,0						
11		9	1000	84	2,2						
12		10	1200	91	3,6						
13		11	1000	74	2,1						
14		12	1150	75	2,8						
15		13	1070	78	2,2						
16		14	1120	72	2,1						
17		15	1250	80	2,4						
18		16	1020	78	2,2						
19		17	1060	85	2,3						
20		18	1550	89	3,9						
21		19	1480	90	3,8						
22		20	1010	83	2,0						

Abbildung 14.24: Mehrere Ergebnisse der Funktion *RGP* im markierten Zellbereich

Noch einmal das Datenanalysetool Regression

Sie verwenden das Datenanalysetool Regression so, wie Sie die Funktionen `TREND` und `RGP` für die multiple Regression verwenden. Geben Sie das entsprechende Datenfeld für die *x*-Variablen an und schon kann's losgehen.

Und so müssen Sie vorgehen:

1. **Geben Sie die Daten in das Tabellenblatt sowie Beschriftungen für die Spalten ein.**

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Wählen Sie im Dialogfeld ANALYSE-FUNKTIONEN die Option REGRESSION aus.**

4. **Klicken Sie auf OK, um das Dialogfeld REGRESSION zu öffnen.**

   In Abbildung 14.15 ist dieses Dialogfeld dargestellt.

5. **Geben Sie im Feld Y-EINGABEBEREICH den Zellbereich ein, in dem sich die Daten für die *y*-Variable befinden.**

Wie bei dem Beispiel für `RGP` befinden sich auch hier die GPAs (mit Beschriftung) in den Zellen $E$2:$E$22. Beachten Sie die Dollarzeichen ($) für den absoluten Bezug.

6. **Geben Sie im Feld X-EINGABEBEREICH den Zellbereich ein, in dem sich die Daten für die *x*-Variable befinden.**

   Die SAT-Werte und die High-School-Durchschnittsnoten (mit Beschriftung) befinden sich in den Zellen $C$2:$D$22.

7. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN.**

   Ich habe die Überschriften in die Zellbereiche mit aufgenommen. Daher habe ich dieses Kontrollkästchen aktiviert.

8. **Im Feld KONFIDENZNIVEAU befindet sich standardmäßig der Wert 95 %, was einem Niveau von Alpha von 0,05 entspricht.**

   Ich habe den Standardwert verwendet.

9. **Wählen Sie im Bereich AUSGABE, wo das Ergebnis ausgegeben werden soll.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, damit das Ergebnis auf einem neuen Arbeitsblatt in der aktuellen Arbeitsmappe angezeigt wird.

10. **Der Bereich RESIDUEN enthält vier Kontrollkästchen, mit denen die Abweichungen zwischen den Datenpunkten und den geschätzten Punkten angezeigt werden können. Aktivieren Sie so viele Kontrollkästchen, wie Sie mögen, oder gar keine.**

    Ich habe alle vier Kontrollkästchen aktiviert.

11. **Aktivieren Sie das Kontrollkästchen QUANTILSPLOT, wenn Sie ein Diagramm mit den Quantilen der *y*-Variable erstellen möchten.**

    Ich habe dieses Kontrollkästchen eingeschaltet.

12. **Klicken Sie auf OK.**

Eine ausführliche Beschreibung des Ergebnisses finden Sie im Abschnitt über das Datenanalysetool Regression. Dieses Ergebnis dasselbe, ist abgesehen von einigen Änderungen und zusätzlichen Angaben aufgrund der neuen Variablen. In Abbildung 14.25 sind die Tabellen *Regressions-Statistik*, *ANOVA* und die *Koeffiziententabelle* dargestellt.

	A	B	C	D	E	F	G	H	I
1	AUSGABE: ZUSAMMENFASSUNG								
2									
3	*Regressions-Statistik*								
4	Multipler Korrelationskoeffizient	0,875529451							
5	Bestimmtheitsmaß	0,76655182							
6	Adjustiertes Bestimmtheitsmaß	0,739087328							
7	Standardfehler	0,346627012							
8	Beobachtungen	20							
9									
10	ANOVA								
11		*Freiheitsgrade (df)*	*Quadratsummen (SS)*	*Quadratsumi*	*Prüfgröße (F)*	*F krit*			
12	Regression	2	6,706945148	3,35347257	27,91065009	4,26206E-06			
13	Residue	17	2,042554852	0,12015029					
14	Gesamt	19	8,7495						
15									
16		*Koeffizienten*	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*Untere 95,0%*	*Obere 95,0%*
17	Schnittpunkt	-3,66871154	1,088105488	-3,37165062	0,003623339	-5,964413448	-1,373009632	-5,964413448	-1,373009632
18	SAT	0,002547602	0,000589753	4,31978097	0,000464766	0,001303333	0,003791872	0,001303333	0,003791872
19	HS-Durchschnitt	0,042599846	0,015811932	2,6941582	0,015361179	0,009239586	0,075960106	0,009239586	0,075960106
20									

Abbildung 14.25: Ein Teil des Ergebnisses aus den Berechnungen des Datenanalysetools REGRESSION: die Regressionsstatistik, die Tabelle ANOVA und die Koeffiziententabelle

In der Tabelle *ANOVA* werden die neuen df-Werte (2, 17 und 19 jeweils für Regression, Residue und Gesamt) angezeigt. Die Koeffiziententabelle enthält zudem Informationen zur Durchschnittsnote an der High School. Sie enthält die Werte aller Koeffizienten sowie Standardfehler und Informationen zum Hypothesentest mittels *t*-Test.

Wenn Sie das Beispiel betrachten, werden Sie die Tabelle mit den Residuen im Ergebnis finden. (Sie ist in Abbildung 14.25 nicht dargestellt.) Wenn Sie die absoluten Werte der Residuen aus der linearen Regression mit denen aus der multiplen Regression

vergleichen, werden Sie feststellen, dass die Werte aus der multiplen Regression im Durchschnitt kleiner sind.

Die grafische Darstellung enthält ebenfalls einige zusätzliche Informationen: ein Streudiagramm mit Durchschnittsnote an der High School und GPA mit geschätzten GPAs sowie ein Diagramm mit Residuen und Durchschnittsnote an der High School.

Kapitel 15

Korrelation: Aufstieg und Fall von Zusammenhängen

In diesem Kapitel

- Verstehen, was es mit der Korrelation auf sich hat
- Korrelation und Regression verbinden
- Rückschlüsse aus Korrelationen ziehen
- Items analysieren

In Kapitel 14 habe ich Ihnen alles über *Regression*, das Tool zum Zusammenfassen von Zusammenhängen zwischen Variablen, erzählt. In diesem Kapitel lernen Sie *Korrelation* kennen, ein weiteres Tool zur Betrachtung von Zusammenhängen.

Ich verwende das Beispiel der SAT-Werte und der GPA aus Kapitel 14 und zeige Ihnen, wie Sie die Daten etwas anders betrachten können. Die neuen Konzepte haben mit dem zu tun, was ich Ihnen im vorherigen Kapitel gezeigt habe, und Sie werden erfahren, inwiefern. Ich werde Ihnen außerdem zeigen, wie Sie Hypothesen über Zusammenhänge testen und wie Sie Excel-Funktionen und Datenanalysetools für Korrelationen verwenden können.

Noch einmal Streudiagramme

Ein *Streudiagramm* ist eine Möglichkeit der grafischen Darstellung von Zusammenhängen zwischen Variablen. In Abbildung 15.1 ist ein Streudiagramm mit den GPAs und SAT-Werten von 20 Studenten an

der fiktiven Universität Sahutsket dargestellt. Die GPAs befinden sich auf einer Skala bis 4,0 und die SAT-Werte sind die Gesamtwerte.

Jeder Punkt steht für einen Studenten. Die Lage eines Punkts in horizontaler Richtung steht für den SAT-Wert des Studenten. Die Lage desselben Punkts in vertikaler Richtung steht für den GPA des Studenten.

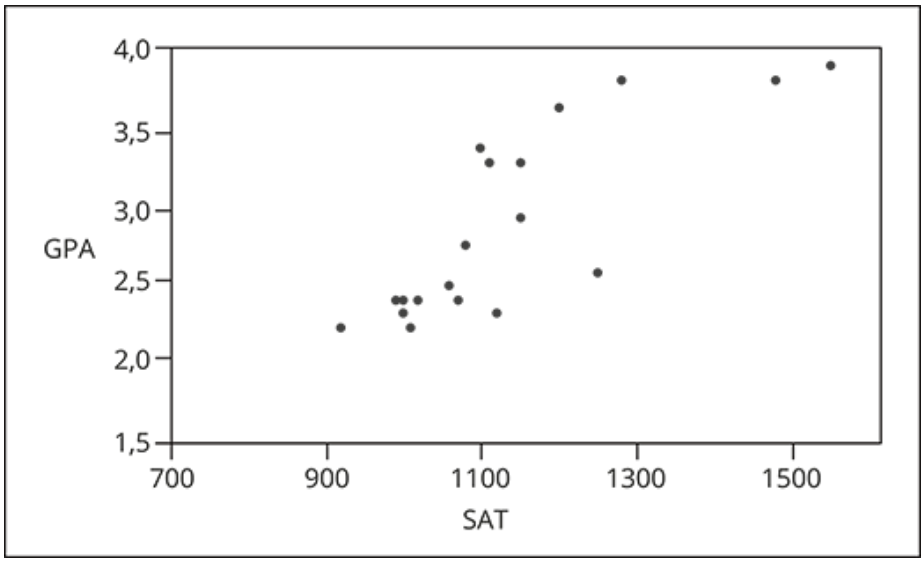


Abbildung 15.1: Ein Streudiagramm mit den SAT-Werten und GPAs von 20 Studenten an der Universität SahutsketGrundlegendes zur Korrelation

In Kapitel 14 habe ich den SAT-Wert als *unabhängige Variable* und GPA als *abhängige Variable* bezeichnet. In Kapitel 14 geht es darum, wie der GPA mit dem SAT-Wert geschätzt werden kann. Ein wichtiger Punkt hierbei ist: Obwohl Werte einer Variablen zum *Schätzen* der Werte der anderen Variablen verwendet werden, bedeutet das *nicht*, dass der Wert der einen Variablen den Wert der anderen *verursacht*. »Zusammenhang« bedeutet nicht zwangsläufig »Kausalität«.

Korrelation stellt eine statistische Möglichkeit zur Betrachtung eines Zusammenhangs dar. Wenn zwei Dinge korrelieren, bedeutet das, dass sie sich gemeinsam verändern. Eine *positive Korrelation* bedeutet, dass hohe Werte bei der einer Variablen mit hohen Werten bei der anderen Variablen und dass niedrige Werte bei der einen Variablen mit niedrigen Werten bei der anderen verknüpft sind. Das

Streudiagramm in Abbildung 15.1 ist ein Beispiel für eine positive Korrelation.

Eine *negative Korrelation* bedeutet dagegen, dass hohe Werte bei der einen Variablen mit *niedrigen* Werten bei der anderen Variablen verknüpft sind. Eine *negative Korrelation* bedeutet auch, dass niedrige Werte bei der einen Variablen mit hohen Werten bei der anderen verknüpft sind. Ein Beispiel hierfür ist die Korrelation zwischen Körpergewicht und Zeit, während der Diät gehalten wird. Wenn die Diät gut ist, ist das Körpergewicht umso geringer, je länger Diät gehalten wird. Und: Je kürzer Diät gehalten wird, umso höher ist das Körpergewicht.

In Tabelle 15.1 (Wiederholung von Tabelle 14.2) sind die Daten aus dem Streudiagramm aufgeführt.

Student	SAT	GPA
1	990	2,2
2	1150	3,2
3	1080	2,6
4	1100	3,3
5	1280	3,8
6	990	2,2
7	1110	3,2
8	920	2,0
9	1000	2,2
10	1200	3,6
11	1000	2,1
12	1150	2,8
13	1070	2,2
14	1120	2,1
15	1250	2,4
16	1020	2,2

Student	SAT	GPA
17	1060	2,3
18	1550	3,9
19	1480	3,8
20	1010	2,0
Mittelwert	1126,5	2,705
Varianz	26171,32	0,46
Standardabweichung	161,78	0,82

Tabelle 15.1: SAT-Werte und GPAs für 20 Studenten der Universität Sahutsket

Wenn wir SAT-Wert und GPA wie in Kapitel 14 verwenden, ist der SAT-Wert die *x*-Variable und GPA die *y*-Variable.

Die Formel für die Berechnung der Korrelation zwischen den beiden lautet wie folgt:

$$r = \frac{\left[\frac{1}{N-1}\right]\sum(x-\bar{x})(y-\bar{y})}{s_x s_y}$$

Der Ausdruck auf der linken Seite, *r*, ist der *Korrelationskoeffizient*. Er wird auch als *Produktmoment-Korrelationskoeffizient nach Pearson* bezeichnet (nach Karl Pearson, der diesen Koeffizienten entwickelt hat).

Bei den beiden Ausdrücken im Nenner auf der rechten Seite handelt es sich um die Standardabweichung der *x*-Variablen und die der *y*-Variablen. Der Ausdruck im Zähler wird als *Kovarianz* bezeichnet. Somit lässt sich diese Formal auch wie folgt formulieren:

$$r = \frac{\text{cov}(x,y)}{s_x s_y}$$

Die Kovarianz gibt an, dass sich *x* und *y* gemeinsam verändern. Bei der Division der Kovarianz durch das Produkt der beiden Standardabweichungen gibt es Grenzen. Die untere Grenze des Korrelationskoeffizienten liegt bei −1,00 und die obere bei +1,00.

Der Korrelationskoeffizient −1,00 stellt eine perfekte negative Korrelation dar (niedrige *x*-Werte sind mit hohen *y*-Werten und umgekehrt verknüpft). Der Korrelationskoeffizient +1,00 stellt eine perfekte positive Korrelation dar (niedrige *x*-Werte sind mit niedrigen *y*-Werten und umgekehrt verknüpft). Eine Korrelation mit dem Wert 0,00 bedeutet, dass zwischen den beiden Variablen kein Zusammenhang besteht.

Setzen wir die Zahlen aus Tabelle 15.1 in die Formel ein:

$$r = \frac{\left[\frac{1}{N-1}\right]\Sigma(x-\bar{x})(y-\bar{y})}{s_x s_y}$$

$$= \frac{\left[\frac{1}{20-1}\right]\left[(990-1126{,}5)(2{,}2-2{,}705)+\ldots+(1010-1126{,}5)\,(2{,}0-2{,}705)\right]}{(161{,}78)\,(0{,}82)} = 0{,}817$$

Was genau bedeutet diese Zahl? Das werden Sie gleich erfahren.

Korrelation und Regression

In Abbildung 15.2 ist das Streudiagramm mit der Geraden dargestellt, die am besten an die Punkte angepasst ist. Es können unendlich viele Geraden durch diese Punkte gezogen werden. Aber welche ist die beste?

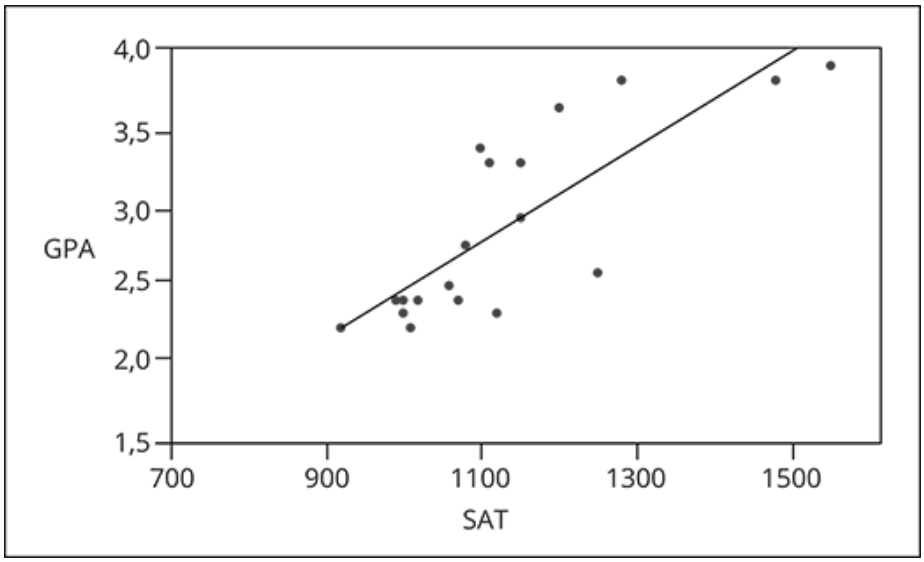


Abbildung 15.2: Streudiagramm der 20 Studenten mit der Regressionsgeraden

Um die »beste« zu sein, muss eine Gerade einem bestimmten Standard entsprechen: Wenn Sie die vertikalen Abstände zwischen den Punkten und der Geraden einzeichnen, die Abstände quadrieren und die Quadrate anschließend addieren, ist die am besten angepasste Gerade diejenige, bei der die Summe dieser quadratischen Abstände am kleinsten ist. Diese Gerade wird als *Regressionsgerade* bezeichnet.

Mit den Regressionsgeraden können Sie Werte schätzen. Wie in Kapitel 14 bereits erwähnt, können Sie den Wert einer *y*-Variablen ohne Regressionsgerade am besten mit dem Mittelwert der *y*-Variablen schätzen. Bei der Regressionsgeraden wird auch die *x*-Variable berücksichtigt, so dass die Schätzung genauer wird. Jeder Punkt auf der Regressionsgeraden stellt einen geschätzten Wert für die *y*-Variable dar. Wenn wir das Symbol aus der Regression verwenden, ist jeder geschätzte Wert ein *y*′.

Warum erzähle ich Ihnen das alles? Weil die Korrelation eng mit der Regression zusammenhängt. In Abbildung 15.3 ist ein Punkt im Streudiagramm und dessen Abstand zur Regressionsgeraden und zum Mittelwert hervorgehoben. (Das ist dieselbe Abbildung wie Abbildung 14.3.)

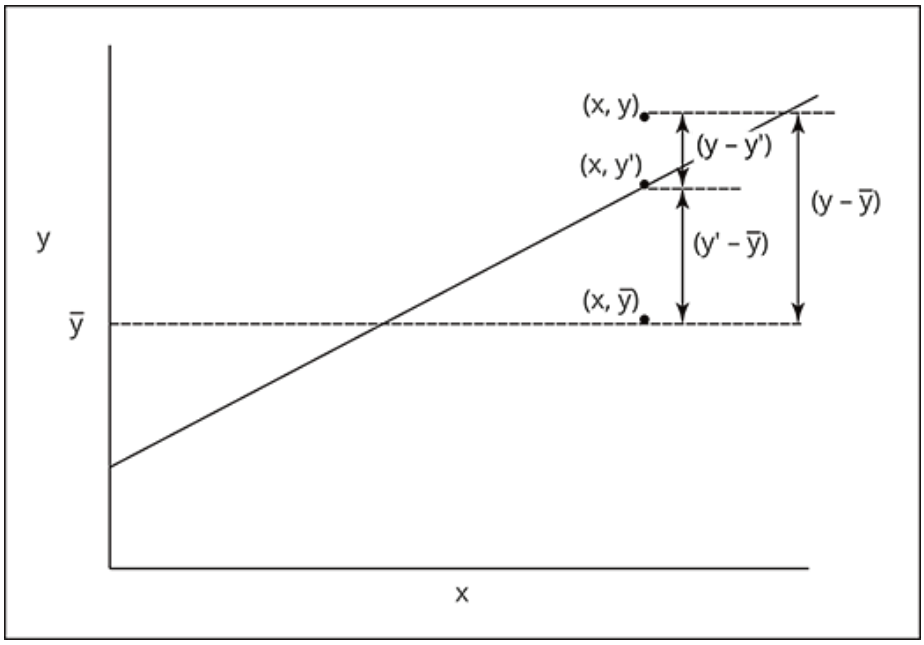


Abbildung 15.3: Ein Punkt im Streudiagramm und die zugehörigen Abstände

Beachten Sie die drei Abstände in der Abbildung. Der mit ($y - y'$) gekennzeichnete Abstand ist die Differenz zwischen dem Punkt und dem durch die Regressionsgeraden ermittelten Schätzwert seiner Lage. (In Kapitel 14 wird dies als *Residue* bezeichnet.) Der mit ($y - \bar{y}$) gekennzeichnete Abstand ist die Differenz zwischen dem Punkt und dem Mittelwert der *y*-Werte. Der mit ($y' - \bar{y}$) gekennzeichnete Abstand stellt die Verbesserung der Schätzung dar, die dadurch erzielt wird, dass der Punkt anstatt mit dem Mittelwert mit der Regressionsgeraden geschätzt wird.

Anhand von Abbildung 15.3 wird ersichtlich, dass die drei Abstände wie folgt zusammenhängen:

$$(y - y') + (y' - \bar{y}) = (y - \bar{y})$$

Wie in Kapitel 14 bereits erwähnt, können diese Residuen quadriert und deren Quadrate addiert, alle Abweichungen der geschätzten Punkte vom Mittelwert quadriert und deren Quadrate addiert und alle Abweichungen der tatsächlichen Punkte vom Mittelwert quadriert und deren Quadrate addiert werden.

Dabei stellt sich heraus, dass zwischen den Summen der Quadrate derselbe Zusammenhang besteht wie zwischen den Abweichungen, die ich Ihnen gerade gezeigt habe:

$$SS_{Residue} + SS_{Regression} = SS_{Gesamt}$$

Wenn $SS_{Regression}$ im Vergleich zu $SS_{Residue}$ groß ist, ist das ein Hinweis darauf, dass zwischen der *x*-Variablen und der *y*-Variablen ein enger Zusammenhang besteht. Das bedeutet, dass im Streudiagramm die Streuung um die Regressionsgerade gering ist.

Wenn dagegen $SS_{Regression}$ im Vergleich zu $SS_{Residue}$ klein ist, ist das ein Hinweis darauf, dass zwischen der *x*-Variablen und der *y*-Variablen kein enger Zusammenhang besteht. In diesem Fall ist die Streuung um die Regressionsgerade im Streudiagramm groß.

Eine Möglichkeit, $SS_{Regression}$ gegen $SS_{Residue}$ zu testen, besteht darin, jede Quadratsumme durch die zugehörigen Freiheitsgrade (1 für $SS_{Regression}$ und $N - 2$ für $SS_{Residue}$) zu dividieren, um so die Varianzschätzwerte zu ermitteln, und anschließend einen Wert durch den anderen zu dividieren, um so den *F*-Quotienten zu ermitteln. Wenn $MS_{Regression}$ signifikant größer ist als $MS_{Residue}$, hängen *x* und *y* eng zusammen. (Ausführlichere Informationen hierzu finden Sie in Kapitel 14.)

Hier ist der Beweis hinsichtlich der Korrelation: Die Größe von $SS_{Regression}$ lässt sich auch durch den Vergleich mit SS_{Gesamt} ermitteln. Dividieren Sie dazu die erste Quadratsumme durch die zweite. Wenn der Quotient groß ist, hängen *x* und *y* eng zusammen. Dieser Quotient hat einen Namen. Er wird als *Bestimmtheitsmaß* oder *Bestimmtheitskoeffizient* bezeichnet. Als Symbol wird r^2 verwendet. Wenn Sie die Wurzel aus diesem Koeffizienten ziehen, ergibt sich … der Korrelationskoeffizient!

$$r = \pm\sqrt{r^2} = \pm\sqrt{\frac{SS_{Regresion}}{SS_{Gesamt}}}$$

Das Plus-Minus-Zeichen ($\pm$) bedeutet, dass *r*, je nachdem, ob die Steigung der Regressionsgeraden positiv oder negativ ist, entweder die positive oder negative Wurzel ist.

Wenn Sie also einen Korrelationskoeffizienten berechnen und schnell wissen möchten, was sein Wert bedeutet, quadrieren Sie ihn einfach. Anhand des Ergebnisses, also des Bestimmtheitsmaßes, können Sie den Prozentsatz von SS_{Gesamt} ermitteln, der sich im Zusammenhang zwischen der *x*- und der *y*-Variablen widerspiegelt. Wenn der Prozentsatz groß ist, weist der Korrelationskoeffizient auf einen engen Zusammenhang hin. Wenn der Prozentsatz klein ist, weist der Korrelationskoeffizient auf keinen engen Zusammenhang hin.

Im Beispiel mit dem GPA und dem SAT-Wert beträgt der Korrelationskoeffizient 0,817. Das Bestimmtheitsmaß beträgt

$$r^2 = (0{,}817)^2 = 0{,}667$$

Bei dieser Stichprobe von 20 Studenten beträgt $SS_{Regression}$ 66,7 % von SS_{Gesamt}. Hört sich wie ein großer Prozentsatz an. Aber was ist groß? Was ist klein? Diese Fragen schreien förmlich nach Hypothesentests.

Hypothesen über Korrelationen testen

In diesem Abschnitt zeige ich Ihnen, wie Sie wichtige Fragen zu Korrelationen beantworten können. Wie bei jedem anderen Hypothesentest geht es darum, mit Stichprobenwerten Rückschlüsse auf die Parameter der Grundgesamtheit zu ziehen. Hier ist der Stichprobenwert *r*, der Korrelationskoeffizient. Für den Parameter der Grundgesamtheit wird üblicherweise ρ (rho), das griechische Pendant zum Buchstaben *r*, verwendet. (Ja, dieser Buchstabe sieht wie unser Buchstabe *p* aus, ist aber tatsächlich der griechische Buchstabe für *r*.)

Im Zusammenhang mit der Korrelation stellen sich zwei wichtige Fragen: 1. Ist ein Korrelationskoeffizient größer als null? 2. Unterscheiden sich zwei Korrelationskoeffizienten voneinander?

Ist ein Korrelationskoeffizient größer als null?

Wenn wir wieder zu unserem Beispiel mit den SAT-Werten und GAPs von der Universität Sahutsket zurückkehren, können Sie mit dem *r* der Stichprobe Hypothesen über das ρ der Grundgesamtheit, den Korrelationskoeffizienten für alle Studenten an der Universität Sahutsket, testen.

Angenommen, Sie wissen im Voraus (vor dem Sammeln der Stichprobendaten), dass die Korrelation zwischen SAT und GPA positiv ist, dann lauten die Hypothesen wie folgt:

$$H_0 : \rho \leq 0$$
$$H_1 : \rho > 0$$

Für α lege ich 0,05 fest.

Der hier geeignete statistische Test ist der *t*-Test. Die Gleichung hierfür lautet:

$$t = \frac{r - \rho}{s_r}$$

Der df-Wert bei diesem Test lautet $N - 2$.

Bei diesem Beispiel werden im Zähler für $r = 0{,}817$ und für ρ (in H_0) null festgelegt. Und was ist mit dem Nenner? Ich werde Sie nicht mit lästigen Details quälen. Nur so viel sei gesagt: Der Nenner setzt sich wie folgt zusammen:

$$\sqrt{\frac{1 - r^2}{N - 2}}$$

Mit ein bisschen Algebra lässt sich die Formel für den *t*-Test wie folgt vereinfachen:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Somit gilt für dieses Beispiel:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0{,}817\sqrt{20-2}}{\sqrt{1-0{,}817^2}} = 6{,}011$$

Wobei df = 18 und α = 0,05 (einseitiger Test) und der kritische Wert von *t* = 2,10 ist. (Sie können die Werte mit der Tabellenblattfunktion `T.INV` überprüfen.) Da der berechnete Wert größer ist als der kritische Wert, wird H_0 verworfen.

Unterscheiden sich zwei Korrelationskoeffizienten voneinander?

In einer Stichprobe von 24 Studenten am Farshimmelt College beträgt der Wert für die Korrelation zwischen SAT und GPA 0,752. Unterscheidet sich dieser Wert vom Wert für die Korrelation (0,817) an der Universität Sahutsket? Wenn Sie nicht davon ausgehen können, dass ein Korrelationswert größer sein sollte als der andere, lauten die Hypothesen wie folgt:

$$H_0 : \rho_{Sahutket} = \rho_{Farshimmelt}$$
$$H_1 : \rho_{Sahutket} \neq \rho_{Farshimmelt}$$

α beträgt wieder = 0,05.

Aus ausgesprochen technischen Gründen kann hierfür kein *t*-Test verwendet werden. Es ist sogar so, dass Sie hier nicht einmal mit den beiden Korrelationskoeffizienten 0,817 und 0,752 arbeiten können.

Stattdessen müssen Sie die Korrelationskoeffizienten in etwas anderes *transformieren* und dann diese beiden »etwas anderes« in eine Formel einsetzen, die – glauben Sie es oder nicht – einen *z*-Test ergibt.

Die Transformation wird als *Fisher-z-Transformation* bezeichnet. Fisher ist der Statistiker, an den das *F* im *F*-Test erinnert. Er transformierte das *r* wie folgt in ein *z*:

$$z_r = \frac{1}{2}[log_e(1+r) - log_e(1-r)]$$

Wenn Sie wissen, was $\log_e$ bedeutet, ist das prima. Wenn nicht, brauchen Sie sich keine Gedanken zu machen. (Ich werde das in Kapitel 22 erläutern.) Excel kümmert sich um all das, wie Sie gleich sehen werden.

Für dieses Beispiel gilt jedenfalls:

$$z_{0,817} = \frac{1}{2}[log_e(1+0,817) - log_e(1-0,817)]$$

$$z_{0,752} = \frac{1}{2}[log_e(1+0,752) - log_e(1-0,752)]$$

Nach dem Transformieren von *r* in *z* lautet die Formel wie folgt:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

Der Nenner stellt sich als einfacher heraus, als Sie vielleicht denken. Er lautet wie folgt:

$$\sigma_{z_1-z_2} = \sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}}$$

Für dieses Beispiel gilt:

$$\sigma_{z_1-z_2} = \sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}} = \sqrt{\frac{1}{20-3}+\frac{1}{20-3}} = 0{,}326$$

Die gesamte Gleichung lautet:

$$Z = \frac{z_1-z_2}{\sigma_{z_1-z_2}} = \frac{1{,}1477-0{,}9775}{0{,}326} = 0{,}522$$

Als Nächstes muss der berechnete Wert mit einer Standardnormalverteilung verglichen werden. Bei einem zweiseitigen Test und bei α = 0,05 beträgt der kritische Wert in einer Standardnormalverteilung 1,96 an der oberen und –1,96 an der unteren Seite. Der berechnete Wert liegt zwischen diesen beiden Werten, so dass H_0 nicht verworfen wird.

Tabellenblattfunktionen für die Korrelation

Excel stellt zwei Tabellenblattfunktionen für die Berechnung von Korrelationen bereit, und beide tun dasselbe auf dieselbe Art und Weise! Warum Excel zwei Funktionen, `KORREL` und `PEARSON`, enthält, weiß ich nicht. Jedenfalls handelt es sich hierbei um die beiden Hauptfunktionen für die Korrelation.

Die anderen Funktionen sind `BESTIMMTHEITSMASS`, `KOVARIANZ.P` und `KOVARIANZ.S`. Mit `BESTIMMTHEITSMASS` wird das Bestimmtheitsmaß (Quadrat des Korrelationskoeffizienten) berechnet. `KOVARIANZ.P` berechnet die Kovarianz, wie ich es im Abschnitt »Grundlegendes zur Korrelation« beschrieben habe. Diese Funktion verwendet *N*–1. `KOVARIANZ.S` verwendet *N*.

KORREL und PEARSON

In Abbildung 15.4 sind die Daten aus dem Beispiel mit den SAT-Werten und GAPs von der Universität Sahutsket und das Dialogfeld FUNKTIONSARGUMENTE für die Funktion KORREL dargestellt.

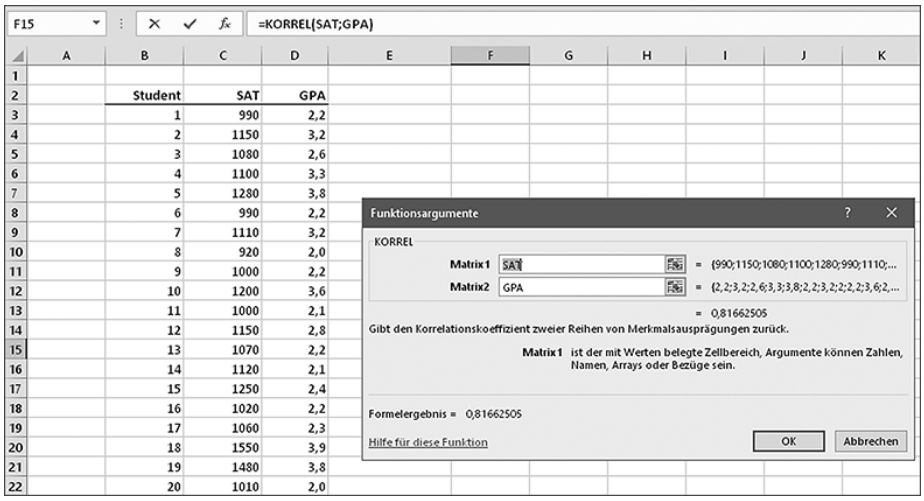


Abbildung 15.4: Das Dialogfeld FUNKTIONSARGUMENTE für KORREL mit den Daten

Gehen Sie wie folgt vor, um diese Funktion zu verwenden:

1. **Geben Sie die Daten in die Zellbereiche ein und markieren Sie die Zelle, in die das Ergebnis der Funktion KORREL eingefügt werden soll.**

   Ich habe die SAT-Werte in die Zellen C3:C22 und die GPA-Daten in die Zellen D3:D22 eingegeben. Dem Zellbereich C3:C22 habe ich den Namen SAT und dem Zellbereich D3:D22 den Namen GPA zugewiesen. (Werfen Sie einen Blick in Kapitel 2, wenn Sie nicht genau wissen, wie Sie in Excel Namen definieren.) Anschließend habe ich Zelle F15 markiert.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf KORREL, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

In das Feld MATRIX1 habe ich SAT eingegeben; dies ist der Name, den ich für den Zellbereich C3:C22 definiert habe. Dieser Zellbereich enthält die Werte für eine der Variablen.

In das Feld MATRIX2 habe ich GPA eingegeben; dies ist der Name, den ich für den Zellbereich D3:D22 definiert habe. Dieser Zellbereich enthält die Werte für die andere Variable.

Nachdem ich die Werte der beiden Argumente eingegeben habe, wird das Ergebnis (0,81662505) im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, damit das Ergebnis in die markierte Zelle eingefügt wird.**

   Wenn Sie anstelle der Option KORREL die Option PEARSON auswählen, erhalten Sie auf genau dieselbe Art und Weise exakt dasselbe Ergebnis.

Itemanalyse: Eine nützliche Anwendung der Korrelation

Lehrer möchten oft gerne wissen, in welchem Zusammenhang die Leistung bei einer bestimmten Prüfungsfrage mit der Prüfungsgesamtleistung steht. Im Idealfall beantwortet jemand, der den Stoff gelernt hat, die Frage richtig, jemand, der nicht gelernt hat, falsch. Wenn die Fragen von allen richtig – oder falsch – beantwortet wird, ist die Frage sinnlos. Diese Beurteilung wird als *Itemanalyse* bezeichnet.

Nehmen wir an, die Prüfungsfrage kann entweder richtig oder falsch beantwortet und die Prüfung mit 0 bis 100 Punkten bewertet werden. Sie können nun beispielsweise einer falschen Antwort auf die Frage den Wert 0 und einer richtigen Antwort den Wert 1 zuweisen und dann einen Korrelationskoeffizienten berechnen, bei dem jedes Wertepaar für die Frage entweder 0 oder 1 und für die Prüfung eine Zahl zwischen 0 und 100 ist. Der Wert für die Prüfung wird als *dichotome Variable* bezeichnet und diese Art von Korrelation als *punktbiseriale Korrelation*.

Wenn die punktbiseriale Korrelation für eine Prüfungsfrage groß ist, sollten Sie diese Frage stellen. Wenn die Korrelation gering ist, ist die Frage aller Wahrscheinlichkeit nach sinnlos.

Da eine der Variablen nur die Werte 0 oder 1 annehmen kann, unterscheidet sich die Formel für den biserialen Korrelationskoeffizienten ein wenig von der Formel für den normalen Korrelationskoeffizienten. Wenn Sie Excel für die Berechnung verwenden, spielt das jedoch keine Rolle. Verwenden Sie einfach die Funktionen KORREL (oder PEARSON) wie beschrieben.

BESTIMMTHEITSMASS

Wenn Sie das Bestimmtheitsmaß (r^2) schnell berechnen müssen, ist BESTIMMTHEITSMASS die richtige Funktion für Sie. Einen wirklichen Bedarf sehe ich für diese Funktion nicht, da es kein Problem ist, einfach KORREL zu verwenden und das Ergebnis zu quadrieren.

Wenn Sie das Dialogfeld FUNKTIONSARGUMENTE für die Funktion BESTIMMTHEITSMASS für dieses Beispiel ausfüllen, sieht die Bearbeitungsleiste folgendermaßen aus:

```
=BESTIMMTHEITSMASS(GPA;SAT)
```

Im Hinblick auf das Dialogfeld unterscheidet sich diese Funktion von KORREL (und PEARSON) nur dadurch, dass die Felder, die Sie ausfüllen, nicht MATRIX1 und MATRIX2, sondern Y_WERTE und X_WERTE heißen.

KOVARIANZ.P und KOVARIANZ.S

Wenn es um die Berechnung von Korrelationen geht, sehe ich für diese Funktionen keine wirkliche Notwendigkeit. Vollständigkeitshalber will ich dennoch erwähnen, dass KOVARIANZ.P die Kovarianz mit dieser Formel berechnet:

$$\text{Kovarianz} = \left[\frac{1}{N}\right]\Sigma(x-\bar{x})(y-\bar{y})$$

KOVARIANZ.S verwendet für die Berechnung die folgende Formel:

$$\text{Kovarianz} = \left[\frac{1}{N-1}\right]\Sigma(x-\bar{x})(y-\bar{y})$$

Das P im Namen der ersten Funktion weist Sie daraufhin, dass die Funktion die Kovarianz für eine Population berechnet; an dem S im Namen der anderen Funktion erkennen Sie, dass die Funktion die Kovarianz einer Stichprobe berechnet (oder genauer gesagt, die Kovarianz in einer Population schätzt).

Sie verwenden diese Funktionen genauso wie die Funktion KORREL. Nachdem Sie beispielsweise im Dialogfeld FUNKTIONSARGUMENTE die Felder für die Funktion KOVARIANZ.P für dieses Beispiel ausgefüllt haben, steht in der Bearbeitungsleiste Folgendes:

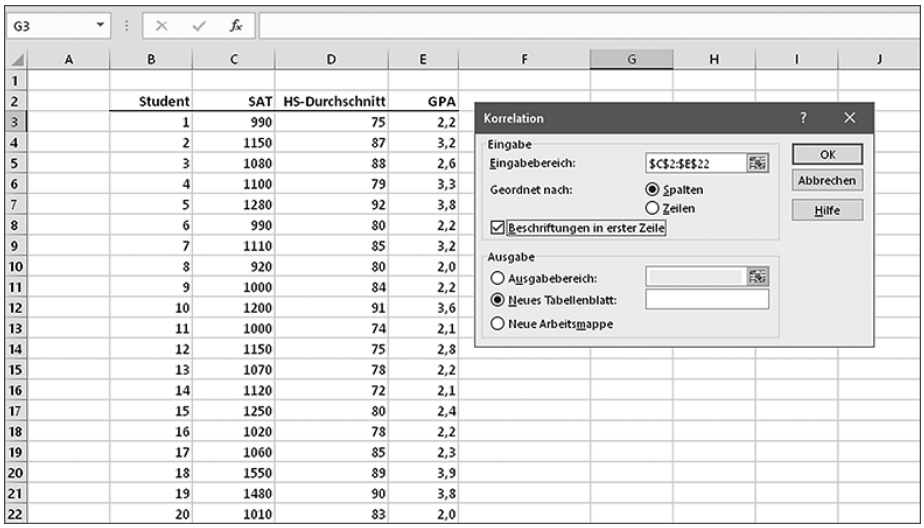
```
=KOVARIANZ.P(SAT;GPA)
```

Wenn Sie mit dieser Funktion *r* berechnen möchten, dividieren Sie das Ergebnis durch das Produkt aus STABW.N(SAT) und STABW.N(GPA). Ich überlasse es Ihnen herauszufinden, wie Sie mit KOVARIANZ.S *r* berechnen. Ich weiß nicht, wozu das alles, wenn Sie auch einfach die Funktion KORREL verwenden können.

Datenanalysetool: Korrelation

Wenn Sie einen einzelnen Korrelationskoeffizienten berechnen müssen, werden Sie feststellen, dass das Datenanalysetool Korrelation von Excel dasselbe bewirkt wie die Funktion KORREL, das Ergebnis jedoch in Form einer Tabelle anzeigt. Dieses Tool ist dann hilfreich, wenn Sie für mehrere Daten mehrere Korrelationen berechnen müssen.

In Abbildung 15.5 sind beispielsweise SAT-Wert, Notendurchschnitt an der High School und GPA für 20 Studenten an der Universität Sahutsket zusammen mit dem Dialogfeld für das Datenanalysetool Korrelation dargestellt.



Student	SAT	HS-Durchschnitt	GPA
1	990	75	2,2
2	1150	87	3,2
3	1080	88	2,6
4	1100	79	3,3
5	1280	92	3,8
6	990	80	2,2
7	1110	85	3,2
8	920	80	2,0
9	1000	84	2,2
10	1200	91	3,6
11	1000	74	2,1
12	1150	75	2,8
13	1070	78	2,2
14	1120	72	2,1
15	1250	80	2,4
16	1020	78	2,2
17	1060	85	2,3
18	1550	89	3,9
19	1480	90	3,8
20	1010	83	2,0

Abbildung 15.5: Das Dialogfeld für das Datenanalysetool Korrelation zusammen mit Daten für den SAT-Wert, den Notendurchschnitt an der High School und den GPA

Und so arbeiten Sie mit diesem Tool:

1. **Geben Sie in das Tabellenblatt die Daten sowie die Beschriftungen für die Spalten ein.**

   In diesem Beispiel befinden sich die Daten (einschließlich der Beschriftung) in den Zellen C2 bis E22.

2. **Klicken Sie im Menüband auf DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

3. **Blättern Sie im Dialogfeld ANALYSE-FUNKTIONEN nach unten und markieren Sie die Option KORRELATION.**

4. **Klicken Sie auf OK, um das Dialogfeld KORRELATION zu öffnen.**

   Dieses Dialogfeld ist in Abbildung 15.5 dargestellt.

5. **Geben Sie im Feld EINGABEBEREICH den Zellbereich ein, in dem sich die Daten befinden.**

   Ich habe $C$2:$E$22 eingegeben. Beachten Sie die Dollarzeichen ($) für den absoluten Bezug.

6. **Legen Sie im Bereich GEORDNET NACH fest, ob die Daten in Spalten oder in Zeilen angeordnet sind.**

   Ich habe hier das Optionsfeld SPALTEN ausgewählt.

7. **Wenn der Zellbereich Spaltenüberschriften enthält, aktivieren Sie das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE.**

   Ich habe die Überschriften in den Zellbereich mit aufgenommen. Daher habe ich auch dieses Kontrollkästchen aktiviert.

8. **Legen Sie im Bereich AUSGABE fest, wo Excel das Ergebnis der Analyse ausgeben soll.**

   Ich habe das Optionsfeld NEUES TABELLENBLATT ausgewählt, so dass das Ergebnis auf einem neuen Arbeitsblatt angezeigt wird.

9. **Klicken Sie auf OK.**

   Da ich das Optionsfeld NEUES TABELLENBLATT ausgewählt habe, wird ein neues Arbeitsblatt mit dem Ergebnis erstellt.

Ergebnis in Tabellen

In Abbildung 15.6 ist das von diesem Tool berechnete Ergebnis in einer Tabelle mit erweiterten Spalten dargestellt. Bei der Tabelle handelt es sich um eine *Korrelationsmatrix*.

	A	B	C	D	E
1		*SAT*	*HS-Durchschnitt*	*GPA*	
2	SAT	1			
3	HS-Durchschnitt	0,552527329	1		
4	GPA	0,81662505	0,714353653	1	
5					

Abbildung 15.6: Das Ergebnis des Datenanalysetools KORRELATION in einer Tabelle

Die einzelnen Zellen in der Matrix repräsentieren die Korrelation der Variablen in der Zeile mit der Variablen in der Spalte. In Zelle B3 ist beispielsweise die Korrelation von SAT mit dem Notendurchschnitt an der High School angegeben. In den Zellen der Hauptdiagonale steht jeweils eine 1. Das kommt daher, weil jede Zelle der Hauptdiagonalen die Korrelation einer Variablen mit sich selbst repräsentiert.

Es muss nur die halbe Matrix ausgefüllt werden. Die Zellen oberhalb der Hauptdiagonalen würden dieselben Werte wie die Zellen unterhalb enthalten.

Was genau besagt diese Tabelle? Lesen Sie weiter …

Multiple Korrelation

Die Korrelationskoeffizienten in dieser Matrix generieren zusammen einen *multiplen Korrelationskoeffizienten*. Hierbei handelt es sich um die Zusammenfassung des Zusammenhangs zwischen der abhängigen Variablen (GPA in diesem Beispiel) und den beiden unabhängigen Variablen (SAT und Notendurchschnitt an der High School).

Damit Sie sehen, wie diese Korrelationskoeffizienten zusammengefasst werden, kürze ich GPA mit *G*, SAT mit *S* und High-School-Notendurchschnitt mit *H* ab. Somit ist r_{GS} der Korrelationskoeffizient für GPA und SAT, r_{GH} ist der für GPA und High-School-Notendurchschnitt und r_{SH} ist der für SAT und High-School-Notendurchschnitt.

Und so sieht die Formel aus, die das alles zusammenfügt:

$$R_{G.SH} = \sqrt{\frac{r_{GS}^2 + r_{GH}^2 - 2 r_{GS} r_{GH} r_{SH}}{1 - r_{GS}^2}}$$

Der Großbuchstabe *R* auf der linken Seite gibt an, dass es sich hierbei, im Gegensatz zu dem Kleinbuchstaben *r* für eine Korrelation zwischen zwei Variablen, um einen multiplen Korrelationskoeffizienten handelt. Das tiefgestellte *G.SH* steht für die

multiple Korrelation zwischen GPA und der Kombination aus SAT und High-School-Notendurchschnitt.

Dies ist die Berechnung, mit der im Abschnitt REGRESSIONS-STATISTIK im Ergebnis des Datenanalysetools Regression das Bestimmtheitsmaß berechnet wird (siehe Kapitel 14).

Für dieses Beispiel gilt:

$$R_{G.SH} = \sqrt{\frac{(0{,}816625)^2 + (0{,}714354)^2 - 2(0{,}816625)(0{,}714354)(0{,}552527)}{1-(0{,}816625)^2}} = 0{,}875529$$

Da ich dieselben Daten zum Erklären der multiplen Regression in Kapitel 14 verwendet habe, finden Sie diesen Wert (mit einigen weiteren Dezimalstellen) auch in Abbildung 14.25 in Zelle B4.

Wenn Sie diese Zahl quadrieren, erhalten Sie das *multiple Bestimmtheitsmaß*. In Kapitel 14 habe ich das Bestimmtheitsmaß erwähnt, und genau das ist es. Es ist ein weiteres Element in der vom Datenanalysetool Regression berechneten Regressionsstatistik. Sie finden es auch im Ergebnis der Funktion RGP, wobei es dort nicht gekennzeichnet ist.

In diesem Beispiel lautet das Ergebnis wie folgt:

$$R^2_{G.SH} = (0{,}875529)^2 = 0{,}766552$$

Diese Zahl finden Sie in Abbildung 14.24 in Zelle H5 (Ergebnis der Funktion `RGP`). Zudem finden Sie sie in Abbildung 14.25 in Zelle B5 (Bericht des Datenanalysetools Regression).

Das Bestimmtheitsmaß anpassen

Hier finden Sie weitere Informationen zum Bestimmtheitsmaß im Hinblick auf Excel. Neben dem Bestimmtheitsmaß berechnet das Datenanalysetool Regression auch das angepasste, oder wie es bei Excel heißt: das *adjustierte Bestimmtheitsmaß*. In Abbildung 14.25 ist dieser Wert in Zelle B6 angegeben. Warum ist es notwendig, das Bestimmtheitsmaß anzupassen?

Bei der multiplen Regression wird die Regressionsgleichung durch die Addition unabhängiger Variablen (wie High-School-Notendurchschnitt) manchmal nicht ganz so genau. Dies wird beim multiplen Bestimmtheitsmaß nicht berücksichtigt. Der Nenner ist SS_{Gesamt} (für die abhängige Variable) und ändert sich nie. Der Zähler kann nur größer werden oder gleich bleiben. Somit spiegelt sich eine Abnahme der Genauigkeit nicht in einem niedrigeren Wert für das Bestimmtheitsmaß wider.

Dieses Manko wird durch die Berücksichtigung der Freiheitsgrade beseitigt. Jedes Mal, wenn eine unabhängige Variable hinzugefügt wird, werden die Freiheitsgrade geändert. Und das macht den ganzen Unterschied aus. Nur der Vollständigkeit halber hier die Anpassung:

$$Angepasstes\ R^2 = 1 - \left(1 - R^2\right)\left[\frac{(N-1)}{(N-k-1)}\right]$$

R^2 steht für Bestimmtheitsmaß und das *k* im Nenner für die Anzahl der unabhängigen Variablen.

Partielle Korrelation

GPA und SAT sind (in diesem Beispiel) mit dem High-School-Notendurchschnitt verknüpft. Aufgrund der Verknüpfung der beiden Werte mit dem High-School-Notendurchschnitt bleibt die tatsächliche Korrelation zwischen den beiden Werten möglicherweise verborgen.

Wie würde die Korrelation zwischen diesen beiden Werten ohne diese Verknüpfung aussehen? Oder anders ausgedrückt: Wie würde der Wert für die Korrelation zwischen GPA und SAT bei konstantem High-School-Notendurchschnitt lauten?

Eine Möglichkeit, den High-School-Notendurchschnitt konstant zu halten, besteht darin, den Wert für die GPA-SAT-Korrelation für eine Stichprobe von Studenten zu ermitteln, die denselben High-School-Notendurchschnitt, beispielsweise 87, haben. In einer Stichprobe wie dieser beträgt der Wert für die Korrelation der einzelnen Variablen mit dem High-School-Notendurchschnitt null. So etwas kommt in der Realität jedoch nicht vor.

Eine andere Möglichkeit besteht darin, den Wert der *partiellen Korrelation* zwischen GPA und SAT zu ermitteln. Hierbei handelt es

sich um ein statistisches Verfahren, bei dem die Verknüpfung der Variablen mit der High-School-Durchschnittsnote aufgehoben wird. Hierfür werden die Korrelationskoeffizienten in der Korrelationsmatrix verwendet:

$$r_{GS.H} = \frac{r_{GS} - r_{GH} r_{SH}}{\sqrt{1 - r_{GH}^2} \sqrt{1 - r_{SH}^2}}$$

Auch hier steht *G* für GPA, *S* für SAT und *H* für High-School-Notendurchschnitt. Das tiefgestellte *GS.H* steht für die Korrelation zwischen GPA und SAT ohne Verknüpfung zum High-School-Notendurchschnitt.

Für dieses Beispiel gilt:

$$r_{GS.H} = \frac{0{,}816625 - (0{,}714353)\,(0{,}552527)}{\sqrt{1 - (0{,}714353)^2} \sqrt{1 - (0{,}552527)^2}} = 0{,}547005$$

Semipartielle Korrelation

Darüber hinaus ist es auch möglich, die Korrelation zwischen High-School-Notendurchschnitt und SAT-Wert ohne Aufhebung der Korrelation zwischen High-School-Notendurchschnitt und GPA aufzuheben. Hierbei spricht man von einer *semipartiellen Korrelation*. Für die Formel werden ebenfalls die Korrelationskoeffizienten aus der Korrelationsmatrix verwendet:

$$r_{G(S.H)} = \frac{r_{GS} - r_{GH} r_{SH}}{\sqrt{1 - r_{SH}^2}}$$

Das tiefgestellte *G(S.H)* steht für die Korrelation zwischen GPA und SAT, wobei die Verknüpfung zum High-School-Notendurchschnitt nur für den SAT-Wert aufgehoben ist.

Setzen wir die Zahlen aus unserem Beispiel in die Formel ein:

$$r_{G(S.H)} = \frac{0{,}816625 - (0{,}714353)\,(0{,}552527)}{\sqrt{1-(0{,}552527)^2}} = 0{,}315714$$

In einigen Statistikbüchern wird die semipartielle Korrelation auch als *Teilkorrelation* bezeichnet.

Datenanalysetool: Kovarianz

Das Datenanalysetool Kovarianz wird genauso wie das Datenanalysetool Korrelation verwendet. Keine Angst: Ich werde Ihnen nun nicht die ganzen Schritte noch einmal erläutern. Stattdessen werde ich Ihnen das Ergebnis in Tabellenform in Abbildung 15.7 zeigen: Die Daten stammen aus Abbildung 15.5.

	A	B	C	D	E
1		*SAT*	*HS-Durchschnitt*	*GPA*	
2	SAT	24862,75			
3	HS-Durchschnitt	512,375	34,5875		
4	GPA	85,1675	2,77875	0,437475	
5					

Abbildung 15.7: Das Ergebnis des Datenanalysetools KOVARIANZ für den SAT-Wert, den Notendurchschnitt an der High School und dem GPA in Tabellenform

Bei der Tabelle handelt es sich um eine *Kovarianzmatrix*. Die einzelnen Zellen in der Matrix stehen für die Kovarianz der Variablen in der Zeile mit der Variablen in der Spalte (und zwar so, wie `KOVARIANZ.P` dieses Maß berechnen würde, also indem in der Formel *N* verwendet wird). In Zelle C4 ist die Kovarianz von GPA mit dem Notendurchschnitt an der High School angegeben. Die Hauptdiagonale in dieser Matrix stellt die Varianz der einzelnen Variablen dar (was der Kovarianz einer Variablen mit sich selbst entspricht). In diesem Beispiel ist die Varianz das, was Sie berechnen, wenn Sie die Funktion `VARIANZEN.P` verwenden.

Auch hier muss nur die halbe Matrix ausgefüllt werden. Zellen oberhalb der Hauptdiagonalen würden dieselben Werte wie die unter ihr enthalten.

Wie bei `KOVAR` sehe ich jedoch keinen Anlass, wozu Sie dieses Tool verwenden sollten. Ich habe es hier nur der Vollständigkeit halber erläutert.

Hypothesen über Korrelationen testen

Excel enthält keine Tabellenblattfunktion zum Testen von Hypothesen über *r*. Wie bereits erwähnt, führen Sie einen *t*-Test durch, dessen Formel wie folgt lautet:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Mit dieser Formel habe ich *t* berechnet, wobei sich in Zelle H12 die Zahl 0,817 befindet:

```
=H12*WURZEL(20-2)/WURZEL(1-H12^2)
```

Anschließend habe ich das Ergebnis (6,011 und ein paar weitere Dezimalstellen) als Eingabe für `T.VERT.RE` (und 18 für die Freiheitsgrade sowie 1 für die Anzahl der Seiten) verwendet, um festzustellen, dass die einseitige Wahrscheinlichkeit des Ergebnisses weit unter 0,05 liegt.

Tabellenblattfunktionen: FISHER, FISHERINV

Excel übernimmt die relativ komplexen Transformationen, mit denen Sie Hypothesen über die Differenz zwischen zwei Korrelationskoeffizienten durchführen können. Mit `FISHER` wird *r* in *z* transformiert, `FISHERINV` bewirkt das Umgekehrte. Wenn Sie sich

noch erinnern: Die transformierten Werte werden in der folgenden Formel verwendet:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

wobei der Nenner wie folgt lautet:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

In dem weiter oben beschriebenen Beispiel (Vergleich: Sahutsket – Farshimmelt) mit den Korrelationskoeffizienten 0,817 und 0,752 habe ich einen zweiseitigen Test ausgeführt. Zunächst einmal werden die einzelnen Korrelationskoeffizienten transformiert. Ich werde 0,817 mit der Tabellenblattfunktion `FISHER` transformieren:

1. **Markieren Sie eine Zelle für das Ergebnis von `FISHER`.**

   Ich habe B3 für den transformierten Wert gewählt.

2. **Klicken Sie im Menü mit den statistischen Funktionen auf FISHER, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

   Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `FISHER` ist in Abbildung 15.8 dargestellt.

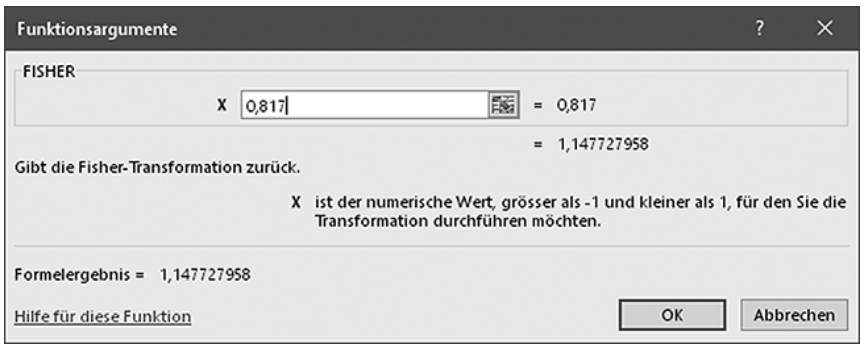


Abbildung 15.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `FISHER`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE den Wert des Arguments ein.**

   Ich habe in das Feld x den Korrelationskoeffizienten 0,817 eingegeben. Das Ergebnis 1,147728 wird im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, damit die Lösung in die markierte Zelle eingefügt wird.**

Ich habe B4 zum Speichern der Transformation von 0,752 ausgewählt. Als Nächstes habe ich mit der folgenden Formel Z berechnet:

```
=(B3-B4)/WURZEL((1/(20-3))+1/(24-3))
```

Schließlich habe ich mit der Funktion `NORM.S.INV` den kritischen Wert von *z* zum Verwerfen von H_0 bei einem zweiseitigen Alpha von 0,05 ermittelt. Da das Ergebnis der Formel (0,521633) kleiner als dieser kritische Wert (1,96) ist, wird H_0 nicht verworfen.

Kapitel 16

Es wird höchste Zeit

In diesem Kapitel

- Zeitreihen verstehen
- Gleitende Durchschnitte
- Die Dinge glätten
- Prognosen mit nur einem Klick

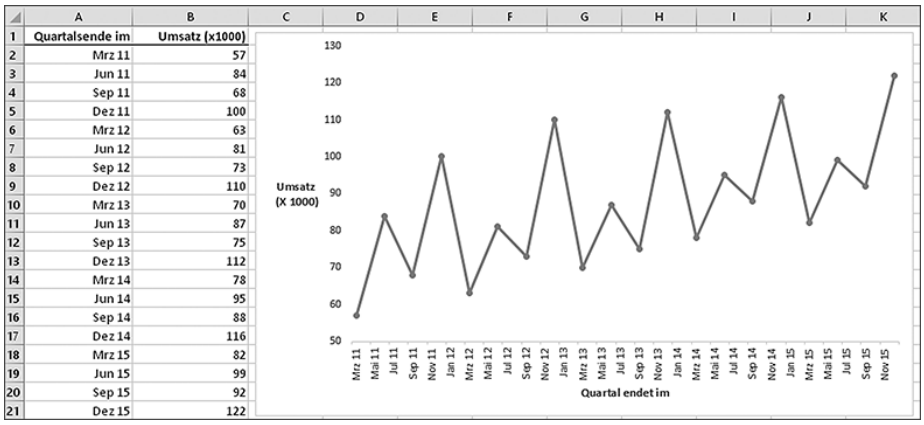
In vielen Bereichen (Wissenschaft, Medizin, Unternehmen), ist es oft erforderlich, Daten über gewissen Zeiträume zu sammeln. Bei dieser Art von Daten handelt es sich um *Zeitreihen*. Dieses Kapitel befasst sich mit Zeitriehen und damit, wie Sie mit Excel diese Zeitreihen verstehen und wie Sie sie verwenden können, um Prognosen zu erstellen.

Eine Zeitreihe und ihre Komponenten

Manager müssen ihre Entscheidungen häufig auf Zeitreihen – wie Umsatzzahlen – basieren. Die Zahlen in einer Zeitreihe zeigen in der Regel eine ständige Auf-und-ab-Bewegung.

Hier ein Beispiel. Die (fiktive) Firma FarDrate Timepiece Corporation verkauft das wunderbare Uhrenmodell MeesKyte. Die Umsätze werden am Ende jedes Quartals erfasst. Abbildung 16.1 zeigt ein Arbeitsblatt und ein Diagramm mit diesen Umsatzzahlen für den Zeitraum von 2011 bis 2015.

Werfen Sie einen Blick auf das Diagramm. Es hat den Anschein, als ob die Spitzen und Täler ziemlich regelmäßig auftreten würden. Im letzten Quartal jedes Jahr lassen sich höhere Umsätze erkennen – möglicherweise werden die Uhren als Weihnachtsgeschenk gekauft. Auch in der Jahresmitte lassen sich Umsatzspitzen erkennen, die jedoch nicht so stark ausfallen – vielleicht kaufen Eltern ihren Kindern die MeesKyte-Uhr für das bestandene Abitur. Dieses Muster wiederholt sich in jedem Jahr; daher wird es die *Saisonkomponente* der Zeitreihe genannt.



	A	B
1	Quartalsende im	Umsatz (x1000)
2	Mrz 11	57
3	Jun 11	84
4	Sep 11	68
5	Dez 11	100
6	Mrz 12	63
7	Jun 12	81
8	Sep 12	73
9	Dez 12	110
10	Mrz 13	70
11	Jun 13	87
12	Sep 13	75
13	Dez 13	112
14	Mrz 14	78
15	Jun 14	95
16	Sep 14	88
17	Dez 14	116
18	Mrz 15	82
19	Jun 15	99
20	Sep 15	92
21	Dez 15	122

Abbildung 16.1: Die Quartalsumsätze der MeesKyte-Uhr

Außerdem lässt sich ein generelles Umsatzwachstum erkennen. Dies mag an einer guten Konjunktur liegen oder an einer größeren Bekanntheit der Marke (oder an beidem). Dies wird die *Trendkomponente* der Zeitreihe genannt. In diesem Beispiel ist der Trend linear, dies ist aber nicht immer der Fall.

Außerdem sehen wir, dass die Spitzen und Täler um den generellen Aufwärtstrend herum variieren. Dies wird die *zyklische Komponente* der Zeitreihe genannt.

Schließlich können auch sporadische, sich nicht wiederholende Einflüsse auf eine Zeitreihe auswirken. Dies wird die *unregelmäßige Komponente* einer Zeitreihe genannt.

Sanftes Gleiten

Durch einen Mittelwert, der sämtliche Höhen und Tiefen in sich vereinigt, könnte das tatsächliche Bild des Gesamttrends verschleiert werden.

Eine Möglichkeit, die Höhen und Tiefen zu glätten und dennoch das Gesamtbild zu betrachten, ist die Berechnung eines *gleitenden Durchschnitts*. Dies ist ein Durchschnitt, der anhand der jüngsten Punktwerte in der Zeitreihe berechnet wird. Er gleitet, weil er im Verlauf der Zeitreihe immer weiter berechnet wird. Wenn Sie am Ende einen Wert hinzufügen, wird ein Wert am Anfang gelöscht.

Wir haben also nun die Umsätze der MeesKyte-Uhr und beschließen, für die jeweils letzten fünf Quartale einen gleitenden Durchschnitt zu berechnen. Beginnen Sie mit dem Mittelwert des 1. bis 5. Quartals (Quartale, die im März 2011 bis März 2012 enden) der 20 Quartale. Dann nehmen Sie die Umsätze des 2. bis 6. Quartals, dann die Quartale 3 bis 7 und so weiter, bis Sie schließlich den Mittelwert der letzten fünf Quartale der Zeitreihe berechnet haben.

Beim gleitenden Durchschnitt handelt es sich um eine Prognose. Es handelt sich um die beste Schätzung, die darauf basiert, den Mittelwert der Umsätze aus den letzten fünf Quartalen zu berechnen.

Excel stellt zwei Möglichkeiten zur Verfügung, um den gleitenden Durchschnitt zu berechnen. Eine Variante fällt in den Bereich »schnelle Lösung« (hierbei wird eine Trennlinie in das Diagramm eingefügt) und die zweite verwendet eines der Werkzeuge aus den Datenanalysetools von Excel.

Den Trend als Linie darstellen

Und so verwenden Sie die Trendlinienvariante:

1. **Geben Sie die Daten in ein Arbeitsblatt ein.**

Ich habe die Daten in Spalten A und B sowie Zeilen 1 bis 21 eingegeben. In Zeile 1 befinden sich Spaltenüberschriften.

2. **Markieren Sie die Daten und fügen Sie ein Liniendiagramm ein.**

   Wählen Sie auf der Registerkarte EINFÜGEN in der Gruppe DIAGRAMME den Befehl LINIE MIT DATENPUNKTEN.

3. **Klicken Sie das Diagramm an, damit neben dem Diagramm die Schaltfläche DIAGRAMMELEMENTE eingeblendet wird (die mit dem Pluszeichen).**

4. **Klicken Sie die Schaltfläche DIAGRAMMELEMENTE an.**

5. **Wählen Sie im Popupmenü den Befehl TRENDLINIE | WEITERE OPTIONEN.**

6. **Wählen Sie im Aufgabenbereich TRENDLINIE FORMATIEREN das Optionsfeld GLEITENDER DURCHSCHNITT aus und ändern Sie den ZEITRAUM.**

   Für dieses Beispiel habe ich den Zeitraum auf **5** festgelegt.

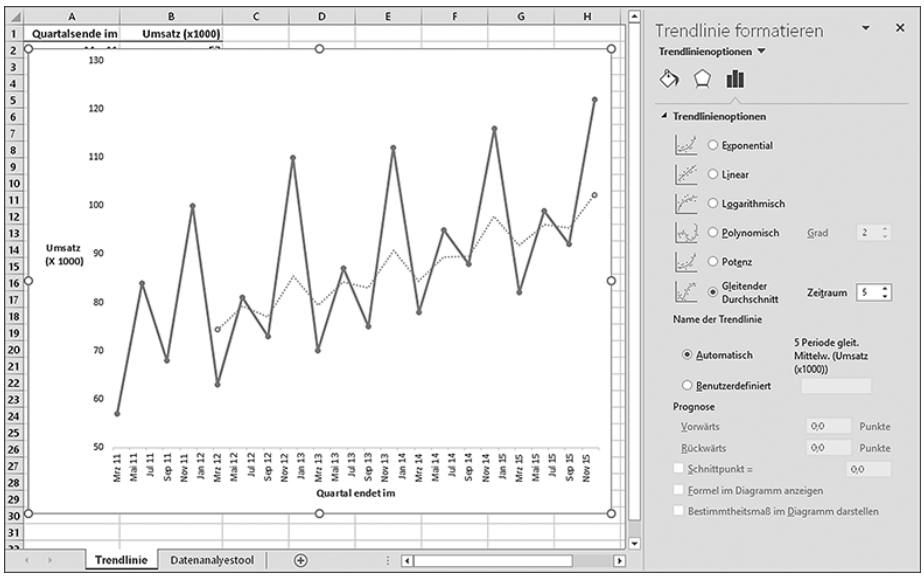


Abbildung 16.2: Die Trendlinie formatieren, um einen gleitenden Durchschnitt zu erstellen

Abbildung 16.2 zeigt den Aufgabenbereich TRENDLINIE FORMATIEREN sowie das Diagramm mit der Linie, die den gleitenden Durchschnitt darstellt. Das Gute an dieser Vorgehensweise ist, dass Sie einfach mit verschiedenen Zeiträumen, auch Intervalle genannt, experimentieren können und im Diagramm sofort sehen, wie sich diese Änderung auswirkt. Leider werden bei dieser Technik die Werte der gleitenden Mittelwerte nicht angezeigt. Um diese Werte zu berechnen, benötigen Sie das …

Datenanalysewerkzeug: Gleitender Durchschnitt

Das Excel-Analysewerkzeug Gleitender Durchschnitt kann den gleitenden Durchschnitt ebenfalls im Diagramm darstellen und außerdem die numerischen Werte anzeigen. Anders als bei der Verwendung der Trendlinie können Sie jedoch nicht so einfach mit verschiedenen Intervallen experimentieren.

Die einzelnen Schritte:

1. **Nachdem Sie die Daten in ein Arbeitsblatt eingegeben haben, klicken Sie im Menüband auf DATEN | DATENANALYSE. Wählen Sie dann im Dialogfeld Analyse-Funktionen GLEITENDER DURCHSCHNITT aus.**

   Hierdurch wird das Dialogfeld GLEITENDER DURCHSCHNITT geöffnet.

2. **Geben Sie in das Feld EINGABEBEREICH den Zellbereich mit den Daten ein.**

   In diesem Beispiel habe ich den Zellbereich B1:B21 eingegeben. Die Spalte mit den Datumswerten verwende ich diesmal nicht.

3. **Nehmen Sie weitere Eingaben im Dialogfeld vor.**

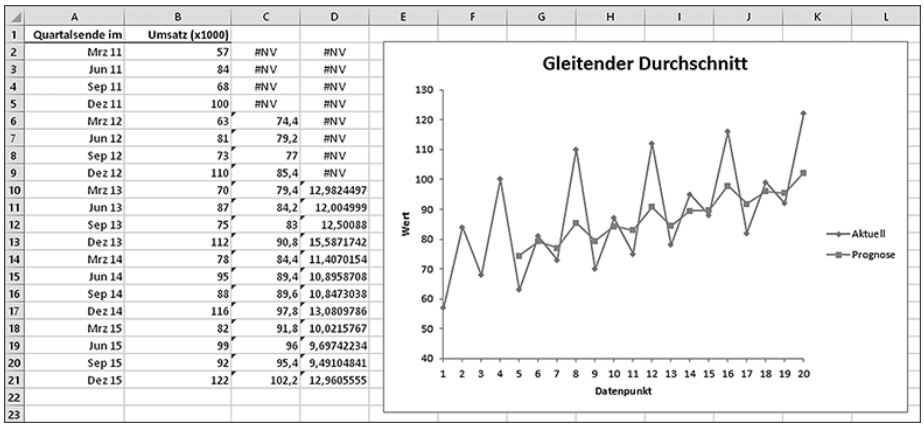
   Ich habe das Kontrollkästchen BESCHRIFTUNGEN IN ERSTER ZEILE eingeschaltet und in das Feld INTERVALL den Wert 5 eingegeben. (Das Feld INTERVALL entspricht dem Feld ZEITRAUM aus dem

Beispiel mit der Trendlinie.) Anschließend habe ich den Zellbereich C2:C22 als AUSGABEBEREICH festgelegt und die Kontrollkästchen DIAGRAMMDARSTELLUNG sowie STANDARDFEHLER aktiviert. Abbildung 16.3 zeigt das fertig ausgefüllte Dialogfeld.

Abbildung 16.3: Das fertig ausgefüllte Dialogfeld GLEITENDER DURCHSCHNITT

4. **Klicken Sie auf OK.**

   Hierdurch werden die Werte des gleitenden Durchschnitts in Spalte C und die Standardfehler in Spalte D eingetragen. Anschließend wird ein Diagramm mit den Daten und den gleitenden Durchschnitten erstellt. Abbildung 16.4 zeigt das fertige Ergebnis.



	A	B	C	D
1	Quartalsende im	Umsatz (x1000)		
2	Mrz 11	57	#NV	#NV
3	Jun 11	84	#NV	#NV
4	Sep 11	68	#NV	#NV
5	Dez 11	100	#NV	#NV
6	Mrz 12	63	74,4	#NV
7	Jun 12	81	79,2	#NV
8	Sep 12	73	77	#NV
9	Dez 12	110	85,4	#NV
10	Mrz 13	70	79,4	12,9824497
11	Jun 13	87	84,2	12,004999
12	Sep 13	75	83	12,50088
13	Dez 13	112	90,8	15,5871742
14	Mrz 14	78	84,4	11,4070154
15	Jun 14	95	89,4	10,8958708
16	Sep 14	88	89,6	10,8473038
17	Dez 14	116	97,8	13,0809786
18	Mrz 15	82	91,8	10,0215767
19	Jun 15	99	96	9,69742234
20	Sep 15	92	95,4	9,49104841
21	Dez 15	122	102,2	12,9605555
22				
23				

Abbildung 16.4: Die Ergebnisse: Gleitende Durchschnitte, Standardfehler und ein Diagramm

Ignorieren Sie die nicht so hübsch aussehenden #NV-Fehlermeldungen. Jede Zahl in Spalte C ist ein gleitender Durchschnitt – bei dem es sich, wie bereits erwähnt, um eine Umsatzprognose auf der Basis der vorhergehenden fünf Quartale handelt.

Jede der Zahlen in Spalte D ist ein Standardfehler. Ein Standardfehler ist in diesem Zusammenhang die Quadratwurzel des Durchschnitts der Quadratdifferenz zwischen dem Umsatz und der Prognose der jeweils vorhergehenden fünf Quartale. Der erste Standardfehler in Zelle D10 ist demnach:

$$\sqrt{\frac{(63-74,4)^2+(81-79,2)^2+(73-77)^2+(110-85,4)^2+(70-79,4)^2}{5}}=12,98245$$

Im Diagramm (das ich im Vergleich zu seinem ursprünglichen Erscheinungsbild etwas gestreckt habe) steht der gleitende Durchschnitt in der Wertereihe, die mit »Prognose« beschriftet ist. Manchmal passt die Prognose zu den tatsächlichen Daten, manchmal nicht. In der Abbildung wird deutlich, dass der gleitende Durchschnitt die Höhen und Tiefen der Kursdaten glättet.

Wie viele Werte sollten Sie im Allgemeinen in die Berechnung einbeziehen? Nun, das ist Ihre Entscheidung. Wenn Sie zu viele Werte aufnehmen, riskieren Sie, dass Ihr Ergebnis durch überflüssige Daten beeinflusst wird. Nehmen Sie hingegen zu wenige Werte auf, besteht die Gefahr, dass Sie etwas Wichtiges übersehen.

Das Analysewerkzeug Gleitender Durchschnitt beschriftet die x-Achse nicht mit den Datumswerten. Sie können diese jedoch ergänzen, nachdem das Diagramm erstellt wurde. Klicken Sie das Diagramm mit der rechten Maustaste an, klicken Sie im Kontextmenü auf DATEN AUSWÄHLEN und bearbeiten Sie die HORIZONTALE ACHSENBESCHRIFTUNGEN (RUBRIK).

Exponentielles Glätten

Das *exponentielle Glätten* ähnelt dem gleitenden Durchschnitt. Es ist ein Verfahren zum Erstellen von Prognosen auf der Grundlage vorheriger Daten. Im Gegensatz zum gleitenden Durchschnitt, bei dem nur eine Sequenz tatsächlicher Werte zum Einsatz kommt, wird beim exponentiellen Glätten eine vorherige Vorhersage mit einbezogen.

Beim exponentiellen Glätten wird ein *Glättungsparameter*, eine Zahl zwischen null und eins, verwendet. Wenn α für den Glättungsparameter steht, lautet die Formel

$$y'_t = (1-\alpha)\,y'_{t-1} + \alpha\,y'_{t-1}$$

Wenn wir wieder das Beispiel mit den Umsatzzahlen heranziehen, steht y'_t für den vorhergesagten Umsatz zum Zeitpunkt *t*. Wenn *t* heute ist, ist *t*−1 gestern. y_{t-1} ist demnach der tatsächliche Umsatz im vorherigen Quartal und y'_{t-1} der für das vorherige Quartal vorhergesagte Umsatz. Die Sequenz der Vorhersagen beginnt mit dem ersten vorhergesagten Wert als Beobachtungswert des vorhergehenden Tages.

Ein größerer Wert für den Glättungsparameter verleiht der Vorhersage für das vorherige Quartal mehr Gewicht. Ein kleinerer Wert hingegen verleiht dem tatsächlichen Umsatz des vorherigen Quartals ein größeres Gewicht. Bei einem Glättungsparameter von 0,5 werden beide gleich stark gewichtet.

In Abbildung 16.5 ist das Dialogfeld für die Analyse-Funktion EXPONENTIELLES GLÄTTEN dargestellt. Abgesehen vom Feld GLÄTTUNGSPARAMETER ähnelt es dem Dialogfeld für den gleitenden Durchschnitt.

Abbildung 16.5: Das Dialogfeld der Analyse-Funktion EXPONENTIELLES GLÄTTEN

Ich habe die Funktion des exponentiellen Glättens insgesamt drei Mal auf die Daten des vorherigen Beispiels angewandt und dabei als Glättungsparameter die Werte 0,1, 0,5 und 0,9 eingesetzt. In Abbildung 16.6 sind die drei Diagramme darstellt.

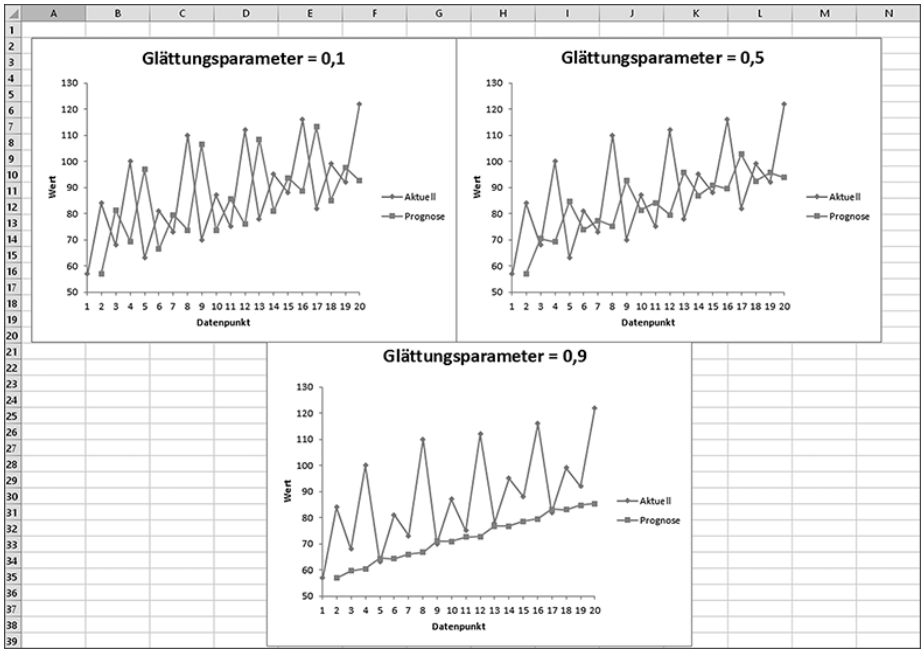


Abbildung 16.6: Exponentielles Glätten mit drei Glättungsparametern

Die glatteste Sequenz von Prognosen ergibt sich beim höchsten Glättungsparameter von 0,9. Der niedrigste Glättungsparameter, 0,1, führt zur Vorhersage der am stärksten ausgeprägten Höhen und

Tiefen. Wie sollten Sie den Glättungsparameter einstellen? Wie schon beim Intervall für den gleitenden Durchschnitt ist das allein Ihre Entscheidung. Ihre Erfahrung und der jeweils spezifische Anwendungsbereich sind dabei die entscheidenden Faktoren.

Möglicherweise ist Ihnen in Abbildung 16.6 aufgefallen, dass die Prognosen, die das exponentielle Glätten liefert, in diesem Beispiel nicht allzu genau sind.

Was tun? Lesen Sie weiter…

Prognosen mit einem Klick

Ein neues Feature in Excel 2016 (nur in der Windows-Version) erlaubt es Ihnen, eine Zeitreihe zu markieren, und dann mit einem Mausklick (genaugenommen ein paar Mausklicks) eine Reihe von extrapolierten Prognosen mit den Konfidenzintervallen jeder Prognose zu erhalten.

Diese neuen Prognosefeatures verwenden die Gruppe der neuen `PROGNOSE.ETS`-Funktionen (die ebenfalls lediglich in der Windows-Version zur Verfügung stehen):

- `PROGNOSE.ETS` berechnet einen zukünftigen Wert mithilfe eines Algorithmus für die dreifache exponentielle Glättung.
- `PROGNOSE.ETS.KONFINT` gibt ein Konfidenzintervall für einen prognostizierten Wert zurück.
- `PROGNOSE.ETS.STAT` gibt einen statistischen Wert zurück, der mit der ETS-Prognose verbunden ist.
- `PROGNOSE.ETS.SAISONALITÄT` gibt die Länge des repetitiven Musters zurück, das Excel in den Daten erkannt hat.

Ich könnte Ihnen viel über diese Funktionen erzählen, aber ich denke nicht, dass Sie diese jemals individuell verwenden werden.

Dreifacher Spaß – exponentiell

Die Grundlage von Excels neuem Prognosetool ist eine viel komplexere Form des exponentiellen Glättens, als die, die Sie bisher kennengelernt haben. Das Konzept, das ich im vorherigen Abschnitt beschrieben habe, heißt *einfaches exponentielles Glätten*.

Wenn Sie eine Trendkomponente haben (wie das generelle Umsatzwachstum in dem Beispiel), verliert das einfache exponentielle Glätten an Genauigkeit. Möglicherweise konnten Sie dies bereits in Abbildung 16.6 erkennen.

Wenn sich in den Daten ein Trend befinden, müssen Sie ein klein wenig zusätzlich addieren (die Steigung):

$$y'_t = (1-\alpha)\,y_{t-1} + \alpha\left(y'_{t-1} + b_{t-1}\right)$$

und:

$$b_t = \gamma\left(y'_t - y_{t-1}\right) + (1-\gamma)\,b_{t-1}$$

Der Buchstabe vor der ersten runden Klammern in der letzten Gleichung, der aussieht wie ein y, ist der griechische Buchstabe Gamma. Diese Gleichungen erweitern die Berechnung der Prognose um den Trend. Diese Art des Glättens wird *doppeltes exponentielles Glätten* genannt.

Wenn in den Daten sowohl ein Trend als auch eine Saisonalität erkennbar sind (wie beim FarDrate-Beispiel der Fall), dann ist das *dreifache exponentielle Glätten* die Rettung. (Die Abkürzung ETS in den Excel-Funktionsnamen steht für *Error, Trend, Seasonal*). Die drei Formeln für das dreifache exponentielle Glätten wurden übrigens von den Statistikern Charles Holt und Peter Winters entwickelt. Daher wird diese Methode auch Holt-Winters-Verfahren genannt.

Die erste Formel ist eine Überarbeitung der Formel für das einfache exponentielle Glätten:

$$y'_t = (1-\alpha)\left(\frac{y_{t-1}}{l_{t-1}}\right) + \alpha\left(y'_{t-1} + b_{t-1}\right)$$

Die zweite ist die Formel, die Sie oben bereits gesehen haben:

$$b_t = \gamma\left(y'_t - y_{t-1}\right) + (1-\gamma)\,b_{t-1}$$

Die dritte Formel ist:

$$I_t = \beta\left(\frac{y_t}{y'_t}\right) + (1-\beta)\, I_{t-L}$$

Die erste Formel ist für das allgemeine Glätten zuständig, die zweite glättet den Trend und die dritte glättet die saisonale Komponente.

I ist der saisonale Index und *L* ist die Anzahl der Intervalle je Saison.

Jeder der drei Koeffizienten (α, β und γ) ist eine Zahl zwischen 0 und 1.

Die Mathematik, die benötigt wird, um diese Koeffizienten auf die Daten anzupassen, ist sehr, sehr komplex. Jedoch können Sie dieses Feature mit dem Excel-Prognosearbeitsblatt (das ist das Ein-Klick-Prognosewerkzeug) sehr einfach einsetzen.

Lassen Sie uns eine ETS-Prognose erstellen. Hier sind die einzelnen Schritte:

1. **Geben Sie Daten ein, wobei sich die Datumswerte in einer Spalte befinden sollten.**

   Wie in unserem bisherigen Beispiel befinden sich die Daten in den Spalten A und B.

2. **Markieren Sie die Daten.**

3. **Klicken Sie auf der Registerkarte DATEN in der Gruppe PROGNOSE auf PROGNOSEBLATT.**

   Excel zeigt das Dialogfeld PROGNOSEARBEITSBLATT ERSTELLEN an, das Abbildung 16.7 zeigt. Wie Sie sehen können, hat Excel schon hart gearbeitet, und für Sie ein paar Dinge berechnet. Die blaue Linie stellt die Daten dar, die dicke, orangefarbene Linie ist die Prognose. Die beiden dünnen orangefarbenen Linien symbolisieren die untere beziehungsweise obere Konfidenzgrenze bei einem Konfidenzintervall von 95 %.

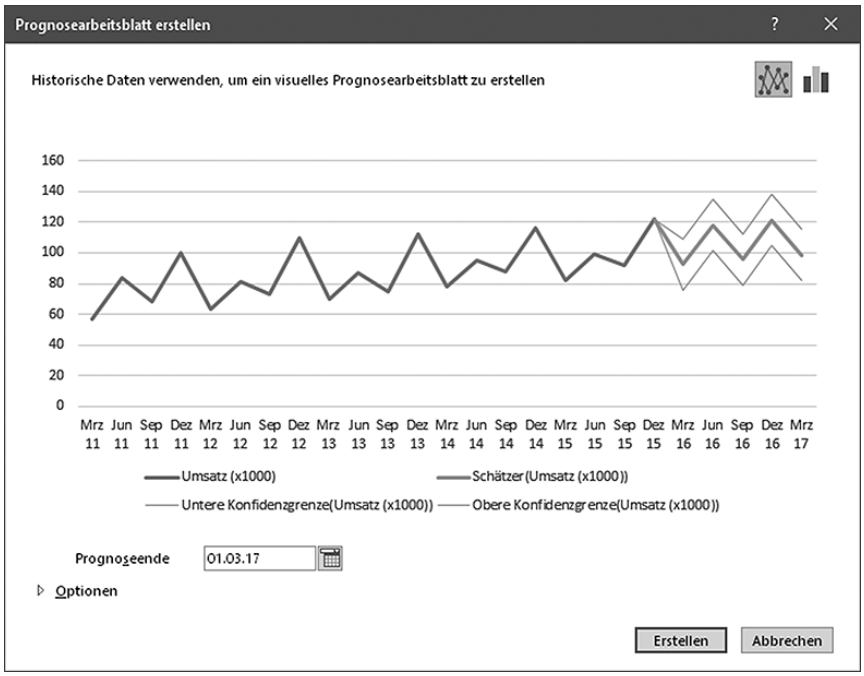


Abbildung 16.7: Das Dialogfeld PROGNOSEARBEITSBLATT ERSTELLEN

4. **Klicken Sie in der unteren linken Ecke des Dialogfeldes PROGNOSEARBEITSBLATT ERSTELLEN auf den Pfeil neben OPTIONEN.**

   Hierdurch wird das Dialogfeld vergrößert (siehe Abbildung 16.8).

Abbildung 16.8: Das erweiterte Dialogfeld PROGNOSEARBEITSBLATT ERSTELLEN

5. **Nehmen Sie im erweiterten Dialogfeld die gewünschten Änderungen vor.**

   Um einen Eindruck von der Prognosegenauigkeit zu erhalten, habe ich das Datum im Feld PROGNOSESTART auf den 31. März 2015 zurückgesetzt. Auf diese Weise kann ich für einige der Prognosen die tatsächlichen Daten sehen.

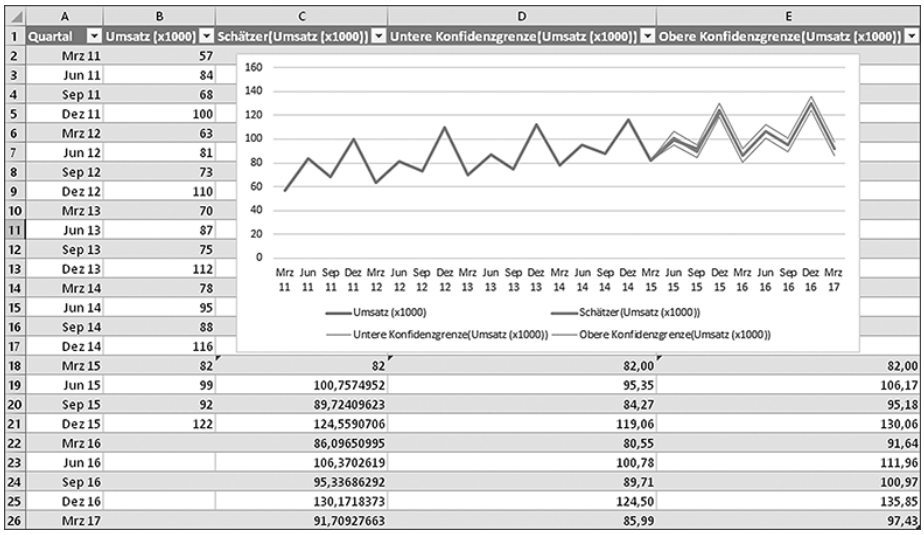
   Im Bereich SAISONALITÄT sehen Sie im grauen Feld neben MANUELL FESTLEGEN eine 2. Dies bedeutet, dass `PROGNOSE.ETS.SAISONALITÄT` ein Muster erkannt hat, das sich alle zwei Quartale wiederholt. Ale Geschäftsführer von FarDrate gehe ich jedoch davon aus, das sich das Muster alle vier Quartale wiederholt. Daher aktiviere ich das Optionsfeld MANUELL FESTLEGEN und ändere den Wert 2 in 4. Das Diagramm

im Dialogfeld wird sofort angepasst und zeigt genauere Prognosen sowie engere 95 %-Konfidenzgrenzen an.

Zusätzlich habe ich das Kontrollkästchen PROGNOSESTATISTIKEN AUFNEHMEN eingeschaltet. Ich kann im Dialogfeld weitere Dinge ändern, wie beispielsweise das Konfidenzintervall, jedoch reicht das, was ich getan habe, aus.

6. **Klicken Sie auf ERSTELLEN.**

   Hierdurch wird das von Excel erstellte Prognosearbeitsblatt geöffnet (siehe Abbildung 16.9).



	A	B	C	D	E
1	Quartal	Umsatz (x1000)	Schätzer(Umsatz (x1000))	Untere Konfidenzgrenze(Umsatz (x1000))	Obere Konfidenzgrenze(Umsatz (x1000))
2	Mrz 11	57			
3	Jun 11	84			
4	Sep 11	68			
5	Dez 11	100			
6	Mrz 12	63			
7	Jun 12	81			
8	Sep 12	73			
9	Dez 12	110			
10	Mrz 13	70			
11	Jun 13	87			
12	Sep 13	75			
13	Dez 13	112			
14	Mrz 14	78			
15	Jun 14	95			
16	Sep 14	88			
17	Dez 14	116			
18	Mrz 15	82	82	82,00	82,00
19	Jun 15	99	100,7574952	95,35	106,17
20	Sep 15	92	89,72409623	84,27	95,18
21	Dez 15	122	124,5590706	119,06	130,06
22	Mrz 16		86,09650995	80,55	91,64
23	Jun 16		106,3702619	100,78	111,96
24	Sep 16		95,33686292	89,71	100,97
25	Dez 16		130,1718373	124,50	135,85
26	Mrz 17		91,70927663	85,99	97,43

Abbildung 16.9: Das von Excel erstellte Prognosearbeitsblatt

Auf dem Prognosearbeitsblatt sind außer dem Diagramm in Spalte C auch die geschätzten Werte zu sehen (die mit `PROGNOSE.ETS` berechnet wurden), sowie in den Spalten D und E die Konfidenzgrenzen (diese wurden mit `PROGNOSE.ETS.KONFINT` berechnet).

In den Zeilen 1 bis 8 der Spalten G und H des Prognosearbeitsblatts sind einige Statistiken zu sehen, die mit `FORECAST.ETS.STAT` berechnet wurden, und die Abbildung 16.10 zeigt.

G	H
Statistik	Wert
Alpha	0,126
Beta	0,001
Gamma	0,001
MASE	0,095
SMAPE	0,026
MAE	2,317
RMSE	2,812

Abbildung 16.10: Die Prognosestatistiken

Alpha, Beta und Gamma sind die Werte, auf die ich mich im Kasten weiter vorne beziehe. Die anderen vier Kennziffern geben an, wie gut die Prognose mit den Daten übereinstimmt. Die ersten beiden sind etwas sehr kompliziert, daher überschlage ich sie.

Die dritte Metrik, MAE (Mean Absolute Error = mittlerer absoluter Fehler) ist der Mittelwert des absoluten Wertes aus der Differenz zwischen den Daten und den Prognose. Der letzte Wert, RMSE (Root Mean Squared Error = Wurzel der mittleren Fehlerquadratsumme) ist der Mittelwert der quadrierten Abweichungen der Daten von den Prognosen. Sie können sich den Standardfehler der Schätzung in der Regression (siehe Kapitel) als die Wurzel der mittleren Fehlerquadratsumme vorstellen, die um die Freiheitsgrade angepasst ist.

Kapitel 17

Parameterfreie Statistik

In diesem Kapitel

- Einführung in die parameterfreie Statistik
- Unabhängige Stichproben testen
- Verbundene Stichproben testen
- Rangkorrelationsanalyse

Die statistischen Verfahren, die ich in den bisherigen Kapiteln beschrieben habe, haben mehrere Merkmale gemeinsam. Erstens gehen wir davon aus, dass es sich um intervallskalierte Daten handelt. (Wenn Sie nicht wissen, was dies bedeutet, lesen Sie in Kapitel 1 den Abschnitt »Arten von Daten«.) Zweitens können wir Stichprobenstatistiken verwenden, um die Parameter der Stichprobenverteilung zu schätzen. Außerdem können wir den zentralen Grenzwertsatz verwenden, um die Merkmale der Verteilung zu bestimmen, damit wir Hypothesen testen können.

Manchmal müssen wir nominalskalierte oder ordinalskalierte Daten analysieren (auch diese Begriffe sind in Kapitel 1 erläutert). Und manchmal können wir die Verteilung der Variablen, mit der wir arbeiten, nicht angeben.

Um mit diesen Fällen umgehen zu können, haben Statistiker die parameterfreien statistischen Tests entwickelt. Die Liste dieser Tests ist sehr lang und sie wird immer länger. Viele dieser Verfahren benötigen für das Testen von Hypothesen spezielle Lookup-Tabellen.

Ich möchte diese speziellen Tabellen vermeiden. Um die Tests auszuwählen, die ich in diesem Kapitel beschreibe, habe ich

folgende Bedingungen verwendet. Der Test muss Hypothesen entweder anhand einer wohlbekannten, in Excel integrierten Verteilung überprüfen können oder eine wohlbekannte Verteilung verwenden, wenn die Stichproben groß sind. Von den parameterfreien Tests, auf die dies zutrifft, habe ich dann sechs klassische Verfahren ausgewählt.

Da Excel für diese Tests weder in den Analyse-Funktionen noch in den eingebauten Funktionen Werkzeuge bereitstellt, besteht die allgemeine Vorgehensweise darin, die Testformeln in Excel-Formeln zu übersetzen und dann eine der statistischen Funktionen von Excel (`NORM.VERT` oder `CHIQU.VERT.RE`) zu verwenden, um den Hypothesentest durchzuführen.

Der einfachste Weg, um diese Tests vorzustellen, besteht darin, Ihnen zuerst das fertige Arbeitsblatt zu zeigen und dann die Schritte zu beschreiben, mit denen ich das Arbeitsblatt erstellt habe.

Unabhängige Stichproben

Die parameterfreien Tests, die ich in diesem Abschnitt vorstelle, sind mit dem t-Test für unabhängige Gruppen und der einfaktoriellen Varianzanalyse vergleichbar.

Zwei Stichproben: Mann-Whitney-U-Test

Der Mann-Whitney-U-Test ist einer der bekanntesten parameterfreien Tests. Sie verwenden ihn, um zu testen, ob zwei unabhängige Stichproben aus der gleichen Grundgesamtheit stammen. Er wird in Situationen verwendet, in denen Sie keine intervallskalierten Daten vorliegen haben und den t-Test nicht einsetzen können.

Wenn es sich um ordinale Daten handelt, arbeiten Statistiker mit den Rängen der Daten und nicht mit den ursprünglichen Datenpunkten. Und das ist genau das, was dieser Test macht.

Angenommen, in einer Studie mit 40 Personen sollen sich diese zufällig entweder einen stressigen Film oder einen Zeichentrickfilm anschauen. Anschließend sollen die Testpersonen auf einer Skala von 1 (am wenigsten gestresst) bis 100 (am meisten gestresst) angeben, wie gestresst sie sich fühlen.

Abbildung 17.1 zeigt ein Arbeitsblatt, bei dem sich die Daten dieser Studie in den Spalten A und B befinden. Die Nullhypothese lautet, dass die beiden Gruppen aus der gleichen Grundgesamtheit stammen. Die alternative Hypothese lautet, dass der Stress nach einem stressigen Film höher ist. Daher kann hier ein einseitiger Test mit $\alpha = 0{,}05$ verwendet werden.

Die erste Aufgabe besteht darin, den Rang jedes Wertes zu bilden (in aufsteigender Reihenfolge), und zwar im Kontext aller Werte. Wenn die Nullhypothese wahr ist, müssten die hohen und die niedrigen Ränge (und alle Ränge dazwischen) in den beiden Gruppen gleichmäßig verteilt sein.

Für die Formel des Mann-Whitney-U-Tests kann ich entweder die Ränge der Werte aus Spalte A oder die Ränge der Werte aus Spalte B verwenden.

Wenn ich Spalte A verwende, sieht die Formel folgendermaßen aus:

$$U = N_A N_B + \frac{N_A \left(N_A + 1\right)}{2} - R_A$$

Wenn ich Spalte B verwende, ist das die Formel:

$$U = N_A N_B + \frac{N_B \left(N_B + 1\right)}{2} - R_B$$

	A	B	C	D	E	F	G	H
1	Stressiger Film	Zeichentrick		Rang Film	Rang Zeichentrick		NA	NB
2	87	32		39	9		20	20
3	55	15		25	2			
4	31	18		8	3		U=	271
5	45	20		19	4		Mittelwert=	200
6	67	43		31	18		Standardabw.=	36,968455
7	68	23		32	6			
8	22	76		5	36			
9	71	56		33	26		p=	0,97260617
10	46	49		20	22			
11	37	34		13	11			
12	39	53		15	24			
13	41	14		16	1			
14	74	47		35	21			
15	58	77		27	37			
16	60	35		28	12			
17	83	62		38	29			
18	33	38		10	14			
19	73	89		34	40			
20	52	29		23	7			
21	64	42		30	17			
22				481	339			

Abbildung 17.1: Der Mann-Whitney-U-Test

R_A ist die Summe der Ränge der Werte aus Spalte A, R_B ist die Summe der Ränge der Werte aus Spalte B, N_A ist die die Anzahl der Werte in Spalte A und N_B die Anzahl der Werte in Spalte B.

Bei großen Stichproben entspricht die Stichprobenverteilung von U ungefähr einer Normalverteilung mit:

$$\mu_U = \frac{N_A N_B}{2}$$

und:

$$\sigma_U = \sqrt{\frac{(N_A)(N_B)(N_A + N_B + 1)}{12}}$$

Zurück zu Abbildung 17.1. Die Ränge der Werte aus den Spalten A und B werden in den Spalten D und E angezeigt. Wie habe ich diese dorthin bekommen? Ganz einfach. In Zelle D2 gebe ich Folgendes ein:

```
=RANG.MITTELWERT(A2;$A$2:$B$21;1)
```

Anschließend habe ich Spalten D und E automatisch ausgefüllt.

Jetzt geht es darum, die U-Formel in eine Excel-Formel zu übersetzen. Damit die Formel deutlicher wird, habe ich dem Wert in Zelle G2 den Namen `NA` zugewiesen – dies ist die Anzahl der Werte in Spalte A – und dem Wert in Zelle H2 den Namen `NB`; dies ist die Anzahl der Werte in Spalte B.

Ich werde die U-Formel anhand der Werte in Spalte A berechnen. Darum gebe ich in Zelle H4 folgende Formel ein:

```
=NA*NB+((NA*(NA+1))/2)-E22
```

Ich drücke die Enter-Taste und erhalte als Ergebnis 271.

Jetzt der Test. Um den Mittelwert der Stichprobenverteilung zu erhalten, gebe ich

```
=(NA*NB)/2
```

in Zelle H5 und

```
=WURZEL((NA*NB*(NA+NB+1))/12)
```

in Zelle H6 ein. Um die mit U verknüpfte Wahrscheinlichkeit zu berechnen, gebe ich

```
=NORM.VERT(H4;H5;H6;WAHR)
```

in Zelle H8 ein. Das Ergebnis 0,97 bedeutet, dass die Wahrscheinlichkeit, dass U kleiner als der berechnete Wert ist, größer ist als 0,95. Damit ist die Wahrscheinlichkeit, dass ein Ergebnis extremer ist, weniger als 0,05. Daher kann ich die Nullhypothese verwerfen.

Mehr als zwei Stichproben: Kruska-Wallis einseitiger ANOVA-Test

FarKlempt Robotics führt unter den Mitarbeitern eine Umfrage über die Arbeitszufriedenheit durch. Entwickler, Manager, Wartungstechniker und technische Redakteure sollen ihre Arbeitszufriedenheit auf einer Skala von 1 (am wenigsten zufrieden) bis 100 (extrem zufrieden) bewerten. In jeder Berufskategorie gibt es sechs Mitarbeiter. Abbildung 17.2 zeigt das Arbeitsblatt, bei dem sich

die Daten in den Spalten A bis D und dort in den Zeilen 1 bis 7 befinden. Die Nullhypothese ist, dass die Stichproben aus der gleichen Grundgesamtheit stammen. Die alternative Hypothese ist, dass dies nicht stimmt.

Der parameterfreie Test, der sich für diese Daten eignet, ist die einseitige Varianzanalyse nach Kruskal-Wallis. Wie bei Mann-Whitney lege ich zuerst für alle 24 Werte deren Rang fest. Falls die Nullhypothese zutrifft, müssten die Ränge innerhalb der Gruppen gleichmäßig verteilt sein.

Die Formel dieser Statistik lautet wie folgt:

$$H = \left[\frac{12}{N(N+1)} \sum \frac{R^2}{n} \right] - 3(N+1)$$

	A	B	C	D	E	F	G
1	Entwickler	Manager	Wartung	Autoren		N_Gesamt	N_Gruppe
2	25	63	54	79		24	6
3	58	70	61	50			
4	57	52	95	77			
5	51	56	80	64			
6	42	62	78	55		H=	7,98
7	59	76	81	53		p=	0,04642686
8							
9	Rentwickler	Rmanager	Rwartung	Rautoren			
10	1	15	7	21			
11	11	17	13	3			
12	10	5	24	19			
13	4	9	22	16			
14	2	14	20	8			
15	12	18	23	6			
16	40	78	109	73			
17							

Abbildung 17.2: Die einseitige Kruskal-Wallis-Varianzanalyse

N ist die Gesamtanzahl der Wert und n ist die Anzahl der Werte in jeder Gruppe. Um das Beispiel einfach zu halten, verwende ich in jeder Gruppe die gleiche Anzahl von Werten, dies ist jedoch für diesen Test nicht erforderlich. *R* ist die Summe der Ränge in einer Gruppe. *H* ist ungefähr wie Chi-Quadrat verteilt, wobei df der Anzahl der Gruppen minus 1 entspricht, wenn jedes *N* größer ist als 5.

Wenn Sie einen Blick auf Abbildung 17.2 werfen, können Sie sehen, dass sich die Ränge der Daten in den Zeilen 9 bis 15 der Spalten A bis D befinden. In Zeile 16 befinden sich die Summen der Ränge der einzelnen Gruppen. Dem Wert in Zelle F2, in der sich die Gesamtanzahl der Werte befindet, habe ich den Namen `N_Gesamt` zugewiesen. Der Wert in G2, die Anzahl der Werte in jeder Gruppe, besitzt den Namen `n_Gruppe`.

Um *H* zu berechnen, gebe ich in Zelle G6 folgende Formel ein:

```
=(12/(N_Gesamt*(N_Gesamt+1)))*
(QUADRATESUMME(A16:D16)/n_Gruppe)-3*(N_Gesamt+1)
```

Für den Hypothesentest gebe ich in Zelle G7 diese Formel ein:

```
=CHIQU.VERT.RE(G6;3)
```

Das Ergebnis ist kleiner als 0,05, daher verwerfe ich die Nullhypothese.

Parallelstichproben

Die parameterfreien Tests, die ich Ihnen diesem Abschnitt vorstelle, sind mit dem t-Test für gepaarte Stichproben und der Varianzanalyse mit Messwiederholung vergleichbar.

Zwei Stichproben :Wilcoxon-Vorzeichen-Rang-Test

Dieser Test verwendet die Differenzen zwischen gepaarten Werten. Er geht jedoch noch ein wenig darüber hinaus. Neben der Richtung der Differenz berücksichtigt der Wilcoxon-Test auch die Größen der Differenzen. Dieser Test bietet sich daher dann an, wenn Sie ordinalskalierte Daten vorliegen haben, deren Genauigkeit präzise genug ist, um die Ränge der Differenzen zu bestimmen.

Hier ein Beispiel. Ein Soziologe untersucht 26 Paare eineiiger Zwillinge. Jedes der Kinder in einem Zwillingspaar besucht entweder eine öffentliche Schule oder wird zu Hause unterrichtet. Am Ende eines Schuljahres bewertet der Soziologe die Kontaktfreudigkeit der

Kinder auf einer Skala von 1 (nicht kontaktfreudig) bis 100 (sehr kontaktfreudig). Die Nullhypothese lautet, dass sich die beiden Gruppen nicht unterscheiden. Die Alternativhypothese ist, dass sie sich unterscheiden.

Abbildung 17.3 zeigt die Daten in den Zeilen 1 bis 27 der Spalten A und B. Die Differenz zwischen den Paaren steht in Spalte C und die Ränge der Differenzen stehen in Spalte D.

	A	B	C	D	E	F	G
1	Zwilling Öffentliche Schule	Zwilling Hausunterricht	Differenz	Rang der Differenz		N	
2	64	54	10	14		26	
3	44	51	-7	6			
4	51	48	3	11		Anzahl mit positiver Differenz=	18
5	34	39	-5	8		Anzahl mit negativer Differenz=	8
6	84	25	59	26			
7	66	46	20	19		Summe der seltener	36
8	61	78	-17	3		vorkommenden Ränge=	
9	22	58	-36	1			
10	39	30	9	13		Mittelwert=	175,5
11	78	38	40	25		Standardabweichung =	39,37321425
12	40	51	-11	4			
13	51	47	4	12		p=	0,000197788
14	54	39	15	16			
15	71	43	28	21			
16	70	43	27	20			
17	86	48	38	23			
18	74	72	2	10			
19	69	53	16	17			
20	63	30	33	22			
21	43	53	-10	5			
22	50	49	1	9			
23	50	56	-6	7			
24	33	51	-18	2			
25	52	39	13	15			
26	45	26	19	18			
27	56	17	39	24			

Abbildung 17.3: Der Wilcoxon-Vorzeichen-Rang-Test

Jetzt wird es interessant. Der nächste Schritt besteht darin, die Anzahl der Paare mit einer negativen Differenz und die Anzahl der Paare mit einer positiven Differenz zu berechnen. Wenn die Nullhypothese stimmt, müsste die Summe der Ränge der positiven Differenzen ungefähr so groß sein wie die Summe der Ränge der negativen Differenzen.

Sie verwenden dann die seltenere Kategorie und addieren deren Ränge. Ich bezeichne diese Summe mit *T*. Wenn die Anzahl der Paare, *N*, größer ist als 25, dann ist *T* normalverteilt mit

$$\mu_T = \frac{N(N+1)}{4}$$

und

$$\sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

Auf dem Arbeitsblatt berechne ich in Zelle G4 mit dieser Formel die Anzahl der positiven Differenzen:

```
=ZÄHLENWENN(C2:C27;">0")
```

Die Anzahl der negativen Differenzen berechne ich in Zelle G5 mit:

```
=ZÄHLENWENN(C2:C27;"<0")
```

Das Ergebnis ist, dass es 8 negative und 16 positive Differenzen gibt. Daher addiere ich anschließend die Summe der negativen Ränge. Hierzu gebe ich in Zelle G7 diese Formel ein:

```
=SUMMEWENN(C2:C27;"<0";D2:D27)
```

Um die Hypothesen zu testen, muss ich den Mittelwert und die Standardabweichung der Stichprobenverteilung berechnen. Ich habe Zelle F2, in der sich die Anzahl der Zwillingspaare befindet, den Namen `N` zugewiesen.

Daher gebe ich

```
=(N*(N+1))/4
```

in Zelle G10 ein, um den Mittelwert zu berechnen, und dann

```
=WURZEL((N*(N+1)*(2*N+1))/24)
```

in Zelle G11, um die Standardabweichung zu erhalten.

Und schließlich verwende ich in G12 diese Formel, um die Hypothese zu testen:

```
=NORM.VERT(G7;G10;G11;WAHR)
```

Der sehr kleine Wert – sehr viel kleiner als 0,05 – zeigt an, dass ich die Nullhypothese verwerfen kann.

Mehr als zwei Stichproben: zweiseitige Friedman-Varianzanalyse

Wenn Sie ordinalskalierte Daten und mehr als zwei verbundene Stichproben haben, ist die zweiseitige Friedman-Varianzanalyse der geeignete parameterfreie Test. Bei diesem Test werden die Ränge der Daten innerhalb der Stichprobe gebildet.

Ein Beispiel sehen Sie in Abbildung 17.4. Zwanzig Personen haben auf einer Skala von 1 (sehr wenig Wissen) bis 10 (sehr viel Wissen) ihre Kenntnisse über Wirtschaft, Geografie und Geschichte eingeschätzt. Die Daten befinden sich in den Zeilen 1 bis 21 der Spalten A bis D. Stellen Sie sich Wirtschaft, Geographie und Geschichte als drei verbundene Stichproben vor. Sie sind verbunden, da die drei Zahlen in einer Zeile die Daten der gleichen Person repräsentieren. Die Ränge innerhalb jeder der Stichproben befinden sich in den Spalten F bis H.

	A	B	C	D	E	F	G	H	I	J	K
1	Person	Wirtschaft	Geographie	Geschichte		Rwirtschaft	Rgeographie	Rgeschichte		Gesamt_N	Gruppen
2	1	9	5	7		3	1	2		20	3
3	2	8	7	9		2	1	3			
4	3	2	7	8		1	2	3			
5	4	4	5	9		1	2	3			
6	5	8	2	7		3	1	2			
7	6	5	4	6		2	1	3		$\chi_r^2 =$	10,9
8	7	1	10	9		1	3	2		p=	0,0042963
9	8	5	1	10		2	1	3			
10	9	9	4	10		2	1	3			
11	10	7	3	8		2	1	3			
12	11	4	7	9		1	2	3			
13	12	6	8	10		1	2	3			
14	13	9	5	10		2	1	3			
15	14	3	6	9		1	2	3			
16	15	5	7	3		2	3	1			
17	16	4	2	3		3	1	2			
18	17	7	8	10		1	2	3			
19	18	3	7	8		1	2	3			
20	19	9	5	7		3	1	2			
21	20	7	9	8		1	3	2			
22						35	33	52			

Abbildung 17.4: Die zweiseitige Varianzanalyse nach Friedman

Die Nullhypothese sagt, dass die drei verschiedenen Wissensbereiche keine Unterschiede aufweisen; die Gegenhypothese lautet, dass es Unterschiede gibt. Wenn die Nullhypothese nicht wahr ist, müsste die Summe der Ränge für die drei Wissensbereiche annähernd gleich sein.

Friedmann verweist auf diese Teststatistik mit x_r^2, daher werde ich dies auch tun. Die Formel lautet:

$$x_r^2 = \left[\frac{12}{Nk(k+1)} \sum R^2 \right] - 3N(K+1)$$

N ist die Anzahl der Individuen und *k* ist die Anzahl der Gruppen (Wirtschaft, Geographie, Geschichte). Wenn die Stichprobe groß genug ist (mehr als 9), ist x_r^2 mit *k*–1 Freiheitsgraden wie Chi-Quadrat verteilt.

Um die Ränge in die Spalten F bis H einzufügen, gebe ich in Zelle F2 diese Formel ein:

```
=RANG.MITTELWERT(B2;$B2:$D2;1)
```

Beachten Sie das Dollarzeichen vor der Angabe der Spalte; dies fehlt vor der Angabe der Zeile. Dies erlaubt es mir, die Zellen von F2 bis H21 automatisch auszufüllen. Anschließend bilde ich in Zeile 22 die Summen der Ränge.

Damit die Excel-Formel besser lesbar ist, weise ich Zelle J2 (dort steht die Anzahl der Personen aus der Studie) den Namen `Gesamt_N` zu. Der Name für Zelle K2 (die Anzahl der Wissensbereiche) lautet `Gruppen`.

Um x_r^2 zu berechnen, gebe ich in Zelle K7 diese Formel ein:

```
=(12/(Gesamt_N*Gruppen*
(Gruppen+1))*QUADRATESUMME(F22:H22))-3*Gesamt_N*
(Gruppen+1)
```

Um die Hypothese zu testen, verwende ich in K8 diese Formel:

```
=CHIQU.VERT.RE(K7;Gruppen-1)
```

Da der Wert kleiner ist als 0,05, verwerfe ich die Nullhypothese.

Mehr als zwei Stichproben: Cochran-Q-Test

Wie bitte? Ein weiterer Test für mehr als zwei Stichproben? Es handelt sich hier um eine ganz andere Sache. Die anderen Tests in diesem Kapitel arbeiten mit ordinalen Daten. Dieser hier verwendet nominalskalierte Daten.

Im Arbeitsblatt in Abbildung 17.5 befinden sich die Daten einer Studie mit 20 Personen, die Anagramme lösen. Die Anagramme sind entweder leicht (Spalte B), mittelschwer (Spalte C) oder schwierig (Spalte D). Wenn eine Person das Anagramm innerhalb einer Minute löst, gilt dies als Erfolg und wird mit 1 gekennzeichnet. Gelingt es nicht, ist dies ein Misserfolg, der mit 0 gekennzeichnet wird.

	A	B	C	D	E	F	G	H
1	Person	Einfach	Mittel	Schwierig	L		k	
2	1	1	0	0	1		3	
3	2	1	1	0	2			
4	3	0	1	1	2			
5	4	0	0	0	0		Q=	10,125
6	5	1	0	0	1		p=	0,006329715
7	6	1	1	0	2			
8	7	1	1	0	2			
9	8	0	1	0	1			
10	9	1	0	0	1			
11	10	0	0	0	0			
12	11	1	1	0	2			
13	12	0	1	0	1			
14	13	1	1	0	2			
15	14	1	1	1	3			
16	15	1	1	0	2			
17	16	1	1	1	3			
18	17	1	1	0	2			
19	18	1	1	0	2			
20	19	1	0	1	2			
21	20	0	1	1	2			
22	G	14	14	5				

Abbildung 17.5: Cochran-Q-Test

Die Nullhypothese lautet, dass es bei den drei Schwierigkeitsgraden keine Unterschiede gibt. Die alternative Hypothese ist, dass diese Unterschiede sehr wohl existieren. Wenn die Nullhypothese wahr ist, ist die Summe der drei Bedingungen (Zellen B22, C22, und D22) annähernd gleich.

Die Formel für diesen Test lautet:

$$Q = \frac{K(K-1)\sum(G-\bar{G})^2}{k\sum L - \sum L^2}$$

In dieser Formel entspricht *G* der Summe einer Spalte, *L* der Summe einer Zeile und *k* entspricht der Anzahl der Bedingungen (in unserem Beispiel also 3). Die Stichprobenverteilung von *Q* ist annähend Chi-Quadrat verteilt mit *k*–1 Freiheitsgraden.

Damit die Excel-Formel besser lesbar ist, habe ich den Zeilensummen in Spalte E den Namen *L* zugewiesen; die Spaltensummen in Zeile 22 besitzen den Namen *G* und *k* ist der Name für den Wert in Zelle G2 (die Anzahl der Bedingungen).

Die Formel für die Berechnung von *Q* in Zelle H5 lautet somit wie folgt:

```
=(k*(k-1)*SUMQUADABW(G))/(k*SUMME(L)-
QUADRATESUMME(L))
```

Kleine Auffrischung zu `SUMQUADABW`? Werfen Sie einen Blick in Kapitel 5.

Um die Hypothese zu testen, gebe ich in H6 diese Formel ein:

```
=CHIQU.VERT.RE(H5;k-1)
```

Der Wert ist kleiner als 0,05. Dies zeigt an, dass ich die Nullhypothese verwerfen sollte.

Spearmans Rangkorrelationskoeffizient

Spearmans Rangkorrelationskoeffizient, r_S, war einer der ersten parameterfreien Tests, die auf Rängen basieren. Das Konzept

besteht darin, bei einer Stichprobe von Individuen, für die jeweils zwei Variablen gemessen wurden, für jeden Wert innerhalb einer Variablen den Rang zu bestimmen. Dann wird für jedes Individuum der eine Rang vom anderen subtrahiert. Falls die Korrelation (in positiver Richtung) perfekt ist, haben alle Differenzen den Wert null.

Abbildung 17.6 zeigt ein Beispiel. Ein Betriebspsychologe hat die Kontaktfreudigkeit von 20 Mitarbeiten von FarDrate TimePiece Corporation bewertet. Die Skala reicht von 1 (nicht kontaktfreudig) bis 100 (außerordentlich kontaktfreudig). Die gleichen Mitarbeiter haben eine Einschätzung über ihre Arbeitszufriedenheit abgegeben, und zwar auf einer Skala von 1 (geringste Zufriedenheit) bis 80 (höchste Zufriedenheit). Die Nullhypothese lautet, dass es keinen Zusammenhang zwischen der Kontaktfreudigkeit und der Arbeitszufriedenheit gibt. Die Gegenhypothese ist, dass diese beiden Variablen korrelieren.

Die Daten befinden sich in den Spalten B und C und die Ränge in den Spalten E und F. Die Differenzen aus den Rängen der Paare befinden sich in Spalte G.

	A	B	C	D	E	F	G	H	I	J
1	Person	Kontaktfreudigkeit (1-100)	Arbeitszufriedenheit (1-80)		Rkontakt	Rarbeits	Differenz		Anzahl der Paare	
2	1	84	78		16	20	-4		20	
3	2	45	52		5	9	-4			
4	3	91	70		18	15	3		rs =	0,571429
5	4	34	38		3	4	-1			
6	5	67	55		11	11	0		t=	2,954196
7	6	74	57		14	13	1		p=	0,008489
8	7	51	77		6	19	-13			
9	8	32	56		1	12	-11			
10	9	59	32		9	2	7			
11	10	71	75		13	17	-4			
12	11	42	42		4	6	-2			
13	12	53	46		8	7	1			
14	13	64	49		10	8	2			
15	14	99	72		20	16	4			
16	15	92	68		19	14	5			
17	16	81	39		15	5	10			
18	17	33	22		2	1	1			
19	18	68	54		12	10	2			
20	19	52	36		7	3	4			
21	20	90	76		17	18	-1			

Abbildung 17.6: Spearmans Rangkorrelationskoeffizient

Die Formel lautet:

$$r_S = 1 - \frac{6\sum d^2}{n^3 - n}$$

wobei *d* der Differenz eines Wertepaares entspricht. Wie beim normalen Korrelationskoeffizienten (siehe Kapitel 15) sollte der Wert von r_S ungefähr dann bei null liegen, wenn die Nullhypothese wahr ist.

Um die Ränge in Spalte E zu berechnen, gebe ich in Zelle E2 diese Formel ein:

```
=RANG.GLEICH(B2;$B$2:$B$21;1)
```

Den Rest der Spalte fülle ich automatisch aus. Für die Ränge in Spalte F gebe ich in Zelle F2 die folgende Formel ein und fülle auch hier die weiteren Zellen automatisch aus:

```
=RANG.GLEICH(C2;$C$2:$C$21;1)
```

Zur Berechnung des Korrelationskoeffizienten brauche ich in Zelle J4 keine komplizierte Formel einzugeben. Warum? Weil Excel und die mathematische Statistik gut zusammenarbeiten und mit einer Überraschung aufwarten. Alles, was ich in Zelle J4 eingeben muss, ist dies:

```
=KORREL(E2:E21;F2:F21)
```

Das war es dann schon. Wenn Sie `KORREL` auf die Ränge anwenden, erhalten Sie das gleiche Ergebnis wie mit der obigen Formel. (Darum ist es auch nicht zwingend erforderlich, in Spalte G die Differenz zwischen den Rangpaaren zu berechnen.)

Der Hypothesentest ist auch ein alter Bekannter, wenn Sie Kapitel 15 gelesen haben. Berechnen Sie:

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

n ist die Anzahl der Paare; der Test besitzt *n*–2 Freiheitsgrade.

Für den Wert in Zelle I2 habe ich den Namen `Anzahl_der_Paare` definiert. Darum gebe ich

```
=J4*WURZEL(Anzahl_der_Paare-2)/WURZEL(1-J4^2)
```

in J6 ein und

```
=T.VERT.2S(J6;Anzahl_der_Paare-2)
```

in J7. Ich habe die Funktion für die zweiseitige t-Verteilung verwendet, da ich die Richtung der Korrelation im Vorhinein nicht kenne. Auch hier zeigt der kleine p-Wert an, dass ich die Nullhypothese verwerfen sollte.

Ein paar Hinweise zum Schluss

Ich möchte Sie noch gern auf ein paar Dinge aufmerksam machen. Erstens habe ich in den Beispielen keine Ränge mit Bindungen *(Ties)* verwendet. Für rangbasierte Statistiken ergeben sich hierdurch kleinere Probleme (bis auf die zweiseitige Friedmann-Varianzanalyse) und ich wollte diese vermeiden.

Zweitens verbergen sich in späteren Kapiteln dieses Buches zusätzliche parameterfreie Tests. In Kapitel 18 lernen Sie Hypothesentests kennen, die auf der Binomialverteilung basieren. In Kapitel 22 beschreibe ich, wie Sie die Unabhängigkeit von zwei nominalen Variablen testen können.

Teil IV

Wahrscheinlichkeit

In diesem Teil …

Mit Zufallsvariablen arbeiten

Zählregeln verstehen

Bedingte Wahrscheinlichkeit einsetzen

Wahrscheinlichkeitsverteilungen visualisieren

Modellieren und simulieren

Kapitel 18

Einführung in die Wahrscheinlichkeit

In diesem Kapitel

- Wahrscheinlichkeit definieren
- Mit Wahrscheinlichkeit arbeiten
- Mit Zufallsvariablen und deren Verteilungen umgehen
- Sich auf die Binomialverteilung konzentrieren

In diesem Buch erwähne ich immer wieder das Konzept der Wahrscheinlichkeit, da es die Grundlage für das Testen von Hypothesen und die Inferenzstatistik ist. Meistens stelle ich die Wahrscheinlichkeit als den Bereich unter einem Teil einer Verteilung dar. So ist die Wahrscheinlichkeit eines Fehlers 1. Art (auch bekannt als Fehler vom Typ α) der Bereich am Ende der Standardnormalverteilung oder der *t*-Verteilung.

In diesem Kapitel werde ich die Wahrscheinlichkeit detaillierter betrachten und Zufallsvariablen, Permutationen und Kombinationen einführen. Ich werde die Grundlagen und Anwendungen der Wahrscheinlichkeit untersuchen und mich dann eingehend mit einer Reihe spezifischer Wahrscheinlichkeitsverteilungen befassen. Nachdem Sie die theoretischen Grundlagen zu den Wahrscheinlichkeitsverteilungen kennen, stelle ich die Excel-Tabellenblattfunktionen vor, die mit der Wahrscheinlichkeit zusammenhängen.

Was ist Wahrscheinlichkeit?

Die meisten von uns haben intuitiv eine Vorstellung davon, worum es bei der Wahrscheinlichkeit überhaupt geht. Wenn man eine Münze wirft, hat man eine 50-50-Chance, dass das Ergebnis »Kopf« lautet. Wenn man würfelt (mit nur einem Würfel), besteht die Wahrscheinlichkeit, dass man eine »2« würfelt, 1/6.

Wenn Sie Ihre Definition etwas formaler gestalten möchten, sprechen Sie höchstwahrscheinlich über alle möglichen Dinge, die passieren können, und darüber, wie groß die Wahrscheinlichkeit für die Dinge ist, die Sie interessieren. Wenn Sie eine Münze werfen, können zwei Dinge passieren, und wenn davon nur eines für Sie interessant ist (Kopf), dann beträgt die Wahrscheinlichkeit, dass das betreffende Ereignis eintritt, 1/2. Wenn Sie würfeln, können sechs Dinge passieren. Wenn davon nur eins für Sie von Interesse ist (die gewürfelte 2), dann beträgt die Wahrscheinlichkeit, dass dieses Ereignis eintritt, 1/6.

Experimente, Versuche, Ereignisse und Stichprobenräume

Statistiker und andere Menschen, die mit Wahrscheinlichkeit arbeiten, bezeichnen den Vorgang des Werfens einer Münze oder eines Würfels vermutlich als *Experiment*. Jedes Mal, wenn Sie den Prozess durchlaufen, ist dies ein *Versuch*.

Das deckt sich zwar vielleicht nicht mit Ihrer persönlichen Definition eines Experiments (oder auch eines Versuchs), aber für einen Statistiker ist ein Experiment ein Prozess, der eins von mindestens zwei unterschiedlichen Ergebnissen liefert (wie Kopf oder Zahl).

Hier ein weiterer Teil der Definition eines Experiments: Sie können das Ergebnis nicht mit Sicherheit vorhersagen. Jedes unterschiedliche Ergebnis wird als *Elementarereignis* bezeichnet. Wenn Sie eine Reihe von Elementarereignissen zusammenführen, haben Sie ein *Ereignis*. Wenn wir beim Beispiel des Würfels bleiben,

ergeben die Elementarereignisse 2, 4 und 6 das Ereignis »gerade Zahl«.

Wenn Sie alle möglichen Elementarereignisse zusammenfassen, erhalten Sie einen so genannten *Wahrscheinlichkeitsraum*. Die Zahlen 1, 2, 3, 4, 5 und 6 ergeben zusammen den Wahrscheinlichkeitsraum für einen Würfel. »Kopf« und »Zahl« bilden den Wahrscheinlichkeitsraum einer Münze.

Wahrscheinlichkeitsräume und Wahrscheinlichkeit

Was hat das alles mit Wahrscheinlichkeit zu tun? Wenn jedes Elementarereignis in einem Wahrscheinlichkeitsraum gleich wahrscheinlich ist, dann ist die Wahrscheinlichkeit eines Ereignisses

$$P(\text{Ereignis}) = \frac{\text{Anzahl der Elementarereignisse im Ereignis}}{\text{Anzahl der Elementarereignisse im Wahrscheinlichkeitsraum}}$$

Das »P« steht für Wahrscheinlichkeit (im Englischen: probability).

Die Wahrscheinlichkeit, beim Würfeln eine gerade Zahl zu erhalten, ist demnach

$$P(\text{gerade Zahl}) = \frac{\text{Anzahl der geradzahligen Elementarereignisse}}{\text{Anzahl der möglichen Würfelergebnisse}} = \frac{3}{6} = 0{,}5$$

Wenn die Elementarereignisse nicht alle die gleiche Wahrscheinlichkeit haben, ermitteln Sie die Wahrscheinlichkeit eines Ereignisses auf andere Art und Weise. Zunächst einmal benötigen Sie eine Vorgehensweise, mit der Sie jedem Elementarereignis eine Wahrscheinlichkeit zuweisen können. Dann addieren Sie die Wahrscheinlichkeiten der Elementarereignisse, die das Ereignis bilden.

Bei der Ergebniswahrscheinlichkeit sind ein paar Dinge zu beachten: Jede Wahrscheinlichkeit muss zwischen null und eins liegen. Die Wahrscheinlichkeiten der Elementarereignisse in einem Wahrscheinlichkeitsraum müssen addiert 1,00 ergeben.

Wie können Sie die Wahrscheinlichkeiten zuweisen? Manchmal verfügen Sie im Vorfeld über Informationen, wie zum Beispiel die, dass beim Werfen von Münzen in 60 Prozent der Fälle Kopf oben liegt. Manchmal müssen Sie die Situation nur gedanklich durchspielen, um die Wahrscheinlichkeit eines Ergebnisses herauszufinden.

Hier schnell ein Beispiel für das »gedankliche Durchspielen«. Nehmen wir einmal an, ein Würfel ist so konfiguriert, dass die Wahrscheinlichkeit eines Ergebnisses proportional zum numerischen Wert der Ergebnisse ist: Eine 6 kommt sechs Mal öfter vor als eine 1, eine 5 fünf Mal öfter als eine 1 usw. Wie hoch ist die Wahrscheinlichkeit der einzelnen Ergebnisse? Alle Wahrscheinlichkeiten müssen zusammenaddiert 1,00 ergeben und alle Augen auf einem Würfel ergeben zusammenaddiert 21 (1 + 2 + 3 + 4 + 5 + 6 = 21). Die Wahrscheinlichkeit ist demnach: $P(1) = 1/21$, $P(2) = 2/21$, …, $P(6) = 6/21$.

Zusammengesetzte Ereignisse

Es gibt einige Regeln für den Umgang mit *zusammengesetzten Ereignissen*, die Ihnen beim »gedanklichen Durchspielen« helfen können. Ein zusammengesetztes Ereignis besteht aus mehr als einem Ereignis. Ereignisse können durch *Vereinigung* oder *Schnitt* (oder durch beide Verfahren) miteinander kombiniert werden.

Vereinigung und Schnitt

Wie hoch ist die Wahrscheinlichkeit, mit einem neutralen Würfel eine 1 oder eine 4 zu würfeln? Mathematiker haben ein Symbol für *oder*. Es sieht so aus $\cup$ und heißt *Vereinigung*. Mit diesem Symbol ist die Wahrscheinlichkeit einer 1 oder einer 4 $P(1 \cup 4)$.

Bei dieser Art von Wahrscheinlichkeit empfiehlt es sich, den Überblick über die Elementarereignisse zu behalten. In jedem Ereignis gibt es ein Elementarereignis. Das Ereignis »1 oder 4« hat demnach zwei Elementarereignisse. Bei einem Wahrscheinlichkeitsraum von sechs Ergebnissen beträgt die Wahrscheinlichkeit 2/6 bzw. 1/3. Eine andere Berechnungsmöglichkeit ist diese:

$$P(1 \cup 4) = P(1) + P(4) = (1/6) + (1/6) = 2/6 = 1/3$$

Hier noch eine etwas komplexere Frage: Wie hoch ist die Wahrscheinlichkeit, eine Zahl zwischen 1 und 3 oder eine Zahl zwischen 2 und 4 zu würfeln?

Eine einfache Addition der Elementarereignisse für die einzelnen Ereignisse bringt uns dieses Mal leider nicht weiter. Im Ereignis »zwischen 1 und 3« gibt es drei Ereignisse, genauso wie im Ereignis »zwischen 2 und 4«. Die Wahrscheinlichkeit kann nicht 3 + 3 geteilt durch die sechs Ereignisse im Wahrscheinlichkeitsraum sein, da dies 1,00 ergibt und somit für $P(5)$ und $P(6)$ nichts mehr übrig bleibt. Aus diesem Grund können die Wahrscheinlichkeiten nicht einfach addiert werden.

Das Problem liegt in der Überlappung der beiden Ereignisse. Die Elementarereignisse für »zwischen 1 und 3« sind 1, 2 und 3. Die Elementarereignisse für »zwischen 2 und 4« sind 2, 3 und 4. Zwei Ereignisse überlappen sich: 2 und 3. Um sie nicht zwei Mal zu zählen, müssen sie von der Gesamtsumme abgezogen werden. Durch ein paar Dinge wird die Sache einfacher, während ich fortfahre. Ich kürze »zwischen 1 und 3« ab und nenne es *A*; »zwischen 2 und 4« nenne ich *B*. Außerdem verwende ich das mathematische Symbol für »überlappen«. Das Symbol ist $\cap$ und es heißt *Schnitt*.

Bei Verwendung der Symbole lautet die Wahrscheinlichkeit für »zwischen 1 und 3« oder »zwischen 2 und 4«.

$$P(A \cup B) = \frac{\left(\begin{array}{l}\text{Anzahl der Ereignisse in A + Anzahl der Ereignisse in B} \\ \text{– Anzahl der Ereignisse in (A} \cap \text{B)}\end{array}\right)}{\text{Anzahl der Ereignisse im Wahrscheinlichkeitsraum}}$$

$$P(A \cup B) = \frac{3+3-2}{6} = \frac{4}{6} = \frac{2}{3}$$

Sie können auch die Wahrscheinlichkeiten verwenden:

$$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

Die allgemeine Formel lautet:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Warum konnten im ersten Beispiel die Wahrscheinlichkeiten noch einfach addiert werden? Weil *P*(1 ∩ 4) null ist: Es ist nicht möglich, bei einem Wurf des Würfels gleichzeitig eine 1 und eine 4 zu würfeln. Wenn *P*(*A* ∩ *B*) = 0 gilt, *schließen sich A* und *B gegenseitig aus*.

Noch mehr zum Schnitt

Stellen Sie sich vor, Sie werfen eine Münze und würfeln gleichzeitig. Diese beiden Experimente sind *unabhängig* voneinander, da das Ergebnis des einen keinen Einfluss auf das Ergebnis des anderen hat.

Wie hoch ist die Wahrscheinlichkeit, Kopf zu werfen und eine 4 zu würfeln? Sie verwenden das Schnittsymbol und notieren die Formel für *P*(Kopf ∩ 4):

$$P(\text{Kopf} \cap 4) = \frac{\text{Anzahl der Elementarereignisse in Kopf} \cap 4}{\text{Anzahl der Elementarereignisse im Wahrscheinlichkeitsraum}}$$

Beginnen Sie mit dem Wahrscheinlichkeitsraum. In Tabelle 18.1 sind alle Elementarereignisse aufgeführt.

Kopf, 1	Zahl, 1
Kopf, 2	Zahl, 2
Kopf, 3	Zahl, 3
Kopf, 4	Zahl, 4
Kopf, 5	Zahl, 5
Kopf, 6	Zahl, 6

Tabelle 18.1: Die Elementarereignisse im Wahrscheinlichkeitsraum beim Werfen einer Münze und beim Würfeln mit nur einem Würfel

Aus der Tabelle wird ersichtlich, dass es zwölf mögliche Ereignisse gibt. Wie viele Ereignisse gibt es im Ereignis »Kopf und 4«? Nur eins. Es folgt

$$P(\text{Kopf} \cap 4) = \frac{\text{Anzahl der Elementarereignisse in Kopf} \cap 4}{\text{Anzahl der Elementarereignisse im Wahrscheinlichkeitsraum}} = \frac{1}{12}$$

Sie können auch die Wahrscheinlichkeiten verwenden:

$$P(\text{Kopf} \cap 4) = P(\text{Kopf}) \times P(4) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

Allgemein ausgedrückt gilt Folgendes, wenn *A* und *B* unabhängig voneinander sind:

$$P(A \cap B) = P(A) \cdot P(B)$$

Bedingte Wahrscheinlichkeit

Unter bestimmten Umständen engen Sie den Wahrscheinlichkeitsraum ein. Nehmen wir einmal an, ich würfele und sage Ihnen, das Ergebnis sei größer als 2. Wie hoch ist die Wahrscheinlichkeit, dass es eine 5 ist?

Normalerweise wäre die Wahrscheinlichkeit, dass es sich um eine 5 handelt, 1/6. In diesem Fall umfasst der Wahrscheinlichkeitsraum

allerdings nicht 1, 2, 3, 4, 5 und 6. Wenn Sie bereits wissen, dass das Ergebnis größer als 2 ist, umfasst der Wahrscheinlichkeitsraum 3, 4, 5 und 6. Die Wahrscheinlichkeit einer 5 beträgt dann 1/4.

Das ist ein Beispiel für eine *bedingte Wahrscheinlichkeit*. Sie ist »bedingt«, weil ich eine »Bedingung« vorgegeben habe – der Wurf führte zu einem Ergebnis größer 2. Die Schreibweise dafür lautet

$P(5 \mid \text{größer als } 2)$

Der vertikale Strich ist das Symbol für *unter der Bedingung*. Sie lesen den Ausdruck wie folgt: »Die Wahrscheinlichkeit einer 5 unter der Bedingung größer als 2.«

Mit Wahrscheinlichkeiten arbeiten

Wenn Sie zwei Ereignisse *A* und *B* haben, gilt im Allgemeinen

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

solange $P(B)$ nicht null ist.

Für den Schnitt im Zähler rechts ist dies *kein* Fall, bei dem man die Wahrscheinlichkeiten einfach miteinander multiplizieren kann. Wäre dies möglich, würde es sich nicht um eine bedingte Wahrscheinlichkeit handeln, da es ja bedeuten würde, dass *A* und *B* unabhängig sind. Wenn sie aber unabhängig sind, kann ein Ereignis nicht die Bedingung für das andere Ereignis sein.

Die Wahrscheinlichkeit des Schnitts müssen Sie gedanklich durchspielen. Wie viele Ergebnisse gibt es bei einem Würfel in dem Ereignis »5 ∩ größer als 2«? Nur eins. Also gilt: $P(5 \cap \text{größer als } 2)$ ist 1/6 und

$$P(5|\text{größer als } 2) = \frac{P(5 \cap \text{größer als } 2)}{P(\text{größer als } 2)} = \frac{\frac{1}{6}}{\frac{4}{6}} = \frac{1}{4}$$

Die Grundlage des Testens von Hypothesen

Das Testen von Hypothesen, das ich in früheren Kapiteln beschrieben habe, beinhaltet die bedingte Wahrscheinlichkeit. Wenn Sie eine Stichprobe berechnen, einen statistischen Test durchrechnen und die Testgröße dann mit einem kritischen Wert vergleichen, suchen Sie nach einer bedingten Wahrscheinlichkeit. Genauer gesagt, suchen Sie nach

$$P(\text{erhaltene Testgröße oder ein extremerer Wert} | H_0 \text{ ist wahr})$$

Wenn die bedingte Wahrscheinlichkeit gering ist (weniger als 0,05 in den Beispielen in den Kapiteln zum Testen von Hypothesen), verwerfen Sie H_0.

Große Wahrscheinlichkeitsräume

Wenn man sich mit Wahrscheinlichkeit befasst, ist es wichtig, den Wahrscheinlichkeitsraum richtig zu verstehen. In den Beispielen, die ich Ihnen gebe, sind die Wahrscheinlichkeitsräume klein. Bei einer Münze oder einem Würfel ist es einfach, die Elementarereignisse aufzulisten.

Im wirklichen Leben ist aber natürlich kaum etwas so einfach strukturiert. In den meisten Fällen gibt es große Wahrscheinlichkeitsräume und deshalb ist es eher unangenehm, jedes einzelne Elementarereignis aufzuführen.

Stellen Sie sich beispielsweise einmal vor, dass Sie zwei Mal würfeln. Wie viele Elementarereignisse gibt es in dem Wahrscheinlichkeitsraum für beide Würfe? Sie können sich natürlich hinsetzen und sie einzeln auflisten, aber es wäre zweifellos besser, die Frage durch Schlussfolgerung zu lösen: Es gibt sechs Möglichkeiten für den ersten Wurf. Jedes dieser Ergebnisse kann mit

sechs Möglichkeiten im zweiten Wurf kombiniert werden. Der Wahrscheinlichkeitsraum umfasst also 6 × 6 = 36 mögliche Elementarereignisse. (Das ähnelt dem Wahrscheinlichkeitsraum des Münze/Würfel-Beispiels in Tabelle 18.1, wo der Wahrscheinlichkeitsraum aus 2 × 6 = 12 Elementarereignissen besteht. Zwölf Ereignisse lassen sich noch leicht in einer Tabelle darstellen. Bei 36 Ergebnissen fängt es langsam an, lästig zu werden.)

Auch Ereignisse erfordern häufig einen gewissen gedanklichen Aufwand. Wie hoch ist die Wahrscheinlichkeit, dass die Gesamtzahl der Augen, wenn man zwei Mal würfelt, fünf beträgt? Sie müssen die Anzahl der Kombinationen zählen, bei denen die Gesamtaugenzahl fünf beträgt, und diese dann durch die Anzahl der Elementarereignisse im Wahrscheinlichkeitsraum (36) teilen. Sie erhalten fünf Augen bei den folgenden Wurfkombinationen: 1 und 4, 2 und 3, 3 und 2 oder 4 und 1. Das sind insgesamt vier Möglichkeiten und die Möglichkeiten überlappen sich nicht gegenseitig (ähm, ich meine, sie schneiden sich nicht). Also gilt:

$$P(5) = \frac{\text{Anzahl der Möglichkeiten, eine 5 zu würfeln}}{\text{Anzahl der möglichen Ergebnisse von zwei Würfen}} = \frac{4}{36} = 0{,}11$$

Sämtliche Elementarereignisse für den Wahrscheinlichkeitsraum aufzulisten, ist häufig ein Albtraum. Zum Glück gibt es Abkürzungen, die ich Ihnen in den folgenden Abschnitten vorstellen möchte. Da eine solche Abkürzung dabei hilft, schnell eine Anzahl von Dingen zu zählen, wird die Abkürzung auch als *Zählregel* bezeichnet.

Ob Sie es glauben oder nicht: Ich habe Ihnen gerade eine Zählregel untergeschoben, ohne dass Sie es bemerkt haben. Ein paar Absätze weiter oben habe ich Ihnen mitgeteilt, dass der Wahrscheinlichkeitsraum bei zwei Würfen des Würfels 6 × 6 = 36 mögliche Ergebnisse umfasst. Das ist die so genannte *Produktregel*: Wenn es im ersten Versuch eines Experiments N_1 mögliche Ergebnisse gibt und im zweiten Versuch N_2 mögliche Ergebnisse, beträgt die Anzahl der möglichen Ergebnisse $N_1 N_2$. Jedes mögliche

Ergebnis des ersten Versuchs kann mit sämtlichen möglichen Ergebnissen des zweiten Versuchs kombiniert werden. Wie wäre es mit drei Versuchen? Das hieße dann $N_1N_2N_3$.

Jetzt zu einigen weiteren Zählregeln.

Permutationen

Nehmen Sie einmal an, Sie müssen fünf Objekte in eine Reihenfolge bringen. Wie viele Möglichkeiten gibt es dafür? Für die erste Position in der Folge gibt es fünf Auswahlmöglichkeiten. Sobald Sie eine Wahl getroffen haben, gibt es für die zweite Position vier Auswahlmöglichkeiten. Für die dritte Position ergeben sich dann drei Möglichkeiten, für die vierte zwei und eine für die fünfte Position. Die Anzahl der Möglichkeiten beträgt daher (5)(4)(3)(2)(1) = 120.

Im Allgemeinen ist die Anzahl der Folgen von *N* Objekten $N(N–1)(N–2)…(2)(1)$. Diese Art der Berechnung ist recht häufig in Wahrscheinlichkeitsanwendungen anzutreffen und hat ihre eigene Schreibweise, nämlich »*N*!«. Das liest man allerdings nicht, indem man mit lauter Stimme »N« ruft. Stattdessen heißt es *N Fakultät*. Aufgrund der Definition gilt 1! = 1 und 0! = 1.

Nun zum angenehmen Teil. Wenn Sie die 26 Buchstaben des Alphabets in einer Folge sortieren müssen, beträgt die Anzahl der möglichen Reihenfolgen 26!, eine enorm große Zahl. Nehmen Sie jetzt aber einmal an, Sie müssen Sequenzen mit jeweils fünf Buchstaben bilden und dabei darauf achten, dass sich innerhalb der Sequenzen kein Buchstabe wiederholt. Wie viele Möglichkeiten gibt es dafür? Sie haben 26 Möglichkeiten für den ersten Buchstaben, 25 für den zweiten, 24 für den dritten, 23 für den vierten, 22 für den fünften und das war's. Das ergibt also (26)(25)(24)(23)(22). Und so hängt dieses Produkt mit 26! zusammen:

$$\frac{26!}{21!}$$

Die Sequenz wird als *Permutation* bezeichnet. Wenn Sie Permutationen von *n* Dingen von jeweils dem Umfang *k* bilden,

lautet die Schreibweise $P_{k,n}$ (*P* steht für Permutation). Die Formel ist:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

Nur der Vollständigkeit halber hier ein weiteres Problemen. Angenommen, ich würde in dieser Sequenz von fünf Objekten Wiederholungen erlauben. Die Folge aaabc wäre daher eine erlaubte Sequenz. In diesem Fall wäre die Anzahl der Möglichkeiten 26 × 26 × 26 × 26 × 26, oder wie Mathematiker es ausdrücken würden: die fünfte Potenz von 26. Oder wie Mathematiker es schreiben würden: 26^5.

Kombinationen

In dem vorhergehenden Beispiel unterscheiden sich die einzelnen Sequenzen: *abcde*, *adbce*, *dbcae* usw. Allein für die Buchstaben a, b, c, d und e kommen Sie auf 5! = 120 dieser unterschiedlichen Sequenzen.

Nehmen Sie einmal an, ich füge die Einschränkung ein, dass sich eine der Folgen nicht von einer anderen unterscheidet. Mich interessiert nur, dass ich Sätze aus sich nicht wiederholenden Buchstaben ohne bestimmte Reihenfolge habe. Ein solcher Satz wird *Kombination* genannt. In diesem Beispiel ist die Anzahl der Kombinationen gleich der Anzahl der Permutationen geteilt durch 5!:

$$\frac{26!}{5!\,(21!)}$$

Im Allgemeinen lautet die Schreibweise für Kombinationen von *n* Dingen von jeweils dem Umfang *k*, $K_{k,n}$ (K steht für »Kombination«). Die Formel ist

$$K_{k,n} \frac{n!}{k!\,(n-k)!}$$

Auch hier sei der Vollständigkeit halber Folgendes erwähnt. Angenommen, ich würde in diesen Sequenzen Wiederholungen erlauben. Auf wie viele unterschiedliche Sequenzen würde ich dann kommen? Dies entspricht *N*+r–1 Dingen von jeweils dem Umfang *N*–1 oder ${}_{N+r+1}C_{N-1}$. In diesem Beispiel wäre das ${}_{30}C_{25}$.

Tabellenblattfunktionen

Excel stellt mehrere Tabellenblattfunktionen zur Verfügung, die Ihnen bei der Arbeit mit Fakultäten, Permutationen und Kombinationen helfen.

FAKULTÄT

Die Funktion FAKULTÄT, mit der Fakultäten berechnet werden, ist überraschenderweise nicht in der Statistikkategorie zu finden. Stattdessen finden Sie sie im Menü der Schaltfläche MATHEMATIK UND TRIGONOMETRIE. Die Funktion ist einfach anzuwenden. Geben Sie einfach eine Zahl ein und Sie erhalten die Fakultät zurück:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion FAKULTÄT angezeigt werden soll.**

2. **Klicken Sie auf der Registerkarte FORMELN auf MATHEMATIK UND TRIGONOMETRIE und wählen Sie im Menü der Schaltfläche FAKULTÄT aus, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE in das Feld ZAHL die Zahl ein, deren Fakultät Sie berechnen möchten.**

   Das Ergebnis wird im Dialogfeld angezeigt. Wenn ich beispielsweise die Zahl 5 eingebe, wird als Ergebnis 120 angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle zu übertragen.**

VARIATIONEN und VARIATIONEN2

Diese Funktionen finden Sie im Menü der statistischen Funktionen. Wie es der Name vermuten lässt, können Sie mit VARIATIONEN die Permutation ${}_{N}P_{r}$ berechnen. Im Folgenden beschreibe ich Ihnen, wie Sie die Funktion anwenden, um ${}_{26}P_{5}$ zu berechnen, das heißt die Anzahl der jeweils fünf Buchstaben umfassenden Sequenzen (ohne sich wiederholende Buchstaben), die Sie aus den 26 Buchstaben des Alphabets bilden können. Bei einer Permutation gilt, dass sich *abcde* von *bcdae* unterscheidet. Führen Sie diese Schritte durch:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion VARIATIONEN angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf VARIATIONEN, um das Dialogfeld FUNKTIONSARGUMENTE für diese Funktion zu öffnen (siehe Abbildung 18.1).**

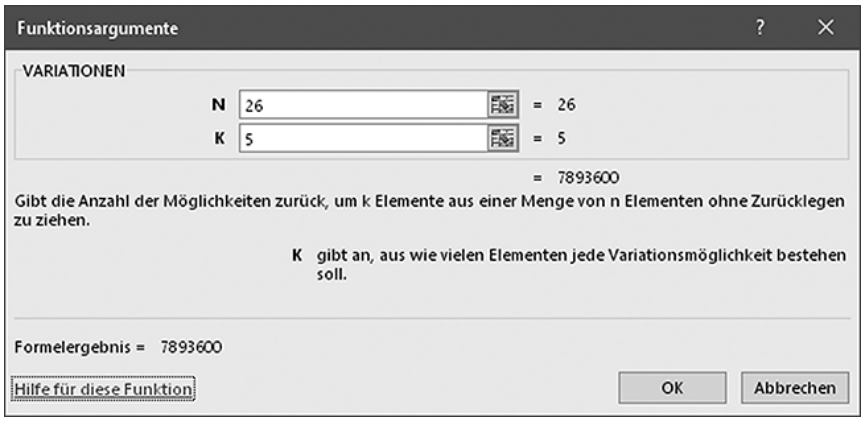


Abbildung 18.1: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion VARIATIONEN

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld habe N ich den Wert von N aus $P_{k,n}$ eingegeben. In diesem Beispiel ist das die Zahl 26.

   In das Feld K habe ich den Wert k aus $P_{k,n}$ eingegeben. In diesem Fall ist das die Zahl 5.

Nachdem beide Argumente in das Dialogfeld eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. Für unser Beispiel lautet es 7893600.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Die Funktionen VARIATIONEN2 führt die gleiche Berechnung durch, jedoch geht sie davon aus, dass Wiederholungen erlaubt sind. Die Funktion verwendet die gleichen Argumente wie VARIATIONEN und daher sieht das Dialogfeld FUNKTIONSARGUMENTE genauso aus. Das Ergebnis der Funktion entspricht N^r. Bei diesem Beispiel ist das Ergebnis übrigens 11881376.

KOMBINATIONEN und KOMBINATIONEN2

Die Funktion KOMBINATIONEN funktioniert ähnlich wie die Funktion VARIATIONEN. In Excel werden KOMBINATIONEN und KOMBINATIONEN2unter der Kategorie MATHEMATIK UND TRIGONOMETRIE aufgeführt.

Im Folgenden beschreibe ich Ihnen, wie Sie die Funktion anwenden können, um $K_{5,26}$ zu berechnen, das heißt die Anzahl der Möglichkeiten, eine fünf Buchstaben umfassende Sequenz (ohne sich wiederholende Buchstaben) aus den 26 Buchstaben des Alphabets zu bilden. Im Fall einer Kombination gilt *abcde* als äquivalent zu *bcdae*.

1. **Wählen Sie eine Zelle aus, in der das Ergebnis der Funktion KOMBINATIONEN angezeigt werden soll.**

2. **Klicken Sie im Menüband auf der Registerkarte FORMELN auf MATHEMATIK UND TRIGONOMETRIE und dann auf KOMBINATIONEN, um das Dialogfeld FUNKTIONSARGUMENTE dieser Funktion zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente der Funktion ein.**

   In das Feld N habe ich den Wert von *n* aus $K_{k,n}$ eingegeben. In diesem Fall ist das erneut die Zahl 26.

   In das Feld K habe ich den Wert von *k* aus $K_{k,n}$ eingegeben. In diesem Fall ist das erneut die Zahl 5.

   Nachdem die Werte eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. Für unser Beispiel lautet das Ergebnis 65870.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Falls Sie Wiederholungen zulassen wollen, verwenden Sie die Funktion `KOMBINATIONEN2`. Das Dialogfeld FUNKTIONSARGUMENTE sieht für diese Funktion wie das für `KOMBINATIONEN` aus, lediglich die Beschriftungen der Eingabefelder lauten anders. In diesem Beispiel entspricht die Antwort $K_{25,30}$ (142506).

Zufallsvariablen: diskret und stetig

Kehren wir zu dem Beispiel des Würfelns mit einem neutralen, »ungezinkten« Würfel zurück, bei dem sechs Elementarereignisse möglich sind. Wenn ich das Ergebnis eines Würfelwurfs mit *x* bezeichne, kann *x* jede beliebige ganze Zahl zwischen 1 und 6 sein. Da *x* also eine Reihe von Werten annehmen kann, ist es eine Variable. Da die möglichen Werte von *x* den Elementarereignissen eines Experiments entsprechen (was bedeutet, dass man die Werte nicht mit absoluter Sicherheit voraussagen kann), wird *x* auch als *Zufallsvariable* bezeichnet.

Es gibt zwei Arten von Zufallsvariablen: Eine Art ist die *diskrete* Zufallsvariable. Würfeln ist ein gutes Beispiel für diese Variablenart. Eine diskrete Zufallsvariable kann nur das annehmen, was Mathematiker gern als *zählbare* Anzahl von Werten bezeichnen – wie die Zahlen 1 bis 6. Werte, die zwischen den ganzen Zahlen 1 bis 6 liegen (z.B. 1,25 oder 3,1416) sind für eine Zufallsvariable, die Würfelergebnisse beschreibt, nicht möglich.

Die andere Art der Zufallsvariablen ist die *stetige* Zufallsvariable. Eine stetige Zufallsvariable kann eine unendliche Anzahl von Werten annehmen. Die Temperatur ist hierfür ein gutes Beispiel. Je nach Genauigkeit des Thermometers können sich Temperaturwerte wie z.B. 34,516 Grad ergeben.

Wahrscheinlichkeitsverteilungen und Dichtefunktionen

Zurück zum Würfeln. Für jeden Wert der Zufallsvariablen *x* (d. h. 1 bis 6) gibt es eine Wahrscheinlichkeit. Wenn der Würfel neutral ist, ist die Wahrscheinlichkeit für jede Zahl 1/6. Wenn man die einzelnen Werte einer diskreten Zufallsvariablen wie *x* mit den jeweiligen Wahrscheinlichkeiten kombiniert, erhält man eine *Wahrscheinlichkeitsverteilung*.

Wahrscheinlichkeitsverteilungen lassen sich gut in Diagrammen darstellen. In Abbildung 18.2 ist die Wahrscheinlichkeitsverteilung von *x* dargestellt.

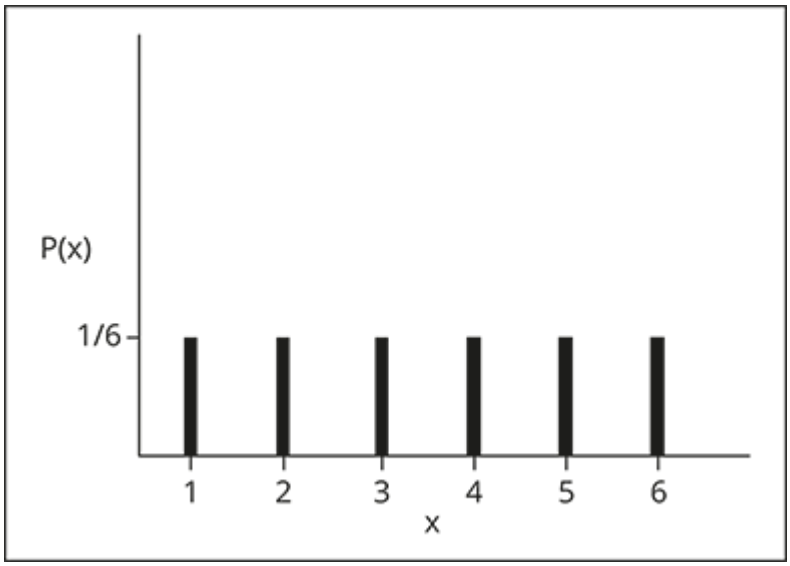


Abbildung 18.2: Die Wahrscheinlichkeitsverteilung für x, einer Zufallsvariablen, die auf den Würfen eines neutralen Würfels basiert

Eine Zufallsvariable hat einen Mittelwert, eine Varianz und eine Standardabweichung. Die Berechnung dieser Parameter ist relativ einfach. In der Welt der Zufallsvariablen heißt der Mittelwert *Erwartungswert*, und der Erwartungswert einer Zufallsvariablen x wird durch *E(x)* abgekürzt. So wird er berechnet:

$$E(x) = \sum x(P(x))$$

Für die Wahrscheinlichkeitsverteilung in Abbildung 18.2 bedeutet das

$$E(x) = \sum x(P(x)) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3{,}5$$

Die Varianz einer Zufallsvariablen wird häufig mit *Var*(*x*) abgekürzt und die entsprechende Formel lautet

$$Var(x) = \sum x^2 \ (P(x)) - \left[E(x)\right]^2$$

Wenn Sie erneut die Wahrscheinlichkeitsverteilung aus Abbildung 18.2 betrachten, ergibt sich:

$$Var(x) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) - 3{,}5^2 = 2{,}917$$

Die Standardabweichung ist die Quadratwurzel der Varianz, in diesem Fall also 1,708.

Bei stetigen Zufallsvariablen wird das Ganze etwas schwieriger. Sie können in diesem Fall einen Wert nicht mit einer Wahrscheinlichkeit kombinieren, da Sie den Wert nicht wirklich greifen können. Stattdessen assoziieren Sie eine stetige Zufallsvariable mit einer mathematischen Regel (einer Gleichung), die eine *Wahrscheinlichkeitsdichte* erzeugt. Die Funktion wird *Wahrscheinlichkeitsdichtefunktion* genannt. Zur Berechnung des Mittelwerts und der Varianz einer stetigen Zufallsvariablen benötigen Sie eine Rechenmethode.

In Kapitel 8 beschreibe ich Ihnen eine Wahrscheinlichkeitsdichtefunktion, die Standardnormalverteilung. Ich zeige Sie Ihnen an dieser Stelle noch einmal in Abbildung 18.3.

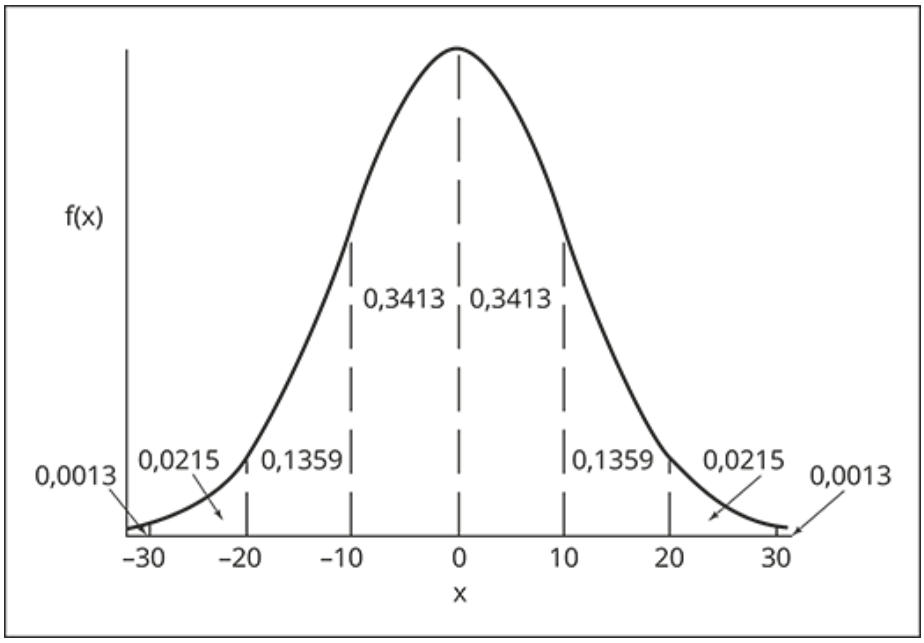


Abbildung 18.3: Die Standardnormalverteilung: eine Wahrscheinlichkeitsdichtefunktion

In der Abbildung steht *f(x)* für die Wahrscheinlichkeitsdichte. Da die Wahrscheinlichkeitsdichte einige komplexe mathematische Konzepte beinhalten kann, werde ich nicht näher darauf eingehen. Wie ich in Kapitel 8 bereits erwähnt habe, kann man sich die Wahrscheinlichkeitsdichte als etwas vorstellen, das den Bereich unterhalb der Kurve in Wahrscheinlichkeit verwandelt.

Auch wenn man nicht von der Wahrscheinlichkeit eines spezifischen Wertes einer stetigen Zufallsvariablen sprechen kann, kann man doch mit der Wahrscheinlichkeit eines Intervalls arbeiten. Um die Wahrscheinlichkeit herauszufinden, mit der die Zufallsvariable einen Wert innerhalb eines Intervalls annimmt, müssen Sie den Teil des Gesamtbereichs unter der Kurve bestimmen, der innerhalb des Intervalls liegt. Das ist in Abbildung 18.3 dargestellt. Die Wahrscheinlichkeit, dass *x* zwischen 0 und 1σ liegt, beträgt 0,3413.

Im Rest dieses Kapitels befasse ich mich nur noch mit diskreten Zufallsvariablen, von denen ich Ihnen gleich eine besondere vorstellen möchte.

Die Binomialverteilung

StellenSie sich ein Experiment mit den folgenden fünf Merkmalen vor:

- Das Experiment umfasst *N* identische Versuche.

  Ein Versuch könnte darin bestehen, zu würfeln oder eine Münze zu werfen.

- Jeder Versuch führt zu einem von zwei möglichen Elementarereignissen.

  Standardmäßig wird ein Ergebnis als *Erfolg*, das andere als *Misserfolg* bezeichnet. Beim Würfeln könnte das Ergebnis 3 beispielsweise als Erfolg gelten und alle anderen Ergebnisse demnach als Misserfolg.

- Die Wahrscheinlichkeit eines Erfolgs bleibt von Versuch zu Versuch gleich.

  Standardmäßig wird die Abkürzung *p* für die Erfolgswahrscheinlichkeit und *1-p* (oder *q*) für die Wahrscheinlichkeit eines Misserfolgs verwendet.

- Die Versuche sind voneinander unabhängig.

- Die diskrete Zufallsvariable *x* ist die Anzahl der Erfolge in den *N* Versuchen.

Diese Art von Experiment wird als *Binomialexperiment* bezeichnet. Die Wahrscheinlichkeitsverteilung für *x* folgt dieser Regel:

$$P(x) = \frac{N!}{x!(N-x)!}\, p^x\,(1-p)^{N-x}$$

Der Teil ganz rechts, $p^x(1-p)^{N-x}$, ist die Wahrscheinlichkeit einer Kombination von *x* Erfolgen in *N* Versuchen. Der Ausdruck links daneben ist ${}_NK_x$, die Anzahl der möglichen Kombinationen von *x* Erfolgen in *N* Versuchen.

Das Ganze nennt man *Binomialverteilung*. Mit ihrer Hilfe können Sie bestimmen, wie hoch die Wahrscheinlichkeit ist, vier Mal eine 3 zu bekommen, wenn Sie zehn Mal würfeln:

$$P(4) = \frac{10!}{4!(6)!}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6 = 0{,}054$$

Die *negative Binomialverteilung* steht damit in einem engen Zusammenhang. Bei einer solchen Verteilung ist die Zufallsvariable die Anzahl der Versuche vor dem x-ten Erfolg. Sie verwenden die negative Binomialverteilung beispielsweise, um die Wahrscheinlichkeit von fünf Würfen zu bestimmen, die zu keiner 3 führen, bevor Sie zum vierten Mal eine 3 würfeln.

Dazu müssen Sie in den acht Würfen vor der vierten 3 fünf Würfe, die keine 3 ergeben, sowie drei Erfolge (Würfe, die 3 Augen ergeben) haben. Dann wird beim nächsten Wurf eine 3 gewürfelt. Die Wahrscheinlichkeit für eine Kombination aus vier Erfolgen und fünf Misserfolgen ist $p^4(1-p)^5$. Die Anzahl der Möglichkeiten für eine Kombination aus fünf Misserfolgen und 4-1 Erfolgen ist ${}_{5+4-1}K_{4-1}$. Die Wahrscheinlichkeit ist somit

$$P(5 \text{ Misserfolge vor dem vierten Erfolg}) = \frac{(5+4-1)!}{4-1!(5)!}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^5 = 0{,}017$$

Die allgemeine Gleichung für die negative Binomialverteilung (manchmal auch *Pascalverteilung* genannt) lautet

$$P(f \text{ Misserfolge vor dem } x \text{ ten Erfolg}) = \frac{(f+x-1)!}{(x-1)!(f)!}p^x(1-p)^f$$

Tabellenblattfunktionen

Diese Verteilungen sind recht rechenintensiv, weshalb ich jetzt direkt zu den zugehörigen Tabellenblattfunktionen komme.

BINOM.VERT und BINOM.VERT.BEREICH

Dies sind die Excel-Tabellenblattfunktionen für die Binomialverteilung. Als Beispiel verwende ich `BINOM.VERT` für die Berechnung der Wahrscheinlichkeit, bei zehn Würfen mit einem ungezinkten Würfel vier Mal eine 3 zu würfeln:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion `BINOM.VERT` angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf BINOM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 18.4).**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld ZAHL_ERFOLGE habe ich die Anzahl der Erfolge eingegeben. In unserem Beispiel ist die Anzahl der Erfolge 4.

   In das Feld VERSUCHE habe ich die Anzahl der Versuche eingegeben. Die Anzahl der Versuche ist 10.

   In das Feld ERFOLGSWAHRSCH habe ich die Erfolgswahrscheinlichkeit eingegeben. Ich habe 1/6 eingegeben, die Wahrscheinlichkeit einer 3 beim Würfeln mit einem neutralen Würfel.

   Geben Sie in das Feld KUMULIERT den Wert `FALSCH` ein, wenn Sie die Wahrscheinlichkeit errechnen wollen, exakt die Anzahl von Erfolgen zu erzielen, die im Feld ZAHL_ERFOLGE angegeben wurde. Geben Sie `WAHR` für die Wahrscheinlichkeit ein, genau diese Anzahl von Erfolgen oder aber auch weniger zu erzielen. Ich habe `FALSCH` eingegeben.

   Nachdem alle Argumente eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Abbildung 18.4: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion BINOM.VERT

Um Ihnen einen besseren Eindruck davon zu geben, wie eine Binomialverteilung aussieht, verwende ich BINOM.VERT (mit der Angabe FALSCH im Feld KUMULIERT), um *P*(0) bis *P*(10) zu bestimmen. Anschließend verwende ich die Diagrammfeatures von Excel (siehe Kapitel 3), um die Ergebnisse zu visualisieren. Die Daten und das Diagramm sind in Abbildung 18.5 dargestellt.

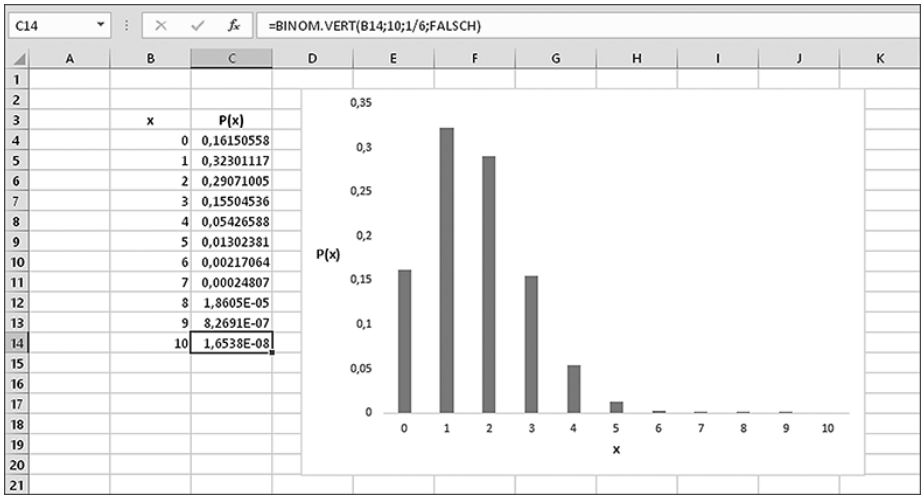


Abbildung 18.5: Die Binomialverteilung für x Erfolge bei zehn Würfen eines Würfels, mit p = 1/6

Wenn Sie übrigens WAHR in das Feld KUMULIERT eingeben, erhalten Sie als Ergebnis 0,984 (und ein paar weitere Dezimalstellen), was *P*(0) + *P*(1) + *P*(2) + *P*(3) + *P*(4) entspricht.

Abbildung 18.5 ist hilfreich, wenn Sie die Wahrscheinlichkeit dafür ermitteln wollen, dass bei zehn Versuchen zwischen 4 und 6 Erfolge eintreten. Ermitteln Sie *P*(4), *P*(5) und *P*(6) und addieren Sie die Ergebnisse.

Ein einfacherer Weg, besonders dann, wenn Ihnen kein Diagramm wie in Abbildung 18.5 zur Verfügung steht oder wenn Sie die Funktion BINOM.VERT nicht drei Mal verwenden wollen, ist die Funktion BINOM.VERT.BEREICH zu verwenden. Abbildung 18.6 zeigt das Dialogfeld für diese Funktion; die Werte für die Argumente

sind bereits eingegeben. Nachdem alle Argumente eingegeben wurden, wird das Ergebnis (0,069460321) im Dialogfeld angezeigt.

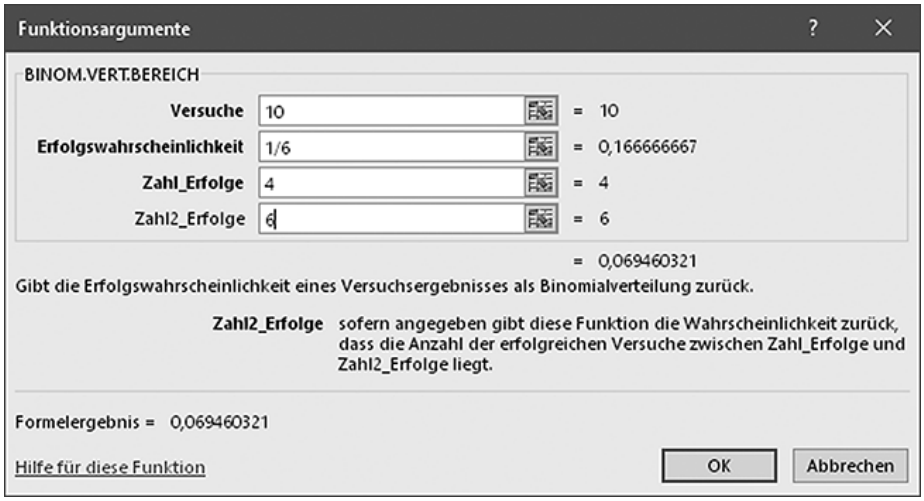


Abbildung 18.6: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `BINOM.VERT.BEREICH`

Wenn Sie in das Feld ZAHL2_ERFOLGE keinen Wert eingeben, gibt `BINOM.VERT.BEREICH` die Wahrscheinlichkeit für die Anzahl der Erfolge zurück, die Sie in ZAHL_ERFOLGE eingegeben haben. Wenn Sie in ZAHL_ERFOLGE keinen Wert eingeben, gibt die Funktion die Wahrscheinlichkeit dafür zurück, dass maximal die Anzahl der in ZAHL2_ERFOLGE durchgeführten Versuche erfolgreich ist (beispielsweise die kumulative Wahrscheinlichkeit).

NEGBINOM.VERT

Wie derName der Funktion bereits andeutet, befasst sich `NEGBINOM.VERT` mit der negativen Binomialverteilung. Ich verwende die Funktion an dieser Stelle, um das oben genannte Beispiel – die Wahrscheinlichkeit von fünf Misserfolgen (Würfen, die nicht zu einer 3 führen) vor dem vierten Erfolg (der vierten 3) – zu bearbeiten. Hier sind die Schritte:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion `NEGBINOM.VERT` angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf NEGBINOM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 18.7).**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld ZAHL_MISSERFOLGE habe ich die Anzahl der Misserfolge eingegeben. In unserem Beispiel ist die Anzahl der Misserfolge **5**.

   In das Feld ZAHL_ERFOLGE habe ich die Anzahl der Erfolge eingegeben. In diesem Beispiel sind dies **4**.

   In das Feld ERFOLGSWAHRSCH habe ich die Erfolgswahrscheinlichkeit eingegeben, diese ist **1/6**.

   In das Feld KUMULIERT habe ich `FALSCH` eingegeben. Hierdurch erhalte ich die Wahrscheinlichkeit für die Anzahl der Erfolge (den Wert der Dichtefunktion). Wenn ich `WAHR` eingebe, ist das Ergebnis die Wahrscheinlichkeit für höchstens diese Anzahl an Erfolgen (der Wert der Verteilungsfunktion).

   Nachdem ich für alle Argumente die Werte eingegeben habe, wird das Ergebnis im Dialogfeld angezeigt. Das Ergebnis lautet 0,017 plus einige weitere Dezimalstellen.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Abbildung 18.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion NEGBINOM.VERT

Hypothesen mit der Binomialverteilung testen

Das Testen von Hypothesen beinhaltet manchmal die Binomialverteilung. Normalerweise haben Sie eine bestimmte Vorstellung von der Wahrscheinlichkeit eines Erfolges und legen diese in einer Nullhypothese nieder. Dann führen Sie *N* Versuche durch und dokumentieren dabei die Anzahl der Erfolge. Zum Schluss berechnen Sie, wie hoch die Wahrscheinlichkeit ist, diese Anzahl von Erfolgen oder eine noch höhere zu erzielen, wenn Ihre Nullhypothese H_0 wahr ist. Ist die Wahrscheinlichkeit gering, verwerfen Sie H_0.

Wenn Sie eine Hypothese auf diese Weise testen, verwenden Sie Stichproben, um eine Schlussfolgerung über einen Grundgesamtheitsparameter zu ziehen. Der Parameter ist in diesem Fall die Wahrscheinlichkeit eines Erfolges in der Grundgesamtheit von Versuchen. Parameter werden standardmäßig durch griechische Buchstaben dargestellt. Statistiker verwenden π (pi), die griechische Entsprechung des Buchstabens *p*, für die Darstellung eines Erfolges in der Grundgesamtheit.

Wir bleiben bei unserem Würfelbeispiel. Stellen Sie sich vor, Sie haben einen Würfel und möchten prüfen, ob er sich neutral verhält oder nicht. Sollte er nicht neutral sein, dann vermuten Sie, dass er tendenziell häufiger eine 3 würfelt. Definieren Sie einen Wurf, der als Erfolg zu einer 3 führt. Wenn Sie zehn Mal würfeln und bei vier Würfen fällt eine 3, sind vier Würfe also als Erfolge zu werten. Drücken wir all das nun als Test einer Hypothese aus:

H_0: $\pi \leq 1/6$
H_1: $\pi > 1/6$

Wie gehabt, setze ich $\alpha = 0{,}05$.

Zum Testen dieser Hypothesen müssen Sie herausfinden, wie hoch die Wahrscheinlichkeit ist, bei zehn Würfen mit $p = 1/6$ mindestens vier Erfolge zu erzielen. Diese Wahrscheinlichkeit beträgt $P(4) + P(5) + P(6) + P(7) + P(8) + P(9) + P(10)$. Wenn das Gesamtergebnis weniger als 0,05 beträgt, verwerfen Sie H_0.

Das alles erfordert großen Rechenaufwand. Sie können stattdessen auch die Funktion `BINOM.VERT` verwenden (wie ich es getan habe, als ich das Arbeitsblatt in Abbildung 18.5 erstellt habe) oder einen ganz anderen Weg einschlagen. Sie können einen kritischen Wert für die Anzahl der Erfolge bestimmen. Wenn die Anzahl der Erfolge dann größer als der kritische Wert ist, verwerfen Sie H_0.

Möchten Sie wissen, wie Sie den kritischen Wert bestimmen können? Na, z. B. mit einer praktischen Tabellenblattfunktion, die ich Ihnen gleich vorstellen werde.

BINOM.INV

Diese Funktion ist perfekt für das Testen von Hypothesen auf Binomialbasis geeignet. Füttern Sie `BINOM.INV` mit der Anzahl der Versuche, der Erfolgswahrscheinlichkeit und der Grenzwahrscheinlichkeit Alpha. `BINOM.INV` gibt den kleinsten Wert von *x* (der Anzahl der Erfolge) zurück, bei dem die kumulierte Wahrscheinlichkeit größer oder gleich der Grenzwahrscheinlichkeit ist.

Hier die Schritte, die Sie für das Testen der Hypothese in unserem Beispiel durchführen müssen:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion `BINOM.INV` angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf BINOM.INV, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 18.8).**

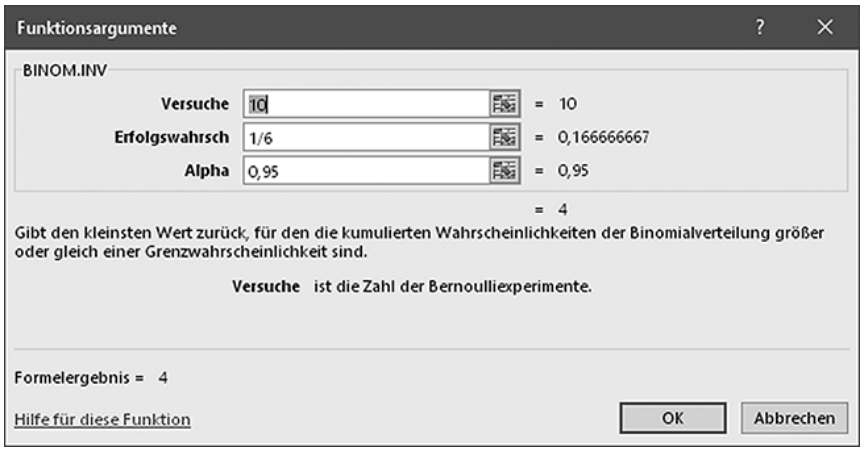


Abbildung 18.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `BINOM.INV`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld VERSUCHE habe ich **10** eingegeben, die Anzahl der Versuche.

   In das Feld ERFOLGSWAHRSCH habe ich die Erfolgswahrscheinlichkeit eingegeben. In diesem Beispiel ist es 1/6, der Wert von π gemäß H_0.

   In das Feld ALPHA habe ich die kumulierte Wahrscheinlichkeit eingegeben, die ich übertreffen möchte. Ich habe 0,95 eingegeben, da ich den kritischen Wert bestimmen möchte, der die oberen 5 Prozent der Binomialverteilung abtrennt.

Nachdem für alle Argumente die Werte eingegeben wurden, wird der kritische Wert, 4, im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Der kritische Wert entspricht der Anzahl der Erfolge in der Stichprobe. Die Entscheidung ist, H_0 zu verwerfen.

Weitere Informationen zum Testen von Hypothesen

In manchen Fällen approximiert die Binomialverteilung die Standardnormalverteilung. Wenn das der Fall ist, nutzen Sie die Statistik der Normalverteilung, um Fragen zur Binomialverteilung zu beantworten.

Diese Statistiken beinhalten so genannte *Z-Scores*, was bedeutet, dass Sie den Mittelwert und die Standardabweichung der Binomialverteilung kennen müssen. Zum Glück können sie einfach berechnet werden. Wenn N die Anzahl der Versuche und π die Erfolgswahrscheinlichkeit ist, dann ist der Mittelwert

$$\mu = N\pi$$

Die Varianz ist

$$\sigma^2 = N\pi(1-\pi)$$

und die Standardabweichung ist

$$\sigma = \sqrt{N\pi(1-\pi)}$$

Die binomiale Näherung an die Normalverteilung ist geeignet, wenn $N\pi \geq 5$ und $N(1-\pi) \geq 5$.

Wenn Sie eine Hypothese testen, ziehen Sie eine Schlussfolgerung über π und müssen mit einem Schätzwert beginnen. Sie führen *N* Versuche durch und erzielen *x* Erfolge. Der Schätzwert ist

$$P = \frac{x}{N}$$

Für die Erstellung eines Z-Scores benötigen Sie eine weitere Information – den Standardfehler von *P*. Das klingt komplizierter, als es ist, denn der Standardfehler ist lediglich

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{N}}$$

Jetzt können Sie eine Hypothese testen.

Hier ein Beispiel: Der Geschäftsführer von FarKlempt Robotics glaubt, dass 50 Prozent der FarKlempt-Roboter für den Einsatz zu Hause gekauft werden. Eine Stichprobe unter 1.000 FarKlempt-Kunden ergibt, dass 550 von ihnen ihre Roboter zu Hause einsetzen. Weicht das signifikant von dem ab, was der Geschäftsführer glaubt? Die Hypothesen lauten:

H_0: π = 0.50
H_1: π ≠ 0.50

Ich setze α = 0,05

$N\pi$ = 500 und $N(1-\pi)$ = 500, also ist die Normalnäherung geeignet.

Berechnen Sie zunächst P:

$$P = \frac{x}{N} = \frac{550}{1000} = 0,55$$

Jetzt bilden Sie einen Z-Wert:

$$z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{N}}} = \frac{0{,}55 - 0{,}50}{\sqrt{\frac{(0{,}50)(1-0{,}50)}{1000}}} = \frac{0{,}05}{\sqrt{\frac{0{,}25}{1000}}} = 3{,}162$$

Ist der Z-Wert von 3,162 mit α = 0,05 groß genug, um H_0 zu verwerfen? Das können Sie auf einfache Art und Weise mit der Tabellenblattfunktion `NORM.S.VERT` (siehe Kapitel 8) überprüfen. Wenn Sie das tun, werden Sie herausfinden, dass dieser Z-Score weniger als 0,01 des Bereichs am oberen Ende der Standardnormalverteilung abtrennt. Die Entscheidung lautet daher, H_0 zu verwerfen.

Die hypergeometrische Verteilung

Hier ist eine weitere Verteilung, die sich mit Erfolgen und Misserfolgen befasst.

Beginnen wir mit einem Beispiel. Ein Satz von 16 Glühbirnen enthält neun funktionierende Glühbirnen und sieben defekte. Wenn Sie zufällig sechs Glühbirnen aus diesen 16 auswählen, wie hoch ist dann die Wahrscheinlichkeit, dass drei dieser sechs Glühbirnen intakt sind? Die Auswahl einer intakten Glühbirne wird als »Erfolg« betrachtet.

Nach Ihrer Auswahl enthält der von Ihnen ausgewählte Satz eine Kombination aus drei der intakten neun Glühbirnen und eine Kombination aus drei der defekten sieben Glühbirnen. Die Wahrscheinlichkeit, drei intakte Glühbirnen zu ziehen, ist daher eine Kombination aus Zählregeln:

$$P(3) = \frac{({}_9K_3)({}_7K_3)}{{}_{16}K_6} = \frac{(84)(35)}{8008} = 0{,}37$$

Jedes Ergebnis der Auswahl der intakten Glühbirnen kann mit jedem Ergebnis der Auswahl der defekten Glühbirnen zusammenhängen,

weshalb im Zähler die Produktregel angewendet werden kann. Der Nenner (der Wahrscheinlichkeitsraum) ist die Anzahl der möglichen Kombinationen von sechs Elementen in einer Gruppe von 16 Elementen.

Dies ist ein Beispiel für eine *hypergeometrische Verteilung*. Bei einer kleinen Grundgesamtheit, die aus N_1 Erfolgen und N_2 Misserfolgen besteht, ist die allgemeine Wahrscheinlichkeit von *x* Erfolgen in einer Stichprobe von *m* Elementen

$$P(x) = \frac{({}_{N_1}K_x)({}_{N_2}K_{m-x})}{{}_{N_1+N_2}K_m}$$

Die Zufallsvariable *x* ist eine *hypergeometrisch verteilte Zufallsvariable*.

HYPGEOM.VERT

Diese Funktion übernimmt für Sie sämtliche Berechnungen, wenn Sie sich mit der hypergeometrischen Verteilung beschäftigen. Im Folgenden beschreibe ich, wie Sie die Funktion anwenden können, um das obige Beispiel zu bearbeiten:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion** `HYPGEOM.VERT` **angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf HYPGEOM.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 18.9).**

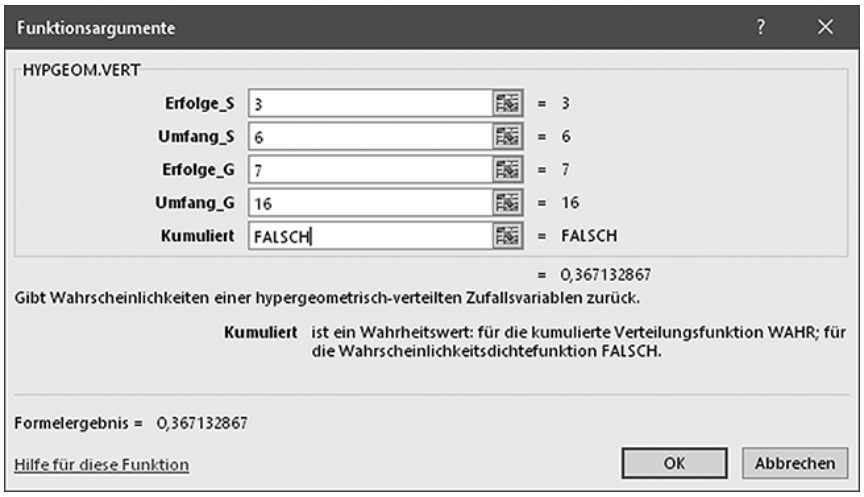


Abbildung 18.9: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion HYPGEOM.VERT

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Argumente für die Funktion ein.**

   In das Feld ERFOLGE_S habe ich die Anzahl der Erfolge in der Stichprobe eingegeben. In diesem Beispiel beträgt die Anzahl 3.

   In das Feld UMFANG_S habe ich die Anzahl der Elemente in der Stichprobe eingegeben. In diesem Beispiel ist der Umfang der Stichprobe 6.

   In das Feld ERFOLGE_G habe ich die Anzahl der Erfolge in der Grundgesamtheit eingegeben. In diesem Beispiel sind es 7, die Anzahl der intakten Glühbirnen.

   In das Feld UMFANG_G habe ich die Anzahl der Elemente in der Grundgesamtheit ein. In diesem Beispiel beträgt die Anzahl 16, das heißt die Gesamtzahl der Glühbirnen.

   In das Feld KUMULIERT habe ich `FALSCH` eingegeben. Hierdurch erhalte ich die Wahrscheinlichkeit für die Anzahl der Erfolge, die ich in das Feld ERFOLGE_S eingegeben habe (den Wert der Wahrscheinlichkeitsdichtefunktion). Wenn ich hier `WAHR` eingebe, gibt die Funktion die Wahrscheinlichkeit für maximal

diese Anzahl der Erfolge zurück (den Wert der kumulierten Verteilungsfunktion).

Nachdem für alle Argumente Werte eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. Es lautet 0,367, gefolgt von einigen weiteren Dezimalstellen.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Wie bei der Binomialverteilung verwende ich die Funktion `HYPERGEOM.VERT` zur Berechnung von *P*(0) bis *P*(6). Dann verwende ich die Diagrammfeatures von Excel (siehe Kapitel 3), um das Ergebnis zu visualisieren. Die Daten und das entsprechende Diagramm sind in Abbildung 18.10 dargestellt. Ich führe Ihnen die hypergeometrische Verteilung nur deshalb bildlich vor, damit Sie sie besser verstehen können.

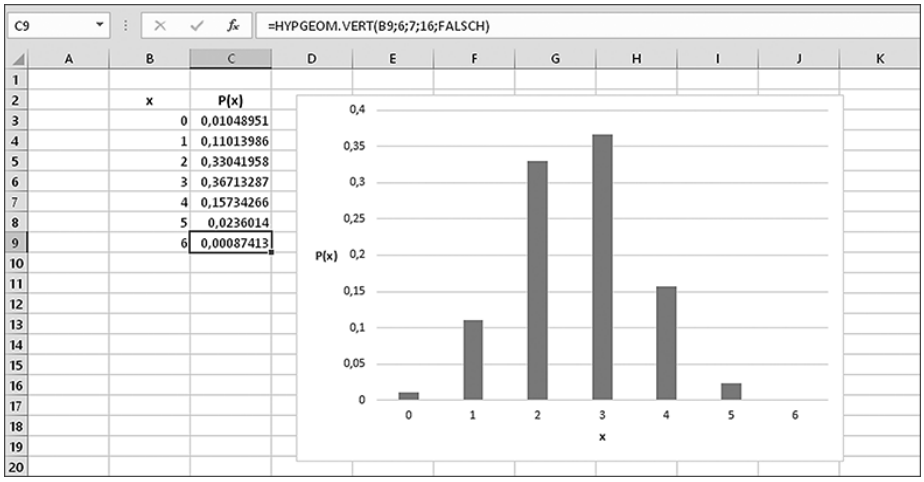


Abbildung 18.10: Die hypergeometrische Verteilung von *x* Erfolgen in einer Stichprobe von sechs Elementen aus einer Grundgesamtheit, die aus sieben Erfolgen und neun Misserfolgen besteht

Kapitel 19

Mehr zur Wahrscheinlichkeit

In diesem Kapitel

- Die Betaverteilung
- Die Poissonverteilung
- Die Gammaverteilung
- Exponentiell gesprochen

In Kapitel 18 habe ich die Wahrscheinlichkeit halb formal beschrieben und die Verteilung von Zufallsvariablen eingeführt. Die Binomialverteilung steht am Anfang. In diesem Kapitel wende ich mich weiteren Verteilungen zu.

Eines der Icons auf den Seiten dieses Buches (und in anderen Büchern der *Dummies*-Reihe) weist Sie darauf hin, dass es um »technischen Kram« geht. Es wäre vielleicht angebracht gewesen, dieses Icon gleich ganz oben noch über der Überschrift dieses Kapitels anzubringen. Hier also eine kleine Warnung: Es folgt eine Menge Mathematik. Ich habe die mathematischen Konzepte eingefügt, damit Sie verstehen, was Sie überhaupt machen, wenn Sie die Dialogfelder der beschriebenen Excel-Funktionen bearbeiten.

Sind diese Funktionen ein wenig esoterisch angehaucht? Nun … ja. Werden Sie je die Gelegenheit haben, sie anzuwenden? Nun … möglicherweise.

Betaverteilung

Die Betaverteilung (nicht zu verwechseln mit »beta«, der Wahrscheinlichkeit eines Typ-2-Fehlers) ist im Reich der Verteilungen so eine Art Chamäleon. Es gibt sie je nach Zusammenhang in den verschiedensten Erscheinungsformen. Ich werde Sie nicht mit sämtlichen mathematischen Konzepten belästigen, die hinter der Betaverteilung stehen, da eine umfassende Behandlung einigen Rechenaufwand erfordert.

Die Betaverteilung und die Binomialverteilung, die ich in Kapitel 18 beschrieben habe, hängen zusammen. Der Zusammenhang mit der Binomialverteilung ist wie folgt: In der Binomialverteilung ist die Zufallsvariable *x* die Anzahl der Erfolge in *N* Versuchen. *p* ist dabei die Erfolgswahrscheinlichkeit. *N* und *p* sind Konstanten. In der Betaverteilung ist die Zufallsvariable *x* die Erfolgswahrscheinlichkeit und *N* und die Anzahl der Erfolge sind Konstanten.

Warum ist das nützlich? Im echten Leben kennen Sie den Wert von *p* für gewöhnlich nicht und versuchen, ihn herauszufinden. Normalweise führen Sie eine Untersuchung durch, bestimmen die Anzahl der Erfolge in einer Reihe von Versuchen und müssen *p* dann schätzen. Die Betaverteilung liefert Ihnen die Wahrscheinlichkeit möglicher Werte von *p* für die Anzahl der Versuche und Erfolge in Ihrer Untersuchung.

Einige der daran beteiligten mathematischen Konzepte sind kompliziert, aber ich kann Ihnen zumindest die Regel zeigen, durch die die Dichtefunktion für *N* Versuche mit *r* Erfolgen erzeugt wird, wenn *N* und *r* ganze Zahlen sind:

$$f(x \mid r, N) = \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1}$$

Der senkrechte Strich in der Klammer links bedeutet »für«. Diese Dichtefunktion steht also für spezifische Werte von *N* und *r*. Rechenintensiv wird es, wenn *N* und *r* keine ganzen Zahlen sind. (Dichtefunktion? »Für?« Werfen Sie einen Blick in Kapitel 18.)

Um Ihnen einen Eindruck davon zu geben, wie diese Funktion aussieht, verwende ich Excel, um die Dichtefunktion für vier Erfolge

in zehn Versuchen zu erzeugen und grafisch darzustellen. In Abbildung 19.1 sind die Daten und das zugehörige Diagramm aufgeführt. Jeder Wert auf der *x*-Achse ist ein möglicher Wert für die Erfolgswahrscheinlichkeit. Die Kurve stellt die Wahrscheinlichkeitsdichte dar. Wie ich in Kapitel 18 bereits dargestellt habe, sorgt die Wahrscheinlichkeitsdichte dafür, dass der Bereich unter der Kurve der Wahrscheinlichkeit entspricht. Der höchste Punkt der Kurve liegt bei $x = 4$. Das ist der Wert, den Sie für vier Erfolge in zehn Versuchen erwarten können. Ich habe Zelle C3 markiert, damit Sie in der Bearbeitungsleiste sehen können, wie ich mit `BETA.VERT` die Wahrscheinlichkeiten berechnet habe.

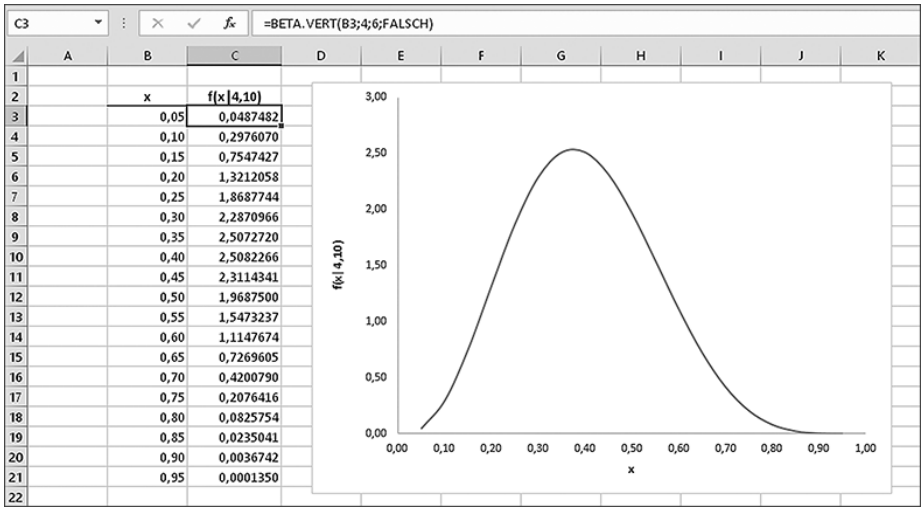


Abbildung 19.1: Die Betadichtefunktion für vier Erfolge in zehn Versuchen

Stellen Sie sich vor, ich würfele (mit nur einem Würfel) und definiere jeden Wurf als Erfolg, der eine 3 ergibt. Ich gehe davon aus, dass sich der Würfel neutral verhält, weshalb ich annehmen kann, dass $p = P(3) = 1/6$ ist. Ich würfele zehn Mal und erhalte vier Mal die 3. Was sagt das über die Annahme aus, der Würfel verhalte sich neutral?

Das Diagramm in Abbildung 19.1 liefert Ihnen einen Hinweis. Der Bereich links von 0,16667 (der Dezimalentsprechung von 1/6) ist ein recht kleiner Teil des Gesamtbereichs, was bedeutet, dass die Wahrscheinlichkeit, dass p 1/6 oder weniger ist, ziemlich gering ist.

Nun, wenn Sie den ganzen Aufwand der Diagrammerstellung auf sich nehmen, um dann Bereichsanteile mehr oder weniger ungenau

abzuschätzen und als Ergebnis herausbekommen, dass die Wahrscheinlichkeit »ziemlich gering« ist, dann rechtfertigt das ungenaue Ergebnis kaum den ganzen Arbeitsaufwand. Zum Glück kann Excel das besser.

BETA.VERT

Die Funktion BETAVERT entlastet Sie von der ganzen Diagrammerstellung und ungenauen Schätzarbeit. Mit ihr können Sie die kumulative Betaverteilung nutzen, um festzustellen, wie hoch die Wahrscheinlichkeit ist, dass *p* kleiner oder gleich einem bestimmten Wert ist. Wenn man bedenkt, wie komplex die Betaverteilung ist, ist die Funktion `BETA.VERT` überraschend einfach in der Anwendung.

Im Dialogfeld FUNKTIONSARGUMENTE der Funktion `BETA.VERT` und in der Hilfe zu `BETA.VERT` sehen Sie die Felder ALPHA und BETA. Im Dialogfeld steht, dass sie Parameter der Verteilung sind. Wie dem auch sei, eine große Hilfe ist diese Definition nicht – zumindest nicht für die Anwendung von Alpha und Beta.

Hier also nun die Grundbestandteile: Für das Beispiel, das ich gerade bearbeite, ist Alpha die Anzahl der Erfolge und Beta die Anzahl der Misserfolge.

Wenn Sie die Dichtefunktion mit Alpha (α) und Beta (β) ausdrücken, lautet sie

$$f(x) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$$

Auch das gilt nur, wenn α und β ganze Zahlen sind. Ist das nicht der Fall, müssen Sie *f(x)* mathematisch bestimmen.

Führen Sie die folgenden Schritte durch:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion BETA.VERT angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf BETA.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 19.2).**

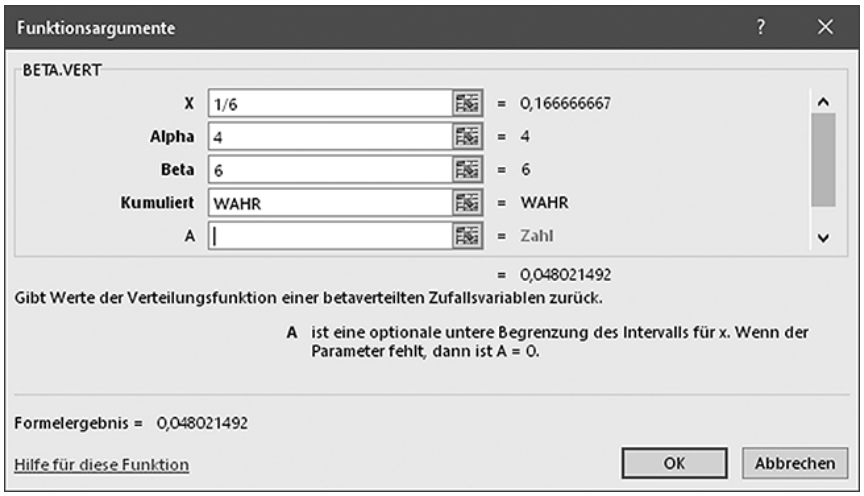


Abbildung 19.2: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion BETA.VERT

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Im Feld x wird die Erfolgswahrscheinlichkeit eingegeben. In unserem Beispiel beträgt die Erfolgswahrscheinlichkeit 1/6.

   Excel bezeichnet ALPHA und BETA (Beta kommt im nächsten Schritt) als »Parameter der Verteilung«. Ich behandele sie als »Anzahl der Erfolge« und »Anzahl der Misserfolge«. Also habe ich **4** in das Feld ALPHA und **6** in das Feld BETA eingegeben.

   In das Feld KUMULIERT habe ich **WAHR** eingegeben. Damit erhalte ich den Wert der Verteilungsfunktion, also den Bereich, der zwischen 0 und 1/6 liegt. Wenn ich hier FALSCH eingebe, erhalte ich die Höhe der Verteilungsfunktion für diesen Wert von X. Wie Sie in Abbildung 19.1 sehen können, habe ich beim Erstellen des Diagramms das Argument FALSCH benutzt.

Das Feld A ist eine Auswertungsgrenze für den Wert im Feld X. Hiermit ist die untere Grenze des Intervalls von X gemeint. Da er für das vorliegende Beispiel nicht relevant ist, habe ich das Feld leer gelassen, wodurch es automatisch mit dem Wert 0 vorbelegt wird. Ein weiteres optionales Argument ist B; hier können Sie die obere Grenze des Intervalls von X eintragen. Wenn dieser Wert fehlt, wird der Standardwert 1 verwendet.

Nachdem alle Werte eingetragen sind, wird das Ergebnis im Dialogfeld angezeigt.

Das Ergebnis in diesem Beispiel ist 0,048021492. Das ist wirklich »ziemlich gering«. Bei vier Erfolgen in sechs Würfen gehen Sie intuitiv davon aus, dass *p* größer ist als 1/6.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Der Anwendungsbereich der Betaverteilung ist natürlich größer, als ich Ihnen hier gezeigt habe. Folglich können Sie nahezu alle Arten von Zahlen in die verschiedenen Felder eingeben. Der Wert, den Sie in das Feld X eingeben, kann zum Beispiel größer als 1,00 sein. Sie können außerdem auch Werte, die keine ganzen Zahlen sind, in die Felder ALPHA und BETA eingeben.

BETA.INV

Diese Funktion ist die Umkehrung von `BETA.VERT`. Wenn Sie eine Wahrscheinlichkeit und Werte für Erfolge und Misserfolge eingeben, gibt die Funktion einen Wert für *p* zurück. Wenn Sie beispielsweise 0,048021492, vier Erfolge und sechs Misserfolge eingeben, lautet das Ergebnis 0,1666667, das Dezimaläquivalent zu 1/6.

`BETA.INV` kann für etwas Nützlicheres verwendet werden. Sie können die Funktion dazu nutzen, die

Vertrauensgrenzen für die Erfolgswahrscheinlichkeit zu bestimmen.

Nehmen wir einmal an, es gab *r* Erfolge in *N* Versuchen. Sie interessieren sich nun für die Vertrauensgrenzen von 95 Prozent für die Wahrscheinlichkeit eines Erfolgs. Die untere Grenze ist:

```
BETA.INV(0,025, r, N-r)
```

Die obere Grenze ist:

```
BETAINV(0,975, r, N-r)
```

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion `BETA.INV` angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf BETA.INV, , um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 19.3).**

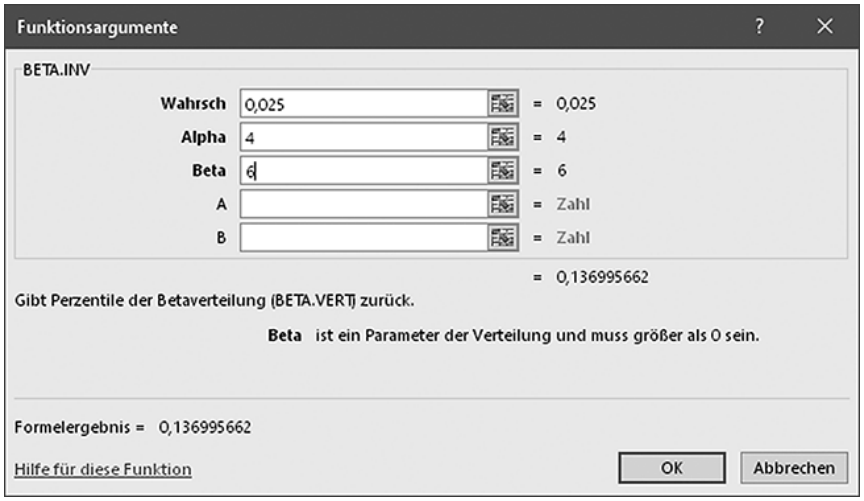


Abbildung 19.3: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `BETA.INV`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Feld WAHRSCH wird die kumulative Wahrscheinlichkeit eingegeben. Für die untere Grenze der Vertrauensgrenzen von 95 Prozent beträgt die Wahrscheinlichkeit 0,025.

In das Feld ALPHA habe ich die Anzahl der Erfolge eingegeben. In diesem Beispiel sind dies 4.

In das Feld BETA habe ich die Anzahl der Misserfolge (*nicht* die Anzahl der Versuche) eingegeben. In diesem Beispiel ist die Anzahl der Misserfolge 6.

Die Felder A und B sind Auswertungsgrenzen für den Wert im Feld X. Da sie für die vorliegende Art von Beispiel nicht relevant sind, habe ich sie leer gelassen, wodurch die Felder automatisch wie folgt vorbelegt werden: A = 0 und B = 1.

Nachdem die Werte für WAHRSCH, ALPHA und BETA eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. In diesem Beispiel ist das Ergebnis 0,13699536.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle zu übertragen.**

Die Eingabe von 0,975 in das Feld WAHRSCH liefert als Ergebnis die Zahl 0,700704575. Die Vertrauensgrenzen von 95 Prozent für die Wahrscheinlichkeit eines Erfolgs sind also 0,137 und 0,701 (gerundet), wenn Sie in zehn Versuchen vier Erfolge erzielen.

Bei einer höheren Anzahl von Versuchen werden die Vertrauensgrenzen enger. Bei 40 Erfolgen in 100 Versuchen betragen die Vertrauensgrenzen 0,307 und 0,497.

Poissonverteilung

Wenn Sie sich mit einem Prozess beschäftigen, der eine Binomialverteilung erzeugt, und Sie eine extrem große Anzahl von Versuchen und eine sehr kleine Anzahl von Erfolgen haben, ist die *Poissonverteilung* eine Näherung der Binomialverteilung. Die Gleichung der Poissonverteilung lautet

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

μ steht im Zähler für die mittlere Anzahl von Erfolgen in den Versuchen und *e* ist 2,71828 (sowie eine unendliche Zahl weiterer Dezimalstellen), eine Konstante, die den Mathematikern lieb und teuer ist.

Hier ein Beispiel: Die FarKlempt Robotics GmbH stellt ein Universalgelenk für die Ellbogen der Roboter des Unternehmens her. Der Produktionsprozess unterliegt strenger Computerkontrolle, so dass die Wahrscheinlichkeit eines defekten Gelenks 0,001 beträgt. Wie hoch ist die Wahrscheinlichkeit, dass in einer Stichprobe von 1.000 Gelenken ein Gelenk defekt ist? Wie hoch ist die Wahrscheinlichkeit, dass zwei Gelenke defekt sind? Oder drei?

Diese Verteilung, die nach dem im 19. Jahrhundert lebenden Mathematiker Siméon-Denis Poisson benannt wurde, ist vom Rechenaufwand her einfacher als die Binomialverteilung – oder das war sie zumindest einmal, als die Mathematiker noch keine Computerhilfen hatten. Mit Excel können Sie die Funktion `BINOM.VERT` einfach für die Durchführung binomialer Berechnungen anwenden.

Zuerst wende ich die Poissonverteilung auf das FarKlempt-Beispiel an. Wenn π = 0,001 und *N* = 1000, dann beträgt der Mittelwert

$\mu = N\pi = (1000)\,(0{,}001) = 1$

(Eine Erklärung zu $\mu = N\pi$ finden Sie in Kapitel 16.)

Nun zur Poissonverteilung. Die Wahrscheinlichkeit, dass in einer Stichprobe von 1.000 Gelenken ein Gelenk defekt ist, beträgt:

$$P(1) = \frac{\mu^x e^{-\mu}}{x!} = \frac{1^1 (2{,}71828)^{-1}}{1!} = 0{,}368$$

Für zwei defekte Gelenke von 1.000 beträgt sie:

$$P(2) = \frac{\mu^x e^{-\mu}}{x!} = \frac{1^2 (2{,}71828)^{-1}}{2!} = 0{,}184$$

Und für drei defekte Gelenke von 1.000 beträgt sie:

$$P(3) = \frac{\mu^x e^{-\mu}}{x!} = \frac{1^3 (2{,}71828)^{-1}}{3!} = 0{,}061$$

Wenn Sie sich das durchlesen, kommt es Ihnen vielleicht komisch vor, dass ich ein defektes Teil als Erfolg bezeichne. Denken Sie daran, dass das nur die Bezeichnung eines spezifischen Ereignisses ist.

POISSON.VERT

Hier die Schritte, die Sie durchführen müssen, um die Excel-Funktion `POISSON.VERT` auf das oben genannte Beispiel anzuwenden:

1. **Wählen Sie die Zelle aus, in die das Ergebnis der Funktion `POISSON.VERT` eingefügt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf POISSON.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 19.4).**

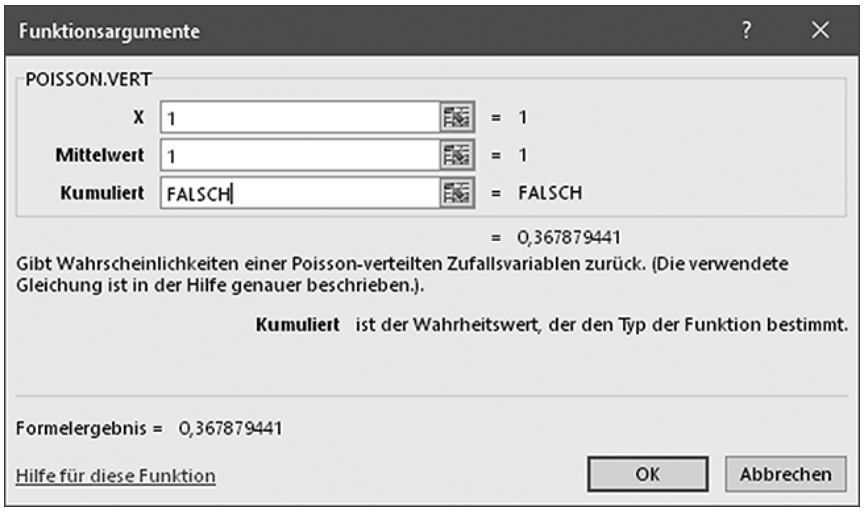


Abbildung 19.4: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `POISSON.VERT`

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Feld X habe ich die Anzahl der Ereignisse eingegeben, für die ich die Wahrscheinlichkeit bestimmen möchte. Da ich nach *P*(1) suche, gebe ich 1 ein.

   In das Feld MITTELWERT habe ich den Mittelwert des Prozesses eingegeben. Das ist $N\pi$, in unserem Beispiel 1.

   In das Feld KUMULIERT kann `WAHR` für die kumulative Wahrscheinlichkeit oder `FALSCH` für genau die Wahrscheinlichkeit der Anzahl der Ereignisse eingegeben werden. Ich habe `FALSCH` eingegeben.

   Nachdem die Werte für X, MITTELWERT und KUMULIERT eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. In diesem Beispiel ist das Ergebnis 0,367879441.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Im Beispiel habe ich Ihnen auch die Wahrscheinlichkeit für zwei und für drei defekte Gelenke von 1.000 vorgeführt. Um die Berechnungen nachzuvollziehen, müssen Sie in Schritt 3 in das Feld X den Wert 2 eingeben, um *P*(2) zu berechnen, und 3, um *P*(3) zu bestimmen.

Wie bereits weiter vorne gesagt, ist es im 21. Jahrhundert ziemlich einfach, Binomialverteilungen direkt zu berechnen. In Abbildung 19.5 habe ich einmal die Poissonverteilung und die Binomialwahrscheinlichkeiten für die Zahlen in Spalte B und die Bedingungen in diesem Beispiel dargestellt. Ich zeige die Wahrscheinlichkeiten grafisch so, dass Sie gut erkennen können, wie nah sie beieinander liegen. Ich habe die Zelle D3 markiert, damit Ihnen in der Bearbeitungsleiste angezeigt wird, wie ich die Funktion `BINOM.VERT` für die Berechnung der Binomialwahrscheinlichkeiten angewendet habe.

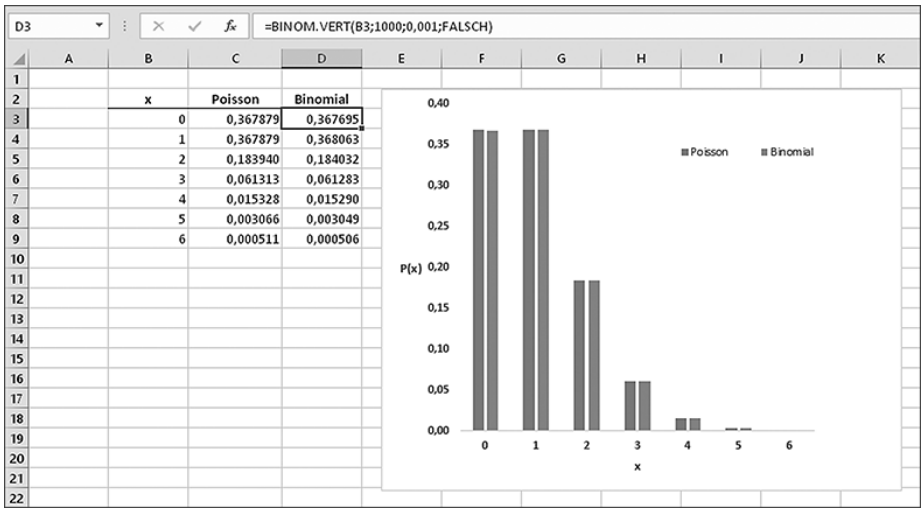


Abbildung 19.5: Poissonwahrscheinlichkeiten und Binomialwahrscheinlichkeiten

Der Nutzen der Poissonverteilung ist als Näherung heute zwar nicht mehr zeitgemäß, sie hat aber ein Eigenleben entwickelt. So unterschiedliche Erscheinungen wie Reaktionszeitdaten in psychologischen Experimenten, die Entartung radioaktiver Substanzen und Spielergebnisse in Profi-Hockey-Spielen scheinen auf Poissonverteilungen zu passen. Aus diesem Grund bauen Unternehmensanalysten und Forscher ihre Modelle gern auf dieser Verteilung auf. (»Modelle auf etwas aufbauen?« Was soll das nun wieder bedeuten? Ich erzähle Ihnen alles darüber in Kapitel 20.)

Gammaverteilung

Sie erinnern sich vermutlich noch aus Kapitel 18, dass die Anzahl der Möglichkeiten, um N Objekte in einer Sequenz anzuordnen, $N!$ (N Fakultät) entspricht. Sie erinnern sich bestimmt auch, dass $N! = N(N-1)(N-2) \ldots (2)(1)$ entspricht. Es sieht so aus, als ob sich die Fakultät nur für natürliche Zahlen berechnen lässt, oder?

Die Gammafunktion und GAMMA

Nicht so schnell. Einige Mathematiker (zu denen auch einige sehr berühmte gehören), haben das Konzept der Fakultät erweitert und verallgemeinert, damit auch rationale Zahlen und negative Ganzzahlen (was dann ziemlich komplex wird) eingeschlossen

werden. Diese Erweiterung wird Gammafunktion genannt. Wenn das Argument von Gamma eine positive Ganzzahl ist – nennen wir diese *N* –, dann ist das Ergebnis (*N*–1)!. Anderenfalls gibt Gamma das Ergebnis einer Gleichung zurück, die auf der Integralrechnung basiert.

Ich will hier nicht in die Details gehen, sondern Ihnen ein Beispiel zeigen: `4! = 24` und `5! = 120`. Daher würde das Ergebnis der Fakultät von 4,3 (was immer das auch genau ist), irgendwo zwischen 24 und 120 liegen. Wegen (*N*–1), das ich oben erwähnt habe, ermitteln Sie die Fakultät, indem Sie die Gammafunktion für 5,3 berechnen lassen (und nicht für 4,3). Das Ergebnis von Gamma(5,3) ist 38,08.

Der Name der entsprechenden Tabellenblattfunktion lautet `GAMMA`. `GAMMA` kann genau ein Argument übergeben werden. Übergeben Sie an `GAMMA` eine Zahl und Sie erhalten als Rückgabewert den Gammafunktionwert. Die Formel:

```
=GAMMA(5,3)
```

beispielsweise liefert als Ergebnis 38,08.

Die Gammaverteilung und GAMMA.VERT

Alldas oben Gesagte ist vor allem für den Bereich der theoretischen Mathematik interessant. Es wird interessanter (und auch nützlicher), wenn Sie Gamma an eine Wahrscheinlichkeitsverteilung knüpfen. Diese Hochzeit wird die Gammaverteilung genannt.

Die Gammaverteilung steht mit der Poissonverteilung ebenso in Zusammenhang, wie die negative Binomialverteilung mit der Binomialverteilung zusammenhängt. Die negative Binomialverteilung nennt Ihnen die Anzahl der Versuche, die notwendig sind, bis in einer Binomialverteilung eine festgelegte Anzahl von Erfolgen erreicht wird. Die Gammaverteilung sagt Ihnen, wie viele Stichproben Sie machen müssen, um eine festgelegte Anzahl von Erfolgen in einer Poissonverteilung zu erzielen. Die einzelnen Stichproben können eine Gruppe von Objekten sein (wie im Beispiel

mit dem Universalgelenk der FarKlempt GmbH) oder auch ein physikalischer Bereich oder ein Zeitintervall.

Die Wahrscheinlichkeitsdichtefunktion der Gammaverteilung lautet:

$$f(x) = \frac{1}{\beta^{\alpha}(\alpha-1)!} x^{\alpha-1} e^{\frac{-x}{\beta}}$$

Das funktioniert wieder nur, wenn α eine ganze Zahl ist. Ist das nicht der Fall, dann benötigen Sie – Sie haben es schon erraten – verschiedene Rechenverfahren. (Übrigens, wenn diese Funktion ausschließlich ganzzahlige Werte von α umfasst, nenn man sie *Erlangverteilung*. Das nur für den Fall, dass Sie mal jemand danach fragt.) Der Buchstabe *e* steht wieder für die Konstante 2,7818, die ich weiter oben schon erwähnt habe.

Machen Sie sich über das exotische Aussehen der mathematischen Formel keine Sorgen. Solange Sie die Bedeutung der einzelnen Symbole verstehen, sind Sie auf der sicheren Seite. Den komplizierten Teil übernimmt Excel für Sie.

Hier jetzt also die Bedeutung der einzelnen Symbole. Im Beispiel der FarKlempt Robotics GmbH ist α die Anzahl der Erfolge und β entspricht dem μ der Poissonverteilung. Die Variable *x* steht für die Anzahl der Stichproben. Wenn *x* den Wert 3 hat, α 2 und β 1 ist, sprechen Sie über die Wahrscheinlichkeitsdichte im Zusammenhang damit, den zweiten Erfolg in der dritten Stichprobe zu finden, wenn die durchschnittliche Anzahl der Erfolge pro Stichprobe (1.000) 1 ist. (Woher kommt noch mal die 1? Sie ergibt sich durch 1.000 Universalgelenke pro Stichprobe geteilt durch 0,001, der Wahrscheinlichkeit eines defekten Gelenks.)

Um die Wahrscheinlichkeit zu bestimmen, müssen Sie den Bereich unter der Dichtefunktion betrachten. Das führt mich zu der Excel-Tabellenblattfunktion, die für die Gammaverteilung entwickelt wurde.

Die Funktion GAMMA.VERT bietet Ihnen eine Reihe von Optionen. Sie können sie für die Berechnung der Wahrscheinlichkeitsdichte oder auch für die Berechnung der Wahrscheinlichkeit verwenden. In

Abbildung 19.6 ist dargestellt, wie ich mit der ersten Option ein Diagramm der Wahrscheinlichkeitsdichte erstellt habe, damit Sie erkennen können, wie die Funktion aussieht. Vor dem Hintergrund des Beispiels, das ich gerade beschrieben habe, setze ich Alpha auf 2, Beta auf 1 und berechne die Dichte der Werte von *x* in Spalte D.

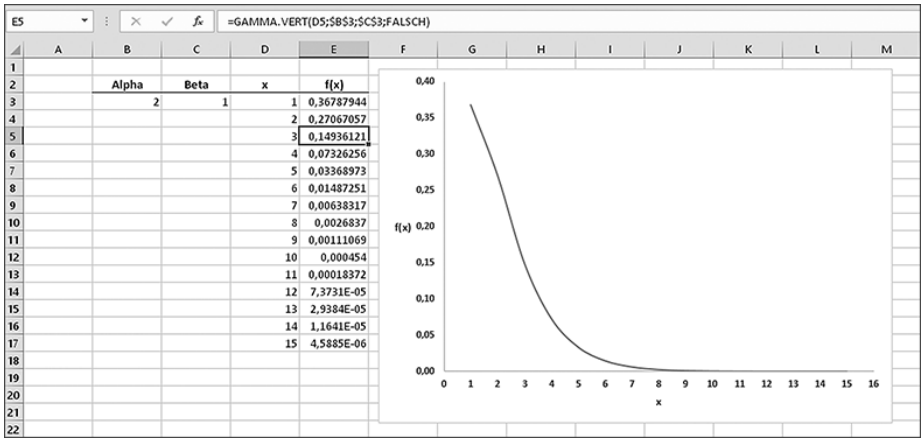


Abbildung 19.6: Die Dichtefunktion für Gamma mit Alpha = 2 und Beta = 1

Die Werte in Spalte E zeigen die Wahrscheinlichkeitsdichten für das Auffinden eines zweiten defekten Universalgelenks in der angegebenen Stichprobenanzahl von 1.000. In der Zelle E5 steht zum Beispiel die Wahrscheinlichkeitsdichte für das Auffinden des zweiten defekten Gelenks in der dritten Stichprobe.

Im echten Leben arbeiten Sie eher mit Wahrscheinlichkeiten als mit Dichten. Im Folgenden zeige ich Ihnen, wie Sie die Funktion `GAMMA.VERT` dazu verwenden können, um die Wahrscheinlichkeit des Auffindens des zweiten defekten Gelenks in der dritten Stichprobe zu bestimmen. So müssen Sie vorgehen:

1. **Wählen Sie die Zelle aus, in der das Ergebnis der Funktion `GAMMA.VERT` angezeigt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf GAMMA.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 19.7).**

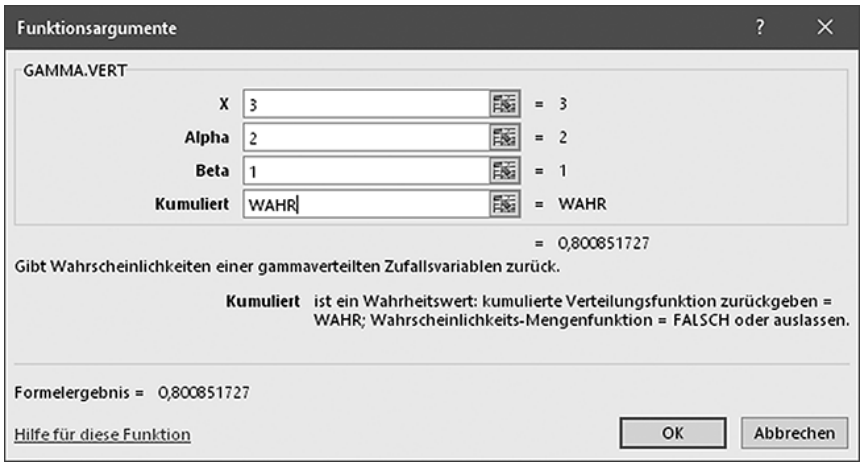


Abbildung 19.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion GAMMA.VERT

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die Werte der Argumente ein.**

   In das Feld X wird die Anzahl der Stichproben eingegeben, für die die Wahrscheinlichkeit bestimmt werden soll. Ich suche nach *P*(3), also habe ich 3 eingegeben.

   In das Feld ALPHA habe ich die die Anzahl der Erfolge eingegeben. Ich möchte den zweiten Erfolg in der dritten Stichprobe auffinden, also habe ich 2 eingegeben.

   In das Feld BETA habe ich die durchschnittliche Anzahl von Erfolgen eingegeben, die sich innerhalb einer Stichprobe einstellen. Für unser Beispiel ist das 1.

   In das Feld KUMULIERT wird für die kumulative Verteilung `WAHR` und für die Wahrscheinlichkeitsdichte `FALSCH` eingegeben. Ich möchte die Wahrscheinlichkeit und nicht die Dichte bestimmen, also habe ich WAHR eingegeben.

   Nachdem die Werte für X, ALPHA, BETA und KUMULIERT eingegeben wurden, wird das Ergebnis – in diesem Beispiel 0,800851727 – im Dialogfeld angezeigt.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

GAMMA.INV

Wenn Sie mit einem gewissen Wahrscheinlichkeitsgrad wissen möchten, wie viele Stichproben benötigt werden, um eine festgelegte Anzahl von Erfolgen zu erzielen, ist dies die richtige Funktion für Sie.

GAMMA.INV ist die Umkehrung von GAMMA.VERT. Geben Sie einen Wert für die Wahrscheinlichkeit sowie Werte für Alpha und Beta ein und die Funktion liefert Ihnen die Anzahl der Stichproben. Im Dialogfeld FUNKTIONSARGUMENTE der Funktion GAMMA.INV gibt es ein Feld WAHRSCH, ein Feld ALPHA und ein Feld BETA. Wenn Sie das Ergebnis des vorhergehenden Abschnitts in das Feld WAHRSCH und die gleichen Zahlen wie zuvor in die Felder ALPHA und BETA eingeben, liefert die Funktion das Ergebnis 3, wie in Abbildung 19.8 zu sehen ist.

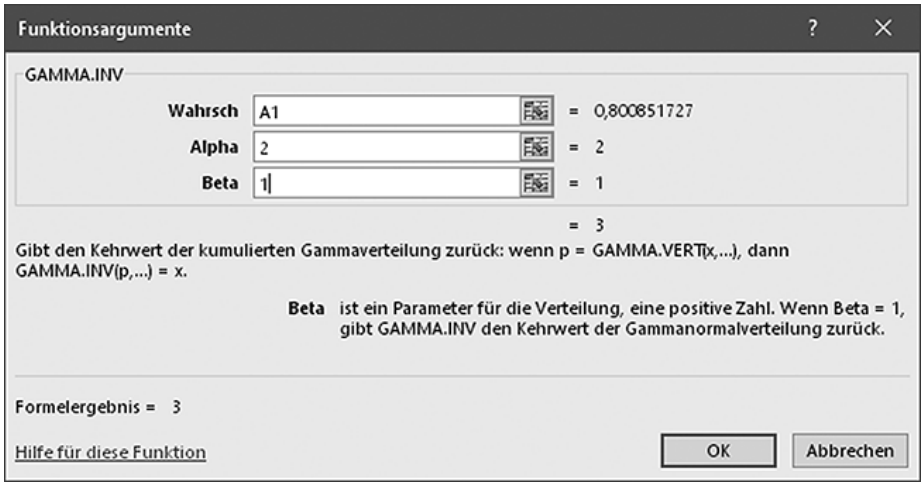


Abbildung 19.8: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion GAMMA.INV

Exponentialverteilung

Wenn Sie die Gammaverteilung nutzen und Alpha = 1 ist, haben Sie eine Exponentialverteilung. Das Ergebnis ist die Wahrscheinlichkeit,

dass eine festgelegte Anzahl von Stichproben benötigt wird, um den ersten Erfolg zu erzielen.

Wie sieht die Dichtefunktion aus? Entschuldigen Sie bitte, aber ich muss mal kurz in die Mathematik abschweifen. Hier noch einmal die Dichtefunktion für Gamma:

$$f(x) = \frac{1}{\beta^{\alpha}(\alpha-1)!} x^{\alpha-1} e^{\frac{-x}{\beta}}$$

Wenn α = 1, dann sieht das Ergebnis so aus:

$$f(x) = \frac{1}{\beta} e^{\frac{-x}{\beta}}$$

Statistiker verwenden gern das Symbol λ (den griechischen Buchstaben Lambda) für $\frac{1}{\beta}$, was uns zur endgültigen Version der Gleichung führt:

$$f(x) = \lambda e^{-\lambda x}$$

Ich erwähne das nur, weil das Dialogfeld FUNKTIONSARGUMENTE für die Excel-Funktion `EXPON.VERT` ein Feld LAMBDA enthält und ich möchte, dass Sie wissen, was es bedeutet.

EXPON.VERT

Verwenden Sie die Funktion `EXPON.VERT`, um die Wahrscheinlichkeit zu bestimmen, dass eine festgelegte Anzahl von Stichproben benötigt wird, um den ersten Erfolg in einer Poissonverteilung zu erzielen. Auch hier verwende ich wieder das Beispiel mit dem Universalgelenk. Ich zeige Ihnen, wie Sie die Wahrscheinlichkeit bestimmen können, dass Sie den ersten Erfolg in der dritten Stichprobe auffinden werden.

1. **Wählen Sie eine Zelle aus, in die das Ergebnis der Funktion `EXPON.VERT` eingefügt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf EXPON.VERT, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 19.9).**

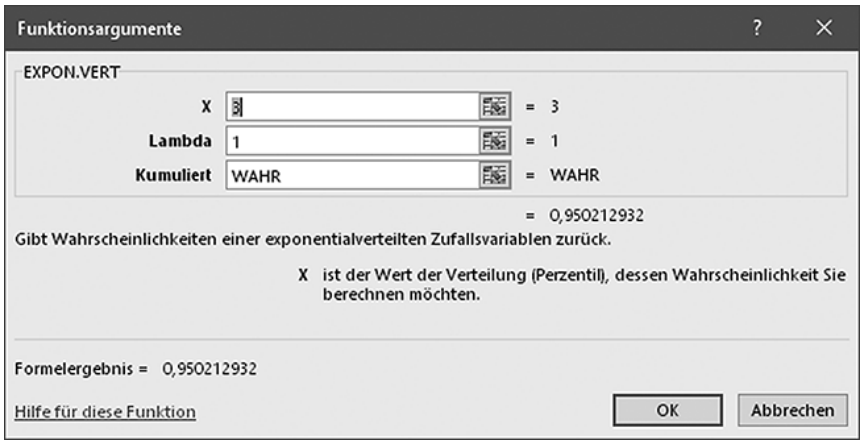


Abbildung 19.9: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion `EXPON.VERT`

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Feld X HABE ICH die Anzahl der Stichproben eingegeben, für die die Wahrscheinlichkeit bestimmt werden soll. Ich suche nach *P*(3), also habe ich 3 eingegeben.

   In das Feld LAMBDA habe ich die durchschnittliche Anzahl von Erfolgen pro Stichprobe eingegeben. Hier benötigen wir wieder die Zahlen, die ich Ihnen für das Beispiel genannt habe – die Erfolgswahrscheinlichkeit (0,001) mal der Anzahl der Universalgelenke pro Stichprobe (1.000). Ich habe demnach 1 eingegeben.

   In das Feld KUMULIERT geben Sie `WAHR` ein, wenn Sie die kumulative Verteilung erhalten möchten, oder `FALSCH`, wenn Sie die Wahrscheinlichkeitsdichte bestimmen möchten. Ich möchte die Wahrscheinlichkeit und nicht die Dichte bestimmen. Daher habe ich `WAHR` eingegeben.

Nachdem die Werte für X, LAMBDA und KUMULIERT eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. In diesem Beispiel ist das Ergebnis 0,950212932.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle einzufügen.**

Kapitel 20

Modelle

In diesem Kapitel

- Modelle erkunden
- Modelle erstellen und anpassen
- Mit der Monte-Carlo-Methode arbeiten

Der Ausdruck Modell begegnet einem heutzutage immer häufiger. Einfach ausgedrückt ist ein *Modell* etwas, das Sie kennen, mit dem Sie arbeiten können und das Ihnen dabei hilft, etwas zu verstehen, über das Sie nur sehr wenig wissen. Ein Modell soll die Sache, die es darstellt, nachahmen. Ein Globus ist zum Beispiel ein Modell der Erde. Eine Straßenkarte ist ein Modell einer Gegend. Eine Blaupause ist ein Modell eines Hauses.

Forscher verwenden Modelle, um natürliche Prozesse und Phänomene zu verstehen. Unternehmensanalysten verwenden Modelle, um Geschäftsprozesse zu verstehen. Die Modelle, die diese Leute anwenden, können Konzepte aus der Mathematik und Statistik beinhalten, also Konzepte, die so bekannt sind, dass sie Licht ins Dunkel bringen können. Der Grundgedanke besteht darin, ein Modell zu erstellen, das aus Konzepten besteht, die Sie verstehen. Dann lassen Sie das Modell zeigen, was es kann, und überprüfen, ob die Ergebnisse den Ergebnissen im echten Leben entsprechen.

In diesem Kapitel beschäftige ich mich mit der Erstellung von Modellen. Mein Ziel ist es, Ihnen zu zeigen, wie Sie die statistischen Möglichkeiten von Excel nutzen können, um Prozesse aus Ihrem Leben zu verstehen.

Die Modellierung einer Verteilung

Eine Möglichkeit für die Erstellung von Modellen besteht darin, Daten zu sammeln und diese dann in Form einer Verteilung zu gruppieren. Dann suchen Sie nach einem Prozess, der genau zu einer solchen Verteilung führt. Anschließend drücken Sie den Prozess statistisch aus, damit eine entsprechende Verteilung erzeugt werden kann. Abschließend prüfen Sie, inwiefern die erzeugte Verteilung der echten Verteilung entspricht. Dieser »Prozess, den Sie suchen und in statistischer Form wiedergeben« ist das Modell.

Wenn die Verteilung, die Sie erzeugen, gut zu den echten Daten passt, heißt das dann, dass Ihr Modell stimmt? Bedeutet es, dass der Prozess, den Sie bestimmt haben, auch der Prozess ist, durch den die Daten erzeugt werden?

Leider nein. So funktioniert es bedauerlicherweise nicht. Sie können aufzeigen, dass ein Modell nicht zutrifft, aber Sie können nicht beweisen, dass es stimmt.

Näheres zur Poissonverteilung

In diesem Abschnitt bearbeite ich beispielhaft die Erstellung eines Modells mit der Poissonverteilung. Ich stelle diese Verteilung in Kapitel 19 vor und merke dort an, dass sie eine ganze Reihe von Prozessen im wahren Leben zu charakterisieren scheint. Wenn ich vom Charakterisieren eines Prozesses spreche, meine ich, dass eine Verteilung von Daten aus dem echten Leben große Ähnlichkeit mit einer Poissonverteilung hat. Wenn das der Fall ist, ist es möglich, dass der Prozess, der eine Poissonverteilung ergibt, auch für die Entstehung der Daten verantwortlich ist.

Was für ein Prozess ist das? Beginnen wir mit einer Zufallsvariablen *x*, die wiedergibt, wie häufig ein spezifisches Ereignis in einem Intervall vorkommt. In Kapitel 19 bestand das Intervall aus einer Stichprobe von 1.000 Universalgelenken. Das spezifische Ereignis

war das Auftreten eines defekten Gelenks. Poissonverteilungen sind außerdem für Ereignisse geeignet, die in bestimmten Zeitintervallen auftreten. Als Ereignis wäre zum Beispiel das Eintreffen eines Fahrzeugs an einer Mautstelle denkbar. Ich stelle Ihnen als Nächstes die Bedingungen für einen *Poissonprozess* vor und verwende zur Illustrierung des Ganzen sowohl defekte Gelenke aus auch das Eintreffen von Fahrzeugen an einer Mautstelle:

- Die Häufigkeiten des Auftretens eines Ereignisses in zwei sich nicht überschneidenden Intervallen sind unabhängig voneinander.

  Die Anzahl der defekten Gelenke in einer Stichprobe ist unabhängig von der Anzahl der defekten Gelenke in einer anderen Stichprobe. Die Anzahl der Fahrzeuge, die innerhalb einer Stunde an einer Mautstelle eintreffen, ist unabhängig von der Anzahl der Fahrzeuge innerhalb einer anderen Stunde.

- Die Wahrscheinlichkeit des Auftretens des Ereignisses ist proportional zur Größe des Intervalls.

  Die Wahrscheinlichkeit des Auftretens eines defekten Gelenks ist in einer 10.000 Stück umfassenden Stichprobe größer als in einer Stichprobe, die nur 1.000 Stück umfasst. Die Wahrscheinlichkeit, dass ein Fahrzeug an einer Mautstelle ankommt, ist in einem Intervall von einer Stunde größer als in einem Intervall von nur einer halben Stunde.

- Die Wahrscheinlichkeit, dass ein Ereignis in einem kleinen Intervall mehr als einmal eintritt, ist 0 oder nahezu 0.

  In einer Stichprobe von 1.000 Universalgelenken ist die Wahrscheinlichkeit, direkt nacheinander zwei defekte Gelenke zu finden, extrem gering. An einer Mautstelle treffen niemals zwei Fahrzeuge gleichzeitig ein.

Wie ich bereits in Kapitel 19 erwähnte, lautet die Formel für die Poissonverteilung

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

In dieser Gleichung steht μ für die durchschnittliche Häufigkeit des Auftretens des Ereignisses in dem betrachteten Intervall. *e* ist die Konstante 2,781828 (gefolgt von unendlich vielen weiteren Dezimalstellen).

Die Poissonverteilung visualisieren

Um die Poissonverteilung besser zu verstehen, zeige ich Ihnen, wie Sie diese Verteilung visualisieren und damit experimentieren können. Abbildung 20.1 zeigt ein Arbeitsblatt mit dem Diagramm und den Werten, die dem Diagramm zugrunde liegen.

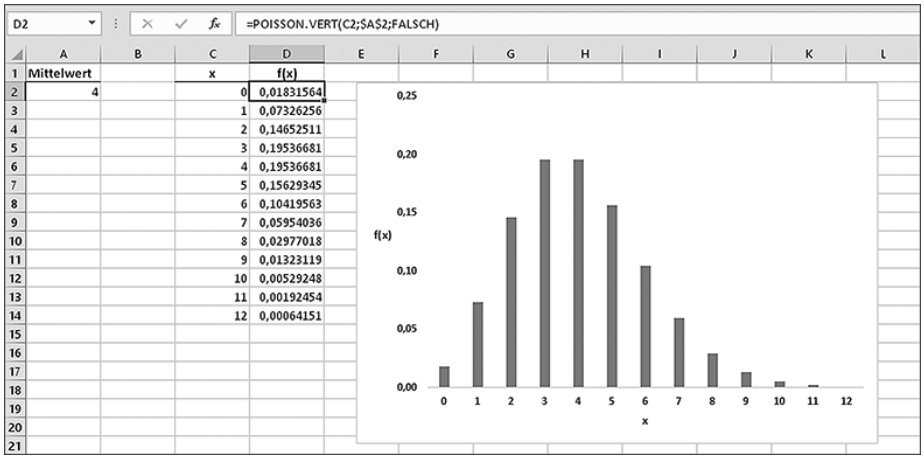


Abbildung 20.1: Die Poissonverteilung visualisieren

Zelle A2 enthält den Mittelwert der Poissonverteilung. Die Werte für *f(x)* habe ich in Spalte D berechnet und hierzu in Zelle D2 folgende Formel eingegeben und diese dann automatisch ausgefüllt:

```
=POISSON.VERT(C2;$A$2;FALSCH)
```

Da ich als drittes Argument `FALSCH` angegeben habe, erhalte ich die mit dem ersten Argument verknüpfte Wahrscheinlichkeit. (`WAHR` gibt

die kumulative Wahrscheinlichkeit zurück. Mehr zu `POISSON.VERT` finden Sie in Kapitel 19.)

Anschließend habe ich ein Säulendiagramm eingefügt und dieses noch leicht modifiziert. Sie erfahren mehr über diese Verteilung, wenn Sie in Zelle A2 unterschiedliche Werte eingeben und dann schauen, wie sich diese Änderung auf das Diagramm auswirkt.

Mit der Poissonverteilung arbeiten

Jetzt ist es Zeit, die Poissonverteilung in einem Modell anzuwenden. Bei der FarBlonJet GmbH beobachten Webdesigner, wie häufig die Intranetseite des Unternehmens pro Stunde aufgerufen wird. Sie überwachen die Seite über einen Zeitraum von 200 aufeinander folgenden Stunden und gruppieren die Daten wie in Tabelle 20.1 dargestellt.

Aufrufe/Stunde	Anzahl der beobachteten Stunden	Aufrufe/Stunde × Anzahl der beobachteten Stunden
0	10	0
1	30	30
2	44	88
3	44	132
4	36	144
5	18	90
6	10	60
7	8	56
Gesamt	200	600

Tabelle 20.1: Aufrufe der Intranetseite der FarBlonJet GmbH pro Stunde

In der ersten Spalte steht die Variable Aufrufe/Stunde. In der zweiten Spalte mit dem Titel »Anzahl der beobachteten Stunden« steht die Anzahl der Stunden, in denen die einzelnen Werte von Aufrufe/Stunde auftraten. Während der insgesamt 200 beobachteten Stunden gab es zehn Stunden ohne jeden Seitenaufruf, 30 Stunden

mit nur einem Seitenaufruf, 44 Stunden mit zwei Aufrufen usw. Aufgrund dieser Daten entscheiden sich die Webdesigner, eine Poissonverteilung als Modell der Aufrufe/Stunde anzuwenden. Anders ausgedrückt: Sie glauben, dass ein Poissonprozess die Anzahl der Aufrufe der Webseite pro Stunde erzeugt.

Wenn man die erste Spalte mit der zweiten Spalte multipliziert, erhält man die dritte Spalte. Bildet man nun die Summe der Werte der dritten Spalte, zeigt sich, dass die Intranetseite im Verlauf der 200 Beobachtungsstunden 600 Mal aufgerufen wurde. Die durchschnittliche Anzahl der Aufrufe/Stunde beträgt daher 3,00.

Wenn man die Poissonverteilung auf dieses Beispiel anwendet, ergibt sich

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{3^x e^{-3}}{x!}$$

Ab hier mache ich in Excel weiter.

Die POISSON.VERT-Funktion nochmal anwenden

In Abbildung 20.2 sind die einzelnen Werte von *x* (Aufrufe/Stunde), die Wahrscheinlichkeit der einzelnen *x*-Werte, wenn die durchschnittliche Anzahl der Aufrufe pro Stunde 3 beträgt, die vorhergesagte Anzahl von Stunden und die beobachtete Anzahl von Stunden (aus der zweiten Spalte in Tabelle 20.1) dargestellt. Ich habe Zelle B3 markiert, damit Sie in der Bearbeitungsleiste sehen können, wie ich die Tabellenblattfunktion `POISSON.VERT` angewendet habe. Spalte B habe ich bis hin zu Zelle B10 automatisch ausgefüllt. (Detaillierte Informationen zur Funktion `POISSON.VERT` finden Sie in Kapitel 19.)

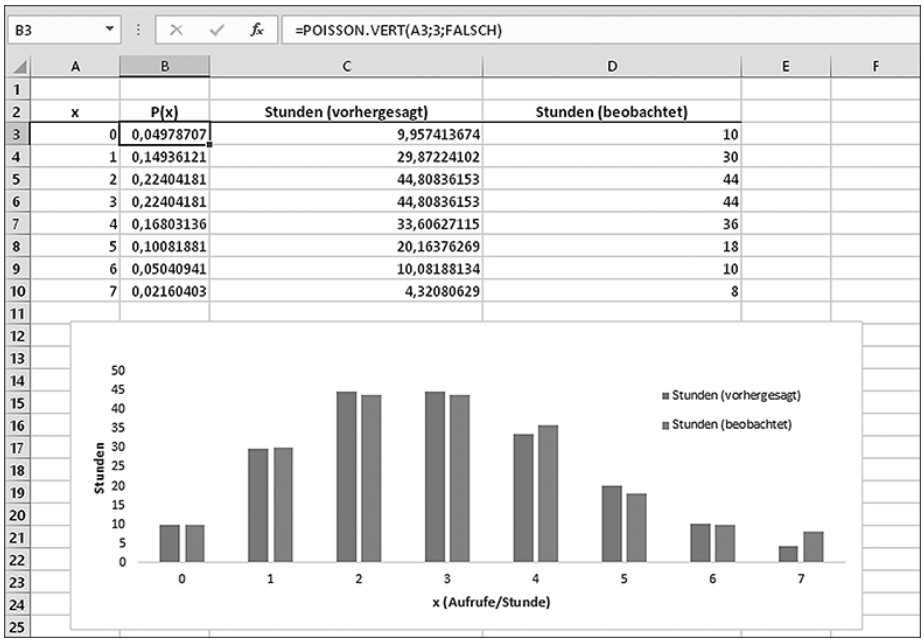


Abbildung 20.2: Aufrufe einer Website pro Stunde – mit der Poissonfunktion vorhergesagte Anzahl (μ = 3) und tatsächliche beobachtete Anzahl

Um die vorhergesagte Anzahl von Stunden zu erhalten, habe ich die einzelnen Wahrscheinlichkeitswerte in Spalte B mit 200 (der Gesamtzahl der beobachteten Stunden) multipliziert. Ich habe anschließend ein Diagramm erstellt (siehe Kapitel 3), um zu zeigen, wie nah die vorhergesagten bei den tatsächlich beobachteten Stunden liegen. Das sieht recht nah aus, oder?

Prüfen, ob das Modell passt

Nun, »recht nah aussehen« reicht für einen Statistiker nicht. Ein statistischer Test ist eine absolute Notwendigkeit. Wie alle statistischen Tests beginnt auch dieser mit einer Nullhypothese und einer Alternativhypothese. Hier sind sie:

H_0: Die Verteilung der beobachteten Aufrufe/Stunde folgt einer Poissonverteilung.

H_1: Nicht H_0

Der geeignete statistische Test umfasst eine Erweiterung der Binomialverteilung. Sie nennt sich *Multinomialverteilung* – »Multi«

deshalb, weil sie mehr Kategorien als nur *Erfolg* und *Misserfolg* beinhaltet. Sie ist schwierig anzuwenden und Excel hat keine Tabellenblattfunktion, die die Berechnungen übernimmt.

Glücklicherweise bemerkte Karl Pearson, ein Pionier auf dem Gebiet der Statistik (und der Erfinder des Korrelationskoeffizienten), dass χ^2 (Chi-Quadrat), eine Verteilung, die ich Ihnen in Kapitel 10 vorstelle, eine Annäherung an die Multinominalverteilung ist. χ^2 war ursprünglich für das Testen von Hypothesen zu Varianzen mit einer Stichprobe gedacht, erlangte jedoch weit größere Bekanntheit für Anwendungen wie jene, die ich Ihnen gleich vorstellen werde.

Pearsons Idee lautete wie folgt: Wenn Sie wissen möchten, wie gut eine angenommene Verteilung (wie die Poissonverteilung) auf eine Stichprobe (wie die beobachtete Anzahl der Stunden) passt, müssen Sie die Verteilung dazu verwenden, eine Stichprobe (zum Beispiel die vorhergesagte Anzahl von Stunden) auf Basis dieser Verteilung zu erzeugen und dann die folgende Formel anwenden:

$$\chi^2 = \sum \frac{(\text{beobachtet} - \text{vorhergesagt})^2}{\text{vorhergesagt}}$$

Normalerweise verwendet man eher den Ausdruck *erwartet* als *vorhergesagt* und sowohl *beobachtet* als auch *erwartet* werden abgekürzt. Die normalerweise anzutreffende Form der Formel lautet:

$$\chi^2 = \sum \frac{(B-E)^2}{E}$$

Für unser Beispiel bedeutet das

$$\chi^2 = \sum \frac{(B-E)^2}{E} = \frac{(10-9{,}9574)^2}{9{,}9574} + \frac{(30-29{,}8722)^2}{29{,}8722} + \ldots + \frac{(8-4{,}3208)^2}{4{,}3208}$$

Wie lautet das Endergebnis des Ganzen? Excel wird das für uns herausfinden. In Abbildung 20.3 sind die gleichen Spalten wie eben

zu sehen. Die zusätzliche Spalte F enthält die Werte für $(B-E)^2/E$. Hier hätte ich die Formel

```
=((D3-C3)^2)/C3
```

für die Berechnung des Wertes in Zelle F3 verwenden und dann diese Formel automatisch bis zur Zelle F10 ausfüllen können.

Ich habe jedoch einen anderen Weg gewählt. Zuerst habe ich dem Zellbereich C3:C10 den Namen `Stunden_erwartet` und dann dem Zellbereich D3:D10 den Namen `Stunden_beobachtet` zugewiesen. Anschließend habe ich eine Arrayformel verwendet (siehe Kapitel 2). Ich habe F3:F10 markiert und dann diese Formel erzeugt:

```
=(Stunden_beobachtet-
Stunden_erwartet)^2/Stunden_erwartet
```

Ich habe dann Strg+Shift+Enter gedrückt, um die Werte in den Zellbereich F3:F10 einzufügen. Diese Tastenkombination fasst die Formel in der Bearbeitungsleiste in geschweifte Klammern ein.

Die Summe der Werte in Spalte F befindet sich in Zelle F11 und das wiederum ist χ^2. Wenn Sie aufzeigen möchten, dass die Poissonverteilung gut zu den Daten passt, muss der Wert von χ^2 klein sein.

F3 {=(Stunden_beobachtet-Stunden_erwartet)^2/Stunden_erwartet}

	A	B	C	D	E	F	G
1							
2	x	P(x)	Stunden (erwartet)	Stunden (beobachtet)		$(B-E)^2/E$	
3	0	0,04978707	9,957413674	10		0,000182	
4	1	0,14936121	29,87224102	30		0,000546	
5	2	0,22404181	44,80836153	44		0,014583	
6	3	0,22404181	44,80836153	44		0,014583	
7	4	0,16803136	33,60627115	36		0,170502	
8	5	0,10081881	20,16376269	18		0,232192	
9	6	0,05040941	10,08188134	10		0,000665	
10	7	0,02160403	4,32080629	8		3,132857	
11					Summe=	3,566111	
12							

Abbildung 20.3: Webseitenaufrufe/Stunde – mit der Poissonverteilung ($\mu = 3$) vorhergesagte Anzahl und beobachtete Anzahl sowie die notwendigen Berechnungen zur Bestimmung von χ^2.

Okay. Und nun? Ist 3,5661 groß oder klein?

Um das herauszufinden, vergleichen Sie den berechneten Wert von χ^2 mit der χ^2-Verteilung. Ziel ist es zu bestimmen, wie hoch die Wahrscheinlichkeit ist, einen Wert zu erhalten, der mindestens so groß wie der berechnete Wert, also 3,5661, ist. Der Trick besteht darin zu wissen, wie hoch die Anzahl der Freiheitsgrade (df) ist. Bei einem Einsatzbereich, bei dem es um die Anpassungsgüte geht, verwenden Sie

$$df = k - m - 1$$

mit k = Anzahl der Kategorien und m = Anzahl der Parameter, die aufgrund der Daten geschätzt werden. Die Anzahl der Kategorien beträgt 8 (0 Aufrufe/Stunde bis 7 Aufrufe/Stunde). Aber was ist mit der Anzahl der Parameter? Ich habe die Anzahl der beobachteten Stunden verwendet, um den Parameter μ zu schätzen, so dass m in unserem Fall 1 beträgt. Daher gilt df = 8–1–1= 6.

Wenden Sie auf den Wert in Zelle F11 die Tabellenblattfunktion `CHIQU.VERT.RE(` an und geben Sie für die Freiheitsgrade den Wert 6 ein. `CHIQU.VERT.RE` gibt den Wert 0,73515 zurück, also die Wahrscheinlichkeit, einen Wert von χ^2 von mindestens 3,5661 zu erhalten, wenn H_0 wahr ist. (Weitere Informationen zur Funktion `CHIQU.VERT.RE` finden Sie in Kapitel 10.) In Abbildung 20.4 ist die χ^2-Verteilung mit 6 Freiheitsgraden und der dunkle Bereich rechts von 3,5661 dargestellt.

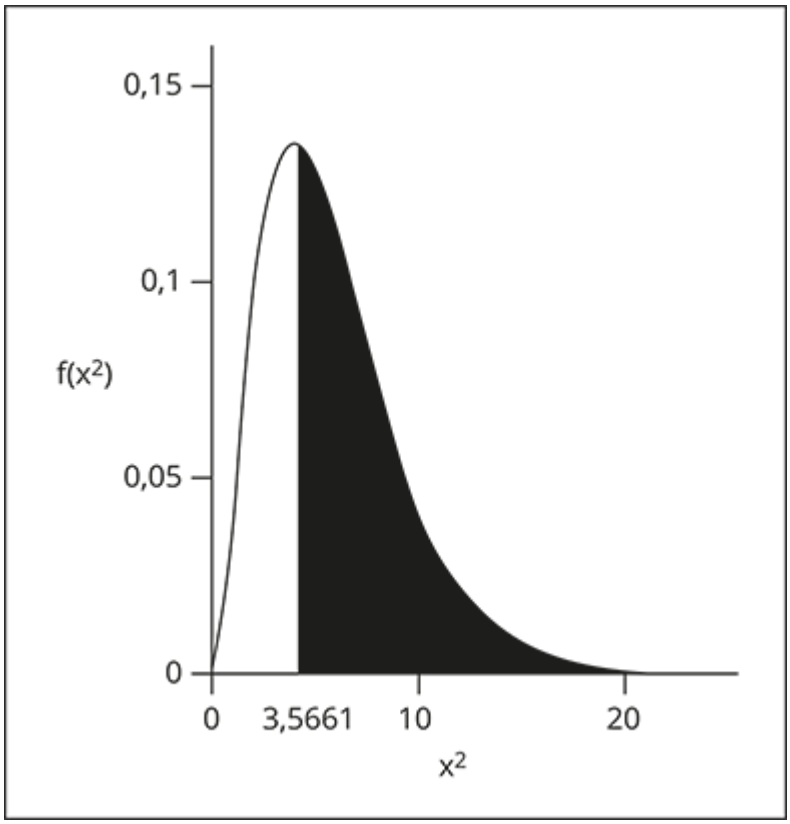


Abbildung 20.4: Die χ^2-Verteilung mit df = 6. Der dunkle Bereich ist die Wahrscheinlichkeit, einen Wert von χ^2 von mindestens 3,5661 zu erhalten, wenn H_0 wahr ist.

Wenn α = 0,05 ist, lautet die Entscheidung, H_0 nicht zu verwerfen, was bedeutet, dass Sie die Hypothese, dass die beobachteten Daten aus einer Poissonverteilung stammen, nicht verwerfen können.

Das ist einer der seltenen Momente, in denen es gut ist, H_0 nicht zu verwerfen – wenn Sie dafür plädieren, dass ein Poissonprozess für die Erzeugung der Daten verantwortlich ist. Wenn die Wahrscheinlichkeit nur ein bisschen größer als 0,05 gewesen wäre, würde es verdächtig aussehen, H_0 nicht zu verwerfen. Durch die große Wahrscheinlichkeit erscheint es hingegen vernünftiger, H_0 – und einen zugrunde liegenden Poissonprozess – nicht zu verwerfen. (Weitere Informationen dazu finden Sie im separaten Textkasten ganz am Ende von Kapitel 10.)

Ein kurzer Hinweis zu CHIQU.TEST

Excel enthält die Tabellenblattfunktion CHIQU.TEST, die auf den ersten Blick so aussieht, als würde sie genau diesen Test durchführen, dabei aber nur ein Zehntel des Arbeitsaufwands verursachen. Im Dialogfeld FUNKTIONSARGUMENTE der Funktion gibt es ein Feld für beobachtete Werte und ein Feld für erwartete Werte.

Ein Problem ist, dass CHIQU.TEST keinen Wert für χ^2 zurückgibt. Die Funktion überspringt diesen Schritt und gibt stattdessen die Wahrscheinlichkeit zurück, dass χ^2 mindestens genauso groß ist wie der Wert, den Sie mit den beobachteten und vorhergesagten Werten errechnen.

Ein anderes Problem besteht darin, dass die Freiheitsgrade von CHIQU.TEST in diesem Fall falsch sind. CHIQU.TEST macht einfach weiter und geht davon aus, dass df = $k-1$ (7) statt $k-m-1$ (6) ist. Sie verlieren einen Freiheitsgrad, da Sie μ auf der Grundlage der Daten schätzen. Bei anderen Modellverfahren verlieren Sie sogar mehr als einen Freiheitsgrad. Nehmen wir beispielsweise einmal an, Sie glauben, dass eine Normalverteilung den zugrunde liegenden Prozess bestimmt. In diesem Fall schätzen Sie μ und σ auf der Basis der Daten und verlieren bereits zwei Freiheitsgrade.

Dadurch, dass das Ergebnis von CHIQU.TEST auf einem falschen Wert für df basiert, ergibt die Funktion einen unangemessen großen (und daher falschen) Wert der Wahrscheinlichkeit.

CHIQU.TEST wäre perfekt, wenn es möglich wäre, einen Wert für df einzugeben, oder wenn die Funktion einen Wert für χ^2 zurückgäbe (den Sie dann mit CHIQU.VERT und dem korrekten Wert von df überprüfen könnten).

Wenn Sie keine Freiheitsgrade verlieren, funktioniert CHIQU.TEST völlig ohne Probleme. Kommt das überhaupt vor? Im nächsten Abschnitt schon.

Modelle für Baseballstatistiken

Baseball ist traditionell ein Spiel, für das umfassende Statistiken erstellt werden, und viele Fans studieren diese Statistiken mit großer Akribie. SABR, die Society for American Baseball Research, wurde von einer Gruppe enthusiastischer Fan-Statistiker ins Leben gerufen, die tief in die Statistiken dieser amerikanischen Sportart eintauchen. Sie nennen ihre Arbeit abgeleitet von der Abkürzung ihrer Gruppe *Sabremetrics* (etwa: Sabremetrie) und sich selbst *Sabremetricians*.

Ich erzähle Ihnen das deshalb, weil die Sabremetrie ein schönes Beispiel für die Modellerstellung liefert. Sie basiert auf der logischen Idee, dass es das Ziel einer Baseballmannschaft ist, Punkte (so genannte »Runs«) zu erzielen und den Gegner davon abzuhalten, seinerseits Punkte zu machen. Je besser die Mannschaft bei der Umsetzung dieser beiden Ziele ist, umso Spiele gewinnt sie. Bill James, der Namensgeber und wichtigste Vertreter der Sabremetrie, hat herausgefunden, dass zwischen der Anzahl der gewonnenen Punkte, der Anzahl der zugelassenen Punkte und der Siegquote der Mannschaft ein enger Zusammenhang besteht. Er bezeichnet das als *Pythagorean Percentage*, was mit *Pythagoreischer Prozentsatz* übersetzt werden könnte:

$$\text{Pythagoreischer Prozentsatz} = \frac{(\text{Punktgewinne})^2}{(\text{Punktgewinne})^2 + (\text{Zugelassene Punkte})^2}$$

Sie können es sich als ein Modell vorstellen, mit dem man die Anzahl der gewonnenen Spiele vorhersagen kann. (Dies ist die Originalformel von Bill James, die ich nachfolgend verwende. Im Laufe der Jahre haben die »Sabremetriker« herausgefunden, dass 1,84 ein genauerer Exponent ist als die 2.) Berechnen Sie diesen Prozentsatz und multiplizieren Sie ihn mit der Anzahl der Spiele, die eine Mannschaft absolviert. Vergleichen Sie das Ergebnis dann mit den tatsächlichen Siegen der Mannschaft. Wie gut kann durch das Modell die Anzahl der Spiele vorhergesagt werden, die die einzelnen Mannschaften in der Saison 2011 gewonnen haben?

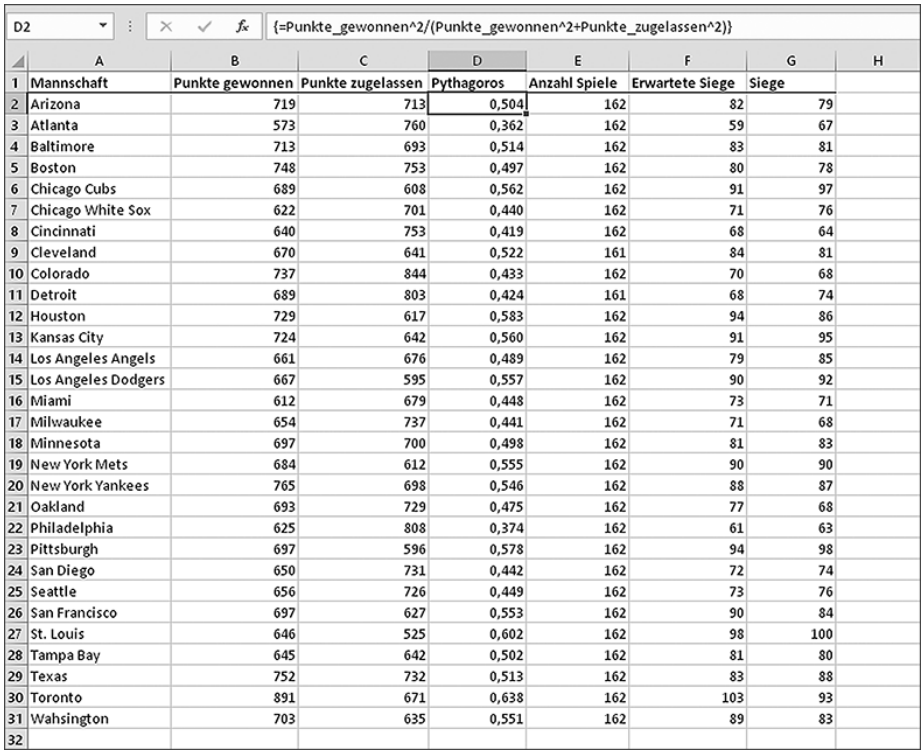
Die von Bill James entwickelte Formel kann übrigens nicht nur für Baseball, sondern auch für andere Sportarten eingesetzt werden. Wenn Sie beispielsweise Fußballfan sind, geben Sie in die Suchmaschine Ihres Vertrauens einmal **Pythagoras-Formel Fußball** ein und stöbern Sie durch die Suchergebnisse. (Dies ist ein Tipp des Übersetzers.)

Um das herauszufinden, habe ich alle relevanten Daten für jede einzelne Mannschaft der amerikanischen Major League für das Jahr 2011 zusammengetragen. (Danke an `www.baseball-reference.com`.) Die Daten habe ich in das in Abbildung 20.5 abgebildete Arbeitsblatt eingetragen.

Wie Sie in Abbildung 20.5 sehen können, habe ich für die Berechnung des Pythagoreischen Prozentsatzes in Spalte D eine Arrayformel verwendet. Zuerst habe ich den Daten in Spalte B den Namen `Punkte_gewonnen` zugewiesen und den Daten in Spalte C den Namen `Punkte_zugelassen`. Anschließend habe ich den Zellbereich D2:D31 markiert und diese Formel eingegeben:

```
=Punkte_gewonnen^2/(Punkte_gewonnen^2+Punkte_zugelassen^2)
```

Die Arrayformel wurde durch Drücken der Tastenkombination Strg+Shift+Enter erstellt, die Werte werden dann in den Zellbereich D2:D31 eingefügt und die Formel in der Bearbeitungsleiste wird in geschweifte Klammern eingefasst.



D2 {=Punkte_gewonnen^2/(Punkte_gewonnen^2+Punkte_zugelassen^2)}

	A	B	C	D	E	F	G	H
1	Mannschaft	Punkte gewonnen	Punkte zugelassen	Pythagoros	Anzahl Spiele	Erwartete Siege	Siege	
2	Arizona	719	713	0,504	162	82	79	
3	Atlanta	573	760	0,362	162	59	67	
4	Baltimore	713	693	0,514	162	83	81	
5	Boston	748	753	0,497	162	80	78	
6	Chicago Cubs	689	608	0,562	162	91	97	
7	Chicago White Sox	622	701	0,440	162	71	76	
8	Cincinnati	640	753	0,419	162	68	64	
9	Cleveland	670	641	0,522	161	84	81	
10	Colorado	737	844	0,433	162	70	68	
11	Detroit	689	803	0,424	161	68	74	
12	Houston	729	617	0,583	162	94	86	
13	Kansas City	724	642	0,560	162	91	95	
14	Los Angeles Angels	661	676	0,489	162	79	85	
15	Los Angeles Dodgers	667	595	0,557	162	90	92	
16	Miami	612	679	0,448	162	73	71	
17	Milwaukee	654	737	0,441	162	71	68	
18	Minnesota	697	700	0,498	162	81	83	
19	New York Mets	684	612	0,555	162	90	90	
20	New York Yankees	765	698	0,546	162	88	87	
21	Oakland	693	729	0,475	162	77	68	
22	Philadelphia	625	808	0,374	162	61	63	
23	Pittsburgh	697	596	0,578	162	94	98	
24	San Diego	650	731	0,442	162	72	74	
25	Seattle	656	726	0,449	162	73	76	
26	San Francisco	697	627	0,553	162	90	84	
27	St. Louis	646	525	0,602	162	98	100	
28	Tampa Bay	645	642	0,502	162	81	80	
29	Texas	752	732	0,513	162	83	88	
30	Toronto	891	671	0,638	162	103	93	
31	Wahsington	703	635	0,551	162	89	83	
32								

Abbildung 20.5: Punktgewinne, zugelassene Punkte, vorhergesagte Siege und tatsächliche Siege der einzelnen Baseballmannschaften der Major League im Jahr 2011

Alternativ hätte ich ohne die Verwendung von Namen und einer Arrayformel auch zuerst diese Formel in Zelle D2 eingeben können:

```
=B2^2/((B2^2)+(C2^2))
```

Anschließend hätte ich die restlichen Zellen in Spalte D automatisch ausfüllen lassen.

Zum Schluss habe ich die einzelnen Pythagoreischen Prozentsätze in Spalte D mit der Anzahl der jeweils absolvierten Spiele multipliziert (28 Mannschaften spielten 162 Mal, zwei Mannschaften nur 161 Mal), um die Anzahl der vorhergesagten Siege in Spalte E zu berechnen. Da die Anzahl der tatsächlichen Siege auf jeden Fall eine ganze Zahl sein muss, habe ich die Tabellenblattfunktion `RUNDEN` verwendet, um die Anzahl der vorhergesagten Siege entsprechend zu runden. Die Formel, die den Wert in Zelle F3 liefert, lautet zum Beispiel:

```
=RUNDEN(D3*E3;0)
```

Die Null innerhalb der Klammern zeigt an, dass ich im Ergebnis der Funktion keine Dezimalstellen habe möchte.

Bevor ich weitermache, weise ich den Daten in Spalte F den Namen `Erwartete_Siege` und den Daten in Spalte G den Namen `Siege` zu.

Wie gut passt das Modell zur Realität? In diesem Fall kann die Funktion `CHIQU.TEST` eine Antwort auf die Frage liefern. Ich verliere keine Freiheitsgrade: Ich habe nicht die Daten der tatsächlichen Siege in Spalte G verwendet, um Parameter wie einen Mittelwert oder eine Varianz zu schätzen, und ich habe diese Parameter nicht dazu verwendet, um die vorhergesagten Siege zu berechnen. Stattdessen basieren die Vorhersagen auf anderen Daten, nämlich den Punktgewinnen und den zugelassenen Punkten. Daher gilt df = $k - m - 1 = 30 - 0 - 1 = 29$.

Und so können Sie die Funktion `CHIQU.TEST` anwenden (wenn sie geeignet ist!):

1. **Wenn Sie alle Daten eingegeben haben, wählen Sie eine Zelle aus, in die das Ergebnis von `CHIQU.TEST` eingefügt werden soll.**

2. **Klicken Sie im Menü mit den statistischen Funktionen auf CHIQU.TEST, um das Dialogfeld FUNKTIONSARGUMENTE zu öffnen (siehe Abbildung 20.6).**

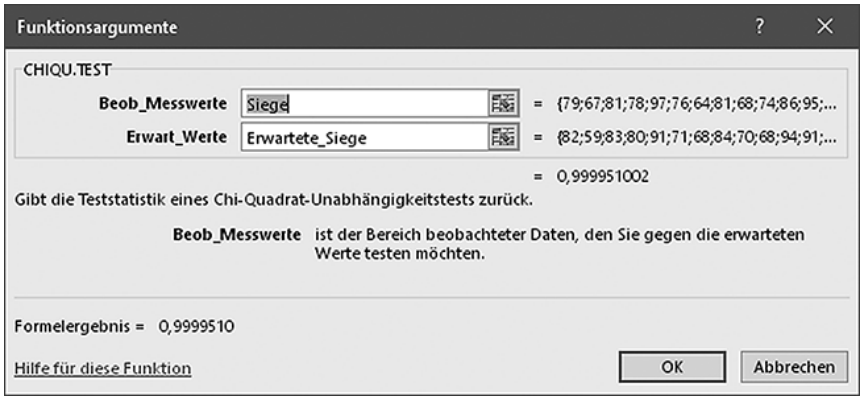


Abbildung 20.6: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion *CHIQU.TEST*

3. **Geben Sie in das Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente der Funktion ein.**

   Geben Sie in das Feld BEOB_MESSWERTE den Zellenbereich ein, der die beobachteten Werte enthält. In diesem Beispiel ist dies der Bereich `Siege` (der Name für den Zellbereich F2:F31).

   Geben Sie in das Feld ERWART_WERTE den Zellenbereich ein, der die vorhergesagten Werte enthält. In diesem Beispiel ist dies `Erwartete_Siege` (der Name für den Zellbereich E2:E31).

   Nachdem die Werte für BEOB_MESSWERTE und ERWART_WERTE eingegeben wurden, wird das Ergebnis im Dialogfeld angezeigt. In diesem Beispiel ist das Ergebnis 0,9999510. Es bedeutet, dass Sie mit 29 Freiheitsgraden mit großer Wahrscheinlichkeit einen Wert von χ^2 finden werden, der mindestens genauso hoch ist wie der, den Sie auf der Grundlage der beobachteten und vorhergesagten Werte errechnet haben. Anders ausgedrückt: Der für χ^2 berechnete Wert ist sehr klein, was bedeutet, dass die Anzahl der erwarteten Siege sehr nahe bei den Anzahl der tatsächlichen Siege liegt. Ergebnis: Das Modell passt sehr gut zu den Daten.

4. **Klicken Sie auf OK, um das Ergebnis in die ausgewählte Zelle zu übertragen.**

Simulationen

Ein anderer Ansatz für die Verwendung von Modellen ist die Simulation eines Prozesses. In diesem Fall definieren Sie das, was ein Prozess bewirkt, so umfangreich wie möglich, verwenden dann Zahlen, um den Prozess darzustellen, und führen ihn aus. Dieses Verfahren ist hervorragend dazu geeignet herauszufinden, was ein Prozess eigentlich bewirkt, wenn andere Analysemethoden zu komplex sind.

Es darauf ankommen lassen: die Monte-Carlo-Methode

Viele Prozesse beinhalten das Element der Zufälligkeit. Das Ergebnis lässt sich nun einmal nicht mit Sicherheit vorhersagen. Um diese Art von Prozess zu simulieren, müssen Sie auch die Zufälligkeit simulieren. Simulationen, die Zufälligkeit beinhalten, werden *Monte-Carlo*-Simulationengenannt. Der Name stammt von der Stadt in Monaco, deren Hauptattraktion das große Spielcasino ist.

In den folgenden Abschnitten gebe ich Ihnen eine Reihe von Beispielen. Diese Beispiele sind nicht so komplex, dass Sie sie nicht analysieren könnten. Genau aus diesem Grund verwende ich sie auch: Sie können die Ergebnisse mit der Analyse vergleichen.

Den Würfel »zinken«

In Kapitel 18 habe ich einen Würfel erwähnt, der dazu neigt, in Abhängigkeit der Augen auf seinen Flächen ein bestimmtes Ergebnis zu liefern: Eine 6 erscheint sechs Mal häufiger als eine 1, eine 5 fünf Mal häufiger usw. Bei einem Wurf beträgt die Wahrscheinlichkeit, eine *n* zu würfeln, *n*/21.

Stellen Sie sich vor, Sie haben zwei derartig »gezinkte« Würfel. Wie würden die Ergebnisse von 200 Würfen dieser Würfel aussehen? Wie wäre der Durchschnittswert dieser 200 Würfe? Was wäre mit der Varianz und der Standardabweichung? Excel kann Monte-Carlo-Simulationen für Sie durchführen und die Antworten auf all diese Fragen liefern.

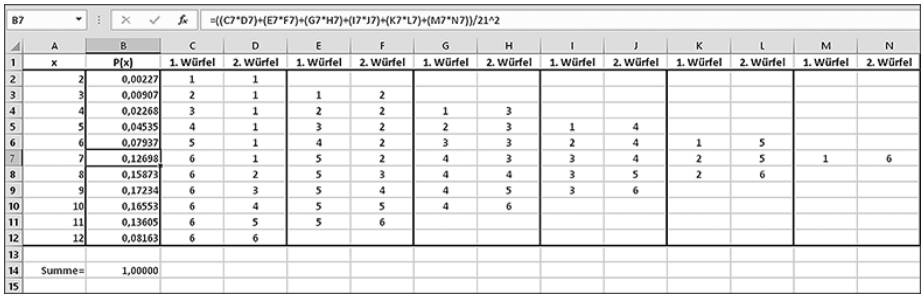
Zu Beginn lasse ich Excel die Wahrscheinlichkeit der einzelnen Ergebnisse berechnen. In Abbildung 20.7 ist dargestellt, wie ich das mache. In Spalte A stehen die möglichen Ergebnisse beim Würfeln mit zwei Würfeln (2 bis 12). In den Spalten C bis N stehen die verschiedenen Kombinationsmöglichkeiten, mit denen man die jeweiligen Ergebnisse erzielen kann. Die Spalten C, E, G, I, K und M enthalten die möglichen Würfelergebnisse des ersten Würfels,

während die Spalten D, F, H, J, L und N die möglichen Ergebnisse des zweiten Würfels enthalten. In Spalte B wiederum steht die Wahrscheinlichkeit der einzelnen Ergebnisse basierend auf den Zahlen in den Spalten C bis M. Ich habe die Zelle B7 markiert, damit Sie in der Bearbeitungsleiste sehen können, dass ich die folgende Formel verwendet habe, um Excel dazu zu bringen, die Wahrscheinlichkeit für das Würfelergebnis 7 zu berechnen:

```
=((C7*D7)+(E7*F7)+(G7*H7)+(I7*J7)+(K7*L7)+
(M7*N7))/21^2
```

Die restlichen Zellen in Spalte B habe ich automatisch aufgefüllt.

Die Summe in Zelle B14 bestätigt, dass ich jede Möglichkeit in Betracht gezogen habe.



B7 =((C7*D7)+(E7*F7)+(G7*H7)+(I7*J7)+(K7*L7)+(M7*N7))/21^2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	x	P(x)	1. Würfel	2. Würfel	1. Würfel	2. Würfel	1. Würfel	2. Würfel	1. Würfel	2. Würfel	1. Würfel	2. Würfel	1. Würfel	2. Würfel
2	2	0,00227	1	1										
3	3	0,00907	2	1	1	2								
4	4	0,02268	3	1	2	2	1	3						
5	5	0,04535	4	1	3	2	2	3	1	4				
6	6	0,07937	5	1	4	2	3	3	2	4	1	5		
7	7	0,12698	6	1	5	2	4	3	3	4	2	5	1	6
8	8	0,15873	6	2	5	3	4	4	3	5	2	6		
9	9	0,17234	6	3	5	4	4	5	3	6				
10	10	0,16553	6	4	5	5	4	6						
11	11	0,13605	6	5	5	6								
12	12	0,08163	6	6										
13														
14	Summe=	1,00000												
15														

Abbildung 20.7: Ergebnisse und Wahrscheinlichkeiten beim Einsatz eines Paares »gezinkter« Würfel

Als Nächstes müssen wir den Prozess des Würfelns simulieren. Jeder Wurf erzeugt einen Wert der Zufallsvariablen *x* gemäß der Wahrscheinlichkeitsverteilung, die durch die Spalten A und N definiert ist. Wie können Sie diese Würfe simulieren?

Datenanalysetool: Zufallszahlengenerierung

Das Zufallszahlengenerierungstool von Excel ist für diese Art der Simulation maßgeschneidert. Teilen Sie dem Programm ganz einfach mit, wie viele Werte Sie erzeugen möchten, geben Sie eine Wahrscheinlichkeitsverteilung vor und schon erzeugt das Programm Zufallszahlen gemäß der Parameter der ausgewählten Verteilung. Jede dieser Zufallszahlen entspricht einem Wurf der Würfel.

Und so verwenden Sie das Tool zur Erzeugung von Zufallszahlen:

1. **Wählen Sie im Menüband DATEN | DATENANALYSE, um das Dialogfeld ANALYSE-FUNKTIONEN zu öffnen.**

2. **Blättern Sie im Dialogfeld ANALYSE-FUNKTIONEN in der Auswahlliste ANALYSE-FUNKTIONEN nach unten und markieren Sie den Eintrag ZUFALLSZAHLENGENERIERUNG.**

3. **Klicken Sie auf OK, um das Dialogfeld ZUFALLSZAHLENGENERIERUNG zu öffnen.**

   Das Dialogfeld ZUFALLSZAHLENGENERIERUNG ist in Abbildung 20.8 dargestellt.

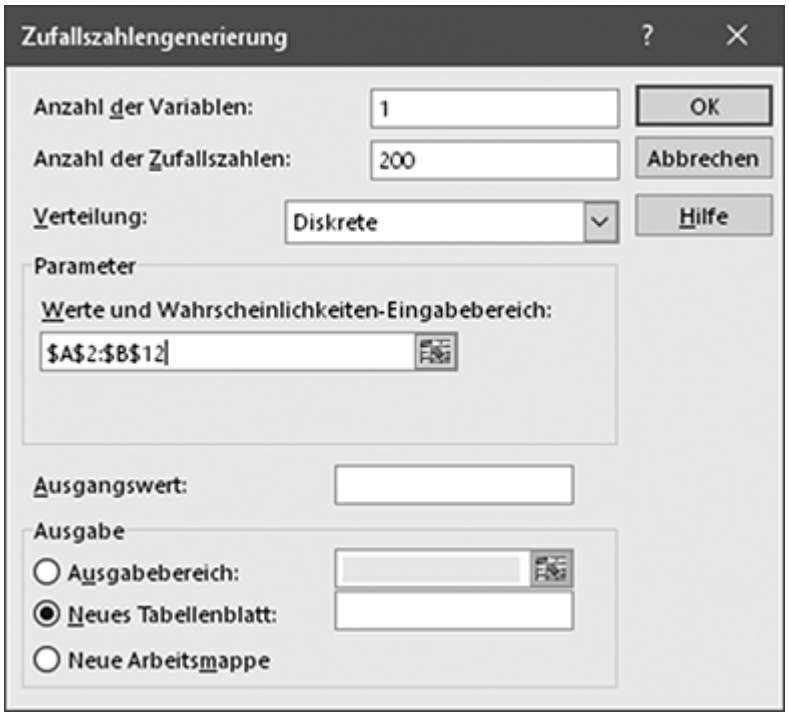


Abbildung 20.8: Das Dialogfeld ZUFALLSZAHLENGENERIERUNG

4. **Geben Sie in das Feld ANZAHL DER VARIABLEN die Anzahl der Variablen ein, für die Sie Zufallszahlen generieren möchten.**

   Ich habe für unser Beispiel die Zahl **1** eingegeben. Ich interessiere mich nur für die Ergebnisse des Würfelns mit zwei Würfeln.

5. **Geben Sie in das Eingabefeld ANZAHL DER ZUFALLSZAHLEN die Anzahl der zu generierenden Zahlen ein.**

   Ich habe die Zahl **200** eingegeben, um 200 Würfe mit den »gezinkten« Würfeln zu simulieren.

6. **Klicken Sie auf den Pfeil im Feld VERTEILUNG, um den Verteilungstyp auszuwählen.**

   Ihnen stehen hier insgesamt sieben Optionen zur Auswahl. Die Wahl, die Sie treffen, bestimmt das Aussehen des Bereichs PARAMETER des Dialogfeldes, da unterschiedliche Verteilungstypen unterschiedliche Typen (und Anzahlen) von Parametern haben. Hier geht es um eine diskrete Zufallsvariable, so dass Sie die Option DISKRETE auswählen sollten.

7. **Da Sie im vorigen Schritt die Option DISKRETE ausgewählt haben, wird das Eingabefeld WERTE UND WAHRSCHEINLICHKEITEN-EINGABEBEREICH im Bereich PARAMETER angezeigt. Geben Sie dort den Zellenbereich ein, in dem sich die Werte der Variablen und der zugehörigen Wahrscheinlichkeiten befinden.**

   Die möglichen Ergebnisse des Würfelns sind im Zellenbereich A2:A12 und die Wahrscheinlichkeiten im Zellenbereich B2:B12 enthalten. Der gesuchte Bereich lautet also A2:B12. Excel fügt die Dollarzeichen ($) für den absoluten Bezug ein.

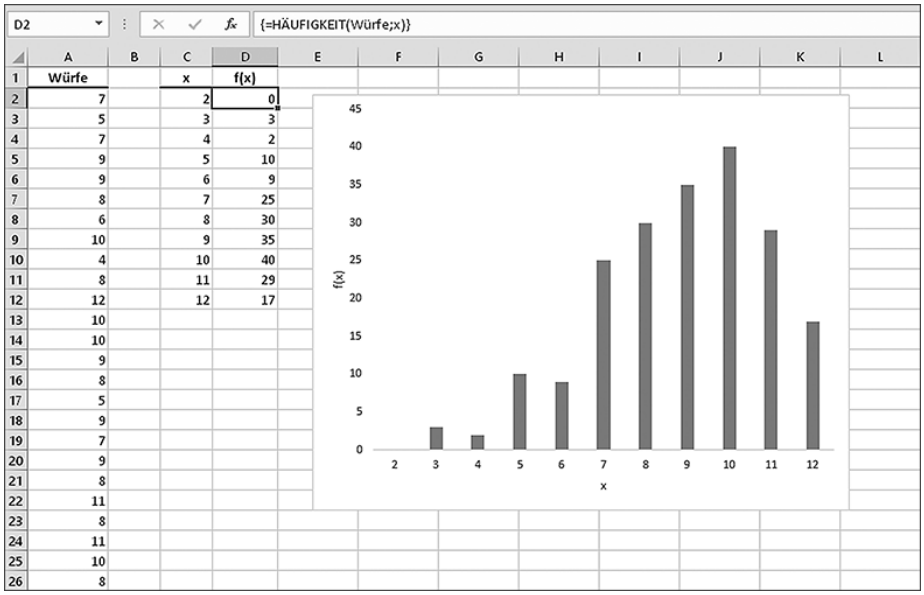
8. **Wählen Sie im Bereich AUSGABE eines der Optionsfelder aus, um festzulegen, wo die Ergebnisse angezeigt werden sollen.**

   Ich habe die Option NEUES TABELLENBLATT ausgewählt, um die Ergebnisse auf einer neuen Seite des Arbeitsblatts anzuzeigen.

9. **Klicken Sie auf OK.**

Da ich die Option NEUES TABELLENBLATT ausgewählt habe, öffnet sich ein neues Arbeitsblatt mit den Ergebnissen. Dieses Arbeitsblatt ist in Abbildung 20.9 dargestellt. Die erzeugten Zufallszahlen stehen in Spalte A. (Ich habe hier noch eine Spaltenüberschrift ergänzt.) Insgesamt sind es 200 Zeilen, also viel zu lang, um sie Ihnen zeigen zu können. Ich hätte sie ausschneiden und in zehn Spalten à 20 Zellen einfügen können, aber dann würden Sie sich auch nur 200 Zufallszahlen ansehen.

Stattdessen verwende ich die Funktion `HÄUFIGKEIT`, um die Zahlen (sie stehen in der Spalte C) nach Häufigkeiten zu sortieren und das Ergebnis in Spalte D anzuzeigen. Dann verwende ich die Diagrammfeatures von Excel, um die Ergebnisse zu visualisieren. Ich habe die Zelle D2 markiert, damit Sie in der Bearbeitungsleiste sehen können, wie ich die Funktion `HÄUFIGKEIT` für diese Zelle eingesetzt habe. Wie Sie sehen können, habe ich dem Zellbereich A2:A201 den Namen `Würfe` und dem Zellbereich C2:C12 den Namen `x` zugewiesen.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Würfe		x	f(x)								
2	7		2	0								
3	5		3	3								
4	7		4	2								
5	9		5	10								
6	9		6	9								
7	8		7	25								
8	6		8	30								
9	10		9	35								
10	4		10	40								
11	8		11	29								
12	12		12	17								
13	10											
14	10											
15	9											
16	8											
17	5											
18	9											
19	7											
20	9											
21	8											
22	11											
23	8											
24	11											
25	10											
26	8											

Abbildung 20.9: Die Ergebnisse der Simulation von 200 Würfen mit zwei »gezinkten« Würfeln

Wie sieht es mit den statistischen Werten für diese simulierten Würfe aus? Mit

```
=MITTELWERT(Würfe)
```

ermitteln Sie, dass der Mittelwert 8,87 beträgt.

```
=VAR.S(Würfe)
```

ergibt den Wert 4,19407035 als Schätzwert der Varianz; und `WURZEL` angewendet auf die Varianz liefert den Wert 2,04794296 als Schätzwert der Standardabweichung.

Wie passen diese Werte zu den Parametern der Zufallsvariablen? Das ist es, was ich meinte, als ich vom Vergleich mit der Analyse gesprochen habe. In Kapitel 18 zeige ich Ihnen, wie Sie den Erwartungswert (den Mittelwert), die Varianz und die Standardabweichung einer diskreten Zufallsvariablen berechnen können.

Der Erwartungswert ist:

$$E(x) = \sum x(P(x))$$

Im Arbeitsblatt in Abbildung 20.7 verwende ich die Tabellenblattfunktion `SUMMENPRODUKT`, um $E(x)$ zu berechnen. Die Formel lautet:

```
=SUMMENPRODUKT(A2:A12;B2:B12)
```

Der Erwartungswert beträgt 8,667.

Die Varianz ist:

$$Var(x) = \sum x^2\ (P(x)) - \left[E(x)\right]^2$$

Mit $E(x)$ hinterlegt in Zelle B16, verwende ich diese Formel

```
=SUMMENPRODUKT(A2:A12;A2:A12;B2:B12)-B16^2
```

Beachten Sie die doppelte Verwendung von A2:A12 in `SUMMENPRODUKT`. Als Ergebnis erhalten die Summe von x^2.

Die Formel gibt als Varianz den Wert 4,444 zurück. Wenn man die Funktion `WURZEL` auf diese Zahl anwendet, ergibt sich eine Standardabweichung von 2,108.

In Tabelle 20.2 zeige ich Ihnen, wie nah die Ergebnisse der Simulation bei den Parametern der Zufallsvariablen liegen.

	Statistische Werte der Simulation	Verteilungsparameter
Mittelwert	8,640	8,667
Varianz	4,573	4,444
Standardabweichung	2,139	2,108

Tabelle 20.2: Statistische Werte der Simulation des Würfelns mit »gezinkten« Würfeln und die Parameter der diskreten Verteilung

Simulation des zentralen Grenzwertsatzes

Es mag für Sie vielleicht überraschend klingen, aber Statistiker verwenden recht häufig Simulationen, um Entscheidungen hinsichtlich ihrer statistischen Werte zu treffen. Das tun sie immer dann, wenn die mathematische Analyse sehr schwierig wird.

Einige statistische Tests hängen zum Beispiel von normalverteilten Grundgesamtheiten ab. Wenn die Grundgesamtheiten nicht normalverteilt sind, was passiert dann mit diesen Tests? Funktionieren sie immer noch so, wie sie sollen? Um diese Frage zu beantworten, können Statistiker nicht-normalverteilte Grundgesamtheiten von Zahlen erzeugen, Experimente mit diesen Zahlen simulieren und die statistischen Tests dann mit den simulierten Ergebnissen vergleichen.

In diesem Abschnitt verwende ich eine Simulation, um ein wichtiges statistisches Element, den zentralen Grenzwertsatz, zu untersuchen. In Kapitel 9 stelle ich Ihnen den zentralen Grenzwertsatz im Zusammenhang mit der Stichprobenverteilung des Mittelwerts vor. Ich simuliere die Stichprobenentnahme aus einer Grundgesamtheit

mit nur drei möglichen Werten, um Ihnen zu zeigen, dass die Stichprobenverteilung sogar bei einem so kleinen Stichprobenumfang anfängt, wie eine Standardnormalverteilung auszusehen.

Ich benutze das Zufallszahlengenerierungstool, um eine normalverteilte Grundgesamtheit zu erzeugen, und ziehe daraus 40 Stichproben mit jeweils 16 Werten. Ich berechne den Mittelwert für die einzelnen Stichproben und erstelle eine Verteilung dieser Mittelwerte. Ich möchte überprüfen, inwiefern diese Verteilung zum zentralen Grenzwertsatz passt.

Die Verteilung für dieses Beispiel hat die Parameter der Grundgesamtheit der Treffer bei einem IQ-Test, eine Verteilung, die ich für Beispiele in mehreren Kapiteln heranziehe. Es handelt sich dabei um eine Standardnormalverteilung mit μ = 100 und σ = 16. Nach dem zentralen Grenzwertsatz sollte der Mittelwert der Mittelwertverteilung 100 und die Standardabweichung (der Standardfehler des Mittelwerts) 4 betragen.

Bei einer Standardnormalverteilung sieht das Dialogfeld ZUFALLSZAHLENGENERIERUNG wie in Abbildung 20.10 dargestellt aus. Durch die ersten beiden Einträge wird Excel dazu veranlasst, 16 Zufallszahlen für eine einzige Variable zu erzeugen. Wenn Sie im Listenfeld VERTEILUNG die Option STANDARD auswählen, werden im Abschnitt PARAMETER die Felder MITTELWERT und STANDARDABWEICHUNG angezeigt. Wie in der Abbildung ersichtlich ist, habe ich den Wert 100 für den Mittelwert und den Wert 16 für die Standardabweichung eingegeben. Im Bereich AUSGABE habe ich die Option AUSGABEBEREICH ausgewählt und eine 16 Zellen umfassende Spalte festgelegt. Dadurch werden die generierten Zufallszahlen im aktuellen Arbeitsblatt in der angegebenen Spalte und (da ich 40 Stichproben erstellen lasse) in den 39 benachbarten Spalten eingefügt. Anschließend habe ich die Funktion `MITTELWERT` verwendet, um den Mittelwert der einzelnen Spalten (hier also der Stichproben) zu berechnen. Anschließend habe ich mit der Funktion `RUNDEN` die Nachkommastellen der Mittelwerte »entfernt«.

Danach habe ich die gerundeten Mittelwerte der Stichproben auf ein anderes Arbeitsblatt kopiert, um deren Verteilung darzustellen. Mit der Funktion `HÄUFIGKEIT` habe ich die Häufigkeitsverteilung der Mittelwerte berechnet und diese Werte dann in einem Diagramm visualisiert. Ist der zentrale Grenzwertsatz geeignet, die Ergebnisse korrekt vorherzusagen? Abbildung 20.11 stellt meine Ergebnisse dar.

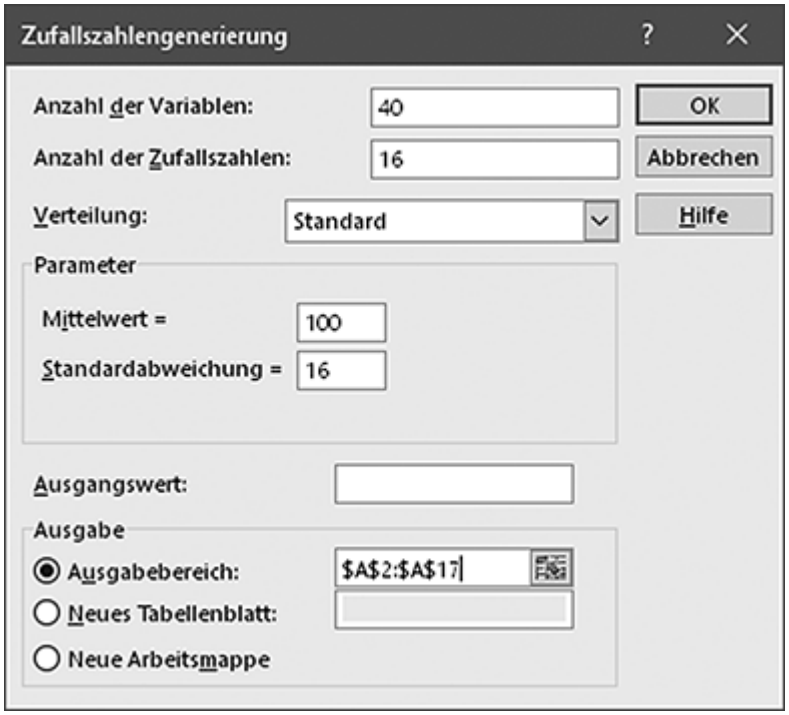
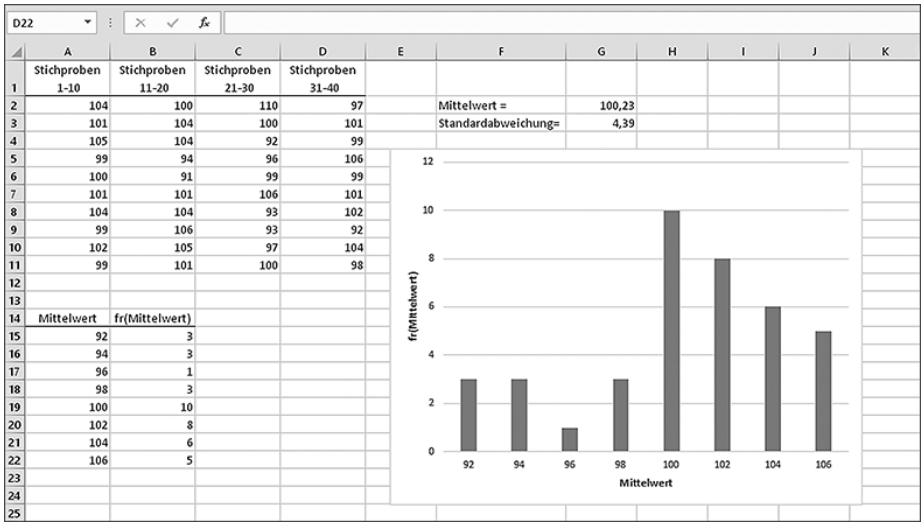


Abbildung 20.10: Das Dialogfeld ZUFALLSZAHLENGENERIERUNG für eine Standardnormalverteilung



	A	B	C	D	E	F	G
1	Stichproben 1-10	Stichproben 11-20	Stichproben 21-30	Stichproben 31-40			
2	104	100	110	97		Mittelwert =	100,23
3	101	104	100	101		Standardabweichung=	4,39
4	105	104	92	99			
5	99	94	96	106			
6	100	91	99	99			
7	101	101	106	101			
8	104	104	93	102			
9	99	106	93	92			
10	102	105	97	104			
11	99	101	100	98			
12							
13							
14	Mittelwert	fr(Mittelwert)					
15	92	3					
16	94	3					
17	96	1					
18	98	3					
19	100	10					
20	102	8					
21	104	6					
22	106	5					
23							
24							
25							

Abbildung 20.11: Die Ergebnisse der Simulation des zentralen Grenzwertsatzes

Der Mittelwert der Mittelwerte beträgt 100,23 und liegt damit nah am vorhergesagten Wert des zentralen Grenzwertsatzes von 100. Die Standardabweichung der Mittelwerte beträgt 4,39 und liegt nah am vorhergesagten Wert des zentralen Grenzwertsatzes von 4 für den Standardfehler des Mittelwerts. Das Diagramm zeigt eine Standardnormalverteilung, die allerdings leicht schief ist. Im Allgemeinen passt die Simulation gut zum zentralen Grenzwertsatz. Falls Sie diese Schritte selbst durchführen, erhalten Sie aufgrund leicht anderer Zufallszahlen andere Ergebnisse als ich, jedoch sollte Ihr Gesamtergebnis ungefähr ähnlich aussehen.

Ein paar Absätze weiter oben beschreibe ich, dass ich »die 40 Mittelwerte in ein anderes Arbeitsblatt kopiert habe«. Das ist allerdings nicht so einfach, wie es sich anhört. Wenn Sie versuchen, eine Zelle mit einer Formel in ein anderes Arbeitsblatt einzufügen, sträubt sich Excel für gewöhnlich und reagiert beim Einfügen mit einer hässlichen Fehlermeldung. Das passiert, wenn sich die Formel auf Zellen bezieht, die im neuen Arbeitsblatt keine Werte enthalten. Außerdem sind die Mittelwerte der Stichproben auf dem einen Blatt in einer Zeile angeordnet; im Zielblatt sollen diese jedoch untereinander, also in einer Spalte, stehen.

Um das zu umgehen, müssen Sie für die Zellen, die Sie kopieren möchten, einen kleinen Trick anwenden. Sie müssen ihre Inhalte mit einer Formel in den Wert umwandeln, der durch die Formel berechnet wird, und die Quellzellen transponieren, damit diese nicht nebeneinander, sondern untereinander eingefügt werde. Dazu müssen Sie folgende Schritte durchführen:

1. **Markieren Sie die Zelle oder die Zellen, die Sie kopieren möchten.**

2. **Klicken Sie die Markierung mit der rechten Maustaste an und wählen Sie im Kontextmenü den Befehl KOPIEREN oder drücken Sie ohne vorherigen Rechtsklick die Tastenkombination Strg+C.**

3. **Klicken Sie die Zelle, in die Sie die Kopie einfügen wollen, mit der rechten Maustaste an.**

4. **Klicken Sie im Kontextmenü auf INHALTE EINFÜGEN und dann unten im Menü ein weiteres Mal auf INHALTE EINFÜGEN, um das gleichnamige Dialogfeld zu öffnen (siehe Abbildung 20.12).**

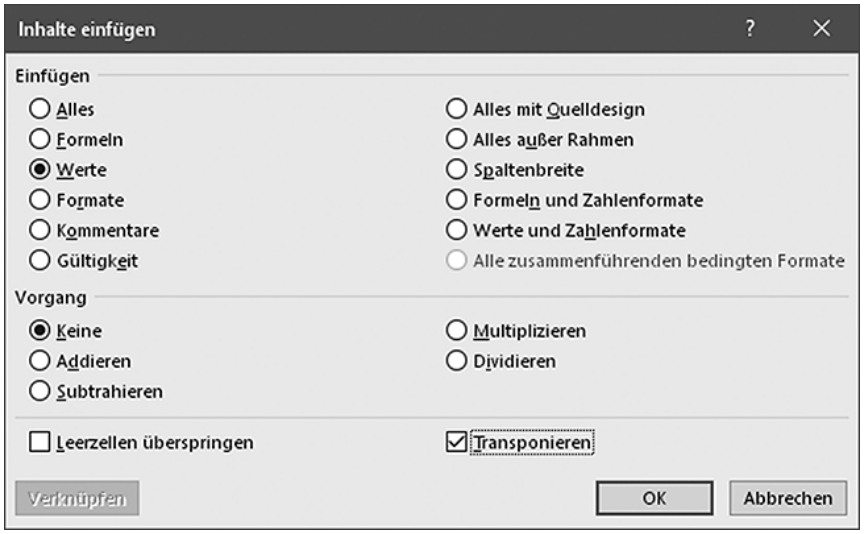


Abbildung 20.12: Das Dialogfeld INHALTE EINFÜGEN

5. **Führen Sie eine oder beide der folgenden Aktionen durch:**

   - Wenn Sie statt der Formeln die Werte einfügen wollen, schalten Sie im Abschnitt EINFÜGEN das Optionsfeld WERTE ein.

   - Wenn Sie eine Zeile in eine Spalte kopieren wollen oder umgekehrt, schalten Sie im unteren Bereich des Dialogfeldes das Kontrollkästchen TRANSPONIEREN ein.

6. **Klicken Sie auf OK, um die Umwandlung durchzuführen.**

Wenn Sie beim Einfügen der kopierten Zellen sowohl die Zellen als Werte einfügen als auch diese transponieren wollen, müssen Sie das Dialogfeld INHALTE EINFÜGEN wählen, wie oben beschrieben. Falls Sie die kopierten Zellen nur als Werte einfügen wollen oder wenn Sie Werte kopiert haben und diese beim Einfügen transponieren wollen, können Sie auch direkt die Schaltflächen im Kontextmenü verwenden (siehe Abbildung 20.13).

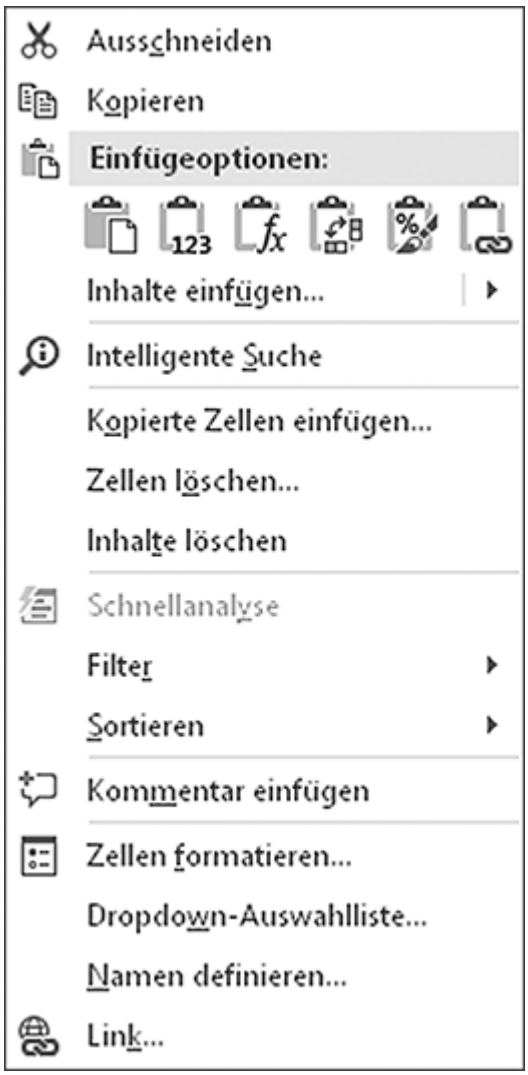


Abbildung 20.13: Das Kontextmenü, das bei einem Rechtsklick angezeigt wird, nachdem Sie mehrere Zellen markiert und kopiert haben

Klicken Sie im Bereich EINFÜGEOPTIONEN auf die Schaltfläche mit dem Klemmbrett und den Ziffern 123, um die kopierten Zellen als Werte einzufügen (die zweite Schaltfläche von links). Um Zeilen/Spalten beim Einfügen zu transponieren, klicken Sie die Schaltfläche mit dem Klemmbrett und dem abgewinkelten Doppelpfeil an (die vierte Schaltfläche von links).

Teil V

Der Top-Ten-Teil

In diesem Teil …

Tipps und Fallstricke bei der Arbeit mit Statistiken und Grafiken

Logarithmen und Exponenten verstehen

Daten sortieren

Kapitel 21

Zehn Tipps und Fallstricke bei der Arbeit mit Statistiken und Grafiken

In diesem Kapitel

- Die Signifikanz bestimmen
- Obacht vor Grafiken
- Nehmen Sie sich vor der Regression in Acht
- Setzen Sie Konzepte sorgfältig ein

Der Welt der Statistik ist voller Fallstricke, aber auch voller Möglichkeiten. Ob Sie Statistiken nur benutzen oder aber interpretieren müssen, es ist jederzeit möglich, in die Falle zu tappen. Sie können sie aber auch umgehen. Hier finden Sie zehn Tipps und Fallstricke aus den Bereichen des Testens von Hypothesen, der Regression, der Korrelation und der Grafiken.

Signifikant bedeutet nicht immer wichtig

Wie schon weiter vorn in diesem Buch gesagt, ist »Signifikanz« unter vielen Gesichtspunkten ein schlecht gewählter Ausdruck. Wenn ein statistischer Test zu einem signifikanten Ergebnis führt und die Entscheidung lautet, H_0 zu verwerfen, dann heißt das noch nicht, dass die Untersuchung, die hinter den Daten steht, wichtig ist.

Statistiken können nur bei der Entscheidungsfindung im Hinblick auf Zahlen und die Schlussfolgerungen über die Prozesse, die diese Zahlen erzeugt haben, helfen. Sie können den Prozessen selbst keine größere Wichtigkeit verleihen. Die Wichtigkeit ist etwas, was Sie für sich selbst abschätzen müssen. Kein statistischer Test kann das für Sie übernehmen.

Der Versuch, eine Nullhypothese nicht zu verwerfen, hat einige Auswirkungen

Erlauben Sie mir, Ihnen eine Geschichte zu erzählen: Vor einigen Jahren versuchte ein Industrieunternehmen zu beweisen, dass es endlich die Umweltschutzrichtlinien einhielt. Das Unternehmen führte zahlreichen Messungen durch, die die Verschmutzung des Wassers um die Fabrik herum zum Inhalt hatten, und es verglich die Messwerte mit einer Reihe von Erwartungswerten, die durch Nullhypothesen erzeugt worden waren. Es stellte fest, dass es H_0 mit $\alpha = 0{,}05$ nicht verwerfen konnte. Die Messungen wichen nicht signifikant (da haben wir das Wort schon wieder) von den Daten von »sauberem« Wasser ab.

Für das Unternehmen war das der Beweis, dass alles im Reinen war. Nähere Untersuchungen deckten jedoch auf, dass die Daten in Richtung Signifikanz strebten, die Verschmutzung jedoch nicht stark genug war, um H_0 zu verwerfen. Bedeutet das, dass das Unternehmen die Umwelt nicht verschmutzt?

Ganz und gar nicht. Bei dem Versuch, eine Nullhypothese zu »beweisen«, hat das Unternehmen die Karten zu seinen Gunsten gemischt. Es legte die Messlatte sehr hoch an, übersprang sie nicht und klopfte sich dann selbst lobend auf die Schulter.

Hin und wieder ist es gut zu versuchen, H_0 nicht zu verwerfen. Wenn Sie das tun, achten Sie darauf, für α einen hohen Wert anzusetzen (ca. 0,20 bis 0,30), damit kleine Abweichungen von H_0 dazu führen, dass H_0 verworfen wird. (Ich bespreche dieses Thema in Kapitel 10 und erwähne es auch in anderen Teilen dieses Buchs. Trotzdem halte ich es für so wichtig, dass ich es hier noch einmal wiederholen möchte.)

Regressionen sind nicht immer linear

Beim Versuch, ein Regressionsmodell auf ein Streuungsdiagramm anzuwenden, ist man schnell geneigt, sofort eine Linie zu verwenden. Das ist das am besten verständliche Regressionsmodell und sobald Sie den Dreh raushaben, verlieren Steigungen und Achsenabschnitte schnell ihren Schrecken.

Doch die lineare Regression ist nicht die einzige Regression. Sie können auch eine Kurve durch ein Streudiagramm legen. Ich will nichts beschönigen: Die statistischen Konzepte, die hinter einer krummlinigen Regression liegen, sind schwerer zu verstehen als die Konzepte, die hinter der linearen Regression stehen.

Es lohnt sich allerdings, sich die Zeit zu nehmen, diese Konzepte zu verinnerlichen. Manchmal passt eine Kurve ganz einfach besser als eine Linie. (Das ist zum Teil bereits eine Vorschau auf Kapitel 22, wo ich mich eingehend mit der krummlinigen Regression und einigen der zugrunde liegenden Konzepte beschäftigen werde.)

Extrapolieren über die Grenzen eines Streuungsdiagramms

hinaus ist eine ganz schlechte Idee

Ob Sie nun mit einer linearen oder einer krummlinigen Regression arbeiten, denken Sie auf jeden Fall daran, dass Verallgemeinerungen über die Grenzen eines Streuungsdiagramms hinaus unangebracht sind.

Nehmen wir einmal an, Sie haben eine solide Vorhersagebeziehung zwischen einem Test der mathematischen Fähigkeit und der Leistung in Mathematikkursen aufgestellt, aber Ihr Streuungsdiagramm deckt nur einen kleinen Bereich der mathematischen Fähigkeit ab. Sie können nicht wissen, ob die Beziehung auch über diesen Bereich hinaus Gültigkeit hat. Vorhersagen außerhalb des Bereichs sind ungültig.

Die beste Vorgehensweise wäre, das Streuungsdiagramm zu vergrößern, indem Sie mehr Personen testen. Möglicherweise werden Sie feststellen, dass die Ursprungsbeziehung nur ein Teil des Gesamtbildes ist.

Untersuchen Sie die Variabilität um eine Regressionsgerade herum

Eine sorgfältige Analyse der Residuen (der Unterschiede zwischen den beobachteten und den vorhergesagten Werten) kann viel darüber aussagen, wie gut die Gerade zu den Daten passt. Grundsätzlich kann man annehmen, dass die Variabilität um eine Regressionsgerade herum überall gleich ist. Ist das nicht der Fall, dann ist das Modell vielleicht doch nicht so vorhersagekräftig, wie Sie dachten. Wenn die Variabilität systematisch ist (das heißt, wenn die Variabilität an einem Ende der Geraden größer ist als am anderen Ende), wäre eine krummlinige Regression vielleicht

passender als eine lineare Regression. Der Standardfehler des Schätzwerts ist dafür nicht unbedingt immer ein Indikator.

Eine Stichprobe kann auch zu groß sein

Ob Sie's nun glauben oder nicht, aber dies passiert manchmal im Zusammenhang mit Korrelationskoeffizienten. Durch eine sehr große Stichprobe kann ein kleiner Korrelationskoeffizient statistisch betrachtet plötzlich signifikant werden. Mit 100 Freiheitsgraden und α = 0,05 beispielsweise führt ein Korrelationskoeffizient von 0,195 dazu, dass die Nullhypothese verworfen wird, dass der Korrelationskoeffizient der Grundgesamtheit gleich null ist.

Aber was bedeutet dieser Korrelationskoeffizient eigentlich? Der Determinationskoeffizient (auch Bestimmtheitsmaß genannt) – r^2 – beträgt lediglich 0,038, was bedeutet, dass die Regressionsquadratsumme $SS_{Regression}$ weniger als 4 Prozent der Gesamtquadratsumme SS_{Gesamt} ausmacht (siehe Kapitel 16). Das ist eine sehr kleine Assoziation.

Im Endeffekt läuft das alles auf Folgendes hinaus: Wenn Sie einen Korrelationskoeffizienten betrachten, müssen Sie stets die Stichprobengröße mit in Betracht ziehen. Ab einer gewissen Größe kann eine triviale Assoziation statistisch betrachtet signifikant sein. (Mmmmh … Signifikanz … da ist sie schon wieder!)

Achse ist nicht gleich Achse

Wenn Sie ein Diagramm betrachten, achten Sie stets darauf, was durch die einzelnen Achsen dargestellt wird. Gehen Sie ganz sicher, dass Sie die Maßeinheiten verstehen. Wissen Sie, was die unabhängige Variable ist? Können Sie all das mit Ihren eigenen Worten beschreiben? Wenn die Antwort auf eine dieser Fragen »Nein« lautet, verstehen Sie das Diagramm, das Sie betrachten, nicht.

Wenn in einer Fernsehwerbung ein Diagramm eingeblendet wird, sollten Sie sofort misstrauisch werden, wenn es so schnell wieder ausgeblendet wird, dass Sie die Achsen nicht richtig erkennen können. Der Auftraggeber des Werbespots versucht möglicherweise, einen falschen Eindruck von einer eher zweifelhaften Beziehung innerhalb des Diagramms zu erwecken. Die grafisch dargestellte Beziehung könnte genauso gültig sein wie das andere große Hauptthema der Fernsehwerbung – der wissenschaftliche Beweis durch einen kleinen Zeichentrickfilm: Wenn kleine Zeichentrickbürstchen kleine Zeichentrickzähnchen putzen, garantiert das noch nicht unbedingt, dass Ihre Zähne weißer werden, wenn Sie das angepriesene Produkt kaufen. (Ich weiß, das gehört nicht zum Thema, aber ich wollte es trotzdem unbedingt erwähnen.)

Eine kategorische Variable grafisch so darzustellen, als sei sie eine quantitative Variable, ist ganz einfach falsch

Sie sind also bereit, an der Stein-Schere-Papier-Weltmeisterschaft teilzunehmen. Als Vorbereitung für dieses internationale Turnier haben Sie über Ihre gesamten Wettkämpfe der letzten zehn Jahre Buch geführt und dabei aufgelistet, wie hoch Ihre prozentuelle Gewinnquote war, wenn Sie die einzelnen Figuren gespielt haben.

Um die ganzen Ergebnisse zusammenzufassen, möchten Sie die Diagramm-Features von Excel verwenden, um die Daten zu visualisieren. Eines ist sicher: Egal, was Sie bei der Stein-Schere-Papier-Frage auch bevorzugen, das Diagramm sollte auf gar keinen Fall wie das Diagramm in Abbildung 21.1 aussehen.

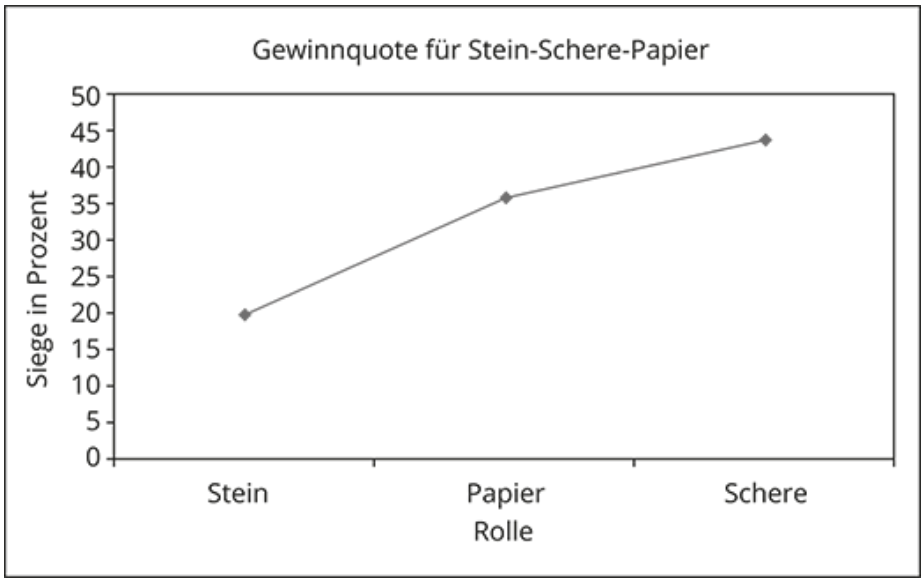


Abbildung 21.1: Der absolute falsche Weg, um kategorische Daten darzustellen

So viele Leute erstellen diese Art von Diagrammen – Leute, die es eigentlich besser wissen sollten! Die Gerade im Diagramm impliziert Kontinuität von einem Punkt zum anderen. Bei diesen Daten ist das aber natürlich vollkommen unmöglich. Was liegt zwischen Stein und Papier? Warum liegen sie gleich weit auseinander? Warum sind die drei Kategorien in dieser Reihenfolge aufgeführt? (Merken Sie, dass mich das wurmt?)

Einfach ausgedrückt, sieht es wie folgt aus: Ein Liniendiagramm ist nicht das richtige Diagramm, wenn mindestens eine Ihrer Variablen aus einer Reihe von Kategorien besteht. Erstellen Sie stattdessen lieber ein Säulendiagramm. Auch ein Kreisdiagramm würde in diesem Fall funktionieren, weil die Daten Prozentzahlen sind und Sie nur ein paar Kreisabschnitte benötigen. (In Kapitel 3 finden Sie Yogi Berras Richtlinien für Kreisabschnitte bzw. Pizzastücke.)

Als ich die erste Auflage dieses Buches schrieb, hatte ich für dieses Beispiel die etwas verrückte Idee der Stein-Schere-Papier-Weltmeisterschaft. Einige Zeit ist seither vergangen und heute gibt es diese Weltmeisterschaft sogar. (Sie wird von der World RPS Society ausgerichtet, mehr Informationen dazu im Web unter worldrps.com).

Wann immer möglich, sollten Sie Variabilität in Ihr Diagramm einschließen

Wenn die Punkte in Ihrem Diagramm Mittelwerte darstellen, sollten Sie darauf achten, dass im Diagramm die Standardfehler der einzelnen Mittelwerte enthalten sind. Dadurch gewinnt der Betrachter einen Eindruck von der Variabilität innerhalb der Daten, was ein wichtiger Aspekt der Daten ist. Hier noch ein kleiner Hinweis: In Kapitel 22 zeige ich Ihnen, wie Sie das mit Excel tun können.

Mittelwerte allein erzählen nicht immer die gesamte Geschichte. Nutzen Sie jede Gelegenheit, um Varianzen und Standardabweichungen zu untersuchen. Manchmal findet man dabei verborgene Schätze. Eine systematische Variation – beispielsweise hohe Varianzwerte kombiniert mit großen Mittelwerten – könnte ein Hinweis auf eine Beziehung sein, die Sie bisher noch nicht erkannt haben.

Seien Sie vorsichtig, wenn Sie Konzepte aus Statistikbüchern in Excel übertragen

Wenn Sie ernsthaft Statistik betreiben möchten, werden Sie vermutlich auch in ein oder zwei Statistikbücher schauen. Denken Sie daran, dass die Symbole in einigen Bereichen der Statistik nicht standardisiert sind. So wird in einigen Texten zum Beispiel eher die Abkürzung *M* als $\bar{x}$ für den Stichprobenmittelwert verwendet. In anderen Fällen wird die Abweichung vom Mittelwert einfach nur mit *x* dargestellt.

Konzepte aus Fachbüchern in die Statistikfunktionen von Excel zu übertragen, kann eine echte Herausforderung sein, was zum einen an den Fachbüchern und zum anderen an Excel liegt. Die Texte in Dialogfeldern oder in den Hilfedateien können Symbole enthalten, die von den Symbolen in den Büchern abweichen. Möglicherweise werden auch die gleichen Symbole verwendet – allerdings mit unterschiedlicher Bedeutung. Diese Diskrepanz kann dazu führen, dass Sie für einen Parameter in einem Dialogfeld etwas Falsches eingeben, was wiederum zu einem Fehler führt, der nur schwer zurückzuverfolgen ist.

Kapitel 22

Zehn Dinge (eigentlich zwölf), die einfach nicht in die anderen Kapitel passten

In diesem Kapitel

- Variabilität visuell darstellen
- Kleinigkeiten zur Wahrscheinlichkeit
- Nach Unabhängigkeit suchen
- Logarithmen
- Sortieren

Ich habe dieses Buch geschrieben, um Ihnen sämtliche statistischen Fähigkeiten von Excel vorzustellen. Ich wollte Ihnen diese Fähigkeiten im Kontext der Welt der Statistik nahe bringen und dafür hatte ich einen festen Plan im Kopf.

Einige dieser Fähigkeiten passten aber leider nicht in den Plan. Trotzdem möchte ich, dass Sie etwas darüber erfahren. Also, hier sind sie.

Diagrammdarstellung des Standardfehlers des Mittelwerts

Wenn Sie ein Diagramm erstellen und Ihre Daten Mittelwerte sind, empfiehlt es sich, die Standardfehler der einzelnen Mittelwerte mit in

das Diagramm aufzunehmen. Der Betrachter erhält dadurch einen Eindruck von der Ausbreitung der Werte um die einzelnen Mittelwerte herum.

In Abbildung 22.1 finden Sie ein Beispiel für eine Situation, in der das eintritt. Die Daten sind (fiktive) Ergebnisse, die von vier Gruppen von Menschen bei einem Test erzielt werden. In den Spaltenüberschriften finden Sie die Vorbereitungszeit, die die acht Personen innerhalb der einzelnen Gruppen aufgebracht haben. Das Diagramm habe ich mit den Diagrammfeatures von Excel (siehe Kapitel 3) erstellt. Da die unabhängige Variable quantitativ ist, konnte ich ein Liniendiagramm verwenden. (Nähere Informationen dazu finden Sie in Kapitel 19.)

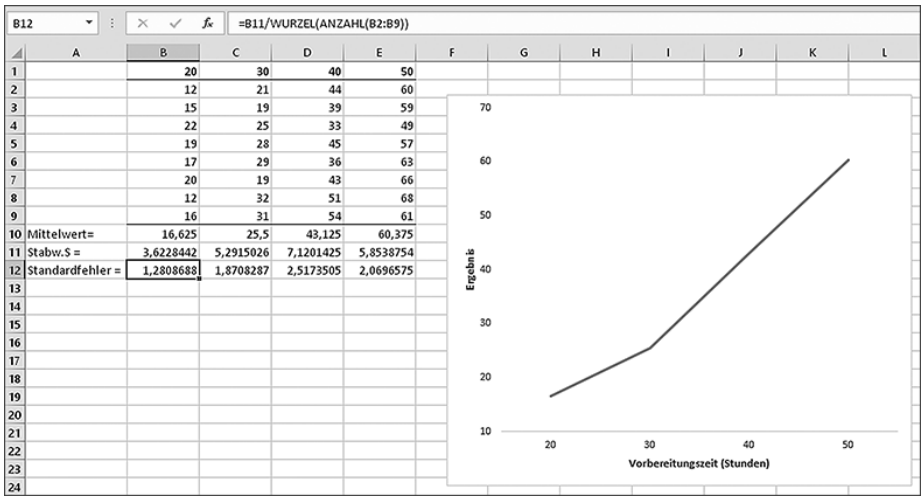


Abbildung 22.1: Vier Gruppen – ihre Mittelwerte, Standardabweichungen und Standardfehler. Im Diagramm sind die Gruppenmittelwerte dargestellt.

Für die jeweiligen Gruppen habe ich die Funktion `MITTELWERT` zur Berechnung des Mittelwerts und die Funktion `STABW.S` zur Berechnung der Standardabweichung angewendet. Darüber hinaus habe ich für die einzelnen Mittelwerte auch den jeweiligen Standardfehler berechnet. In der Abbildung habe ich Zelle B12 markiert, damit Sie in der Bearbeitungsleiste sehen können, dass ich den Standardfehler für Spalte B mit dieser Formel berechnet habe:

```
=B11/WURZEL(ANZAHL(B2:B9))
```

Der Trick besteht nun darin, die einzelnen Standardfehler in das Diagramm zu integrieren. In Excel 2016 lässt sich dies einfach erledigen und die Vorgehensweise unterscheidet sich von früheren Excel-Versionen. Markieren Sie zuerst das Diagramm. Excel blendet dann die kontextbezogenen Registerkarten DIAGRAMMTOOLS/ENTWURF und DIAGRAMMTOOLS/FORMAT ein. Wählen Sie dann den Befehl ENTWURF | DIAGRAMMELEMENT HINZUFÜGEN | FEHLERINDIKATOREN | WEITERE FEHLERINDIKATORENOPTIONEN. Abbildung 22.2 zeigt, was ich meine.

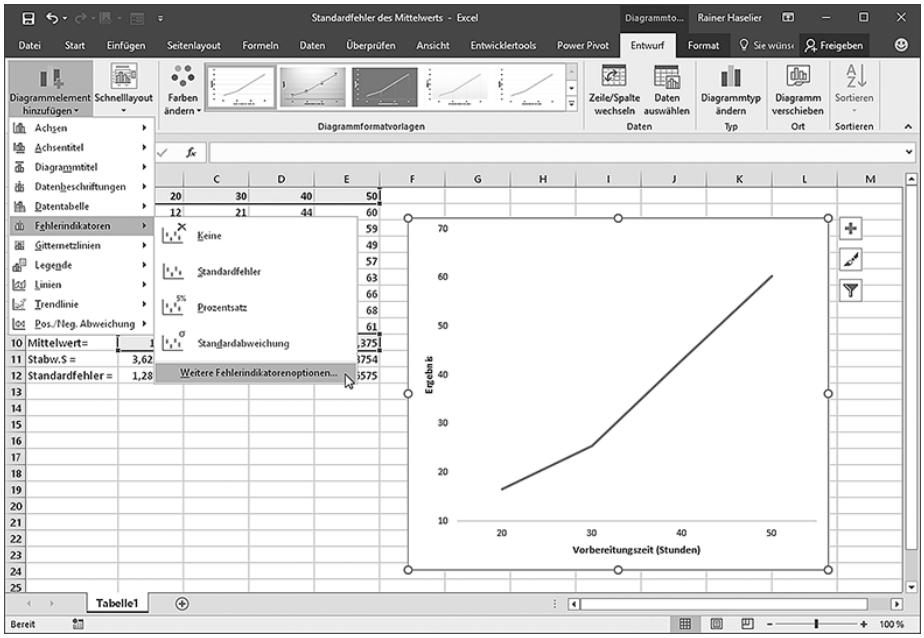


Abbildung 22.2: Der Befehlspfad zum Einfügen von Fehlerindikatoren

Im Menü FEHLERINDIKATOREN müssen Sie aufpassen. Einer der möglichen Befehle lautet STANDARDFEHLER. Vermeiden Sie ihn. Wenn Sie glauben, dass Sie Excel durch Anklicken dieses Befehls dazu bewegen, den Standardfehler der einzelnen Mittelwerte in das Diagramm einzufügen, kann ich Ihnen versichern, dass Excel absolut keine Ahnung hat, wovon Sie überhaupt reden. Wenn Sie die Option auswählen, berechnet Excel den Standardfehler der Gesamtgruppe der vier Mittelwerte und nicht den Standardfehler der einzelnen Gruppen.

Verwenden Sie stattdessen den Befehl WEITERE FEHLERINDIKATORENOPTIONEN. Hierdurch wird der Aufgabenbereich FEHLERINDIKATOREN FORMATIEREN geöffnet (siehe Abbildung 22.3).

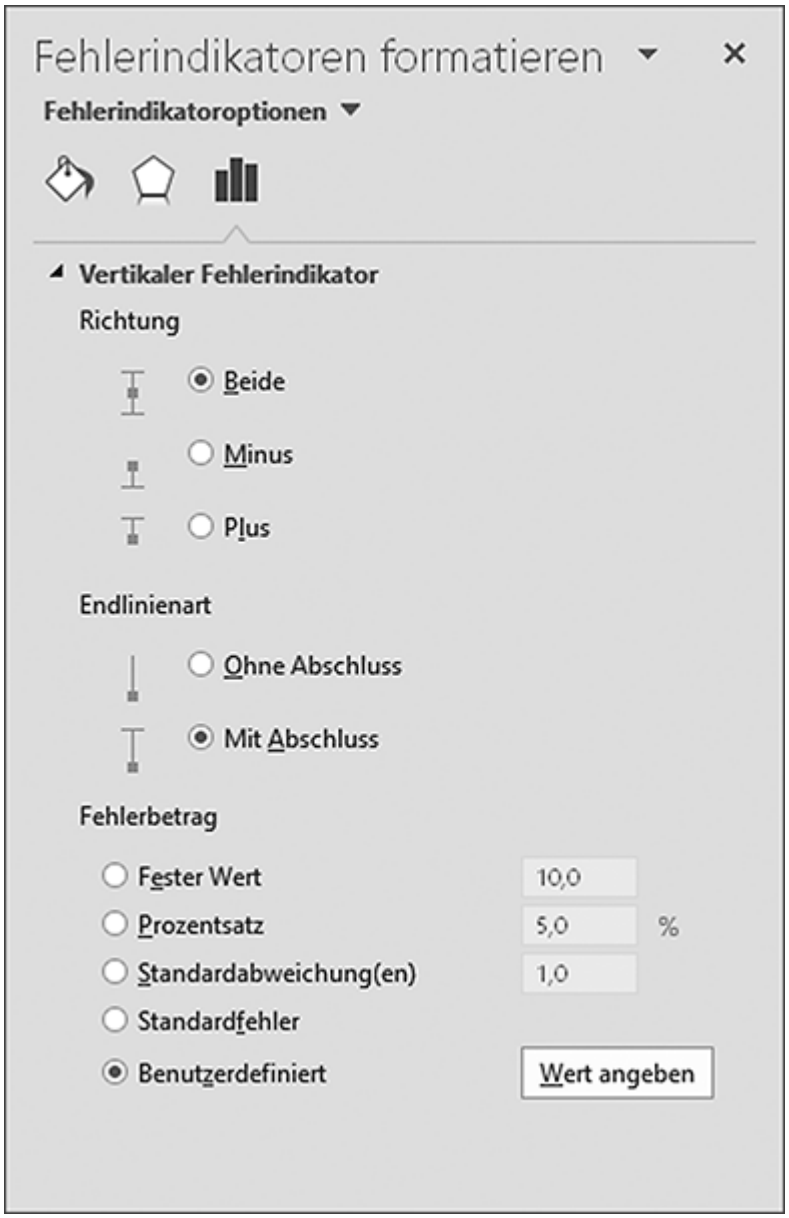


Abbildung 22.3: Der Aufgabenbereich FEHLERINDIKATOREN FORMATIEREN

Wählen Sie im Bereich RICHTUNG des Aufgabenbereichs das Optionsfeld BEIDE und im Bereich ENDLINIENART das Optionsfeld MIT ABSCHLUSS.

Erinnern Sie sich noch an die Warnung, die ich ein paar Absätze weiter vorne aussprach? Hier eine ähnliche Warnung. Ein Optionsfeld im Bereich FEHLERBEITRAG ist mit STANDARDFEHLER beschriftet. Vermeiden Sie ebenfalls diese Option. Hiermit weisen Sie Excel nicht an, den Standardfehler der Mittelwerte in das Diagramm einzufügen.

Blättern Sie im Bereich FEHLERBEITRAG nach unten und schalten Sie das Optionsfeld neben BENUTZERDEFINIERT ein. Hierdurch wird die Schaltfläche WERT ANGEBEN aktiviert. Klicken Sie diese Schaltfläche an, um das Dialogfeld BENUTZERDEFINIERTE FEHLERINDIKATOREN angezeigt (siehe Abbildung 22.4). Markieren Sie, während sich die Einfügemarke im Feld POSITIVER FEHLERWERT befindet, den Zellbereich, in dem sich die Standardfehler befinden (`$B$12:$E$12`). Drücken Sie Tab, um zum Feld NEGATIVER FEHLERWERT zu gelangen, und tragen Sie hier den gleichen Zellbereich ein.

Abbildung 22.4: Das Dialogfeld »Benutzerdefinierte Fehlerindikatoren«

Möglicherweise verursacht das Feld NEGATIVER FEHLERWERT kleine Probleme. Achten Sie darauf, dass das Feld leer ist, bevor Sie den Zellbereich eingeben.

Klicken Sie im Dialogfeld BENUTZERDEFINIERTE FEHLERINDIKATOREN auf OK und schließen Sie den Aufgabenbereich FEHLERINDIKATOREN FORMATIEREN. Das resultierende Diagramm sieht wie in Abbildung 22.5 dargestellt aus.

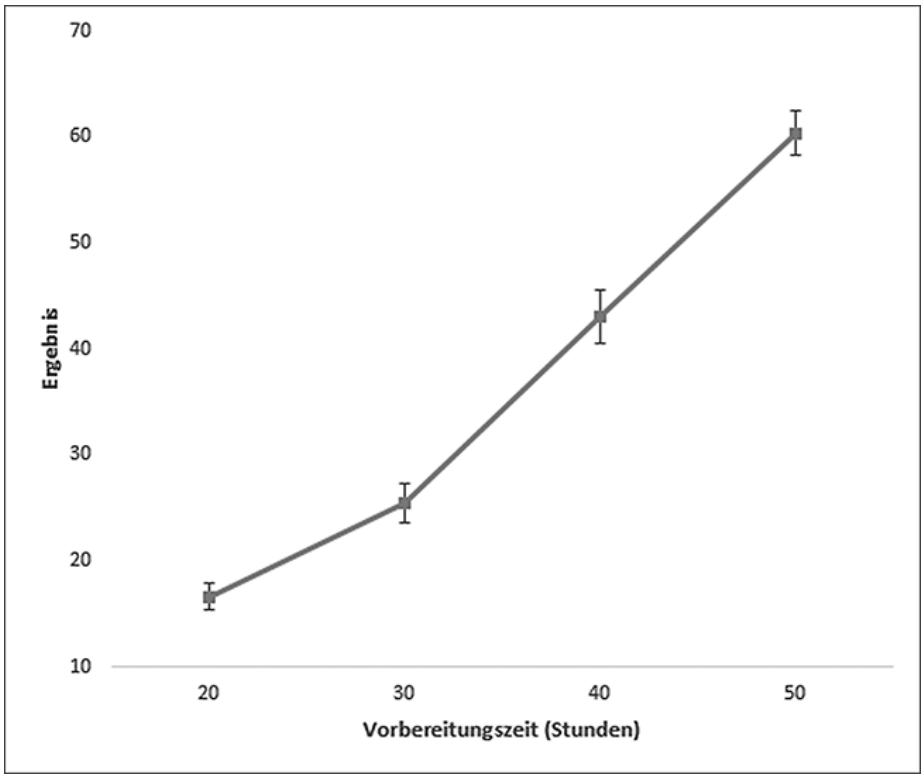


Abbildung 22.5: Das Diagramm der Gruppenmittelwerte mit den Standardfehlern der einzelnen Mittelwerte

Wahrscheinlichkeiten und Verteilungen

Im Folgenden stelle ich Ihnen einige Tabellenblattfunktionen vor, die mit der Wahrscheinlichkeit in Zusammenhang stehen. Sie sind ein wenig exotisch, aber vielleicht haben Sie ja trotzdem Verwendung für sie.

WAHRSCHBEREICH

Wenn Sie eine Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen haben und die Wahrscheinlichkeit bestimmen

möchten, dass die Variable einen bestimmten Wert annimmt, ist die Funktion WAHRSCHBEREICH genau das Richtige für Sie. In Abbildung 22.6 ist das Dialogfeld FUNKTIONSARGUMENTE der Funktion WAHRSCHBEREICH zusammen mit einer Verteilung dargestellt.

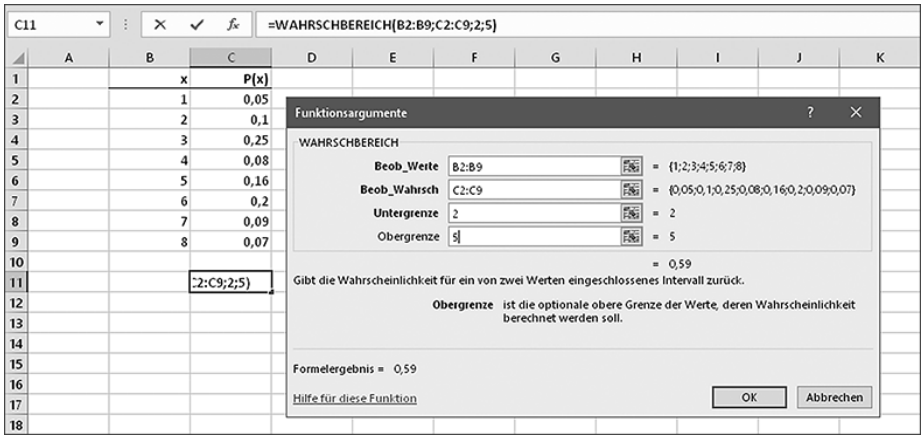


Abbildung 22.6: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion WAHRSCHBEREICH und eine Wahrscheinlichkeitsverteilung

Sie müssen Werte für die Zufallsvariable (BEOB_WERTE), die Wahrscheinlichkeiten (BEOB_WAHRSCH), eine UNTERGRENZE und eine OBERGRENZE eingeben. Die Funktion WAHRSCHBEREICH gibt die Wahrscheinlichkeit zurück, dass die Zufallsvariable einen Wert innerhalb dieser Grenzen – einschließlich der Grenzwerte – annimmt.

Wenn Sie das Feld OBERGRENZE leer lassen, gibt WAHRSCHBEREICH die Wahrscheinlichkeit des Wertes zurück, den Sie bei UNTERGRENZE angegeben haben. Wenn Sie das Feld UNTERGRENZE leer lassen, gibt WAHRSCHBEREICH die Wahrscheinlichkeit zurück, dass die Zufallsvariable höchstens den als OBERGRENZE angegeben Wert annimmt (beispielsweise die kumulative Wahrscheinlichkeit).

WEIBULL.VERT

Diese Wahrscheinlichkeitsdichtefunktion ist besonders für das Ingenieurswesen geeignet. Sie dient als Modell für die Zeit, die vergeht, bis ein physikalisches System ausfällt. Ingenieure wissen, dass die Anzahl der Ausfälle bei einigen Systemen im Verlauf der Zeit gleich bleibt, da Erschütterungen des Systems den Ausfall hervorrufen. Bei anderen Systemen, wie zum Beispiel im Fall von Mikroelektronikbauteilen, nimmt die Anzahl der Ausfälle im Laufe der Zeit ab. Bei wieder anderen Systemen hingegen kommt es mit der Zeit durch Verschleiß zu einem Anstieg der Ausfälle.

Durch die beiden Parameter der Weibull-Verteilung können all diese Möglichkeiten berücksichtigt werden. Ein Parameter, ALPHA, bestimmt, wie weit oder eng die Verteilung gefasst ist. Der andere Parameter, BETA, gibt an, wo die Verteilung auf der X-Achse zentriert ist.

Die Weibull-Wahrscheinlichkeitsdichtefunktion ist eine ziemlich komplizierte Gleichung. Dank Excel müssen Sie sich darüber aber keine Gedanken machen. In Abbildung 22.7 ist das Dialogfeld FUNKTIONSARGUMENTE der Funktion `WEIBULL.VERT` dargestellt.

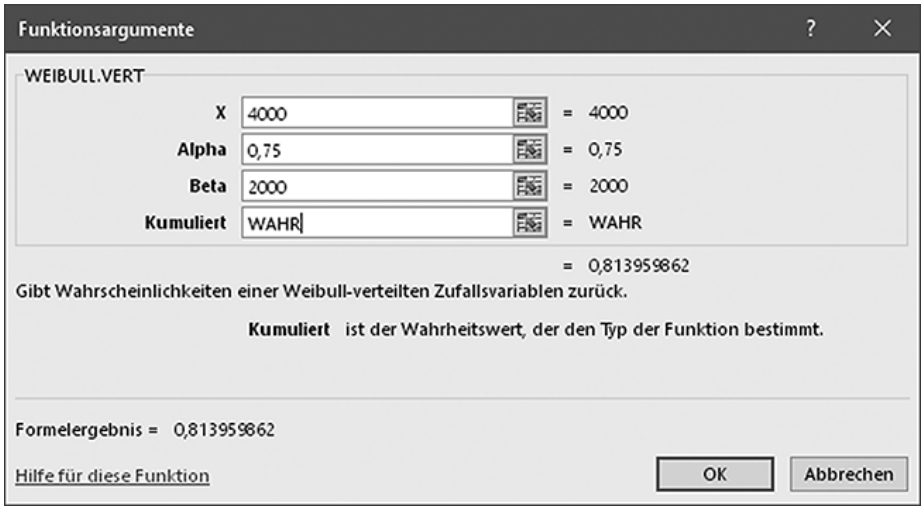


Abbildung 22.7: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `WEIBULL.VERT`

Das Dialogfeld in der Abbildung liefert die Antworten auf alle Fragen, die ein Ingenieur stellen würde: Nehmen wir an, die Zeit bis zum Ausfall einer Glühbirne in einem LCD-Projektor folgt einer Weibull-

Verteilung mit = 0,75 und = 2000 Stunden. Wie hoch ist die Wahrscheinlichkeit, dass die Glühbirne höchstens 4000 Stunden hält? Im Dialogfeld ist angegeben, dass die Wahrscheinlichkeit 0,814 beträgt.

Stichproben ziehen

Die Excel-Analyse-Funktion Stichprobenziehung ist sehr nützlich, wenn man Stichproben erstellen muss. Sie können die Funktion auf unterschiedliche Art und Weise nutzen. Wenn Sie beispielsweise eine Fokusgruppe zusammenstellen und die Teilnehmer aus einer Gruppe von Personen auswählen müssen, könnten Sie jeder Person eine Nummer zuweisen und dann die Analyse-Funktion Stichprobenziehung verwenden, um die Gruppen zusammenzustellen.

Eine Möglichkeit ist die *periodische* Auswahl. Sie geben *n* vor und Excel zieht dann jede *n*-te Zahl als Stichprobe. Eine andere Möglichkeit ist die *zufällige* Auswahl. Hier geben Sie vor, wie viele Einzelelemente Sie benötigen, und Excel übernimmt den Rest.

In Abbildung 22.8 sind das Dialogfeld STICHPROBENZIEHUNG, drei Gruppen, aus denen die Stichproben gezogen werden sollen, sowie zwei Ausgabespalten dargestellt.

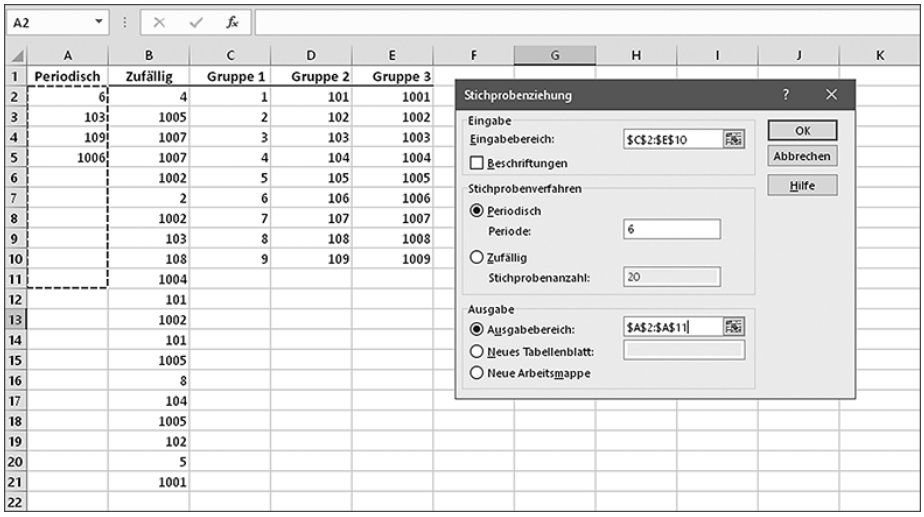


Abbildung 22.8: Das Dialogfeld der Analyse-Funktion STICHPROBENZIEHUNG, drei Gruppen, die die Grundlage für die Stichprobenziehung bilden, und die

entsprechenden Ergebnisse

Die erste Ausgabespalte, Spalte A, zeigt die Ergebnisse der periodischen Stichprobenziehung mit der Periode 6. Die Stichprobenziehung beginnt mit dem sechsten Wert in Gruppe 1. Excel zählt dann die einzelnen Einträge weiter und gibt den jeweils sechsten Eintrag aus. Das wird so lange wiederholt, bis das Ende der letzten Gruppe erreicht ist. Wie Sie sehen können, wiederholt sich beim periodischen Verfahren der Zyklus nicht. Als Ausgabebereich habe ich einen Bereich bis zur Zelle A11 angegeben, aber Excel stoppt schon nach vier Zahlen.

Die zweite Ausgabespalte, Spalte B, enthält die Ergebnisse der zufälligen Stichprobenziehung. Ich hatte vorgegeben, dass ich zwanzig Werte benötige, und genau das hat Excel mir auch geliefert. Wenn Sie sich die Zahlen in Spalte B genau ansehen, werden Sie feststellen, dass Zahlen bei diesem Verfahren mehrfach ausgewählt werden können.

Seien Sie vorsichtig: Wenn Sie das Kontrollkästchen BESCHRIFTUNGEN aktivieren, scheint es so, als hätte das keinerlei Auswirkungen. Wenn ich dann aber einen Eingabebereich definiere, der die Zellen C1, D1 und E1 enthält, erhalte ich eine Fehlermeldung mit dem Hinweis, dass der Eingabebereich nicht-numerische Daten umfasst. Kein Drama, aber doch ein wenig störend.

Unabhängigkeit testen: die wahre Verwendung von CHIQU.TEST

In Kapitel 20 beschreibe ich, wie Sie die Funktion `CHIQU.TEST` verwenden können, um zu prüfen, wie gut ein Modell zu einer

Gruppe von Daten passt. In diesem Kapitel warne ich Sie auch vor den Fallstricken, die mit der Verwendung der Funktion in diesem Zusammenhang einhergehen, und ich erwähne auch, dass die Funktion eigentlich einem anderen Zweck dient.

Und hier ist er nun, der andere Zweck. Stellen Sie sich vor, Sie haben im Rahmen einer Umfrage 200 Personen befragt. Jede dieser Personen lebt entweder auf dem Land, in der Stadt oder in einem Vorort. In der Umfrage werden sie nach ihrem bevorzugten Filmgenre gefragt. Als Auswahlmöglichkeiten gibt es dabei die Optionen Drama, Komödie oder Zeichentrickfilm. Jetzt möchten Sie wissen, ob das bevorzugte Filmgenre unabhängig von der Umgebung ist, in der die Personen leben.

Die Ergebnisse sind in Tabelle 22.1 dargestellt.

	Drama	Komödie	Zeichentrickfilm	Gesamt
Land	40	30	10	80
Stadt	20	30	20	70
Vorort	10	20	20	50
Gesamt	70	80	50	200

Tabelle 22.1: Wohnsituation und bevorzugtes Filmgenre

Die Zahlen in den einzelnen Tabellenzellen stehen für die Anzahl der Personen mit dem in der jeweiligen Zeile aufgeführten Wohnort, die das in der jeweiligen Spalte genannte Filmgenre bevorzugen.

Zeigen die Daten an, dass das bevorzugte Filmgenre unabhängig von der Wohnsituation ist? Diese Frage schreit geradezu nach einem Hypothesentest:

> H_0: Das bevorzugte Filmgenre ist unabhängig von der Wohnsituation.
> H_1: Nicht H_0
> $\alpha = 0{,}05$

Um den Test erfolgreich durchzuführen, müssen Sie erst einmal wissen, was Sie erwartet, wenn es unabhängig voneinander ist. Dann vergleichen Sie die Daten mit den erwarteten Zahlen und prüfen, ob sie zueinander passen. Ist das der Fall, dann können Sie H_0 nicht verwerfen. Wenn sie nicht zueinander passen, verwerfen Sie H_0.

Konzepte aus dem Bereich der Wahrscheinlichkeit helfen Ihnen bei der Bestimmung der erwarteten Daten. In Kapitel 18 erläutere ich, dass Sie bei zwei voneinander unabhängigen Ereignissen ihre jeweiligen Wahrscheinlichkeiten miteinander multiplizieren, um die Wahrscheinlichkeit zu bestimmen, dass beide Ereignisse gleichzeitig eintreten. In diesem Fall hier können Sie die Tabellenwerte wie Anteile und die Anteile wiederum wie Wahrscheinlichkeiten behandeln.

In Ihrer Stichprobe beträgt die Wahrscheinlichkeit, dass eine Person in einer ländlichen Umgebung lebt, zum Beispiel 80/200. Die Wahrscheinlichkeit, dass eine Person das Filmgenre Drama bevorzugt, beträgt 70/200. Wie hoch ist die Wahrscheinlichkeit, dass eine Person zur Kategorie »kommt vom Land und mag Dramen« gehört? Wenn die Wohnsituation und das bevorzugte Filmgenre unabhängig voneinander sind, gilt (80/200) × (70/200). Um aus dieser Wahrscheinlichkeit eine erwartete Anzahl von Personen zu erhalten, multiplizieren Sie sie mit der Gesamtzahl der Personen in der Stichprobe, also mit 200. Die erwartete Anzahl Personen ist demnach (80 × 70)/200, also 28.

Im Allgemeinen gilt

$$\text{Erwartete Zahl in einer Zelle} = \frac{\text{ZeileGesamt} \times \text{SpalteGesamt}}{\text{Gesamt}}$$

Wenn Sie die erwarteten Zahlen berechnet haben, vergleichen Sie sie mit der folgenden Formel mit den tatsächlich beobachteten Zahlen (den Daten):

$$\chi^2 = \Sigma\left[\frac{(\text{Beobachtet} - \text{Erwartet})^2}{\text{Erwartet}}\right]$$

Sie prüfen das Ergebnis anhand einer χ^2-Verteilung (Chi-Quadrat-Verteilung) mit df = (Anzahl der Zeilen – 1) × (Anzahl der Spalten – 1), was in diesem Fall das Ergebnis 4 liefert.

Die Funktion `CHIQU.TEST` führt den Test durch. Sie geben die beobachteten Werte und die erwarteten Werte ein und `CHIQU.TEST` gibt die Wahrscheinlichkeit zurück, dass ein χ^2 herauskommt, das mindestens so groß wie das Ergebnis der vorstehenden Formel ist, wenn die beiden Kategorietypen unabhängig voneinander sind. Ist die Wahrscheinlichkeit gering (unter 0,05), verwerfen Sie H_0. Wenn nicht, verwerfen Sie die Hypothese nicht. `CHIQU.TEST` gibt keinen Wert für χ^2 zurück, sondern nur die Wahrscheinlichkeit (unter einer χ^2-Verteilung mit dem korrekten Wert für df).

In Abbildung 22.9 ist ein Arbeitsblatt mit den beobachteten Daten und den Erwartungswerten dargestellt. Ebenfalls zu sehen ist das Dialogfeld FUNKTIONSARGUMENTE für die Funktion `CHIQU.TEST`. Bevor ich das Dialogfeld öffnete, habe ich dem Zellbereich C3:E5 den Namen `Beobachtet` und dem Zellbereich C10:E12 den Namen `Erwartet` zugewiesen. (Werfen Sie einen Blick in Kapitel 2, wenn Sie nicht mehr genau wissen, wie das geht.)

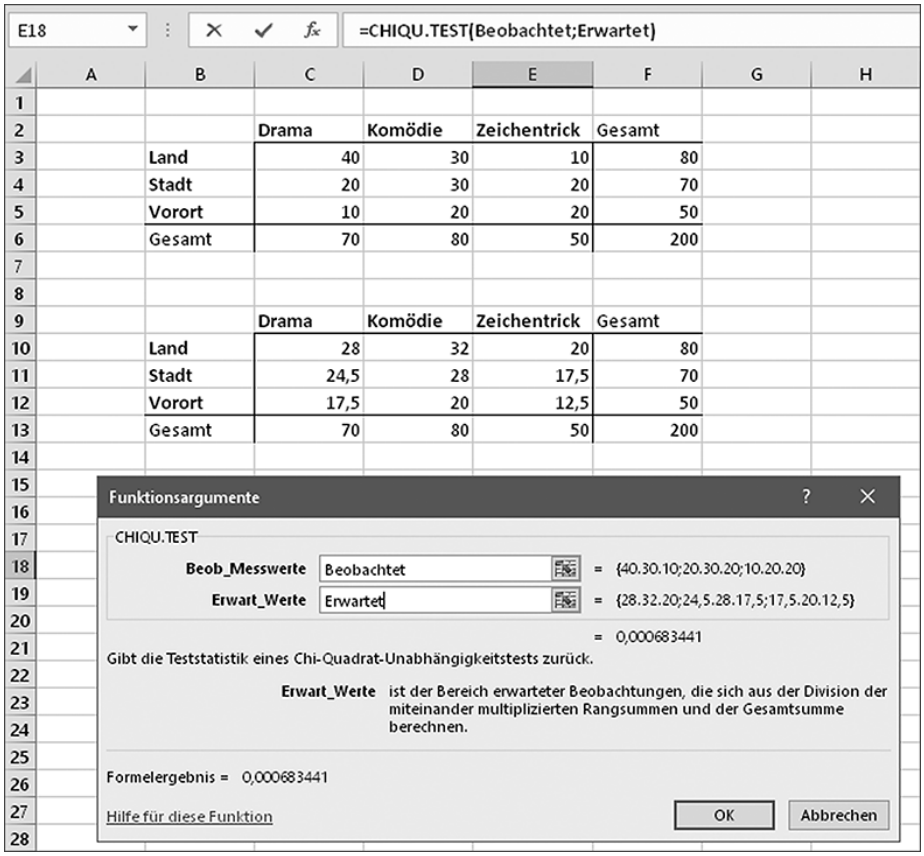


Abbildung 22.9: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion `CHIQU.TEST` mit den beobachteten und erwarteten Werten

In der Abbildung ist zu sehen, dass ich `Beobachtet` in das Feld BEOB_MESSWERTE und `Erwartet` in das Eingabefeld ERWART_WERTE eingegeben habe. Im Dialogfeld wird eine sehr geringe Wahrscheinlichkeit, nämlich 0,00068, angezeigt, so dass die Entscheidung lauten muss, H_0 zu verwerfen. Die Daten stimmen mit der Annahme überein, dass das bevorzugte Filmgenre von der Wohnsituation abhängt.

Logarithmica Esoterica

Die Funktionen in diesem Abschnitt sind *wirklich* abgehoben. Deshalb habe ich sie mir auch bis zum Schluss aufgehoben.

Wenn Sie nicht zufällig ein totaler Technik-Freak sind, werden Sie sie vermutlich niemals anwenden. Ich stelle sie Ihnen aber trotzdem aus Gründen der Vollständigkeit vor. Sie könnten Ihnen begegnen, wenn Sie sich die Statistikfunktionen von Excel ansehen, und dann würden Sie sich vielleicht fragen, was sich hinter diesen exotischen Funktionen eigentlich verbirgt.

Sie basieren auf Phänomenen, die Mathematiker als *natürliche Logarithmen* bezeichnen und die wiederum auf *e* basieren, der Konstanten, die ich bereits mehrfach in diesem Buch erwähnt habe. Ich beginne nun mit einer kurzen Erklärung der Logarithmen und wende mich dann *e* zu.

Was ist ein Logarithmus?

Ein Logarithmus ist schlicht und einfach ein *Exponent* – eine Potenz, in die Sie eine Zahl erheben. In der Gleichung

$$10^2 = 100$$

ist 2 ein Exponent. Bedeutet das, dass 2 auch ein Logarithmus ist? Nun … ja. Logarithmisch ausgedruckt heißt das

$$\log_{10}100 = 2$$

Das ist einfach nur eine andere Art und Weise auszudrücken, dass $10^2 = 100$ ist. Mathematiker lesen das so: »Der Logarithmus von 100 zur Basis 10 ist gleich 2.« Es bedeutet Folgendes: Wenn Sie 10 in eine Potenz erheben wollen, um als Ergebnis 100 zu erhalten, ist diese Potenz 2.

Wie sieht das mit 1000 aus? Wie Sie wissen, ist

$$10^3 = 1000$$

also gilt

$$\log_{10}1000 = 3$$

Und wie steht's mit 453? Äh … mmmmh … das ist ja so, als würde man versuchen, die folgende Aufgabe zu lösen:

$$10^x = 453$$

Wie könnte die Antwort lauten? 10^2 bedeutet 10 10 und das wiederum ergibt 100. 10^3 bedeutet 10 10 10 und das ergibt 1000. Aber 453?

Und genau an dieser Stelle müssen Sie sich gedanklich außerhalb des Dialogfeldes bewegen. Sie müssen sich Exponenten vorstellen, die keine ganzen Zahlen sind. Ich weiß, ich weiß … wie kann man eine Zahl nur bruchstückweise mit sich selbst multiplizieren? Wenn Sie das irgendwie könnten, würde die Zahl in der 435er-Gleichung irgendwo zwischen 2 (was 100 ergibt) und 3 (was 1000 ergibt) liegen.

Im sechzehnten Jahrhundert demonstrierte der Mathematiker John Napier, wie es geht. Das war die Geburtsstunde der Logarithmen. Warum hat sich Napier überhaupt damit auseinander gesetzt? Ein Grund war, dass es eine große Hilfe für Astronomen darstellte. Astronomen müssen sich mit … nun ja … astronomisch großen Zahlen beschäftigen. Durch Logarithmen verringert sich der Rechenaufwand gleich mehrfach. Zum einen können große Zahlen durch kleine Zahlen ersetzt werden: Der Logarithmus von 1.000.000 ist 6 und der Logarithmus von 100.000.000 ist 8. Außerdem ermöglichen Logarithmen eine Reihe nützlicher Abkürzungsmöglichkeiten. Vor der Erfindung von Rechenmaschinen und Computern war das ein echter Fortschritt.

Zufällig gilt

$$10^{2,6560982} = 453$$

was bedeutet, dass

$$\log_{10}453 = 2,6560982$$

ist. Sie können das gern mit Excel überprüfen, wenn Sie mir nicht glauben. Markieren Sie eine Zelle und geben Sie

```
=LOG(453,10)
```

ein, drücken Sie Enter und sehen Sie, was passiert. Anschließend sollten Sie, um den Kreis zu schließen, den Prozess umkehren. Wenn Sie beispielsweise ursprünglich Zelle D3 ausgewählt hatten, markieren Sie nun eine andere Zelle und geben Sie

```
=POTENZ(10;D3)
```

oder

```
=10^D3
```

ein. Egal, für welche Variante Sie sich entscheiden, das Ergebnis lautet 453.

Zehn, die Zahl, die in den Exponent erhoben wird, wird als *Basis* bezeichnet. Da sie auch die Basis unseres Zahlensystems ist und wir alle damit vertraut sind, werden Logarithmen mit der Basis 10 als *Zehnerlogarithmen* bezeichnet.

Heißt das, dass man auch andere Basen haben kann? Selbstverständlich. *Jede* Zahl (mit Ausnahme von 0 und 1 oder einer negativen Zahl) kann eine Basis sein. Beispielsweise gilt

$6{,}4^2 = 40{,}96$

also gilt auch

$\log_{6{,}4} 40{,}96 = 2$

Wenn Sie je den Ausdruck *log* ohne Basis sehen, gilt die 10 als Basis. Daher gilt

$\log 100 = 2$

Im Zusammenhang mit der Basis spielt eine Zahl eine besondere Rolle …

Was ist e?

Das führt mich zu *e*, einer Konstanten, die mit Wachstum im Zusammenhang steht. Bevor ich mich wieder den Logarithmen zuwende, berichte ich Ihnen erst einmal über *e*.

Stellen Sie sich vor, dass auf einem Bankkonto die fürstliche Summe von einem Euro hinterlegt ist. Nehmen wir einmal an, der Zinssatz beträgt 2 Prozent pro Jahr. (Viel Glück *damit*.) Wenn es ein einfacher Zinssatz ist, überweist die Bank jedes Jahr 0,02 Euro auf das Konto, so dass Sie nach 50 Jahren über 2 Euro verfügen.

Wenn es sich aber um Zinseszins handelt, haben Sie nach 50 Jahren $(1 + 0{,}02)^{50}$, was etwas mehr als 2,68 Euro ist, vorausgesetzt, die Bank zahlt die Zinsen einmal im Jahr.

Wenn sie die Zinsen allerdings zwei Mal im Jahr auszahlt, beläuft sich jede Einzelzahlung auf 0,01 Euro und die Bank hat die Zinsen nach 50 Jahren insgesamt 100 Mal ausgezahlt. Das bedeutet $(1 + 0{,}01)^{100}$ bzw. knapp über 2,70 Euro. Was passiert, wenn die Zinsen vier Mal pro Jahr ausgezahlt werden? Nach 50 Jahren – und 200 Auszahlungen – haben Sie $(1 + 0{,}005)^{200}$, also 2,71 Euro und ein paar Zerquetschte. Aber geben Sie nicht alles auf einmal aus!

Wenn wir uns auf das »etwas mehr« und »knapp über« konzentrieren und es ein wenig auf die Spitze treiben, sieht es wie folgt aus: Nach 100.000 Verzinsungen verfügen Sie über 2,718268 Euro und nach 100 Millionen Verzinsungen sage und schreibe über 2,718282 Euro.

Wenn Sie die Bank dazu bewegen könnten, das Geld in den 50 Jahren noch viel, viel öfter zu verzinsen, würde sich der Geldbetrag einer *Grenze* annähern – einer Summe, an die er nah herankommt, aber die er dennoch nie ganz erreicht. Diese Grenze ist *e*.

So, wie ich das Beispiel aufgezogen habe, lautet die Regel für die Berechnung des Betrags

$$(1 + (1/n))\ n$$

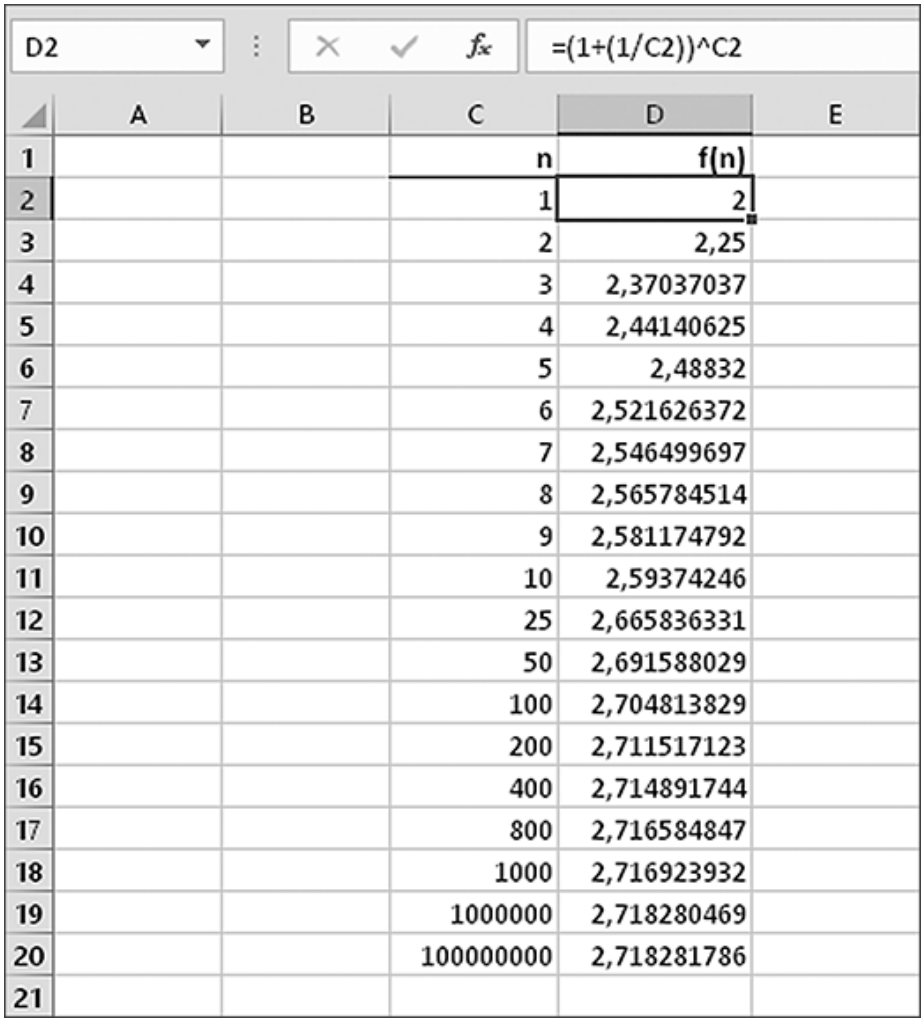
mit *n* als Anzahl der Zahlungen. Zwei Cent sind 1/50tel eines Euros und ich habe zunächst 50 Jahre festgelegt, also 50 Zahlungen. Dann habe ich zwei jährliche Zahlungen festgelegt (und die Zahlungen

eines Jahres müssen insgesamt jeweils 2 Prozent ergeben). In diesem Fall erfolgen in 50 Jahren 100 Zahlungen von 1/100tel Euro usw.

Um das live in Aktion zu sehen, müssen Sie die Zahlen in eine Spalte eines Arbeitsblatts eingeben, wie ich das in Abbildung 22.10 getan habe. In die Zellen C2 bis C20 habe ich die Zahlen 1 bis 10 eingegeben. Dann habe ich schrittweise bis hin zu 100 Millionen weitergemacht. In Zelle D2 habe ich die Formel

```
=(1+(1/C2))^C2
```

eingegeben und die restlichen Zellen bis D20 automatisch damit gefüllt. Der Wert in Zelle D20 liegt sehr nah bei *e*.



	A	B	C	D	E
1			n	f(n)	
2			1	2	
3			2	2,25	
4			3	2,37037037	
5			4	2,44140625	
6			5	2,48832	
7			6	2,521626372	
8			7	2,546499697	
9			8	2,565784514	
10			9	2,581174792	
11			10	2,59374246	
12			25	2,665836331	
13			50	2,691588029	
14			100	2,704813829	
15			200	2,711517123	
16			400	2,714891744	
17			800	2,716584847	
18			1000	2,716923932	
19			1000000	2,718280469	
20			100000000	2,718281786	
21					

Abbildung 22.10: Der Weg zu e

Mathematiker kennen noch einen anderen Weg hin zu *e*:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots$$

Die Ausrufezeichen bedeuten *Fakultät*. 1! = 1, 2! = 2 1, 3! = 3 2 1. (Weitere Informationen zu Fakultäten finden Sie in Kapitel 16.)

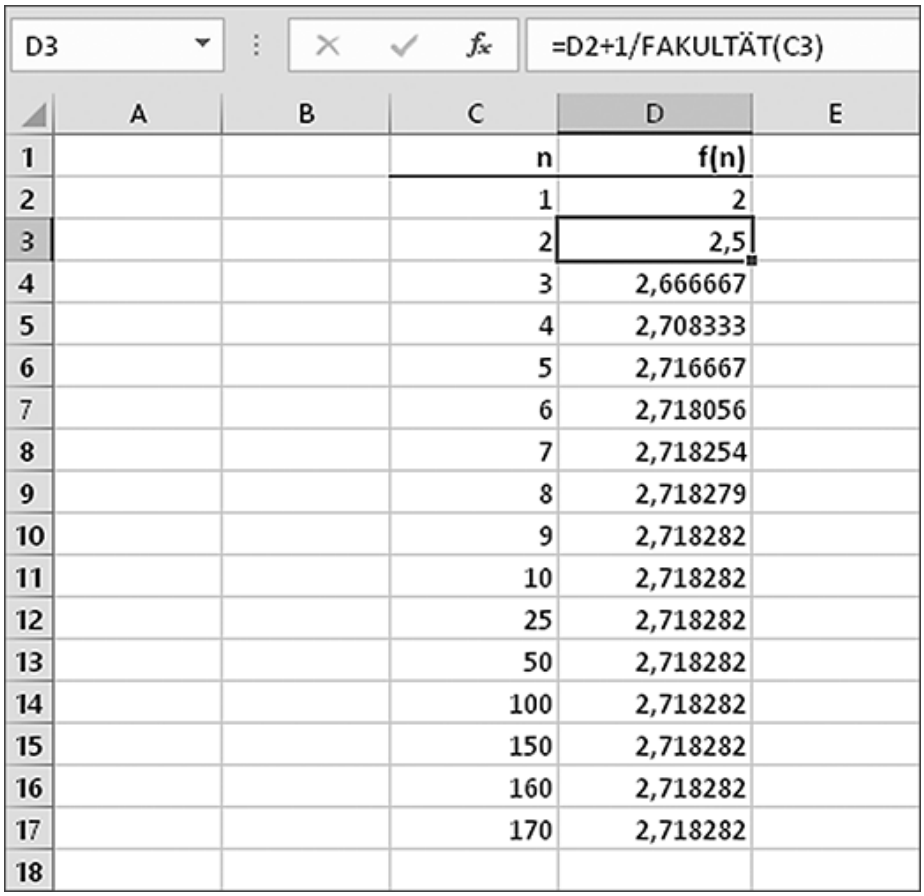
Auch das können Sie mit Excel visuell darstellen. In Abbildung 22.11 ist ein Arbeitsblatt dargestellt, das Zahlen bis zur Zahl 170 in Spalte C enthält. In Spalte D2 habe ich die folgende Formel eingegeben:

```
=1+1/FAKULTÄT(C2)
```

In Zelle D3 habe ich hingegen die folgende Formel eingegeben, wie Sie in der Bearbeitungsleiste sehen können:

```
=D2+1/FAKULTÄT(C3)
```

Anschließend habe ich die restlichen Zellen bis zu Zelle D17 automatisch aufgefüllt. Der Eintrag in Zelle D17 liegt sehr nah bei *e*. Ab Zelle D11 ist keine Änderung mehr erkennbar, selbst wenn Sie die Anzahl der anzuzeigenden Dezimalstellen erhöhen.



D3 | =D2+1/FAKULTÄT(C3)

	A	B	C	D	E
1			n	f(n)	
2			1	2	
3			2	2,5	
4			3	2,666667	
5			4	2,708333	
6			5	2,716667	
7			6	2,718056	
8			7	2,718254	
9			8	2,718279	
10			9	2,718282	
11			10	2,718282	
12			25	2,718282	
13			50	2,718282	
14			100	2,718282	
15			150	2,718282	
16			160	2,718282	
17			170	2,718282	
18					

Abbildung 22.11: Ein anderer Weg zu e

Warum habe ich mit der Zahl 170 aufgehört? Weil dies das Maximum ist, das Excel leisten kann. Ab 171 wird eine Fehlermeldung angezeigt.

e hängt also mit Wachstum zusammen. Der Wert der Konstanten ist 2,781828… Die drei Punkte bedeuten, dass man den genauen Wert niemals ganz erreicht (wie im Fall von π, der Konstanten, mit der Sie die Fläche eines Kreises bestimmen können).

Diese Zahl taucht an allen möglichen Stellen auf. Sie findet sich in der Formel für die Normalverteilung (siehe Kapitel 8) und auch in den verschiedenen Verteilungen, die ich in Kapitel 17 beschreibe. Viele Naturphänomene hängen mit *e* zusammen.

Sie ist so wichtig, dass Wissenschaftler, Mathematiker und Unternehmensanalysten sie als Basis für Logarithmen verwenden.

Logarithmen zur Basis *e* heißen *natürliche Logarithmen*. Die Abkürzung für einen natürlichen Logarithmus lautet *ln*.

Tabelle 22.2 enthält einen Vergleich (bis auf drei Dezimalstellen gerundet) zwischen Zehnerlogarithmen und natürlichen Logarithmen:

Zahl	Log	Ln
	0,434	1,000
10	1,000	2,303
50	1,699	3,912
100	2,000	4,605
453	2,656	6,116
1000	3,000	6,908

Tabelle 22.2: Zehnerlogarithmen (log) und natürliche Logarithmen (ln)

Eines noch. In vielen Formeln und Gleichungen muss man *e* in eine Potenz erheben. Manchmal ist die Potenz ein ziemlich komplizierter mathematischer Ausdruck. Da hochgestellte Potenzen häufig in einer recht kleinen Schriftgröße gedruckt werden, kann es sehr anstrengend sein, wenn man sie ständig lesen muss. Um die Augen zu entlasten, haben die Mathematiker eine besondere Schreibweise erfunden: *exp*. Immer, wenn Sie *exp* gefolgt von etwas in Klammern sehen, bedeutet das, dass Sie *e* mit dem, was in Klammern steht, potenzieren müssen. Beispielsweise gilt

$$exp\ (1{,}6) = e^{1{,}6} = 4{,}953$$

In Excel übernimmt die Funktion `EXP` die Berechnung für Sie.

Wo wir gerade beim Potenzieren und der Konstanten *e* sind: Als das Management von Google, Inc. anlässlich des geplanten Börsengangs des Unternehmens Informationen zur Neuemission veröffentlichte, gab es an, dass das Unternehmen 2.718.281.828 US-Dollar einnehmen wolle, was nichts anderes ist als *e* mal eine Milliarde US-Dollar bis auf den nächsten Dollar gerundet.

Wenden wir uns nun den Excel-Funktionen zu.

LOGNORM.VERT

Eine Zufallsvariable gilt als *lognormalverteilt*, wenn ihr natürlicher Logarithmus normalverteilt ist. Der Name ist vielleicht ein wenig missverständlich, da ich ja gerade erst erklärt habe, dass *log* für »Zehnerlogarithmus« und *ln* für »natürlicher Logarithmus« steht.

Anders als bei der Normalverteilung kann bei der Lognormalverteilung der mögliche Wert der Variablen keine negative Zahl sein. Ebenfalls anders als die Normalverteilung ist die Lognormalverteilung nicht symmetrisch, sondern rechtsschief.

Wie die Weibull-Verteilung, die weiter vorne beschrieben ist, wird auch die Lognormalverteilung von Ingenieuren verwendet, um das Versagen eines physikalischen Systems, besonders in Hinblick auf Verschleiß und Reißen, modellhaft darzustellen. Hier wird wieder die Eigenschaft der Logarithmen, große Zahlen durch kleine Zahlen darzustellen, ausgenutzt. Wenn der Lebenszyklus eines Systems über eine große Anzahl von Stunden betrachtet wird, ist es wesentlich einfacher, sich die Verteilung von Logarithmen statt einer Verteilung von Stunden vorzustellen.

In der Excel-Funktion `LOGNORM.VERT` kommt die Lognormalverteilung zum Einsatz. Sie legen einen Wert, einen Mittelwert und eine Standardabweichung für die Lognormalverteilung fest. `LOGNORM.VERT` gibt die Wahrscheinlichkeit zurück, dass die Variable höchstens diesen Wert hat.

Die FarKlempt Robotics GmbH hat zum Beispiel umfassende Daten darüber gesammelt, wie viele Stunden vergehen, bis ein Universalgelenkbauteil ausfällt, das in ihre Roboter eingebaut wird. Das Unternehmen stellt fest, dass die Anzahl der Stunden bis zum Ausfall mit einem Mittelwert von 10 und einer Standardabweichung von 2,5 lognormalverteilt ist. Wie hoch ist die Wahrscheinlichkeit, dass das Bauteil innerhalb von höchstens 10.000 Stunden ausfällt?

In Abbildung 22.12 ist das Dialogfeld FUNKTIONSARGUMENTE der Funktion `LOGNORM.VERT` für dieses Beispiel dargestellt. In das Feld X habe ich `ln(10000)` eingegeben. In das Feld MITTELWERT habe ich

10, in das Feld STANDARDABWN den Wert 2,5 und in das Feld KUMULIERT WAHR eingegeben. Im Dialogfeld wird das Ergebnis 0,000929 (mit ein paar weiteren Dezimalstellen) angezeigt. Wenn ich in das Feld KUMULIERT den Wert FALSCH eingebe, gibt die Funktion die Wahrscheinlichkeitsdichte (die Höhe der Funktion) für den Wert im Feld x zurück.

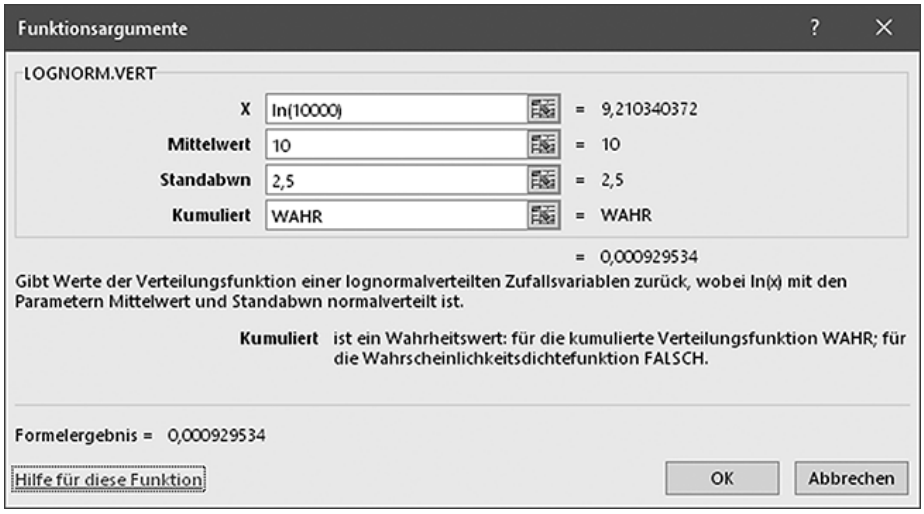


Abbildung 22.12: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion LOGNORM.VERT

LOGNORM.INV

LOGNORM.INV ist die Umkehrung von LOGNORM.VERT. Sie geben eine Wahrscheinlichkeit, einen Mittelwert und eine Standardabweichung für eine Lognormalverteilung ein. LOGNORM.INV liefert Ihnen dann den Wert der Zufallsvariablen zurück, der diese Wahrscheinlichkeit abtrennt.

Um den Wert zu bestimmen, der in der Verteilung des vorstehenden Beispiels 0,001 abtrennt, verwende ich das Dialogfeld FUNKTIONSARGUMENTE der Funktion LOGNORM.INV (siehe Abbildung 22.13). Nachdem ich die Werte für die Argumente eingegeben habe, wird im Dialogfeld das Ergebnis 9,722 (mit einigen weitere Dezimalstellen) angezeigt.

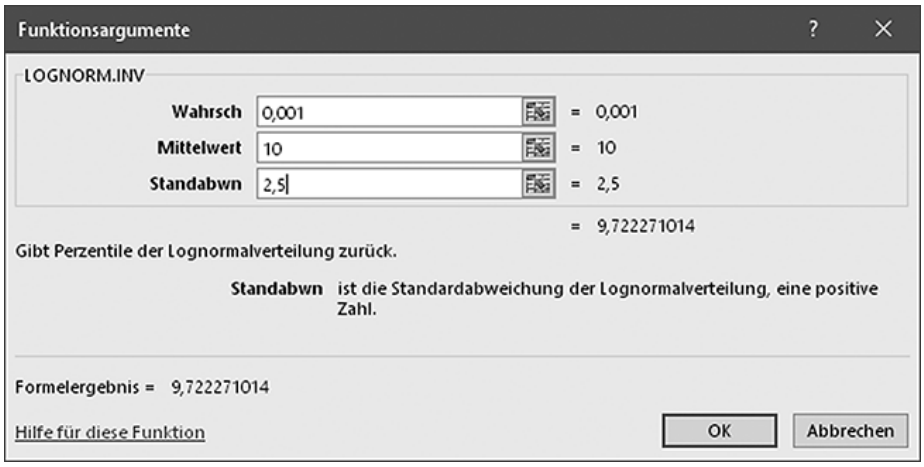


Abbildung 22.13: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion `LOGNORM.INV`

In Stunden ausgedrückt sind das übrigens 16.685 Stunden – und das nur für 0,001.

Arrayfunktion: RKP

In Kapitel 14 gebe ich Ihnen umfassende Informationen zur linearen Regression. Es ist aber auch möglich, dass zwischen zwei Variablen eine Beziehung besteht, die krummlinig und nicht linear ist.

Die Gleichung für eine Gerade, die zu einem Streudiagramm passt, lautet

$$y' = a + bx$$

Eine Möglichkeit, eine Kurve durch ein Streudiagramm zu legen, ist die folgende Gleichung:

$$y' = aebx$$

Die Funktion `RKP` schätzt *a* und *b* für diese krummlinige Gleichung. In Abbildung 22.14 sind das Dialogfeld FUNKTIONSARGUMENTE der Funktion `RKP` und die Daten für dieses Beispiel dargestellt. Außerdem wird die Matrix für die Ergebnisse angezeigt. Vor der Verwendung dieser Funktion habe ich dem Zellbereich B2:B12 den Namen `x` und dem Zellbereich C2:C12 den Namen `y` zugewiesen.

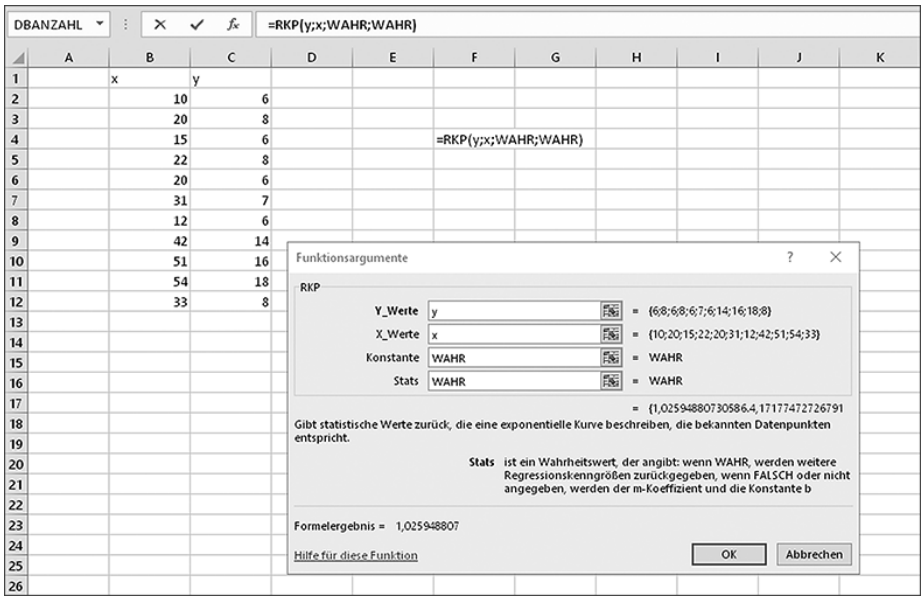


Abbildung 22.14: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion `RKP` mit Daten und der für die Ergebnisse ausgewählten Matrix

Hier die einzelnen Schritte für die Verwendung dieser Funktion:

1. **Wenn Sie die Daten eingegeben haben, markieren Sie eine Zellenmatrix, die fünf Zeilen und zwei Spalten umfasst und das Ergebnis der Funktion `RKP` aufnehmen soll.**

   Ich habe den Bereich F4:G8 markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion RKP aus, um das Dialogfeld FUNKTIONSARGUMENTE der Funktion `RKP` zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Geben Sie in das Feld Y_WERTE den Zellbereich ein, der die Werte für die *y*-Variable enthält. In diesem Beispiel ist dies `y`, der Name, den ich dem Zellbereich C2:C12 zugewiesen habe.

   Geben Sie in das Eingabefeld X_WERTE den Zellbereich ein, der die Werte für die *x*-Variable enthält. In diesem Beispiel ist dies `x`, der Name, den ich dem Zellbereich B2:B12 zugewiesen habe

Geben Sie in das Eingabefeld KONSTANTE den Wert `WAHR` ein (oder lassen Sie das Feld leer), um den Wert von *a* in der krummlinigen Gleichung, die ich Ihnen gezeigt habe, zu berechnen. Wenn Sie `FALSCH` eingeben, wird *a* auf 1 gesetzt. Ich habe `WAHR` eingegeben.

Im Dialogfeld wird statt *a* die Angabe *b* verwendet. Symbole werden nun einmal leider nicht standardisiert verwendet.

Geben Sie in das Eingabefeld STATS den Wert `WAHR` ein, um zusätzlich zu *a* und *b* auch die Regressionsstatistik zurückzugeben. Wenn Sie `FALSCH` eingeben (oder das Feld leer lassen), werden nur *a* und *b* zurückgegeben.

Im Dialogfeld werden die Ausdrücke *b* (statt *a* wie bei mir) und *m-Koeffizient* verwendet.

4. **WICHTIG: Klicken Sie NICHT auf OK. Da es sich hier um eine Matrixfunktion handelt, müssen Sie Strg+Shift+Enter drücken, um das Ergebnis der Funktion `RKP` in die markierte Matrix zu übertragen.**

In Abbildung 22.15 sind die Ergebnisse der Funktion `RKP` dargestellt. Da sie unbeschriftet angezeigt werden, habe ich die Beschriftungen für Sie in das Arbeitsblatt eingefügt. In der linken Spalte steht *exp*(*b*) (Näheres dazu gleich), der Standardfehler von *b*, R Quadrat, F und $SS_{Regression}$. In der rechten Spalte steht *a,* der Standardfehler von *a,* der Standardfehler des Schätzwertes, die Freiheitsgrade und $SS_{Residue}$. Weitere Informationen zu diesen statistischen Werten finden Sie in Kapitel 14 und 15.

	A	B	C	D	E	F	G	H
1		x	y					
2		10	6					
3		20	8					
4		15	6		exp	1,02594881	4,17177473	Abschnitt
5		22	8		Standardfehler der Steigung	0,00310866	0,0986791	Standardfehler des Abschnitts
6		20	6		R Quadrat	0,88298132	0,15062209	Standardfehler des Schätzwerts
7		31	7		F	67,9107979	9	df
8		12	6		SSRegression	1,54069328	0,20418313	SSResidue
9		42	14					
10		51	16					
11		54	18					
12		33	8					
13								
14								

Abbildung 22.15: Die Ergebnisse der Funktion `RKP` in der ausgewählten Matrix

Zu *exp*(*b*). Leider gibt die Funktion `RKP` nicht den Wert von *b*, also den Exponenten der krummlinigen Gleichung, zurück. Um diesen Exponenten zu bestimmen, müssen Sie den natürlichen Logarithmus dessen berechnen, was die Funktion zurückgibt. Die Tabellenblattfunktion `LN` von Excel gibt in diesem Fall die Zahl 0,0256 als Wert des Exponenten zurück.

Die krummlinige Regressionsgleichung für die Stichprobendaten lautet also

$$y' = 4{,}1718e^{0{,}0256x}$$

oder in der *exp*-Schreibweise, von der ich Ihnen berichtet habe

$$y' = 4{,}1718\ exp\ (0{,}0256x)$$

Eine gute Möglichkeit, all das besser zu verstehen, ist die Verwendung der Diagrammfeatures von Excel, mit denen Sie ein Streudiagramm erstellen können (siehe Kapitel 3). Klicken Sie anschließend mit der rechten Maustaste auf einen Datenpunkt im Diagramm und wählen Sie im

Kontextmenü den Befehl TRENDLINIE HINZUFÜGEN aus. Hierdurch wird eine Trendlinie in das Diagramm eingefügt und gleichzeitig – was noch wichtiger ist – der Aufgabenbereich TRENDLINIE FORMATIEREN geöffnet (siehe Abbildung 22.16). Schalten Sie das Optionsfeld neben EXPONENTIAL ein, wie ich es in der Abbildung auch getan habe. Blättern Sie dann an das Ende des Aufgabenbereichs und schalten Sie das Kontrollkästchen FORMEL IM DIAGRAMM ANZEIGEN ein.

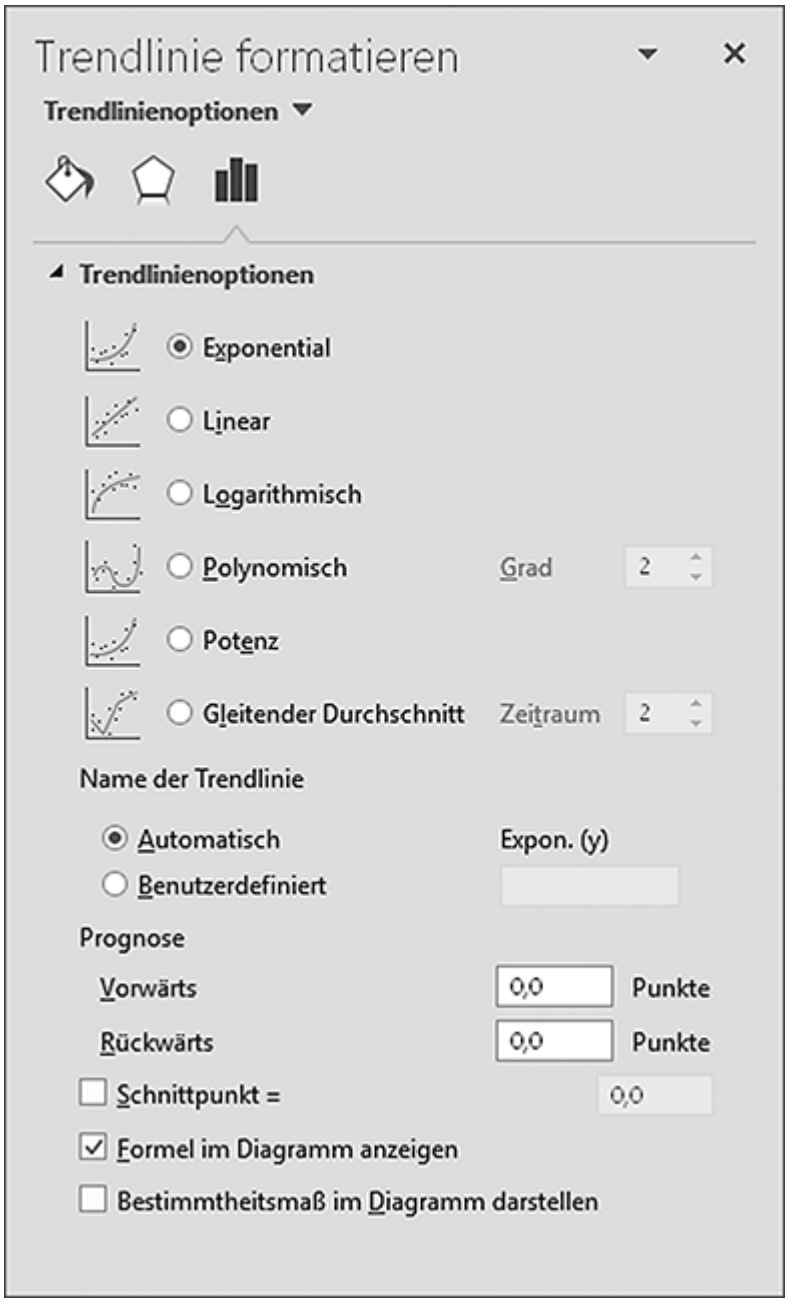


Abbildung 22.16: Der Aufgabenbereich TRENDLINIE FORMATIEREN

Klicken Sie auf SCHLIESSEN und schon haben Sie ein Streudiagramm komplett mit Kurve und Formel. Ich habe meins mehrfach umformatiert, um es gedruckt besser darstellen zu können. Das Ergebnis ist in Abbildung 22.17 dargestellt.

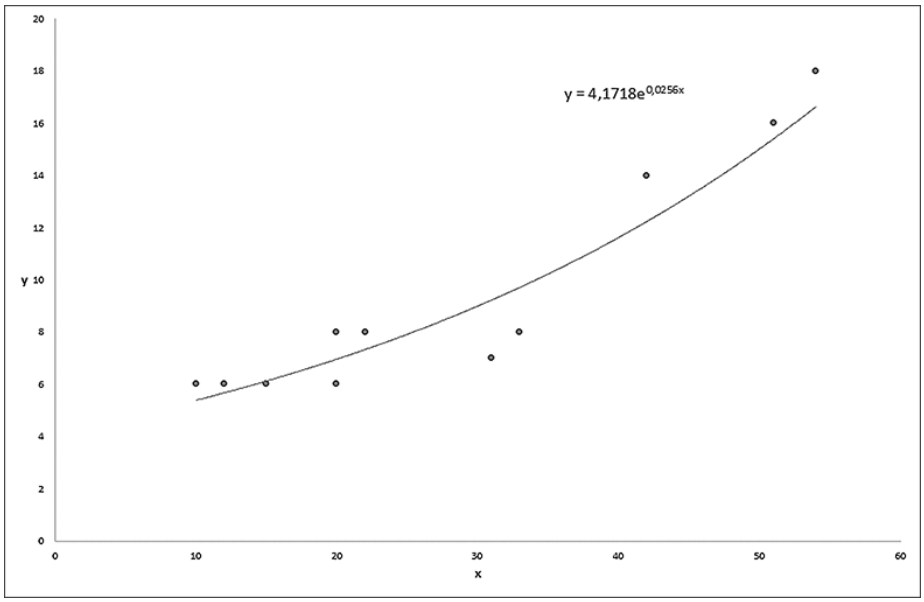


Abbildung 22.17: Das Streudiagramm mit Kurve und Formel

Arrayfunktion: VARIATION

`VARIATION` ist die Antwort einer krummlinigen Regression auf `TREND` (siehe Kapitel 14). Sie können diese Funktion auf zwei Arten verwenden – um eine Reihe von *y*-Werten für die *x*-Werte in Ihrer Stichprobe oder eine Reihe von *y*-Werten für eine neue Reihe von *x*-Werten vorherzusagen.

y-Werte für die *x*-Werte in Ihrer Stichprobe vorhersagen

In Abbildung 22.18 ist die Funktion `VARIATION` für die Berechnung der *y*-Werte für die *x*-Werte dargestellt, die ich bereits habe. Ich habe die Bearbeitungsleiste in die Abbildung mit aufgenommen, damit Sie sehen können, wie die Formel bei dieser Verwendung der Funktion `VARIATION` aussieht.

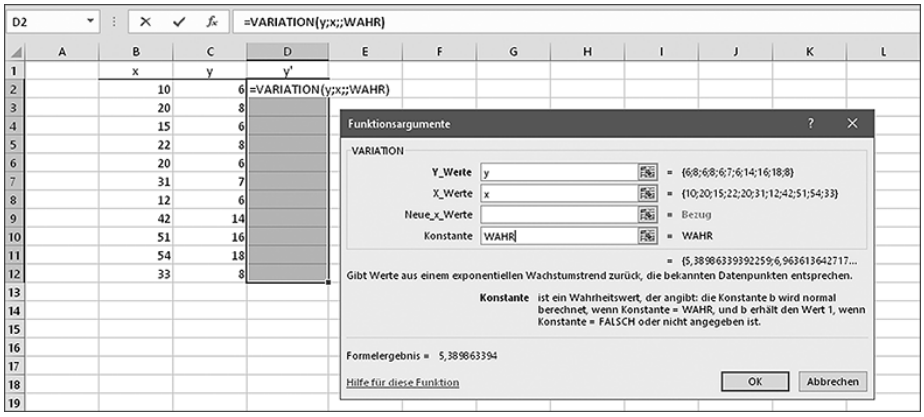


Abbildung 22.18: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion VARIATION mit den Stichprobendaten. VARIATION wird hier verwendet, um die y-Werte für die x-Werte der Stichprobe vorherzusagen.

Hier die einzelnen Schritte:

1. **Nachdem Sie die Daten eingegeben haben, markieren Sie einen Zellbereich für die Ergebnisse der Funktion VARIATION.**

   Ich habe D2:D12 markiert, um die vorhergesagten y-Werte direkt neben den y-Werten der Stichprobe einzublenden.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion VARIATION aus, um das Dialogfeld FUNKTIONSARGUMENTE der Funktion VARIATION zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Geben Sie in das Feld Y_WERTE den Zellbereich ein, der die Werte für die *y*-Variable enthält. In diesem Beispiel habe ich `y` eingegeben, dies ist der für den Zellbereich C2:C12 definierte Name.

   Geben Sie in das Feld X_WERTE den Zellbereich ein, der die Werte für die x-Variable enthält. In diesem Beispiel habe ich `x` eingegeben, dies ist der für den Zellbereich B2:B12 definierte Name.

Da ich hier keine neuen x-Werte berechne, lasse ich das Feld NEUE_X_WERTE leer.

Geben Sie in das Eingabefeld KONSTANTE den Wert `WAHR` ein (oder lassen Sie das Feld leer), um *a* zu berechnen. Wenn Sie `FALSCH` eingeben, wird *a* auf 1 gesetzt. Ich habe `WAHR` eingegeben. (Ich weiß wirklich nicht, warum man `FALSCH` eingeben sollte.) Im Dialogfeld wird wieder statt *a* die Angabe *b* verwendet.

4. **WICHTIG: Klicken Sie NICHT auf OK. Da es sich hier um eine Matrixfunktion handelt, müssen Sie Strg+Shift+Enter drücken, um das Ergebnis der Funktion `VARIATION` in die markierte Matrix zu übertragen.**

   In Abbildung 22.19 sind die Ergebnisse in den Zellen D2:D12 dargestellt.

D2 | {=VARIATION(y;x;;WAHR)}

	A	B	C	D	E
1		x	y	y'	
2		10	6	5,389863394	
3		20	8	6,963613643	
4		15	6	6,126412185	
5		22	8	7,329697464	
6		20	6	6,963613643	
7		31	7	9,230330843	
8		12	6	5,673213661	
9		42	14	12,23488433	
10		51	16	15,40746133	
11		54	18	16,63826968	
12		33	8	9,715578152	
13					

Abbildung 22.19: Die Ergebnisse der Funktion `VARIATION`: vorhergesagte y-Werte für die x-Werte der Stichprobe

Eine neue Reihe von y-Werten für eine neue Reihe von x-Werten vorhersagen

In diesem Fall verwende ich die Funktion VARIATION, um *y*-Werte für eine neue Reihe von *x*-Werten vorherzusagen. In Abbildung 22.20 ist auch die für die Ergebnisse ausgewählte Matrix dargestellt. Auch hier füge ich die Bearbeitungsleiste mit in die Abbildung ein, um Ihnen die Anwendung der Funktion zu verdeutlichen.

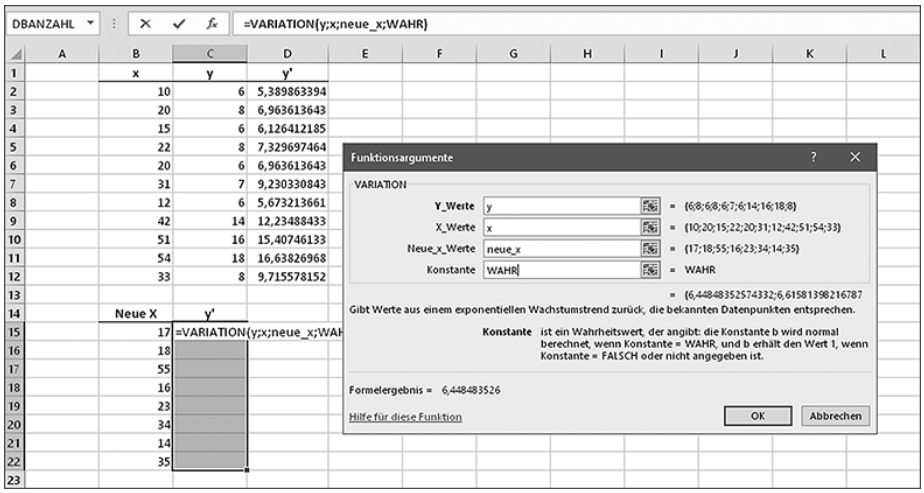


Abbildung 22.20: Das Dialogfeld FUNKTIONSARGUMENTE für die Funktion VARIATION mit den zugehörigen Daten. VARIATION wird hier verwendet, um y-Werte für eine neue Reihe von x-Werten vorherzusagen.

Gehen Sie folgendermaßen vor, um diese Funktion so zu verwenden:

1. **Nachdem Sie die Daten eingegeben haben, markieren Sie einen Zellbereich für die Ergebnisse der Funktion VARIATION.**

   Ich habe C15:C22 markiert.

2. **Wählen Sie im Menü mit den statistischen Funktionen die Funktion VARIATION aus, um das Dialogfeld FUNKTIONSARGUMENTE der Funktion VARIATION zu öffnen.**

3. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   Geben Sie in das Feld Y_WERTE den Zellbereich ein, der die Werte für die *y*-Variable enthält. In diesem Beispiel habe ich `y` eingegeben, dies ist der für den Zellbereich C2:C12 definierte Name.

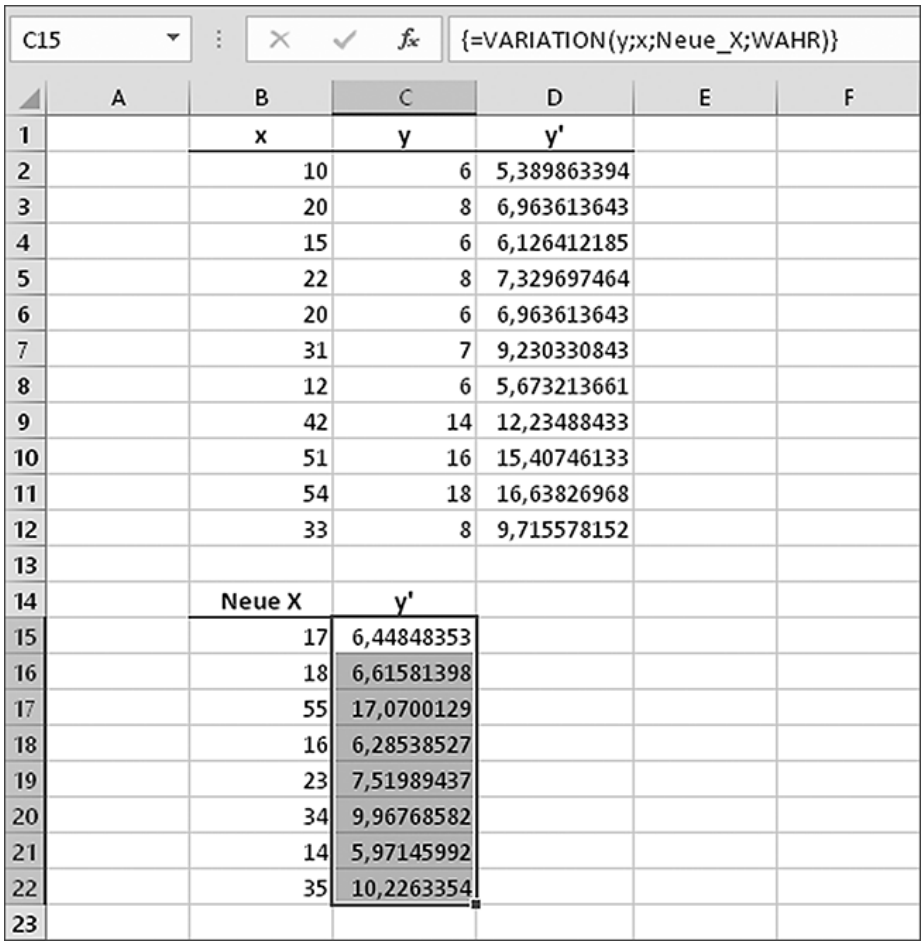
   Geben Sie in das Feld X_WERTE den Zellbereich ein, der die Werte für die x-Variable enthält. In diesem Beispiel habe ich `x` eingegeben, dies ist der für den Zellbereich B2:B12 definierte Name.

   Geben Sie in das Feld NEUE_X_WERTE den Zellbereich ein, der die neuen Werte für die x-Variable enthält. In diesem Beispiel habe ich `neue_x` eingegeben, dies ist der für den Zellbereich B15:B22 definierte Name.

   Geben Sie in das Eingabefeld KONSTANTE den Wert `WAHR` ein (oder lassen Sie das Feld leer), um *a* zu berechnen. Wenn Sie `FALSCH` eingeben, wird *a* auf 1 gesetzt. Ich habe `WAHR` eingegeben. (Um es noch einmal zu wiederholen: Ich weiß wirklich nicht, warum man `FALSCH` eingeben sollte.)

4. **WICHTIG: Klicken Sie NICHT auf OK. Da es sich hier um eine Matrixfunktion handelt, müssen Sie Strg+Shift+Enter drücken, um das Ergebnis der Funktion `VARIATION` in die markierte Matrix zu übertragen.**

In Abbildung 22.21 sind die Ergebnisse im Zellbereich C15:C22 dargestellt.



C15 | {=VARIATION(y;x;Neue_X;WAHR)}

	A	B	C	D	E	F
1		x	y	y'		
2		10	6	5,389863394		
3		20	8	6,963613643		
4		15	6	6,126412185		
5		22	8	7,329697464		
6		20	6	6,963613643		
7		31	7	9,230330843		
8		12	6	5,673213661		
9		42	14	12,23488433		
10		51	16	15,40746133		
11		54	18	16,63826968		
12		33	8	9,715578152		
13						
14		Neue X	y'			
15		17	6,44848353			
16		18	6,61581398			
17		55	17,0700129			
18		16	6,28538527			
19		23	7,51989437			
20		34	9,96768582			
21		14	5,97145992			
22		35	10,2263354			
23						

Abbildung 22.21: Die Ergebnisse der Funktion `VARIATION`: vorhergesagte y-Werte für eine neue Reihe von x-Werten

Die Logs von Gamma

Hört sich doch an wie ein Science-Fiction-Thriller, oder?

Die Gammafunktion, die ich in Kapitel 19 beschreibe, erweitert die Fakultäten auf den Bereich der Zahlen, die keine Ganzzahlen sind. Da es hier um Fakultäten geht, werden die Zahlen sehr schnell sehr groß. Logarithmen sind ein geeignetes Gegenmittel.

In einer früheren Version stellte Excel bereits die Funktion `GAMMALN` zur Verfügung, mit der der natürliche Logarithmus der Gammafunktion für das Argument x berechnet werden kann. (Diese Funktion gab es bereits, bevor `GAMMA` zur Verfügung stand.)

In Excel 2013 wurde `GAMMALN` einem Facelift unterzogen und die neue Funktion liefert (vermutlich) genauere Ergebnisse. Daher heißt

die neue Tabellenblattfunktion `GAMMALN.GENAU`.

Die neue Funktion wird folgendermaßen verwendet:

```
=GAMMALN.GENAU(5,3)
```

Diese Formel entspricht:

```
=LN(GAMMA(5,3))
```

Das Ergebnis lautet übrigens 3,64.

Nur damit Sie es wissen: Ich habe für das Ergebnis 14 Dezimalstellen anzeigen lassen und konnte bei diesem Beispiel keinen Unterschied zwischen den Ergebnissen von `GAMMALN` und `GAMMALN.GENAU` feststellen.

Daten sortieren

In verhaltenswissenschaftlichen Experimenten präsentieren die Forscher den Teilnehmern mehrere Aufgaben, die diese abarbeiten sollen. Die Bedingungen dieser Aufgaben sind die unabhängigen Variablen und die bei der Durchführung der Aufgaben erhobenen Messergebnisse stellen die abhängigen Variablen dar.

Aus methodologischen Gründen werden die Bedingungen und die Reihenfolge der Aufgaben zufällig angeordnet, damit die Probanden die Aufgaben in unterschiedlicher Reihenfolge durchführen. Die Daten spiegeln diese Reihenfolge wider. Um die Daten analysieren zu können, ist es erforderlich, die Daten aller Versuchsteilnehmer in die gleiche Reihenfolge zu bringen.

Das Arbeitsblatt in Abbildung 22.22 enthält die Daten eines Teilnehmers in einem Experiment. *Breite* und *Abstand* sind die unabhängigen Variablen; *Schritte* und *Fehler* sind die abhängigen Variablen. Das Ziel besteht darin, die Daten zuerst in aufsteigender Reihenfolge nach der Breite und dann in aufsteigender Reihenfolge nach dem Abstand zu sortieren.

	A	B	C	D	E	F
1		Breite	Abstand	Schritte	Fehler	
2		60	500	18	0	
3		30	1000	13	0	
4		240	500	23	1	
5		60	1000	15	1	
6		120	500	20	1	
7		240	1000	18	1	
8		30	500	12	2	
9		120	1000	16	1	
10						

Abbildung 22.22: Die unsortierten Daten

Dies erledigen Sie so:

1. **Markieren Sie den Zellbereich, in dem sich die Daten befinden.**

   In diesem Beispiel habe ich B2:E9 markiert.

2. **Wählen Sie DATEN | SORTIEREN.**

   Hierdurch wird das Dialogfeld SORTIEREN geöffnet, das Abbildung 22.23 zeigt. Nach dem Öffnen des Dialogfeldes wird unterhalb von SPALTE lediglich eine Zeile angezeigt. Diese Zeile ist mit SORTIEREN NACH beschriftet. Da meine Daten Spaltenüberschriften besitzen, habe ich das Kontrollkästchen DATEN HABEN ÜBERSCHRIFTEN eingeschaltet.

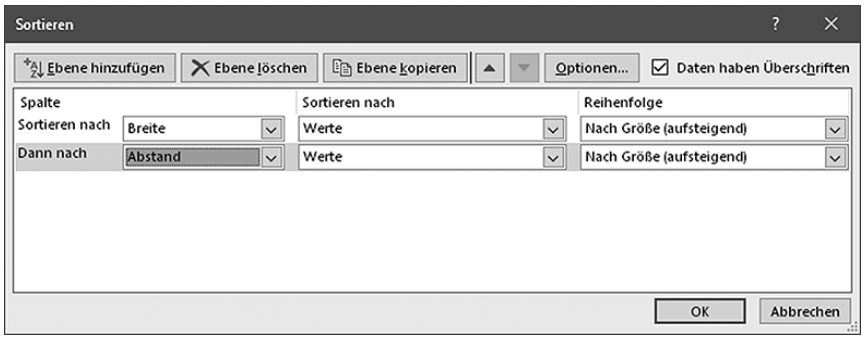


Abbildung 22.23: Das Dialogfeld »Sortieren«

3. **Wählen Sie in der Dropdownliste neben SORTIEREN NACH das erste Feld aus, nach dem sortiert werden soll. Passen Sie die Optionen SORTIEREN NACH und REIHENFOLGE an.**

   Ich hab BREITE ausgewählt und die Standardoptionen für SORTIEREN NACH (Werte) und REIHENFOLGE (Nach Größe aufsteigend) übernommen.

4. **Klicken Sie die Schaltfläche EBENE HINZUFÜGEN an.**

   Hierdurch wird eine weitere Zeile eingefügt; diese ist mit DANN NACH beschriftet.

5. **Wählen Sie in der Dropdownliste neben DANN NACH das nächste Feld aus, nach dem sortiert werden soll. Passen Sie die Optionen SORTIEREN NACH und REIHENFOLGE an.**

   Ich hab ABSTAND ausgewählt und die Standardoptionen übernommen.

6. **Nachdem Sie das letzte Sortierfeld festgelegt haben, klicken Sie auf OK.**

   Die sortierten Daten zeigt Abbildung 22.24.

	A	B	C	D	E	F
1		Breite	Abstand	Schritte	Fehler	
2		30	500	12	2	
3		30	1000	13	0	
4		60	500	18	0	
5		60	1000	15	1	
6		120	500	20	1	
7		120	1000	16	1	
8		240	500	23	1	
9		240	1000	18	1	
10						

Abbildung 22.24: Die Daten wurden nach den Feldern Breite und Abstand sortiert.

Anahng A

Wenn Ihr Arbeitsblatt eine Datenbank ist

In diesem Kapitel

- Datenbanken in Excel
- Statistiken in Datenbanken
- PivotTable-Berichte

Die Hauptfunktion von Excel im täglichen Leben ist die Durchführung von Berechnungen. Wie die Kapitel in diesem Buch zeigen, drehen sich viele dieser Berechnungen um eingebaute Statistikfunktionen.

Darüber hinaus können Sie ein Arbeitsblatt erstellen, um Daten wie in einer Datenbank zu speichern, obwohl Excel dabei natürlich einen eingeschränkteren Funktionsumfang als speziell dafür entwickelte Datenbankpakete hat. Excel bietet Ihnen Datenbankfunktionen, die seinen Statistikfunktionen ähneln, weshalb ich Sie Ihnen gern vorstellen würde.

Einführung in Excel-Datenbanken

Wenn man es ganz genau nimmt, liefert Excel als Ergebnis eine *Datenliste*. Dabei handelt es sich um eine Reihe von Arbeitsblattzellen, in die Sie in einem einheitlichen Format Daten eingeben, die miteinander in Beziehung stehen. Sie ordnen die Daten in Spalten ein und benennen die einzelnen Spalten. In der

Datenbankterminologie ist eine benannte Spalte ein *Feld*. Eine Zeile ist ein eigener *Datensatz*.

Diese Strukturierung ist nützlich, wenn man Inventarlisten führt, aber auch nur dann, wenn diese nicht zu umfangreich werden. Sie würden eine Excel-Datenbank also beispielsweise nicht für die Warenbestandsführung eines Lagers einer großen Firma verwenden. Bei kleineren Unternehmen kann Excel aber durchaus das geeignete Programm sein.

Die Trabanten-Datenbank

In Abbildung A.1 sehen Sie ein Beispiel. Es handelt sich dabei um eine Liste der klassischen Trabanten in unserem Sonnensystem. Mit »klassisch« meine ich, dass die meisten von ihnen vor dem zwanzigsten Jahrhundert mit herkömmlichen Teleskopen entdeckt wurden. Das Licht der drei Trabanten in der Liste, die im zwanzigsten Jahrhundert entdeckt wurden, war so schwach, dass die Astronomen sie erst entdeckten, als sie Fotoplatten untersuchten. Heutige Superteleskope und Raumsonden haben noch wesentlich mehr Trabanten aufgespürt, die ich in der Liste aber nicht aufführe.

	A	B	C	D	E	F	G
1		Name	Planet	Umlaufzeit (Tage)	Durchschnittliche Entfernung (x1000 km)	Entdeckungsjahr	Entdecker
2		*io	Saturn	1,26	>150	>1877	Galileo
3				>20			Cassini
4							
5							
6							
7							
8							
9							
10		Name	Planet	Umlaufzeit (Tage)	Durchschnittliche Entfernung (x1000 km)	Entdeckungsjahr	Entdecker
11		Amalthea	Jupiter	0,50	181,30	1892	Barnard
12		Ariel	Uranus	2,52	191,24	1851	Lassell
13		Callisto	Jupiter	16,69	1883,00	1610	Galileo
14		Charon	Pluto	6,39	19,64	1978	Christy
15		Deimos	Mars	1,26	23,46	1877	Hall
16		Dione	Saturn	2,74	377,40	1684	Cassini
17		Enceladus	Saturn	1,37	238,02	1789	Herschel
18		Europa	Jupiter	3,55	670,90	1610	Galileo
19		Ganymed	Jupiter	7,15	1070,00	1610	Galileo
20		Hyperion	Saturn	21,28	1481,00	1848	Bond
21		Iapetus	Saturn	79,33	3561,30	1671	Cassini
22		Io	Jupiter	1,77	421,60	1610	Galileo
23		Mimas	Saturn	9,42	185,52	1789	Herschel
24		Miranda	Uranus	1,41	129,78	1948	Kuiper
25		Mond	Erde	27,32	384,40	entf.	entf.
26		Nereid	Neptun	360,14	5513,40	1949	Kuiper
27		Oberon	Uranus	13,46	582,50	1787	Herschel
28		Phobos	Mars	0,32	9,38	1877	Hall
29		Phoboe	Saturn	-550,48	12952,00	1898	Pickering
30		Rhea	Saturn	4,52	527,04	1672	Cassini
31		Tethys	Saturn	1,89	294,66	1684	Cassini
32		Titan	Saturn	15,94	1221,85	1655	Huygens
33		Titania	Uranus	8,71	435,84	1787	Herschel
34		Triton	Neptun	-5,88	354,80	1846	Lassell
35		Umbriel	Uranus	4,14	265,97	1851	Lassell

Abbildung A.1: Die Trabanten-Datenbank

Die Datenbank befindet sich in den Zellen B10:G35. Diesem Zellbereich habe ich den Namen `Trabanten` zugewiesen. Beachten Sie, dass die Feldnamen (Spaltenüberschriften) Bestandteil dieses Zellbereichs sind.

Im Feld NAME stehen die Namen der Trabanten, im Feld PLANET stehen die Planeten, um die sich die Trabanten drehen.

Im Feld UMLAUFZEIT (TAGE) ist angegeben, wie lange die einzelnen Trabanten brauchen, um ihren jeweiligen Planeten einmal komplett zu umrunden. Unser Mond zum Beispiel braucht dazu etwas mehr als 27 Tage. Einige Datensätze haben an dieser Stelle negative Werte. Das bedeutet, dass die Richtung, in der sie sich um ihren Planeten drehen, entgegengesetzt zu der Drehrichtung des Planeten ist.

DURCHSCHNITTLICHE ENTFERNUNG (×1000 KM) ist die durchschnittliche Entfernung zwischen Planet und Trabant in tausend Kilometern. In den letzten beiden Feldern stehen das Jahr der Entdeckung und der

Name des Astronomen, der den Trabanten entdeckt hat. Bei unserem Mond sind diese beiden Daten natürlich unbekannt.

Nachdem Sie die einzelnen Felder beschriftet haben, weisen Sie der Zelle, in der der Feldname steht, einen definierten Namen zu. Wichtig ist, dass der Zellbereich, für den Sie den Namen definieren, lediglich aus der einen Zelle besteht, in der sich der Feldname befindet. Definieren Sie den Namen **nicht** für die gesamte Spalte mit den Daten. Zelle B10 beispielsweise habe ich den definierten Namen `Name` zugewiesen.

Okay, ich gebe es zu: Mit all den Namen im vorigen Satz ist dies vielleicht nicht leicht nachzuvollziehen. Daher hier zwei weitere Beispiele, die einfacher sind. Zelle C10 habe ich den Namen `Planet` zugewiesen, Zelle D10 den Namen `Umlaufzeit_Tage` und so weiter. Nun kann ich diese Namen in den Datenbankfunktionen von Excel verwenden.

Der Kriterienbereich

Ich habe die Spaltenüberschriften – Entschuldigung, Feldnamen – in die obere Zeile kopiert. Darüber hinaus habe ich einige Daten in nahe gelegene Zellen eingetragen. Dies ist der *Kriterienbereich*. Sie können diesen Kriterienbereich verwenden, um mit den Excel-Datenbankfunktionen Fragen über Ihre Daten zu stellen; die Funktionen liefern dann die Antworten. Datenbankfreaks kennen diesen Vorgang als *Abfrage*. Die Kriterien sind ein unverzichtbarer Bestandteil jeder Datenbankfunktion. (Kriterien ist die Plural- oder Mehrzahlform. In der Einzahl bzw. im Singular heißt es Kriterium.)

Dieser Bereich muss sich nicht zwingend oben auf dem Arbeitsblatt befinden. Sie können jeden beliebigen Bereich im Arbeitsblatt als Kriterienbereich verwenden.

Wenn Sie eine Excel-Datenbankfunktion anwenden, dann hat sie dieses Format:

```
=FUNKTION(Datenbank;Feld;Kriterium)
```

Die Funktion nutzt Daten der angegebenen Datenbank, wird im angegebenen Feld aktiv und befolgt dabei die angegebenen Kriterien.

Wenn Sie beispielsweise wissen möchten, wie viele Trabanten zum Saturn gehören, markieren Sie eine Zelle und geben die folgende Formel ein:

```
=DBANZAHL(Trabanten;
Durchschnittliche_Entfernung_x1000_km;C1:C2)
```

Und das bedeutet die Formel: In der Datenbank (B10:G35) zählt `DBANZAHL` die Anzahl der Zellen im Feld DURCHSCHNITTLICHE ENTFERNUNG (×1000 KM), die Zahlen enthalten, eingeschränkt durch das im Zellenbereich C1:C2 festgelegte Kriterium. Das Kriterium entspricht Planet = Saturn. Beachten Sie bitte, dass das Kriterium mindestens eine Spaltenüberschrift … äh … einen Feldnamen aus dem Kriterienbereich und mindestens eine Zeile umfassen muss. Bitte beachten Sie, dass Sie den Feldnamen selbst im Kriterium nicht verwenden dürfen; stattdessen verwenden Sie die Zelladresse (beispielsweise C1).

Wenn Sie mehr als eine Zeile in die Formel aufnehmen, bedeutet das »oder«. Wenn Ihr Kriterium z.B. G1:G3 ist, legen Sie Trabanten fest, die von Galileo oder von Cassini entdeckt wurden.

Wenn Sie mehr als eine Spalte in ein Kriterium einschließen, bedeutet das »und«. Wenn Ihr Kriterium beispielsweise E1:F2 ist, legen Sie Trabanten fest, die mehr als 150.000 km von ihren jeweiligen Planeten entfernt sind und die erst nach 1877 entdeckt wurden.

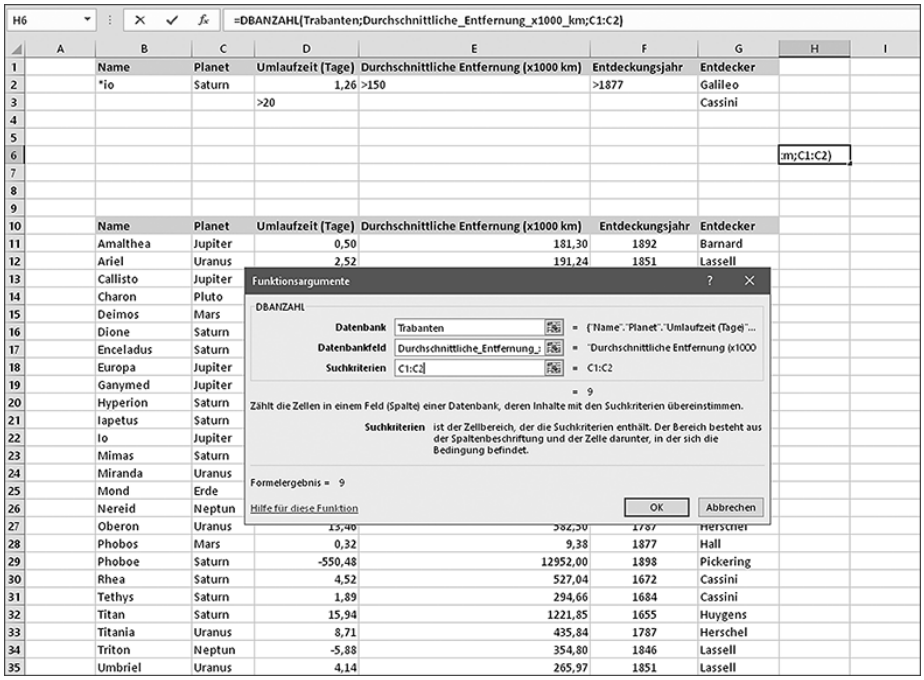
Das Format einer Datenbankfunktion

Die Formel

```
=DBANZAHL(Trabanten;Durchschnittliche_Entfernung_x
1000_km;C1:C2)
```

können Sie wie alle anderen Formeln auch im Dialogfeld FUNKTIONSARGUMENTE eingeben. In Abbildung A.2 ist das entsprechende Dialogfeld der vorstehenden Formel vor dem Hintergrund der Datenbank und des Kriterienbereichs dargestellt.

Wie öffnen Sie dieses Dialogfeld? Anders als bei den statistischen Funktionen und den Funktionen der Kategorie MATHEMATIK UND TRIGONOMETRIE befindet sich auf der Registerkarte FORMELN keine eigene Schaltfläche für die Datenbankfunktionen. Stattdessen klicken Sie die Schaltfläche FUNKTION EINFÜGEN an (Sie finden diese sowohl in der Bearbeitungsleiste aus auch auf der Registerkarte FORMELN in der Gruppe FUNKTIONSBIBLIOTHEK), um das Dialogfeld FUNKTION EINFÜGEN zu öffnen. Wählen Sie dann im Listenfeld KATEGORIE AUSWÄHLEN den Eintrag DATENBANK aus und klicken Sie abschließend in der Liste mit den Funktionen den Namen der Datenbankfunktion an, die Sie verwenden wollen.



	A	B	C	D	E	F	G	H	I
1		Name	Planet	Umlaufzeit (Tage)	Durchschnittliche Entfernung (x1000 km)	Entdeckungsjahr	Entdecker		
2		*io	Saturn	1,26	>150	>1877	Galileo		
3				>20			Cassini		
4									
5									
6								:m;C1:C2)	
7									
8									
9									
10		Name	Planet	Umlaufzeit (Tage)	Durchschnittliche Entfernung (x1000 km)	Entdeckungsjahr	Entdecker		
11		Amalthea	Jupiter	0,50	181,30	1892	Barnard		
12		Ariel	Uranus	2,52	191,24	1851	Lassell		
13		Callisto	Jupiter						
14		Charon	Pluto						
15		Deimos	Mars						
16		Dione	Saturn						
17		Enceladus	Saturn						
18		Europa	Jupiter						
19		Ganymed	Jupiter						
20		Hyperion	Saturn						
21		Iapetus	Saturn						
22		Io	Jupiter						
23		Mimas	Saturn						
24		Miranda	Uranus						
25		Mond	Erde						
26		Nereid	Neptun						
27		Oberon	Uranus	13,46	582,50	1787	Herschel		
28		Phobos	Mars	0,32	9,38	1877	Hall		
29		Phoboe	Saturn	-550,48	12952,00	1898	Pickering		
30		Rhea	Saturn	4,52	527,04	1672	Cassini		
31		Tethys	Saturn	1,89	294,66	1684	Cassini		
32		Titan	Saturn	15,94	1221,85	1655	Huygens		
33		Titania	Uranus	8,71	435,84	1787	Herschel		
34		Triton	Neptun	-5,88	354,80	1846	Lassell		
35		Umbriel	Uranus	4,14	265,97	1851	Lassell		

Abbildung A.2: Das Dialogfeld FUNKTIONSARGUMENTE der Funktion DBANZAHL

Hier ein Beispiel:

1. **Markieren Sie eine Zelle des Arbeitsblatts.**

Ich habe die Zelle H6 markiert.

2. **Klicken Sie auf die Schaltfläche FUNKTION EINFÜGEN (sie ist mit FX beschriftet), um das Dialogfeld FUNKTION EINFÜGEN zu öffnen.**

3. **Wählen Sie im Dialogfeld FUNKTION EINFÜGEN eine Funktion aus, um das entsprechende Dialogfeld FUNKTIONSARGUMENTE zu öffnen.**

   Aus der Kategorie DATENBANK habe ich die Funktion DBANZAHL ausgewählt. Das entsprechende Dialogfeld ist in Abbildung A.2 dargestellt.

4. **Geben Sie im Dialogfeld FUNKTIONSARGUMENTE die Werte für die Argumente ein.**

   In das Eingabefeld DATENBANK habe ich den Namen des Zellbereichs, `Trabanten`, eingegeben. In das Feld DATENBANKFELD habe ich `Durchschnittliche_Entfernung_x1000_km` eingegeben. Dies ist nicht so viel Tipparbeit, wie es erscheinen mag. Um die Namen einzugeben, können Sie einfach im Arbeitsblatt den betreffenden Zellbereich oder die betreffende Zelle markieren. Für Trabanten habe ich den Zellbereich der Datenbank markiert und für das Feld DATENBANKFELD die Zelle E10. Abschließend habe ich für das Feld SUCHKRITERIEN den Zellbereich C1:C2 markiert.

   Das Ergebnis, 9, wird im Dialogfeld angezeigt.

5. **Klicken Sie auf OK, um das Ergebnis in die markierte Zelle einzufügen.**

Alle Datenbankfunktionen haben das gleiche Format. Sie können sie auf die gleiche Art und Weise aktivieren und Sie geben die gleiche Art von Informationen in die entsprechenden Dialogfelder ein. Die Beschreibung der einzelnen Schritte lasse ich daher weg, wenn ich

die diversen Funktionen beschreibe, und stelle Ihnen stattdessen die entsprechenden Arbeitsblattformeln vor.

Datensätze zählen und Auszüge aus der Datenbank

Eine wesentliche Fähigkeit von Datenbanken ist es, Sie wissen zu lassen, wie viele Datensätze ein bestimmtes Kriterium erfüllen. Eine weitere wichtige Fähigkeit ist das Abrufen von Datensätzen. Hier nun die entsprechenden Excel-Versionen.

DBANZAHL und DBANZAHL2

Wie ich soeben gezeigt habe, können Sie mit der Funktion `DBANZAHL` Datensätze zählen. Dabei gibt es allerdings die Einschränkung, dass in dem festgelegten Feld Zahlen stehen müssen. Ist dies nicht der Fall, erhalten Sie null als Ergebnis, wie zum Beispiel bei

```
=DBANZAHL(Trabanten;Name;C1:C2)
```

da keiner der Datensätze im Feld `Name` Zahlen enthält.

`DBANZAHL2` zählt Datensätze anderes. Diese Funktion funktioniert mit jeder Art von Feld. Sie zählt die Anzahl der nicht leeren Datensätze im Feld, die das Suchkriterium erfüllen. Die folgende Formel liefert daher das Ergebnis 9:

```
=DBANZAHL2(Trabanten;Name;C1:C2)
```

Die Anwendung von »oder«

Hier eine Formel, die ein »oder« enthält:

```
=DBANZAHL2(Trabanten;Name;D1:D3)
```

Das Suchkriterium D1:D3 spezifiziert Trabanten, deren Umlaufzeit 1,26 Tage oder mehr als 20 Tage beträgt. Wie ich bereits erwähnte, bedeutet die Verwendung mehrerer Zeilen »oder«. Dieses Suchkriterium trifft auf fünf Trabanten zu: Deimos, Hyperion, Iapetus, auf unseren Mond und Nereid.

Platzhalter

Wenn Sie sich Abbildung A.1 einmal genauer ansehen, werden Sie in Zelle B2 den kryptischen Eintrag `*io` entdecken. Ich habe diesen Eintrag einfügt, um Ihnen zu zeigen, dass die Excel-Datenbankfunktionen mit Platzhaltern umgehen können. Die Formel

```
=DBANZAHL2(Trabanten;Name;B1:B2)
```

liefert das Ergebnis 3, die Anzahl der Trabanten, in deren Name an beliebiger Stelle die Zeichenfolge *io* enthalten ist (Dione, Io und Hyperion).

DBAUSZUG

`DBAUSZUG` ruft exakt einen Datensatz ab. Wenn das Suchkriterium auf mehr als einen (oder keinen) Datensatz zutrifft, gibt `DBAUSZUG` eine Fehlermeldung zurück.

Diese Formel

```
=DBAUSZUG(Trabanten;Name;D1:D2)
```

liefert als Ergebnis Deimos, den Namen des Trabanten, dessen Umlaufzeit 1,26 Tage beträgt.

Diese Formel

```
=DBAUSZUG(Trabanten;Name;E1:E2)
```

führt hingegen zu einer Fehlermeldung, da das Suchkriterium auf mehr als einen Datensatz zutrifft.

Arithmetik

Ohne seine Berechnungsmöglichkeiten wäre Excel nicht Excel. Hier einige der Möglichkeiten für Datenbanken.

DBMAX und DBMIN

Wie ihre Namen schon andeuten, liefern `DBMAX` und `DBMIN` den größten bzw. kleinsten Wert aus den Datenbankeinträgen. Die Formel

```
=DBMAX(Trabanten;Umlaufzeit_Tage;E1:E2)
```

liefert als Ergebnis 360,14. Dies ist die maximale Umlaufzeit eines Trabanten, der mehr als 150.000 km von seinem Planten entfernt ist.

Für den kleinsten Wert, auf den das Suchkriterium zutrifft, ergibt die Formel

```
=DBMIN(Trabanten;Umlaufzeit_Tage;E1:E2)
```

das Ergebnis -550,48. Das ist Phoebe, ein Trabant, der sich entgegen der Drehrichtung seines Planeten dreht.

DBSUMME

`DBSUMME` addiert die Werte in einem Feld. Um alle Umlaufzeiten von Trabanten, die von Galileo oder Cassini entdeckt wurden, zu summieren, verwenden Sie diese Formel:

```
=DBSUMME(Trabanten;Umlaufzeit_Tage;G1:G3)
```

Die Summe lautet 117,64.

Sie möchten alle Umlaufzeiten summieren? (Ich weiß … `=SUMME(D11:D35)`, aber lassen Sie mich trotzdem weitermachen.)

Diese Formel erledigt das für Sie:

```
=DBSUMME(Trabanten;Umlaufzeit_Tage;C1:C3)
```

Warum? Das Suchkriterium ist der Schlüssel. C1:C2 bedeutet Planet = Saturn oder … irgendetwas anderes, da C2 leer ist. Die Summe ist übrigens 35,46. Also: Seien Sie vorsichtig, wenn Sie eine Leerzeile in Ihr Suchkriterium einbinden.

DBPRODUKT

Hier nun eine Funktion, die es vermutlich nur deshalb gibt, weil die Entwickler von Excel sie erstellen konnten. Sie legen die Datenwerte fest und `DBPRODUKT` multipliziert sie.

Die Formel

```
=DBPRODUKT(Trabanten;Umlaufzeit_Tage;G1:G2)
```

ergibt das Produkt (749,832) der Umlaufzeiten der Trabanten, die von Galileo entdeckt wurden – eine Berechnung, von der ich sicher bin, dass Galileo nie darauf gekommen wäre.

Statistiken

Damit sind wir bei den Statistikfunktionen für Datenbanken angelangt. Sie funktionieren genauso wie die ähnlich benannten Tabellenblattfunktionen.

DBMITTELWERT

Hier kommt die Formel für den Mittelwert der Umlaufzeiten von Trabanten, die nach 1887 entdeckt wurden:

```
=DBMITTELWERT(Trabanten;Umlaufzeit_Tage;F1:F2)
```

Der Mittelwert ist negativ (-36,4080), da die Spezifikation auch die beiden Trabanten mit negativer Umlaufzeit umfasst.

DBVARIANZ und DBVARIANZEN

`DBVARIANZ` ist die Datenbankentsprechung von `VAR.S`, der Funktion, die die Summe von *N* Abweichungsquadraten durch *N*–1 dividiert. Das nennt man *Stichprobenvarianz*.

`DBVARIANZEN` ist die Datenbankentsprechung von `VAR.P`, der Funktion, die die Summe von *N* Abweichungsquadraten durch *N* dividiert. Das ist die *Grundgesamtheitsvarianz*. (Ausführliche Informationen zu den Funktionen `VAR.S` und `VAR.P`, der Stichprobenvarianz und der Grundgesamtheitsvarianz sowie zu den Zusammenhängen von *N* und *N*–1 finden Sie in Kapitel 5.)

Hier folgt nun die Stichprobenvarianz für die Umlaufzeit von Trabanten, die mehr als 150.000 km von ihren jeweiligen Planeten entfernt sind und die nach 1877 entdeckt wurden:

```
=DBVARIANZ(Trabanten;Umlaufzeit_Tage;E1:F2)
```

Das Ergebnis lautet 210.358,1.

Die Grundgesamtheitsvarianz eben dieser Gruppe von Trabanten ist

```
=DBVARIANZEN(Trabanten;Umlaufzeit_Tage;E1:F2)
```

Das Ergebnis lautet 140.238,7.

Auch hier gilt wieder Folgendes: Wenn Sie in Ihr Suchkriterium mehrere Spalten einschließen, bedeutet das »und«.

DBSTDABW und DBSTDABWN

Diesebeiden Funktionen geben Standardabweichungen zurück. Die Standardabweichung ist die Quadratwurzel der Varianz (siehe Kapitel 5). `DBSTDABW` gibt die Stichprobenstandardabweichung zurück, also die Quadratwurzel des Wertes, der durch die Funktion `DBVARIANZ` zurückgegeben wird. `DBSTDABWN` gibt die Grundgesamtheitsstandardabweichung zurück, also die Quadratwurzel des Wertes, der durch die Funktion `DBVARIANZEN` zurückgegeben wird.

Wenn man die Spezifikationen des vorstehenden Beispiels erneut anwendet, lautet die Formel für die Stichprobenstandardabweichung

```
=DBSTDABW(Trabanten;Umlaufzeit_Tage;E1:F2)
```

Das Ergebnis lautet 458,6481.

Die Formel für die Grundgesamtheitsstandardabweichung lautet

```
=DBSTDABWN(Trabanten;Umlaufzeit_Tage;E1:F2)
```

Das Ergebnis ist 374,4846.

Datenmasken verwenden

In Excel 2016 gibt es eine Datenmaske, die Ihnen die Arbeit mit Datenbanken erleichtert. Leider finden Sie im Menüband keine Schaltfläche, um die Datenmaske zu öffnen. Sie können diese Schaltfläche jedoch in die Symbolleiste für den Schnellzugriff einfügen.

Gehen Sie dazu folgendermaßen vor:

1. **Klicken Sie auf den Pfeil an der rechten Seite der Symbolleiste für den Schnellzugriff, um das Menü SYMBOLLEISTE FÜR DEN SCHNELLZUGRIFF ANPASSEN zu öffnen.**

2. **Klicken Sie in diesem Menü auf WEITERE BEFEHLE, um das Dialogfeld EXCEL-OPTIONEN zu öffnen.**

3. **Klicken Sie in der Dropdownliste BEFEHLE AUSWÄHLEN auf NICHT IM MENÜBAND ENTHALTENE BEFEHLE.**

4. **Blättern Sie im Listenfeld auf der linken Seite nach unten und markieren Sie den Befehl MASKE.**

5. **Klicken Sie auf die Schaltfläche HINZUFÜGEN, um den Befehl MASKE in das Listenfeld auf der rechten Seite einzufügen.**

6. **Klicken Sie auf OK, um das Dialogfeld EXCEL-OPTIONEN zu schließen.**

   Der Befehl MASKE ist nun in der Symbolleiste für den Schnellzugriff sichtbar.

Um die Datenmaske zu öffnen, markieren Sie eine beliebige Zelle in der Tabelle mit den Daten. Es ist nicht erforderlich, den gesamten Zellbereich zu markieren. Klicken Sie dann in der Symbolleiste für den Schnellzugriff auf MASKE.

In Abbildung A.3 sehen Sie die Eingabemaske. Excel trägt die Feldnamen automatisch ein und in den einzelnen Feldern stehen automatisch die Werte des ersten Datensatzes. Sie können die Maske dazu verwenden, innerhalb der Datenbank zu navigieren

oder einen Datensatz hinzuzufügen. Sie können mit einem Datensatz anfangen und dann einen neuen Datensatz hinzufügen und so weiter, aber ich finde es persönlich einfacher, die einzelnen Datensätze einfach in das Arbeitsblatt einzugeben.

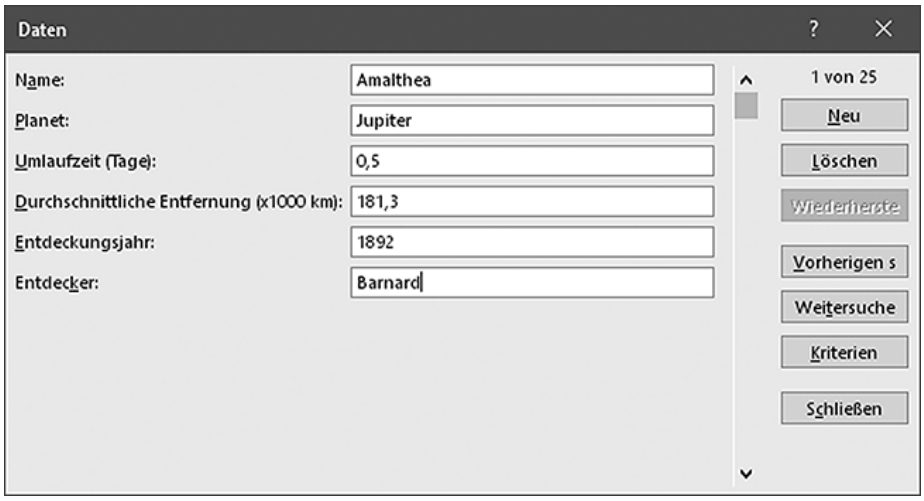


Abbildung A.3: Die Datenmaske, die Sie für Excel-Datenbanken verwenden können

Immer wenn Sie Datensätze hinzufügen (gleich auf welche Art und Weise), müssen Sie den Befehl FORMELN | NAMENS-MANAGER verwenden, um den Zellbereich, zu dem der Datenbankname gehört, entsprechend zu erweitern.

In der Titelleiste wird der Name des Arbeitsblatts angezeigt, wie er im Blattregister zu sehen ist. Daher ist es eine gute Idee, für die Blattregisterkarte einen aussagekräftigen Namen zu verwenden. Es ist deutlicher, wenn die Maske einen Titel wie *Daten* oder *Trabanten* besitzt, statt einfach nur *Blatt1*.

PivotTable-Berichte

Ein so genannter *PivotTable-Bericht* (auch einfach nur *PivotTable* genannt) ist eine Kreuztabelle – eine andere Art, Daten zu betrachten. Sie können die Datenbank reorganisieren und dabei im wahrsten Sinne des Wortes »umstülpen«. Dafür gibt es sogar mehrere Möglichkeiten.

So können Sie zum Beispiel eine PivotTable erstellen, bei der die Trabanten in mehreren Zeilen und die Planeten in jeweils eigenen Spalten stehen. Die Daten für die Umlaufzeiten stehen dann in den sich jeweils überschneidenden Zellen. In Abbildung A.4 ist dargestellt, was ich meine.

	A	B	C	D	E	F	G	H	I	J
1										
2										
3	Summe von Umlaufzeit (Tage)	Spaltenbeschriftungen								
4	Zeilenbeschriftungen	Erde	Jupiter	Mars	Neptun	Pluto	Saturn	Uranus	Gesamtergebnis	
5	Amalthea		0,5						0,5	
6	Ariel							2,52	2,52	
7	Callisto		16,69						16,69	
8	Charon					6,39			6,39	
9	Deimos			1,26					1,26	
10	Dione						2,74		2,74	
11	Enceladus						1,37		1,37	
12	Europa		3,55						3,55	
13	Ganymed		7,15						7,15	
14	Hyperion						21,28		21,28	
15	Iapetus						79,33		79,33	
16	Io		1,77						1,77	
17	Mimas						9,42		9,42	
18	Miranda							1,41	1,41	
19	Mond	27,32							27,32	
20	Nereid				360,14				360,14	
21	Oberon							13,46	13,46	
22	Phoboe						-550,48		-550,48	
23	Phobos			0,32					0,32	
24	Rhea						4,52		4,52	
25	Tethys						1,89		1,89	
26	Titan						15,94		15,94	
27	Titania							8,71	8,71	
28	Triton				-5,88				-5,88	
29	Umbriel							4,14	4,14	
30	Gesamtergebnis	27,32	29,66	1,58	354,26	6,39	-413,99	30,24	35,46	

Abbildung A.4: Eine PivotTable der Trabantendaten mit Trabanten, Planeten und Umlaufzeiten

In Abbildung A.5 ist eine PivotTable dargestellt, die eine andere Sicht auf die Daten bietet. In diesem Fall treten die einzelnen Trabanten in den Hintergrund und die Planeten stehen im Vordergrund. Die Zeilen der einzelnen Planeten sind wiederum in zwei Zeilen unterteilt – eine für die Umlaufzeit und eine für die durchschnittliche Entfernung. Die Zahlen sind die Summen für die Trabanten der jeweiligen Planeten. Wenn man zum Beispiel die Umlaufzeit von allen Trabanten des Jupiters addiert, lautet das Ergebnis 29,66.

	A	B	C
1			
2			
3	Zeilenbeschriftungen	Summe von Umlaufzeit (Tage)	Summe von Durchschnittliche Entfernung (x1000 km)
4	Erde	27,32	384,4
5	Jupiter	29,66	4226,8
6	Mars	1,58	32,84
7	Neptun	354,26	5868,2
8	Pluto	6,39	19,64
9	Saturn	-413,99	20838,79
10	Uranus	30,24	1605,33
11	Gesamtergebnis	35,46	32976

Abbildung A.5: Eine andere PivotTable der Trabantendaten mit Planeten, Umlaufzeiten und der durchschnittlichen Entfernung

In diesem Beispiel zeige ich exemplarisch, wie Sie die PivotTable in Abbildung A.4 erstellen. Gehen Sie dazu wie folgt vor:

1. **Öffnen Sie das Arbeitsblatt, das die Datenbank enthält.**

   In diesem Beispiel ist es das Arbeitsblatt Daten.

2. **Klicken Sie eine beliebige Zelle in der Datenbank an.**

3. **Wählen Sie EINFÜGEN | PIVOTTABLE, um das Dialogfeld PIVOTTABLE ERSTELLEN zu öffnen.**

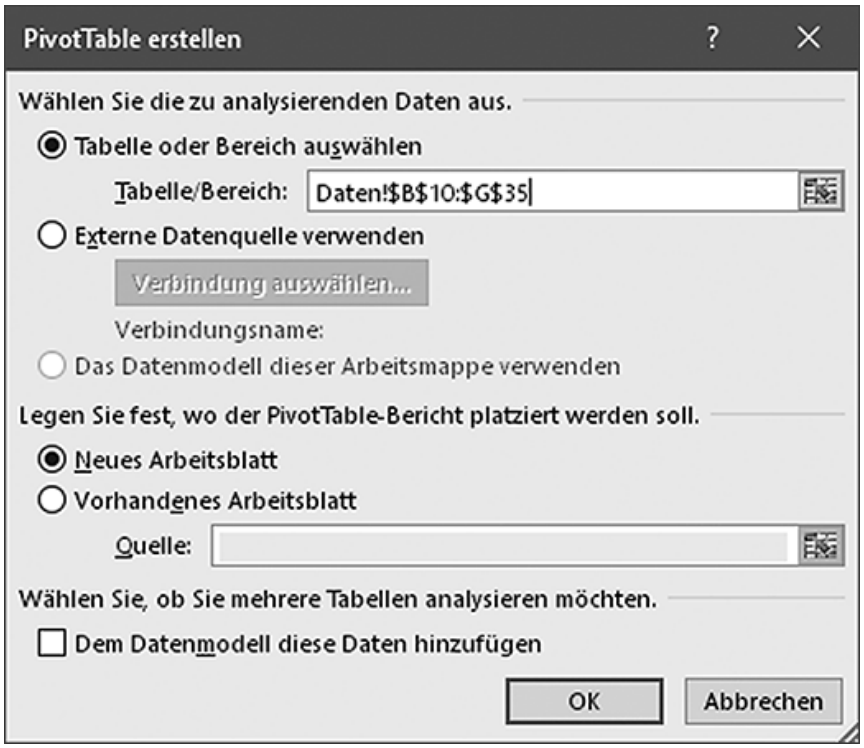


Abbildung A.6: Das Dialogfeld PIVOTTABLE ERSTELLEN

4. **Nehmen Sie Ihre Eingaben in diesem Dialogfeld vor.**

   Da ich vor dem Öffnen des Dialogfeldes bereits eine Zelle in dem Bereich mit der Datenbank markiert habe, ist das erste Optionsfeld bereits markiert und das Feld TABELLE/BEREICH ist ausgefüllt.

   Ich habe das Optionsfeld NEUES ARBEITSBLATT markiert, damit die PivotTable auf einem neuen Blatt platziert wird.

5. **Klicken Sie auf OK.**

   Excel öffnet ein neues Blatt mit dem Layout der PivotTable und dem Aufgabenbereich PIVOTTABLE-FELDER (siehe Abbildung A.7).

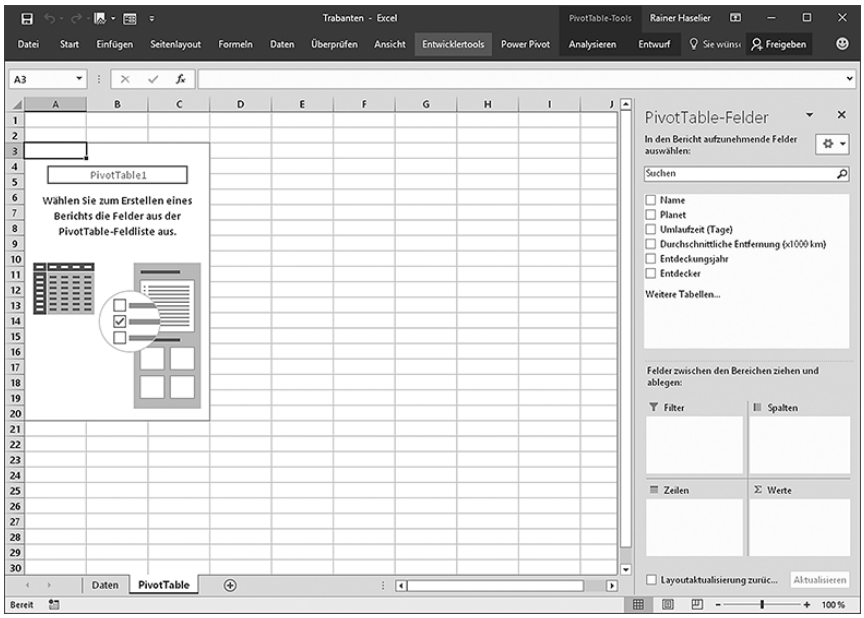


Abbildung A.7: Ein neues Arbeitsblatt mit dem PivotTable-Layout

6. **Füllen Sie das PivotTable-Layout aus.**

   Um das PivotTable-Layout mit Daten zu füllen, verwenden Sie den Aufgabenbereich PIVOTTABLE-FELDER. Ziehen Sie eines der Felder in das betreffende Kästchen unten im Aufgabenbereich und legen Sie es dort ab.

   Ich beginne mit den Zeilen. Damit die Namen der Trabanten in den Zeilen erscheinen, wähle ich das Feld NAME aus und ziehe es in das Kästchen ZEILEN. Das Ergebnis ist in Abbildung A.8 dargestellt. Zusätzlich zu den Trabantennamen in den Zeilen erscheint NAME in der Feldliste in Fettdruck und das betreffende Kontrollkästchen ist aktiviert. Diese Informationen zeigen an, dass sich das Feld in der PivotTable befindet. Wenn Sie das Kontrollkästchen ausschalten, wird das Feld aus der PivotTable entfernt.

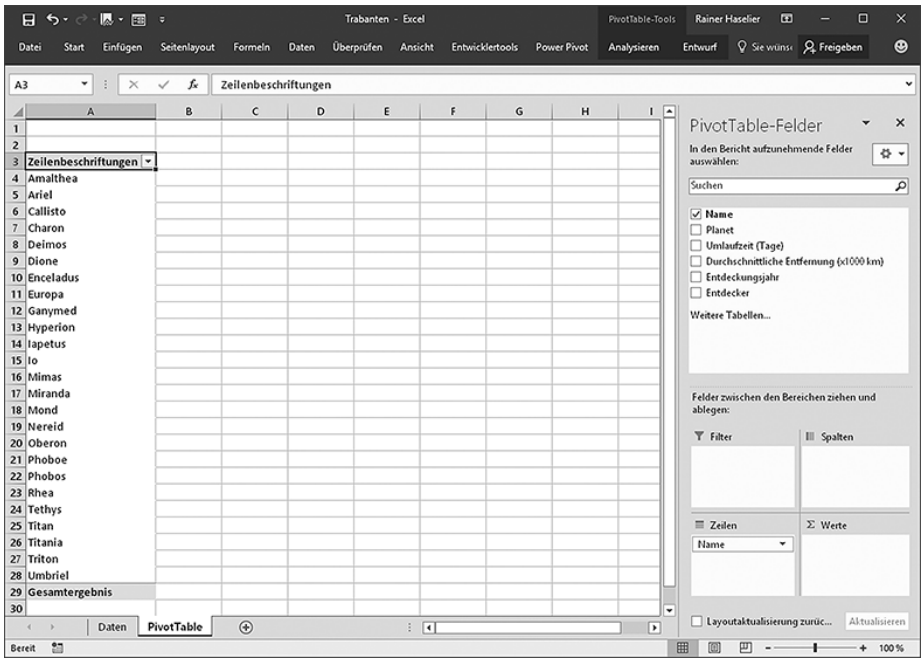


Abbildung A.8: Das PivotTable-Layout, nachdem das Feld »Name« in das mit »Zeilen« beschriftete Kästchen gezogen wurde

Anschließend habe ich das Feld PLANET in das Kästchen SPALTEN gezogen (siehe Abbildung A.9).

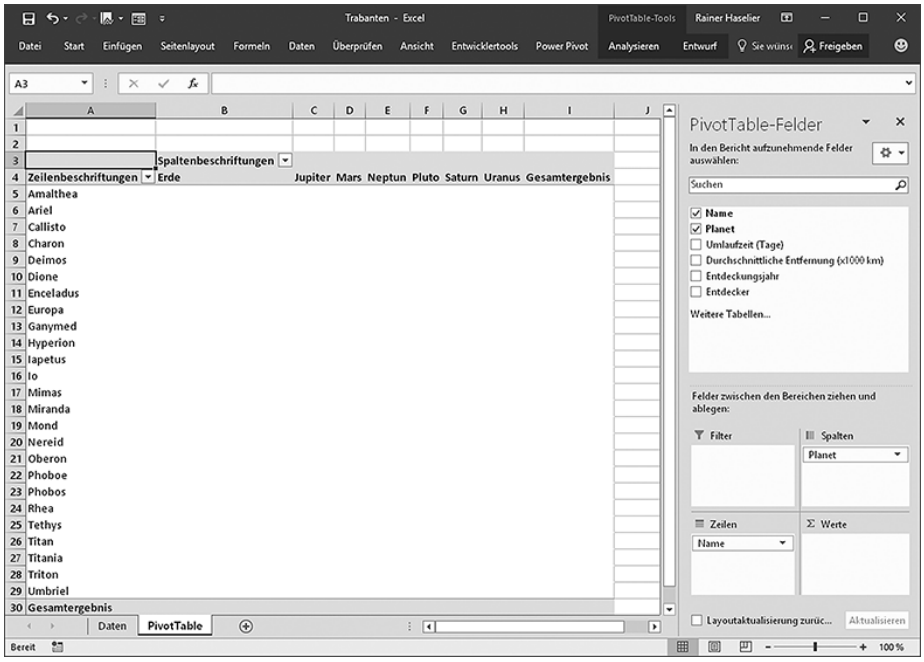


Abbildung A.9: Das PivotTable-Layout, nachdem das Feld »Planet« in das mit »Spalten« beschriftete Kästchen gezogen wurde

Wenn man das Feld UMLAUFZEIT (TAGE) aus der Feldliste in das Kästchen Σ WERTE zieht, ergibt sich die in Abbildung A.4 dargestellte Tabelle.

Wenn Sie ein Feld in das Kästchen BERICHTSFILTER ziehen, erhalten Sie eine Art mehrseitige Version der Tabelle. Wenn Sie beispielsweise das Feld ENTDECKER in den Bereich BERICHTSFILTER ziehen, erhalten Sie ein Dropdownmenü, mit dem Sie lediglich die von einer bestimmten Person entdeckten Trabanten anzeigen lassen können.

Der kleine nach unten zeigende Pfeil neben den Feldnamen öffnet ein Menü, mit dessen Befehlen Sie nach diesem Feld sortieren und filtern können.

PivotTables sind deshalb so wichtig, weil Sie darüber direkt mit den Daten arbeiten können. Wenn Sie ein Feld in die Tabelle ziehen oder wieder entfernen, können Sie Zusammenhänge erkennen und Analysen durchführen, an die Sie nicht einmal denken würden, wenn Sie nur mit der ursprünglichen Datenbank arbeiten.

Anahng B

Die Kovarianzanalyse

In diesem Kapitel

- Kovarianz unter der Lupe
- Warum und wie wird die Kovarianz analysiert
- Kovarianzanalyse (ANCOVA) mit Excel

Eine der wichtigsten Erkenntnisse dieses Buches ist, dass Excel mit einer überraschenden Vielzahl von analytischen Werkzeugen und Funktionen ausgestattet ist. Obwohl die Werkzeugsammlung nicht so allumfassend ist wie in einer speziellen Statistiksoftware, so ist sie dennoch beeindruckend.

Bestimmte Analysen stehen jedoch in Excel nicht zur Verfügung. Und manche dieser Verfahren könnten für Sie wichtig sein. In bestimmten Fällen können Sie mit ein wenig Einfallsreichtum die bestehenden Möglichkeiten von Excel verwenden, um diese Analysen dennoch durchzuführen. In diesem Anhang konzentriere ich mich auf eine dieser Analysen, die Kovarianzanalyse.

Kovarianz unter die Lupe genommen

In Kapitel 15 habe ich die Kovarianz im Zusammenhang mit der Korrelation kurz erwähnt, und zwar als Zähler des Korrelationskoeffizienten. Außerdem habe ich darauf hingewiesen, dass die Kovarianz angibt, dass sich zwei Variablen gemeinsam verändern.

Was genau bedeutet dies?

Stellen Sie sich eine Gruppe Personen vor, deren mathematischen Fähigkeiten und Kontaktfreudigkeit Sie messen. (Nehmen Sie hierbei einfach an, dass Ihnen valide und zuverlässige Verfahren zur Verfügung stehen, beides zu messen.) Wenn Sie feststellen, dass die Leute mit den besten mathematischen Fähigkeiten die kontaktfreudigsten sind, dann besitzt dieses Ding namens Kovarianz einen hohen numerischen Wert, der positiv ist. Diese Art von positiver Beziehung wird *direkte* Beziehung genannt.

Es ist jedoch auch ein anderes Ergebnis möglich: Es kann sich auch ergeben, dass die Leute mit den besten mathematischen Fähigkeiten die am wenigsten kontaktfreudigen sind. In diesem Fall hat die Kovarianz einen hohen numerischen Wert, der jedoch negativ ist. Diese Art von negativer Beziehung wird *inverse* Beziehung genannt.

Und es gibt noch eine dritte Variante, nämlich, dass es keinen Zusammenhang zwischen den mathematischen Fähigkeiten und der Kontaktfreudigkeit gibt. In diesem Fall sind die beiden Variablen *unabhängig* und der numerische Wert der Kovarianz ist nahe null. Der Wert kann leicht positiv oder leicht negativ sein, aber er ist bei dieser Variante immer nahe null.

Hört sich ein wenig an wie Korrelation? So sollte es auch sein. Wie weiter vorne erwähnt, ist die Kovarianz der Zähler des Korrelationskoeffizienten. Sie können sich dies auch so vorstellen, dass der Korrelationskoeffizient die untere und obere Grenze der Kovarianz festlegt. Mit ein wenig Erfahrung können Sie verstehen, was ein Korrelationskoeffizient von 0,98 oder 0,62 bedeutet. Ohne ausreichend Informationen über den Kontext lässt sich nicht verstehen, was sich hinter einer Kovarianz von 473,5 verbirgt.

Warum Sie die Kovarianz analysieren

Das war also ein kleiner Auffrischungskurs zum Thema Kovarianz. Und warum wollen Sie diese analysieren?

Lassen Sie uns mit einer typischen Studie beginnen. Sie weisen verschiedenen Personen unterschiedliche Bedingungen zu und messen dann unter diesen Bedingungen deren Leistung. Beispielsweise können Sie 15 Kinder einer von drei Gruppen zuweisen, die sich darin unterscheiden, wie sich die Kinder auf eine Prüfung vorbereiten. Eine Gruppe wird von einem echten Lehrer unterrichtet. Eine andere bereitet sich mit einem interaktiven Lehrprogramm auf einem Computer vor. Die dritte Gruppe liest ein Lehrbuch. Danach legen sie die Prüfung ab.

Das Prüfungsergebnis des Kindes ist die abhängige Variable. Die drei verschiedenen Bedingungen zur Vorbereitung stellen zusammen die unabhängige Variable dar. Das Ziel besteht darin herauszufinden, ob die unterschiedlichen Bedingungen der unabhängigen Variable Auswirkungen auf die abhängige Variable haben. Kapitel 12 beschreibt, dass Sie hierfür einen Hypothesentest benötigen, der folgendermaßen aussieht:

$$H_0 : \mu_{Lehrer} = \mu_{Computer} = \mu_{Lehrbuch}$$
$$H_1 : Nicht\ H_0$$

In Kapitel 12 habe ich außerdem darauf hingewiesen, dass das geeignete statistische Verfahren die Varianzanalyse (ANOVA) darstellt.

So weit, so gut. Was ist nun mit der Kovarianz?

Neben der abhängigen und unabhängigen Variablen kann ein dritter Variablentyp ins Spiel kommen. Und zwar so: Angenommen, Sie besitzen eine weitere, relevante Maßzahl für jedes der 15 Kinder: die mathematische Begabung. Neben der Art der Vorbereitung könnte auch diese das Ergebnis der Prüfung beeinflussen. Diese dritte Variable wird Kovariate genannt. Die Beziehung zwischen der abhängigen Variablen und der Kovariaten ist die Kovarianz.

Die ganz Großen im Bereich Forschungsdesign und Analyse haben für das zufällige Zuweisen von Individuen an unterschiedliche

Bedingungen der unabhängigen Variablen und dem Konstanthalten der anderen Bedingungen (wie der Tageszeit, zu der die Prüfung durchgeführt wurde, der Zeit, die ein Kind für die Prüfungsvorbereitung aufbringt, der Zeit, die einem Kind für die Prüfung zur Verfügung steht) einen Namen. Sie nennen dies *experimentelle Kontrolle*.

Auch für die Bewertung der Effekte einer Kovariate, also der Kovarianz mit der abhängigen Variablen, gibt es einen Namen. Diese wird *statistische Kontrolle* genannt. Beide sind wertvolle Werkzeuge im Repertoire eines Analytikers.

Die entscheidende Frage ist: Warum benötigen Sie die statistische Kontrolle? Angenommen, Sie führen eine Studie durch und können zwischen den unterschiedlichen Vorbereitungsgruppen keine signifikanten Unterschiede feststellen. Dies kann bedeuten, dass die experimentelle Kontrolle nicht mächtig genug war, um einen Unterschied auszumachen, der in der Art der Vorbereitung bedingt ist. An dieser Stelle kann die statistische Kontrolle helfen. Angenommen, die mathematische Begabung hat das Prüfungsergebnis so beeinflusst, dass hierdurch die Effekte der Vorbereitungsart verdeckt werden. Die Frage lautet, ob eine mögliche Korrelation zwischen dem Prüfungsergebnis und der Begabung die Ergebnisse beeinflusst hat.

Indem die experimentelle und die statistische Kontrolle miteinander kombiniert werden, kann die Kovarianzanalyse (ANCOVA) diese Frage beantworten.

Wie Sie die Kovarianz analysieren

Wie kombinieren Sie diese beiden Arten der Kontrolle?

In Kapitel 12 habe ich ausgeführt, dass ANOVA SS_{Gesamt} in $SS_{Innerhalb}$ und $SS_{Zwischen}$ aufteilt. Wenn Sie jede diese Quadratsummen (SS = Sum of Square) durch die Freiheitsgrade dividieren, erhalten Sie drei MS-Werte (Varianzen, MS = mittlere Quadratsumme). $MS_{Zwischen}$

spiegelt die Unterschiede zwischen den Gruppenmittelwerten wider. $MS_{Innherhalb}$ schätzt die Varianz der Grundgesamtheit. Diesem Wert liegen die gepoolten Varianzen innerhalb der Gruppen zugrunde. Wenn $MS_{Zwischen}$ signifikant größer ist als $MS_{Innerhalb}$, können Sie die Nullhypothese verwerfen. Falls nicht, dann können Sie es nicht. (Sollte Ihnen das alles ein wenig merkwürdig vorkommen, finden Sie in Kapitel 12 ausführliche Informationen.)

Bei der Kovarianzanalyse (ANCOVA) verwenden Sie die Beziehung zwischen der abhängigen Variablen und der Kovariate, um $SS_{Zwischen}$ und $SS_{Innerhalb}$ zu justieren. Falls es sich um eine starke Beziehung handelt, ist es wahrscheinlich, dass die Anpassung $SS_{Zwischen}$ vergrößert und $SS_{Innerhalb}$ reduziert. Wie im Rest der Welt gibt es auch in der Statistik den Lunch nicht umsonst: Durch die Anpassung wird $df_{Innerhalb}$ um 1 verkleinert. Vielleicht hilft es, sich das Ganze so vorzustellen:

ANOVA:

$$F = \frac{SS_{Zwischen} / df_{Zwischen}}{SS_{Innerhalb} / df_{Innerhalb}}$$

ANCOVA:

$$F = \frac{(SS_{Zwischen} + etwas, das\ auf\ der\ Beziehung\ mit\ der\ Kovariaten\ basiert) / df_{Zwischen}}{(SS_{Innerhalb} + etwas, das\ auf\ der\ Beziehung\ mit\ der\ Kovariaten\ basiert / df_{Innerhalb-1}}$$

Ein mögliches Ergebnis der Anpassung kann (wenn die Beziehung stark ist) darin bestehen, dass der F-Wert bei ANCOVA größer ist als bei ANOVA. Praktisch bedeutet dies, dass Sie durch Verwendung der statistischen Kontrolle eine aussagekräftigere Studie erhalten (also eher die Möglichkeit haben, einen Effekt zu sehen), als wenn Sie lediglich die experimentelle Kontrolle einsetzen.

Hier noch ein weiterer Ansatz, um die Unterschiede zwischen ANOVA und ANCOVA zu verstehen. ANOVA hilft, eine Nadel im Heuhaufen zu finden. ANCOVA dient dem gleichen Zweck, jedoch

entfernt diese Analyse etwas Heu aus dem Haufen und macht gleichzeitig die Nadel ein wenig größer.

ANCOVA mit Excel

Obwohl Excel keine fertigen Tools für ANCOVA enthält, so können Sie das verwenden, was Excel bietet, um die Kovarianzanalyse erheblich einfacher durchzuführen, als es in Statistikbüchern aussieht.

Was stellt Excel zur Verfügung? Im Hinblick auf die Kovarianz erinnern Sie sich vielleicht daran, dass die Tabellenblattfunktionen `KOVARIANZ.P` und `KOVARIANZ.S` Ihnen die Rechenarbeit abnehmen und dass Sie vom Datenanalyse-Tool Kovarianz Unterstützung erhalten.

Es hört sich so an, als ob sich das Datenanalyse-Tool Kovarianz hervorragend für die Kovarianzanalyse (ANCOVA) eignen würde. Seltsamerweise ist dies nicht der Fall. Die Tabellenblattfunktionen helfen Ihnen hier auch nicht weiter.

Stattdessen stelle ich nachfolgend zwei Ansätze vor, die andere Excel-Features verwenden, um die ANCOVA durchzuführen. Ein Ansatz verwendet Anova: einfaktorielle Varianzanalyse in Kombination mit ein paar Tabellenblattfunktionen. Der andere Ansatz verwendet das Analysewerkzeug Regression, und zwar ebenfalls in Kombination mit einigen Tabellenblattfunktionen.

Beide Ansätze nehmen sich die Daten vor, die Sie in Tabelle B.1 sehen. In beiden Verfahren verwende ich die Steigungen der Regressionsgeraden, um die Beziehung zwischen der abhängigen Variablen und der Kovariaten auszudrücken.

Lehrer			Computer			Fachbuch		
Math.	Begabung	Punkte	Math.	Begabung	Punkte	Math.	Begabung	Punkte
10		6	7		9	7		9
9		9	7		5	9		12

Lehrer		Computer		Fachbuch	
Math. Begabung	Punkte	Math. Begabung	Punkte	Math. Begabung	Punkte
8	7	8	14	4	9
6	2	11	10	11	18
9	10	11	15	7	11

Tabelle B.1: Daten für das Ergebnis einer Prüfung bei drei verschiedenen Arten der Vorbereitung und unterschiedlicher mathematischer Begabung

Diese Methoden sind für die Arten von Forschungsdesigns geeignet, die ich in Kapitel 12 beschreibe (einfaktorielle Anova). Sie können ANCOVA bei jedem Forschungsdesign verwendet, bei dem die Varianzanalyse verendet wird, jedoch eignen sich die beiden in diesem Anhang beschriebenen Methoden nicht bei komplizierten Designs (wie beispielsweise den in Kapitel 13 beschriebenen).

Methode 1: ANOVA

In den meisten Fachbüchern, die das Thema Kovarianzanalyse behandeln, stoßen Sie auf eine Vielzahl mysteriös aussehender Berechnungsformeln, die entwickelt wurden, um die noch viel komplizierter aussehenden Berechnungen zu vermeiden, die auf Regression basieren. Dies führt oft dazu, dass sie verbergen, worum es bei ANCOVA eigentlich geht.

Mit der in diesem und der im folgenden Abschnitt vorgestellten Methode zeige ich Ihnen, wie Sie die Möglichkeiten von Excel verwenden können, um all dies zu umschiffen. Mein Ziel ist, ANCOVA viel einfacher zu machen, als sie in statistischen Fachbüchern aussieht.

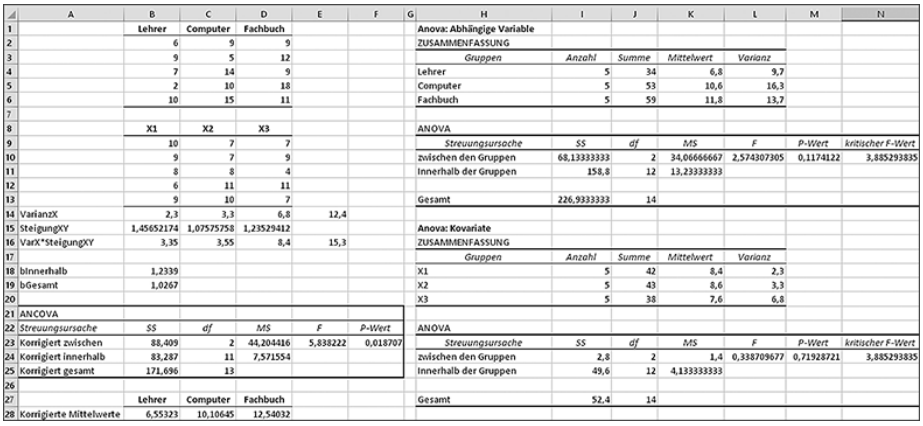
Abbildung B.1 zeigt das ANCOVA-Arbeitsblatt für diese Methode. Nachfolgend beschreibe ich die Schritte, die ich beim Erstellen des Arbeitsblatts durchgeführt habe.

Die Daten aus Tabelle B.1 befinden sich im Zellbereich B1:D13. Dort finden Sie eine Tabelle mit den Daten der anhängigen Variablen und eine weitere mit den Daten der Kovariaten. Ich habe die Daten darum so angeordnet, da ich das Analysewerkzeug *Anova:*

Einfaktorielle Varianzanalyse einmal für die abhängige Variable und ein weiteres Mal für die Kovariate verwenden werden. Der Eingabebereich für das Anova-Tool muss ein zusammenhängender Zellbereich sein. Daher kann das Layout aus Tabelle B.1 nicht verwendet werden.

Zuerst muss ich die Voraussetzungen für die Berechnung von $b_{Innerhalb}$ schaffen; dies ist ein Maß, das ich verwende, um die Quadratsummen (SS) und die Gruppenmittelwerte für die postanalytischen Tests zu korrigieren.

Was ist $b_{Innerhalb}$? Stellen Sie sich für jede der drei Gruppen ein Scatterplot vor, in dem jeweils auch eine Regressionsgrade eingezeichnet ist. Jede Regressionsgrade besitzt eine Steigung. Der Wert von $b_{Innerhalb}$ ist der Mittelwert der Steigungen der Gruppen, wobei jede Steigung anhand der Varianz der Kovariate innerhalb der Gruppe gewichtet wird.



	A	B	C	D	E	F
1		Lehrer	Computer	Fachbuch		
2		6	9	9		
3		9	5	12		
4		7	14	9		
5		2	10	18		
6		10	15	11		
7						
8		X1	X2	X3		
9		10	7	7		
10		9	7	9		
11		8	8	4		
12		6	11	11		
13		9	10	7		
14	VarianzX	2,3	3,3	6,8	12,4	
15	SteigungXY	1,45652174	1,07575758	1,23529412		
16	VarX*SteigungXY	3,35	3,55	8,4	15,3	
17						
18	bInnerhalb	1,2339				
19	bGesamt	1,0267				
20						
21	ANCOVA					
22	*Streuungsursache*	*SS*	*df*	*MS*	*F*	*P-Wert*
23	Korrigiert zwischen	88,409	2	44,204416	5,838222	0,018707
24	Korrigiert innerhalb	83,287	11	7,571554		
25	Korrigiert gesamt	171,696	13			
26						
27		Lehrer	Computer	Fachbuch		
28	Korrigierte Mittelwerte	6,55323	10,10645	12,54032		

	H	I	J	K	L	M	N
1	Anova: Abhängige Variable						
2	ZUSAMMENFASSUNG						
3	*Gruppen*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
4	Lehrer	5	34	6,8	9,7		
5	Computer	5	53	10,6	16,3		
6	Fachbuch	5	59	11,8	13,7		
7							
8	ANOVA						
9	*Streuungsursache*	*SS*	*df*	*MS*	*F*	*P-Wert*	*kritischer F-Wert*
10	zwischen den Gruppen	68,13333333	2	34,06666667	2,574307305	0,1174122	3,885293835
11	Innerhalb der Gruppen	158,8	12	13,23333333			
12							
13	Gesamt	226,9333333	14				
14							
15	Anova: Kovariate						
16	ZUSAMMENFASSUNG						
17	*Gruppen*	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*		
18	X1	5	42	8,4	2,3		
19	X2	5	43	8,6	3,3		
20	X3	5	38	7,6	6,8		
21							
22	ANOVA						
23	*Streuungsursache*	*SS*	*df*	*MS*	*F*	*P-Wert*	*kritischer F-Wert*
24	zwischen den Gruppen	2,8	2	1,4	0,338709677	0,71928721	3,885293835
25	Innerhalb der Gruppen	49,6	12	4,133333333			
26							
27	Gesamt	52,4	14				

Abbildung B.1: Das Arbeitsblatt für ANCOVA, Methode 1, für die Daten aus Tabelle B.1

Dies ist vermutlich einer der seltenen Fälle, in denen eine Formel aussagekräftiger ist als viele Worte. Hier ist sie:

$$b_{Innerhalb} = \frac{\sum s_{xi}^2 \; b_i}{\sum s_{xi}^2}$$

Zuerst habe ich mit `VAR.S` die Varianz jeder Gruppe innerhalb der Kovariaten berechnet. Diese Varianzen befinden sich in Zeile 14 und sind mit *VarianzX* beschriftet. Ich hätte auch das Anova-Tool verwenden können, um diese Varianzen zu erhalten. Der Einsatz der Funktion macht das Ganze jedoch deutlicher. Die Summe dieser Varianzen befindet sich in Zelle E14; sie beträgt 12,4.

Anschließend habe ich mit der Funktion `STEIGUNG` die Steigung innerhalb der Gruppen berechnet. Die Ergebnisse befinden sich in Zeile 15 und sind mit *SteigungXY* beschriftet.

Zeile 16, die mit *VarX*SteigungXY* beschriftet ist, enthält das Produkt aus der Steigung der einzelnen Gruppen und der Varianz in dieser Gruppe. In Zelle E16 befindet sich die Summe dieser Produkte.

Der Wert von $b_{Innerhalb}$ befindet sich in Zelle B18. Die Formel in dieser Zelle lautet:

```
=E16/E14
```

In Zelle B19 befindet sich eine weitere Steigung, die ich für die Anpassung der Quadratsummen SS verwende. Diesen Wert nenne ich b_{Gesamt} und er repräsentiert die Steigung der Regressionsgeraden, die in einem Scatterplot mit allen Werten, unabhängig von der Gruppe, in der sie sich befindet, gezeichnet wird. Die Formel für B19 lautet:

```
=STEIGUNG(B2:D6;B9:D13)
```

Anschließend habe ich in den Zellen A21:F25 die schick aussehende ANCOVA-Tabelle erstellt. Das Ziel ist, diese Tabelle zu füllen. Ich habe die Zellen und Rahmen so formatiert, damit die Tabelle so ähnlich aussieht wie die Tabelle, die das Datenanalysetool Anova: Einfaktorielle Varianzanalyse erstellt. Anschließend habe ich in den Zellen A23:A25 die Streuungsursachen eingegeben und in C23:C25 die Freiheitsgrade. Die df-Werte sind mit denen von ANOVA identisch, jedoch verlieren Sie einen Freiheitsgrad wegen $df_{Innerhalb}$; dies spiegelt sich dann auch in df_{Gesamt} wieder.

Die nächste Aufgabe besteht darin, das Tool Anova: Einfaktorielle Varianzanalyse zu verwenden. Ich habe es einmal für die abhängige Variable und ein weiteres Mal für die Kovariate ausgeführt. Warum für die Kovariate? Schließlich führe ich für die mathematische Begabung keine Hypothesentests durch. Der Grund für die Varianzanalyse über die Kovariate besteht darin, dass die Ausgabe von ANOVA die SS-Werte liefert, die ich für die Vervollständigung von ANCOVA brauche.

Wie Sie Anova: Einfaktorielle Varianzanalyse verwenden, habe ich in Kapitel 12 beschrieben. Der einzige Unterschied besteht darin, dass ich die Ergebnisse diesmal in das aktuelle und nicht in ein neues Tabellenblatt ausgeben lasse.

Ich habe ein paar Zeilen aus der Ausgabe des Anova-Tools entfernt, damit alles gut in einen Bildschirmausschnitt passt. Außerdem habe ich der besseren Übersichtlichkeit wegen die Tabelle ein wenig bearbeitet.

Mit den Werten aus den ANOVA-Tabellen kann ich die Werte in der ANCOVA-Tabelle erstellen. Die korrigierte Quadratsumme SS_{Gesamt} wird folgendermaßen berechnet:

$$Korrigierte\ SS_{GesamtY} = SS_{GesamtY} - b^2_{Gesamt} SS_{GesamtX}$$

$$Korrigierte\ SS_{GesamtY} = 226,93 - (1,03)^2\ (52,4) = 171,70$$

Also muss ich in Zelle B25 der ANCOVA-Tabelle diese Formel eingeben:

```
=I13-B19^2*I27
```

Übrigens habe ich die Werte hier im Buch auf zwei Dezimalstellen abgerundet, damit die Formeln besser lesbar sind. Im Arbeitsblatt selbst finden Sie mehr Dezimalstellen.

Weiter geht's mit der korrigierten Quadratsumme $SS_{Innerhalb}$. Die Formel für die Korrektur lautet:

$$Korrigierte\ SS_{InnerhalbY} = SS_{InnerhalbY} - b^2_{Innerhalb} SS_{GesamtY}$$

$$Korrigierte\ SS_{InnerhalbY} = 158,80 - (1,23)^2\ (49,6) = 83,29$$

Darum füge ich in Zelle B24 diese Formel ein:

```
=I11-B18^2*I25
```

Was ist mit der korrigierten Quadratsumme $SS_{Zwischen}$? Dies wird in Zelle B23 mit einer einfachen Subtraktion berechnet:

```
=B25-B24
```

Ich habe die ANCOVA-Tabelle fertiggestellt, indem ich jede korrigierte Quadratsumme durch die Freiheitsgrade dividiere und dann die korrigierte mittlere Quadratsumme $MS_{Zwischen}$ durch die korrigierte mittlere Quadratsumme $MS_{Innerhalb}$ dividiere, um so *F* zu berechnen.

Die letzte Spalte in der ANCOVA-Tabelle, der P-Wert, verwendet einen kleinen Trick und die Funktion `F.VERT.RE`. Die Formel in Zelle F23 lautet:

```
=F.VERT.RE(E23;C23,C24)
```

Direkt unterhalb der ANCOVA-Tabelle habe ich die korrigierten Mittelwerte für die abhängige Variable eingefügt. Diese werden bei den postanalytischen Tests benötigt, zu denen ich weiter hinten in diesem Abschnitt komme. Die verschiedenen korrigierten Mittelwerte werden anhand dieser Formel berechnet:

$$Korrigiertes\ \bar{y}_i = \bar{y}_i - b_{Innerhalb} \left(\bar{x}_i - \bar{X} \right)$$

Für unser Beispiel sind dies die korrigierten Mittelwerte:

$$Korrigiertes\ \bar{y}_{Lehrer} = 6,8 - 1,23(8,4 - 8,2) = 6,55$$

$$Korrigiertes\ \bar{y}_{Computer} = 10,6 - 1,23(8,6 - 8,2) = 10,11$$

$$Korrigiertes\ \bar{y}_{Fachbuch} = 11{,}8 - 1{,}23(7{,}6 - 8{,}2) = 12{,}54$$

Diese Korrekturen vergrößern die Streuung zwischen den Mittelwerten.

In diesem Arbeitsblatt befinden sich die Gruppenmittelwerte in den ANOVA-Tabellen. Genau gesagt, befinden sie sich für die abhängige Variable in K4:K6 und für die Kovariate in K18:K20. Daher lautet die Formel für den korrigieren Mittelwert Lehrer (in Zelle B28):

```
=K4-B18*(K18-MITTELWERT(B9:D13))
```

Die Formel für den korrigieren Mittelwert Computer (in Zelle C28) lautet:

```
=K5-B18*(K19-MITTELWERT(B9:D13))
```

Die Formel für den korrigieren Mittelwert Fachbuch (in Zelle D28) lautet:

```
=K6-B18*(K20-MITTELWERT(B9:D13))
```

Was ist der Vorteil von ANCOVA? Werfen Sie nochmal einen Blick auf Abbildung B.1. Vergleichen Sie die Werte der abhängigen Variablen in ANOVA ($F_{2,12}$ = 2,57) mit dem Ergebnis in ANCOCA ($F_{2,11}$ = 5,84). Obwohl Sie einen Freiheitsgrad geopfert haben, zeigt der *P*-Wert an, dass das ANCOVA-Ergebnis signifikant ist. Das ANOVA-Ergebnis ist es nicht. Durch die ANCOVA wurde $SS_{Zwischen}$ nach oben korrigiert und $SS_{Innerhalb}$ nach unten.

Die Moral von der Geschichte: Die Beziehung zwischen der abhängigen Variablen und der Kovariaten ermöglich es Ihnen, einen signifikanten Effekt aufzudecken, den Sie anderenfalls nicht gesehen hätten. In unserem Beispiel vermeidet ANCOVA einen Fehler zweiter Art.

Methode 2: Regression

Wenn die oben beschriebene Methode funktioniert, warum behellige ich Sie dann damit, Ihnen eine weitere Methode vorzustellen? Noch

wichtiger ist die Frage, warum Sie sich die Mühe machen sollten, diese zu erlernen und zu meistern?

Der Grund, warum ich es für sinnvoll halte weiterzulesen, ist: Die Methode, die ich Ihnen nun zeigen werde, erfordert es, dass Sie Ihre Denkweise über die Daten ändern. Falls Sie dazu in der Lage sind, eröffnen sich Ihnen neue Sichtweisen auf statistische Analysen und Sie können gleichzeitig einige neue statistische Konzepte verstehen.

Hier nun konkret, was ich damit meine. Das Arbeitsblatt in Abbildung B.2 zeigt die Daten aus Tabelle B.1 in einem anderen Layout. Die Abbildung zeigt gleichzeitig den Status, nachdem die Arbeiten an der Kovarianzanalyse abgeschlossen sind.

In Spalte A befinden sich die Prüfungsergebnisse und in Spalte B die Daten zur mathematischen Begabung. Es fällt auf, dass diese nicht in drei verschiedene Gruppen eingeteilt sind. Was ist da los?

Mein Plan besteht darin, das als multiple Regression zu behandeln. Das Prüfungsergebnis ist hierbei die abhängige Variable und die mathematische Begabung ist eine unabhängige Variable. Sie ist jedoch nicht die einzige unabhängige Variable.

	A	B	C	D	E	F	G	H	I	J	K
1	Note	Math. Begabung	Kennung1	Kennung2							
2	6	10	1	0							
3	9	9	1	0		*Regressions-Statistik*					
4	7	8	1	0		Multipler Korrelationskoeffizient	0,795605872				
5	2	6	1	0		Bestimmtheitsmaß	0,632988704				
6	10	9	1	0		Adjustiertes Bestimmtheitsmaß	0,532894714				
7	9	7	0	1		Standardfehler	2,751645735				
8	5	7	0	1		Beobachtungen	15				
9	14	8	0	1							
10	10	11	0	1		ANOVA					
11	15	10	0	1			*df*	*SS*	*MS*	*F*	*F krit*
12	9	7	0	0		Regression	3	143,6462366	47,88207885	6,323943177	0,00944713
13	12	9	0	0		Residue	11	83,28709677	7,571554252		
14	9	4	0	0		Gesamt	14	226,9333333			
15	18	11	0	0							
16	11	7	0	0			*Koeffizienten*	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*
17						Schnittpunkt	2,422580645	3,214266308	0,753696307	0,466858375	-4,651971799
18	SSGesamtX =	52,4				Math. Begabung	1,233870968	0,39070744	3,158043186	0,00911065	0,373929691
19						Kennung1	-5,987096774	1,768140033	-3,386098761	0,006076656	-9,878746747
20	bgesamt=	1,026717557				Kennung2	-2,433870968	1,783612627	-1,364573748	0,199650778	-6,359575892
21											
22	Mittelwerte					ANCOVA					
23	Gruppe	Abh. Variable	Kovariate	Korrigiert		*Streuungsursache*	*SS*	*df*	*MS*	*F*	*F krit*
24	Lehrer	6,8	8,4	6,5532		Korrigiert zwischen	88,409	2	44,204416	5,838222	0,018707
25	Computer	10,6	8,6	10,3532		Korrigiert innerhalb	83,287	11	7,571554		
26	Fachbuch	11,8	7,6	12,7871		Korrigiert gesamt	171,696	13			

Abbildung B.2: Das Arbeitsblatt für ANCOVA, Methode 2, für die Daten aus Tabelle B.1

Der Schlüssel besteht darin, auf irgendeine Art die Gruppe darzustellen, zu der die einzelnen Individuen gehören. Hierum kümmern sich *Kennung1* und *Kennung2* in den Spalten C und D. Wenn ein Kind zur Gruppe Lehrer gehört, dann ist Kennung1=1 und

Kennung2=0. Gehört es zur Gruppe Computer, dann ist Kennung1=0 und Kennung2=1. Gehört es zur Gruppe Fachbuch, dann ist Kennung1=0 und Kennung2=0. Allgemein ausgedrückt können Sie bei *k* Gruppen mit *k*–1 Spalten auf diese Weise die Gruppenmitgliedschaft angeben. Nur damit Sie es wissen: In der Welt der Statistik werden die Spalten B, C und D Vektoren genannt (ich gehe hier aber nicht weiter darauf ein).

Dieses Kategorisierungsschema funktioniert dann, wenn sich in jeder Gruppe die gleiche Anzahl Individuen befindet. Sollte dies nicht der Fall sein, werden die Dinge ein wenig knifflig.

Ich habe zwischen den Gruppen Linien eingefügt, diese dienen jedoch lediglich der Lesbarkeit.

Ich habe hier also eine multiple Regression eingerichtet, und zwar mit einer abhängigen Variablen (Note) und drei unabhängigen Variablen (mathematische Begabung, Kennung1 und Kennung2). Ich sage damit, dass die Note von der mathematischen Begabung, von Kennung1 und von Kennung2 abhängt.

Unterhalb der Datentabelle befindet sich in Zelle B18 die Quadratsumme $SS_{GesamtX}$. Die Formel in dieser Zelle lautet:

```
=ANZAHL(B2:B16)*VAR.P(B2:B16)
```

Ich habe also die Anzahl der Werte in Spalte B (15) mit der Varianz in Spalte B multipliziert (wobei ich diese als Population mit 15 Werten behandele und daher VAR.P verwende). So erhalte ich $SS_{GesamtX}$, den Zähler einer Varianz.

In Zelle B20 habe ich b_{Gesamt} berechnet:

```
=STEIGUNG(A2:A16;B2:B16)
```

Was ist mit $b_{Innerhalb}$? Bitte einen Moment Geduld.

Unterhalb dieser Werte befindet sich eine Tabelle der Mittelwerte und der korrigierten Mittelwerte, auf die ich später noch zu sprechen

komme.

Sie glauben es kaum, aber die Analyse ist fast fertig.

Alles, was noch zu tun bleibt, ist, das Analyse-Werkzeug Regression auszuführen und dessen Ergebnisse zu verwenden, um die ANCOVA-Tabelle zu vervollständigen.

In Kapitel 14 beschreibe ich, wie Sie das Analyse-Werkzeug Regression verwenden. Der Unterschied hier besteht darin (wie auch bei ANOVA in Methode 1), dass ich die Ergebnisse in das aktuelle und nicht in ein neues Tabellenblatt ausgeben lasse. Für unser Beispiel wird im Dialogfeld REGRESSION in das Feld Y-EINGABEBEREICH A1:A16 eingegeben. Da es sich um eine multiple Regression handelt, ist der Zellbereich für X-EINGABEBEREICH B1:D16.

Die ANCOVA-Tabelle befindet sich unterhalb der Ausgabe des Tools Regression. Wie im vorigen Beispiel habe ich die gesamte Tabelle samt Linien, Beschriftungen, Zeilen, Spalten und so weiter erstellt und formatiert. Anschließend habe ich die Freiheitsgrade *(df)* eingegeben.

Dann mache ich wie bei Methode 1 mit der korrigierten Quadratsumme SS_{Gesamt} weiter. Das ist diese Formel:

$$Korrigierte\ SS_{GesamtY} = SS_{GesamtY} - b^2_{Gesamt} SS_{GesamtX}$$
$$Korrigierte\ SS_{GesamtY} = 226,93 - (1,03)^2\ (52,4) = 171,70$$

Diesmal gebe ich in Zelle G26 die folgende Formel ein:

```
=H14-(B20^2*B18)
```

Und jetzt? Die Werte für die Zeile *Korrigiert innerhalb* befinden sich in der ANOVA-Tabelle für die Regressionsausgabe. Sie befinden sich dort in der Zeile, die mit *Residue* beschriftet ist. Die korrigierte Quadratsumme $SS_{Innerhalb}$ entspricht $SS_{Residue}$ und die korrigierte mittlere Quadratsumme $MS_{Innerhalb}$ entspricht $MS_{Residue}$.

Das bedeutet, dass

```
=H13
```

in Zelle G25 eingegeben wird und

```
=I13
```

in Zelle I25.

Die korrigierte Quadratsumme $SS_{Zwischen}$ wird in Zelle G24 folgendermaßen berechnet:

```
=G26-G25
```

Ich habe die korrigierten Quadratsummen durch die Freiheitsgrade dividiert, um die korrigieren mittleren Quadratsummen zu berechnen. Dann habe ich die korrigierte mittlere Quadratsumme $MS_{Zwischen}$ durch die korrigierte mittlere Quadratsumme $MS_{Innerhalb}$ geteilt, um *F* zu erhalten. Der Eintrag ganz rechts, *F krit*, wird wie in Methode 1 mit `F.VERT.RE` bestimmt.

Habe ich $b_{Innerhalb}$ vergessen? Nein. Er befindet sich in der Ausgabe des Tools Regression in Zelle G18 bei den Koeffizienten. Es ist der Regressionskoeffizient für die mathematische Begabung. Ich habe diesen Wert verwendet, um in der Tabelle mit den Mittelwerten die korrigierten Mittelwerte zu berechnen. Nachdem ich mit `MITTELWERT` die Mittelwerte für die abhängige Variable und die Kovariate berechnet habe, habe ich diese Formel in Zelle D24 eingegeben:

```
=B24-$G$18*(C24-MITTELWERT($B$2:$B$16))
```

und anschließend die Zellen D25 und D26 automatisch ausgefüllt.

Am Anfang dieses Abschnitts haben Sie sich vielleicht gefragt, warum ich Sie mit einem weiteren Ansatz behelligen will. Nachdem Sie diesen Abschnitt gelesen haben, fragen Sie sich vielleicht, warum ich Ihnen überhaupt den ersten Ansatz vorgestellt habe.

Nach der ANCOVA

Wie ich auch in Kapitel 12 erwähne, deutet ein kritischer F-Wert darauf hin, dass irgendwo in den Daten ein Effekt verborgen ist. Es ist jedoch immer noch erforderlich, hineinzuzoomen, um diese Stelle zu finden.

Postanalytisches Tests gibt es in zwei Varianten: einer, die Sie vorher planen, und einer, die nicht geplant ist. Die erste Variante, *geplante Vergleiche*, sind in den Vorstellungen darüber begründet, was Sie erwarten, bevor Sie die Daten sammeln. Die zweite Variante, Post-Hoc-Tests, entsteht durch das, was Sie in den gesammelten Daten interessant finden.

Bei ANOVA führen Sie diese Tests anhand der Gruppenmittelwerte durch. Bei ANCOVA korrigieren die Sie Gruppenmittelwerte (der abhängigen Variablen), so wie Sie es auch für die Quadratsummen (SS) und mit den mittleren Quadratsummen (MS) machen. Außerdem passen Sie den Fehlerterm (den Nenner) der Tests an.

Hier habe ich es mit geplanten Vergleichen zu tun. Nach der Anpassung der Mittelwerte (die in beiden Methoden vorgenommen wird), besteht der nächste Schritt darin, die mittlere Quadratsumme MS anzupassen, die bei den geplanten Vergleichen als Nenner verwendet wird. Ich nenne die korrigierte mittlere Quadratsumme MS'_{Fehler}. Sie wird folgendermaßen berechnet:

$$MS'_{Fehler} = Korrigierte\ MS_{Innerhalb}\left(1+\frac{SS_{ZwischenX}\,(k-1)}{SS_{InnerhhalbX}}\right)$$

k entspricht der Anzahl der Gruppen.

In unserem Beispiel ergibt dies:

$$MS'_{Fehler} = 7{,}57\left(1+\frac{2{,}8(2)}{49{,}6}\right) = 8{,}43$$

Im Arbeitsblatt der Methode 1 wähle ich eine Zelle aus und gebe dort diese Formel ein:

```
=D24*(1+(I24*2)/I25)
```

Ich kann nun, wie in Kapitel 12 gezeigt, mit den geplanten Vergleichen weitermachen.

Nach allem, was ich Ihnen hier gezeigt habe, sieht es so aus, als ob Methode 1 gegenüber Methode 2 Vorteile besitzt. Wenn Sie das

Analysewerkzeug ANOVA verwenden, erhalten Sie alle Werte, die Sie für die geplanten Vergleiche benötigen, und Sie erhalten die ANOVA der abhängigen Variablen, die Sie mit ANCOVA vergleichen können.

Mit den Daten, die so wie bei Methode 2 angeordnet sind, erhalten Sie all diese Informationen nicht, oder?

Doch, Sie bekommen sie. Und damit ist die Bühne dafür bereitet, eine bestimmte statistische Analyse (ANOVA) aus einem neuen Blickwinkel zu betrachten, so wie ich es am Anfang der Beschreibung von Methode 2 bereits erwähnte.

Lesen Sie bitte weiter.

Eins hab ich noch

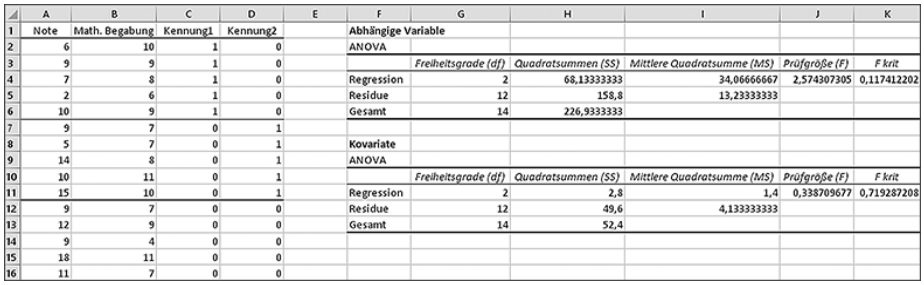
In mehreren Kapiteln dieses Buches können Sie das Wechselspiel und Zusammenspiel von ANOVA und Regression betrachten. Nach jeder Regressionsanalyse wird ANOVA verwendet, um Hypothesen über Regressionskonzepte zu testen.

In diesem Abschnitt drehe ich die Dinge um. Ich verwende eine Aufgabe, die normalerweise mit ANOVA gelöst wird, mache daraus ein Regressionsproblem und verwende dann die Regression, um die Varianzanalyse durchzuführen. Dieser Ansatz wird multiple Regression/Korrelation oder abgekürzt MRC (multiple regression/correlation) genannt. Dieser Ansatz basiert auf der Arbeit des Psychologen und Statistikers Jacob Cohen, der in den späten 1960er Jahren das Konzept der multiplen Regression als grundlegendes und allgemeines System für die Datenanalyse formulierte.

Ohne hier auf alle Details einzugehen, nur so viel: Cohens Konzept ist, dass viele Daten sich in einem Format darstellen lassen können, auf die die Regression angewendet werden kann. Die Regressionsanalyse ist dann alles, was erforderlich ist. Hierdurch werden dann verschiedene andere statistische Techniken zu Sonderfällen des allgemeinen Systems.

Die abhängige Variable in Tabelle B.1 bietet die Möglichkeit, diesen Ansatz zu verdeutlichen. Durch die Gruppen, die unter *Kennung1* und *Kennung2* definiert sind, und wenn die Kennungen als unabhängige Variable betrachtet werden, ist ANOVA das Gleiche wie die multiple Regression.

Vielleicht sind Sie mir schon einen Schritt voraus. Abbildung B.3 zeigt die Daten, die so aufgebaut sind wie in Abbildung B.2, jedoch mit anderen Analysen. Abbildung B.3:



	A	B	C	D	E	F	G	H	I	J	K
1	Note	Math. Begabung	Kennung1	Kennung2		Abhängige Variable					
2	6	10	1	0		ANOVA					
3	9	9	1	0			*Freiheitsgrade (df)*	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
4	7	8	1	0		Regression	2	68,13333333	34,06666667	2,574307305	0,117412202
5	2	6	1	0		Residue	12	158,8	13,23333333		
6	10	9	1	0		Gesamt	14	226,9333333			
7	9	7	0	1							
8	5	7	0	1		Kovariate					
9	14	8	0	1		ANOVA					
10	10	11	0	1			*Freiheitsgrade (df)*	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
11	15	10	0	1		Regression	2	2,8	1,4	0,338709677	0,719287208
12	9	7	0	0		Residue	12	49,6	4,133333333		
13	12	9	0	0		Gesamt	14	52,4			
14	9	4	0	0							
15	18	11	0	0							
16	11	7	0	0							

Abbildung B.3: Das Arbeitsblatt ANCOVA, Methode 2, mit zwei verschiedenen Regressionsanalysen

Ich habe das Analysewerkzeug Regression zweimal verwendet und die Ausgabe beide Male an dieses Arbeitsblatt geschickt. Von der Ausgabe habe ich bis auf die ANOVA-Tabelle alles gelöscht.

Beim ersten Mal habe ich dem Tool Regression als Y-EINGABEBEREICH den Bereich A2:A16 übergeben (Note) und als X-EINGABEBEREICH C2:D16 (Kennung1 und Kennung2). Das Ergebnis ist die ANOVA-Tabelle oben rechts, direkt unter der Überschrift *Abhängige Variable*.

Beim zweiten Mal war der Y-EINGABEBEREICH B2:B16 (mathematische Begabung) und der X-EINGABEBEREICH wiederum C2:D16 (Kennung1 und Kennung2). Das Ergebnis ist die ANOVA-Tabelle rechts unten, direkt unter der Überschrift *Kovariate*.

Kommen Ihnen die Zahlen bekannt vor? Sie sollten es. Sie sind mit denen der beiden ANOVAs identisch, die ich bei Methode 1 verwende. Der einzige Unterschied sind jedoch die Namen in der Spalte *Streuungsursache*. Anstelle von *Zwischen* und *Innerhalb* wird hier *Regression* beziehungsweise *Residue* verwendet.

Um die postanalytischen Tests für ANCOVA Methode 2 vorzunehmen, benötigen Sie lediglich ANOVA für die Kovariate. Ich habe beide gezeigt, um den Ansatz der multiplen Regression/Korrelation zu verdeutlichen.

Sie können den MRC-Ansatz auch auf komplexere Hypothesentests erweitern, etwa für die, die ich in Kapitel 13 behandle. In diesem Fall werden die Dinge etwas komplizierter: Sie müssen dann mehr über multiple Regression wissen und darüber, wie sie auf Dinge wie die Wechselwirkung angewendet wird. Auf jeden Fall ist es lohnenswert, sich intensiver mit dem MRC-Ansatz auseinanderzusetzen.

Indexverzeichnis

A

A-posteriori-Test

A-priori-Test

Abfrage

Abhängige Stichproben

- Datenanalysetool t-Test
- T.TEST, Funktion

Abhängige Variable 1 2

- Diagramme

Absolute Abweichungen

Absolute Zellbezüge

- in relative umwandeln

Absoluter Nullpunkt

Absolutwerte

Abweichung

absolute

Grundgesamtheit 1 2 3 4 5 6

MITTELABW, Funktion

mittlere

mittlere absolute

quadratische

Stichprobe 1 2 3 4

Summe der quadratischen Abweichungen vom Mittelwert

SUMQUADABW, Funktion

Achse formatieren, Aufgabenbereich

Achsen

verstehen

Achsenabschnitt 1 2

ACHSENABSCHNITT, Funktion

Achsenbeschriftungen, Dialogfeld

Achsentitel einfügen

Add-Ins, Dialogfeld

Adjustiertes Bestimmtheitsmaß

Aktienportfolio, Kursdiagramm

Alpha

- hohen Wert ansetzen

Alpha (α)

- Prognose
- Zwei-Stichproben-Hypothesentest

Alphafehler

Alternativhypothese

- Ein-Stichproben-Hypothesentest
- Zwei-Stichproben-Hypothesentest

Am besten angepasste Gerade 1 2

Analyse-Funktionen

- aktivieren
- beschreibende Statistik
- Populationskenngrößen
- Übersicht

ANCOVA

- F-Test
- Friedman zweiseitige Varianzanalyse
- gemischtes Design
- Überblick
- zweifaktorielle Varianzanalyse ohne Messwiederholung
- Interaktion
- Kruska-Wallis, einseitiger ANOVA-Test
- Einfaktoriell

Einfaktorielle Varianzanalyse:für ANCOVA

Zweifaktoriell mit Messwiederholung

angepasste Gruppenmittelwerte

in Excel

Kovarianz

postanalytische Tests

Zweifaktoriell ohne Messwiederholung

QUADRATESUMME, Funktion

Spaltenfaktor

SUMMENPRODUKT, Funktion

Trendanalyse 1 2

wiederholte Messungen

Zweifaktorielle Varianzanalyse mit Messwiederholung, Datenanalysetool

Anpassen der Statusleiste

Anpassung

testen

Anteil

kumulierter

Anzahl addierter Zahlen (N)

ANZAHL, Funktion 1 2

Arbeitsmappen, Beispiele herunterladen

Argument

Arrayformeln

erstellen

Arrayfunktionen

HÄUFIGKEIT

MODUS.VIELF

RGP 1 2

TREND

VARIATION

Asymptotisch

Ausfüllkästchen

Ausprägung

Ausreißer

Automatisches Ausfüllen von Zellen

- Dialogfeld Reihe
- Popupmenü Füllbereich
- ziehen mit der Maus

B

Balkendiagramme

- Achsen
- Achsentitel einfügen
- Daten auswählen
- gruppierte Balken, Diagrammtyp

Bearbeitungsleiste

Bedingte Wahrscheinlichkeit 1 2

Beispieldateien herunterladen

Benutzerdefinierte Fehlerindikatoren, Dialogfeld

Berichtsfilter

Berra, Yogi

Beschreibende Statistik 1 2

Beschreibende Statistiken

- ANZAHL, Funktion
- ANZAHL2, Funktion
- ANZAHLLEERZELLEN, Funktion
- HÄUFIGKEIT, Funktion
- Häufigkeitsverteilung
- KURT, Funktion
- Kurtosis
- MAX, Funktion

MAXA, Funktion

MIN, Funktion

MINA, Funktion

Populationskenngrößen

Schiefe

SCHIEFE, Funktion

SCHIEFE.P, Funktion

ZÄHLENWENN, Funktion

ZÄHLENWENNS, Funktion

Bestimmtheitskoeffizient

Bestimmtheitsmaß 1 2 3 4 5

Bestimmtheitsmaß im Diagramm darstellen, Trendlinie

adjustierte

anpassen

multiples

Beta (β)

BETA.VERT, Funktion 1 2

Beta

Prognose

Betafehler

BETAINV (alte Funktion)

BETAVERT (alte Funktion)

Betaverteilung

BETA.INV, Funktion

BETA.VERT, Funktion

Betadichtefunktion 1 2

Beziehungen

direkte

inverse

Bimodaler Wert

BINOM.INV, Funktion

Binomialexperiment

Binomialverteilung 1 2 3 4

- BINOM.INV, Funktion
- BINOM.VERT, Funktion
- Hypothesen testen
- negative
- NEGBINOM.VERT, Funktion
- Pascalverteilung
- Überblick

BINOMVERT (alte Funktion)

Biquadratische Komponente

Blasendiagramme

Blöcke mit zufälliger Zuteilung

Bonds, Barry

Box and Whisker, Kastengrafik

Breitgipfliges Histogramm

C

C-Wert

Chi-Quadrat

Chi-Quadrat-Verteilung 1 2 3 4

CHIQU.INV, Funktion

CHIQU.INV.RE, Funktion

CHIQU.TEST, Funktion 1 2

CHIQU.VERT, Funktion

CHIQU.VERT.RE, Funktion 1 2

Diagramm erstellen

Formel

Unabhängigkeit testen

CHIINV (alte Funktion)

CHIQU-TEST, Funktion

CHIQU.INV, Funktion

CHIQU.INV.RE, Funktion

CHIQU.TEST, Funktion 1 2

CHIQU.VERT.RE, Funktion

- Chi-Quadrat-Verteilung
- Cochran-Q-Test
- Friedmann zweiseitige Varianzanalyse
- Kruska-Wallis, einseitiger ANOVA-Test

CHITEST (alte Funktion)

CHIVERT (alte Funktion)

Cochran-Q-Test

- Einsatzgebiet
- Formel
- Schritte für Verwendung

Cohen, Jacob

D

3D-Karten

Daten sortieren

- intervallskalierte
- nominale
- nominalskalierte
- ordinale
- ordinalskalierte
- verhältnisskalierte

Datenanalysetool

- Zweifaktorielle Varianzanalyse ohne Messwiederholung

Datenanalysetools

- Gleitender Durchschnitt
- Histogramm

- Korrelation
- Kovarianz
- Populationskenngrößen
- Regression
- Stichprobenziehung
- Übersicht
- verwenden
- Zufallszahlengenerierung
- Zweifaktorielle Varianzanalyse mit Messwiederholung
- Zweistichproben t-Test: Gleicher Varianzen
- Zweistichproben t-Test: Unterschiedliche Varianzen

Datenart

Datenbalken

- Schnellanalyse

Datenbank

- Abfrage
- Datenmaske
- Datensatz
- Feld

Datenbankfunktionen

- Berechnungen
- Datensätze abrufen
- Datensätze zählen
- DBANZAHL
- DBANZAHL2
- DBAUSZUG
- DBMIN
- DBSTDABW
- DBSTDABWN

DBSUMME

DBVARIANZ

DBVARIANZEN

Format

Kriterienbereich

Statistikfunktionen

Datenbeschriftungen formatieren, Aufgabenbereich

Datenmaske

Datenquelle auswählen, Dialogfeld 1 2

Datensatz

DBANZAHL, Funktion

DBANZAHL2, Funktion

DBAUSZUG, Funktion

DBMITTELWERT, Funktion

DBPRODUKT, Funktion

DBSTDABW, Funktion

DBSTDABWN, Funktion

DBSUMME, Funktion

DBVARIANZ, Funktion

DBVARIANZEN, Funktion

Definierte Namen

In Formel verwenden, Dropdownliste

Deskriptive Statistik 1 2

Determinationskoeffizient

df, Freiheitsgrade 1 2 3 4 5 6 7 8

Diagramm einfügen, Dialogfeld 1 2

Diagramm verschieben, Dialogfeld

Diagrammblatt

Diagramm

3D-Karten

abhängige Variable

Achse formatieren, Aufgabenbereich

Achsen

Achsen verstehen

Achsentitel einfügen

ANOVA gemischtes Design

Balkendiagramm

Betaverteilung

Blasendiagramm

Box and Whisker

Chi-Quadrat-Verteilungen

Datenbeschriftungen formatieren, Aufgabenbereich

Datenquelle auswählen, Dialogfeld

dynamisch

einfügen

empfohlene

Eröffnungs-Höchst-Tiefst-Schlusskurs

Fehlerindikatoren 1 2

Formatvorlagen

gestapelte Säulen

Histogramme

Kastengrafikdiagramm

kategorische Variable als quantitative Variable darstellen

kontextbezogene Registerkarten

Kreisdiagramme

Kursdiagramm

Legende formatieren, Aufgabenbereich

Legende verschieben

lineare Regression

Linie mit Datenpunkten

Linie mit Datenpunkten, Diagrammtyp

Liniendiagramme

Netzdiagramm

Oberflächendiagramm

Pareto-Diagramm

Poissonverteilung

Punktdiagramm 1 2

Punkte mit interpolierten Linien

Sparklines

Standardfehler des Mittelwerts

Standardnormalverteilung

statistische

Streudiagramme 1 2

Sunburst-Diagramm

t-Verteilung

Treemap

Trendlinie hinzufügen 1 2

unabhängige Variable

verschieben

Vorschau

y-Achse

Diagrammelemente, Schaltfläche

Diagrammformatvorlagen

Diagramme, gruppierte Balken, Diagrammtyp

Dichotome Variable

Dichtefunktion

Gamma 1 2

Differenz zwischen den Mittelwerten der Grundgesamtheit

Diskrete Zufallsvariable

$ (Dollarzeichen)

Doppeltes exponentielles Glätten

Dreifaches exponentielles Glätten

Durchschnitt siehe Mittelwert

gleitender

E

e, Konstante 1 2 3

Ein-Stichproben-Hypothesentest

CHIQU.INV, Funktion

CHIQU.INV.RE, Funktion

CHIQU.VERT.RE, Funktion

Überblick

visualisieren

Einfache Haupteffekte

Einfaktorielle Varianzanalyse

Einseitiger Hypothesentest 1 2 3

Elementarereignis

Epsilon, Fehler in der Beziehung

Ereignis

zusammengesetztes 1 2 3 4

Ergebniswahrscheinlichkeit

Erlangverteilung

Error, Trend, Seasonal

Erwartungswert 1 2

ETS (Error, Trend, Seasonal)

Prognose

Excel Add-Ins aktivieren

EXP, Funktion

Experiment

Binomialexperiment

Münze werfen 1 2

Experimentelle Kontrolle

Exponent 1 2

Exponentialverteilung

Formel

Exponentielles Glätten 1 2

doppeltes

dreifach

Glättungsparameter

EXPONVERT (alte Funktion)

Exzess

F

f(x), Wahrscheinlichkeitsdichte

F-Quotient

F-Test

F-Verteilung und F-Test

ANOVA 1 2

F.TEST, Funktion

F.VERT, Funktion

F.VERT.RE, Funktion

zusammen mit t verwenden

Zwei-Stichproben F-Test (Datenanalysetool)

F.VERT, Funktion 1 2

F.VERT.RE, Funktion

Zwei-Stichproben-Hypothesentest

Faktor 1 2 3

Fakultät 1 2

FAKULTÄT, Funktion 1 2

Fehler 1. Art 1 2 3

Fehler 2. Art 1 2

in der Grundgesamtheit

Fehlerindikatoren

benutzerdefinierte

formatieren

hinzufügen

Fehlerterm 1 2 3

Feld

FINV (alte Funktion)

Fisher, Sir Ronald 1 2 3

Fisher-z-Transformation

Fläche

kumulierte

Formel im Diagramm anzeigen, Trendlinien

Formel-Generator (Mac)

Formeln

absoluter Zellbezug

Bearbeitungsleiste

Gleichheitszeichen

relativer Zellbezug

Freiheitsgrade 1 2 3 4 5 6 7 8 9

Friedman, zweiseitige Varianzanalyse

Formel

Schritte

FTEST (alte Funktion)

Funktion einfügen

Dialogfeld

Funktionen

ACHSENABSCHNITT

ANZAHL 1 2

Argumente

Arrayfunktion

BESTIMMTHEITSMASS

BETA.VERT

BINOM.INV

BINOM.VERT.BEREICH

CHIQU.INV

CHIQU.INV.RE

CHIQU.TEST 1 2 3

CHIQU.VERT.RE 1 2

DBANZAHL

DBANZAHL2

DBAUSZUG

DBSTDABW

DBSTDABWN

DBSUMME

DBVARIANZ

DBVARIANZEN

EXP

F.TEST

F.VERT

F.VERT.RE 1 2

FAKULTÄT 1 2

FISHER

FISHERINV

G.TEST

GAMMA

GAMMA.VERT 1 2

GAMMALN

GAMMALN.GENAU

GEOMITTEL

GESTUTZTMITTEL

HARMITTEL

HÄUFIGKEIT 1 2 3 4

KGRÖSSTE

KKLEINSTE

KOMBINATIONEN

KOMBINATIONEN2

KONFIDENZ.T

KORREL

KOVARIANZ.P

KOVARIANZ.S

KURT

LOG

LOGNORM.INV

Matrixfunktion

MAX 1 2

MAXA

MEDIAN

MIN 1 2

MINA

MITTELABW

MITTELWERT 1 2 3

MITTELWERTA

MITTELWERTWENN

MITTELWERTWENNS

MODUS.EINF

MODUS.VIELF

NEGBINOM.VERT

NORM.INV

NORM.S.INV 1 2 3

NORM.S.VERT

NORM.VERT

PEARSON

PHI

PI

POISSON.VERT 1 2

POTENZ

PROGNOSE.ETS

PROGNOSE.ETS.KONFINT

PROGNOSE.ETS.SAISONALITÄT

PROGNOSE.ETS.STAT

PROGNOSE.LINEAR

QUADRATESUMME 1 2

QUANTIL.EXKL

QUANTIL.INKL

QUANTILSRANG.EXKL

RANG.GLEICH

RANG.MITTELW

RGP

RUNDEN 1 2

SCHIEFE

SCHIEFE.P

STABW.N

STABW.S 1 2 3

STABWA

STABWNA

STANDARDISIERUNG

statistische

STEIGUNG

STFEHLERYX

SUMME

SUMMENPRODUKT 1 2 3

SUMMEWENN

SUMMEWENNS

SUMQUADABW

T.INV.2S

T.TEST

T.VERT

T.VERT.2S

T.VERT.RE 1 2

Tabellenblattfunktionen

TREND 1 2

UND

VAR.P

VAR.S

VARIANZA

VARIANZENA

VARIATIONEN2

WAHRSCHBEREICH

WURZEL

ZÄHLENWENNS

zu Symbolleiste für den Schnellzugriff hinzufügen

ZUFALLSBEREICH

zugreifen auf 1 2

Funktionsargumente

FVERT (alte Funktion)

G

G.TEST, Funktion

GAMMA, Funktion

GAMMA.VERT, Funktion 1 2

Gamma

- Prognose

Gammafunktion

GAMMAINV (alte Funktion)

GAMMALN, Funktion

GAMMALN.GENAU, Funktion

GAMMAVERT (alte Funktion)

Gammaverteilung

- Wahrscheinlichkeitsdichtefunktion

GAUSS, Funktion

Gegenhypothese

Gemischtes Design ANOVA

- Überblick
- zweifaktorielle Varianzanalyse ohne Messwiederholung

GEOMITTEL, Funktion

Geplante Vergleiche

Gepoolter Schätzwert

Gerade

am besten angepasste Gerade 1 2 3

Geradengleichung

Gesamtmittelwert 1 2

Gesamtvarianz 1 2

Geschweifte Klammern ({ })

Gestapelte Säulen

GESTUTZTMITTEL, Funktion

Glätten

doppeltes exponentielles

dreifaches exponentielles

exponentielles

Glättungsparameter

Glättungsparameter

Gleiche Varianzen

Gleichheitszeichen =

Gleitender Durchschnitt 1 2

Gleitender Durchschnitt, Datenanalysetool

Glockenkurve

Grundgesamtheiten

- Standardabweichung 1 2
- Varianz 1 2 3
- z-Wert-Formel

Grundgesamtheitsstandardabweichung

Grundgesamtheitsvarianz

Gruppenmittelwerte

- ANCOVA

Gruppierte Balken, Diagrammtyp

GTEST (alte Funktion)

H

H0

nicht verwerfen

verwerfen

H1

HARMITTEL, Funktion

Häufigkeit

HÄUFIGKEIT, Funktion 1 2 3

Häufigkeitsverteilung 1 2

Histogramm (Diagrammtyp)

Haupteffekte, einfache

Histogramme 1 2

breitgipfliges

Containerbreite

Datenanalysetool

leptokurtisches

platykurtisches

Schiefe

schmalgipfliges

sortiert

Holt, Charles

Holt-Winters-Verfahren

Horizontale Achsenbeschriftungen (Rubrik)

Hypergeometrische Verteilung

HYPGEOM.VERT, Funktion

HYPGEOMVERT (alte Funktion)

Hypothesen

Alternativhypothese 1 2

Gegenhypothese

nicht verwerfen

- Nullhypothese
- Regression
- testen 1 2
- testen 1 2
- über Korrelationen testen
- verwerfen

Hypothesentest

Hypothesentest mit mehr als zwei Stichproben

- mittlere Quadratsumme
- Quadratsumme
- Schwierigkeiten
- Überblick

Hypothesentest, Regression

- F-Test
- Fehler in der Grundgesamtheit

mittlere Quadratsumme

Quadratsumme

Standardfehler der Steigung

Korrelationen

abhängige Stichproben

Binomialverteilung

Ein-Stichproben-Hypothesentest

einseitig

Zwei-Stichproben-Hypothesentest

zweiseitig

I

In Formel verwenden, Dropdownliste

Inferenz

Inferenzielle Statistik

Inferenzstatistik

Fehler 1. Art

Fehler 2. Art

Innergruppenvariable

Innersubjekt-Design

Interaktion

Intervalle

Häufigkeitsverteilung

Intervallskalierte Daten

IQ-Wert

Itemanalyse

J

James, Bill 1 2

K

k-größter Wert

k-kleinster Wert

Kastengrafikdiagramme

Median

Mittelwert

Whisker

Kenngrößen

Statusleiste

KGRÖSSTE, Funktion

KKLEINSTE, Funktion

Klassen, Häufigkeitsverteilung

Kombinationen

KOMBINATIONEN, Funktion

KOMBINATIONEN2, Funktion 1 2

Komponenten

biquadratische

kubische

lineare

quadratische

quintische

signifikant lineare

KONFIDENZ (alte Funktion)

Konstante

Kontrast

Kontrolle

statistische

Konzepte aus Statistikbüchern in Excel übertragen

KORREL, Funktion

Spearmans Rangkorrelationskoeffizient

Korrelation 1 2

Korrelation, Datenanalysetool

partielle Korrelation

Schritte

semipartielle Korrelation

tabellarische Ausgabe

Teilkorrelation

Bestimmtheitskoeffizient

Bestimmtheitsmaß

BESTIMMTHEITSMASS, Funktion

FISHER, Funktion

FISHERINV, Funktion

KORREL, Funktion

Korrelationskoeffizient

Kovarianz 1 2

negative

PEARSON, Funktion

positive

punktbiseriale

Residuen

und Regression

Korrelationskoeffizient 1 2 3

Korrelationskoeffizient nach Pearson

multipler

Korrelationsmatrix

KOVAR (alte Funktion)

KOVARIANZ.P, Funktion

KOVARIANZ.S, Funktion

Kovarianz 1 2 3

KOVARIANZ.P, Funktion

KOVARIANZ.S, Funktion

Kovarianzanalyse

Gründe

in Excel 1 2

Kovarianz

postanalytische Tests

Kovarianzmatrix

Kovariate

statistische Kontrolle

Kreisdiagramme

bearbeiten

Daten auswählen

Datenbeschriftungen

Kreuztabelle

KRITBINOM (alte Funktion)

Kriterienbereich

Kritischer Wert

Krummlinige Regression

Kruska-Wallis, einseitiger ANOVA-Test

Kubische Komponente

Kumulierte Fläche

Kumulierte Häufigkeit

Kumulierte Wahrscheinlichkeit

Kumulierter Anteil

Kursdiagramme

- Daten auswählen
- Eröffnungs-Höchst-Tiefst-Schlusskurs

KURT, Funktion

Kurtosis 1 2

- Formel
- Histogramme
- Tool Populationskenngrößen

Kurvenanpassung

L

Lagemaße

Legende

- formatieren, Aufgabenbereich
- verschieben

Leptokurtisches Histogramm

Lineare Komponente

Lineare Regression 1 2

Lineare Trendlinie

Linearer Trend

Linie mit Datenpunkten, Diagrammtyp

Liniendiagramm

- Standardfehler des Mittelwerts

Liniendiagramme

- Achsentitel einfügen

Daten auswählen 1 2

erstellen

Farben ändern

horizontale Achsenbeschriftungen (Rubrik)

Linie mit Datenpunkten

Linie mit Datenpunkten, Diagrammtyp

Linksschiefe Verteilung

Linkssteile Verteilung

ln, natürlicher Logarithmus 1 2

LOG, Funktion

Logarithmen

natürlicher Logarithmus

Logarithmus

Basis

Definition

e, Konstante

GAMMLN.GENAU, Funktion

LOGNORM.INV, Funktion

LOGNORM.VERT, Funktion

natürlicher 1 2

RXP, Funktion

Zehnerlogarithmus

LOGINV (alte Funktion)

Logische Funktionen

UND

WENN

LOGNORM.INV, Funktion

LOGNORM.VERT, Funktion

Lognormalverteilung

LOGNORMVERT (alte Funktion)

M

MAE (Mean Absolute Error)

Mann-Whitney-U-Test

Matrixfunktion 1 2

MAX, Funktion

Mean Absolute Error

Mean Square (MS) siehe Mittlere Quadratsumme

Median

- Kastengrafikdiagramm
- Tool Populationskenngrößen

Messwiederholung 1 2

MIN, Funktion

Minimum

- MINA, Funktion

MITTELABW, Funktion

Mitteln ohne Ausreißer

Mittelwert 1 2

MITTELWERT, Funktion 1 2 3

Simulationen

Mittelwert

der Stichprobenverteilung des Mittelwerts

aus der Grundgesamtheit

berechnen

der Stichprobenverteilung des Mittelwerts

GEOMITTEL, Funktion

Gesamt- 1 2

GESTUTZTMITTEL, Funktion

HARMITTEL, Funktion

Kastengrafikdiagramm

MITTELWERT, Funktion

MITTELWERTA, Funktion

MITTELWERTWENNS, Funktion

Schnellanalyse

Standardfehler 1 2

Standardfehler des Mittelwerts in Diagramm darstellen

Standardnormalverteilung

Statusleiste

Stichprobenverteilung

Summe der quadratischen Abweichungen

Trend

vergleichen

von quadratischen Abweichungen

MITTELWERTA, Funktion

Zellen mit Text

MITTELWERTWENN, Funktion

MITTELWERTWENNS, Funktion

Mittlere absolute Abweichung

Mittlere Abweichung 1 2

Mittlere Quadratsumme

- Definition
- Gesamtvarianz
- innerhalb der Gruppen
- Regression
- Varianz der Stichprobenmittelwerte

Mittlerer absoluter Fehler

Modalwert

- ermitteln
- MODUS.EINF, Funktion
- MODUS.VIELF, Funktion

Modelle

prüfen

Modellierung

CHIQU.TEST, Funktion

Monte-Carlo-Methode

POISSON.VERT, Funktion

Poissonprozess

visualisieren

zentraler Grenzwertsatz

MODUS (alte Funktion)

MODUS.EINF, Funktion

MODUS.VIELF, Funktion

Modus

Tool Populationskenngrößen

Monte-Carlo-Simulation

MRC. Multiple Regression/Correlation

MS, Mean Square

MSFehler

MSG

MSGesamt

MSI

MSInnerhalb

MSZ

MSZwischen

Multinominalverteilung 1 2

Multiple Korrelation

Multiple Regression

Multiple Regression/Korrelation

- am besten angepasste Gerade
- RGP, Funktion
- TREND, Funktion

Multiples Bestimmtheitsmaß

Münze werfen

- Wahrscheinlichkeit

N

N

Namen definieren

- Namen in Funktionen verwenden
- Namens-Manager
- Regeln
- Schritte
- und Beschriftungen
- Vorteile

Namenfeld

Natürlicher Logarithmus 1 2 3

Negative Binomialverteilung

Negative Korrelation

Negative Steigung

NEGBINOM.VERT, Funktion

NEGBINOMVERT (alte Funktion)

Netzdiagramme

Neuer Name, Dialogfeld

Neues Tabellenblatt

Nominale Daten

Nominalskalierte Daten

NORM.INV, Funktion

NORM.S.INV, Funktion 1 2 3

NORM.S.VERT, Funktion

NORM.VERT, Funktion

Mann-Whitney-U-Test

Wilcoxon-Vorzeichen-Rang-Test

Normalverteilung

GAUSS, Funktion 1 2

Glockenkurve

NORM.INV, Funktion

NORM.S.INV, Funktion

NORM.S.VERT, Funktion

NORM.VERT, Funktion

Parameter

PHI, Funktion

Überblick

z-Werte

NORMINV (alte Funktion)

NORMSINV (alte Funktion)

NORMSVERT (alte Funktion)

NORMVERT (alte Funktion)

Nullhypothese

Ein-Stichproben-Hypothesentest

nicht verwerfen

Zwei-Stichproben-Hypothesentest

Nullpunkt

O

Obere Vertrauensgrenze

Oberflächendiagramme

Ordinalskalierte Daten

P

Parameter

Parameterfreie statistische Tests

Korrelation

Parallelstichproben 1 2 3

unabhängige Stichproben 1 2

Pareto

Pareto-Diagramm

Partielle Korrelation

Pascalverteilung

- Formel
- NEGBINOM.VERT, Funktion

PEARSON, Funktion

Pearson, Karl 1 2

Pearson, Korrelationskoeffizient

Permutation

PHI, Funktion

PI, Funktion

Pi, mathematische Konstante

PivotTable

PivotTable erstellen, Dialogfeld

PivotTable-Bericht

- Berichtsfilter
- einfügen
- Felder
- Layout

Platykurtisches Histogramm

Platzhalter

POISSON (alte Funktion)

Poisson, Siméon-Denis

POISSON.VERT, Funktion 1 2

Poissonprozess 1 2

Poissonverteilung

- Modell erstellen
- POISSON.VERT, Funktion

Pooling der Varianzen

- mehr als zwei Stichproben
- Zwei-Stichproben-Hypothesentest

Population

Populationskenngrößen 1 2

- Datenanalysetool

Positive Korrelation

Positive Steigung

Post-hoc-Test 1 2

Postanalytische Tests

Potenz

POTENZ, Funktion

Power View

Produktmoment-Korrelationskoeffizient nach Pearson

Produktregel 1 2

PROGNOSE.ETS, Funktion

PROGNOSE.ETS.KONFINT, Funktion

PROGNOSE.ETS.SAISONALITÄT, Funktion

PROGNOSE.ETS.STAT, Funktion

Prognose

- mittlerer absoluter Fehler
- PROGNOSE.ETS, Funktion
- PROGNOSE.ETS.KONFINT, Funktion
- PROGNOSE.ETS.SAISONALITÄT, Funktion
- PROGNOSE.ETS.STAT, Funktion
- Prognosearbeitsblatt
- Trendlinie
- Wurzel der mittleren Fehlerquadratsumme

Prognosearbeitsblatt erstellen, Dialogfeld

- Prognosestatistiken
- Saisonalität

Prüfungsergebnisse

Punktbiseriale Korrelation

Punktdiagramme 1 2

- Achsentitel einfügen
- Punkte mit geraden Linien und Datenpunkten
- Punkte mit interpolierten Linien
- Trendlinie hinzufügen
- Variabilität

Punkte mit interpolierten Linien, Diagrammtyp

Pythagoreischer Prozentsatz

Q

QUADRATESUMME, Funktion 1 2

- Cochran-Q-Test
- Friedmann zweiseitige Varianzanalyse
- Kruska-Wallis, einseitiger ANOVA-Test

Quadratische Abweichung

Quadratische Komponente

Quadratsumme

- mittlere
- Regression

Quadratsummenzerlegung

QUANTIL (alte Funktion)

QUANTIL.EXKL, Funktion

QUANTIL.INKL, Funktion

Quantile 1 2

- Kastengrafikdiagramm
- QUANTIL.EXKL, Funktion
- QUANTIL.INKL, Funktion

QUANTILRANG (alte Funktion)

Quantilsplot 1 2

QUARTILE (alte Funktion)

Quintische Komponente

R

r2, Bestimmtheitskoeffizient

RANG (alte Funktion)

Rang und Quantil

Datenanalysetool

RANG.GLEICH, Funktion

Spearmans Rangkorrelationskoeffizient

RANG.MITTELW, Funktion

RANG.MITTELWERT, Funktion

Friedman, zweiseitige Varianzanalyse

Ränge mit Bindungen

KGRÖSSTE, Funktion

KKLEINSTE, Funktion

QUANTILSRANG.EXKL, Funktion

QUANTILSRANG.INKL, Funktion

RAND.MITTELW, Funktion

Rang und Quantil, Datenanalysetool

RANG.GLEICH, Funktion

Rangkorrelationskoeffizient

parameterfreier Test

Rechtsschiefe Verteilung

Rechtssteile Verteilung

Regression 1 2 3

Regression, Datenanalysetool

Regression, lineare

Regression, multiple

am besten angepasste Gerade

Excel-Tools

RGP, Funktion

TREND, Funktion

Achsenabschnitt

Achsenabschnitt testen

ACHSENABSCHNITT, Funktion

am besten angepasste Gerade

ANCOVA

Anpassung

Datenanalysetool 1 2

F-Test

Fehler in der Beziehung

Geradengleichung

Gesamtvarianz

krummlinige

lineare

mittlere Quadratsumme

Quadratsumme

Regressionsgerade 1 2

Regressionsgleichung

Regressionskoeffizient

Residue

Restvarianz

RGP, Funktion

Schätzwert des Standardfehlers des Achsenabschnitts

Standardfehler der Steigung

Steigung

Steigung testen

STFEHLERYX, Funktion

Streudiagramme

Tabellenblattfunktionen

TREND, Funktion

und Korrelation

Regressionsgerade 1 2 3

Regressionsgleichung

Regressionskoeffizient

Regressionslinie

Relative Position eines Werts

Relative Zellbezüge

in absolute umwandeln

Relativer Zellbezug

Replikationen

Residuen 1 2 3 4 5 6

standardisierte

Residuenplot

Restvarianz

Reziproker Wert

RGP

RGP, Funktion

lineare Regression

multiple Regression

Ringsegmente, Sunburst-Diagramm

RKP, Arrayfunktion

RMSE (Root Mean Squared Error)

rS, Spearmans Rangkorrelationskoeffizient

RUNDEN, Funktion 1 2

Ruth, Babe

S

SABR, Society for American Baseball Research

Sabremetrie

Saisonkomponente einer Zeitreihe

SAT-Wert

Säulendiagramme

- Achse formatieren, Aufgabenbereich
- Achsen
- Achsentitel einfügen
- bearbeiten
- Daten auswählen
- erstellen
- Farbe der Achse ändern
- gestapelte Säulen
- Legende verschieben
- Schnellanalyse

SCHÄTZER (alte Funktion)

Schätzung

- KONFIDENZ.NORM, Funktion

KONFIDENZ.T, Funktion

Konfidenzintervall

Stichprobenverteilung des Mittelwertshäufig

t-Verteilung

zentraler Grenzwertsatz

Schätzwert

gepoolter

Scheffé, Henry

Scheffé-Test

Schiefe

SCHIEFE, Funktion

SCHIEFE.P, Funktion

Schiefe

Tool Populationskenngrößen

Schmalgipfliges Histogramm

Schnellanalyse

- Datenbalken
- Mittelwert
- Säulendiagramm
- Sparklines
- Tabelle
- Überblick

Schnitt 1 2 3

Semipartielle Korrelation

Sigma 1 2

Signifikant lineare Komponente

Signifikanz 1 2 3

Simulationen

- Monte-Carlo-Simulation
- Zufallszahlengenerierung

Society for American Baseball Research (SABR)

Solver aktivieren

Sortieren

- Daten
- Dialogfeld
- mehrere Ebenen verwenden

Sortiertes Histogramm

Spaltenfaktor

Spannweite

- Tool Populationskenngrößen

Sparklines

Sparklines erstellen, Dialogfeld

- erstellen
- Gewinn/Verlust
- in Word verwenden

Linie

löschen

Säule

Schnellanalyse

Spearmans Rangkorrelationskoeffizient

Formel

Hypothesentest

SS (Sum of Squares) siehe Quadratsumme

STABW (alte Funktion)

STABW.N, Funktion

STABW.S, Funktion 1 2 3

STABWA, Funktion 1 2

STABWN

STABWNA, Funktion 1 2

STABWWENN, nicht vorhandene Funktion

STABWWENNS, nicht vorhandene Funktion

Standardabweichung (Stichprobe) 1 2 3 4

Formel

der Stichprobenverteilung

einer Grundgesamtheit

einer Stichprobe

Formel (Grundgesamtheit)

Normalverteilung 1 2

STABW.N, Funktion

STABW.S, Funktion

STABWA, Funktion

STABWNA, Funktion

STABWWENN, nicht vorhandene Funktion

STABWWENNS, nicht vorhandene Funktion

Standardfehler der Differenz zwischen Mittelwerten

Standardnormalverteilung

TABW.S, Funktion

Tool Populationskenngrößen

Standardfehler 1 2 3 4 5

Standardfehler des Mittelwerts

in Diagramm darstellen

der Differenz zwischen Mittelwerten

des Mittelwerts 1 2

Schätzwert

Tool Populationskenngrößen

Standardisieren

Standardisierte Residue

STANDARDISIERUNG, Funktion

Standardnormalverteilung 1 2 3 4

als Diagramm darstellen

Formel

GAUSS, Funktion

NORM.S.VERT, Funktion

NORM.VERT, Funktion

PHI, Funktion

z-Werte 1 2

Standardwerte

Formel

Stanford-Binet-Wert

Statistik

beschreibende 1 2

deskriptive 1 2

inferenzielle

Statistisch signifikant

Statistische Funktion

Statistische Kontrolle

Statusleiste

- anpassen
- Anzahl
- Mittelwert

Steigung

- negative
- positiv

Steigungen

Stetige Zufallsvariable

STFEHLERYX, Funktion

Stichprobe, Varianz

- Varianz berechnen

Stichproben

- abhängige 1 2

große

Parallel-

z-Wert-Formel

ziehen

Stichprobengröße

Stichprobenmittelwert

Stichprobenstandardabweichung

Stichprobenumfang

Stichprobenvarianz

Formel

Tool Populationskenngrößen

Stichprobenverteilung des Mittelwerts 1 2 3 4 5

t-Verteilung

der Differenz zwischen Mittelwerten

Stichprobenziehung 1 2

Streudiagramme 1 2 3

Achsentitel einfügen

Daten auswählen

extrapolieren

Trendlinie hinzufügen

Streuung

berechnen

um Regressionsgerade

Sum of Squares (SS) siehe Quadratsumme

Summe von

SUMME, Funktion

SUMMENPRODUKT

SUMMENPRODUKT, Funktion 1 2 3

SUMMEWENN, Funktion

Argumente

Wilcoxon-Vorzeichen-Rang-Test

SUMMEWENNS, Funktion

Argumente

mit definierten Namen

SUMQUADABW, Funktion

Cochran-Q-Test

Sunburst-Diagramm

T

t-Verteilung und t-Test

ANOVA 1 2 3

Diagramm

KONFIDENZ.T, Funktion

Spearmans Rangkorrelationskoeffizient

T.INV.2S, Funktion

T.VERT.2S, Funktion

T.VERT.RE, Funktion

Überblick

T-Wert 1 2

T.TEST, Funktion

abhängige Stichproben

T.VERT.2S, Funktion

Spearmans Rangkorrelationskoeffizient

T.VERT.RE, Funktion 1 2

Tabellen

Schnellanalyse

Tabellenblattfunktion 1 2

Korrelation

Teilkorrelation

Testen

Hypothesen 1 2

Varianz

zwei Varianzen

Testgröße

kritischer Wert

Teststatistik

Ties

TINV (alte Funktion)

Transponieren

Treemap

TREND, Funktion

Trend

in Balkendiagramm darstellen

linearer

Trendanalyse 1 2

Trendkomponente einer Zeitreihe

Trendlinie 1 2

Trendlinie formatieren, Aufgabenbereich

Bestimmtheitsmaß im Diagramm darstellen

formatieren, Aufgabenbereich

hinzufügen 1 2

TTEST (alte Funktion)

Tufte, Edward

TVERT 1 2

U

Unabhängige Experimente

Unabhängige Variable 1 2

Diagramme

Faktor

UND, Funktion 1 2

Ungeplante Vergleiche

unregelmäßige Komponente einer Zeitreihe

Untere Vertrauensgrenze

Unterschiedliche Varianzen

V

VAR.P, Funktion

VAR.S, Funktion

Simulationen

Variablen

abhängige 1 2 3

dichotome

unabhängige 1 2 3 4

x

y

Varianz 1 2 3 4 5

VARIANZ (alte Funktion)

Varianz, Grundgesamtheiten

VAR.P, Funktion

VARIANZENA, Funktion

Varianz, Stichprobe

VAR.S, Funktion

VARIANZA, Funktion

berechnen

Symbol

VARIANZA, Funktion

Varianzanalyse 1 2

Varianzanalyse, gemischtes Design

Ergebnisse visualisieren

Überblick

zweifaktorielle Varianzanalyse ohne Messwiederholung 1 2

angepasste Gruppenmittelwerte

einfaktorielle 1 2

geplante Vergleiche

Interaktion

Kruska-Wallis, einseitiger ANOVA-Test

nicht geplante Vergleiche

QUADRATESUMME, Funktion

Spaltenfaktor

SUMMENPRODUKT, Funktion

Trendanalyse

Zeilenfaktor

zweifaktorielle

Zweifaktorielle Varianzanalyse mit Messwiederholung, Datenanalysetool

Zweifaktorielle Varianzanalyse ohne Messwiederholung

VARIANZEN (alte Funktion)

Varianzen

zwei testen

Pooling 1 2 3

Zufallsvariable

VARIANZENA, Funktion

VARIATION, Arrayfunktion

Variationsberechnung siehe Abweichung und Standardabweichung

Vereinigung 1 2

Vergleiche

geplante

ungeplante

Vergleichskoeffizient

Verhältnisskalierte Daten

Versuch

Verteilung

Betaverteilung

Chi-Quadrat

Erlangverteilung

Exponentialverteilung

Gammaverteilung

hypergeometrische

linksschiefe

linkssteile

Lognormalverteilung

modellieren

Multinomialverteilung

rechtsschiefe

rechtssteile

Weibull-Verteilung

Vertrauensgrenzen 1 2

für Mittelwert

KONFIDENZ.NORM, Funktion

obere

untere

Viertelwerte

Von Paar zu Paar konstant gehalten

W

WAHRSCHBEREICH, Funktion

Wahrscheinlichkeit 1 2

bedingte

BETA.VERT, Funktion

Betaverteilung

Binomialverteilung 1 2 3 4 5 6

diskrete Zufallsvariable

Ergebniswahrscheinlichkeit

Erwartungswert

EXPON.VERT, Funktion

FAKULTÄT, Funktion

Formel

GAMMA.INV, Funktion

GAMMA.VERT, Funktion

große Wahrscheinlichkeitsräume

hypergeometrische Verteilung

Hypothesen testen

Kombinationen

KOMBINATIONEN, Funktion

KOMBINATIONEN2, Funktion

kumulierte

Pascalverteilung

Permutationen

POISSON.VERT, Funktion

Poissonverteilung

Produktregel

stetige Zufallsvariable

Tabellenblattfunktionen

unabhängige Experimente

VARIATIONEN, Funktion

VARIATIONEN2, Funktion

Verteilung

Wahrscheinlichkeitsdichte

Zählregel

Zufallsvariable

zusammengesetzte Ereignisse

Wahrscheinlichkeit;Wahrscheinlichkeitsräume

Wahrscheinlichkeitsdichte 1 2 3 4

Wahrscheinlichkeitsdichtefunktion 1 2 3

Gammaverteilung

Wahrscheinlichkeitsräume 1 2

große

Wahrscheinlichkeitsverteilung 1 2

WEIBULL (alte Funktion)

Weibull-Verteilung

WEIBULL.VERT, Funktion

WENN, Funktion

Wertebereich

Tool Populationskenngröße

Whisker, Kastengrafikdiagramm

Wiederholte Messungen

Wilcoxon-Vorzeichen-Rang-Test

Winters, Peter

Wölbung 1 2

Word

Sparklines einfügen

World RPS Society

Wortgrafiken

Würfel

zinken

Wurzel der mittleren Fehlerquadratsumme

WURZEL, Funktion

X

X

x-Achse 1 2

x-Variable

Y

y-Achse

y-Achsenabschnitt

y-Variablen

Z

Z-Scores

z-Test

z-Werte 1 2 3

Eigenschaften

Formel

Normalverteilung 1 2 3

ZÄHLENWENN, Funktion

Wilcoxon-Vorzeichen-Rang-Test

Zählregel 1 2

Zehnerlogarithmus 1 2

Zeilenfaktor

Zeitreihen

doppeltes exponentielles Glätten

dreifaches exponentielles Glätten

einfaches exponentielles Glätten

ETS (Error, Trend, Seasonal)

exponentielles Glätten

Glättungsparameter

gleitender Durchschnitt

Prognose 1 2 3 4

Saisonkomponente

Trendkomponente

unregelmäßige Komponente

zyklische Komponente

Zellbezug

absoluter

Zellbezüge, absolute

definierte Namen

relative

Zellen zählen

ANZAHL, Funktion

ANZAHL2, Funktion

ANZAHLLEEREZELLEN, Funktion

ZÄHLENWENN, Funktion

ZÄHLENWENNS, Funktion

Zentrale Tendenz

Ausreißer

geometrisches Mittel

GEOMITTEL, Funktion

GESTUTZTMITTEL, Funktion

HARMITTEL, Funktion

harmonisches Mittel

- Median
- Mittelwert
- MITTELWERT, Funktion
- MITTELWERTA, Funktion
- MITTELWERTWENN, Funktion
- MITTELWERTWENNS, Funktion
- Modalwert
- MODUS.EINF, Funktion
- MODUS.VIELF, Funktion

Zentraler Grenzwertsatz

- Ein-Stichproben-Hypothesentest
- Mittelwert der Stichprobenverteilung des Mittelwerts
- Simulation
- Standardfehler der Differenz zwischen den Mittelwerten

Standardfehler des Mittelwerts

Stichprobenverteilung des Mittelwerts

Überblick

Zwei-Stichproben-Hypothesentest

ZUFALLSBEREICH, Funktion

Zufallsfehler 1 2

Zufallsvariable, diskrete 1 2 3

Erwartungswert

hypergeometrisch verteilte

Standardabweichung

Varianz

Zufallsvariablen

stetige 1 2

Zufallszahlengenerierung 1 2

Zufallszahlengenerierung, Datenanalysetool

Simulationen

Zusammengesetzte Ereignisse

Schnitt

Vereinigung

Zwei-Stichproben-F-Test für Varianzen

Zwei-Stichproben-Hypothestentest

F.INV, Funktion

F.INV,RE, Funktion

F.TEST, Funktion

T.TEST, Funktion

Zwei-Stichproben-F-Test (Datenanalysetool)

Zweistichproben t-Test bei gleichen Varianzen

Zweistichproben t-Test bei unterschiedlichen Varianzen

Zweistichproben-Test bei bekannten Varianzen

α (Alpha)

F.VERT, Funktion

F.VERT.RE, Funktion

zentraler Grenzwertsatz

Zweistichproben t-Test bei abhängigen Stichproben

Zwei-Stichproben-t-Test

Zweifaktorielle Varianzanalyse

Zweifaktorielle Varianzanalyse mit Messwiederholung, Datenanalysetool

Zweifaktorielle Varianzanalyse ohne Messwiederholung, Datenanalysetool

mit Messwiederholung, Datenanalysetool

ohne Messwiederholung, Datenanalysetool

Spaltenfaktor

Wechselwirkungen

Zeilenfaktor

Zweiseitiger Hypothesentest 1 2 3 4

Zweistichproben t-Test bei abhängigen Stichproben (Datenanalysetool)

- Gleicher Varianzen

Zweistichproben-Test bei bekannten Varianzen (Datenanalysetool)

Zwischengruppenvariable

Zyklische Komponente einer Zeitreihe

WILEY END USER LICENSE AGREEMENT

Besuchen Sie www.wiley.com/go/eula, um Wiley’s E-Book-EULA einzusehen.