Echte Daten auswerten

1

In diesem Kapitel

- Einführung in Begriffe der Statistik
- Von Stichproben Rückschlüsse auf Grundgesamtheiten ziehen
- Wahrscheinlichkeiten kennen lernen
- Entscheidungen treffen
- Grundlegendes zu Excel

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?

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 verwandten) Vorstellungen, 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 Rechnen geht. 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 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.

Daneben erfahren Sie außerdem Grundlegendes zu Excel.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 Analysten 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 Wahlen 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 in den USA 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*.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 ist die Beziehung zwischen Grundgesamtheiten und Stichproben sowie zwischen Parametern und Statistiken dargestellt.

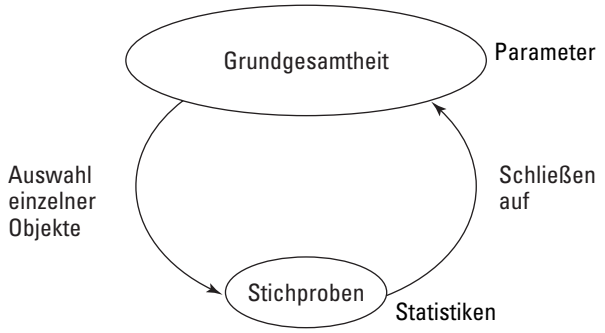


Abbildung 1.1: Die Beziehung zwischen Grundgesamtheiten, Stichproben, Parametern und Statistiken

Abhängige und unabhängige Variablen

Einfach ausgedrückt, ist eine *Variable* etwas, das mehrere Werte annehmen kann. (Etwas, das nur einen Wert annehmen kann, wird als *Konstante* bezeichnet.) Einige Variablen, die Sie bereits kennen gelernt haben, 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.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 immer 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 Mohs'sche 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 Mohs'schen 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 diesselbe 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. Das kommt daher, weil 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.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

Jede dieser Datenarten kann 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 gehörenden 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 1/6 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 1/4.

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

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

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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, d.h. 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 Untersuchung 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 mal. Die Münze kann symmetrisch sein, und Sie können dennoch 99 Mal Kopf und 1 Mal Zahl werfen, richtig? Absolut. Das weiß man nie so recht. Sie müssen Stichprobendaten sammeln (das Ergebnis aus 100 Würfen) und Rückschlüsse ziehen. Die Rückschlüsse können richtig sein oder auch nicht.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 nicht schuldig 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 wird. Dieser Ausgangspunkt wird als *Nullhypothese* bezeichnet. In der Statistik wird für die Nullhypothese H_0 geschrieben. 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 als *Alternativhypothese* bezeichnet. In der Statistik wird für die Alternativhypothese H_1 geschrieben.

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 oben 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.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

Ist Ihnen außerdem 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 kennen lernen.

Zwei Arten von Fehlern

Beim Auswerten der Daten aus einer Untersuchung und 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* und *Fehler 2. Art*.

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 der 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.

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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.

Einige Dinge über Excel, die Sie unbedingt wissen müssen

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 mit Excel-Formeln effizienter arbeiten.

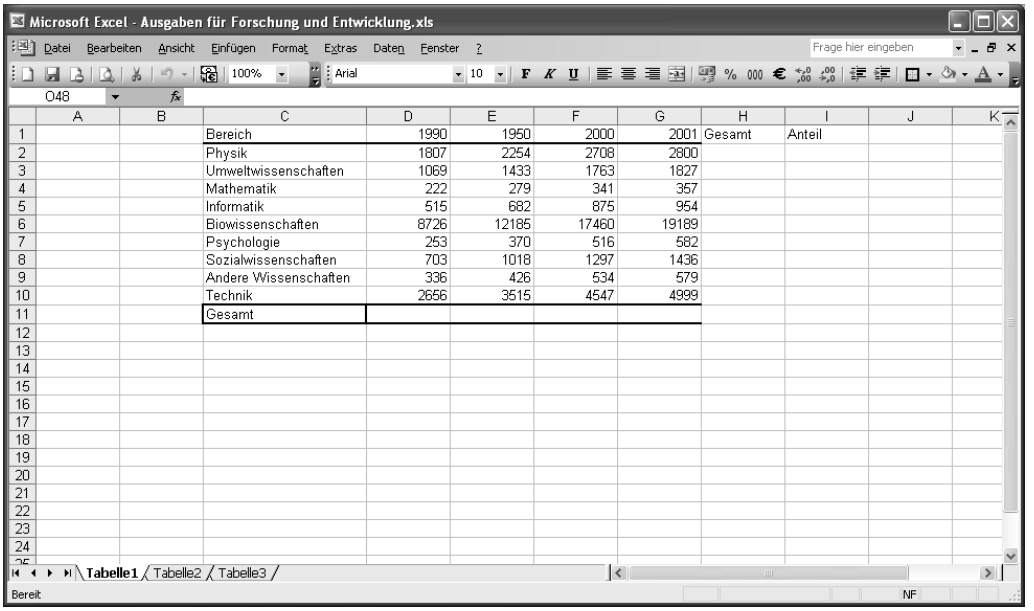
Automatisches Ausfüllen von Zellen

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 1.2 ist ein Arbeitsblatt mit den Ausgaben für Forschung und Entwicklung im Bereich Wissenschaft und Technik an Hochschulen und Universitäten für die angegebenen

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

Jahre dargestellt. Die Daten, die vom Bericht der U.S. National Science Foundation stammen, sind in Millionen US-Dollar angegeben. Ich habe rechts eine Spalte für den Gesamtbetrag aus den Bereichen und unten eine Zeile für den Gesamtbetrag aus den Jahren freigelassen.



	A	B	C	D	E	F	G	H	I	J	K
1			Bereich	1990	1950	2000	2001	Gesamt	Anteil		
2			Physik	1807	2254	2708	2800				
3			Umweltwissenschaften	1069	1433	1763	1827				
4			Mathematik	222	279	341	357				
5			Informatik	515	682	875	954				
6			Biowissenschaften	8726	12185	17460	19189				
7			Psychologie	253	370	516	582				
8			Sozialwissenschaften	703	1018	1297	1436				
9			Andere Wissenschaften	336	426	534	579				
10			Technik	2656	3515	4547	4999				
11			Gesamt								

Abbildung 1.2: Ausgaben für Forschung und Entwicklung im Bereich Wissenschaft und Technik

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 ↵-Taste drücken, wird in H2 der Gesamtbetrag angezeigt.

Um nun diese Formel in die Zellen H3 bis H10 einzufügen, zeigen Sie mit dem Cursor auf die untere rechte Ecke von H2, bis aus dem Cursor 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

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 auf das Menü BEARBEITEN zu klicken. Klicken Sie im Menü BEARBEITEN auf AUSFÜLLEN|REIHE. Daraufhin wird das Dialogfeld REIHE angezeigt (siehe Abbildung 1.3). Wenn Sie in diesem Dialogfeld auf das Optionsfeld AUTOAUSFÜLLEN und anschließend auf OK klicken, haben Sie dasselbe erreicht wie mit dem Drag&Drop-Verfahren.

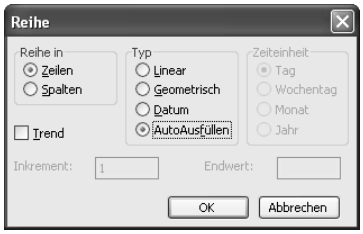


Abbildung 1.3: Das Dialogfeld REIHE

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 diese nicht zig Mal neu eingeben wollen. Daher lohnt es sich zu wissen, wie das automatische Ausfüllen funktioniert.

Zellbezüge

Die zweite wichtige Grundlage ist die Art und Weise, wie Excel Zellbezüge in Arbeitsblättern herstellt. Betrachten wir noch einmal das Arbeitsblatt in Abbildung 1.2. 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 genau. Ich möchte den Gesamtbetrag in jeder Zeile, also passt Excel die Formel beim automatischen Ausfüllen der einzelnen Zellen entsprechend an. Diese Art der 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.

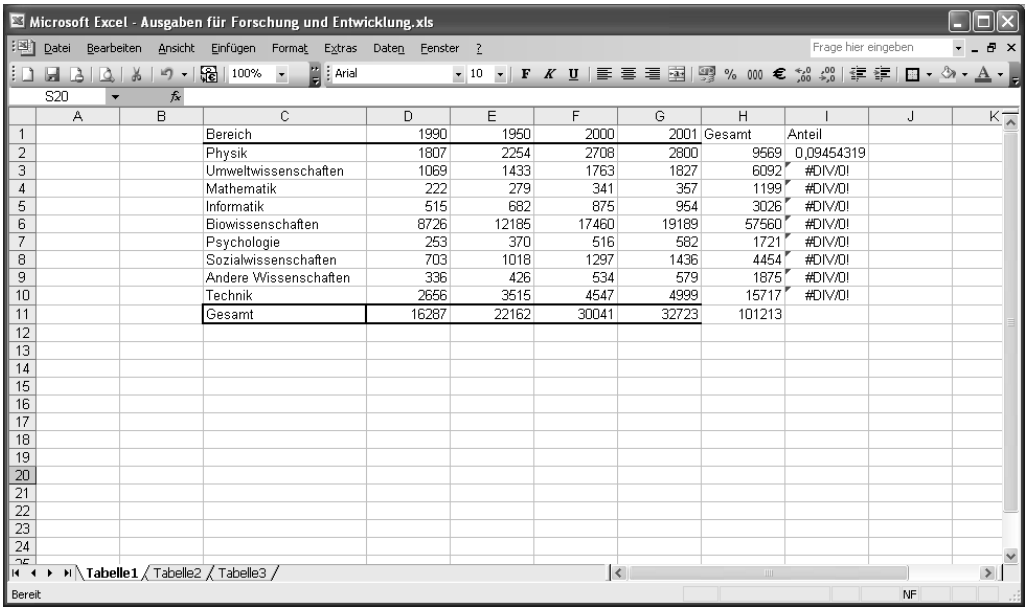
http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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 gebe ich zunächst die folgende Formel in I2 ein:

= H2/H11

Wenn Sie die [↵]-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 ... oje, oje! In Abbildung 1.4 ist das misslungene Ergebnis dargestellt. In den Zellen I3 bis I10 befindet sich dieses extrem hässliche #/DIV0! Was ist hier nur passiert?



	A	B	C	D	E	F	G	H	I	J	K
1			Bereich	1990	1950	2000	2001	Gesamt	Anteil		
2			Physik	1807	2254	2708	2800	9569	0,09454319		
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											
17											
18											
19											
20											
21											
22											
23											
24											

Abbildung 1.4: Whoops! 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

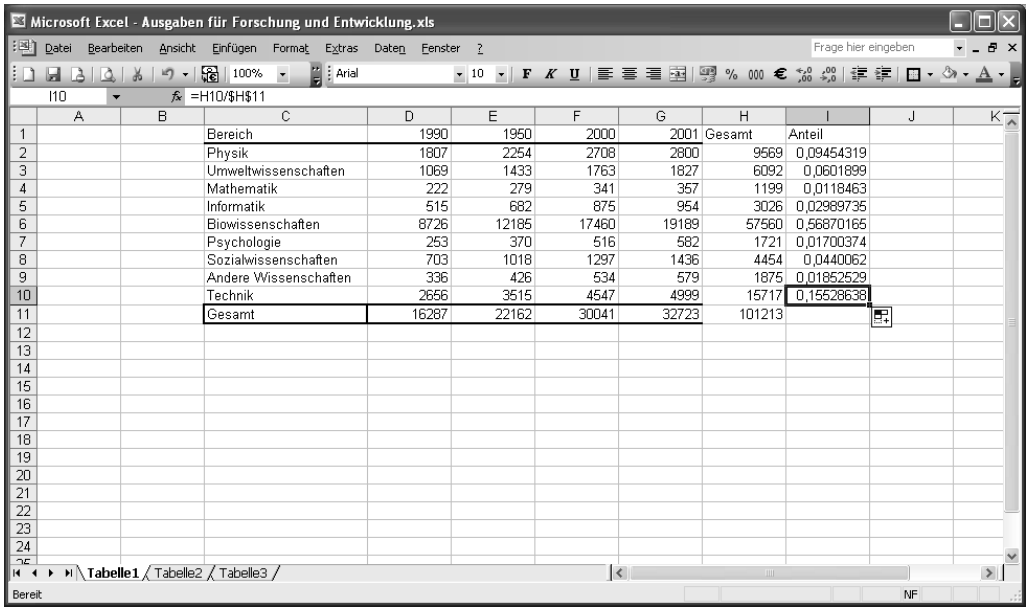
http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9

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/H11.

So weiß Excel, dass beim automatischen Ausfüllen weder die Spalte noch die Zeile angepasst werden soll. In Abbildung 1.5 ist das Arbeitsblatt mit den jeweiligen Anteilen dargestellt. In der Abbildung ist Zelle I10 markiert. Beachten Sie die Formel in der Bearbeitungsleiste, das längliche weiße Feld neben der Schaltfläche, die mit f_x gekennzeichnet ist. (Die Bearbeitungsleiste und diese Schaltfläche werden in Kapitel 2 beschrieben.)



	A	B	C	D	E	F	G	H	I	J	K
1			Bereich	1990	1950	2000	2001	Gesamt	Anteil		
2			Physik	1807	2254	2708	2800	9569	0,09454319		
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	0,02989735		
6			Biowissenschaften	8726	12185	17460	19189	57560	0,56870165		
7			Psychologie	253	370	516	582	1721	0,01700374		
8			Sozialwissenschaften	703	1018	1297	1436	4454	0,0440062		
9			Andere Wissenschaften	336	426	534	579	1875	0,01852529		
10			Technik	2656	3515	4547	4999	15717	0,15528638		
11			Gesamt	16287	22162	30041	32723	101213			

Abbildung 1.5: Automatisches Ausfüllen mit absoluten Zellbezügen

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

http://www.wiley-vch.de/publish/dt/books/ISBN3-527-70169-9