IN DIESEM KAPITEL

- Ihre Rolle als Datenanalyst erkennen
- Lernen, statistische Fehler zu vermeiden
- Die Sprache der weiterführenden Statistik kennen lernen

Kapitel 1
Datenanalyse als Kunst und Wissenschaft

Sie lesen dieses Buch, also sind Sie sehr wahrscheinlich bereits vertraut mit den Grundlagen der Statistik. Jetzt werden Sie das Ganze etwas erweitern. Auf der nächsten Stufe nutzen Sie Ihre bisherigen Kenntnisse, fügen ein paar weitere Werkzeuge und Techniken der fortgeschrittenen Stufe hinzu und setzen schließlich alles zusammen, um anhand echter Daten realistischere Fragen beantworten zu können.

In der Sprache der Statistik werden Sie jetzt die Welt der *Datenanalysten* betreten. Diese Welt ist sehr interessant. Sie können zahlreiche Möglichkeiten kennen lernen, und es stehen Ihnen viele Werkzeuge zur Verfügung. Aber Sie haben es vielleicht schon erkannt: Sie sollten sich in dieser Welt sehr vorsichtig bewegen und immer die richtigen Methoden für jede Situation auswählen. In diesem Buch werden Sie merken, dass ich die zugrunde liegenden Theorien und die Konzepte hinter den Methoden erkläre, wo es erforderlich ist, um Ihnen zu helfen, gute Entscheidungen zu treffen – und nicht nur in das Zeig&Klick-Verhalten zu verfallen, das so viele Softwarepakete heute unterstützen.

In diesem Kapitel geht es um die Begriffe aus der Statistik, die ein Datenanalyst auf der fortgeschrittenen Stufe benötigt. Sie werden ahnen, welchen Einfluss Ihre Ergebnisse haben können, wenn Sie erkennen, was diese Analysetechniken leisten. Außerdem erfahren Sie mehr über die häufigsten Fehlanwendungen innerhalb der Datenanalyse und ihre Auswirkungen.

Datenanalyse: Nicht mehr nur für Statistiker

Bis vor einiger Zeit konnten nur Statistiker eine Datenanalyse durchführen. Der Grund dafür ist, dass die einzigen Computerprogramme, die damals zur Verfügung standen, sehr kompliziert waren und sehr viel Wissen über Statistik erforderten, wenn man sie anwenden und ausführen wollte. Die Berechnungen waren mühsam, und manchmal sogar unvorhersehbar, und man benötigte ein fundiertes Verständnis für die Theorien und Methoden hinter den Berechnungen, um richtige und zuverlässige Antworten zu erhalten.

Heute kann jeder auf einfache Weise Daten analysieren. Viele benutzerfreundliche Statistik-Softwarepakete wurden genau auf diesen Zweck ausgelegt – Microsoft Excel, Minitab, R, SAS und SPSS, um nur ein paar wenige zu nennen. Es gibt sogar verschiedene kostenlose Online-Programme, wie etwa Stat Crunch, die Ihnen helfen, Zahlen zu verarbeiten und Ergebnisse zu erhalten. Wie Sie in diesem Abschnitt erfahren werden, sind die modernen, einfach zu bedienenden Statistik-Pakete einerseits gut, andererseits schlecht.

Bei der Anwendung statistischer Techniken zur Datenanalyse ist es von entscheidender Bedeutung, zu wissen, was bei der Zahlenverarbeitung passiert, damit Sie (und nicht der Computer) die Analyse überwachen können. Aus diesem Grund sind die Kenntnisse aus der fortgeschrittenen Statistik so wichtig.

Die gute alte Zeit

In der guten alten Zeit musste man ein Computerprogramm schreiben, um festzustellen, welche Methoden welche Ergebnisse erzeugten. Dazu mussten Sie erst einmal lernen, zu programmieren. Sie mussten Ihre Daten auf ganz spezielle Weise eingeben, wie das Programm sie eben benötigte, und dann mussten Sie Ihr Programm an einen Großrechner weitergeben und darauf warten, dass Ihre Ergebnisse auf dem Drucker erschienen. Das alles war zeitaufwändig und ganz allgemein unangenehm.

Ich erinnere mich noch genau an den Tag meines völligen Zusammenbruchs an der Uni. Ich hatte gerade gelernt, diese komplizierten Programme zu schreiben, die man selbst für die einfachsten Analysen benötigte. Egal, wie sehr ich versuchte, ein perfektes Programm zu schaffen, der Computer spuckte meine Arbeit wieder aus, ohne irgendeine Analyse durchgeführt zu haben, weil er bei der Eingabe immer wieder Fehler feststellte. Den Rest gab er mir, als ich mein Programm zum x-ten Mal eingab, und der Computer mir schließlich in der allerletzten Zeile mitteilte: »Fehler #34410: Zu viele Fehler.«

Denken Sie aber jetzt nicht, Ihre Autorin wüsste nicht, was sie tut. Meine statistischen Methoden waren richtig, ich war nur sehr schlecht im Programmieren. Wenn Sie also je von einem Computer frustriert wurden, dann kann ich es Ihnen nachfühlen und ich werde versuchen, Ihre Probleme mit diesem Buch aus der Welt zu schaffen.

Genug gejammert. Man kann sagen, dass Statistik-Software in den letzten 10 bis 15 Jahren eine unglaubliche Entwicklung durchgemacht hat. Heute können Sie Ihre Daten schnell und einfach und in fast jedem Format eingeben. Außerdem sind die Auswahlmöglichkeiten

für die Datenanalyse übersichtlicher und in Pulldown-Menüs angeordnet. Heute kann fast jeder (auch ich) schnell erkennen, wie er die richtige Verfahrensweise auswählt, und dem Computer mitteilen, was zu tun ist. Die Ergebnisse kommen sofort und erfolgreich, und Sie können sie im Handumdrehen kopieren und in ein Dokument aus Ihrem Textverarbeitungsprogramm einfügen. Beispielsweise brauchen Sie für einen Vergleich des Gewichtsverlusts von Menschen in verschiedenen Diätprogrammen heute nur noch drei Mausklicks, was für Leute wie mich wirklich hochinteressant ist.

Es gibt zahlreiche sehr praktische und effiziente Softwarepakete, wie unter anderem SAS, SPSS, Data Desk, Stat Crunch, MS Excel, Minitab und R. Alle haben ihre Vor- und Nachteile (und manche benutzen sie gerne, andere lehnen sie ab). Meine Software der Wahl, anhand der ich dieses Buch geschrieben habe, ist ursprünglich Minitab beziehungsweise in der Neuauflage insbesondere SPSS – letzteres, weil es sehr einfach zu bedienen ist, weil es korrekte Ergebnisse erzeugt und weil die Ausgabe sehr klar und professionell aussieht. Außerdem werden in diesem Programm grundsätzlich alle Datenanalysetechniken unterstützt, die in der weiterführenden Statistik verwendet werden und die auch in diesem Buch vorgestellt werden.

Der Nachteil der heutigen Statistik-Software

Sie fragen sich vielleicht, wo hier ein Nachteil liegen soll. Wenn es einst mühselig war, Daten zu analysieren, ist es jetzt so unkompliziert geworden wie der Abruf von E-Mail über das Handy – zu gut, um wahr zu sein. Ja und nein. Ja, es ist zu gut, um wahr zu sein, dass die Software praktisch alles für Sie macht – wenn es Ihnen egal ist, was die Programme eigentlich tun. Ja, es ist zu gut, um wahr zu sein, wenn Sie nicht wissen, dass die Bedingungen jedes Mal geprüft werden müssen, bevor eine Analyse durchgeführt werden sollte. Ja, es ist zu gut, um wahr zu sein, wenn Sie die Ergebnisse als vollständig und wunderbar betrachten (wie es viele Möchtegern-Statistiker machen).

Fazit: Die heutigen Softwareprogramme sind zu gut, um wahr zu sein, wenn Sie kein klares und fundiertes Wissen über weiterführende Statistik besitzen, um ihre Arbeitsweise zu verstehen.

Es gibt jedoch auch gute Nachrichten. Wenn Sie dieses Buch lesen, eignen Sie sich das Wissen an, das Sie brauchen, um erfolgreich mit Statistiken arbeiten zu können. Sie lernen so viele Konzepte der fortgeschrittenen Statistik kennen, dass Sie inspiriert werden und sich nicht mehr in gefährliche Situationen manövrieren. Sie werden herausfinden, welche Bedingungen für die Daten überprüft werden müssen, bevor Sie eine Analyse darauf anwenden können, und wie sie geprüft werden. Sie werden ein Gefühl dafür entwickeln, welche Analysen für die Beantwortung Ihrer Fragen zu verwenden sind (und welche Probleme verursachen könnten), und Sie werden abschätzen können, welche Art Ergebnisse Sie erwarten können. Vor allem werden Sie aber entdecken, was alles möglich und zulässig ist, um Schlüsse aus Ihrer Analyse zu ziehen, und welche Einschränkungen Sie treffen und welche Vorbehalte Sie haben müssen.

Regel Nr.1: Informieren Sie sich VOR der Verarbeitung!

Viele Leute begreifen nicht, dass einem die Statistik-Software nicht sagen kann, wann eine bestimmte statistische Technik anzuwenden ist – und wann nicht. Dafür sind Sie selbst verantwortlich. Aus diesem Grund glauben die Leute, dass sie korrekte Analysen durchführen, aber irgendwann machen sie alle möglichen Fehler. Statistik-Softwarepakete basieren auf mathematischen Formeln, und mathematische Formeln sind nicht intelligent genug, zu erkennen, wie Sie sie anwenden, oder Sie zu warnen, wenn Sie irgendetwas falsch machen (und hier kommt dieses Buch ins Spiel).

In diesem Abschnitt zeige ich Ihnen einige Beispiele für die wichtigsten Situationen, wo naiv durchgeführte Datenanalysen schieflaufen, und warum es so wichtig ist, zu wissen, was vom statistischen Standpunkt her hinter den Kulissen abläuft, bevor Sie mit der Datenverarbeitung beginnen.

Nichts ist ewig (nicht einmal eine Gerade)

Nachdem Sie eine statistische Gleichung oder ein Modell gefunden haben, um bestimmte zufällige Phänomene zu erklären oder vorherzusagen, müssen Sie angeben, für welche Werte die Gleichung gilt und für welche Werte sie nicht gilt. Gleichungen erkennen nicht, wann sie funktionieren und wann sie nicht funktionieren. Dies muss der Datenanalyst erkennen. Dasselbe gilt für die Anwendung der Ergebnisse einer bereits durchgeführten Datenanalyse.

Franz Schauinsland ist ein Statistikstudent, der die Auswirkung der Lernzeit auf das Prüfungsergebnis untersucht. Basierend auf seiner Erfahrung und auf der Erfahrung von ein paar Freunden kommt Franz schließlich zu der Gleichung $y = 10x + 30$, wobei y das Prüfungsergebnis darstellt, das man erhält, wenn man eine bestimmte Anzahl Stunden (x) lernt. Diese Gleichung ist das Modell, das Franz für die Vorhersage der Prüfungsergebnisse anhand der Lernzeit verwendet. Beachten Sie, dass dieses Modell die Gleichung einer Geraden darstellt, mit einem y-Achsenschnittpunkt von 30 und einer Steigung von 10.

Franz sagt also anhand seines Modells voraus, dass Sie, wenn Sie überhaupt nicht lernen, das Ergebnis 30 für die Prüfung erhalten (Sie setzen $x = 0$ in die Gleichung ein und lösen nach y auf; dieser Punkt stellt den y-Achsenschnittpunkt der Geraden dar). Außerdem sagt er anhand dieses Modells voraus, dass Sie, wenn Sie fünf Stunden lernen, ein Prüfungsergebnis von $y = 10 * 5 + 30 = 80$ erhalten. Der Punkt (5, 80) liegt also auf dieser Geraden. (Ich will hier nicht genauer darauf eingehen, wie gut das Modell von Franz für die Vorhersage von Prüfungsergebnissen funktioniert, aber ich kann zumindest sagen, dass er noch daran zu arbeiten hat, und belasse es für den Moment dabei.)

Ich bin sicher, Sie stimmen mir zu, dass, wenn x die Lernzeit darstellt, es nie kleiner 0 sein kann. Wenn Sie eine negative Zahl für x einsetzen, beispielsweise $x = -10$, erhalten Sie $y = 10 * -10 + 30 = -70$, was keinen Sinn hat. Das schlechteste Ergebnis, das auftreten kann, ist nach dem Modell von Franz 30. Es tritt ein, wenn x gleich 0 ist. Außerdem kann man keine negative Stundenzahl lang lernen, deshalb ist eine negative Zahl für x sinnlos.

Andererseits sollte es auch nicht vorkommen, dass x eine Zahl im zweistelligen Bereich (10 oder mehr) ist. Warum? Angenommen, jemand hat zehn Stunden für diese Prüfung gelernt. Wenn man in die Gleichung von Franz jetzt 10 für x einsetzt, erhält man $y = 10 * 10 + 30$, also 130. Weil die meisten Prüfungen maximal 100 Punkte vergeben, ist eine Bewertung von 130 nicht möglich. (Ich bin für Bonuspunkte in Prüfungen, aber 30 Zusatzpunkte sind sogar für mich zu viel.)

Der Punkt ist hier, dass es Grenzen für die Werte von x gibt, die in dieser Gleichung sinnvoll sind. Die eigentliche Gleichung, $y = 10x + 30$, weiß davon nichts, und wenn Sie diese Gerade zeichnen, verläuft diese endlos sowohl in positiver als auch in negativer Richtung (siehe Abbildung 1.1).

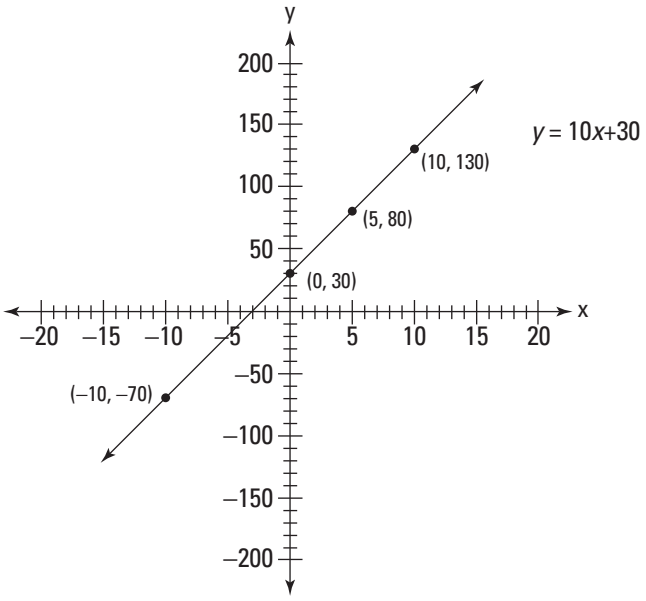

Abbildung 1.1: Die Gerade $y = 10x + 30$ für alle möglichen Werte von x

Datenschnüffeln ist nicht cool!

Statistiker gebrauchen ein Sprichwort, das Sie sicher schon gehört haben: »Zahlen lügen nicht. Zahlen werden gelogen.« Sorgen Sie dafür, dass Sie alle Analysen herausfinden, die für eine Datenmenge durchgeführt wurden, und nicht nur diejenigen, die als statistisch relevant angegeben wurden.

Angenommen, Franz Schauinsland versucht, sein einfaches Modell (aus dem vorigen Abschnitt) anzuwenden, um die Prüfungsergebnisse für seine ganze Klasse vorherzusagen, indem ihm seine Mitschüler jeweils ihre Lernzeit mitteilen. Leider stellt er fest, dass seine Ergebnisse nicht zutreffen. Er findet heraus, dass er mehr Informationen braucht, deshalb versucht er zu erforschen, welche anderen Faktoren außer der Lernzeit helfen könnten, die Prüfungsergebnisse für eine Statistikprüfung vorherzusagen, Franz misst alles, von

der Suppe bis zur Schokolade. Seine Menge möglicher Variablen beinhaltet Lernzeit, Notendurchschnitt, bereits vorhandene Erfahrung in Statistik, Mathematikvorbildung, Interesse an Statistik, ob die Schüler beim Lernen klassische Musik hören, Schuhgröße, ob während der Prüfung Kaugummi gekaut wird, und sogar die Lieblingsfarbe (man kann nie wissen). Um das Ganze zu vervollständigen, nimmt er weitere 11 Variablen auf, so dass er insgesamt 20 mögliche Faktoren hat, von denen er denkt, sie könnten sich auf das Prüfungsergebnis auswirken.

Franz sucht zuerst nach Beziehungen zwischen jeder dieser Variablen und dem Prüfungsergebnis, deshalb braucht er 20 Korrelationen. (*Korrelation* ist ein Maß für die lineare Beziehung zwischen zwei Variablen; weitere Informationen finden Sie im Abschnitt über Korrelation später in diesem Kapitel.) Er stellt fest, dass vier Variablen eine statistisch signifikante Beziehung zum Prüfungsergebnis besitzen (das heißt, die Ergebnisse können mit einer Wahrscheinlichkeit von 95 Prozent korrekt sein – aber nur, wenn die Daten korrekt gesammelt wurden und die Analyse korrekt ausgeführt wurde).

Die Variablen, für die Franz eine Beziehung zum Prüfungsergebnis festgestellt hat, waren die Lernzeit, die Mathematikvorbildung, der Notendurchschnitt und ob die Person während der Prüfung Kaugummi kaut. Es stellt sich heraus, dass dieses Modell relativ gut passt (nach Kriterien, die ich in Kapitel 5 bei der Beschreibung von Modellen für die multiple lineare Regression erläutern werde). Franz ist stolz auf sich und beantwortet die alles entscheidende Frage: Wie kann ich eine bessere Note in der Statistikprüfung schreiben?

Aber man kennt es schon von Apollo 13: »Houston, wir haben ein Problem.« Durch die Betrachtung aller möglichen Korrelationen zwischen seinen 20 Variablen und dem Prüfungsergebnis führt Franz letztlich 20 separate statistische Analysen durch. Unter typischen Bedingungen (die ich in Kapitel 3 beschreiben werde) besteht für jede statistische Analyse eine Wahrscheinlichkeit von 5 Prozent, dass sie falsch ist (man spricht dabei auch vom *Signifikanzniveau* des Tests).

Weil 5 Prozent bei 20 Analysen gleich 1 sind, kann man davon ausgehen, dass eine von 20 Analysen das falsche Ergebnis erzeugt – im Durchschnitt und über lange Zeit betrachtet. Sicher können Sie raten, welche der Korrelationen von Franz in diesem Fall ein falsches Ergebnis verursacht hat. Natürlich hat die Lernzeit *nichts* mit der Prüfungsnote zu tun, und der Kaugummi ist die Lösung für alle unsere Probleme – oder? (Wenn das der Fall wäre, gäbe es keine Statistiker mehr, weil sie alle für Kaugummifabriken arbeiten würden.)

Franz macht das, was man in der Datenanalyse als *Data Snooping* bezeichnet. Franz sucht so lange, bis er etwas findet, und glaubt dann, er hätte das Ergebnis. Diese Strategie ist gefährlich, wird aber in der Realität allzu häufig eingesetzt. Einer der Gründe, warum das Data Snooping heute so zunimmt, ist, dass einfach alle Daten sammeln und sie analysieren – und jeder will irgendwelche Erkenntnisse daraus ableiten. Man verwendet Statistik-Software, die erlaubt, mit ein paar Klicks beliebig viele Analysen durchzuführen, ohne jegliches Gefühl für etwas, was von den Statistikern als *Gesamtfehlerrate* bezeichnet wird (das heißt die Wahrscheinlichkeit, innerhalb eines beliebigen Schritts in der Analyse einen zufälligen Fehler zu machen, und nicht nur die Wahrscheinlichkeit, einen zufälligen Fehler innerhalb einer Analyse zu machen).

(Daten-)Fischen verboten!

Die Wiederholung von Analysen auf verschiedene Arten, um die gewünschten Ergebnisse zu erzielen, wird in der Statistik auch als *Data Fishing* bezeichnet. Die Statistiker betrachten es als großes No-Go (obwohl viele Leute es im Namen der Forschung leider allzu häufig tun).

Maria Besserwisser beispielsweise ist davon überzeugt, dass Zucker im Wasser Schnittblumen länger leben lässt. Sie führt ein Experiment durch, um ihre Hypothese zu beweisen. Sie schneidet zwei Dutzend Rosen und stellt jeweils eine Rose in eine Vase. Sie füllt jede Vase mit 3 Tassen Wasser, aber in 12 der Vasen gibt sie auch noch einen Teelöffel Zucker (die anderen 12 Vasen bilden die Kontrollgruppe, das heißt, Maria behandelt sie nicht weiter, um zu zeigen, was passiert, wenn sie dem Wasser nichts hinzufügt). In den nächsten Abschnitten werden Sie Maria bei ihrem Experiment begleiten. Haben Sie dabei ein Auge auf die statistischen Analysen, die Ihnen begegnen!

Überprüfung der Daten von Maria

Maria zählt, wie viele Tage die Blumen noch gut aussehen, und wendet auf jede Blume dieselben Kriterien an. Nach 10 Tagen sind alle Blumen so weit verwelkt, dass sie entsorgt werden müssen, das Experiment ist also abgeschlossen. Die Daten von Maria sehen Sie in Tabelle 1.1.

Beobachtung	Haltbarkeitstage: Nur Wasser	Haltbarkeitstage: Zuckerwasser
1	3	5
2	3	5
3	4	5
4	4	4
5	4	4
6	4	4
7	3	3
8	3	4
9	2	3
10	4	3
11	4	5
12	4	5

Tabelle 1.1: Tage, die die Rosen im Zuckerwasser im Vergleich zu normalem Wasser (Kontrollgruppe) haltbar waren

Die Hypothese aufstellen

Maria will die beiden Methoden, Wasser und Wasser mit Zucker, vergleichen, um festzustellen, ob die Rosen, die Zucker erhalten haben, länger hielten als die Gruppe in

normalem Wasser. Sie muss einen Hypothesentest durchführen, dessen Nullhypothese Ho lautet: Die Rosen in der Kontrollgruppe sind mindestens solange haltbar wie die Rosen in der Zuckergruppe. Die Alternativhypothese Ha, die sie zeigen will: Die Rosen in der Zuckergruppe wa-ren länger haltbar als die in der Kontrollgruppe. Sie geht davon aus, dass ein *t*-Test für zwei Stichproben hier ausreichend ist. (Um Hypothesentests wird es in Kapitel 3 gehen.)

Prüfung der Bedingungen

Maria hat bereits Statistikkurse besucht und weiß, dass sie vor einer Analyse die korrekten Voraussetzungen prüfen muss. Für einen Vergleich von zwei Gruppen muss sie die Daten aus jeder Gruppe in einem *Histogramm* ausgeben (ein Balkendiagramm, das auf der x-Achse die Anzahl der Tage zeigt, wie lange die Blumen haltbar waren, gruppiert in numerischer Reihenfolge, und auf der y-Achse die Anzahl der Blumen, die die jeweilige Anzahl an Tagen gehalten haben). Gemäß dem, was sie über einen *t*-Test für zwei Stichproben weiß, müssen die Daten in jeder Gruppe normalverteilt sein, bevor sie anfangen kann. Das bedeutet, die Daten müssen bei Betrachtung des Balkendiagramms einer Glockenkurve ähnlich sein. Maria trägt die Daten in die Balkendiagramme für die beiden Gruppen ein und erhält die folgenden Ergebnisse (siehe Abbildung 1.2 und 1.3).

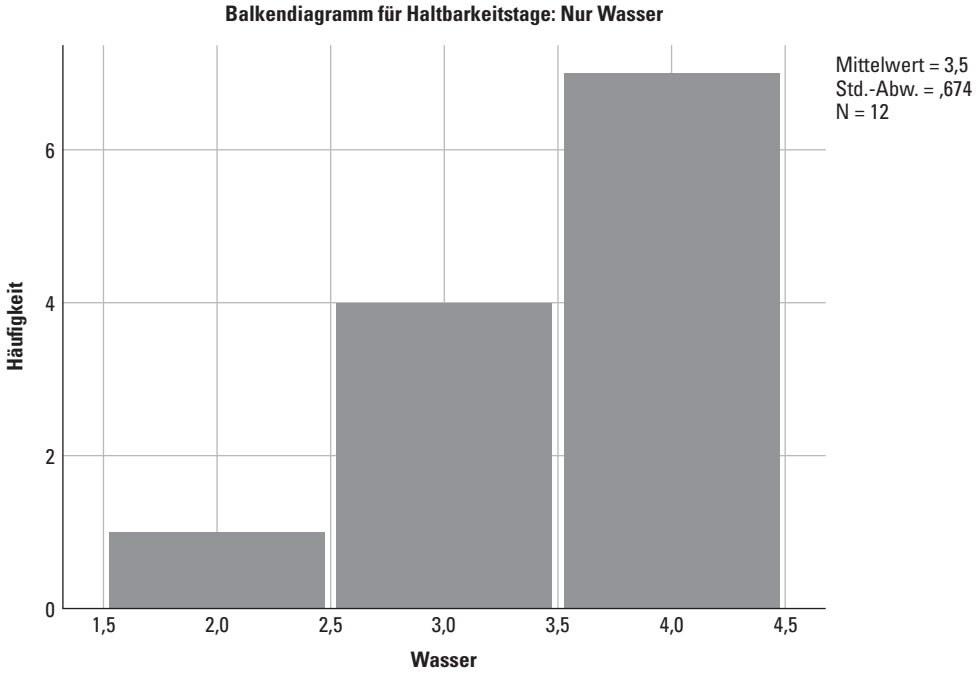

Abbildung 1.2: Histogramm, das die Anzahl der Tage, die die Rosen haltbar waren, zeigt, wobei nur Wasser verwendet wurde

Abbildung 1.3: Histogramm, das die Anzahl der Tage, die die Rosen haltbar waren, zeigt, wobei nur Wasser mit Zucker verwendet wurde

Und jetzt die schlechten Nachrichten

Wie Sie in Abbildung 1.2 und 1.3 sehen, weisen die Daten von Maria nicht die typische glockenförmige Kurve auf. Eines der Probleme ist, dass ihre Daten nur positive ganze Zahlen als Werte annehmen können, deshalb sind Zahlen wie 1,2; 2,3 und so weiter nicht möglich. (Bei Normalverteilungen geht man von vielen möglichen Werten aus.) Das andere Problem ist, dass die Daten außerhalb des typischen 2-, 3-, 4- oder 5-Tage-Bereichs keine Werte enthalten, das Balkendiagramm hat also gar keine Chance, eine Glockenkurve zu bilden. Mehr Daten hätten dieses Problem möglicherweise behoben. In jedem Fall weiß Maria, dass die Bedingungen für einen *t*-Test für zwei Stichproben nicht erfüllt sind, weil die Daten keine Normalverteilung haben. Sie erscheinen viel mehr verzerrt (das heißt, sie verlaufen ansteigend von einer Seite zur anderen).

Nichtparametrischer Versuch

Unerschrocken von dieser Wendung der Ereignisse wendet Maria einen nichtparametrischen Test ihrer Daten an, was das Richtige ist. In Situationen, wo die Voraussetzungen für die typischen Analysen nicht erfüllt sind, verwenden Statistiker *nichtparametrische Statistiken* (wenn beispielsweise keine Normalverteilung vorliegt). Die nichtparametrische Statistik erzeugt häufig *konservativere* (wenngleich exaktere) Ergebnisse als die typischen (parametrischen) Verfahrensweisen, an die Sie gewöhnt sind. (Ich werde im letzten Abschnitt dieses

Kapitels etwas mehr auf nichtparametrische Statistik eingehen; nichtparametrische Verfahren sind detailliert in den Kapiteln 16 bis 19 beschrieben.)

Weil die Daten von Maria nicht normalverteilt sind und nicht einmal eine *symmetrische Verteilung* aufweisen (das heißt eine Verteilung, die auf jeder Seite gleich aussieht, wenn man sie in der Mitte auseinanderschneidet), ist der Mittelwert (oder der Durchschnitt) kein gutes Maß für die Mitte der Daten, deshalb ist ein *t*-Test für zwei Stichproben nicht möglich. Als Alternative kann sie testen, ob die beiden Histogramme gleich sind oder nicht, wenn sie die Histogramme der beiden betreffenden Populationen vergleicht (alle Rosen mit Wasser im Vergleich zu allen Rosen mit Zuckerwasser).

Weil Maria zwei Gruppen vergleicht, verwendet sie einen Wilcoxon-Rangsummentest, auch als Mann-Whitney-Test bezeichnet (siehe Kapitel 19). Der Wilcoxon-Rangsummentest prüft, ob zwei Populationen dieselbe Verteilung haben (das heißt, ob die beiden Histogramme gleich aussehen) oder ob eine der Populationen nach links oder rechts verschoben ist. Maria hat die Theorie, dass die Zuckergruppe länger haltbar ist, deshalb testet sie Ho: »Zuckergruppe und Kontrollgruppe haben dieselbe Verteilung« gegenüber Ha: »Die Zuckergruppe ist gegenüber der Kontrollgruppe nach rechts verschoben«.

Maria scheitert

Zu allem Unglück kann der Wilcoxon-Rangsummentest die Nullhypothese von Maria leider nicht widerlegen. Sie hat nicht bewiesen, was sie durch ihr Experiment bestätigen wollte. Nicht genügend Rosen in der Zuckergruppe waren länger haltbar als die Rosen in der Kontrollgruppe. Sie sehen den Grund für dieses Ergebnis, indem Sie die Mediane der beiden Gruppen vergleichen. Wenn Sie den Median jeder der Datenmengen in Tabelle 1.1 suchen, erhalten Sie in jedem Fall den Wert 4. Weil die Mediane der beiden Datenmengen gleich sind, ist es unwahrscheinlich, dass Maria mit Hilfe dieses Tests ein statistisch signifikantes Ergebnis findet.

Die Regeln werden gebrochen

Nach den Regeln, die alle guten Statistiker befolgen, sollte die Geschichte von Maria hier zu Ende sein. Sie kann immer noch davon überzeugt sein, dass Zucker Rosen hilft, länger zu halten. Sie kann in ihrem weiteren Leben immer Zucker in das Rosenwasser geben und ihren Freunden entsprechende Tipps geben. Aber sie darf nicht sagen, dass Zuckerwasser statistisch gesehen andere Ergebnisse als Wasser alleine erzeugt. Ihre Analyse konnte dies nicht zeigen.

Aber Sie wissen ja, dass Maria mit Nachnamen Besserwisser heißt, deshalb will sie unbedingt diese Ergebnisse erhalten. Sie weiß, dass nichtparametrische Tests im Allgemeinen konservativere Ergebnisse erzeugen als reguläre Tests, und trotz der Tatsache, dass die Bedingungen nicht erfüllt sind, beschließt sie, ihre Daten erneut zu analysieren, diesmal mit dem *t*-Test für zwei Stichproben.

Die Daten in einen *t*-Test für zwei Stichproben einzusetzen, dauert nur zwei Mausklicks, und die Ergebnisse von Maria erbringen einen *p*-Wert von 0,043. Unter Verwendung des üblichen Signifikanzniveaus für Hypothesentests, 0,050, ist ihr *p*-Wert kleiner als diese Zahl, sie kann

also Ho widerlegen. (In einem *t*-Test mit zwei Stichproben besagt Ho, dass der Mittelwert der Kontrollgruppe mindestens so groß ist, wie der Mittelwert der Zuckergruppe. Und ihr Ha ist in diesem Fall, dass der Mittelwert der Zuckergruppe größer als der Mittelwert der Kontrollgruppe ist.) Triumphierend gratuliert sich Maria selbst zu den Ergebnissen, die sie wollte, und sagt sich, dass es niemandem wehtut, wenn man eine andere Analyse ausprobiert, nachdem alles andere fehlgeschlagen ist.

Der Fehler in Marias Vorgehensweise

Noch einmal: »Houston, ... « – den Rest kennen Sie. Marias Problem ist, dass sie geschummelt hat, um ein Ergebnis zu erhalten, das falsch ist. Sie wusste, dass die Bedingungen für den *t*-Test für zwei Stichproben nicht erfüllt waren, aber wenn die korrekte Analyse schon nicht die Ergebnisse erbrachte, die sie wollte, hat sie eine Analyse gesucht, die das geschafft hat. Das Problem ist, dass die Ergebnisse des *t*-Tests für zwei Stichproben trügerisch sind.

Es geht vielleicht nicht um Leben oder Tod, wenn man prüft, ob Rosen mit Zucker im Wasser ein bisschen länger halten oder nicht. (Übrigens sagen die Gärtner, dass das nicht der Fall ist und dass Zucker sogar das Wachstum von Bakterien am Anschnitt fördern kann, wodurch die Blume überhaupt kein Wasser mehr aufnehmen kann.) Aber stellen Sie sich eine Situation vor, bei der Ärzte versuchen zu testen, ob ein bestimmtes Medikament hilft, eine Krankheit schneller zu überwinden, oder ob ein bestimmtes Verfahren Krebspatienten hilft, länger zu leben. Diese Ergebnisse haben wirklich ernsthafte Auswirkungen.

Die Verwendung der falschen Datenanalyse, um die gewünschten Ergebnisse zu erzielen, führt zu zwei großen Problemen:

- ✔ Sie täuschen Ihre Zuhörerschaft, weil diese denkt, Ihre Hypothese sei korrekt, was möglicherweise nicht der Fall ist.
- ✔ Früher oder später wird jemand versuchen, diese Ergebnisse nachzuvollziehen, und stellt dann fest, dass sie nicht nachvollzogen werden können. Diese Entdeckung bewirkt, dass Sie unglaubwürdig werden. Leider haben Sie in der Zwischenzeit auch viele Leute getäuscht.

Das große Ganze: Ein Überblick über weiterführende Statistik

Aufgrund der Gefahren und der bleibenden Wirkungen, die entstehen, wenn für die Beantwortung von Fragestellungen eine Datenanalyse durchgeführt wird und dabei in den falschen Situationen die falschen Techniken ausgewählt werden, ist es sehr wichtig, zu wissen, was hinter den Kulissen der Datenanalyse passiert. Man sollte die Regeln für die sinnvolle Auswahl von Techniken und geeigneten Verfahrensweisen beherzigen. Mit anderen Worten, es ist wichtig für Sie, dass Sie Ihr Statistikwissen auf die nächste Stufe bringen.

Die weiterführende Statistik ist eine Erweiterung der einführenden Statistik, deshalb bleibt die Terminologie gleich und die Techniken bauen auf Ihrem bereits vorhandenen Wissen auf.

Wenn Sie die Konzepte aus dem ersten Kurs verstanden haben, werden Sie keine Probleme mit der Terminologie für die weiterführende Statistik haben. Wenn Sie noch Unsicherheiten in Hinblick auf bestimmte Begriffe aus der einführenden Statistik haben, können Sie Ihr Lehrbuch aus dem ersten Kurs zu Rate ziehen oder in meinem anderen Buch *Statistik für Dummies* nachlesen.

Im nächsten Abschnitt erhalten Sie eine Einführung in die Terminologie der fortgeschrittenen Statistik, ebenso wie einen allgemeinen Überblick über die Techniken, die die Statistiker für die Datenanalyse anwenden, und wie sich das Ganze zusammenfügt.

Populationsparameter

Ein *Populationsparameter* ist eine Zahl, die die Population beschreibt (also die ganze Gruppe, die Sie untersuchen wollen). Beispiele für Parameter sind unter anderem der Mittelwert einer Population, der Median einer Population oder der Anteil der Population, der in eine bestimmte Kategorie einzuordnen ist.

Angenommen, Sie wollen die durchschnittliche Länge eines Handy-Gesprächs zwischen Teenagern (Alter von 13 bis 18) ermitteln. Sie sind nicht an Vergleichen interessiert, sondern wollen nur eine gute Abschätzung der durchschnittlichen Zeit. Sie wollen also einen Populationsparameter schätzen (wie etwa den Mittelwert oder Durchschnittswert). Die Population besteht aus allen Handy-Benutzern zwischen 13 und 18 Jahren. Der Parameter ist die durchschnittliche Länge eines Handy-Gesprächs, das diese Population führt.

Stichprobenkenngröße

Normalerweise können Sie nicht jedes Mitglied einer ganzen Population genauer betrachten (wie wollen Sie die Länge jedes einzelnen Handy-Gesprächs von allen Teenagern messen und aufzeichnen?). Es ist also nicht möglich, die Populationsparameter exakt zu bestimmen – Sie können sie nur schätzen. Aber es ist noch nicht alles verloren. Sie nehmen eine Stichprobe (eine Untermenge der Einzelmitglieder) aus der Population, untersuchen sie und können damit eine gute Schätzung für den Populationsparameter abgeben, wenn Sie Ihre Aufgabe richtig machen. Eine Untermenge dieser Population wird als *Stichprobe* bezeichnet. Eine *Stichprobenkenngröße* (auch *Stichprobenstatistik* genannt) ist eine einzelne Zahl, die diese Untermenge der Population beschreibt.

Im oben beschriebenen Handy-Szenario beispielsweise wählen Sie eine Stichprobe aus den Teenagern aus und messen die Länge ihrer Handy-Gespräche über einen bestimmten Zeitraum (oder sehen ihre Handy-Aufzeichnungen an, falls Sie auf legalem Wege drankommen können). Anschließend berechnen Sie den Durchschnitt der Gesprächslängen. Die durchschnittliche Länge von 120 Handy-Gesprächen könnte beispielsweise 12,2 Minuten betragen – dieser Durchschnittswert ist eine Kenngröße. Man spricht hier auch vom *Stichprobenmittel*, weil es den Durchschnittwert aus Ihren Stichprobendaten darstellt.

Es gibt auch eine Kenngröße namens *Stichprobenanteil* (der Anteil der Einzelmitglieder in der Stichprobe, die eine bestimmte Eigenschaft aufweisen – zum Beispiel der Prozentsatz der weiblichen Teenager, die Handys benutzen). Es gibt viele verschiedene Kenngrößen (die Sie

möglicherweise schon in der einführenden Statistik kennen gelernt haben), um verschiedene Eigenschaften einer Stichprobe zu betrachten, wie beispielsweise Median, Varianz und Standardabweichung.

Vertrauensintervall

Ein *Vertrauensintervall* ist ein Wertebereich, der sinnvolle Schätzungen für einen Populationsparameter zulässt. Ein Vertrauensintervall basiert auf einer Stichprobe und einer Kenngröße, die aus dieser Stichprobe stammt. Der wichtigste Grund dafür, einen Bereich möglicher Werte statt eines einzelnen Werts zu verwenden, ist, dass sich die Stichprobenergebnisse von Stichprobe zu Stichprobe unterscheiden.

Angenommen, Sie wollen den Prozentsatz der Leute schätzen, die Schokolade essen. Laut einem Marktforschungsinstitut berichten 78 Prozent aller Erwachsenen, dass sie Schokolade essen, und 18 Prozent von ihnen haben angegeben, dass sie häufig Süßigkeiten essen. Was fehlt an diesen Ergebnissen? Diese Zahlen sind nur eine einzige Stichprobe aus allen Menschen, und diese Stichprobenergebnisse variieren garantiert von Stichprobe zu Stichprobe. Sie brauchen ein Maß dafür, mit welcher Verschiebung dieser Ergebnisse Sie rechnen müssen, wenn Sie die Studie wiederholen.

Diese erwartete Verschiebung in Ihrer Statistik wird durch den *Fehlerspielraum* angegeben, der ein Vielfaches an Standardabweichungen Ihrer Statistik reflektiert, das Sie addieren oder subtrahieren, um ein bestimmtes Vertrauen in Ihre Ergebnisse zu erhalten (in Kapitel 3 finden Sie genauere Informationen über den Fehlerspielraum). Wenn die Ergebnisse für die Schokoladenesser auf 1.000 Menschen basieren, dann läge der Fehlerspielraum zum Beispiel bei 3 Prozent, das heißt, der tatsächliche Prozentsatz der Menschen, die innerhalb der gesamten Population Schokolade essen, kann mit 78 Prozent plus/minus 3 Prozent angenommen werden. Mit anderen Worten, er liegt irgendwo zwischen 75 Prozent und 81 Prozent. Wenn Sie diese Ergebnisse nur auf einer Stichprobe von 100 Menschen basieren lassen, bläht sich der Fehlerspielraum auf 10 Prozent auf, das heißt, der Prozentsatz der Schokoladenesser kann nur als zwischen 68 und 88 Prozent liegend angegeben werden. Beachten Sie, wie viel breiter das Intervall wird, wenn eine kleinere Stichprobengröße verwendet wird. Dieses Ergebnis bestätigt, dass mehr Daten für mehr Genauigkeit in Ihren Ergebnissen sorgen (vorausgesetzt, die Daten wurden korrekt gesammelt).

Hypothesentest

Ein *Hypothesentest* ist eine statistische Verfahrensweise, die Sie anwenden, um eine bestehende Behauptung über die Population anhand Ihrer Daten zu testen. Die Behauptung wird durch Ho (die Nullhypothese) festgehalten. Wenn Ihre Daten die Behauptung stützen, können Sie Ho nicht widerlegen. Wenn Ihre Daten die Behauptung nicht stützen, widerlegen Sie Ho und schließen auf eine alternative Hypothese, Ha. Der gebräuchlichste Grund, warum Menschen einen Hypothesentest durchführen, ist, nicht zu zeigen, dass ihre Daten eine vorhandene Behauptung stützen, sondern viel mehr, dass die vorhandene Behauptung falsch ist – zugunsten der alternativen Hypothese.

Ein Marktforschungsinstitut hat den Prozentsatz der Menschen untersucht, die sich die Sportnachrichten im Radio anhören. Ihre Statistik basiert auf einer Umfrage bei etwa 1.000

Menschen und hat erbracht, dass im Jahr 2000 23 Prozent der Menschen sagten, sie hören Radio, während im Jahr 2004 nur 20 Prozent angaben, Radio zu hören. Die Frage lautet: Stellt diese Verringerung um 3 Prozentpunkte von 2000 auf 2004 einen relevanten Trend dar, um den sich die Radiosender kümmern sollten?

Um diese Unterschiede formal zu testen, können Sie einen Hypothesentest einrichten. Sie legen Ihre Nullhypothese als das Ergebnis fest, das Sie ohne Ihre Studie annehmen müssen, Ho = es gibt keinen Unterschied zwischen den Daten für das Radio-Publikum zwischen 2000 und 2004.

Hier wird nun sehr allgemein beschrieben, was bei einem Hypothesentest passiert. Sie haben die Stichprobendaten und ermitteln die relevanten Statistiken. In diesem Fall haben Sie zwei Stichprobenprozentwerte, einen für 2000 und einen für 2004. Sie berechnen die Differenz zwischen den beiden Stichproben (3 Prozentpunkte) und dividieren sie durch den Standardfehler für die Differenz. Der Standardfehler gibt an, wie sehr sich die Differenz der Statistik absehbar zwischen den Stichproben ändern wird. In diesem Fall liegt der Standardfehler bei etwa 1,8 Prozent (wie das berechnet wird, erfahren Sie in Kapitel 3).

Berechnet man die Differenz zwischen den Statistiken (3 Prozentpunkte $= 0{,}03$) dividiert durch den Standardfehler ($1{,}8$ Prozent $= 0{,}018$), erhält man den Wert 1,67 (die so genannte *Teststatistik*). Dieser Wert repräsentiert die Differenz zwischen den beiden Statistiken bezogen auf den Standardfehler. Dieses Ergebnis kann ganz universell interpretiert werden. Allgemein ausgedrückt, wenn Ihre Teststatistik zwischen $-2{,}00$ und $+2{,}00$ liegt, bedeutet das, dass sich die ermittelten Ergebnisse nicht ausreichend unterscheiden, um sich Sorgen machen zu müssen, weil dieses Ergebnis in 95 Prozent aller Fälle nur zufällig zustande kommt. (Und das Beispiel entspricht genau dieser Situation.) Nachdem Sie die Variabilität der Stichprobenergebnisse berücksichtigt haben, überträgt sich die Differenz in den speziellen Stichproben nicht auf die von ihnen dargestellten Populationen. Weil Sie Ho nicht widerlegen können, müssen Sie also sagen, dass sich der Prozentsatz der Radiohörer innerhalb der gesamten Population von 2000 auf 2004 wahrscheinlich nicht geändert hat.

Weil Sie ein 95-prozentiges Vertrauensniveau haben, verwendet dieser Test ein Signifikanzniveau (α-Niveau) von $1 - 0{,}95 = 0{,}05$ oder 5 Prozent. Dieser Prozentsatz gibt an, wie wahrscheinlich es ist, dass Ihre Ergebnisse nur durch Zufall entstanden sind.

Das Problem ist, dass die Leute häufig nur die Stichprobenstatistik angeben und die erwartete Veränderung bei einer neuen Stichprobe nicht berücksichtigen. Dieses Versäumnis führt zu großen Fehlern beim Ziehen von Schlussfolgerungen (weitere Informationen über Hypothesentests finden Sie in Kapitel 3).

Varianzanalyse (ANOVA, Analysis of Variance)

ANOVA ist das Akronym für *Analysis of Variance* – Varianzanalyse. Sie wenden die ANOVA in Situationen an, wo man die Mittelwerte von mehr als zwei Populationen vergleichen will. Angenommen, Sie wollen die Lebensdauer von vier Reifenmarken vergleichen, abhängig von den gefahrenen Kilometern. Sie nehmen eine zufällige Stichprobe von 50 Reifen aus jeder Gruppe, betrachten also insgesamt 200 Reifen, und richten ein Experiment ein, um die

Lebensdauer aller Reifen zu vergleichen und aufzuzeichnen. Sie erhalten vier Mittelwerte und vier Standardabweichungen – eine für jede Datenmenge. Innerhalb der Gesamtdatenmenge von 200 Reifen gibt es aber unterschiedliche Arten von Schwankungen, die jeweils unter Verwendung verschiedener Quadratsummen angegeben werden. (Aus der Einführung in die Statistik wissen Sie, dass die Varianz von Datenmengen die Summe aller quadrierten Distanzen zwischen den Daten und dem Mittelwert dividiert durch $n - 1$ ist.)

Einer der Schwankungstypen innerhalb Ihrer Gesamtdatenmenge ist die Schwankung zwischen den Messreihen (auch als SST bezeichnet, »sums of squares for treatment« die Quadratsummen zwischen den Messreihen). SST gibt die Abweichung in der durchschnittlichen Lebensdauer jeder Reifenmarke im Vergleich zu der Gesamtdurchschnittslebensdauer an. Ist SST groß, dann ist es sehr wahrscheinlich, dass es aufgrund der jeweiligen Messreihe (in diesem Fall der Reifenmarke) eine Differenz in Hinblick auf die Lebensdauer gibt.

Als Nächstes haben Sie die Schwankung innerhalb der Messreihen (auch als SSE bezeichnet, »sums of squares for error« die Fehler-Quadratsumme). SSE gibt Schwankung der Reifenlebensdauern innerhalb jeder einzelnen Marke an (schließlich werden nicht alle Reifen gleich gut hergestellt, auch wenn sie von derselben Marke sind). Ist SSE groß, haben Sie so große Schwankungen innerhalb der einzelnen Reifenmarken, dass es schwieriger wird, echte Differenzen zwischen den Marken zu erkennen, selbst wenn es diese gibt.

Und schließlich haben Sie die Gesamtschwankung innerhalb der Datenwerte, wenn Sie sie einfach alle zu einer großen Datenmenge zusammenfassen. Diese Schwankung wird auch als SSTO bezeichnet, »sums of squares total«. Die ANOVA unterteilt die Gesamtschwankung (SSTO) in die Schwankung zwischen den Gruppen (SST) plus die Schwankung innerhalb der Gruppen (SSE).

Um auf Differenzen der durchschnittlichen Lebensdauer für die vier Reifenmarken zu testen, vergleichen Sie die mittleren Quadratsummen zwischen den Messreihen (MST) mit den mittleren Fehler-Quadratsummen (MSE), wozu Sie ein Verhältnis namens *F-Statistik* verwenden. Ist dieses Verhältnis groß, ist die Schwankung zwischen den Marken größer als die Schwankung innerhalb der Marken, was beweist, dass nicht alle Mittelwerte für die verschiedenen Reifenmarken gleich sind. Ist die *F*-Statistik klein, heißt das, dass nicht genügend Differenz zwischen den Messreihenmittelwerten vorlag – im Vergleich zu der allgemeinen Schwankung innerhalb der eigentlichen Messreihen. In diesem Fall kann man nicht sagen, dass sich die Mittelwerte für die Gruppen unterscheiden. (Weitere Informationen über die ANOVA finden Sie in den Kapiteln 9 und 10.)

Multiple Vergleiche

Angenommen, Sie führen eine Varianzanalyse durch und stellen eine Differenz in den durchschnittlichen Lebensdauern der vier Reifenmarken fest (siehe voriger Abschnitt). Ihre nächsten Fragen wären möglicherweise, welche Marken eine Differenz aufweisen und wie groß die Differenz ist. Um diese Fragen beantworten zu können, verwenden Sie Verfahrensweisen der multiplen Vergleiche.

Eine Verfahrensweise der multiplen Vergleiche ist eine statistische Technik, die Mittelwerte miteinander vergleicht und feststellt, welche eine Differenz aufweisen und welche nicht. Anschließend sind Sie in der Lage, die Gruppen in eine bestimmte Reihenfolge zu

bringen, von denjenigen mit dem größten Mittelwert bis zu denjenigen mit dem kleinsten Mittelwert, wobei berücksichtigt wird, wenn manchmal zwei oder mehr Gruppen zu nahe beieinanderliegen, um eine Differenz feststellen zu können, so dass sie in derselben Gruppe angeordnet werden.

Angenommen, Sie vergleichen die Prüfungsnoten von vier verschiedenen Klassen (wir bezeichnen sie als Klasse 1, Klasse 2, Klasse 3 und Klasse 4), und mit Ihrem Verfahren zur Varianzanalyse stellen Sie fest, dass nicht alle Mittelwerte gleich waren. Das bedeutet, die *F*-Statistik ist groß. Anschließend wenden Sie Verfahren für multiple Vergleiche an, um separate Vergleiche durchzuführen und festzustellen, welche Klassen etwa gleich waren und welche eine Differenz aufwiesen, und erhalten damit eine Reihenfolge der einzelnen Klassen. Es könnte beispielsweise der Fall sein, dass Klasse 4 einen statistisch größeren Wert als alle anderen Klassen besitzt, die Klassen 1 und 2 statistisch äquivalent sind, aber beide kleiner als Klasse 4 sind. Und Klasse 1 hat ganz hinten eine eigene Gruppe erhalten. Die Reihenfolge lautet: Klasse 4 (höchster Mittelwert), Klassen 2 und 3 (beide mit dem zweithöchsten Mittelwert) und Klasse 1 (niedrigster Mittelwert).

Führen Sie nie diesen zweiten Schritt aus, um die Mittelwerte der Gruppen zu vergleichen, wenn das Verfahren für die Varianzanalyse im ersten Schritt keine signifikanten Ergebnisse erbracht hat! (Weitere Informationen finden Sie in Kapitel 11.)

Es gibt viele verschiedene Verfahrensweisen für multiple Vergleiche, um einzelne Mittelwerte zu vergleichen und eine Reihenfolge einzuführen, falls Ihre *F*-Statistik feststellt, dass es Unterschiede gibt. Verfahrensweisen für multiple Vergleiche sind unter anderem der Tukey-Test, LSD und paarweise *t*-Tests. (Vielleicht erscheinen Ihnen die Namen dieser Tests etwas seltsam, aber machen Sie sich keine Sorgen – sie sind völlig legal.) Je nach Voraussetzungen und Ihrem Ziel als Datenanalytiker können bestimmte Verfahrensweisen besser als andere geeignet sein. In Kapitel 11 werden die Verfahrensweisen für multiple Vergleiche genauer beschrieben.

Interaktionseffekte

Ein *Interaktionseffekt* in der Statistik verhält sich genau wie in der Medizin. Wenn Sie zwei Medikamente gleichzeitig einnehmen, kann es sein, dass der Kombieffekt sehr viel anders ist, als wenn Sie die beiden Medikamente separat einnehmen würden.

Interaktionseffekte entstehen, wenn Sie ein Modell haben, das zwei oder mehr Variablen umfasst, und Sie diese Variablen verwenden, um Differenzen zu erklären oder Vergleiche in Hinblick auf ein bestimmtes Ergebnis durchzuführen. Wenn Sie zwei oder mehr Variablen in einem Modell haben, können Sie nicht automatisch den Effekt der einzelnen Variablen separat untersuchen. Sie müssen auch berücksichtigen, wie diese Variablen in Hinblick auf das Ergebnis zusammenarbeiten. Mit anderen Worten, Sie müssen prüfen, ob ein Interaktionseffekt vorliegt.

Angenommen, Medizinforscher untersuchen ein neues Medikament gegen Depressionen und wollen wissen, wie dieses Medikament bei Gabe einer geringen Dosis im Vergleich zur Gabe einer hohen Dosis den Blutdruck beeinflusst. Außerdem prüfen sie die Auswirkungen auf Kinder im Vergleich zu Erwachsenen. Insgesamt hat das untersuchte Modell eine Antwortvariable, die Steigerung des Blutdrucks, sowie zwei Faktoren, die möglicherweise Änderungen am Ergebnis erklären, nämlich die Altersgruppe (Erwachsene im Vergleich zu Kindern) und die Verabreichungsdosis (niedrig gegenüber hoch). Es könnte sein, dass die Verabreichungsdosis den Blutdruck von Erwachsenen anders beeinflusst als den Blutdruck bei Kindern. Dieser Modelltyp wird als *Zweifache Varianzanalyse* bezeichnet, mit einem möglichen Interaktionseffekt zwischen den beiden Faktoren (Altersgruppe und Verabreichungsdosis). Weitere Informationen finden Sie in Kapitel 11.

Wenn Statistiker eine Zweifache Varianzanalyse durchführen, zeichnen sie als Erstes die Mittelwertergebnisse für alle verglichenen Gruppen auf und suchen nach Mustern. Man spricht auch von einem Interaktionsdiagramm. Abbildung 1.4 zeigt ein Interaktionsdiagramm für das Medikamentenszenario.

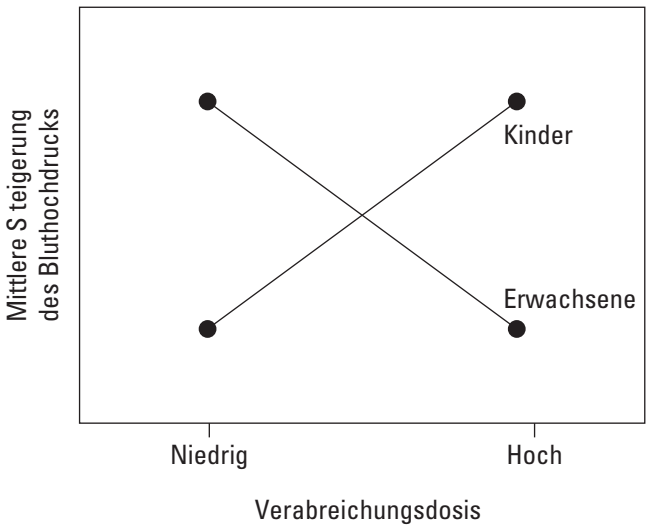

Abbildung 1.4: Interaktionsdiagramm

Wie Sie in Abbildung 1.4 erkennen, schneiden sich die Linien. An der Linie für die Kinder erkennen Sie, dass die mittlere Steigerung des Blutdrucks für geringe Verabreichungsdosen gering ist, aber für die hohe Verabreichungsdosis steigt. Die Blutdrucksteigerung nimmt zu. Bei Erwachsenen dagegen ist die Reaktion genau umgekehrt. Bei der geringen Verabreichungsdosis ist die mittlere Blutdrucksteigerung sehr hoch, aber für die hohe Verabreichungsdosis ist die Steigerung sehr gering. Würden die Ärzte die Studie nicht sowohl für Kinder als auch für Erwachsene durchführen, könnten die Ergebnisse der Studie sehr schädlich für Kinder sein, wenn die Ärzte später die Regeln für die Erwachsenen auch auf die Kinder anwenden. Dieses Beispiel zeigt, wie wichtig Interaktionseffekte sind.

Abbildung 1.5 zeigt eine Situation, in der es keine Interaktionseffekte für dieses Medikament gibt. Die Linien sind parallel, woran Sie erkennen, dass der mittlere Blutdruck bei einer

höheren Verabreichungsdosis des Medikaments sowohl für Erwachsene als auch für Kinder mehr steigt. Die Linie für die Erwachsenen ist weiter oben angesiedelt als die für Kinder, was bedeutet, dass die Steigerung des Blutdrucks für Erwachsene unabhängig von der Höhe der Verabreichungsdosis höher als die Steigerung des Blutdrucks für Kinder ist.

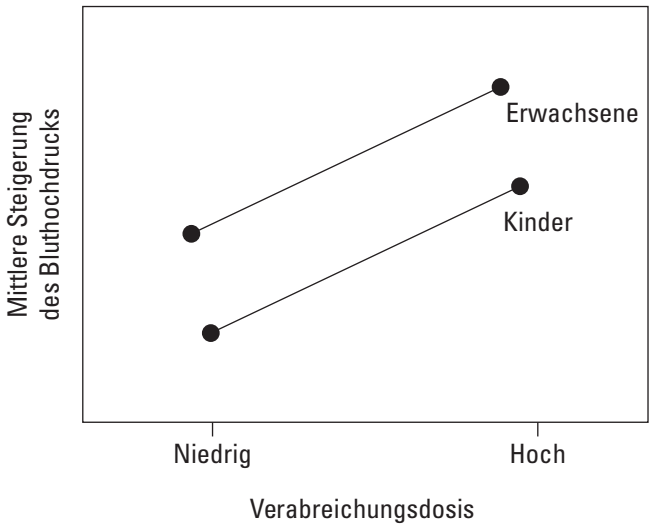

Abbildung 1.5: Keine Interaktionseffekt

Korrelation

Der Begriff *Korrelation* wird häufig falsch benutzt. Statistisch ausgedrückt gibt der Korrelationskoeffizient die Stärke und die Richtung der linearen Beziehung zwischen zwei quantitativen Variablen an (das sind Variablen, die nur Zählungen oder Messungen darstellen).

Das Wort *Korrelation* sollte nicht auf Beziehungen anderer Art angewendet werden. Beispielsweise ist es falsch zu sagen, dass es zwischen der Augenfarbe und der Haarfarbe eine Korrelation gibt. Diese Variablen können in einer Beziehung zueinander stehen, aber es handelt sich dabei nicht um quantitative Variablen, deshalb kann man ihre Beziehung zueinander nicht als Korrelation bezeichnen. (In diesem Fall würden Sie den Begriff Assoziation verwenden. In Kapitel 14 erfahren Sie, wie man auf eine Assoziation von zwei kategorialen Variablen testet.)

Zu einer Korrelation ist Folgendes zu sagen: Der Korrelationskoeffizient ist eine Zahl zwischen $-1{,}0$ und $+1{,}0$. $+1{,}0$ ist eine perfekte positive lineare Beziehung, mit anderen Worten, wenn Sie eine Variable erhöhen, erhöht sich die andere in perfekter Synchronisation. Ein Korrelationskoeffizient von $-1{,}0$ dagegen beschreibt eine perfekte negative lineare Beziehung zwischen den Variablen. Wenn eine Variable zunimmt, sinkt die andere in perfekter Synchronisation. Ein Korrelationskoeffizient von 0 bedeutet, dass es überhaupt keine lineare Beziehung zwischen den Variablen gibt. Die meisten Korrelationen in der realen Welt sind nicht genau $+1{,}0$, $-1{,}0$ oder 0 – sie liegen irgendwo dazwischen. Je näher eine Beziehung an $+1{,}0$ oder $-1{,}0$ liegt, desto stärker ist sie. Je näher sie an 0 liegt, desto schwächer ist sie.

Abbildung 1.6 zeigt ein Beispiel für ein Diagramm mit der Anzahl der Kaffees, die bei einem Football-Spiel in Buffalo, New York, verkauft wurden, ebenso wie die Lufttemperatur (in Fahrenheit) bei jedem Spiel. Diese Datenmenge scheint relativ gerade nach unten zu verlaufen, was eine negative Korrelation repräsentiert. Wenn Sie den Korrelationskoeffizienten berechnen, erhalten Sie den Wert −0,741. Dieser Wert besagt, dass die verkauften Kaffees eine relativ starke negative Beziehung zu der Temperatur bei dem Fußballspiel haben. Das ist sinnvoll, denn nur an Tagen, an denen niedrige Temperaturen herrschen, ist den Menschen kalt und sie wollen mehr Kaffee trinken. An Tagen, an denen die Temperatur höher ist, trinken die Menschen eher weniger Kaffee und vielleicht lieber kalte Getränke. Weitere Informationen über Korrelationen finden Sie in Kapitel 4, wo es um die Modellbildung geht.

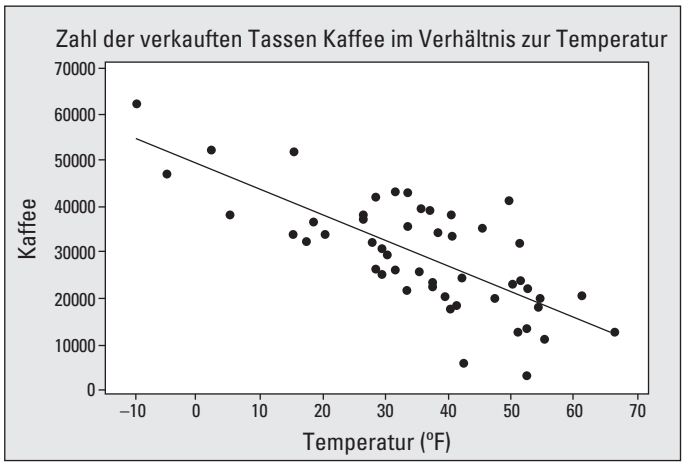

Abbildung 1.6: Bei unterschiedlichen Temperaturen verkaufter Kaffee bei einem Football-Spiel

Lineare Regression

Nachdem Sie festgestellt haben, dass zwei Variablen eine relativ starke lineare Beziehung haben, wollen Sie vielleicht versuchen, Vorhersagen für eine Variable abhängig vom Wert der anderen Variablen zu treffen. Wenn Sie beispielsweise wissen, dass zwischen dem verkauften Kaffee und der Lufttemperatur bei einem Football-Spiel eine relativ starke negative lineare Beziehung vorliegt, könnten Sie diese Information nutzen wollen, um nur anhand der Temperatur vorherzusagen, wie viel Kaffee für ein Spiel benötigt wird. Diese Methode, die am besten mit allen Datenpunkten übereinstimmende Gerade zu finden, wird auch als *lineare Regression* bezeichnet.

In dem Beispiel mit dem Kaffee und der Temperatur (siehe Abbildung 1.6) hat die am besten übereinstimmende Gerade die Gleichung $y = 49337 - 554 * x$, wobei x die Temperatur ist, und y die Anzahl der verkauften Kaffees. Wenn die Temperatur (x) also 0 Grad beträgt, können Sie erwarten, dass 49.337 Kaffees verkauft werden (auf diese Weise interpretieren Sie den y-Achsenabschnitt der Geraden). Um die Steigung dieser Geraden zu interpretieren, stellen Sie sich −554 als −554 dividiert durch 1 vor und wenden das ganz normale Höhe-durch-Länge-Konzept unter Verwendung von Kaffees und Temperaturgraden an. Bei

unserem Beispiel bedeutet das, dass für jedes Grad Temperatursteigerung erwartet werden kann, dass der Kaffeeverkauf um 554 sinkt. Diese Gerade können Sie heranziehen, um Vorhersagen für sinnvolle Temperaturwerte (x) zu treffen. Liegt die Temperatur beispielsweise bei kalten 20 Grad Fahrenheit, können Sie vorhersagen, dass die Anzahl der verkauften Kaffees bei etwa $49337 - 554 * 20 = 38257$ liegt.

Wenn Sie für die Vorhersage der Antwort nur eine Variable verwenden, wird die Regressionsmethode als *einfache lineare Regression* bezeichnet. (Ich wiederhole die Grundlagen der einfachen linearen Regression in Kapitel 4. Aber es gibt noch zahlreiche andere Arten von Regressionen dort draußen, von denen ich einige in diesem Buch vorstelle.)

Die meisten Forscher verwenden mehr als eine Variable zur Vorhersage, diese Technik wird multiple lineare Regression genannt. (Weitere Informationen über die multiple lineare Regression finden Sie in Kapitel 5.) Die multiple lineare Regression ist auf vielerlei Arten problematisch, weil einige Variablen, die Sie in dem Modell verwenden können, miteinander in Beziehung stehen können, weshalb sie überlappende Beiträge zu der Antwort leisten. Diese Möglichkeit der Überlappung macht ihre einzelnen Beiträge schwer nachzuvollziehen. Außerdem müssen Sie auf Interaktionseffekte achten, wenn Sie für die Vorhersage einer Antwort mehr als eine Variable verwenden.

Einfache und multiple lineare Regression gehen davon aus, dass die Antwortvariable (die abhängige Variable) quantitativ ist (das heißt, sie misst oder zählt etwas). Sie könnten jedoch auch daran interessiert sein, Vorhersagen über eine Variable zu treffen, für die es nur zwei Ergebnisse gibt: Ja oder Nein. Beispiele dafür sind etwa, ob ein bestimmtes Pferd ein Rennen gewinnt oder nicht, ob ein Baby ein Junge oder ein Mädchen sein wird oder ob ein Tropensturm ein Hochwasser verursacht oder nicht. Für diese Situationen ist eine andere Regression erforderlich, die so genannte *logistische Regression* (weitere Informationen dazu finden Sie in Kapitel 8).

Und schließlich könnten Sie auch noch daran interessiert sein, ein Modell zu erstellen, für das eine Gerade nicht geeignet ist. Angenommen, Sie wollen anhand der Geschwindigkeit des Autos die Kilometer pro Liter vorhersagen. Während hohe Geschwindigkeiten geringe Kilometerzahlen pro Liter ergeben, können auch niedrige Geschwindigkeiten geringe Kilometerzahlen pro Liter ergeben. Die Beziehung zwischen Geschwindigkeit und Kilometer/Liter folgt also einer Parabel (in diesem Fall einer auf dem Kopf stehenden Kurve). Diese Art Beziehung wird auch als quadratische Beziehung bezeichnet. Allgemein ausgedrückt, Beziehungen, die keiner Geraden folgen, sind *nichtlineare Beziehungen*, und die Techniken für den Umgang mit diesen Situationen werden (Überraschung!) als *nichtlineare Regression* bezeichnet. Weitere Informationen über diese Technik finden Sie in Kapitel 7.

Chi-Quadrat-Tests

Korrelations- und Regressionstechniken gehen davon aus, dass die am detailliertesten untersuchte Variable (die Antwortvariable) quantitativ ist. Das bedeutet, die Variable misst oder zählt irgendetwas. Es gibt jedoch zahlreiche Situationen, wo die untersuchten Daten nicht quantitativ, sondern vielmehr qualitativ sind. Mit anderen Worten, die eigentlichen Daten stellen Kategorien dar, und keine Messungen oder Zähler.

Stellen Sie sich vor, Sie wollen die Meinungen über die Politik der Kanzlerin abhängig von der politischen Orientierung vergleichen. Angenommen, es ist Angela Merkel von der CDU, und Sie wählen eine zufällige Stichprobe von 150 CDU-Wählern, 150 SPD-Wählern und 150 sonstigen Wählern aus, um ihre Meinung über die Kanzlerin zu ermitteln. Die Daten könnten wie in Tabelle 1.2 gezeigt aussehen.

	Zustimmung	Neutral	Ablehnung
CDU-Wähler	100	40	10
SPD-Wähler	40	10	100
sonstige	50	50	50

Tabelle 1.2: Meinungen zu der (CDU-)Kanzlerin nach politischer Richtung

Wenn Sie betrachten, wie sich die Zahlen in den verschiedenen Zeilen in Tabelle 1.2 verteilen, könnten Sie etwas vermuten. Es hat den Anschein, dass die CDU-Wähler dazu tendieren, der Kanzlerin zuzustimmen, während die SPD-Wähler sie ablehnen und die sonstigen Wähler in der Mitte gespalten sind. (So viel zum Geist des mündigen Bürgers ...)

Gilt nun diese Assoziation, die Sie für die Datenmenge dieser Stichprobe von 450 Personen festgestellt haben, für die gesamte Population? Um diese Frage beantworten zu können, müssen Sie einen Hypothesentest durchführen. Und zwar nicht irgendeinen Hypothesentest – einen *Chi-Quadrat-Test auf Unabhängigkeit*. Sie testen, ob die beiden qualitativen Variablen, politische Orientierung und Meinungen zur Kanzlerin, miteinander in Beziehung stehen oder nicht. Wenn sie in einer Beziehung stehen, gelten die Variablen als nicht unabhängig. Stehen sie in keiner Beziehung, sind die Variablen unabhängig.

Ein Chi-Quadrat-Test erledigt im Grunde genommen Folgendes: Er ermittelt die Anzahl der Werte, die Sie in jeder Zelle der Tabelle erwarten, wenn die Variablen unabhängig sind (diese Werte werden brillanterweise als *erwarteter Zellenwert* bezeichnet). Der Chi-Quadrat-Test vergleicht dann diese erwarteten Zellenwerte mit dem, was in den Daten tatsächlich vorhanden war (als *beobachtete Zellenwerte* bezeichnet), und vergleicht sie in einer Chi-Quadrat-Statistik miteinander (siehe Kapitel 14).

Wenn die Chi-Quadrat-Test-Statistik sehr groß ist, werden Sie sehr wahrscheinlich eine Beziehung zwischen den beiden Variablen finden, weil die Gesamtdifferenzen zwischen den beobachteten und den erwarteten Zellenwerten groß sind. Mit anderen Worten, die Variablen sind nicht unabhängig, und Sie können die beobachteten Zellenwerte betrachten, um die erkannte Beziehung zu beschreiben. Ist die Chi-Quadrat-Test-Statistik klein, können Sie nicht schließen, dass Sie eine Beziehung gefunden haben, und die beiden Variablen sind unabhängig.

Im Fall der politischen Orientierung und den Meinungen zur Kanzlerin ist die Chi-Quadrat-Test-Statistik riesig, und Sie können daraus schließen, dass irgendwo eine Beziehung vorliegt. Sie können sagen, dass in der Population die CDU-Wähler dazu tendieren, die Kanzlerin zu unterstützen, dass die SPD-Wähler dazu tendieren, Opposition gegen die Kanzlerin zu machen, und dass die sonstigen Wähler in der Mitte gespalten sind. (Weitere

Informationen darüber, wie Sie die erwarteten Werte finden und den Chi-Quadrat-Test durchführen, finden Sie in Kapitel 14.)

Sie können den Chi-Quadrat-Test auch anwenden, um zu prüfen, ob Ihre Ansichten darüber, wie viel Prozent jeder Gruppe in eine bestimmte Kategorie fallen, zutreffen oder nicht. Können Sie beispielsweise schätzen, welcher Prozentsatz an M&Ms in die einzelnen Farbkategorien fällt? Mehr zu diesen Chi-Quadrat-Varianten und zur M&M-Frage finden Sie in Kapitel 15.

Nichtparametrische Statistik

Die nichtparametrische Statistik ist ein ganzer Bereich innerhalb der Statistik, der Analysetechniken bereitstellt, die verwendet werden können, wenn die Bedingungen für die traditionelleren und häufiger verwendeten Methoden nicht zutreffen. Um beispielsweise einen *t*-Test durchzuführen, müssen die Daten aus einer Population mit Normalverteilung stammen (das heißt, sie müssen eine glockenförmige Kurve aufweisen). Um einen Hypothesentest für zwei Mittelwerte durchzuführen, müssen die Daten jeder Gruppe aus einer eigenen normalverteilten Population stammen. Tatsächlich geben die meisten aller häufig verwendeten Datenanalyseverfahren Bedingungen vor, die erfüllt sein müssen, wenn die Analyseverfahren verwendet werden sollen.

Das Problem bei diesen Bedingungen ist, dass die Menschen häufig vergessen oder sich einfach nicht darum kümmern, diese Bedingungen zu überprüfen. Wenn die Bedingungen nicht erfüllt sind, ist die ganze Analyse ungültig, und der Forscher weiß es noch nicht einmal. Es kann auch vorkommen, dass jemand erkennt, dass die Bedingungen nicht erfüllt sind, und trotzdem weitermacht und die Verfahrensweisen trotzdem anwendet (mehr zu diesem Fehler finden Sie im Abschnitt *(Daten-)Fischen verboten*) in diesem Kapitel).

Viele der traditionellen Methoden werden von den Statistikern als *robust* bezeichnet, was die Verletzung ihrer Voraussetzungen betrifft (eine andere Ausdrucksweise dafür, dass sie relativ nachsichtig sind), aber Sie können es nicht übertreiben. Die Anwendung einer statistischen Verfahrensweise, die nicht für die Situation geeignet ist, verursacht große Probleme in Hinblick auf Korrektheit der Schlussfolgerungen und die Glaubwürdigkeit des Forschers.

Aber keine Angst: Die nichtparametrische Statistik eilt Ihnen zu Hilfe. Wenn die Voraussetzungen für ein Verfahren der Datenanalyse nicht erfüllt sind, besteht durchaus die Wahrscheinlichkeit, dass es eine äquivalente nichtparametrische Verfahrensweise gibt, die schon in den Startlöchern steht. Und das Gute daran ist, dass sie im Allgemeinen relativ zahm in Hinblick auf die Formeln ist und die meisten Statistik-Softwarepakete sie genauso einfach durchführen wie die regulären (parametrischen) Verfahrensweisen.

Vor der Durchführung einer Datenanalyse führen die Statistik-Softwarepakete nicht automatisch eine Überprüfung der Voraussetzungen durch. Es bleibt dem Benutzer überlassen, alle zutreffenden Bedingungen zu überprüfen, und, falls sie ernsthaft verletzt sind, einen anderen Weg einzuschlagen. Häufig ist eine nichtparametrische Verfahrensweise die Lösung. Weitere Informationen über die verschiedenen nichtparametrischen Verfahrensweisen finden Sie in den Kapiteln 16 bis 19.