Kapitel 1
Die Grundlagen

Der schwierigste Teil jeder statistischen Arbeit ist der Anfang – und eine der größten Herausforderungen ist die Wahl der richtigen statistischen Analyse. Die Wahl hängt von der Art Ihrer Daten und der jeweiligen Fragestellung ab. Leider gibt es dabei keinen Ersatz für Erfahrung. Um zu wissen, was zu tun ist, muss man etwas bereits mehrfach richtig gemacht haben.

Entscheidend sind zwei Punkte: Welche Art von **Zielvariable** liegt vor und wie ist die Beschaffenheit der **erklärenden Variable**? Die Zielvariable ist das, woran Sie arbeiten – die Variable, die es zu erklären gilt. Diese Variable wird auf der *y*-Achse eines Diagramms (der Ordinate) abgetragen. Die erklärende Variable wird auf der *x*-Achse (der Abszisse) des Diagramms abgetragen. Sie sind daran interessiert, inwieweit eine Variation der Zielvariable mit einer Variation der erklärenden Variablen zusammenhängt. Ein kontinuierliches Maß ist eine Variable wie Höhe oder Gewicht, die jeden realen zahlenmäßigen Wert annehmen kann. Eine kategoriale Variable ist ein **Faktor** mit zwei oder mehr **Ausprägungen**: Geschlecht ist ein Faktor mit zwei Ausprägungen (männlich und weiblich) und ein Regenbogen kann ein Faktor mit sieben Ausprägungen sein (rot, orange, gelb, grün, blau, indigoblau, violett).

Es ist deshalb entscheidend, dass Sie wissen:

- welche Ihrer Variablen die Zielvariable ist,
- welche die erklärenden Variablen sind,
- ob die erklärenden Variablen kontinuierlich oder kategorial sind oder eine Mischung aus beidem,
- welche Art von Zielvariable Sie haben – ist es ein kontinuierliches Maß, eine Anzahl, ein Verhältnis, ein Todeszeitpunkt oder eine Kategorie?

Diese einfachen Schlüssel führen zur geeigneten statistischen Methode.

1. Die erklärenden Variablen
   (a) Alle erklärenden Variablen sind kontinuierlich — **Regression**
   (b) Alle erklärenden Variablen sind kategorial — **Varianzanalyse (ANOVA)**
   (c) Erklärende Variablen sind kontinuierlich und kategorial — **Kovarianzanalyse (ANCOVA)**
2. Die Zielvariable
   (a) Kontinuierlich — **Normalregression (ANOVA oder ANCOVA)**
   (b) Verhältniszahl — **Logistische Regression**
   (c) Anzahl — **Log-Lineares Modell**
   (d) Binär — **Binäre logistische Analyse**
   (e) Todeszeitpunkt — **Überlebensanalyse**

Es gibt ein paar entscheidende Punkte, die von Anfang an verstanden werden müssen. Wir behandeln diese, bevor wir in Details zu den verschiedenen Arten statistischer Modelle einsteigen.

Alles variiert

Wenn Sie dieselbe Sache zweimal messen, werden Sie zwei unterschiedliche Ergebnisse erhalten. Wenn Sie dieselbe Sache zu verschiedenen Zeitpunkten messen, werden Sie unterschiedliche Ergebnisse erhalten, weil die Sache gealtert ist. Wenn Sie verschiedene Individuen messen, werden sich diese genetisch und umweltbedingt unterscheiden (Anlage und Umwelt). Heterogenität ist universell: Räumliche Heterogenität bedeutet, dass sich Orte immer unterscheiden, und zeitliche Heterogenität bedeutet, dass sich Zeiten stets unterscheiden.

Da alles variiert, ist es nicht von Interesse, genau das herauszufinden. Wir brauchen einen Weg, um zwischen einer Variation, die wissenschaftlich interessant ist, und einer Variation, die nur eine Heterogenität des Hintergrunds widerspiegelt, zu unterscheiden. Dafür brauchen wir die Statistik. Und genau darum geht es in diesem Buch.

Es geht um die Menge an Variation, deren rein zufälliges Auftreten wir erwarten würden, wenn nichts wissenschaftlich Interessantes geschieht. Wenn wir größere Differenzen messen, als wir durch Zufall erwarten, bezeichnen wir das Ergebnis als statistisch signifikant. Wenn wir nicht mehr Variation messen, als wir durch Zufall bedingt erwarten, bezeichnen wir unser Ergebnis als statistisch nicht signifikant. Das bedeutet jedoch keineswegs, dass das Ergebnis nicht wichtig sei. Nicht signifikante Unterschiede in der Le-

bensspanne eines Menschen bei der Behandlung mit zwei verschiedenen Medikamenten können extrem wichtig sein (insbesondere wenn es sich bei dem Patienten um Sie handelt). Nicht signifikant bedeutet nicht dasselbe wie »nicht unterschiedlich«. Der Mangel an Signifikanz kann einfach an der Tatsache liegen, dass unsere Reproduzierbarkeit zu niedrig ist.

Andererseits wollen wir wissen, ob es wirklich keinen Zusammenhang gibt. Es macht das Leben einfacher, wenn wir halbwegs sicher sein können, dass es keine Beziehung zwischen y und x gibt. Einige Studenten denken, dass »nur ein signifikantes Ergebnis ein gutes Ergebnis« sei. Sie halten ihre Untersuchung für gescheitert, wenn sie aufzeigt, dass »A keine signifikante Wirkung auf B hat«. Diese Denkweise ist zwar eine verständliche Schwäche der menschlichen Natur, aber nicht wissenschaftlich. Der Punkt ist doch, dass wir die Wahrheit wissen wollen, wie sie auch beschaffen sein mag. Und wir sollten versuchen, uns nicht zu viele Gedanken darüber zu machen, was am Ende dabei herauskommt. Das ist keine amoralische Haltung, sonderlich lediglich die Art und Weise, wie Wissenschaft am besten funktioniert. Natürlich ist es hoffnungslos idealistisch, so zu tun, als würden sich Wissenschaftler tatsächlich so verhalten. Wissenschaftler hoffen zumeist inständig, dass sich ein bestimmtes experimentelles Ergebnis als statistisch signifikant erweisen wird, sodass sie einen Beitrag in der Zeitschrift *Nature* veröffentlichen können und gefördert werden. Aber das rechtfertigt es trotzdem nicht.

Signifikanz

Was meinen wir, wenn wir sagen, dass ein Ergebnis signifikant sei? Die üblichen Definitionen im Wörterbuch besagen: »eine Bedeutung haben oder befördern« oder »aussagekräftig sein; eine tiefere oder ungesagte Bedeutung nahelegen oder implizieren«. In der Statistik verstehen wir darunter jedoch etwas Spezifisches. Wir gehen davon aus, dass ein »Ergebnis höchstwahrscheinlich nicht zufällig aufgetreten ist«. Das zufällige Auftreten eines Ergebnisses halten wir »insbesondere dann für unwahrscheinlich, wenn sich die Nullhypothese als wahr erweist«. Demnach gibt es zwei Elemente: Wir müssen uns darüber klar sein, was wir mit »unwahrscheinlich« meinen und was wir unter der »Nullhypothese« verstehen. Unter Statistikern herrscht Übereinstimmung darüber, was »unwahrscheinlich« bedeutet. Sie sagen, dass ein Ereignis unwahrscheinlich ist, wenn es in weniger als 5 Prozent der

Gesamtzeit auftritt. Im Allgemeinen sagt die »Nullhypothese«, dass »nichts geschieht« – im Unterschied dazu, dass »etwas *geschieht*«.

Gute und schlechte Hypothesen

Karl Popper wies als Erster darauf hin, dass eine Hypothese dann gut ist, wenn sie **verworfen** werden kann. Er vertrat den Standpunkt, dass *eine gute Hypothese eine falsifizierbare* **Hypothese** ist. Betrachten Sie die folgenden zwei Aussagen.

1. Es gibt Geier im örtlichen Park.
2. Es gibt keine Geier im örtlichen Park.

Beide Aussagen beinhalten dieselbe Idee, aber eine ist widerlegbar und die andere nicht. Fragen Sie sich selbst, wie Sie Option 1 widerlegen würden. Sie gehen in den Park und suchen nach Geiern, aber Sie sehen keinen einzigen. Natürlich bedeutet das nicht, dass es keine gibt. Sie könnten sich vor Ihnen versteckt haben. Egal wie lange oder intensiv Sie schauen, Sie werden die Hypothese nicht widerlegen können. Alles, was Sie sagen können, ist: »Ich habe keinen einzigen Geier gesehen.« Eine der wichtigsten wissenschaftlichen Ansichten ist, dass **das Nichtvorhandensein von Beweisen kein Beweis für das Nichtvorhandensein** ist. Option 2 ist vollkommen anders. Sie verwerfen Hypothese 2, wenn Sie das erste Mal einen Geier im Park sehen. Bis zu dem Zeitpunkt, an dem Sie den ersten Geier im Park sehen, arbeiten Sie mit der Annahme, dass die Hypothese richtig ist. Aber wenn Sie einen Geier sehen, entpuppt sich die Hypothese als eindeutig falsch und Sie verwerfen sie.

Nullhypothesen

Die Nullhypothese besagt, dass »nichts geschieht«. Wenn wir zum Beispiel zwei Stichprobenmittelwerte vergleichen, lautet die Nullhypothese, dass die Mittelwerte der zwei Proben dieselben sind. Auch hier gilt: Beim Arbeiten mit einer Kurve von y gegen x in einer Regressionsanalyse ist die Nullhypothese, dass die Steigung der Beziehung gleich null ist, das heißt: y ist keine Funktion von x – oder anders ausgedrückt: y ist unabhängig von x. Der wesentliche Punkt ist, dass die Nullhypothese falsifizierbar ist. Wir weisen die Nullhypothese zurück, wenn unsere Daten zeigen, dass die Nullhypothese hinreichend unwahrscheinlich ist.

p-Werte

Ein *p-Wert* ist eine Abschätzung der Wahrscheinlichkeit, dass ein bestimmtes Ergebnis – oder ein extremeres Ergebnis als das beobachtete – zufällig aufgetreten sein könnte, wenn die Nullhypothese richtig wäre. Kurz gesagt, der *p-Wert* ist ein Maß für die Plausibilität der Nullhypothese. Wenn etwas wohl kaum zufällig aufgetreten sein kann, sagen wir, dass es statistisch signifikant ist, beispielsweise $p > 0.001$[1]. Vergleicht man zum Beispiel zwei Stichprobenmittelwerte, bei denen die Nullhypothese davon ausgeht, dass die Mittelwerte identisch sind, bedeutet ein niedriger *p-Wert*, dass die Hypothese wohl kaum richtig sein wird und der Unterschied statistisch signifikant ist. Ein hoher *p-Wert* (zum Beispiel $p = 0.23$) bedeutet, dass es keine zwingenden Beweise gibt, auf deren Basis die Nullhypothese zurückgewiesen werden kann. Natürlich sind die Aussagen »Wir weisen die Nullhypothese nicht zurück« und »Die Nullhypothese ist richtig« zwei verschiedene Dinge. Zum Beispiel könnten wir nicht in der Lage sein, eine falsche Nullhypothese zurückzuweisen, weil die Stichprobengröße zu klein oder unser Messfehler zu groß war. Folglich sind *p-Werte* interessant, aber sie erzählen nicht die ganze Geschichte; Wirkungsgrößen und Stichprobengrößen sind für das Ziehen von Schlüssen ebenso wichtig.

Interpretation

Es sollte an diesem Punkt klar sein, dass wir zwei Arten von Fehlern bei der Interpretation unserer statistischen Modelle machen können:

- Wir können die Nullhypothese zurückweisen, obwohl sie richtig ist, oder
- wir können die Nullhypothese akzeptieren, obwohl sie falsch ist.

Diese Fehler werden als **Typ I** und **Typ II** bezeichnet. Angenommen, wir würden die tatsächliche Sachlage kennen (was wir natürlich selten tun), dann sieht das in tabellarischer Form wie folgt aus:

1 Wichtige Anmerkung: Im Folgenden wird statt des in der deutschen Schreibweise für Dezimalstellen üblichen Kommas durchgängig der im Englischen übliche Punkt verwendet, da auch die Eingaben in R in der englischen Schreibweise erfolgen müssen. Kommas in den Befehlen müssen auch als solche eingegeben werden, da sie eine andere Funktion erfüllen, z. B. Abtrennung von Zahlen.

	Tatsächliche Sachlage	
Nullhypothese	*Wahr*	*Falsch*
Akzeptiert	Richtige Entscheidung	Typ II
Zurückgewiesen	Typ I	Richtige Entscheidung

Statistisches Modellieren

Das Ziel besteht darin, die Werte der Parameter in einem spezifischen Modell zu bestimmen, um das Modell für die Daten maßzuschneidern. Die Daten sind unantastbar. Sie sagen uns, was unter den gegebenen Umständen wirklich geschah. Häufig wird fälschlicherweise gesagt, dass »die Daten an das Modell angepasst wurden«, als ob die Daten etwas Flexibles wären und wir ein klares Bild der Struktur des Modells hätten. Im Gegenteil – wir suchen das minimal entsprechende Modell, um die Daten zu beschreiben. Das Modell wird an die Daten angepasst, nicht andersherum. Das beste Modell ist das Modell, das die geringste nicht erklärte Variation produziert (die **minimale restliche Abweichung**), vorbehaltlich der Einschränkung, dass alle Parameter im Modell statistisch signifikant sein sollten.

Sie müssen das Modell bestimmen. Es verkörpert Ihr mechanistisches Verständnis der beteiligten Faktoren und die Art, wie diese mit der Zielvariable zusammenhängen. Wir wollen, dass das Modell wegen des Sparsamkeitsprinzips **minimal** ist und dass es **angemessen** ist, weil es keinen Sinn macht, ein unzulängliches Modell zu erstellen, das einen bedeutenden Bruchteil der Schwankung in den Daten nicht beschreibt. Es ist wichtig, zu verstehen, dass es nicht nur ein Modell gibt; dies ist einer der verbreiteten impliziten Fehler bei der traditionellen Regression und Varianzanalyse (Anova), wo dieselben Modelle häufig kritiklos immer wieder verwandt werden. Unter den meisten Umständen wird es eine Vielzahl von verschiedenen, mehr oder weniger plausiblen Modellen geben, die für jeden gegebenen Satz von Daten geeignet sein könnten. Ein Teil der Aufgabe der Datenanalyse ist es, zu bestimmen, welche der – falls überhaupt – möglichen Modelle passend sind und aus diesem Satz passender Modelle das minimal angemessene Modell herauszusuchen. Manchmal gibt es nicht das eine, beste Modell, sondern eine Reihe verschiedener Modelle beschreibt die Daten gleich gut (oder gleich schlecht, wenn die Varianz groß ist).

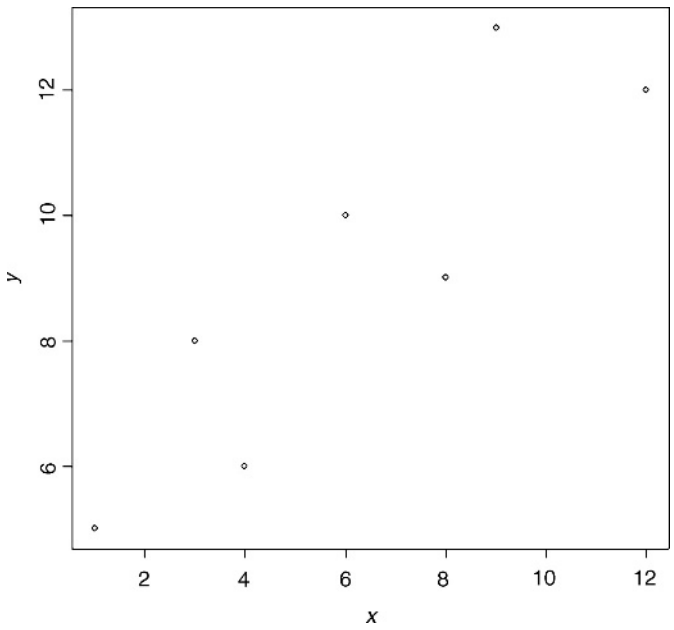
Maximale Wahrscheinlichkeit

Was genau meinen wir, wenn wir sagen, dass die Parameterwerte »die beste Anpassung des Modells an die Daten« erzeugen sollten? Wir gehen davon aus, dass unsere Techniken zu **unvoreingenommenen, Varianz mindernden Schätzwerten** führen sollten. Wir definieren »beste« in Bezug auf die **maximale Wahrscheinlichkeit**. Diese Vorstellung ist Ihnen vermutlich nicht vertraut, deshalb lohnt es sich, etwas Zeit darauf zu verwenden, um ein Gefühl dafür zu bekommen. Es funktioniert wie folgt:

- Angesichts der gegebenen Daten
- und angesichts unserer Modellwahl:
- Welche Werte für die Parameter dieses Modells machen die beobachteten Daten am wahrscheinlichsten?

Hierzu ein Beispiel: *y* ist die Zielvariable und *x* ist die erklärende Variable. Weil sowohl *x* als auch *y* kontinuierliche Variable sind, ist das passende Modell eine Regression.

```
x<-c (1,3,4,6,8,9,12)
y<-c(5,8,6,10, 9, 13, 12)
plot(x,y)
```

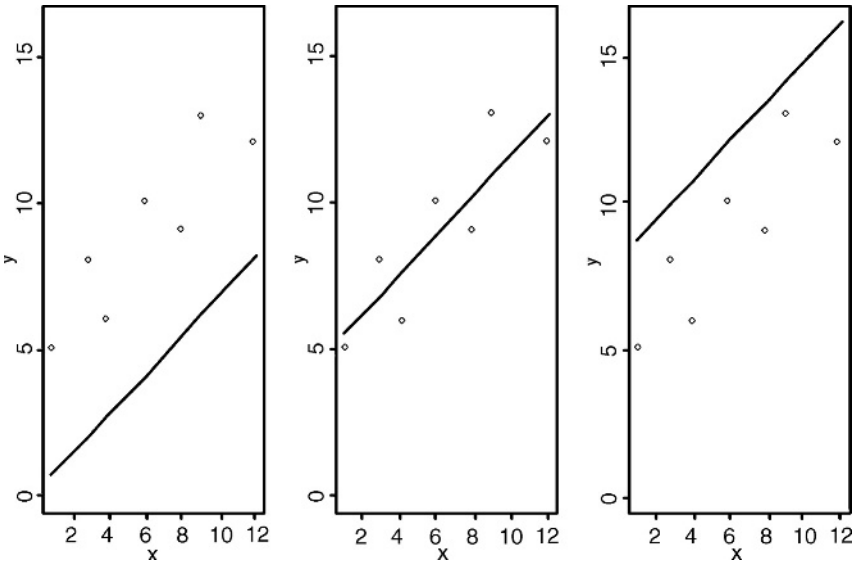


Jetzt müssen wir aus der enormen Bandbreite von möglichen verfügbaren Modellen ein Regressionsmodell auswählen, um diese Daten zu beschreiben. Wir wollen das einfachste Modell wählen, die Gerade:

$$y = a + bx.$$

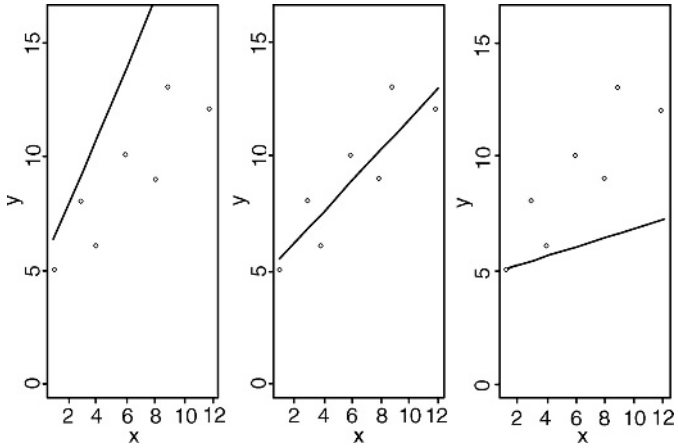
Das ist ein Zwei-Parameter-Modell; der erste Parameter, a, ist der Achsenabschnitt (der Wert von y, wenn x gleich 0 ist) und der zweite, b, die Steigung (die Änderung von y in Abhängigkeit einer Änderung von x um eine Einheit). Die Zielvariable y ist eine lineare Funktion der erklärenden Variable x. Nehmen wir jetzt an, wir wüssten, dass die Steigung 0.68 beträgt, dann kann die Frage der maximalen Wahrscheinlichkeit auf den Abschnitt a angewandt werden.

Wenn der Achsenabschnitt 0 wäre (linkes Diagramm unten), wären die Daten dann wahrscheinlich? Die Antwort lautet natürlich nein. Wenn der Abschnitt 8 wäre (rechtes Diagramm), würden die Daten dann wahrscheinlich? Wieder lautet die Antwort nein. Die maximale Wahrscheinlichkeitsschätzung des Abschnitts wird im mittleren Diagramm gezeigt (sein Wert erweist sich als 4.827).



Wir könnten eine ähnliche Diskussion über die Steigung führen. Angenommen, wir wüssten, dass der Achsenabschnitt 4.827 wäre, wären die

Daten dann wahrscheinlich, wenn das Diagramm eine Steigung von 1.5 hätte (linkes Diagramm unten)?



Die Antwort lautet natürlich nein. Wie steht es mit einer Steigung von 0.2 (rechtes Diagramm)? Wieder sind die Daten unwahrscheinlich, wenn das Diagramm eine so sanfte Steigung hat. Die maximale Wahrscheinlichkeit der Daten angesichts des Modells wird mit einer Steigung von 0.679 (mittleres Diagramm) erreicht. So geht man natürlich nicht vor, aber es macht deutlich, dass wir das Modell auf der Grundlage beurteilen, wie wahrscheinlich die Daten wären, wenn das Modell richtig wäre. In der Praxis werden selbstverständlich beide Parameter gleichzeitig geschätzt.

Versuchsanordnung

Es gibt nur zwei Schlüsselkonzepte:

- Wiederholung und
- Randomisierung.

Man kann wiederholen, um die Reliabilität zu erhöhen. Man kann randomisieren, um die Verzerrung zu beeinflussen. Wenn Sie sorgfältig wiederholen und ordentlich randomisieren, kann nicht viel schiefgehen.

Es gibt eine Reihe anderer Aspekte, deren Beherrschung die Wahrscheinlichkeit erhöht, dass Sie Ihre Daten auf die richtige und nicht etwa auf die falsche Weise analysieren:

- das Sparsamkeitsprinzip
- die Wirksamkeit eines statistischen Tests
- Kontrollen
- das Erkennen von Pseudowiederholungen und das Wissen, wie damit zu verfahren ist
- der Unterschied zwischen experimentellen und beobachtbaren Daten (Nichtorthogonalität).

Es spielt keine große Rolle, wenn Sie nicht Ihre eigene fortgeschrittene statistische Analyse durchführen können. Wenn Ihr Versuch ordentlich konzipiert ist, werden Sie in den meisten Fällen jemanden finden, der Ihnen bei der Statistik hilft. Wenn Ihr Versuch jedoch nicht ordentlich konzipiert oder gründlich randomisiert ist oder angemessene Kontrollen vermissen lässt, dann können Sie in Statistik noch so gut sein, einige (wenn nicht gar alle!) Ihrer Bemühungen werden dann umsonst sein. Auch noch so leistungsstarke Statistik kann eine schlechte Versuchsanordnung nicht in eine gute verwandeln. R ist zwar gut, aber nicht so gut.

Das Sparsamkeitsprinzip (Ockhams Rasiermesser)

Eines der wichtigsten Themen, die sich durch dieses Buch ziehen, betrifft die Modellvereinfachung. Das Sparsamkeitsprinzip wird dem Philosophen William von Ockham zugeschrieben, einem Hauptvertreter des Nominalismus. Ockham bestand darauf, dass bei Vorhandensein mehrerer gleich guter Erklärungen für ein Phänomen **die einfachste Theorie die richtige** sei. Dieses Prinzip erhielt den Namen Ockhams Rasiermesser, weil er die Erklärungen bis auf ein absolutes Minimum »herunterrasierte«. Bei der statistischen Modellierung bedeutet das Sparsamkeitsprinzip:

- Das Modell sollte so wenig Parameter wie möglich haben.
- Lineare Modelle sollten gegenüber nichtlinearen bevorzugt werden.
- Versuchsanordnungen, die auf wenigen Annahmen basieren, sind gegenüber jenen, die auf vielen Annahmen beruhen, zu bevorzugen.
- Modelle sollten abgespeckt werden, bis sie *minimal passend* sind.
- Einfache Erklärungen sind gegenüber komplexen Erklärungen zu bevorzugen.

Der Prozess der Modellvereinfachung ist ein wesentlicher Teil der Hypothesenüberprüfung in R. Im Allgemeinen wird eine Variable nur dann in einem Modell beibehalten, wenn sie eine signifikante Zunahme der Abwei-

chung verursacht, sobald sie aus dem aktuellen Modell entfernt wird. Suche Einfachheit und misstraue ihr dann.

Wenn wir mit Feuereifer Modelle vereinfachen, müssen wir jedoch aufpassen, dass wir das Kind nicht mit dem Bade ausschütten. Einstein nahm eine für ihn typische subtile Modifizierung an Ockhams Rasiermesser vor. Er sagte: »Ein Modell sollte so einfach wie möglich sein. Aber nicht noch einfacher.«

Beobachtung, Theorie und Versuch

Zweifellos besteht der beste Weg, wissenschaftliche Probleme zu lösen, in einer wohlüberlegten Mischung aus Beobachtung, Theorie und Versuch. In den meisten Situationen gibt es jedoch Beschränkungen, welche Dinge getan und auf welche Weise sie getan werden können. Das bedeutet, dass einer oder mehrere Punkte dieser Trilogie geopfert werden müssen. Es gibt zum Beispiel jede Menge Fälle, in denen es ethisch oder logistisch unmöglich ist, manipulative Versuche durchzuführen. In diesen Fällen ist es doppelt wichtig, sicherzustellen, dass die statistische Analyse zu Schlussfolgerungen führt, die so kritisch und eindeutig wie möglich sind.

Kontrollen

Ohne Kontrollen keine Schlussfolgerungen.

Wiederholungen: Die *n*s rechtfertigen die Mittelwerte

Die Notwendigkeit der Wiederholung entsteht aus folgendem Grund: Wenn wir dieselbe Sache mit verschiedenen Individuen ausprobieren, erhalten wir unterschiedliche Ergebnisse. Die Gründe für die Heterogenität bei den Ergebnissen sind zahlreich und vielfältig (Erbgut, Alter, Geschlecht, Zustand, Geschichte, Substrat, Mikroklima und so weiter.) Die Wiederholungen sollen die Verlässlichkeit der Parameterschätzungen erhöhen und uns erlauben, die Variabilität zu quantifizieren, auf die wir stoßen. Um die Wiederholungen zu qualifizieren, gilt für die wiederholten Messungen:

- Sie müssen unabhängig voneinander sein.
- Sie dürfen nicht Bestandteil einer zeitlichen Serie sein (Daten, die an aufeinanderfolgenden Zeitpunkten am selben Ort gesammelt werden, sind nicht unabhängig voneinander).
- Sie dürfen nicht an einem Ort gruppiert sein (die Replikate anzuhäufen bedeutet, dass sie räumlich nicht unabhängig voneinander sind).
- Sie müssen sich in einem angemessenen räumlichen Umfang bewegen.

Idealerweise werden Blöcke mit jeweils einer Replik aus jedem Verfahren gebildet und jedes Verfahren wird in vielen verschiedenen Blocks wiederholt. Wiederholte Messungen (zum Beispiel vom selben Individuum) sind keine Replikate (dies ist vermutlich der häufigste Grund von Pseudowiederholung bei statistischen Arbeiten).

Wie viele Wiederholungen?

Die übliche Antwort lautet: »So viele, wie Sie erzeugen können.« Eine alternative Antwort ist: 30. Eine gute Faustregel besagt: Eine Stichprobe von 30 oder mehr ist eine große Stichprobe, aber eine Stichprobe von weniger als 30 ist eine kleine. Die Regel funktioniert nicht immer. Zum Beispiel ist 30 als Stichprobe bei einer Meinungsumfrage lächerlich klein. Unter anderen Umständen kann es jedoch unerschwinglich sein, einen Versuch 30 Mal zu wiederholen. Trotzdem ist diese Regel von hohem praktischem Nutzen, und wenn es Sie nur darauf bringt – während Sie Ihren Versuch mit 300 Wiederholungen entwerfen –, dass das vielleicht ein bisschen übertrieben sein könnte. Oder wenn Sie glauben, Sie könnten mit lediglich fünf Wiederholungen davonkommen.

Es gibt Wege, die nötige Anzahl für das Testen einer bestimmten Hypothese herauszufinden (diese werden im Folgenden noch erklärt). Manchmal wissen wir gar nichts oder wenig über die Varianz oder die Zielvariable, wenn wir einen Versuch planen. Dafür ist Erfahrung wichtig – und Pilotstudien. Diese sollten einen Hinweis auf die Varianz zwischen den Ausgangseinheiten liefern, bevor die Versuche aufgebaut werden, sowie auf die geschätzte Größenordnung der Ergebnisse, mit der bei dem Versuch zu rechnen ist. Manchmal kann es notwendig sein, den Umfang und die Komplexität der Versuche zu verringern und sich auf die zwangsläufig begrenzten Ressourcen von Arbeitskraft und Geld zu beschränken oder eine eindeutigere Antwort auf eine einfachere Frage zu erarbeiten. Es ist höchst ärgerlich, drei Jahre auf ein großes Experiment zu verwenden und dann festzustellen, dass

die Signifikanz des Ergebnisses lediglich bei $p = 0.08$ liegt. Eine Reduzierung der Anzahl von Versuchsaufbauten hätte vielleicht eine Steigerung der Wiederholung bis zu dem Punkt erlaubt, an dem dasselbe Ergebnis unzweideutig signifikant gewesen wäre.

Teststärke

Die Stärke (Power) eines Tests ist die Wahrscheinlichkeit des Verwerfens der Nullhypothese, wenn diese falsch ist. Das hat mit Fehlern vom Typ II zu tun: β ist die Wahrscheinlichkeit des Akzeptierens der Nullhypothese, wenn diese falsch ist. In einer idealen Welt würden wir β so klein wie möglich machen, aber es gibt einen Haken. Je weiter wir die Wahrscheinlichkeit reduzieren, einen Fehler vom Typ II zu begehen, desto größer wird die Wahrscheinlichkeit, einen Fehler vom Typ I zu begehen und die Nullhypothese zu verwerfen, obwohl sie zutrifft. Ein Kompromiss ist notwendig. Die meisten Statistiker arbeiten mit $\alpha = 0.05$ und $\beta = 0.2$. Unter den Standardvoraussetzungen wird die Stärke eines Tests als $1 - \beta = 0.8$ definiert. Damit werden die Stichprobengrößen berechnet, die nötig sind, um einen spezifizierten Unterschied aufzuspüren, wenn die Fehlerabweichung bekannt ist (oder geschätzt werden kann). Nehmen wir an, dass für eine einzelne Stichprobe die Größe des Unterschieds, den Sie aufspüren wollen, ∂ ist und die Varianz der Ergebnisse s^2 beträgt (wie beispielsweise aus einer Pilotstudie bekannt oder aus der Literatur entnommen), dann brauchen Sie n Replikate, um die Nullhypothese mit der Stärke = 80 Prozent zu verwerfen:

$$n \approx \frac{8 \times s^2}{\partial^2}$$

Dies ist eine vernünftige Faustregel, Sie sollten jedoch auf Nummer sicher gehen, indem Sie größere und nicht kleinere Stichproben als diese nehmen. Angenommen, der Mittelwert ist nahe an 20 und die Abweichung beträgt 10, wir wollen jedoch eine 10-prozentige Veränderung (zum Beispiel $\partial = \pm 2$) mit der Wahrscheinlichkeit 0.8 ermitteln, dann ist $n = 8 \times 10/2^2 = 20$.

Bei dem soeben betrachteten Fall ist die integrierte power.t.test-Funktion in Aktion. Wir müssen spezifizieren, dass der Typ »eine Stichprobe« ist und die Stärke, die wir erhalten wollen, 0.8 beträgt. Der aufzuspürende Unterschied (genannt Delta) beträgt 2.0 und die Standardabweichung (sd) ist $\sqrt{10}$.

```
power.t.test(type="one.sample",power=0.8,sd=sqrt(10),delta=2)

     one-sample t test power calculation

              n = 21.62146
          delta = 2
             sd = 3.162278
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

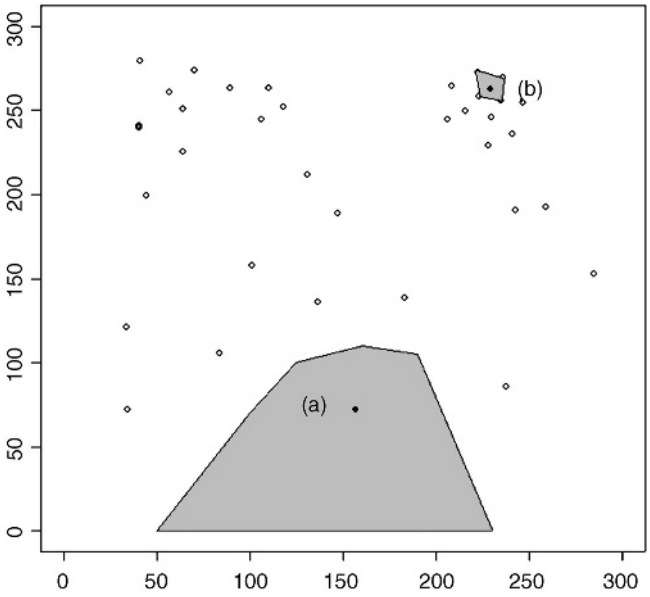
Weitere in R verfügbare Power-Funktionen sind unter anderem power.anova.test und power.prop.test.

Randomisierung

Randomisierung ist etwas, von dem jeder sagt, dass er es tut, aber kaum jemand tut es richtig. Nehmen wir ein einfaches Beispiel. Wie wähle ich aus einem Wald einen Baum aus, um an ihm die Photosyntheseleistung zu messen? Um das Ergebnis nicht zu beeinflussen, möchte ich den Baum nach dem Zufallsprinzip auswählen. Ich könnte zum Beispiel versucht sein, einen Baum auszuwählen, dessen Belaubung nicht zu hoch über dem Erdboden einsetzt, der nahe am Labor steht, der gesund aussieht, dessen Blätter nicht von Insekten befallen sind und so weiter. Ich überlasse es Ihnen, die möglichen Beeinflussungen bei der Wahl eines Baumes aufzulisten. Eine verbreitete Vorgehensweise bei der »zufälligen« Wahl des Baumes besteht darin, eine Karte des Waldes zu nehmen, sich ein Koordinatenpaar auszudenken (sagen wir mal vom Referenzpunkt aus 157 Meter Ost und 68 Meter Nord). Dann messen Sie die Koordinaten ab, und sobald Sie den entsprechenden Punkt im Wald erreicht haben, nehmen Sie den Baum, der diesen Koordinaten am nächsten steht. Aber ist das wirklich ein zufällig ausgewählter Baum?

Wenn er zufällig ausgewählt worden wäre, hätte er die exakt gleich große Chance, ausgewählt zu werden, wie jeder andere Baum in diesem Wald. Lassen Sie uns darüber nachdenken. Sehen Sie sich die folgende Abbildung an, die einen Plan der Verteilung der Bäume auf dem Gelände zeigt. Selbst wenn die Bäume ursprünglich in geraden Reihen angepflanzt wurden, so haben Baumfällungen und Heterogenität des Untergrunds schon bald zu einer haufenartigen Verteilung der Bäume geführt. Jetzt fragen Sie sich einmal, wie viele unterschiedliche Zufallspunkte zur Wahl eines bestimmten

Baumes führen. Beginnen Sie mit Baum (a). Dieser Baum würde bei allen Punkten gewählt werden, die in dem dunkel schraffierten Bereich liegen.



Jetzt betrachten Sie bitte Baum (b). Er wird nur ausgewählt, wenn der zufällige Punkt in den winzigen Bereich fällt, der diesen Baum umgibt. Baum (a) hat eine wesentlich größere Chance, ausgewählt zu werden, als Baum (b). Von daher ist der einem zufällig gewählten Punkt am nächsten stehende Baum kein zufällig ausgewählter Baum. In einem räumlich heterogenen Waldgebiet haben isoliert stehende Bäume und Bäume am Rand von Baumgruppen immer eine höhere Wahrscheinlichkeit, ausgewählt zu werden, als die Bäume innerhalb von Baumgruppen.

Die Lösung sieht folgendermaßen aus: Um einen Baum nach dem Zufallsprinzip auszuwählen, müssen alle Bäume in diesem Wald durchnummeriert werden (alle 24 683 oder wie viele es auch immer sind) und dann muss eine Zahl zwischen 1 und 24 683 aus einem Hut gezogen werden. Es gibt keine Alternative. Alles andere ist keine Randomisierung.

Und jetzt fragen Sie sich bitte, wie oft das wohl in der Praxis so gehandhabt wird, und Sie werden erkennen, was ich meine, wenn ich sage, dass Randomisierung ein klassisches Beispiel des »Tu, was ich sage – und nicht, was ich selbst tue« ist. Als Beispiel dafür, wie wichtig eine ordentliche Randomisierung sein kann, betrachten Sie bitte einmal das folgende Experiment, das

konzipiert wurde, um die Giftigkeit von fünf Kontaktgiften für Insekten zu testen, indem man Gruppen von Mehlkäfern den Chemikalien auf Filterpapier in Petrischalen aussetzte. Die Tiere liefen herum und nahmen das Gift mit ihren Beinen auf. Das Glas mit den Mehlkäfern wurde mitsamt dem Mehl auf ein großes Tablett gekippt und die Käfer wurden eingesammelt, als sie aus dem Mehl krabbelten. Die Tiere wurden den fünf Chemikalien der Reihe nach zugeteilt; vier gleiche Petrischalen wurden mit der ersten Chemikalie behandelt und in jede Petrischale wurden zehn Käfer gesetzt. Erkennen Sie die Quelle für Beeinflussung bei diesem Verfahren?

Es ist absolut plausibel, dass sich Mehlkäfer bei ihrem Aktivitätslevel unterscheiden (Unterschiede bei Geschlecht, Gewicht, Alter und so weiter). Die aktivsten Käfer kommen vielleicht als Erste aus dem Mehl heraus. All diese Käfer enden bei dem Versuch mit dem ersten Insektizid. Wenn wir schließlich Käfer für den Durchgang mit dem letzten Pestizid brauchen, müssen wir vielleicht schon im Mehl suchen, um die letzten verblieben *Tribolium*-Exemplare zu finden. Das spielt durchaus eine Rolle, denn die Menge an Pestizid, die die Käfer aufnehmen, hängt von deren Aktivitätslevel ab. Je aktiver die Käfer sind, desto mehr Chemikalien sammeln sie ein und desto höher ist die Wahrscheinlichkeit, dass sie sterben. Von daher wird der Fehler beim Randomisieren das Ergebnis zugunsten des ersten Insektizids beeinflussen, weil hierbei die aktivsten Käfer eingesetzt werden.

Wir hätten so vorgehen sollen: Füllen Sie $5 \times 4 = 20$ Petrischalen mit jeweils zehn Käfern, wobei die Käfer der Reihe nach abwechselnd in die Schalen gegeben werden. Ordnen Sie dann eine Versuchsanordnung (eines der fünf Pestizide) nach dem Zufallsprinzip jeweils einer Petrischale zu und setzen Sie die Käfer oben auf das vorbehandelte Filterpapier. Sie ordnen die Petrischalen ganz einfach den Versuchsanordnungen zu, indem Sie jeweils eine Nummer je Versuchsanordnung auf einen Zettel schreiben und alle 20 Zettel in einen Beutel geben. Dann ziehen Sie einen Zettel nach dem anderen. Dadurch erhalten Sie eine Versuchsanordnung, die der betreffenden Petrischale jeweils zugewiesen wird. Das mag sich geradezu lächerlich umständlich anhören, ist jedoch – glauben Sie mir – unerlässlich.

Alles andere ist eine faule Ausrede, die in Wahrheit besagt: »Ich gebe zu, dass ich nicht randomisiere, aber du kannst mir glauben, dass dadurch keine wesentliche Beeinflussung entsteht.« Ziehen Sie daraus Ihre eigenen Schlüsse.

Starke Inferenz

Eines der überzeugendsten Mittel, das zur Demonstration der Richtigkeit einer Idee zur Verfügung steht, ist die experimentelle Bestätigung einer als sorgfältig formulierte Hypothese getroffenen Vorhersage. Um eine starke Inferenz zu protokollieren, sind zwei Schritte wesentlich (Platt 1964):

- Formulieren Sie eine klare Hypothese und
- entwickeln Sie einen angemessenen Test.

Keins von beiden bringt etwas ohne das andere. Die Hypothese sollte zum Beispiel nicht zu Voraussagen führen, die sehr wahrscheinlich durch andere extrinsische Mittel auftreten können. Umgekehrt sollte der Test zweifelsfrei demonstrieren, ob die Hypothese richtig oder falsch ist.

Offenbar werden viele wissenschaftliche Experimente durchgeführt, ohne dass eine Hypothese dahintersteckt. Man will einfach nur sehen, was passiert. Diese Vorgehensweise mag ja im frühen Stadium einer Studie empfehlenswert sein. Solche Experimente neigen jedoch dazu, als Selbstzweck zu enden, weil es jede Menge plausibler Erklärungen für die Ergebnisse gibt. Ohne Nachdenken gibt es keine testbaren Vorhersagen; ohne testbare Vorhersagen wird es keinen experimentellen Einfallsreichtum geben; ohne experimentellen Einfallsreichtum wird die Kontrolle wahrscheinlich unzulänglich ausfallen; kurz gesagt: eine mehrdeutige Interpretation. Die Ergebnisse können auf unzählige plausible Gründe zurückzuführen sein. Die Natur hat keinerlei Ambitionen, von den Wissenschaftlern verstanden zu werden. Das müssen wir uns selbst erarbeiten. Ohne Wiederholung, Randomisierung und gute Kontrollen werden wir nur geringe Fortschritte machen.

Schwache Inferenz

Die Formulierung »schwache Inferenz« wird häufig (und oft abfällig) benutzt, um die Interpretation von beobachtenden Studien sowie die Analyse der sogenannten »natürlichen Experimente« zu beschreiben. Es ist albern, diese Daten herabzuwürdigen, weil sie oft die einzigen Daten sind, die wir haben. Das Ziel guter statistischer Analyse besteht im Erhalten maximaler Information über einen vorgegebenen Datensatz, wobei wir die beschränkte Aussagekraft der Daten im Hinterkopf behalten.

Natürliche Experimente finden dann statt, wenn ein Ereignis (das oftmals als ungewöhnliches Ereignis gesehen wird, obwohl es häufig keine Rechtfertigung dafür gibt, was das Ungewöhnliche daran sein soll) passiert, das wie

ein experimenteller Versuch gestaltet ist (ein Wirbelsturm fegt Teile eines Waldes um; ein Erdrutsch trägt Bodenschichten ab; ein Einbruch am Aktienmarkt lässt viele Menschen plötzlich verarmen und so weiter). Hairston (1989) sagt: »Die Erfordernis adäquater Kenntnisse über die Ausgangsbedingungen hat wichtige Auswirkungen auf die Validität vieler natürlicher Experimente. Insofern, als die ‚Experimente' erst dann erkannt werden, wenn sie abgeschlossen sind, oder frühestens wenn sie in Gang sind, kann man unmöglich sicher sein, was die Bedingungen angeht, die existierten, bevor das ‚Experiment' begann. Dann müssen Vermutungen hinsichtlich dieser Bedingungen angestellt werden und jegliche Schlussfolgerungen, die auf der Basis des natürlichen Experiments erlangt werden, sind dahingehend geschwächt, dass es sich nur um Hypothesen handelt, die auch als solche angegeben werden sollten.«

Wann ist es genug?

Idealerweise wird die Dauer eines Experiments im Vorfeld festgelegt, damit man nicht einer der beiden folgenden Versuchungen anheimfällt:

- das Experiment zu beenden, sobald ein zufriedenstellendes Ergebnis erreicht ist
- mit dem Experiment fortzufahren, bis das »richtige« Ergebnis erzielt wurde (der »Gregor-Mendel-Effekt«)

In der Praxis laufen die meisten Experimente wegen der Forschungsförderung wahrscheinlich zu kurz. Dieses Vorgehen ist vor allem in der Medizin und bei den Umweltwissenschaften gefährlich, weil sich die kurzfristigen Dynamiken, die sich nach Pulse-Experimenten zeigen, gänzlich von den langfristigen Dynamiken desselben Systems unterscheiden können. Nur durch langfristige Pulse- and Press-Experimente (press = langfristiger Druck, pulse = plötzliche Katastrophe) wird die ganze Bandbreite der Dynamiken verstanden. Der andere große Vorteil langfristiger Experimente besteht darin, dass eine große Bandbreite von Mustern (zum Beispiel »Arten von Jahren«) ausgetestet wird.

Pseudowiederholung

Pseudowiederholung tritt auf, wenn Sie die Daten so analysieren, als hätten Sie mehr Freiheitsgrade, als Ihnen tatsächlich zur Verfügung stehen. Es gibt zwei Arten von Pseudowiederholung:

- zeitweilige Pseudowiederholungen, zu denen wiederholte Messungen desselben Individuums gehören, und
- räumliche Pseudowiederholungen, zu denen wiederholte Messungen in unmittelbarer Nähe zählen.

Pseudowiederholung ist ein Problem, weil eine der wichtigsten Voraussetzungen der standardmäßigen statistischen Analyse die **Fehlerunabhängigkeit** ist. Wiederholte Messungen bei ein und demselben Individuum über die Zeit führen zu nichtunabhängigen Fehlern, da sich Besonderheiten des Individuums in allen Messungen widerspiegeln (die wiederholten Messungen stehen zeitlich in Beziehung miteinander). Aus unmittelbarer Nähe entnommene Proben werden nichtunabhängige Fehler mit sich bringen, weil alle Stichproben denselben Ort als Grundlage haben (die Erträge aus einem guten Landstück werden alle hoch und die aus einem schlechten Landstück werden alle niedrig sein).

Pseudowiederholung ist in der Regel leicht zu erkennen. Man muss sich lediglich die Frage stellen, wie viele Freiheitsgrade für Fehler das Experiment tatsächlich hat. Wenn ein Feldexperiment scheinbar viele Freiheitsgrade aufweist, ist es wahrscheinlich pseudorepliziert. Nehmen wir ein Beispiel aus der Schädlingsbekämpfung bei Pflanzen. Es gibt 20 Parzellen, zehn werden besprüht und zehn nicht. Auf jeder Parzelle stehen 50 Pflanzen. Jede Pflanze wird während der Wachstumsperiode fünfmal gemessen. Dieses Experiment generiert $20 \times 50 \times 5 = 5\,000$ Werte. Es gibt zwei Sprühbehandlungen, daher muss es einen Freiheitsgrad für das Sprühen und 4 998 Freiheitsgrade für Fehler geben. Oder? Addieren Sie die Wiederholungen in diesem Experiment. Wiederholte Messungen an denselben Pflanzen (bei fünf Stichprobengelegenheiten) sind sicher keine Replikate. Die 50 einzelnen Pflanzen innerhalb jeder Parzelle sind auch keine Wiederholungen. Der Grund dafür ist, dass die Bedingungen innerhalb jeder Parzelle ziemlich identisch sind, sodass alle 50 Pflanzen mehr oder weniger dieselben Voraussetzungen mitbringen, ungeachtet der Sprühbehandlung, die sie erhalten. Tatsächlich gibt es in diesem Experiment zehn Replikate. Es gibt zehn besprühte Parzellen und zehn nicht besprühte Parzellen. Und jede Parzelle wird lediglich ein unabhängiges Faktum für die Zielgröße hervorbringen (zum Beispiel die Proportion des Laubbereichs, der von Insekten verzehrt wird). Folglich gibt

es neun Freiheitsgrade innerhalb jeder Versuchsanordnung und $2 \times 9 = 18$ Freiheitsgrade für Irrtümer in dem ganzen Experiment. Es ist nicht schwierig, in der Literatur Beispiele für Pseudowiederholungen in dieser Größenordnung zu finden (Hurlbert 1984). Das Problem ist, dass es fälschlicherweise zur Berichterstattung über Unmengen signifikanter Ergebnisse führt (mit 4 998 Freiheitsgraden für Fehler ist es nahezu unmöglich, keine signifikanten Unterschiede zu haben). Die erste Fähigkeit, die der angehende Experimentator sich aneignen muss, ist die Fähigkeit, ein Experiment zu planen, das ordentlich repliziert ist.

Wenn Ihre Daten pseudorepliziert sind, gibt es eine Vielzahl von Dingen, die Sie tun können:

- Mitteln Sie die Pseudowiederholung weg und führen Sie Ihre statistische Analyse mit den Mittelwerten durch.
- Führen Sie für jede Zeitphase separate Analysen durch.
- Verwenden Sie eine ordentliche Zeitreihenanalyse oder Mixed-Effects-Modelle.

Ausgangsbedingungen

Viele ansonsten ausgezeichnete Experimente werden verdorben, weil keine Informationen über die Ausgangsbedingungen vorliegen. Woher sollen wir wissen, dass sich etwas verändert hat, wenn wir nicht wissen, wie es am Anfang beschaffen war? Häufig wird implizit davon ausgegangen, dass experimentelle Einheiten am Anfang des Experiments gleich waren. Das muss jedoch gezeigt und nicht einfach geglaubt werden. Eine der wichtigsten Verwendungen von Daten bei Ausgangswerten ist das Überprüfen der wirksamen Randomisierung. Zum Beispiel sollten Sie in der Lage sein, bei einem Experiment, bei dem es um Wachstum geht, mittels Ihrer statistischen Analyse nachzuweisen, dass die einzelnen Organismen zu Beginn des Experiments keine signifikanten Größenunterschiede aufwiesen. Ohne Messungen der Anfangsgröße ist es stets möglich, das Endergebnis auf Unterschiede bei den Ausgangswerten zurückzuführen. Ein weiterer Grund für das Messen der Ausgangswerte besteht darin, dass diese Information häufig genutzt werden kann, um die Lösung der abschließenden Analyse durch die Analyse der Kovarianz zu verbessern (siehe Kapitel 10).

Orthogonales Design und nichtorthogonale Beobachtungsdaten

Die Daten in diesem Buch entfallen auf zwei verschiedene Kategorien. Im Fall geplanter Experimente sind sämtliche Versuchskombinationen gleichmäßig repräsentiert und es fehlen, abgesehen von Durchführungsfehlern, keine Werte. Solche Experimente bezeichnet man als orthogonal. Im Fall beobachtender Studien dagegen haben wir keine Kontrolle über die Zahl der Individuen, über die wir Daten haben, oder über die Kombinationen von beobachteten Gegebenheiten. Viele der erklärenden Variablen korrelieren vermutlich miteinander und mit der Zielgröße. Fehlende Behandlungskombinationen sind gang und gäbe und die Daten werden als nichtorthogonal bezeichnet. Dies erzeugt einen wichtigen Unterschied für unsere statistische Modellierung, weil beim orthogonalen Design die einem gegebenen Faktor zugeschriebene Abweichung konstant ist und nicht von der Reihenfolge abhängt, in welcher der Faktor aus dem Modell entfernt wird. Im Unterschied dazu stellen wir bei nichtorthogonalen Daten fest, dass die einem gegebenen Faktor zugeschriebene Abweichung von der Reihenfolge abhängt, in welcher der Faktor aus dem Modell entfernt wird. Deshalb müssen wir gewissenhaft sein und die Bedeutung von Faktoren in nichtorthogonalen Studien bewerten, wenn sie vom maximalen Modell (das heißt von dem Modell, das sämtliche Faktoren und Interaktionen beinhaltet, mit denen sie möglicherweise verwechselt werden) entfernt werden. Denken Sie daran, dass bei nichtorthogonalen Daten die Reihenfolge eine Rolle spielt.