Thomas Schäfer

Statistik II

Inferenzstatistik

LEHRBUCH

BASISWISSEN PSYCHOLOGIE

VS VERLAG

Thomas Schäfer

Statistik II

Basiswissen Psychologie

Die neue Reihe im VS Verlag: Das Basiswissen ist konzipiert für Studierende und Lehrende der Psychologie und angrenzender Disziplinen, die Wesentliches in kompakter, übersichtlicher Form erfassen wollen.

Eine ideale Vorbereitung für Vorlesungen, Seminare und Prüfungen: Die Bücher bieten Studierenden in aller Kürze einen fundierten Überblick über die wichtigsten Ansätze und Fakten. Sie wecken so Lust am Weiterdenken und Weiterlesen.

Neue Freiräume in der Lehre: Das Basiswissen bietet eine flexible Arbeitsgrundlage. Damit wird Raum geschaffen für individuelle Vertiefungen, Diskussion aktueller Forschung und Praxistransfer.

Thomas Schäfer

Statistik II

Inferenzstatistik

VS VERLAG

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

1. Auflage 2011

Lektorat: Kea S. Brahms / Eva Brechtel-Wahl

VS Verlag für Sozialwissenschaften ist eine Marke von Springer Fachmedien.
Springer Fachmedien ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.vs-verlag.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: Ten Brink, Meppel
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier
Printed in the Netherlands

ISBN 978-3-531-16940-8

Inhaltsverzeichnis

Zusatzmaterialien (Prüfverteilungstabellen) unter www.vs-verlag.de – Online-PLUS zu Thomas Schäfer, Statistik II. (http://www.vs-verlag.de/Privatkunden/Zusatzmaterial/978-3-531-16940-8/Statistik-II.html)

Einleitung

Dieses Buch trägt den Titel „Statistik II“, denn es korrespondiert in seinen Inhalten mit dem, was an vielen Universitäten im 2. Semester Methodenlehre bzw. Statistik gelehrt wird. Im Wesentlichen geht es hier um die Frage, wie gut sich Ergebnisse aus einzelnen Studien auf größere Gruppen von Personen verallgemeinern lassen. Die Idee dieser „Inferenzstatistik“, ihre unterschiedlichen Analyseverfahren sowie die wichtigsten Grundlagen der Wahrscheinlichkeitstheorie sind Gegenstand dieses Buches. Schon am Titel ist zu erkennen, dass es sich um einen Fortsetzungsband handelt. Und tatsächlich werden hier gewisse statistische Grundkenntnisse vorausgesetzt. All diejenigen, die „Statistik I“ aus dieser Reihe gelesen und verstanden haben, können unbesorgt weiter lesen. Diejenigen Leserinnen und Leser, die bisher mit anderer Literatur gearbeitet haben, sollten mit grundlegenden Begriffen und Konzepten vertraut sein, insbesondere mit der Grundidee des Messens, dem Skalenproblem, dem Stichprobenziehen, der Berechnung und Interpretation von Lage- und Streuungsmaßen, Häufigkeitsverteilungen sowie Korrelation und Regression.

Ziel des Buches ist es, Leserinnen und Lesern, die bisher keine Kenntnisse von Inferenzstatistik haben, einen Überblick über die wichtigsten Begriffe und Analysemöglichkeiten zu geben – und dies in möglichst verständlicher Form. Ich hoffe, dies ist gelungen. Natürlich können in diesem Buch aus Platzgründen nicht alle Detailfragen oder auch komplexe Analyseverfahren behandelt werden. Auf weiterführende Literatur wird daher an den entsprechenden Stellen hingewiesen.

Für tatkräftige Unterstützung, wertvolle Hinweise und konstruktive Kritik, die zum Gelingen dieses Buches beigetragen haben, danke ich ganz herzlich Doreen Drechsler, Juliane Eberth und Frederik Haarig. Mein besonderer Dank für eine angenehme Zusammenarbeit und tatkräftige Unterstützung geht auch an den Herausgeber, Jürgen Kriz, sowie Kea S. Brahms vom VS Verlag.

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_1,

1 Was ist Inferenzstatistik?

1.1 Die Idee der Inferenzstatistik

Das Anliegen einer jeden wissenschaftlichen Aussage ist, dass sie möglichst allgemeingültig sein soll. Das heißt, wissenschaftliche Erkenntnisse über Sachverhalte, Zusammenhänge und Gesetzmäßigkeiten bekommen erst dadurch Gewicht, dass sie einen großen Geltungsbereich haben und damit für eine große Masse von Menschen oder für eine große Zahl von Sachverhalten zutreffen. Wie man aufgrund der gesammelten Daten von wenigen Personen Schlüsse (Inferenzen) über sehr große Gruppen von Menschen machen kann, werden wir uns im Zuge der Inferenzstatistik anschauen.

Wie wir wissen, können wir aus Einzelfällen keine gültigen Schlüsse für die Mehrheit ziehen. Wenn Johann Wolfgang von Goethe täglich bis zu zwei Flaschen Wein getrunken hat und dennoch relativ alt geworden ist, können wir daraus nicht die Schlussfolgerung ableiten, dass ein solcher Weinkonsum für alle Menschen gänzlich unbedenklich sei. Warum können wir solche Schlussfolgerungen nicht ziehen? Der Grund liegt auf der Hand: die Anzahl unserer „Untersuchungsteilnehmer" ist zu klein. Von einem Einzelfall eine Aussage über die Allgemeinheit abzuleiten, gelingt deswegen nicht, weil ein Einzelfall immer eine Ausnahme darstellen kann – eine Ausnahme, die durch Zufall ganz bestimmte Merkmale aufweist. Hätten wir durch eine Recherche herausgefunden, dass die Mehrzahl der großen deutschen Dichter der Vergangenheit täglich zwei Flaschen Wein getrunken hat, dann könnten wir es schon eher wagen, eine Aussage über die Unbedenklichkeit von Weinkonsum zu machen. Allerdings könnten wir auch dann nicht wirklich sicher sein, ob diese Aussage eventuell nur für Schriftsteller gilt, nicht aber für den Rest der Bevölkerung.

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_2,

Wir sehen also, Verallgemeinerungen zu treffen, stellt ein nicht zu unterschätzendes Problem dar. Und wem könnte dieses Problem mehr zu schaffen machen als Forschern und Wissenschaftlern? Im Speziellen sind es in der Tat Sozialwissenschaftler, z. B. Psychologen, denen sich dieses Problem stellt. Denn diese versuchen praktisch immer, Aussagen über die Allgemeinheit zu treffen. Solche Aussagen über die Allgemeinheit sind dann problemlos möglich, wenn man *alle* Personen, die die Allgemeinheit bilden, tatsächlich untersuchen kann. Wenn wir zum Beispiel den Anteil von Ehepaaren mit einem Kind in Deutschland herausfinden wollten, könnten wir beim statistischen Bundesamt nachfragen und die genaue Antwort erhalten. Damit erübrigt sich eine „Schlussfolgerung" über diesen Anteil in der Bevölkerung. Stattdessen haben wir den echten Wert bereits bestimmt. Ganz anders sieht es bei psychologischen Fragestellungen aus. Besonders dann, wenn es um bestimmte Effekte geht. Wirkt ein neu entwickeltes Training A besser als ein altes Training B? Werden attraktivere Menschen als erfolgreicher eingeschätzt? Gibt es einen Zusammenhang zwischen Computerspielen und Gewalt bei Jugendlichen? Die Liste ließe sich unendlich fortsetzen. Was all diese Fragen gemeinsam haben, ist, dass sie in wissenschaftlichen Studien nur an einer *Stichprobe* untersucht werden können. Ausgehend vom Ergebnis in dieser Stichprobe versucht man anschließend, diese Ergebnisse auf die Grundgesamtheit zu verallgemeinern.

Grundgesamtheit (Population)

Mit *Grundgesamtheit* oder *Population* ist immer die Gruppe von Menschen gemeint, für die eine Aussage zutreffen soll. In der Regel besteht die Grundgesamtheit in der Gruppe aller Menschen. Denn psychologische Erkenntnisse sollten möglichst allgemeingültig sein. Besonders wenn es um allgemeinpsychologische Gesetzmäßigkeiten (wie bei Wahrnehmungs- oder Gedächtnisprozessen) geht, sollte man davon ausgehen können, dass diese bei allen Menschen in gleicher Weise funktionieren. Manchmal ist die Grundgesamtheit aber kleiner und betrifft eine spezifische Gruppe. Bei der Frage, ob es einen Zusammenhang zwischen Computerspielen und Gewalt bei Jugendlichen gibt, ist die betreffende Population nur die der Jugendlichen. Aussagen über die Wirksamkeit von Antidepressiva gelten hingegen nur für die Population der Depressiven, usw. Egal wie groß die Population ist, für die eine verallgemeinernde Aussage getroffen

werden soll – es stellt sich immer die Frage, wie es machbar ist, solche Schlüsse ausgehend von einer Stichprobe auf eine Population zu ziehen.

Inferenzen und Inferenzstatistik

Solche Schlüsse werden auch Inferenzen genannt (*infero*, lateinisch: schließen), woraus sich die Bezeichnung *Inferenzstatistik* (manchmal auch *Schließende Statistik*) erklärt. Im Prozess der Erkenntnisgewinnung stellt die Inferenzstatistik einen entscheidenden Schritt dar (siehe Abbildung 1.1). Wenn die in einer Stichprobe erhobenen Daten deskriptiv aufbereitet und dargestellt wurden bzw. explorativ nach Mustern oder Zusammenhängen untersucht wurden, ist die Datenanalyse damit im Prinzip beendet. Ob es Unterschiede oder Zusammenhänge zwischen Variablen gibt, ist damit beantwortet und kann entsprechend mit Hilfe von Tabellen, Abbildungen oder statistischen Kennwerten dargestellt werden.

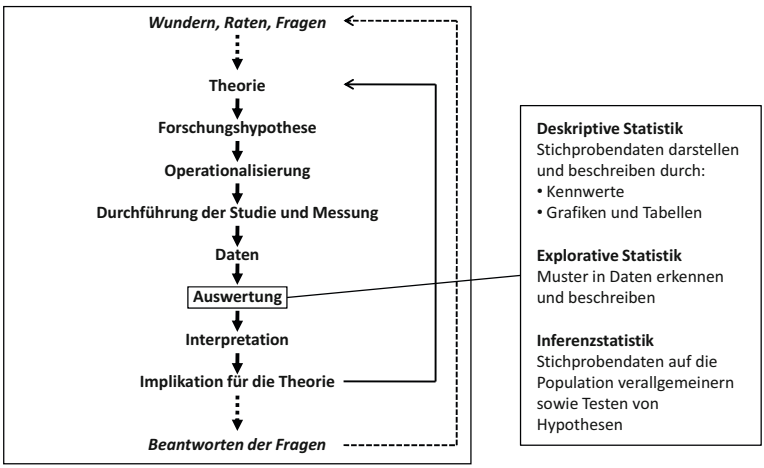

Abbildung 1.1 Die Rolle der Inferenzstatistik im Prozess der Erkenntnisgewinnung

Der nächste Schritt besteht nun aber in der Beantwortung der Frage, ob sich diese gefundenen Ergebnisse auf die jeweils interessierende Population verallgemeinern lassen. Das zu prüfen, ist die Aufgabe der Inferenzstatistik.

> Ziel der Inferenzstatistik sind Schlüsse von einer Stichprobe auf eine Population sowie Aussagen über die Güte dieser Schlüsse.

Als Sie sich mit der Idee des Messens in der Psychologie auseinandergesetzt haben, sind Sie schon auf den Zusammenhang von Stichprobe und Population gestoßen. Dabei haben Sie gesehen, dass Verallgemeinerungen von Stichprobenergebnissen auf die Population nur dann überhaupt möglich sind, wenn die Stichprobe *repräsentativ* für die Population ist. Die Stichprobe muss die Verhältnisse in der Population (zum Beispiel eine bestimmte Geschlechterverteilung, Altersverteilung, Intelligenzverteilung) möglichst gut widerspiegeln. Wie Sie auch bereits wissen sollten, kann man die Repräsentativität von Stichproben versuchen dadurch sicherzustellen, dass man Zufallsstichproben aus der Population zieht. Der Zufall sollte dafür sorgen, dass die eben genannten Verhältnisse in der Population genauso auch in der Stichprobe auftauchen. Und das wiederum gelingt umso besser, je größer die gewählte Stichprobe ist. Große Stichproben repräsentieren die Population besser als kleine Stichproben. Das Prinzip der Zufallsziehung ist in Abbildung 1.2 dargestellt. In der Abbildung ist auch zu sehen, dass die wahren Verhältnisse in der Population immer unbekannt sind. Sie sollen ja durch die Ergebnisse aus der Stichprobe mit Hilfe der Inferenzen geschätzt werden. Man kann sich daher das Ziehen von Stichproben aus der Population wie das Ziehen von Kugeln aus einer Urne vorstellen, deren Inhalt man nicht kennt. Jede Kugel steht für einen Wert, zum Beispiel den IQ von Person X. Die Verhältnisse in der Population sind genauso unbekannt wie der Inhalt der Urne. Und natürlich kann man besser auf den wahren Inhalt schließen, je mehr Kugeln man zieht – bzw. je größer die Stichprobe ist.

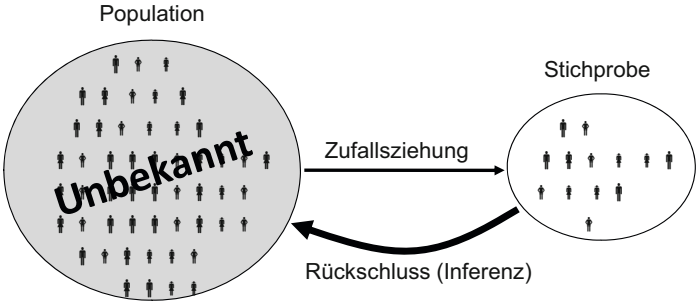

Abbildung 1.2 Inferenzen von der Stichprobe auf die Population

Die Qualität der Stichprobe – zufällig gezogen, möglichst groß – ist die grundlegende Voraussetzung für Inferenzen auf die Population. Dennoch bleibt eine Stichprobe immer nur ein Ausschnitt aus der Population und kann deshalb fehlerbehaftet sein. Mit anderen Worten: es besteht immer eine gewisse Wahrscheinlichkeit, dass wir eine Stichprobe gezogen haben, die nicht exakt die Population widerspiegelt und dass damit auch die in der Stichprobe gefundenen Ergebnisse nicht für die Population gelten, sondern nur in unserer Stichprobe aufgetreten sind. Die gesamte Inferenzstatistik dreht sich also um eine entscheidende Frage: Sind die Ergebnisse in meiner Stichprobe eine zufällige Besonderheit oder kann ich sie auf die Population verallgemeinern?

Wie kann man diese Frage beantworten? Hier gibt es zwei Möglichkeiten. Die erste besteht darin, dass man zu einer bestimmten Forschungsfrage nicht nur eine Studie macht, sondern gleich mehrere. Man würde also viele Stichproben ziehen und immer wieder die gleiche Frage untersuchen. Die Idee dahinter ist simpel: Wenn per Zufall in einer Stichprobe ein Effekt auftritt (zum Beispiel ein Mittelwertsunterschied), der in der Population gar nicht da ist, dann sollte er in einer zweiten Stichprobe nicht mehr auftauchen oder kleiner sein oder gar in die andere Richtung zeigen. Bei mehreren Stichproben sollten sich also zufällig beobachtete Effekte wieder ausmitteln. Falls sich jedoch über diese vielen Stichproben hinweg immer wieder derselbe Effekt zeigt, kann man davon ausgehen, dass dieser tatsächlich auch in der Population vorliegt. Das Problem dabei ist, dass kaum ein Forscher die gleiche Studie mehrmals wiederholt. Das ist zeitlich, finanziell und auch aus Gründen der Publizierbarkeit nicht möglich. Was jedoch

vorkommen kann, ist, dass verschiedene Forscher zu ein und derselben oder zu ähnlichen Fragestellungen Studien durchgeführt haben. Die Ergebnisse dieser Studien kann man sammeln und daraus einen durchschnittlichen Effekt bestimmen. Ein solches Vorgehen wird tatsächlich manchmal angewendet und wird als *Metaanalyse* bezeichnet (also als Analyse aus vielen Analysen). Metaanalysen liefern eine gute Schätzung für die wahren Verhältnisse in der Population. Und die Güte der Schätzung steigt natürlich mit der Anzahl der Studien, die in eine solche Metaanalyse aufgenommen werden.

Metaanalysen sind allerdings die Ausnahme, wenn es um die Schätzung von Populationswerten aus Stichprobenergebnissen geht. Der weitaus häufigere Fall ist der, dass man die Wahrscheinlichkeit, mit der man sich bei der Bestimmung eines Ergebnisses aus einer Stichprobe geirrt hat, bei der Darstellung dieses Ergebnisses einfach mit angibt. Dafür sind einige Überlegungen zu Wahrscheinlichkeiten nötig, mit denen wir uns jetzt beschäftigen werden. Im Anschluss sehen wir uns die verschiedenen Möglichkeiten an, mit denen man konkrete Aussagen über diese Wahrscheinlichkeiten machen kann.

Literaturempfehlung

Stichproben, Populationen und Inferenzstatistik:

Bortz, J. und Döring, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer. (Kapitel 7)
Bühner, M. und Ziegler, M. (2009). S*tatistik für Psychologen und Sozialwissenschaftler*. München: Pearson. (Kapitel 4)
Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 10)

Metaanalyse:

Bortz, J. und Döring, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer. (Kapitel 10)
Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 22)

1.2 Wahrscheinlichkeiten und Verteilungen

Um zu verstehen, welche Rolle die Wahrscheinlichkeit bei der Verallgemeinerung von Stichprobenergebnissen spielt, schauen wir uns zunächst an, was theoretisch passieren würde, wenn wir es mit dem vorhin beschriebenen Idealfall zu tun hätten – nämlich wenn wir nicht nur eine, sondern mehrere Stichproben ziehen und jedes Mal den Effekt bestimmen würden. Wie Sie bereits wissen, liefern einzelne Stichproben sogenannte Häufigkeitsverteilungen. In Abbildung 1.3 sehen wir eine Häufigkeitsverteilung als Ergebnis einer Studie, in der 50 Personen auf einer Skala von 1 bis 5 angeben sollten, wie sehr sie Klassik mögen.

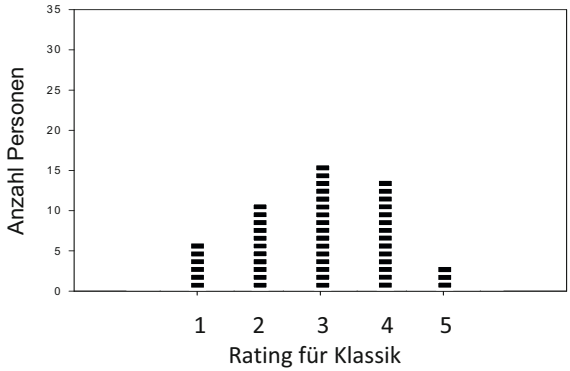

Abbildung 1.3 Häufigkeitsverteilung für die Vorliebe für Klassik

Der Mittelwert beträgt 2,9. Was würde nun passieren, wenn wir die Studie wiederholen und eine neue Stichprobe von Befragten ziehen würden? Wie wir eben argumentiert haben, würde der Zufall beim Ziehen dafür sorgen, dass wir wahrscheinlich etwas andere Werte von den Personen erhalten und damit auch einen anderen Mittelwert. Eine entsprechende Verteilung könnte dann so aussehen:

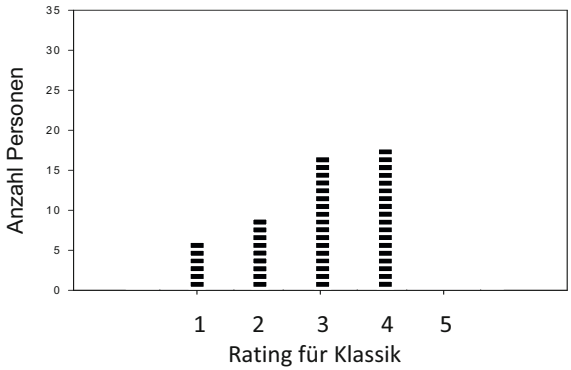

Abbildung 1.4 Alternative Häufigkeitsverteilung für die Vorliebe für Klassik

Der Mittelwert wäre jetzt 3,0, weicht also etwas vom ersten Mittelwert ab. Auf diese Weise könnten wir sehr viele Stichproben ziehen und jedes Mal das Ergebnis ermitteln. Diese Ergebnisse (also die Mittelwerte) können wir wieder in einer Verteilung abtragen. Was wir dabei erhalten, ist eine sogenannte *Stichprobenverteilung*.

Von der Häufigkeitsverteilung zur Stichprobenverteilung

In einer Stichprobenverteilung sind nicht mehr die Werte einzelner Personen abgetragen, sondern die Kennwerte (z. B. Mittelwerte) aus einzelnen Stichproben! Wie sieht eine solche Verteilung aus? Sehen wir uns ein mögliches Beispiel an:

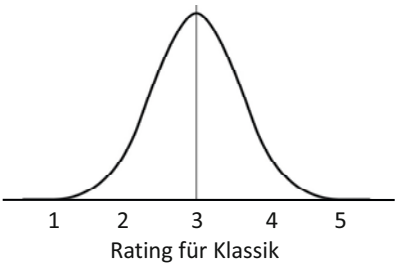

Abbildung 1.5 Stichprobenverteilung für Mittelwerte aus vielen Stichproben

Im Vergleich zu einer Häufigkeitsverteilung weist eine Stichprobenverteilung einige Unterschiede auf. Diese sollten Sie gut im Gedächtnis behalten, denn Stichprobenverteilungen sind die wichtigste Grundlage der Inferenzstatistik. Der auffälligste Unterschied ist der, dass die Werte in der Verteilung nicht mehr nur in die 5 Kategorien der 5 möglichen Werte fallen. Das liegt daran, dass wir es jetzt mit Mittelwerten zu tun haben. Während eine einzelne Person nur einen konkreten Wert haben kann (also 1, 2, 3, 4 oder 5), können Mittelwerte aus Studien beliebige Werte annehmen (wie zum Beispiel den Wert 2,9 von oben). Und da man Mittelwerte beliebig genau berechnen kann, also auf beliebig viele Stellen hinter dem Komma, werden in Stichprobenverteilungen auch nicht alle möglichen Mittelwerte als einzelne Balken dargestellt. Stattdessen sieht man in diesen Verteilungen meist nur eine durchgehende Kurve, die alle Werte enthält (das Integral).

Der zweite Unterschied zur Häufigkeitsverteilung besteht darin, dass auf der Y-Achse nun nicht mehr die Anzahl der Personen steht, sondern die relative Anzahl der Studien, die einen bestimmten Mittelwert geliefert haben. (Manchmal wird – wie in Abb. 1.5 – auf die Darstellung der Y-Achse verzichtet, da die genauen Y-Werte in der Regel gar nicht interessant sind.) Wie wir sehen, sind Studien mit einem Mittelwert von 2,9 vergleichsweise häufig vertreten. Ebenso wie Studien, die einen Mittelwert von 3,0 geliefert haben, wie die aus Abbildung 1.4. Wie die Stichprobenverteilung zeigt, hat es aber auch Studien gegeben, die sehr kleine und sehr große Mittelwerte erbracht haben. Solche Studien sind allerdings seltener zu finden. Das führt uns gleich zum nächsten Unterschied. Wenn wir aus Häufigkeitsverteilungen den Mittelwert berechnen, so liegt dieser – wie der Name schon andeutet – irgendwo in der Mitte der Daten und nicht etwa am Rand. Während es also in der Häufigkeitsverteilung natürlich Personen gegeben hat, die die Werte 1 oder 5 angekreuzt haben, wird der Mittelwert in aller Regel nicht in der Nähe von 1 oder 5 liegen. Für unsere Stichprobenverteilung bedeutet das nun, dass Werte am Rand des Wertebereiches sehr selten vorkommen sollten und Werte in der Mitte viel häufiger. Genau das spiegelt unsere Verteilung auch wider. Sie enthält nur sehr wenige Studien mit einem Mittelwert in der Nähe von 1,0 oder 5,0.

Was ist nun das Entscheidende an solchen Stichprobenverteilungen? Sie geben uns darüber Auskunft, was passieren würde, wenn wir tatsächlich sehr sehr viele Stichproben ziehen würden. Auf der X-Achse sind stets die möglichen Mittelwerte abgetragen, die wir in diesen Stichproben finden könnten. Jeder mögliche Mittelwert schneidet dabei eine bestimmte Fläche von der Verteilung

ab. Diese Fläche entspricht immer der Wahrscheinlichkeit, mit der alle möglichen Mittelwerte *innerhalb dieser Fläche* zu erwarten sind. Nehmen wir noch einmal die Stichprobenverteilung für die Vorliebe für Klassik. Wenn wir davon ausgehen (hypothetisch!), dass der tatsächliche Mittelwert für diese Variable in der Population 3,0 beträgt, so sagt uns diese Verteilung, dass wir – wenn wir immer wieder Studien machen und diesen Mittelwert erheben würden – auch Werte finden würden, die mehr oder weniger stark von 3,0 abweichen. Die Wahrscheinlichkeit, einen bestimmten Wert zu „ziehen", sinkt natürlich mit seinem Abstand vom wahren Wert in der Population, also 3,0. Wie wahrscheinlich wäre es etwa, in einer Studie einen Mittelwert von 3,8 oder einen noch größeren Wert zu finden? Dafür sehen wir uns an, welche Fläche von diesem Wert von der Verteilung „abgeschnitten" wird:

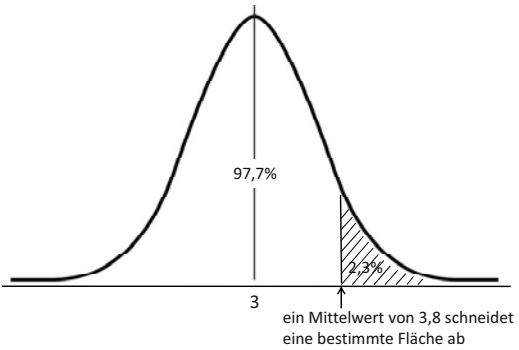

Abbildung 1.6 Die Lage eines empirischen Wertes in einer Stichprobenverteilung

Die Abbildung zeigt, dass die Wahrscheinlichkeit, einen Wert zu ziehen, der kleiner als 3,8 ist, 97,7% beträgt. Alle diese möglichen kleineren Werte sind in der linken nicht-schraffierten Seite der Verteilung enthalten. Werte von 3,8 oder größer würde man nur noch mit einer Wahrscheinlichkeit von 2,3% ziehen. Und natürlich ist die Wahrscheinlichkeit, in einer Studie den wahren Wert von 3,0 zu ziehen, auf lange Sicht am höchsten. Beachten Sie aber, dass wir in diesem Beispiel davon ausgegangen sind, dass wir den wahren Wert in der Population kennen und um ihn herum eine Stichprobenverteilung konstruiert haben. Das ist in der Forschung aber praktisch nie der Fall, sondern der wahre Wert ist genau das, was wir eigentlich suchen. Wir werden gleich darauf zurück kommen.

Die Form der Verteilung in Abbildung 1.6 führt uns zu einer weiteren Besonderheit von Stichprobenverteilungen aus Mittelwerten: sie folgen einer Normalverteilung (also der typischen Glockenform). Diese wird besonders dann deutlich, wenn sehr viele einzelne Stichproben in die Verteilung aufgenommen werden. Da wir ja davon ausgegangen sind, dass sich die Ergebnisse aus einzelnen Stichproben rein zufällig voneinander unterscheiden, sollten sie sich – wenn wir sie in einer Stichprobenverteilung abtragen – auch symmetrisch um einen Mittelwert verteilen. Dieses Prinzip wird durch den sogenannten Zentralen Grenzwertsatz beschrieben.

> Zentraler Grenzwertsatz: Die Verteilung der Mittelwerte einer großen Anzahl von Stichprobenergebnissen folgt immer einer Normalverteilung.

Für die Form der Stichprobenverteilung bleibt aber noch eine andere Frage zu klären: Wie hängt sie von der Stichprobengröße der einzelnen Stichproben ab, die in sie einfließen? An dieser Frage können Sie sehr gut Ihr statistisches Verständnis überprüfen. Überlegen Sie sich, was passieren würde, wenn die Stichproben, die Sie in die Verteilung aufnehmen, alle sehr groß sind. Würde die Verteilung dann unverändert aussehen, würde sie schmaler oder breiter werden? Hier kommt die Antwort: die Verteilung wird schmaler. Das liegt daran, dass größere Stichproben genauere Schätzungen liefern und damit näher am wahren Wert liegen (Sie erinnern sich). Ergebnisse aus solchen Studien streuen also viel weniger um den wahren Wert, was nichts anderes bedeutet, als dass die Stichprobenverteilung schmaler wird. Sie können sich das sehr einfach an einem Extrembeispiel verdeutlichen: Stellen Sie sich vor, Sie würden in vielen Stichproben jedes Mal die gesamte Population untersuchen und immer den Mittelwert eines Merkmals bestimmen. Wie sähe dann die Stichprobenverteilung aus? Ganz einfach: es gibt keine Verteilung in diesem Fall. Denn wenn Sie jedes Mal die gesamte Population untersuchen, werden Sie jedes Mal denselben Mittelwert finden (denn in der Population gibt es nur einen Mittelwert!). Ihre Stichprobenverteilung würde quasi nur noch aus einem Strich über dem wahren Mittelwert bestehen, der also eine Varianz von 0 aufweist. (Und es wäre natürlich sinnlos, eine Studie, bei der man die gesamte Population untersucht hat, zu wiederholen, denn das Ergebnis muss immer dasselbe sein, von Messfehlern einmal abgesehen.)

Mit steigender Stichprobengröße der einzelnen Studien sinkt die Streuung der resultierenden Stichprobenverteilung.

Fassen wir noch einmal zusammen, welchen Nutzen uns die Stichprobenverteilungen bringen. Wenn wir tatsächlich mehrere Studien zu einer Fragestellung gemacht und das Ergebnis jedes Mal in eine solche Verteilung abgetragen haben, dann repräsentiert der Mittelwert dieser Verteilung eine sehr viel bessere Schätzung für den wahren Wert in der Population als der Wert aus einer einzigen Studie. Zum anderen brauchen wir die Stichprobenverteilungen, um eine Angabe darüber machen zu können, wie gut wir Stichprobenergebnisse auf die Population verallgemeinern können. Wie das genau geht, werden wir uns in den nächsten beiden Kapiteln ansehen. Überlegen wir aber noch kurz, für welche Arten von Werten wir überhaupt inferenzstatistische Aussagen machen wollen. Wir haben bisher immer von Stichprobenergebnissen gesprochen. Was kann das alles sein? Das können alle Kennwerte sein, denen Sie schon begegnet sind: also solche, die etwas über die Lage und Streuung eines Merkmals aussagen (wie Mittelwerte, von denen wir eben immer gesprochen haben), solche, die einen Unterschied zwischen verschiedenen Gruppen beschreiben (also Mittelwertsunterschiede) und solche, die einen Zusammenhang zwischen Variablen beschreiben (wie die Korrelation). Im nächsten Kapitel sehen wir uns zunächst an, wie man inferenzstatistische Aussagen über die Lage und die Streuung von Merkmalen treffen kann.

Literaturempfehlung

Bühner, M. und Ziegler, M. (2009). Statistik für Psychologen und Sozialwissenschaftler. München: Pearson. (Kapitel 3)
Sedlmeier, P. & Köhlers, D. (2001). *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln*. Braunschweig: Westermann.

2 Inferenzstatistische Aussagen für Lage- und Streuungsmaße

In Abbildung 1.3 hatten wir in unserer Studie zur Vorliebe für Klassik einen Mittelwert von 2,9 gefunden. Die entsprechende inferenzstatistische Frage lautet nun: wie zuverlässig kann ich diesen gefundenen Mittelwert auf die Population verallgemeinern und sagen, dass die durchschnittliche Vorliebe *aller Deutschen* für Klassik 2,9 beträgt? Mit anderen Worten: wie sehr kann ich meinem gefundenen Mittelwert trauen? Zur Beantwortung dieser Frage gibt es zwei alternative Möglichkeiten. Einerseits können wir versuchen, den Fehler zu schätzen, den wir bei einer solchen Verallgemeinerung „durchschnittlich" machen werden. Das ist der sogenannte *Standardfehler*. Die andere Möglichkeit besteht darin, nicht einfach unseren gefundenen Mittelwert als Schätzung anzugeben, sondern einen Bereich um den Mittelwert herum, der den wahren Mittelwert in der Population wahrscheinlich enthält. Man spricht bei diesem Bereich von einem *Konfidenzintervall*. Wir werden uns nun beide Möglichkeiten ansehen.

2.1 Der Standardfehler

Die Idee hinter dem Standardfehler ist relativ einfach, und sie ist Ihnen schon aus der deskriptiven Statistik bekannt. Wie Sie wissen, ist es nicht sinnvoll, einen Mittelwert ohne seine Streuung anzugeben. Denn Mittelwerte mit kleinen Streuungen sind viel aussagekräftiger als solche mit großen Streuungen. (Sie erinnern sich: kleine Streuungen bedeuten, dass die Häufigkeitsverteilung sehr schmal ist, sich also eng um ihren Mittelwert verteilt und der Mittelwert damit sehr repräsentativ ist für die Daten dieser Verteilung.)

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_3,

Dieses Prinzip kann man sich bei Aussagen über die Güte einer Schätzung von Stichprobenergebnissen auf die Population zunutze machen. Im einfachsten Fall kann man die Standardabweichung der Daten schon als Gütemaß benutzen. Und das ist auch der häufigste Fall: in Publikationen werden in der Regel Mittelwerte (*M*) und ihre Standardabweichungen (*SD*) als Lage- und Streuungsmaße berichtet. Doch obwohl dieses Vorgehen so häufig ist, hat es einen kleinen Haken, der gern übersehen wird: Wie wir behauptet haben, sind Ergebnisse aus einzelnen Stichproben vom Zufall abhängig. Das trifft allerdings nicht nur auf den Mittelwert zu, über den wir die ganze Zeit gesprochen haben, sondern natürlich auch auf die *Streuung* in unserer Stichprobe. Diese kann in ihrer Größe – je nach der gezogenen Stichprobe – ebenso variieren. Daher ist es im Prinzip nicht ganz korrekt, die Standardabweichung einer Stichprobe als Schätzung für die Güte des Mittelwertes zu verwenden. Viel sinnvoller und informativer wäre es zu wissen, wie die Standardabweichung aussehen würde, wenn wir nicht eine, sondern sehr viele Stichproben gezogen hätten. Sie sehen, jetzt kommt die Idee der Stichprobenverteilungen ins Spiel, über die wir vorhin so ausführlich gesprochen haben. Wir würden dann nämlich wissen, wie der gefundene Mittelwert in einer Vielzahl von Stichproben variieren würde, und das wäre eine sehr viel genauere Schätzung für die Streuung. Anders ausgedrückt: das, was wir hierbei suchen – der Standardfehler – ist nichts anderes als die Standardabweichung der Stichprobenverteilung. Betrachten wir noch einmal die Verteilung aus Abbildung 1.5. Hier verteilen sich alle möglichen Mittelwerte, die wir in Stichproben hätten finden können. Die Standardabweichung dieser Verteilung repräsentiert den Standardfehler. Er beziffert die Ungenauigkeit, wenn wir ein Stichprobenergebnis auf die Population verallgemeinern. Nehmen wir an, die Verteilung in Abbildung 1.5 hätte eine Standardabweichung von 0,6. Der *Standardfehler des Mittelwertes* wäre also 0,6. Das würde bedeuten, dass Mittelwerte, die wir aus der Population mit Hilfe einer Stichprobe ziehen, „im Durchschnitt" um 0,6 Einheiten vom wahren Mittelwert abweichen. Beim Schätzen des wahren Wertes würden wir demnach einen „durchschnittlichen Fehler" von 0,6 Einheiten machen. (Für dieses Beispiel würde das bedeuten, dass wir auf der Skala von 1 bis 5 im Schnitt nur um 0,6 Einheiten daneben liegen würden. Das ist ein relativ gutes Ergebnis und bedeutet, dass die Güte unserer Schätzung akzeptabel ist. Wir könnten unseren Mittelwert also guten Gewissens auf die Population verallgemeinern.)

Die Bedeutung des Standardfehlers s_e ist in Abbildung 2.1 noch einmal dargestellt. Wir haben um unseren gefundenen Mittelwert von 3 eine Stichprobenverteilung konstruiert (mit Hilfe einer Computersimulation), die eine Standardabweichung von 0,6 besitzt. Wie wir wissen, umfasst die Standardabweichung – wenn man sie zu beiden Seiten des Mittelwertes aufspannen würde – ungefähr 68% der Fläche der Verteilung. Wenn wir diese Verteilung in eine Standardnormalverteilung transformieren würden, dann würde der Wert 3,6 also einem *z*-Wert von 1 entsprechen und der Wert 2,4 einem *z*-Wert von -1. Die Standardabweichung dieser Stichprobenverteilung entspricht dem Standardfehler des Mittelwertes. Mit dieser Ungenauigkeit ist die Schätzung des wahren Mittelwertes in der Population durch den Mittelwert der Stichprobe behaftet.

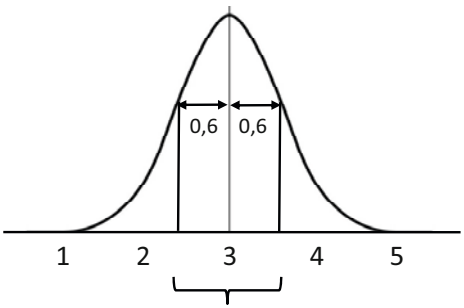

Abbildung 2.1 Darstellung des Standardfehlers anhand der Stichprobenverteilung für die Variable „Mögen von Klassik“

Nun werden Sie sich aber vielleicht schon gefragt haben, was uns die bisher angestellten Überlegungen bringen sollen. Denn schließlich haben wir ja nur eine und nicht mehrere Studien gemacht und damit auch nur eine Stichprobe zur Verfügung. Das ist richtig, allerdings kann man den Standardfehler aus der Standardabweichung des gefundenen Mittelwertes errechnen. Dafür benötigen wir allerdings nicht die Standardabweichung *s*, die wir schon aus der deskriptiven Statistik kennen – denn die gilt nur für eine konkrete Stichprobe – son-

dern wir müssen die Standardabweichung *für die Population schätzen*. Sie wird entsprechend als $\hat{\sigma}$ (Sigma, mit einem Dach für die Schätzung) bezeichnet. Diese weicht allerdings nur minimal von s ab; genauer gesagt, muss für ihre Berechnung nur der Nenner der bekannten Formel um 1 vermindert werden:

$$\hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}} = s\sqrt{\frac{n}{n-1}}$$

Es lässt sich zeigen, dass mit dieser Formel die Standardabweichung in der Population exakter geschätzt werden kann als mit der Formel für s. Daher taucht in der Inferenzstatistik – die sich ja stets auf Populationen bezieht – immer die geschätzte Streuung für die Population $\hat{\sigma}$ auf. Aus dieser geschätzten Populationsstreuung können wir nun den Standardfehler berechnen:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Wie Sie sehen, wird hier der Standardfehler mit dem Symbol $\hat{\sigma}_{\bar{x}}$ gekennzeichnet. Das Sigma mit dem Dach bedeutet auch hier wieder, dass wir es mit einem geschätzten Wert für die Population zu tun haben. Und der Index zeigt an, dass sich die Streuung auf den Mittelwert $\bar{x}$ bezieht. Man spricht daher auch oft vom *Standardfehler des Mittelwertes*. Oft wird er auch in lateinischen Buchstaben dargestellt als s_e oder SE oder S.E. (für *standard error*), manchmal auch als SEM (für *standard error of mean*).

> Der Standardfehler des Mittelwertes quantifiziert den Unterschied zwischen den aus einer einzelnen Stichprobe geschätzten Mittelwerten $\bar{x}$ und dem tatsächlichen, wahren Mittelwert μ. Er entspricht der Standardabweichung der entsprechenden Stichprobenverteilung.

Die Formel spiegelt außerdem wider, dass der Standardfehler immer kleiner ist als die Standardabweichung aus einer Stichprobe. Warum das so sein sollte, haben wir bei der Stichprobenverteilung schon diskutiert: ihre Streuung kann gar nicht so groß werden wie die Streuung einer einzelnen Häufigkeitsverteilung.

Wie wir gesagt hatten, wird bei Mittelwerten in Publikationen oft einfach die Standardabweichung *SD* angegeben und nicht der Standardfehler. Das gilt allerdings umgekehrt für Abbildungen mit sogenannten Fehlerbalken. Wie Sie wissen,

kann man in Balkendiagrammen die Streuung der Mittelwerte mit Hilfe von Fehlerbalken darstellen. Dafür kann man zwar auch die Standardabweichung benutzen; allerdings ist es hier üblicher, den Standardfehler zu verwenden. Manchmal wird auch nur der Mittelwert mit seinem Standardfehler abgetragen. Man spricht dann von *Fehlerplots*. Alle drei Möglichkeiten sind in Abbildung 2.2 dargestellt. Hier wurden jeweils 50 Männer und 50 Frauen nach ihrer Vorliebe für Klassik befragt. Auch hier sieht man sehr schön, dass der Standardfehler viel kleiner ist als die Standardabweichung.

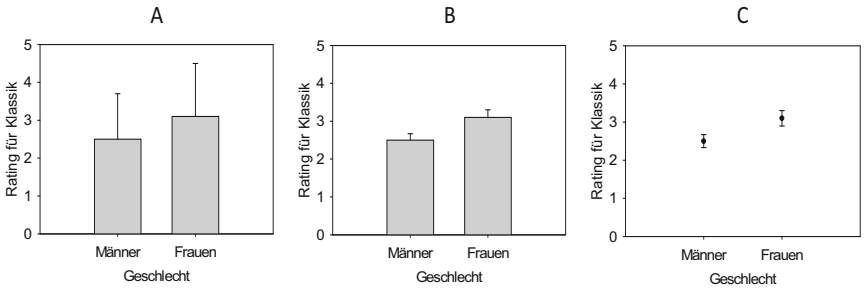

Abbildung 2.2 Drei verschiedene Darstellungsmöglichkeiten für Mittelwerte und ihre Streuungen (A mit Standardabweichungen, B mit Standardfehlern, C Fehlerplot mit Mittelwerten und Standardfehlern; als Beispieldaten wurden verwendet: Männer: $M = 2{,}5$; $SD = 1{,}2$; $\hat{\sigma}_{\bar{x}} = 0{,}17$; Frauen: $M = 3{,}1$; $SD = 1{,}4$; $\hat{\sigma}_{\bar{x}} = 0{,}2$)

Fassen wir noch einmal zusammen: Um einen in einer Stichprobe gefundenen Mittelwert auf eine Population verallgemeinern zu können, müssen wir eine Schätzung für den Mittelwert in der Population angeben sowie ein Gütemaß für diese Schätzung. Die Schätzung des Populationsmittelwertes (oft als $\hat{\mu}$ bezeichnet) erfolgt einfach durch unseren in der Stichprobe gefundenen Mittelwert. Diesen müssen wir auch als den Wert ansehen, den wir in der Population erwarten würden. Daher wird der Mittelwert auch manchmal als *Erwartungswert* bezeichnet. Und die Güte dieser Schätzung können wir durch den Standardfehler angeben. Dieser sollte möglichst klein sein. (Ob ein Standardfehler groß oder klein ist, hängt immer von der Skalierung der jeweiligen Variable

und der Fragestellung ab. In Abbildung 2.2 ist es augenscheinlich so, dass der Fehler – relativ zur Skala, die von 1 bis 5 reicht – recht klein ist.)

Literaturempfehlung

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 6)

Wenden wir uns nun der zweiten Möglichkeit zu, die Güte der Verallgemeinerung eines Mittelwertes auf die Population anzugeben: dem Konfidenzintervall.

2.2 Konfidenzintervalle

Beim Standardfehler haben wir gesehen, dass Schätzungen von Populationsmittelwerten immer mit einem gewissen Fehler behaftet sind, den wir angeben können. Manchmal reicht diese Information aber eventuell nicht aus, oder aber man möchte Schätzungen machen, die noch aussagekräftiger sind. Das kann man tun, wenn man nicht den genauen Mittelwert schätzen will, sondern einen Bereich um den Mittelwert angibt, in dem der wahre Wert in der Population wahrscheinlich liegt. Dieser Bereich kann natürlich viel zuverlässiger geschätzt werden als ein einziger Mittelwert, da er viel mehr mögliche Werte zulässt. Im Prinzip kann man die Verlässlichkeit solcher Bereiche beliebig erhöhen. Man muss sie nur genügend groß wählen. Man spricht bei solchen Bereichen von *Konfidenzintervallen* (oder auch *Vertrauensintervallen*).

> Ein Konfidenzintervall ist ein Wertebereich, bei dem wir darauf vertrauen können (*konfident* sein können), dass es den wahren Wert in der Population mit einer gewissen Wahrscheinlichkeit (der Vertrauenswahrscheinlichkeit) überdeckt.

Konfidenzintervalle sind auf den ersten Blick nicht so leicht zu verstehen. Wenn man sich aber ansieht, wie sie konstruiert werden, wird ersichtlich, dass sie einer sehr einfallsreichen Logik folgen und sehr interessante Aussagen zulassen. Konfidenzintervalle bauen ebenfalls auf der Idee der Stichprobenver-

teilungen auf – genau wie der Standardfehler. Sehen wir uns an, welche grundlegende Idee hinter der Konstruktion von Konfidenzintervallen steckt.

Konstruktion eines Konfidenzintervalls

(1) Zunächst legt man die gewünschte Güte des Intervalls fest. Damit ist die Vertrauenswahrscheinlichkeit gemeint. Wie groß soll die Wahrscheinlichkeit sein, dass unser Intervall den wahren Wert in der Population tatsächlich enthält? Natürlich wollen wir hier eine sehr große Wahrscheinlichkeit. Aber Vorsicht! Je größer wir diese Wahrscheinlichkeit wählen, desto breiter wird später auch unser Intervall werden; es wird also einen breiteren Wertebereich abdecken. Wenn der Wertebereich allerdings zu groß ist, ist das uninformativ für uns. Wir werden gleich noch sehen, warum. Gebräuchliche Wahrscheinlichkeiten liegen bei 90, 95 oder 99 Prozent.

(2) Im nächsten Schritt erheben wir unsere Stichprobe und bestimmen den Mittelwert. Dieser Mittelwert aus unserer Stichprobe ist die beste Schätzung dafür, wie auch der Mittelwert in der Population ausfallen wird. Daher benutzen wir ihn als *Erwartungswert*, um den herum das Intervall „aufgespannt" werden soll.

(3) Überlegen wir nun, wo der wahre Mittelwert überhaupt liegen kann. Das sollten Sie nun bereits beantworten können. Diese Information steckt natürlich in der Stichprobenverteilung. Sie gibt an, welche Mittelwerte man bei wiederholten Ziehungen aus der Population erwarten kann und wie wahrscheinlich diese sind. Wir müssen nun also eine Stichprobenverteilung um unseren Erwartungswert herum konstruieren. (Diese Konstruktion machen wir nur aus Gründen der Verständlichkeit. Wir werden gleich sehen, dass wir auf einem rechnerischen Weg etwas einfacher vorgehen können.) Die Streuung für die Stichprobenverteilung (d. h. den Standardfehler) können wir aus der Streuung unserer Stichprobe berechnen.

(4) Anschließend müssen wir diejenige Fläche der Verteilung um den Mittelwert herum markieren, die wir oben als Vertrauenswahrscheinlichkeit festgelegt haben. Sie erinnern sich: bei Stichprobenverteilungen entspricht die Fläche unter

der Verteilung der Wahrscheinlichkeit der einzelnen Werte, die man „ziehen" kann. Wenn wir von einer Vertrauenswahrscheinlichkeit von 90% ausgehen, müssen wir also die mittleren 90% der Verteilung markieren.

(5) Im letzten Schritt schauen wir, welche Werte auf der X-Achse von diesem Intervall „abgeschnitten" werden. Das sind die Werte unseres Konfidenzintervalls. Jedes Intervall hat eine *untere* und eine *obere Grenze*. Der Bereich zwischen diesen beiden Werten oder Grenzen überdeckt den wahren Wert in der Population nun also mit einer Wahrscheinlichkeit von 90%.

Dieses Vorgehen wird leichter verständlich, wenn wir uns ein Beispiel ansehen. Nehmen wir an, wir hätten an einer Stichprobe von 100 Personen mit Hilfe eines Tests die emotionale Intelligenz erhoben und einen Mittelwert von 105 gefunden. Wir wollen nun das 90%-Konfidenzintervall angeben. Abbildung 2.3 zeigt die entsprechende Stichprobenverteilung und das Konfidenzintervall.

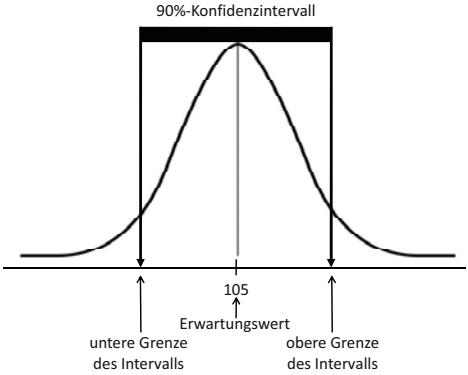

Abbildung 2.3 90%-Konfidenzintervall um einen Mittelwert von 105

Die Abbildung zeigt schematisch, wie um den gefundenen Mittelwert von 105, der uns als Erwartungswert für die Population dient, eine Stichprobenverteilung konstruiert wurde. (Man kann eine solche *theoretische* Stichprobenverteilung am Computer simulieren, wenn man Mittelwert und Streuung der Stichprobe erhoben hat.) Das Konfidenzintervall ist durch den schwarzen Bal-

ken dargestellt. Er sollte – symmetrisch um den Erwartungswert herum – 90% der Fläche dieser Verteilung „abschneiden“. Die entsprechenden Werte, die auf der X-Achse abgeschnitten werden, sind die untere und obere Grenze des Konfidenzintervalls.

Nehmen wir an, die untere Grenze liegt bei einem Wert von 98, die obere bei einem Wert von 112. Wie lautet nun die exakte Interpretation dieses Intervalls? Diese kann man sich wieder aus der Idee der Stichprobenverteilung herleiten: Die mittleren 90% der Verteilung enthalten Werte, die bei wiederholten Ziehungen von Stichproben aus der Population in 90 von 100 Fällen gezogen würden, wenn der wahre Mittelwert in der Population tatsächlich 105 beträgt. In den restlichen 10 von 100 Fällen würden wir Werte außerhalb des Intervalls ziehen. Damit können wir unser Ergebnis so interpretieren: Die Wahrscheinlichkeit, dass das Intervall von 98 bis 112 Punkten den wahren Mittelwert der emotionalen Intelligenz in der Population überdeckt, beträgt 90%. Der Vorteil von Konfidenzintervallen liegt damit auf der Hand: wir können Wahrscheinlichkeiten für die Güte einer Schätzung angeben, unter denen wir uns auch etwas vorstellen können. Das ist gegenüber dem abstrakteren Standardfehler ein großer Vorteil. Beachten Sie allerdings, dass Konfidenzintervalle keine Aussagen über die Wahrscheinlichkeit der Lage des wahren Wertes zulassen, sondern immer nur dafür, dass das Intervall den wahren aber unbekannten Parameter in der Population überdeckt.

Konfidenzintervalle mit Hilfe der t-Verteilung

Vielleicht haben Sie schon gedacht, dass es relativ mühselig wäre, wenn man die oben beschriebene Prozedur zur Bestimmung des Konfidenzintervalls jedes Mal durchlaufen müsste. Als Alternative kann man die Grenzen des Intervalls auch ausrechnen, ohne erst die Stichprobenverteilung konstruieren zu müssen. Man geht dabei eine Art „Umweg“, indem man eine Verteilung benutzt, die bereits bekannt ist. Einer solchen Verteilung sind Sie schon begegnet: der Standardnormalverteilung, die immer einen Mittelwert von 0 und eine Streuung von 1 hat. Im Prinzip kann man die Standardnormalverteilung auch für die Berechnung von Konfidenzintervallen benutzen, da sich die Verteilung von Merkmalen in großen Stichproben (und damit auch die Verteilung von Mittelwerten vieler Stichproben), wie Sie schon wissen, der Normalverteilung annähert. Aller-

dings nur bei großen Stichproben (ab ca. 30 Personen). Bei kleineren Stichproben hat sich gezeigt, dass sie für die Verteilung von Mittelwerten nicht so gut geeignet ist. Stattdessen benutzt man eine andere standardisierte Verteilung: die sogenannte *t-Verteilung*. Während in der Standardnormalverteilung *z*-Werte abgetragen sind, sind es in der *t*-Verteilung *t*-Werte. Auch diese Verteilung hat einen Mittelwert von 0 und eine Streuung von 1. Allerdings ist die Form der *t*-Verteilung – anders als die Form der Standardnormalverteilung – von der Stichprobengröße abhängig. Bei großen Stichproben geht die *t*-Verteilung in die Standardnormalverteilung über – dann ist es egal, welche der beiden Verteilungen man benutzt. Bei kleineren Stichproben ist der Gipfel der *t*-Verteilung etwas niedriger.

Die *t*-Verteilung wird nun folgendermaßen für die Berechnung von Konfidenzintervallen benutzt. Im oben dargestellten Fall haben wir nach den Grenzen des Intervalls anhand der Stichprobenverteilung gesucht. Das heißt, bei einem 90%-Konfidenzintervall haben wir von der Mitte aus geschaut, welche Werte auf der X-Achse vom Intervall abgeschnitten werden. Diese Schnittpunkte suchen wir jetzt nicht in der Stichprobenverteilung unserer Rohdaten, sondern entsprechend in der *t*-Verteilung (siehe Prüfverteilungs-Tabellen). Wir benötigen die *t*-Werte, die von einem 90%-Intervall abgeschnitten werden. Um diese in der *t*-Verteilung zu bestimmen, benötigt man außerdem die Stichprobengröße (da die Form der *t*-Verteilung ja davon abhängig ist). Diese wird in standardisierten Verteilungen – und auch für viele Arten von Berechnungen – als sogenannte *Freiheitsgrade* oder *df* (degrees of freedom) ausgedrückt. Freiheitsgrade beschreiben die Anzahl von Werten, die in einem statistischen Ausdruck frei variieren können. Ein Beispiel: Wenn man weiß, dass die Summe der Klausurnoten von 4 Studierenden 12 ist und man sich nun fragt, wer welche Note hat, so brauchen wir nur 3 der Studierenden befragen. Die Note des vierten Studierenden steht dann fest, weil wir die Summe schon kennen. Diese letzte Note kann daher nicht mehr frei variieren, sondern es gibt hier nur drei Werte, die „frei" sind. Die Freiheitsgrade für dieses Beispiel wären damit: $df = 4 - 1 = 3$. Entsprechend bestimmen sich auch die Freiheitsgrade für Mittelwerte immer nach der folgenden Formel:

$$df = n - 1$$

Warum man beim Umgang mit Verteilungen die Freiheitsgrade benutzt und nicht die Stichprobengröße, ist eher ein mathematisches Problem, weshalb wir hier nicht näher darauf eingehen wollen.

Wir erhalten schließlich einen *t*-Wert, $t_{df,Konf.}$, der für eine bestimmte Höhe der Konfidenz und eine bestimmte Zahl von Freiheitsgraden gilt. (Dieser *t*-Wert gilt gleichzeitig für die obere und für die untere Grenze des Intervalls, da die *t*-Verteilung ja symmetrisch und ihr Mittelwert 0 ist. Beide Werte unterscheiden sich lediglich durch ihr Vorzeichen.) Um diesen *t*-Wert mit der ursprünglichen Verteilung der Rohwerte in Zusammenhang zu bringen (die wir uns dafür nun gar nicht direkt ansehen müssen), wird er einfach mit dem Standardfehler $\hat{\sigma}_{\bar{x}}$ multipliziert. Dieses Produkt beschreibt dann den Abstand, den die Grenzen des Intervalls *in Rohwerten ausgedrückt* von ihrem Mittelwert haben müssen. Daher müssen wir dieses Produkt einmal von unserem Mittelwert $\bar{x}$ abziehen und einmal hinzuaddieren:

$$untere\ Grenze: \bar{x} - \hat{\sigma}_{\bar{x}} \cdot t_{df,Konf.}$$

$$obere\ Grenze: \bar{x} + \hat{\sigma}_{\bar{x}} \cdot t_{df,Konf.}$$

Sehen wir uns dazu ein Beispiel an. Wir haben in einer Studie gefunden, dass der Mittelwert der Restaurantbesuche pro Jahr bei 20 zufällig ausgesuchten Deutschen 47 beträgt, mit einer Standardabweichung von 8,6. Für die Berechnung des 95%-Konfidenzintervalls benötigen wir zunächst den Standardfehler des Mittelwertes, den wir aus der Standardabweichung schätzen können. Die Populationsstreuung beträgt:

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = 8{,}6\sqrt{\frac{20}{19}} = 8{,}82$$

Der Standardfehler des Mittelwertes beträgt dann:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{8{,}82}{\sqrt{20}} = 1{,}97$$

Nun benutzen wir die Tabelle der *t*-Verteilung, um den kritischen *t*-Wert für eine Vertrauenswahrscheinlichkeit von 95% zu finden. Dabei müssen wir bei $df = 20 - 1 = 19$ Freiheitsgraden nachschauen. Bei einer Vertrauenswahrscheinlichkeit von 95% müssen wir die *mittleren* 95% der *t*-Verteilung abschneiden, sodass auf jeder Seite der Verteilung 2,5% übrig bleiben. Der Flächenanteil, bei dem wir nachschauen müssen, ist also 0,975. Die Tabelle liefert dafür einen *t*-Wert von 2,093. Damit können wir die beiden Grenzen des Intervalls ausrechnen:

$$untere\ Grenze: \bar{x} - \hat{\sigma}_{\bar{x}} \cdot\ t_{df,Konf.} = 47 - 1{,}97 \cdot 2{,}093 = 42{,}9$$

$$obere\ Grenze: \bar{x} + \hat{\sigma}_{\bar{x}} \cdot\ t_{df,Konf.} = 47 + 1{,}97 \cdot 2{,}093 = 51{,}1$$

Wir können demnach zu 95% sicher sein, dass der wahre Mittelwert von Restaurantbesuchen der Deutschen von unserem Intervall, das von 42,9 bis 51,1 reicht, überdeckt wird.

Hätten wir die gleichen Werte an einer größeren Stichprobe von 50 Personen gefunden, könnten wir statt der *t*-Verteilung nun die *z*-Verteilung benutzen. Zunächst berechnen wir wieder die Populationsstreuung:

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = 8{,}6\sqrt{\frac{50}{49}} = 8{,}69$$

Der Standardfehler des Mittelwertes beträgt dann:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{8{,}69}{\sqrt{50}} = 1{,}23$$

Den *z*-Wert suchen wir in der *z*-Tabelle bei einem Flächenanteil von 0,975. Die Freiheitsgrade spielen in der *z*-Verteilung keine Rolle. Wir erhalten einen *z*-Wert von 1,96 zur Berechnung des Konfidenzintervalls:

$$untere\ Grenze: \bar{x} - \hat{\sigma}_{\bar{x}} \cdot\ z_{Konf.} = 47 - 1{,}23 \cdot 1{,}96 = 44{,}6$$

$$obere\ Grenze: \bar{x} + \hat{\sigma}_{\bar{x}} \cdot\ z_{Konf.} = 47 + 1{,}23 \cdot 1{,}96 = 49{,}4$$

Wir sehen also, dass unter Verwendung einer größeren Stichprobe die Grenzen des Intervalls näher zusammenrücken und damit eine viel informativere Aussage liefern. Das Intervall erstreckt sich nun nur noch auf einen Wertebereich, dessen Grenzen ca. 5 Punkte auseinander liegen.

Die Höhe der Konfidenz

Wie schon erwähnt, können wir die Vertrauenswahrscheinlichkeit im Prinzip beliebig erhöhen, um noch verlässlichere Schätzungen zu machen. Was würde passieren, wenn wir im eben betrachteten Beispiel nicht 95, sondern 99% Vertrauenswahrscheinlichkeit festgelegt hätten? Wenn wir mehr Vertrauen darin haben wollen, dass unser Intervall den wahren Wert tatsächlich überdeckt,

sollte das offenbar mit einem größeren möglichen Wertebereich einhergehen. Die z-Tabelle liefert für einen Flächenanteil von 0,995 (womit also 0,5% auf jeder Seite abgeschnitten werden) einen Wert von 2,575 (das ist der Mittelwert der z-Werte der beiden Flächen 0,9949 und 0,9951 , denn nur diese sind in der Tabelle abgetragen):

$$untere\ Grenze: \bar{x} - \hat{\sigma}_{\bar{x}} \cdot z_{Konf.} = 47 - 1{,}23 \cdot 2{,}575 = 43{,}8$$

$$obere\ Grenze: \bar{x} + \hat{\sigma}_{\bar{x}} \cdot z_{Konf.} = 47 + 1{,}23 \cdot 2{,}575 = 50{,}2$$

Da das Intervall nun 99% der Verteilung abschneiden muss, liegen die beiden Grenzen natürlich weiter auseinander. Wir können jetzt zu 99% sicher sein, dass dieses Intervall den wahren Wert in der Population überdeckt. Damit haben wir zwar die Vertrauenswahrscheinlichkeit erhöht, allerdings ist das resultierende Intervall nun etwas weniger informativ, weil es mehr mögliche Werte einschließt. Die Präzision unserer Mittelwertschätzung ist also geringer geworden. Überlegen Sie, was passieren würde, wenn Sie die Wahrscheinlichkeit auf 100% erhöhen! Die Antwort ist trivial: Sie könnten dann zu 100% sicher sein, dass der wahre Wert irgendwo zwischen dem kleinstmöglichen und dem größtmöglichen Wert liegt. Ganz einfach deshalb, weil das Intervall nun den gesamten Wertebereich abdeckt. Das wussten Sie aber vorher schon, und daher können Sie mit dieser Information nichts anfangen. Die Festlegung der Vertrauenswahrscheinlichkeit ist also immer ein Kompromiss: sie sollte zwar hoch sein, aber die Grenzen des Intervalls sollten für Sie immer noch eine relevante Information liefern. Daher liegen gebräuchliche Wahrscheinlichkeiten bei 90, 95 oder seltener bei 99 Prozent.

Man kann diesen Kompromiss allerdings noch etwas abmildern, indem man größere Stichproben verwendet. Wie Sie sich erinnern, würde eine größere Stichprobe dazu führen, dass die resultierende Stichprobenverteilung, deren Standardfehler wir für das Konfidenzintervall benötigen, schmaler wird. Das heißt aber, dass die mittleren 90% (oder beliebige X%) nun einen viel kleineren Wertebereich umfassen und damit – bei gleicher Wahrscheinlichkeit – die Grenzen enger zusammenrücken:

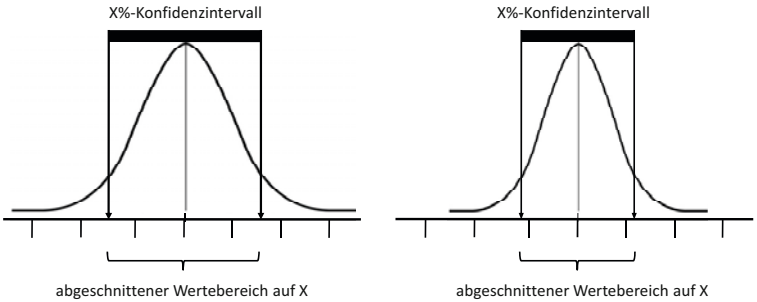

Abbildung 2.4 Dasselbe X%-Konfidenzintervall bei kleiner (links) und großer Stichprobengröße (rechts)

Wie man sehen kann, verteilen sich die wahrscheinlichen Mittelwerte in der Stichprobenverteilung nun auf einen viel kleineren Wertebereich. Folglich schneidet auch das Konfidenzintervall einen viel kleineren Wertebereich ab. Es ist also wesentlich informativer als das Intervall, das aus der kleineren Stichprobe resultierte. Damit gilt also auch hier: je größer die untersuchte Stichprobe, desto besser.

2.3 Standardfehler und Konfidenzintervalle für Anteile

Wir haben uns die Berechnungen für den Standardfehler und für die Konfidenzintervalle bisher für Mittelwerte angesehen und wollten daraus eine Aussage über die Güte der Schätzung von Stichprobenmittelwerten auf Populationsmittelwerte ableiten. Dieses Vorgehen kann nun ebenso auf *Anteile* (bzw. *Prozentwerte*) angewendet werden. Auch für empirisch gefundene Anteile kann man sich die Frage stellen, inwieweit diese auf die Population verallgemeinerbar sind. Nehmen wir an, wir möchten herausfinden, wie groß der Anteil der Vegetarier in Deutschland ist. Wir ziehen eine zufällige Stichprobe von 100 Personen und finden einen Anteil von 15% Vegetariern. Dieser Anteil dient uns auch hier als *Erwartungswert* für den *wahren Anteil* in der Population. Wir würden also schätzen (verallgemeinern), dass 15% der Deutschen Vegetarier sind. Und wir möchten auch hier wieder angeben, wie sehr wir dieser Schätzung trauen können. Dafür können wir – genau wie bei Mittelwerten – entweder den Standardfehler oder ein Konfidenzintervall benutzen.

Auch für Anteile kann man Stichprobenverteilungen konstruieren. Dabei werden in der Verteilung nicht einzelne Mittelwerte abgetragen, sondern einzelne Anteile. Im Vegetarier-Beispiel müsste die Stichprobenverteilung also alle möglichen Anteile von 0 bis 100 Prozent enthalten. Und sie müsste um den erwarteten Anteil von 15% konstruiert werden. Die Stichprobenverteilung, die dabei entsteht, nennt man *Binomialverteilung*. Dieser Name kommt daher, dass das untersuchte Merkmal in nur zwei möglichen Ausprägungen vorliegt (*binom*: lateinisch = in zwei Formen): hier also Vegetarier und Nicht-Vegetarier. Die Binomialverteilung geht aber schon bei sehr kleinen Stichproben in die Standardnormalverteilung über, sodass so gut wie immer diese Verteilung benutzt wird. Das Vorgehen ist dabei identisch mit dem bei der Bestimmung von Konfidenzintervallen für Mittelwerte. Was wir dafür aber noch brauchen, ist der Standardfehler. Dieser ist bei Anteilen nur von der Stichprobengröße n und der Wahrscheinlichkeit p abhängig, mit der ein bestimmtes Ereignis zu erwarten ist. Das „Ereignis" ist in unserem Fall die Merkmalsausprägung „Vegetarier" (denn deren Anteil wollten wir bestimmen). Wie groß ist die Wahrscheinlichkeit, aus der Population einen Vegetarier zu ziehen? Da wir in unserer Stichprobe einen Anteil von 15% gefunden haben und diesen als Schätzung verwenden müssen, beträgt die Wahrscheinlichkeit hier 15 von 100, also 15% bzw. $p = 0{,}15$. Mit diesen beiden Größen kann der Standardfehler für Anteile σ berechnet werden:

$$\sigma = \sqrt{np(1-p)}$$

Für unser Beispiel beträgt er: $\sigma = \sqrt{100 \cdot 0{,}15 \cdot (1-0{,}15)} = 3{,}57$. Wenn wir den Anteil von Vegetariern in Deutschland mit 15% schätzen, so ist diese Schätzung also mit einem Standardfehler von 3,57% behaftet. Mit diesem Standardfehler können wir nun wieder unser Konfidenzintervall bestimmen. Wir entscheiden uns für eine Vertrauenswahrscheinlichkeit von 95% und können daher wieder den z-Wert von oben benutzen – der betrug 1,96:

$$untere\ Grenze: Anteil - \sigma \cdot z_{Konf.} = 15\% - 3{,}57 \cdot 1{,}96 = 8{,}0\%$$

$$obere\ Grenze: Anteil + \sigma \cdot z_{Konf.} = 15\% + 3{,}57 \cdot 1{,}96 = 22{,}0\%$$

Wir können also zu 95 Prozent sicher sein, dass das Intervall von 8% bis 22% den wahren Anteil von Vegetariern in Deutschland überdeckt.

Literaturempfehlung:

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 11)

3 Inferenzstatistische Aussagen für Zusammenhangs- und Unterschiedshypothesen

3.1 Hypothesentesten

Mittelwerte und Anteile sind relativ einfache Angaben, die man über empirisch gewonnene Daten macht. Im vorigen Kapitel haben wir gesehen, wie man die Güte einer Verallgemeinerung von einer Stichprobe auf eine Population einschätzen und mit Hilfe von Zahlen ausdrücken kann. Neben der bloßen Angabe von Mittelwerten oder Anteilen ist ein Forscher jedoch fast immer bestrebt, auch Aussagen über bestimmte Hypothesen zu treffen. Solche Hypothesen beziehen sich immer entweder auf Zusammenhänge zwischen Variablen oder auf Unterschiede zwischen bestimmten Gruppen. Wie Sie aus dem Prozess der Erkenntnisgewinnung wissen, sind Hypothesen (oder Annahmen) einfache Aussagen, die sich aus einer Theorie ableiten. Und manchmal hat man es statt mit einer Hypothese einfach mit einer Fragestellung zu tun, die sich nicht direkt aus einer Theorie ableitet aber dennoch entweder eine Fragestellung zu einem Unterschied oder zu einem Zusammenhang ist.

Unterschiedshypothesen (oder im weiteren Sinne *Unterschiedsfragestellungen*) beziehen sich dabei in der Regel auf einen Unterschied in der Lage zweier Gruppen bzw. Stichproben, und mit der Lage ist in der Regel immer der Mittelwert gemeint. Beispielsweise kann man sich fragen, ob es einen Unterschied im Ergebnis eines Leistungstests gibt zwischen Personen, die ein Training durchlaufen haben und Personen, die nicht am Training teilgenommen haben. Für beide Gruppen von Personen würde man den Mittelwert des Testergebnisses bestimmen und diese beiden Mittelwerte dann miteinander vergleichen. Sie betrachten also den Mittelwertsunterschied.

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_4,

Zusammenhangshypothesen (oder *Zusammenhangsfragestellungen*) beziehen sich hingegen auf den Zusammenhang zweier Variablen, und mit diesem Zusammenhang ist in der Regel die Korrelation gemeint. Für den Zusammenhang von Aggressivität und dem Spielen von gewaltorientierten Computerspielen würde man zum Beispiel die Ausprägung beider Variablen an ein und derselben Stichprobe von Personen untersuchen und prüfen, ob beide Variablen in einem statistischen Zusammenhang stehen.

Für beide Arten von Hypothesen oder Fragestellungen stellt sich im Zuge der Inferenzstatistik nun ebenfalls die Frage, wie man einen empirisch gefundenen Mittelwertsunterschied oder einen empirisch gefundenen Zusammenhang auf die Population verallgemeinern kann. Es besteht stets das Risiko, dass die Effekte in einer Stichprobe nur durch Zufall zustande gekommen und daher nicht auf die Population übertragbar sind. Wir müssen also auch für gefundene Mittelwertsunterschiede und Zusammenhänge angeben, wie sehr wir einer Verallgemeinerung auf die Population trauen können. Dafür gibt es wieder mehrere Möglichkeiten: die Berechnung von Standardfehlern und Konfidenzintervallen, die Sie nun schon kennen, und außerdem die Durchführung von Signifikanztests. Wir werden uns in diesem Kapitel mit der grundlegenden Idee dieser drei Vorgehensweisen beschäftigen und in späteren Kapiteln genauer auf die Testverfahren eingehen, die für spezifische Hypothesen und Fragestellungen entwickelt wurden. Wir werden dabei nicht ständig von Mittelwertsunterschieden und Zusammenhängen sprechen, sondern allgemeiner von *Effekten*. Sehen wir uns zunächst an, was mit einem Effekt gemeint ist.

Was ist ein Effekt?

Der Begriff Effekt bezieht sich immer auf eine unabhängige Variable, die eine bestimmte Wirkung (einen Effekt) auf eine abhängige Variable ausüben soll. Oben haben wir von einem Training gesprochen, das einen Effekt auf das Ergebnis in einem Leistungstest ausüben soll. Im Idealfall sollten die Teilnehmer des Trainings nachher bessere Testwerte erreichen als Personen, die nicht teilgenommen haben. Was auch immer als unabhängige Variable in Frage kommt (ein Training, eine Therapie, eine Manipulation jeglicher Art) – wir sind immer an ihrem Effekt interessiert, den wir anhand der abhängigen Variable messen können. Mit „Manipulation“ kann dabei auch gemeint sein, dass man einfach

verschiedene Personengruppen untersucht. In einer Studie zum Glücksempfinden haben van Boven und Gilovich (2003) zum Beispiel Personen, die ihr Geld eher in materiellen Dingen wie Autos anlegen, mit Personen verglichen, die ihr Geld eher in Erfahrungen wie Urlaubsreisen investiert haben. Das Ergebnis sehen Sie in Abbildung 3.1.

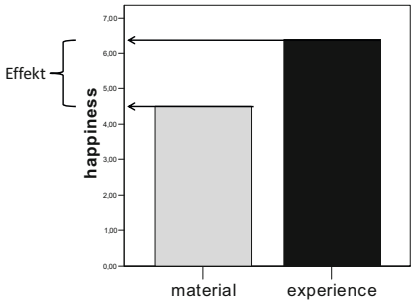

Abbildung 3.1 Glücksempfinden bei Personen, die ihr Geld eher in materielle Dinge oder in Erfahrungen investieren (nach van Boven & Gilovich, 2003)

Der Effekt besteht hier also einfach in dem Unterschied des Glücksempfindens beider untersuchter Gruppen. In gleicher Weise gelten auch Zusammenhänge als Effekte. Wenn zwei Variablen miteinander korrelieren, so beschreibt diese Korrelation den Effekt. Alternativ kann man sich den Zusammenhang auch so vorstellen, dass die eine Variable einen „statistischen Effekt" (nicht zwangsläufig einen kausalen!) auf die andere Variable ausübt, da sie diese – wenn ein Zusammenhang vorliegt – natürlich vorhersagen kann. Im Folgenden wollen wir stets Aussagen über Effekte in der Population treffen, die wir aus den Effekten, die wir in einer Stichprobe gefunden haben, schätzen wollen.

Abhängige und unabhängige Messungen

Beim Hypothesentesten geht es nahezu immer um *mindestens zwei* Messungen. Bei einer Unterschiedshypothese geht es um den Unterschied zweier Mittelwerte, und bei einer Zusammenhangshypothese geht es um die Korrelation

zweier Variablen. Entscheidend für alle weiteren Berechnungen ist nun, ob diese Messungen abhängig oder unabhängig voneinander vorgenommen wurden. Was bedeutet das? Unabhängig sind Messungen dann, wenn jede Messung an einer eigenen Stichprobe bzw. in einer eigenen Gruppe vorgenommen wurde. Dieses Prinzip kennen Sie schon von den verschiedenen Designs von Studien. Beim *between-subjects* Design wurden die Studienteilnehmer randomisiert auf die verschiedenen Versuchsgruppen aufgeteilt. Wenn wir beispielsweise untersuchen wollen, ob Bilder länger im Gedächtnis bleiben als Töne, könnten wir zwei Versuchsgruppen bilden und die uns zur Verfügung stehenden Personen randomisiert diesen Gruppen zuweisen. Da die Personen dieser beiden Gruppen nichts miteinander zu tun haben, sind sie *unabhängig* voneinander. So sind auch beide Messungen – also die Messung der Gedächtnisleistung in Gruppe 1 und 2 – unabhängig voneinander.

Alternativ hätten wir aber auch dieselben Personen beide Bedingungen durchlaufen lassen können. Bei einem solchen *within-subjects* Design hätten die Personen erst den Gedächtnistest für das Bild und später den für den Ton absolviert. Da es sich hier um dieselben Personen handelt, liegt eine *Messwiederholung* vor: das Merkmal wird an derselben Stichprobe wiederholt gemessen *(repeated measurement)*. Messwiederholungen sind immer *abhängige* Messungen, da stets dieselben Personen die Messwerte generieren. Es gibt noch einen zweiten Fall von abhängigen Messungen, bei dem man versucht, so nahe wie möglich an ein Messwiederholungsdesign heran zu kommen. Zur Kontrolle von Störvariablen wie Alter, Geschlecht oder Intelligenz versucht man oft, diese in den verschiedenen Versuchsgruppen konstant zu halten. Das heißt, es befinden sich zwar unterschiedliche Personen in den Gruppen, diese sind aber so ausgewählt, dass sie jeweils gleiche Ausprägungen in den potenziellen Störvariablen haben. Zu jeder Person in Gruppe 1 versucht man also eine entsprechende Person in Gruppe 2 zu bekommen, die gleich alt ist, das gleiche Geschlecht hat usw. Dieser Vorgang wird als *matching* bezeichnet und führt zu gematchten oder *gepaarten* Stichproben bzw. Messungen (da immer ein Paar von Personen mit gleichen Ausprägungen der Störvariablen gesucht wird). Auch gepaarte Stichproben führen zu abhängigen Messungen, da die Personen nicht mehr rein zufällig in den Gruppen landen.

Messwiederholungen und gepaarte Stichproben (matching) führen zu abhängigen Messungen bzw. Stichproben. Sind die Versuchsteilnehmer rein zufällig den verschiedenen Messungen bzw. Stichproben zugeordnet, sind diese unabhängig.

Bei Zusammenhangshypothesen hat man es folglich immer mit abhängigen Messungen zu tun, denn hier werden zwei Merkmale immer an denselben Personen untersucht. Verständlicherweise würde es keinen Sinn machen, die eine Variable an der einen und die andere Variable an einer anderen Stichprobe zu erheben, da dann die Messwerte nicht für beide Variablen gleichzeitig zugeordnet werden könnten. Anders ausgedrückt: man könnte kein Streudiagramm konstruieren, da man nicht weiß, welchem x-Wert man jeweils welchen y-Wert zuordnen soll.

Stichprobenverteilungen bei Unterschieden und Zusammenhängen

Auch bei der Schätzung von Unterschieden und Zusammenhängen in der Population – wir sprechen allgemein von *Populationseffekten* – benötigen wir wieder die Überlegung, was passieren würde, wenn wir nicht nur einmal eine Stichprobe ziehen und den Effekt bestimmen, sondern gleich mehrmals. Diese Überlegung steckt in der Stichprobenverteilung, die alle möglichen Effekte enthält, die wir bei Ziehungen von Stichproben aus der Population finden könnten. Eine solche Stichprobenverteilung können wir auch für Unterschieds- und Zusammenhangsmaße konstruieren. Im Vergleich zur Konstruktion von Stichprobenverteilungen für Mittelwerte ändert sich dabei wenig. Der einzige Unterschied besteht darin, dass nun nicht mehr einzelne Mittelwerte in der Verteilung abgetragen werden, sondern Mittelwertsunterschiede oder Zusammenhänge. Betrachten wir ein Beispiel: Wir wollen untersuchen, ob ein Reaktionszeittraining einen Effekt auf die Reaktionszeit in einer Entscheidungsaufgabe hat, in der es auf schnelle Entscheidungen ankommt. Wir bilden dafür zwei Gruppen: eine, die das Training durchläuft und eine Kontrollgruppe, die das Training nicht durchläuft. Wir wollen nun den Effekt des Trainings auf die Leistung in dem Test untersuchen und bestimmen dafür die beiden Mittelwerte der Testleistung. Die Differenz der beiden Mittelwerte x_{diff} ist genau der Effekt, der uns interes-

siert. In unserer Stichprobenverteilung müssen daher alle möglichen Differenzen abgetragen sein, die man hätte finden können. Wenn wir eine Differenz in der Reaktionszeit von 300 Millisekunden gefunden haben, würden wir diese Differenz wieder als Erwartungswert für die Differenz in der Population benutzen. Die Stichprobenverteilung, die um diesen erwarteten Mittelwertsunterschied aufgespannt wird, kann dann etwa so aussehen wie in Abbildung 3.2.

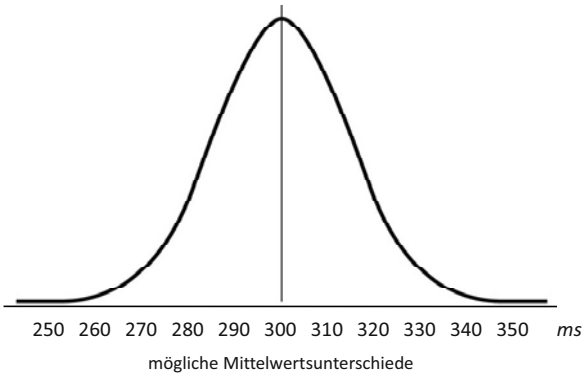

Abbildung 3.2 Mögliche Stichprobenverteilung für einen Mittelwertsunterschied von 300 ms

Auf der X-Achse sind übrigens nicht „alle möglichen" Mittelwertsunterschiede dargestellt, sondern nur diejenigen, die eine gewisse Wahrscheinlichkeit besitzen. Da sich die Kurve der Verteilung asymptotisch der X-Achse nähert, erstrecken sich die „möglichen" Mittelwertsunterschiede im Prinzip von minus bis plus unendlich. Wir sehen aber an der Kurve, dass wir mit sehr hoher Wahrscheinlichkeit Werte finden werden, die – in diesem Beispiel – irgendwo zwischen 250 und 350 ms liegen.

In gleicher Weise könnten wir auch den Zusammenhang zweier Variablen bestimmen und diesen mit Hilfe des Korrelationskoeffizienten *r* ausdrücken. Und auch für den Korrelationskoeffizienten gibt es eine entsprechende Stichprobenverteilung, aus der man ablesen kann, wie wahrscheinlich es ist, einen bestimmten Wert zu ziehen. Die Standardabweichung dieser Stichprobenverteilungen – die uns als Standardfehler dienen soll – kann auch bei Unterschieden und Zusammenhängen aus den Streuungen in den einzelnen Stichproben be-

rechnet werden. Sehen wir uns also zunächst an, wie man den Standardfehler bestimmen kann.

Literaturempfehlung

Bortz, J. und Döring, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer. (Kapitel 8)
Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 9)

3.2 Der Standardfehler

Um einschätzen zu können, wie sehr man der Verallgemeinerung eines in einer Stichprobe gefundenen Effektes auf die Population trauen kann, ist der Standardfehler wieder die einfachste Möglichkeit. Um es jedoch gleich vorweg zu nehmen: anders als bei Mittelwerten ist es bei Zusammenhängen und Unterschieden nicht üblich, allein den Standardfehler anzugeben. Stattdessen wird hier oft auf die beiden anderen Möglichkeiten zurückgegriffen: Konfidenzintervalle und Signifikanztests, mit denen wir uns gleich beschäftigen. Der Grund dafür liegt darin, dass es beim Testen von Hypothesen um Entscheidungen geht. Das heißt, dass man sich entweder für oder gegen eine Hypothese entscheiden will bzw. muss. Soll ich Medikament A gegenüber Medikament B bevorzugen, weil es besser wirkt? Soll ich Vitamin C nehmen, weil es die Immunabwehr verbessert? Solche Fragen haben immer mit Entscheidungen zu tun, die aufgrund von Studien getroffen werden müssen. Dafür ist es oft nicht ausreichend, zum Beispiel für einen Mittelwertsunterschied den Standardfehler anzugeben, da er – besonders für Laien – schwer zu interpretieren ist. Dagegen liefern Konfidenzintervalle und Signifikanztests eher praktische Entscheidungshilfen.

Dennoch wird auch für diese beiden Methoden der Standardfehler benötigt. Sehen wir uns an, wie er bei den verschiedenen Fragestellungen im Prinzip berechnet wird.

Standardfehler für Mittelwertsunterschiede bei unabhängigen Messungen

Für Mittelwertsunterschiede bei unabhängigen Messungen wird der Standardfehler wieder aus der Streuung der einzelnen Stichproben berechnet. Hier muss natürlich die Streuung einer jeden Stichprobe einbezogen werden, die in die Untersuchung eingeflossen ist. Gemeint ist damit die Fehlerstreuung, die in jeder Gruppe dafür sorgt, dass die Messwerte variieren, ohne dass es dafür einen systematischen Grund gibt. Abbildung 3.3 zeigt, wie die Gesamtstreuung aller Personen in einen systematischen und einen unsystematischen Anteil aufgeteilt wird.

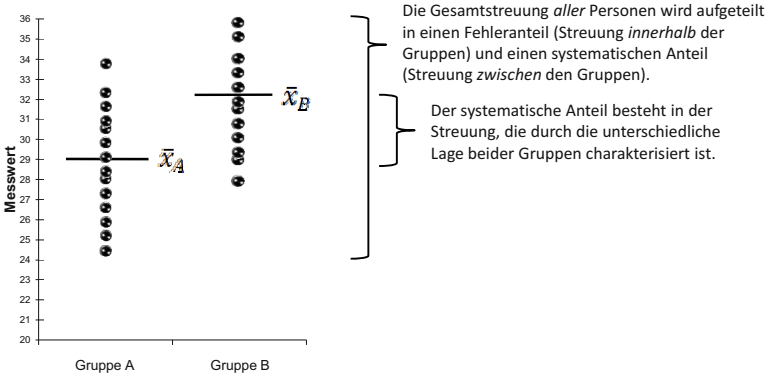

Abbildung 3.3 Veranschaulichung unabhängiger Messungen

Der Anteil an der Gesamtstreuung, der uns interessiert, ist die Streuung *zwischen den Gruppen.* Diese kommt durch den Mittelwertsunterschied $\bar{x}_A - \bar{x}_B$ zustande. Das ist genau der Effekt, um den es uns geht. Wie wir sehen, ist dieser Effekt aber von einem „Rauschen" umgeben, nämlich der Fehlerstreuung (oder auch Streuung *innerhalb der Gruppen*). Diese ist für uns nicht erklärbar und schmälert natürlich die Bedeutsamkeit des gefundenen Effektes. Denn bei sehr großer Fehlerstreuung kann ein solcher Mittelwertsunterschied auch zufällig zustande kommen. Die Fehlerstreuung kann in beiden Gruppen unterschiedlich groß sein. Wenn wir also zwei Gruppen vergleichen, müssen wir die Streuungen der Messwerte in beiden Gruppen berücksichtigen: für Gruppe A ($\hat{\sigma}^2_{\bar{x}_A}$) und für Gruppe B ($\hat{\sigma}^2_{\bar{x}_B}$). Der Mittelwertsunterschied ergibt sich aus der einfachen Differenz der Mittelwerte: $\bar{x}_A - \bar{x}_B$.

Bei gleichen Gruppengrößen ($n_A = n_B$) berechnet sich der Standardfehler dieses Mittelwertsunterschiedes wie folgt:

$$\hat{\sigma}_{\bar{x}_A - \bar{x}_B} = \sqrt{\hat{\sigma}^2_{\bar{x}_A} + \hat{\sigma}^2_{\bar{x}_B}}$$

Dieser Standardfehler gibt hier an, mit welchem „durchschnittlichen" Fehler die Schätzung eines Mittelwertsunterschiedes in der Population behaftet ist. Auch hier gilt, dass er umso kleiner ist, je größer die untersuchten Stichproben sind.

Standardfehler für Mittelwertsunterschiede bei abhängigen Messungen

Für abhängige Messungen sieht die Berechnung des Standardfehlers etwas anders aus. Der Grund dafür liegt darin, dass die Berechnung der Streuung nicht für beide Messungen einzeln erfolgt, sondern für die *Differenz* der Messwerte. Was damit gemeint ist, sehen wir an Abbildung 3.4.

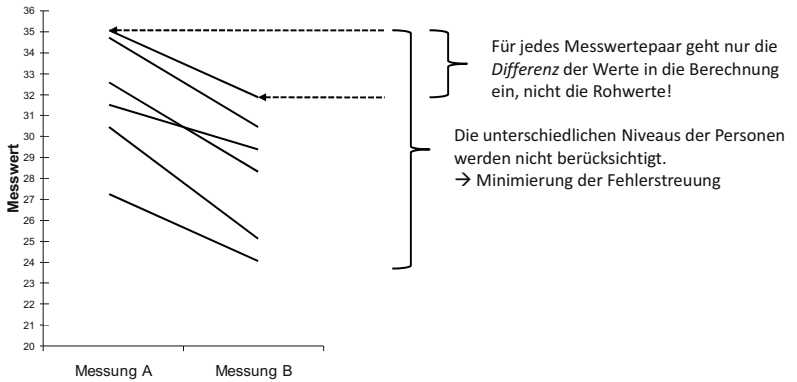

Abbildung 3.4 Veranschaulichung abhängiger Messungen

Wenn wir davon ausgehen, dass dieselben Personen zu zwei Messzeitpunkten A und B getestet werden, dann interessiert uns am Ende nur, ob sich *pro Person* ein Unterschied zwischen der ersten und der zweiten Messung ergeben hat. Wir prüfen also, ob sich für jede einzelne Person der Messwert vergrößert oder

verkleinert hat oder ob er gleich geblieben ist. Uns interessiert jetzt sozusagen die Streuung *innerhalb* der Personen. Danach bilden wir den Durchschnitt aller Differenzen über alle Personen hinweg. Allein diese Differenzen sind für den Effekt der Messwiederholung entscheidend. Es ist dabei völlig irrelevant, auf welchem Niveau sich diese Veränderungen abspielen. Damit ist auch die Streuung *zwischen* den Personen innerhalb eines jeden Messzeitpunktes nicht von Bedeutung. Das bedeutet aber, dass die Fehlerstreuung, die den Effekt umgibt, nun nur in der Streuung der Differenzen besteht. Die Idee dahinter ist relativ einfach: Wenn die Messwiederholung einen Effekt haben soll – also zum Beispiel in Messung B niedrigere Werte erwartet werden – dann sollten alle Differenzen in die gleiche Richtung gehen und möglichst auch gleich groß sein. Wenn bei einigen Personen die Differenz nicht so groß ist oder die Messwerte sogar steigen, vergrößert das die Fehlervarianz. Das wiederum macht einen gefundenen Effekt weniger bedeutsam. Diese Streuung der Differenzen lässt sich berechnen, indem man zunächst von jedem Differenzwert $diff_i$ den Mittelwert aller Differenzwerte abzieht und diese Differenzen quadriert: $(x_{diff_i} - \bar{x}_{diff})^2$. Das macht man für jedes Messwertpaar, also n Mal (n bezeichnet hier die Anzahl der Messwertpaare) und summiert alle Werte auf. Anschließend teilt man wieder durch $n - 1$:

$$\hat{\sigma}_{diff} = \sqrt{\frac{\sum(x_{diff_i} - \bar{x}_{diff})^2}{n - 1}}$$

Der Standardfehler des Mittelwertsunterschiedes kann wieder aus dieser Streuung berechnet werden:

$$\hat{\sigma}_{\bar{x}_{diff}} = \frac{\hat{\sigma}_{diff}}{\sqrt{n}}$$

Standardfehler für Zusammenhänge

Bei Zusammenhängen von zwei Variablen beschreibt der Korrelationskoeffizient r die Enge des Zusammenhangs. Dieser ist dann groß, wenn sich die Werte in einem Streudiagramm um eine Gerade konzentrieren. Damit ist in diesem Koeffizienten die Streuung der Werte bereits enthalten. Sie ist groß, wenn r

klein ist und umgekehrt. Der Standardfehler für den Korrelationskoeffizienten r berechnet sich daher wie folgt:

$$\hat{\sigma}_r = \frac{(1-\rho^2)}{\sqrt{n-1}}$$

Als Erwartungswert für den Korrelationskoeffizienten in der Population ρ wird wiederum der gefundene Korrelationskoeffizient r benutzt. An der Formel ist erkennbar, dass Korrelationskoeffizienten nahe 1 (bzw. -1) zu einem Standardfehler führen, der gegen 0 geht.

Wenn zwischen zwei Variablen eine Korrelation besteht, dann können wir diese benutzen, um die eine Variable aus der anderen vorherzusagen. Das war die Aufgabe der Regressionsrechnung. Die entscheidende Größe, die bei der Regression die Stärke der Vorhersagbarkeit der einen Variable für die andere angibt, war das Regressionsgewicht b bzw. dessen standardisierte Variante β. Auch für das Regressionsgewicht können wir einen Standardfehler angeben. Im Gegensatz zum Korrelationskoeffizienten beschreibt das Regressionsgewicht aber nicht die *absolute* Enge des Zusammenhangs, sondern den *relativen Einfluss* einer Variable auf die andere. Mit relativem Einfluss ist gemeint, dass es weitere Variablen geben kann, die für die Vorhersage der Variable Y in Frage kommen. Jede dieser Vorhersagevariablen (wir hatten sie Prädiktoren genannt) würde ein eigenes Regressionsgewicht erhalten. Aus einem Regressionsgewicht allein können wir daher nicht sofort die Güte der gesamten Vorhersage ableiten. Wie Sie wissen, beschreibt ein Regressionsgewicht den Anstieg der Regressionsgerade. Wie gut die Vorhersage von Y durch X funktioniert, können wir durch die Streuung der Werte um die Gerade herum feststellen. Bei der Regression hatten wir diese Streuung auch als Residuen bezeichnet. Diese durchschnittliche Streuung äußert sich im sogenannten *Standardschätzfehler*.

> Der Standardschätzfehler bei der Regression gibt an, wie stark die Werte um die Regressionsgerade streuen. Er ist damit ebenso ein Gütemaß für die Vorhersage von Y aus X.

Der Standardschätzfehler kann daher als alternatives Gütemaß zum bereits bekannten Determinationskoeffizient r^2 benutzt werden. Er beschreibt die Ungenauigkeit, die entsteht, wenn man Y-Werte aus X-Werten mit Hilfe der Regressionsgeraden schätzen möchte. Diese ist natürlich umso kleiner, je näher die

Werte an der Geraden liegen. Der Standardschätzfehler berechnet sich folgendermaßen:

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{n \cdot s_y^2 - n \cdot b_{yx}^2 \cdot s_x^2}{n-2}}$$

Aus diesem Standardschätzfehler kann man nun wiederum den Standardfehler des Regressionsgewichtes *b* berechnen:

$$\hat{\sigma}_{b_{yx}} = \frac{\hat{\sigma}_{(y|x)}}{s_x \cdot \sqrt{n}}$$

Dieser Standardfehler wird von Statistikprogrammen zu jedem Regressionsgewicht *b* mit angegeben, meist unter der Bezeichnung SE.

3.3 Konfidenzintervalle

Wie eben schon erwähnt, geht es beim Testen von Hypothesen meist um Entscheidungen für oder gegen eine Hypothese. Um statistische Entscheidungshilfen zu erhalten, ist der Standardfehler weniger geeignet. Konfidenzintervalle liefern hier mehr Information, und erst hier wird auch der große Vorteil von Konfidenzintervallen richtig deutlich. Denn sie liefern eine relativ leicht verständliche Angabe darüber, ob ein Effekt möglicherweise durch Zufall gefunden wurde oder ob er „von statistischer Bedeutung" ist – man sagt auch *signifikant* oder *substanziell* oder *systematisch* (diese Begriffe werden meist synonym verwendet).

Konfidenzintervalle für Mittelwertsunterschiede bei unabhängigen Stichproben

Beginnen wir mit Konfidenzintervallen für Mittelwertsunterschiede bei unabhängigen Messungen. Deren Konstruktion ist im Prinzip identisch zur Konstruktion solcher Intervalle bei Mittelwerten. Als Stichprobenverteilung verwenden wir aber nicht die Verteilung einzelner Mittelwerte, sondern wieder die Stichprobenverteilung für Mittelwertsunterschiede. Alternativ können wir auch hier wieder den „Umweg" über die *t*-Verteilung gehen, denn auch Mittelwertsunterschiede sind *t*-verteilt (bzw. standardnormalverteilt bei Stichproben

ab ca. 30 Personen). Wenn wir einen Mittelwertsunterschied und dessen Standardfehler berechnet haben, können wir diese Informationen – zusammen mit dem in der *t*-Verteilung abgelesenen Wert für die Grenzen des Intervalls – nutzen, um die beiden Intervallgrenzen zu berechnen:

$$untere\ Grenze: (\bar{x}_A - \bar{x}_B) - \hat{\sigma}_{\bar{x}_A - \bar{x}_B} \cdot t_{df,Konf.}$$

$$obere\ Grenze: (\bar{x}_A - \bar{x}_B) + \hat{\sigma}_{\bar{x}_A - \bar{x}_B} \cdot t_{df,Konf.}$$

Das Intervall gibt wieder die Wahrscheinlichkeit an, mit der der Bereich zwischen den beiden Grenzen den wahren Mittelwertsunterschied in der Population überdeckt. Beim Testen von Hypothesen ist diese Information nun außerordentlich wertvoll. Denn die spannende Frage ist hier, ob das Intervall den Wert 0 beinhaltet. Der Wert 0 würde bedeuten, dass es in der Population keinen Mittelwertsunterschied gibt. In diesem Fall müssten wir also unsere Hypothese verwerfen und stattdessen davon ausgehen, dass der Mittelwertsunterschied, den wir gefunden haben, nur zufällig zustande kam. Sehen wir uns diese Überlegung an einem Beispiel an. Oben haben wir von einem Training gesprochen, bei dem wir die Hypothese hatten, dass es zu einer Verbesserung in der Testleistung führt. Nehmen wir an, wir hätten hinsichtlich der Testleistung von Trainingsgruppe und Kontrollgruppe einen Mittelwertsunterschied von 10 Punkten gefunden (bei einer Höchstpunktzahl von 100 Punkten). Dabei geht der Mittelwertsunterschied bereits in die richtige Richtung, das heißt, die Trainingsgruppe hat 10 Punkte mehr als die Kontrollgruppe und nicht umgekehrt. Wir würden nun den Standardfehler bestimmen und beispielsweise ein 90%-Konfidenzintervall um den Mittelwertsunterschied herum konstruieren (siehe Abbildung 3.5).

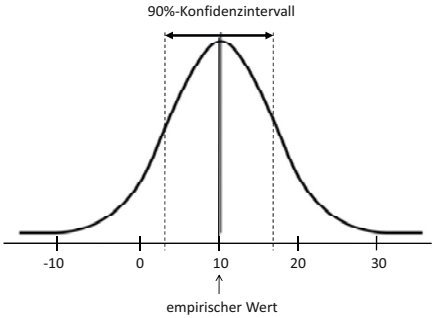

Abbildung 3.5 90%-Konfidenzintervall um einen Mittelwertsunterschied von 10 Punkten

Betrachten wir zunächst die entsprechende Stichprobenverteilung, die um den Mittelwertsunterschied von 10 Punkten herum entstanden ist. Das Entscheidende an dieser Stichprobenverteilung ist, dass sie den Wert 0 beinhaltet und auch über ihn hinausgeht – in den negativen Bereich hinein. Negative Werte – und damit negative Mittelwertsunterschiede – bedeuten nichts anderes, als dass der Mittelwertsunterschied in die entgegengesetzte Richtung zeigt. Bei der Berechnung von Mittelwertsunterschieden zieht man normalerweise den vermeintlich kleineren Wert vom größeren ab. Das ist allerdings keine Regel, sondern ist dem Anwender überlassen. In unserem Beispiel haben wir den Mittelwert der Kontrollgruppe vom Mittelwert der Trainingsgruppe abgezogen und eine Differenz von 10 erhalten. Hätten wir mit unserer Hypothese aber daneben gelegen und die Kontrollgruppe hätte besser abgeschnitten, hätten wir eine negative Differenz erhalten. Die Stichprobenverteilung in Abbildung 3.5 sagt uns nun, dass wir auch solche negativen Differenzen hätten ziehen können. Das Ziehen einer negativen Differenz müsste uns natürlich dazu veranlassen, unsere Hypothese zu verwerfen. Hier wird die Kernfrage der Inferenzstatistik beim Hypothesentesten deutlich: Kann ich meinen gefundenen Mittelwertsunterschied guten Gewissens auf die Population verallgemeinern und behaupten, dass meine Hypothese richtig war? Das Konfidenzintervall soll uns helfen, diese Entscheidung zu treffen. Wie wir sehen, überdeckt unser Konfidenzintervall den Wert 0 nicht. Wir können also zu 90% sicher sein, dass unser Intervall, das den Wert 0 nicht enthält, den wahren Wert in der Population überdeckt. Unser Effekt (der gefundene Mittelwertsunterschied) kann damit als ein bedeutsamer oder

signifikanter Effekt (wohlbemerkt bei einer Vertrauenswahrscheinlichkeit von 90%) betrachtet werden. Hier wird auch deutlich, dass beim Hypothesentesten die tatsächlichen Werte der Intervallgrenzen gar nicht so sehr von Interesse sind, sondern vielmehr die Frage, ob das Intervall die 0 enthält oder nicht. In unserem Beispiel liefert damit das Konfidenzintervall eine schnelle Entscheidung, die darin besteht, dass wir unsere Hypothese annehmen und auf die Population verallgemeinern können.

Was passiert nun, wenn das Intervall den Wert 0 enthält? Dafür müssen wir uns zunächst fragen, wie es dazu überhaupt kommen kann. Zwei Möglichkeiten kommen dafür in Frage. Die erste besteht darin, dass eine Vergrößerung der Vertrauenswahrscheinlichkeit zu einer Verbreiterung des Intervalls führt, und damit steigt natürlich das Risiko, dass das Intervall den Wert 0 überdeckt. Hätten wir in unserem Beispiel als Vertrauenswahrscheinlichkeit nicht 90, sondern 95% gefordert, wären die Grenzen entsprechend weiter auseinander gerückt (siehe Abbildung 3.6).

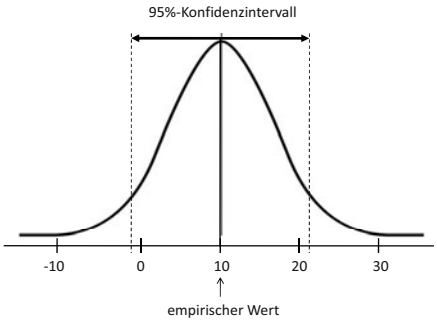

Abbildung 3.6 95%-Konfidenzintervall um einen Mittelwertsunterschied von 10 Punkten

Die höhere Vertrauenswahrscheinlichkeit hat dazu geführt, dass das Intervall nun den Wert 0 enthält. Das bedeutet, dass wir bei einer Vertrauenswahrscheinlichkeit von 95% unsere Hypothese verwerfen müssen, weil wir nicht mehr „genügend sicher" sein können, dass unser Effekt auch in der Population vorhanden ist. Bei dieser großen Vertrauenswahrscheinlichkeit ist der Effekt also nicht mehr signifikant oder bedeutsam. Auch beim Testen von Hypothesen gilt damit, dass die Festlegung einer geeigneten Vertrauenswahrscheinlichkeit eine

Kompromissentscheidung ist. Wir werden uns später noch genauer damit beschäftigen, wie man solche sogenannten Signifikanzniveaus festlegen kann.

Die zweite Möglichkeit dafür, dass ein Konfidenzintervall die 0 enthält, besteht darin, dass der gefundene Mittelwertsunterschied ohnehin nahe 0 liegt. Hätten wir im Beispiel nur eine Differenz von 5 Punkten gefunden, dann läge die Mitte der Stichprobenverteilung näher an 0 und vermutlich hätte dann auch das 90%-Konfidenzintervall die 0 überdeckt. Je kleiner also der gefundene Effekt, desto eher wird ein um ihn herum konstruiertes Konfidenzintervall den Wert 0 beinhalten. Kleine Effekte können sich also viel schwerer als signifikant erweisen. Wie man auch bei kleinen Effekten erreichen kann, dass ihre Konfidenzintervalle die 0 weniger wahrscheinlich überdecken, sollten Sie allerdings auch schon wissen: Wenn Sie die Stichprobengröße für Ihre Studie erhöhen, erhalten Sie eine schmalere Stichprobenverteilung, und das Konfidenzintervall zieht sich auf einen engeren Wertebereich zurück. Damit sinkt die Wahrscheinlichkeit dafür, dass es die 0 überdeckt.

Konfidenzintervalle für Mittelwertsunterschiede bei abhängigen Stichproben

Die Konstruktion von Konfidenzintervallen bei abhängigen Mittelwertsunterschieden ist im Prinzip identisch mit der bei unabhängigen Stichproben. Nur verwenden wir hier als Mittelwertsunterschied die mittleren Differenzen der Werte $\bar{x}_{diff}$ und den Standardfehler dieser Differenzen $\hat{\sigma}_{\bar{x}_{diff}}$ für die Berechnung des Intervalls:

$$untere\ Grenze: \bar{x}_{diff} - \hat{\sigma}_{\bar{x}_{diff}} \cdot t_{df,Konf.}$$

$$obere\ Grenze: \bar{x}_{diff} + \hat{\sigma}_{\bar{x}_{diff}} \cdot t_{df,Konf.}$$

Für die Interpretation ändert sich gegenüber den unabhängigen Stichproben nichts. Die entscheidende Frage ist wieder die, ob das Intervall bei einer festgelegten Vertrauenswahrscheinlichkeit den Wert 0 überdeckt oder nicht. Wenn das nicht der Fall ist, können wir von einem bedeutsamen Effekt der Messwiederholung ausgehen und unsere Hypothese, die von einem Unterschied ausging, annehmen.

Konfidenzintervalle für Zusammenhänge

Als statistische Kennwerte für den Zusammenhang zweier Variablen haben Sie zwei Maße kennengelernt: den Korrelationskoeffizienten *r* und das Regressionsgewicht β, das den relativen Einfluss einer Variable auf eine andere Variable beschreibt. Im Fall einer einfachen linearen Regression – wenn es also nur einen Prädiktor gibt – sind beide Maße identisch. Bei der Konstruktion von Konfidenzintervallen für Zusammenhänge kommt uns hier der Umstand zugute, dass beide Maße ebenfalls einer *t*-Verteilung folgen. Das Vorgehen ist damit prinzipiell wieder das gleiche wie bisher. Allerdings gibt es eine Besonderheit beim Korrelationskoeffizienten *r*. Für ihn ist die *t*-Verteilung nur dann symmetrisch, wenn $r = 0$. Das liegt daran, dass *r* bei 1 bzw. -1 seine Grenze hat und die Verteilung dort jeweils aufhören muss. Bei Koeffizienten größer 0 – und mit denen haben wir es fast immer zu tun – ist die *t*-Verteilung daher nicht mehr symmetrisch. Das Konfidenzintervall muss aber symmetrisch zu beiden Seiten des gefundenen Koeffizienten *r* aufgespannt werden. Bei einer unsymmetrischen Verteilung würde das natürlich zu verzerrten Werten führen. Aus diesem Grund werden Konfidenzintervalle für Korrelationskoeffizienten meist gar nicht berichtet. Wenn man sie dennoch berechnen möchte, kann man statt der *t*-Verteilung auf die *z*-Verteilung zurückgreifen. Diese ist auch für Korrelationskoeffizienten ungleich 0 symmetrisch. Das ist deswegen möglich, weil es eine Transformation (eine Berechnung) gibt, mit der sich Korrelationskoeffizienten in normalverteilte *z*-Werte umrechnen lassen. Die z_r-Werte, die sich bei dieser sogenannten *Fisher's z-Transformation* aus bestimmten Korrelationskoeffizienten ergeben, finden Sie ebenfalls in der Tabellensammlung. Nun spannt man um diesen Wert z_r das Konfidenzintervall auf, indem man den kritischen Wert für die Intervallgrenzen bei einer festgelegten Vertrauenswahrscheinlichkeit aus der *z*-Verteilung abliest. Der Standardfehler, der hier verwendet werden muss, beträgt $\frac{1}{\sqrt{N-3}}$ (das folgt rein rechnerisch aus der Transformation). Die Grenzen des Intervalls berechnen sich dann wie folgt:

$$untere\ Grenze: z_{r_{unten}} = z_r - \frac{1}{\sqrt{N-3}} \cdot z_{Konf}$$

$$obere\ Grenze: z_{r_{oben}} = z_r + \frac{1}{\sqrt{N-3}} \cdot z_{Konf}$$

Die beiden Werte für $z_{r_{unten}}$ und $z_{r_{oben}}$ können nun mit Hilfe der Fisher's-*z*-Tabelle wieder in Korrelationskoeffizienten zurückgerechnet werden. Diese sind die beiden Grenzen des Konfidenzintervalls für *r*.

Die Besonderheit unsymmetrischer Verteilungen gilt nicht für das Regressionsgewicht β. Hier kann das Konfidenzintervall wieder mit Hilfe der *t*-Verteilung bestimmt werden. Für die Berechnung benötigen wir das standardisierte Regressionsgewicht β, dessen Standardfehler $\hat{\sigma}_{\beta_{yx}}$ und den *t*-Wert der entsprechenden Vertrauenswahrscheinlichkeit:

$$untere\ Grenze\ f\ddot{u}r\ \beta: \beta - \hat{\sigma}_{\beta_{yx}} \cdot t_{df,Konf.}$$

$$obere\ Grenze\ f\ddot{u}r\ \beta: \beta + \hat{\sigma}_{\beta_{yx}} \cdot t_{df,Konf.}$$

Literaturempfehlung

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 11)

Abschließende Bemerkung zu den Berechnungen

Sie haben in den letzten beiden Kapiteln relativ viele Formeln kennengelernt, die oft nicht so leicht zu durchschauen sind und sicher den Eindruck erwecken, dass es viel Arbeit wäre, mit Zettel und Stift die entsprechenden Berechnungen durchzuführen. Das ist natürlich richtig und gleichzeitig der Grund dafür, warum man heute den Computer benutzt, um diese Berechnungen anzustellen. Statistikprogramme berechnen Standardfehler und Konfidenzintervalle sozusagen nebenbei (oft muss nur ein Häkchen im entsprechenden Optionsfeld eines Mittelwertsvergleiches oder einer Regression gesetzt werden). Das ist gut so, denn es spart viel Arbeit und hilft, Fehler zu vermeiden. Kaum jemand wird ein Konfidenzintervall per Hand ausrechnen. Allerdings sollten Sie die grundlegenden Ideen, die hinter all diesen Berechnungen stehen, verstanden haben. Deswegen haben wir auch nicht darauf verzichtet, uns die jeweiligen Formeln anzusehen. Nur wenn Sie verstehen, was in einem Statistikprogramm überhaupt passiert, können Sie flexibel und vor allem ohne Fehler zu machen mit dem gelieferten Output umgehen. Und Sie sollten einschätzen können, wann eine bestimmte Berechnung überhaupt sinnvoll ist und welche Aussage sie liefert. Ein blindes An-

klicken von Analysen führt in der Regel zu Ergebnissen, die der Anwender selbst nicht versteht. Vermeiden Sie das und greifen Sie – vor allem bei sehr einfachen Formeln – auch mal zu Zettel und Stift!

3.4 Der Signifikanztest

Wir wenden uns jetzt der dritten Möglichkeit zu, mit der man die Güte der Schätzung von Unterschieden und Zusammenhängen von einer Stichprobe auf die Population beurteilen kann. Es ist die zweifellos bekannteste und am weitesten verbreitete, allerdings auch die schwierigste Methode: der *Signifikanztest*. Signifikanz bedeutet so viel wie Bedeutsamkeit. Der Signifikanztest soll – genau wie das Konfidenzintervall – eine *Entscheidungshilfe* sein, wenn es um (widerstreitende) Hypothesen geht. Wenn sich in einer Stichprobe ein Effekt zeigt, stellt sich die Frage, ob die Bedeutsamkeit dieses Effektes groß genug ist, um ihn auf die Population zu verallgemeinern und die entsprechende Hypothese anzunehmen. Im Gegensatz zu Konfidenzintervallen ist der Signifikanztest ein relativ altes Verfahren. Der erste Gebrauch wird dem Engländer John Arbuthnot in einer Veröffentlichung im Jahre 1710 zugeschrieben. Später (Anfang des 20. Jahrhunderts) wurde der Signifikanztest zu einem universal verwendbaren Verfahren, und zwar unter dem Einfluss eines genialen Statistikers namens Ronald Fisher. Fisher führte grundlegende Begriffe in die Statistik ein wie Freiheitsgrade, Randomisierung und Nullhypothese (die wir uns gleich ansehen werden). Er unterschied als Erster zwischen Stichprobenkennwerten und Populationsparametern und entwickelte die oben behandelten Formeln, mit denen man das Eine aus dem Anderen schätzen kann. Seit der Zeit Fishers ist der Signifikanztest zu einer Art Ritual in den Sozialwissenschaften geworden. Und meist wird nur danach gefragt, ob ein gefundenes Ergebnis signifikant ist oder nicht. Damit geht allerdings eine zu starke Vereinfachung in der Anwendung und Interpretation dieser Methode einher, die von Fisher sicher nicht beabsichtigt war. Doch bevor wir uns kritisch mit dem Signifikanztest auseinandersetzen, sehen wir uns zunächst an, wie er funktioniert und welche Aussagen er zulässt (und welche nicht).

Nullhypothese und Alternativhypothese

Die alles entscheidende Grundlage für den Signifikanztest sind Stichprobenverteilungen. Von denen haben wir in den letzten Kapiteln viel gehört, und hier erhalten sie noch einmal eine zusätzliche Bedeutung. Bisher sind wir bei der Konstruktion solcher Verteilungen stets von unserem gefundenen Effekt (Mittelwert, Anteil, Mittelwertsunterschied, Zusammenhang) ausgegangen und haben um diesen Effekt herum die Verteilung konstruiert (bzw. haben den Computer diese Verteilungen simulieren lassen). Bei Signifikanztests müssen wir uns von dieser sehr anschaulichen Konstruktion von Stichprobenverteilungen ein Stück weit verabschieden. Stattdessen handelt es sich hier nur noch um Verteilungen, die aus theoretischen Überlegungen erwachsen. Und das ist genau die Tatsache, die dazu führt, dass Signifikanztests eine sehr theoretische Angelegenheit sind, die man gut durchschauen muss, um ihre Logik zu verstehen. Dafür betrachten wir zunächst zwei neue Arten von Stichprobenverteilungen, die von nun an sehr zentral sein werden: die Verteilungen für die Nullhypothese und für die Alternativhypothese.

Beim Signifikanztest werden immer Hypothesen *gegeneinander getestet*. Das bedeutet, dass wir mindestens zwei Hypothesen benötigen. Die eine Hypothese kennen wir schon: das ist diejenige, die den erhofften Effekt beschreibt (also einen Mittelwertsunterschied oder einen Zusammenhang in der Population). Sie ergibt sich in aller Regel aus der Forschungshypothese, also der eigentlichen Forschungsfrage, die man beantworten möchte. Sie wird als *Alternativhypothese* oder auch *H1* bezeichnet. Aber wozu ist sie eine „Alternative"? Hier kommt die zweite Hypothese ins Spiel, gegen die die Alternativhypothese getestet wird: die sogenannte *Nullhypothese* oder *H0*. Die Nullhypothese ist die zentrale Idee des Signifikanztests. Sie behauptet nämlich, dass es in der Population gar keinen Effekt gibt.

> Die Nullhypothese geht davon aus, dass es in der Population keinen Effekt (Unterschied, Zusammenhang) gibt. Die Alternativhypothese unterstellt einen solchen Effekt in der Population.

Die Idee dahinter ist die folgende: Wenn wir davon ausgehen (hypothetisch), dass es in der Population keinen Effekt gibt, dann sollten alle Studien zu einer

bestimmten Fragestellung auch keine Effekte finden oder aber Effekte, die sich zufällig um die 0 herum verteilen. Die Wahrscheinlichkeit, sehr große Effekte zu finden, ist dann sehr klein. Ein Effekt, der groß genug ist, kann demzufolge nicht mehr als wahrscheinlich angesehen werden unter der Prämisse, dass diese Nullhypothese zutrifft. In diesem Fall würden wir die Nullhypothese als falsch ablehnen (verwerfen) und stattdessen die Alternativhypothese akzeptieren, die von einem Effekt ausgegangen ist. Dieses Vorgehen sehen wir uns an den entsprechenden Verteilungen an.

Die Tatsache, dass sich in Stichproben gefundene Ergebnisse bei Gültigkeit der Nullhypothese zufällig um den Wert 0 verteilen sollten, beschreibt natürlich nichts anderes als eine Stichprobenverteilung um den Wert 0. Das ist die Verteilung der Nullhypothese. Wir haben es also hier mit einer Stichprobenverteilung zu tun, die als Mittelwert *immer* 0 hat:

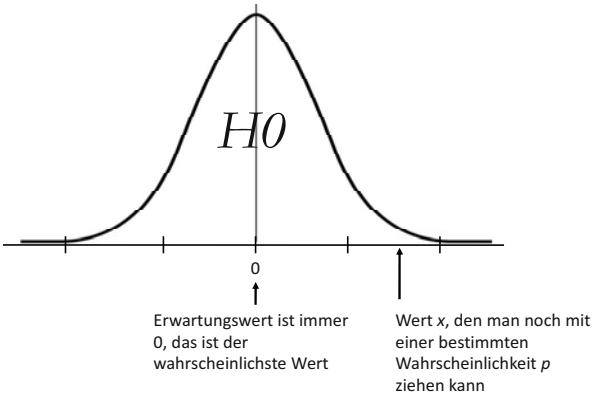

Abbildung 3.7 Stichprobenverteilung der Nullhypothese

Die Nullhypothese macht den Signifikanztest zu einer Art konservativem Verfahren, das heißt, sie behauptet immer, dass es in der Population keinen Effekt gibt und dass ein in einer Stichprobe gefundener Effekt nur auf Zufall beruht. Wie groß dieser Zufall ist, kann man aus der Stichprobenverteilung ablesen. Hätten wir in einer Stichprobe einen Effekt der Größe x gefunden, der von 0 verschieden ist, könnten wir in der Stichprobenverteilung ablesen, wie wahrscheinlich es war, diesen oder einen noch größeren Effekt zu finden, wenn die Nullhypothese gilt. In Abbildung 3.7 beträgt die Wahrscheinlichkeit p, den

Wert x oder noch größere Werte zu finden, vielleicht 0,05, also 5%. Die Nullhypothese würde nun „behaupten“, dass dieses Ergebnis zwar schon sehr unwahrscheinlich war, dass wir aber in einer nächsten Studie einen viel kleineren Effekt finden würden oder einen, der auf der anderen Seite der 0 liegt. Im Schnitt – so die Annahme der Nullhypothese – sollten wiederholte Studien zu Effekten führen, die sich zufällig um die 0 verteilen.

> Der p-Wert ist die Wahrscheinlichkeit dafür, dass in einer Stichprobe der gefundene oder ein noch größerer Effekt auftritt unter der Annahme, dass die Nullhypothese gilt.

Die Logik des Signifikanztests: Wahrscheinlichkeiten und Irrtümer

Fisher argumentierte nun folgendermaßen zur Frage, wie man eine Entscheidung bezüglich der Nullhypothese treffen kann: Bevor man einen Signifikanztest durchführt, legt man eine sogenannte *Irrtumswahrscheinlichkeit* α fest. Die Irrtumswahrscheinlichkeit entspricht einem Wert für p, ab dem man nicht mehr bereit ist, die Nullhypothese zu akzeptieren. Empirisch gefundene Werte, die in diesen *Ablehnungsbereich* fallen, werden als signifikant bezeichnet und führen zur Ablehnung der Nullhypothese. Die Irrtumswahrscheinlichkeit wird daher auch als *Signifikanzniveau* oder *Alpha-Niveau* bezeichnet. Werte, die in diesen Bereich fallen, sind so unwahrscheinlich, dass wir die Nullhypothese ablehnen und dabei natürlich das Risiko eingehen, dass wir uns mit dieser Entscheidung *irren* (daher der Name Irrtumswahrscheinlichkeit). Die Irrtumswahrscheinlichkeit (oder einfach nur Alpha) entspricht also einer Fehlerwahrscheinlichkeit. Der Fehler besteht in der fälschlichen Ablehnung der Nullhypothese. Sehen wir uns diese Überlegung noch einmal an der Stichprobenverteilung an. Abbildung 3.8 zeigt die Festlegung eines Signifikanzniveaus.

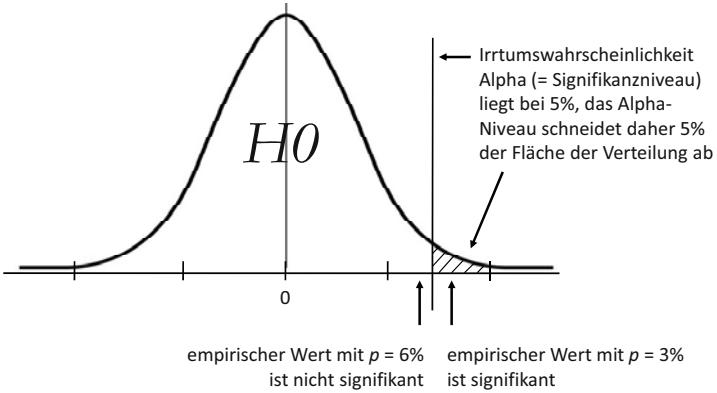

Abbildung 3.8 Die Logik des Signifikanztests

In der Abbildung haben wir uns für ein Signifikanzniveau von 5% entschieden. Das heißt, alle empirischen Ergebnisse aus Stichproben, die mit einer Wahrscheinlichkeit p von kleiner als 5% korrespondieren, veranlassen uns zur Ablehnung der Nullhypothese. Wir sprechen also immer dann von einem signifikanten Ergebnis, wenn $p < \alpha$, also p in den vorher festgelegten Ablehnungsbereich fällt. Wenn wir das tun, besteht gleichzeitig das Risiko p, dass der gefundene Wert doch zur Nullhypothese gehört hat, also nur durch Zufall zustande kam. Um das Risiko eines solchen Irrtums – wir sprechen vom sogenannten *Alpha-Fehler* – zu minimieren, wird die Wahrscheinlichkeit für Alpha meist auf 5% oder 1% festgelegt. Bei einem Alpha von 1% sind natürlich noch viel größere Effekte nötig, um ein signifikantes Ergebnis zu erhalten. Wenn wir also hören, dass ein Ergebnis auf dem 5%-Niveau signifikant geworden ist, dann wissen wir, dass die Entscheidung gegen die Nullhypothese mit einer Irrtumswahrscheinlichkeit von *maximal* 5% behaftet ist.

> Der Alpha-Fehler legt das Niveau der Irrtumswahrscheinlichkeit (Signifikanzniveau) fest. Das Ergebnis eines Signifikanztests ist signifikant, wenn $p < \alpha$.

Die entscheidende Frage ist nun, woher der p-Wert eigentlich kommt. Die Antwort darauf ist relativ einfach, da Sie alles, was Sie dazu benötigen, bereits wissen. Wir hatten oben behauptet, dass zum Beispiel Mittelwertsunterschiede

einer *t*-Verteilung folgen. Alles was wir also für den Signifikanztest tun müssen, ist, wieder die *t*-Verteilung für unsere Stichprobe heranzuziehen. Wie wir wissen, ist die *t*-Verteilung eine Stichprobenverteilung von Mittelwertsunterschieden mit einem Mittelwert von 0. Sie entspricht daher bereits der Verteilung der Nullhypothese! Wenn wir also einen *t*-Wert kennen, müssen wir aus der Verteilung nur den entsprechenden *p*-Wert ablesen. Wie man den *t*-Wert (bzw. andere Prüfgrößen) ausrechnet, werden wir später bei den einzelnen Signifikanztests sehen. Denn neben dem *t*-Wert gibt es noch andere Werte, die man ausrechnen kann – je nachdem, um welche Fragestellung es sich handelt. Während für Mittelwertsunterschiede die *t*-Verteilung gilt, sind es für Anteile die Binomialverteilung, für Häufigkeiten die sogenannte Chi-Quadrat-Verteilung usw. Das Prinzip ist aber immer dasselbe: diese Verteilungen repräsentieren die Nullhypothese und werden zum Prüfen des Signifikanztestergebnisses benutzt. Sehen wir uns diese *Prüfverteilungen* im Überblick an, bevor wir die Überlegungen zum Signifikanztest fortsetzen.

Prüfverteilungen

Da der Signifikanztest für alle Arten von Fragestellungen berechnet werden soll, brauchen wir für alle Fragestellungen eine entsprechende Verteilung für die Nullhypothese. Man spricht hier von sogenannten Prüfverteilungen, da in ihnen der *p*-Wert abgelesen wird, der zum Prüfen auf Signifikanz benötigt wird. Zwei Prüfverteilungen kennen wir schon: die *z*-Verteilung (Standardnormalverteilung) und die *t*-Verteilung. Die *z*-Verteilung kann man benutzen, um innerhalb *einer* Stichprobe für *einen einzigen Wert* zu bestimmen, ob er sich signifikant vom Durchschnitt unterscheidet. Einen solchen *z-Test* kennen Sie schon von der *z*-Standardisierung. Dort wurde für jede Person in einer Stichprobe ein *z*-Wert berechnet, der angibt, wo diese Person relativ zu allen anderen in der Verteilung aller Werte liegt. Für diesen *z*-Wert müssen wir nur in der *z*-Verteilung nachsehen, welchem *p*-Wert er entspricht. Die Tabellen für alle wichtigen Prüfverteilungen finden Sie in der Tabellensammlung. Betrachten wir ein Beispiel. In einer Klausur ist von 50 Studierenden ein Mittelwert von $M = 34$ Punkten erreicht worden. Die Standardabweichung beträgt $SD = 3{,}7$. Wir wollen nun wissen, ob Paul, der 41 Punkte hatte, signifikant besser war als

alle anderen. Dafür bestimmen wir den *z*-Wert für Pauls Ergebnis. Das ist nichts anderes als die *z*-Standardisierung seines Punktwertes:

$$z = \frac{x_{Paul} - \bar{x}}{SD} = \frac{41 - 34}{3{,}7} = 1{,}89$$

Für diesen *z*-Wert (oder genauer: für alle Werte, die kleiner oder gleich diesem *z*-Wert sind) suchen wir die entsprechende Wahrscheinlichkeit aus der Tabelle. In der Tabelle sehen wir, dass ein *z*-Wert von 1,89 eine Fläche von 97,06% abschneidet. Welchem *p*-Wert entspricht das? Dafür ziehen wir einfach diese Fläche von 100% ab: 100 – 97,06 = 2,94. Wenn wir von einem Signifikanzniveau von 5% ausgehen, dann ist $p < \alpha$, was bedeutet, dass Paul signifikant besser ist als der Durchschnitt (seine bessere Leistung ist also wirklich von Bedeutung und nicht nur zufällig entstanden). Nur 2,94% der Studierenden sind besser als Paul. Die Nullhypothese beim *z*-Test hätte behauptet, dass Paul sich nicht vom Durchschnitt unterscheidet. Diese Hypothese haben wir damit verworfen.

Die zweite bekannte Verteilung, die *t*-Verteilung, kann überall da angewendet werden, wo es um Mittelwertsunterschiede, Korrelationskoeffizienten und Regressionsgewichte geht. All diese Maße sind *t*-verteilt. Darüber hinaus gelten für andere Signifikanztests jeweils andere Prüfverteilungen. Wir werden uns hier mit den wichtigsten auseinandersetzen. Dazu gehören beispielsweise die *F*-Verteilung, die die Verteilung von Varianzen widerspiegelt oder die Chi-Quadrat-Verteilung, die die Verteilung von Häufigkeitsdaten enthält. Bei der Betrachtung der einzelnen Signifikanztests werden wir auf diese Verteilungen näher eingehen. Das Nachschauen von *p*-Werten in den Prüfverteilungen praktizieren wir hier im Wesentlichen zur Veranschaulichung der Vorgehensweise beim Signifikanztesten. Denn auch bei diesen Testverfahren wird der *p*-Wert immer von den Statistikprogrammen ausgegeben.

Einseitiges und zweiseitiges Testen

Kommen wir zurück zur Durchführung des Signifikanztests. Bei den bisherigen Überlegungen haben wir gesehen, dass ein gefundener *p*-Wert kleiner als die festgelegte Alpha-Wahrscheinlichkeit sein soll. Bei einem Alpha-Niveau von 5% haben wir 5% der Verteilung abgeschnitten und geschaut, ob der *p*-Wert in diesem abgeschnittenen Teil liegt oder nicht. Allerdings haben wir die

Fläche auf der rechten Seite der Verteilung abgeschnitten. Das ist aber nicht zwingend. Kommen wir noch einmal auf das Beispiel des Trainings zurück, bei dem wir gehofft hatten, dass die Trainingsgruppe in einem Test höhere Werte erreicht als eine Kontrollgruppe, die das Training nicht absolviert hat. Wenn wir eine solche Hypothese haben, die in eine bestimmte Richtung geht, dann sprechen wir von *einseitigem Testen*. Wir postulieren dann nämlich, dass der Effekt auf der rechten Seite der Verteilung unter Annahme der Nullhypothese zu finden ist. Denn auf der linken Seite der Nullhypothesen-Verteilung stehen *t*-Werte, die kleiner 0 sind. Das sind all die Effekte, die entstehen würden, wenn die Kontrollgruppe bessere Werte im Test erreicht hätte. Davon gehen wir aber nicht aus. Wir schließen damit von vorn herein aus, dass wir einen Alpha-Fehler auf der linken Seite machen könnten. Das führt natürlich dazu, dass wir gar nicht erst einen Signifikanztest machen würden, wenn die Kontrollgruppe die höheren Werte erzielen würde. Denn damit wäre unsere Hypothese bereits widerlegt und ein Test wäre nutzlos.

Es gibt nun eine zweite Möglichkeit von Signifikanztests – nämlich solche, bei denen man vorher keine Idee darüber hat, in welche Richtung der Effekt gehen wird, wenn es überhaupt einen gibt. In der Forschung ist es gar nicht so selten, dass man sozusagen „ins Blaue hinein" testet, ob es irgendeinen Effekt gibt. (Obwohl das nicht der Idealfall ist!) Oft gibt es auch verschiedene Hypothesen, die sich konträr gegenüber stehen. Beispielsweise könnte man sich fragen, ob Frauen als attraktiver eingeschätzt werden, wenn sie eine Brille tragen. Die Brille könnte sie erfolgreicher erscheinen lassen und damit attraktiver machen. Andererseits könnte das Fehlen einer Brille mehr Weiblichkeit ausstrahlen. Wenn man keine Hypothese über die Richtung des Effektes hat, spricht man von *zweiseitigem Testen*. Denn der Effekt (ein möglicher Mittelwertsunterschied) kann sich nun auf beiden Seiten der *t*-Verteilung ereignen. Das wiederum führt dazu, dass der Alpha-Fehler auf beide Seiten der Verteilung aufgeteilt werden muss, weil wir nicht wissen, wo genau wir suchen. Bei einem Alpha von 5% müssen wir demnach auf jeder Seite 2,5% der Verteilung abschneiden. Der Effekt hat es damit natürlich „schwerer", signifikant zu werden – egal auf welcher Seite er liegt. Denn er muss nun weiter von 0 entfernt sein, um unter das Niveau von Alpha zu fallen. Dafür haben wir aber den Vorteil, dass wir ein signifikantes Ergebnis auf beiden Seiten akzeptieren würden. Abbildung 3.9 veranschaulicht dieses Prinzip.

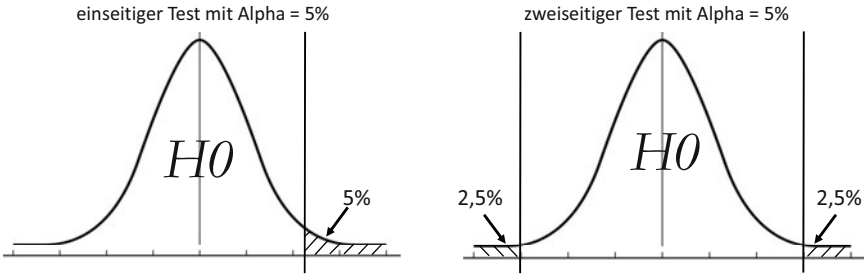

Abbildung 3.9 Einseitiges und zweiseitiges Testen

Für einige Signifikanztests muss man sich vorher überlegen, ob man einseitig oder zweiseitig testen möchte. Bei Mittelwertsunterschieden von zwei Messungen ist es immer möglich, eine Hypothese über die Richtung des Unterschiedes anzugeben. Wenn man sich für einseitiges Testen entschieden hat, hat man es leichter, ein signifikantes Ergebnis zu bekommen. In Statistikprogrammen kann man bei solchen Tests angeben, ob sie ein- oder zweiseitig testen sollen. Es gibt aber auch Signifikanztests, die immer einseitig testen. Das sind zum Beispiel Tests für Häufigkeiten. Da Häufigkeiten nie negativ sein können (im Gegensatz zu Mittelwertsunterschieden), kann man hier auch nur in eine Richtung testen. Ebenso verhält es sich mit Varianzen. Auch Varianzen können nicht negativ sein, und die entsprechenden Tests sind daher immer einseitig. (Später dazu mehr.)

Alternativhypothese und Beta-Fehler

Die oben beschriebene Vorgehensweise beim Signifikanztesten nach Fisher ist im Prinzip die einfachste und gebräuchlichste. Und sie ist diejenige Vorgehensweise, nach denen auch Statistikprogramme arbeiten. Sie geben einen *p*-Wert aus, der angibt, wie wahrscheinlich das gefundene oder ein noch extremeres Ergebnis unter der Annahme der Nullhypothese war. Die Entscheidung über die Höhe des Signifikanzniveaus und über Annahme oder Ablehnung von Hypothesen bleibt dabei dem Anwender überlassen. Bei Fisher blieb allerdings eine Frage offen. Bei einem signifikanten Ergebnis empfiehlt er, die Nullhypothese zu verwerfen. Er schlägt aber keine Handlungsoption vor für den Fall, dass das Er-

gebnis nicht signifikant ist. Ist dann die Nullhypothese „richtig"? Darüber können wir keine Aussage machen. Um dieses Problem zu lösen, haben Jerzy Neyman und Egon Pearson die Alternativhypothese ins Spiel gebracht. Die Alternativhypothese erwächst in aller Regel aus der Forschungshypothese. Sie steht dann der Nullhypothese gegenüber und behauptet (oder „hofft"), dass es einen Effekt gibt. Sehen wir uns an, welche Rolle die Alternativhypothese beim Signifikanztest spielt.

Nach Neyman und Pearson sollte man beim Testen von Hypothesen nicht primär nur von der Nullhypothese ausgehen, sondern sich vielmehr fragen, welchen Effekt man eigentlich untersuchen möchte. Das heißt, man sollte sich vor dem Test überlegen, wie groß der Effekt, der einen interessiert, (mindestens) sein sollte. Um diesen Effekt zu charakterisieren, benötigen wir die Alternativhypothese. Sie muss sich um die Größe des erhofften Effektes von der Nullhypothese unterscheiden. Abbildung 3.10 zeigt die beiden Hypothesen und den Effekt.

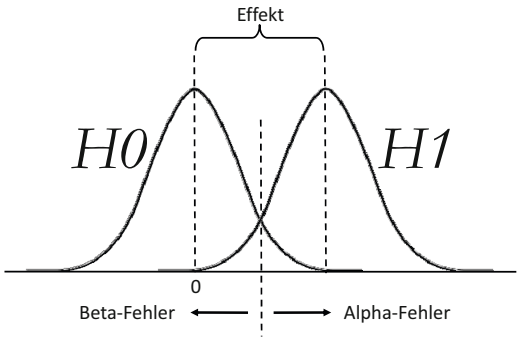

Abbildung 3.10 Nullhypothese, Alternativhypothese und Effekt

Die Nullhypothese erstreckt sich wieder um den Wert 0, die Alternativhypothese hingegen wurde um einen Wert herum konstruiert, der den erhofften Effekt in der Population widerspiegelt. Da dieser Effekt ein theoretischer, erhoffter Wert ist, handelt es sich damit auch bei der Alternativhypothese wieder um eine theoretische Verteilung (die wir wieder simuliert haben). Wie groß soll der Effekt nun sein, um den herum wir diese Verteilung konstruieren sollen? Darauf gibt es zwei Antworten. Zum einen kann man sich fragen, welcher Effekt mindestens vorhan-

den sein soll, damit er für uns von Interesse ist. Sehr kleine Effekte haben in der Regel keine große Bedeutung. Wie groß genau ein solcher Mindesteffekt sein soll, hängt natürlich von der Fragestellung ab. Ein Medikament, welches Krebspatienten auch nur minimal helfen kann, ist schon bei sehr kleinen Effekten sehr interessant. Die Veränderung von Arbeitsplatzbedingungen in einem Unternehmen zur Steigerung der Mitarbeiterzufriedenheit sollte hingegen einen deutlichen Effekt ausmachen, da eine solche Umstellung mit Zeit und Geld verbunden ist. Die Größe des interessanten Mindesteffektes ist daher eine Abwägungsfrage. Die andere Möglichkeit, diesen Effekt zu bestimmen, besteht darin, dass man sich entsprechende Effekte aus bereits durchgeführten Studien anschaut. In der Forschung ist es praktisch immer der Fall, dass es schon Untersuchungen zu einem Thema gegeben hat. Deren Effekte kann man als erwarteten oder als Mindesteffekt benutzen.

Nun fragen Sie sich sicher, wofür man den Effekt und die Alternativhypothese überhaupt braucht. Die Antwort auf diese Frage verbirgt sich hinter einem interessanten Aspekt in Abbildung 3.10: die beiden Verteilungen überschneiden sich. Überlegen wir, was das bedeutet. Wir haben den Überschneidungsbereich in zwei Hälften geteilt. Beide Seiten des Überschneidungsbereiches enthalten Werte, die zu beiden Hypothesen gehören können. Das bedeutet, dass man bei der Entscheidung für oder gegen eine Hypothese verschiedene Fehler machen kann. Nehmen wir zunächst an, die Nullhypothese sei tatsächlich richtig (wir können das nie wissen, aber wir können es der Überlegung halber einmal unterstellen). Dann würden wir, wenn wir in unserer Studie einen Wert im rechten Überschneidungsbereich gefunden haben und die Nullhypothese beibehalten, keinen Fehler begehen. Wenn wir uns aufgrund des gefundenen Wertes allerdings für die Alternativhypothese entscheiden, hätten wir einen Fehler gemacht. Dieser Fehler wird auch *Fehler erster Art* genannt. Er führt dazu, dass wir die Nullhypothese fälschlicherweise verwerfen. Und er ist natürlich identisch mit dem Fehler, den Sie schon kennen: dem Alpha-Fehler.

Gehen wir nun in einer weiteren Überlegung davon aus, dass die Alternativhypothese zutrifft und es tatsächlich einen Effekt in der Population gibt. Dann wird die linke Seite des Überschneidungsbereiches interessant. Wenn wir einen Wert finden, der in diesem Bereich liegt, und dennoch die Nullhypothese ablehnen, haben wir keinen Fehler gemacht. Wenn wir allerdings die Nullhypothese beibehalten, begehen wir einen Fehler. Dieser wird *Fehler zweiter Art* oder *Beta-Fehler* genannt. Er äußert sich darin, dass wir die Alternativhypothese

fälschlicherweise ablehnen. Die vier möglichen Entscheidungen und die möglichen Fehler sind in Tabelle 3.1 zusammengefasst.

Tabelle 3.1 Fehler beim Signifikanztesten

		Entscheidung aufgrund der Stichprobe	
		Entscheidung für die H_0	***Entscheidung für die H_1***
Verhältnisse in der Population (unbekannt)	***In der Population gilt die H_0***	Korrekte Entscheidung	**α-Fehler („Fehler erster Art")**
	In der Population gilt die H_1	**β-Fehler („Fehler zweiter Art")**	Korrekte Entscheidung

Neyman und Pearson haben nun argumentiert, dass man vor einem Signifikanztest *beide* Fehler abwägen sollte, um hinterher eine begründete Entscheidung für oder gegen eine Hypothese treffen zu können. Es genügt dann also nicht mehr, nur den Alpha-Fehler festzulegen, sondern es sollte immer eine Abwägung der beiden Fehler geben, die sich an inhaltlichen Gesichtspunkten orientiert. Denn beide Fehler haben eine andere inhaltliche Bedeutung und ihre Wichtigkeit hängt von der jeweiligen Fragestellung ab. Oben haben wir das Beispiel des Krebsmedikamentes betrachtet. Wenn die Entscheidung ansteht, ob das Medikament eingesetzt werden soll oder nicht – auf welchen Fehler sollte man dann besonders achten? Hier ist natürlich der Beta-Fehler der relevante. Denn der würde bedeuten, dass man fälschlicherweise die Nullhypothese beibehält und das Medikament – obwohl es wirkt – nicht einsetzt. Diesen Fehler sollte man versuchen zu minimieren, vor allem dann, wenn mit der Neueinführung des Medikamentes keine neuen Nebenwirkungen oder erhebliche Preissteigerungen verbunden sind. Im anderen Fall hatten wir überlegt, ob ein Unternehmen die Arbeitsplätze seiner Mitarbeiter neu gestalten sollte. Eine solche Umstellung würde viel Geld und Zeit kosten und soll daher nur eingesetzt werden, wenn sie auch tatsächlich einen Effekt auf die Arbeitszufriedenheit hat. Hier wollen wir daher den Alpha-Fehler minimieren und

nicht fälschlicherweise die Nullhypothese verwerfen, was mit viel Aufwand verbunden wäre, der aber gar keinen Effekt hätte.

Die Abwägung von Alpha- und Beta-Fehler hängt also immer von der jeweiligen Fragestellung ab. Sehen wir uns zunächst an, was bei den beiden Verteilungen aus Abbildung 3.10 passiert, wenn wir das übliche Alpha-Niveau von 5% anwenden. Dieses würde 5% der Fläche der Nullhypothese abschneiden:

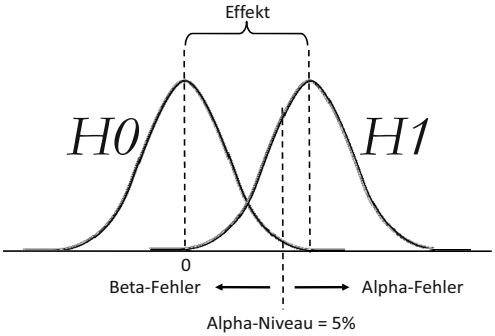

Abbildung 3.11 Abwägung von Alpha- und Beta-Fehler

In diesem Fall würden wir uns erst bei empirischen Effekten, die rechts vom Alpha-Niveau liegen, für die Ablehnung der H0 entscheiden. Das bedeutet, dass alle Werte, die links davon liegen und zur H1 gehören, zur Ablehnung von H1 führen würden. Der Beta-Fehler wäre damit relativ groß. Er entspricht der Fläche der H1-Verteilung, die links von Alpha liegt – in unserem Beispiel also etwa 3%.

Bei der Fragestellung, ob das Krebsmedikament eingeführt werden soll, sollten wir hingegen den Beta-Fehler vermindern und das Alpha-Niveau weiter nach links verschieben. Wir entscheiden uns für einen Beta-Fehler von 1%. Wir würden also das linke 1% der H1-Verteilung abschneiden. Der entsprechende Wert läge dann nur knapp über dem Mittelwert der Nullhypothese. Das würde schließlich einem Alpha-Fehler von fast 50% entsprechen. Aber wir haben argumentiert, dass bei einer solchen Entscheidung ein Alpha-Fehler nicht so dramatisch ist. Denn schlimmstenfalls entscheiden wir uns fälschlicherweise für ein Medikament, das auch nicht besser wirkt als das alte und sonst keine Nachteile hat.

Die Abwägung von Alpha und Beta ist vor allem dann relevant, wenn sich die Verteilungen stark überschneiden. Wenn der erhoffte Effekt in der Population sehr groß ist, dann liegen die beiden Verteilungen weiter voneinander entfernt und ihr Überschneidungsbereich wird kleiner. Dann sind beide Fehler klein. Eine zweite Möglichkeit, wie sich beide Fehler minimieren lassen, ist das altbekannte Prinzip, größere Stichproben zu verwenden. Denn diese führen zu einer geringeren Streuung der Stichprobenverteilungen für H0 und H1 (kleinerer Standardfehler). Damit werden die Verteilungen schmaler und überlappen sich weniger. Beide Effekte sind in Abbildung 3.12 zusammengefasst.

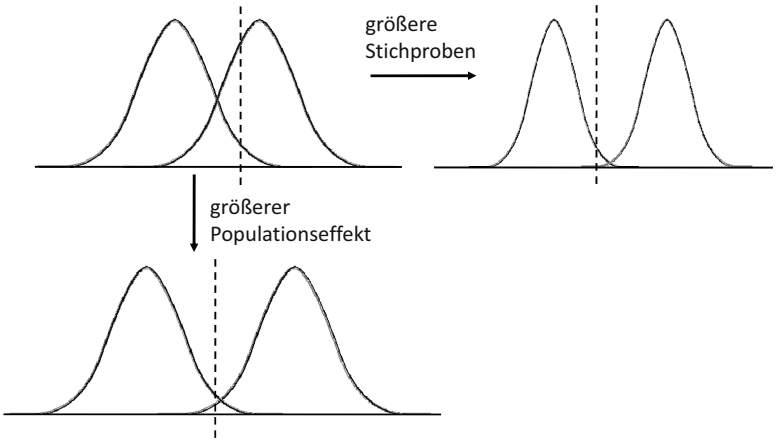

Abbildung 3.12 Der Einfluss von Stichprobengröße und Populationseffekt auf die Fehler beim Signifikanztest (Alpha = 5%)

Fassen wir zusammen: Wenn wir die Alternativhypothese und die Abwägung der beiden Fehler mit in die Durchführung des Signifikanztests einbeziehen, dann sieht das generelle Vorgehen folgendermaßen aus: (1) Formuliere eine Nullhypothese und konstruiere die entsprechende Stichprobenverteilung. (2) Formuliere eine Alternativhypothese und konstruiere die entsprechende Stichprobenverteilung. (3) Mache eine Abwägung der Wichtigkeit der Fehler erster und zweiter Art. (4) Prüfe, ob der *p*-Wert größer oder kleiner als der Fehler erster Art ist. (5) Ist der *p*-Wert kleiner Alpha, verhalte dich so, als ob die

Alternativhypothese stimmt. Ist der *p*-Wert größer als Alpha, verhalte dich so, als ob die Nullhypothese stimmt.

Das sind die Schritte, die nach dem Vorgehen von Neyman und Pearson zu berücksichtigen sind. Wie wir aber oben schon angemerkt hatten, wird von Statistikprogrammen nur der *p*-Wert ausgegeben. Die Alternativhypothese wird dabei nicht beachtet. Was wir aus dieser vorgeschlagenen Vorgehensweise aber lernen sollten ist, dass man eine Abwägung von Alpha und Beta macht und nicht einfach „blind" auf ein festgesetztes Alpha-Niveau vertraut. Besonders bei Fragestellungen, bei denen es auf einen kleinen Beta-Fehler ankommt, sollte man diesen auf 1% oder 5% festlegen und den korrespondierenden Alpha-Fehler als Signifikanzniveau benutzen. Die Berechnung von Alpha bei gegebenem Beta erfolgt beispielsweise mit dem Programm von Sedlmeier und Köhlers (2003).

Da beim Signifikanztesten aber (leider!) oft die Schritte 2 und 3 übersprungen werden, spricht man manchmal von einer *Hybrid*-Vorgehensweise, weil sich die beiden Ansätze von Fisher und Neyman/Pearson dabei vermischen. Immerhin erhalten wir bei diesem Vorgehen nun auch eine Handlungsoption für den Fall, dass das Ergebnis nicht signifikant ist. Wir „verhalten uns dann so", als ob die Nullhypothese stimmt.

Wir haben damit die grundlegende Idee des Signifikanztests und die prinzipielle Vorgehensweise für die Bestimmung des *p*-Wertes vorgestellt. Für jede Art von Fragestellungen gibt es spezielle Signifikanztests, denen wir uns später genauer zuwenden werden. Sie unterscheiden sich im Prinzip nur durch die Prüfverteilung, die sie verwenden, um den *p*-Wert auf Signifikanz zu prüfen.

Angemerkt sei auch hier nochmal, dass die Simulation von Verteilungen (wie für die H0) nicht von Hand erledigt werden muss, sondern im Computer im Hintergrund abläuft. Der Anwender bekommt in der Regel nur den *p*-Wert zu Gesicht. Und es gibt einen weiteren Aspekt, der manchmal zu Verwirrungen führt: Der Signifikanztest lässt keine (!) Aussagen über die Wahrscheinlichkeit von Hypothesen zu. Wir erfahren also nicht, wie wahrscheinlich es ist, dass eine Hypothese zutrifft oder nicht. Wir erfahren stets nur etwas über die Wahrscheinlichkeit des gefundenen Effektes, falls eine Hypothese (in der Regel die H0) zutrifft. Das macht nochmals deutlich, dass der Signifikanztest ein sehr theoretisches Verfahren ist, das zwar schnelle und einfache Entscheidungen herbeiführen kann, das aber nicht ohne weiteres zu verstehen und korrekt zu interpretieren ist. Da oft sehr unkritisch mit dem Signifikanztest umgegangen

wird, sehen wir uns noch an, welche Größen das Ergebnis eigentlich beeinflussen und welche „Manipulationsmöglichkeiten“ damit auch bestehen.

Einflussgrößen auf das Ergebnis des Signifikanztests

Es gibt drei Größen, die das Ergebnis eines Signifikanztests beeinflussen: die Größe des Populationseffektes, die Stichprobengröße und das Alpha-Niveau. Auf welche Art diese Beeinflussung geschieht, haben wir im Prinzip schon gesehen (siehe Abb. 3.12). Wir können sie hier noch einmal zusammenfassen. Beginnen wir mit der Größe des Populationseffektes. Die Größe dieses Effektes kennen wir natürlich nicht – sonst müssten wir keinen Signifikanztest machen – aber wir können überlegen, wie er sich auf das Signifikanztestergebnis auswirkt. Je größer er ist, desto eher werden wir in unserer Stichprobe Werte finden, die sehr weit von der Nullhypothese entfernt sind. Sie werden also zu kleinen *p*-Werten führen. An der Größe des Populationseffektes können wir jedoch nichts ändern.

Anders sieht das bei den anderen beiden Größen aus. Die Stichprobengröße hat, wie wir in Abbildung 3.12 gesehen haben, einen Einfluss auf die Breite der Stichprobenverteilungen. Verwenden wir also größere Stichproben in unserer Studie, erhöht das immer die Wahrscheinlichkeit eines signifikanten Ergebnisses. Gelegentlich hat dieser Effekt zu einer Kritik am Signifikanztest geführt, die bemängelt, dass man mit genügend großen Stichproben jeden noch so kleinen Effekt signifikant „machen“ kann. Diese Kritik trifft allerdings nicht den Signifikanztest selbst, denn wenn sich bei sehr großen Stichproben immer noch ein Effekt zeigt, dann ist der natürlich statistisch umso bedeutsamer. Die Kritik richtet sich vielmehr gegen den sorglosen Umgang mit Signifikanztestergebnissen, nämlich dann, wenn Signifikanz mit inhaltlicher Wichtigkeit verwechselt wird. Wir kommen gleich darauf zurück. Festzuhalten bleibt, dass größere Stichproben eher zu signifikanten Ergebnissen führen können.

Die dritte Einflussgröße ist trivial. Das festgelegte Alpha-Niveau entscheidet darüber, ob ein Ergebnis signifikant ist oder nicht. Begnügen wir uns mit einem Alpha von 20%, bekommen wir eher ein signifikantes Ergebnis als bei einem strengeren Alpha von 1%. Wie gesagt, sind in der Psychologie die Vorstellungen von einem geeigneten Alpha leider relativ festgefahren und liegen meist bei 5 bzw. 1%.

Kritische Betrachtung des Signifikanztests

Aus den zuletzt genannten Punkten können wir direkt einige Kritikpunkte am Signifikanztest ableiten. Wie eben erwähnt, trifft das Argument, dass man mit sehr großen Stichproben auch sehr kleine Effekte signifikant machen kann, eher den Umgang mit der Bedeutung von Signifikanztestergebnissen. Ein *signifikanter* Effekt sollte eben nicht mit einem *wichtigen* Effekt verwechselt werden. Ob ein Effekt nämlich auch *inhaltlich von Interesse* ist, hängt von seiner Größe und der Fragestellung ab. Was aber erfahren wir über die Größe des Effektes, wenn wir einen Signifikanztest gemacht haben? Die Antwort ist: gar nichts. Der *p*-Wert allein ist kein Indikator für die Größe des Effektes, denn wie wir eben gezeigt haben, hängt er von noch anderen Einflussgrößen ab. Wir können daher aus einem Signifikanztestergebnis nicht die Größe des wahren Effektes in der Population schätzen. Damit erfahren wir auch nichts über die inhaltliche Bedeutsamkeit des Signifikanztestergebnisses. Das zeigt wieder, wie sehr theoretisch und wie wenig pragmatisch der Signifikanztest im Grunde ist.

Um solchen Schwierigkeiten bei der Interpretation und dem ritualisierten Umgang mit Signifikanztestergebnissen entgegenzuwirken, vollzieht sich seit einigen Jahren eine Trendwende in der Art und Weise, wie Forschungsergebnisse publiziert werden. Neben den Signifikanztestergebnissen sollen Angaben gemacht werden, die besser interpretierbar sind und auch eine Aussage über die Größe von Effekten zulassen. Eine solche Alternative kennen Sie schon: das Konfidenzintervall. Konfidenzintervalle geben konkrete Wertebereiche in Rohwerten an und sind damit viel aussagekräftiger und besser zu verstehen als ein abstrakter *p*-Wert. Die zweite Alternative zu den Signifikanztestergebnissen sind Schätzungen für den tatsächlichen Effekt in der Population, denen wir uns im nächsten Kapitel widmen werden: die sogenannten *Effektgrößen*. Diese werden heute zunehmend als Ergänzung zu Signifikanztestergebnissen mit angegeben.

Die H0 als Forschungshypothese

Vorher sehen wir uns aber noch kurz einen Spezialfall des Hypothesentestens an. Bei einigen Fragestellungen kann es vorkommen, dass man hofft, dass es *keinen* Effekt – beispielsweise *keinen* Mittelwertsunterschied – gibt. Ein typisches

Beispiel sind Studien, die von Tabakunternehmen durchgeführt werden und zeigen sollen, dass Rauchen nicht schädlich ist. Dabei kann man alle möglichen Fragen untersuchen, aber die Hoffnung ist immer, dass sich Raucher und Nichtraucher eben nicht unterscheiden. Das bedeutet, dass die Forschungshypothese mit der Nullhypothese korrespondiert und dass man ein nicht signifikantes Ergebnis dann als Bestätigung seiner Forschungshypothese wertet. Alle oben genannten Interpretationen gelten dann natürlich umgekehrt. Je strenger man hier das Alpha-Niveau wählt, desto wahrscheinlicher wird man seine Hypothese beibehalten können. Und hier könnte die Verwendung einer *kleinen* Stichprobe dazu führen, dass man ein nicht-signifikantes Ergebnis erhält. Ein nicht-signifikantes Ergebnis anzuvisieren, ist also immer „leichter".

Bei solchen Fragestellungen ist deshalb darauf zu achten, dass man dem Effekt – wenn er denn in der Population da ist – auch eine Chance gibt, sich zu zeigen. Ob Rauchen zu erhöhtem Blutdruck führt, sollte daher in einer „fairen" Studie so untersucht werden, dass ein erhöhter Blutdruck auch sichtbar werden kann. Das heißt, man sollte hier als Alternativhypothese einen Effekt bestimmen, ab dem man sagen würde, dass Rauchen *nicht* mehr unbedenklich ist. Mit Hilfe dieser Alternativhypothese kann man auch wieder eine Abschätzung der Fehler erster und zweiter Art machen und hier den Beta-Fehler nicht zu groß wählen.

Hinzu kommt noch, dass das Nichtvorhandensein eines signifikanten Ergebnisses auch auf schlecht designte oder schlecht kontrollierte Studien zurückzuführen sein kann. Auch solche Fehlerquellen sollte man ausschließen, um bei nicht-signifikanten Ergebnissen tatsächlich argumentieren zu können, dass die Nullhypothese zutrifft.

Literaturempfehlung:

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 12)

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research – online, 1*. [Internet: http://www.mpr-online.de/].

4 Effektgrößen

4.1 Der Sinn von Effektgrößen

Rufen wir uns noch einmal kurz die Idee der Inferenzstatistik ins Gedächtnis: Wir wollen Aussagen darüber machen, wie sehr wir der Schätzung eines Populationseffektes aufgrund eines Stichprobeneffektes trauen können. Dafür haben wir drei Möglichkeiten kennengelernt. Der Standardfehler gibt an, mit welchem durchschnittlichen Fehler bei einer solchen Schätzung zu rechnen ist. Konfidenzintervalle geben einen Bereich von Werten auf der abhängigen Variable an, der den wahren Wert in der Population mit einer bestimmten Wahrscheinlichkeit enthält. Und Signifikanztests fragen nach der Wahrscheinlichkeit, mit der ein Effekt auftreten konnte, wenn die Nullhypothese zutrifft. Konfidenzintervalle und Signifikanztests liefern dabei einfache und schnelle Entscheidungshilfen: Konfidenzintervalle deshalb, weil man dort meist nur schaut, ob sie den Wert 0 beinhalten oder nicht, und Signifikanztests, weil dort lediglich geprüft wird, ob der *p*-Wert kleiner oder größer als Alpha ist.

Die drei genannten Verfahren sind Möglichkeiten, um die Güte der Schätzung zu beurteilen. Allerdings lassen sie alle drei eine wichtige Frage außen vor, die zwar nicht unmittelbar eine inferenzstatistische Frage ist, die aber eigentlich die größte inhaltliche Bedeutung hat: die Frage nach der tatsächlichen Größe des Effektes. Was haben wir bisher gemeint, wenn wir von Effekten gesprochen haben? Wir haben damit Mittelwertsunterschiede und Zusammenhänge beschrieben. In einem Beispiel haben wir etwa den Unterschied in der Reaktionszeit zweier Gruppen verglichen, von denen eine ein Reaktionszeittraining durchlaufen hatte. Der Unterschied betrug 300 ms. Das ist unser Effekt. Anschließend haben wir danach gefragt, wie gut man diesen Effekt auf die Population ver-

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_5,

allgemeinern kann und haben dazu die drei Möglichkeiten der Inferenzstatistik herangezogen. Bei einem signifikanten Ergebnis (oder einem kleinen Standardfehler oder einem Konfidenzintervall, das nicht die 0 überdeckt) würden wir demnach schlussfolgern, dass dieser Effekt von 300 ms nicht zufällig entstanden ist, sondern dass auch in der Population ein Effekt vorliegt.

Für einen Anwender, der lediglich nach dem praktischen Nutzen fragt, ist dieses Ergebnis im Prinzip das, was er wissen wollte. Vorausgesetzt dass er eine Vorstellung darüber hat, ob 300 msec ein *großer* oder ein *kleiner* Vorteil sind. In der Forschung verschärft sich diese Frage nach der Größe des Effektes jedoch erheblich. Denn als Forscher würden und müssten wir uns die Frage stellen, wie groß dieser gefundene Effekt *im Vergleich* ausfällt. Damit ist gemeint, dass wir schlichtweg nicht wissen, ob 300 msec tatsächlich ein bedeutsames Ergebnis sind oder nicht. Das knüpft an die oben genannte Kritik des Signifikanztests an. Denn aus einem Signifikanztestergebnis erfahren wir nichts über die Größe des Effektes, da dieses Ergebnis vor allem von der Stichprobengröße beeinflusst wird (die wir beliebig verändern können). Und besonders dann, wenn ein Forscher seine Studienergebnisse publizieren möchte, sollte er angeben, wie groß sein gefundener Effekt ist. Denn er kann in aller Regel nicht davon ausgehen, dass alle anderen so gute Kenntnisse im jeweiligen Fachbereich haben, dass sie aufgrund eines Mittelwertsunterschiedes einschätzen könnten, ob dieser relativ groß oder relativ klein ist. Vielmehr müsste ein Maß angegeben werden, mit dem man die Größe eines Effektes unabhängig von einer bestimmten Studie, einer bestimmten Stichprobe und einem bestimmten Themenbereich beurteilen kann. Ein solches Maß bietet die *Effektgröße* oder *Effektstärke*.

Bei der Bestimmung der Effektgröße wird dabei ein sehr einfaches Prinzip angewandt, dem wir schon oft begegnet sind. Es besteht darin, dass der in einer Stichprobe gefundene Effekt *standardisiert* wird. Das bedeutet, dass man ihn durch seine Streuung teilt. Da in der Streuung natürlich die Größe der Stichprobe N steckt, wird der Effekt dadurch *unabhängig von der Stichprobengröße* ausgedrückt. Die Standardisierung führt außerdem dazu, dass eine solche Effektgröße unabhängig von der jeweiligen Studie (bzw. der in verschiedenen Studien verwendeten Maßeinheiten oder Skalen) interpretiert werden kann. Das heißt, dass Effektgrößen über verschiedene Themenbereiche und Untersuchungen hinweg vergleichbar sind – genauso wie z-standardisierte Werte.

Effektgrößen sind standardisierte Effekte, die die Stichprobengröße berücksichtigen. Sie sind über Stichproben und Themenbereiche hinweg vergleichbar.

Bevor wir uns die Berechnung von Effektgrößen ansehen, betrachten wir eine Art Ausnahme, die uns beim Verständnis dafür helfen soll, was mit Effektgrößen genau gemeint ist. Sie erinnern sich an die Verteilung des Intelligenzquotienten. Diese folgt einer Normalverteilung. Und da man sie in der Psychologie so oft benötigt, wurde für sie eine eigene Standardisierung vorgenommen: ihr Mittelwert wurde auf 100 Punkte festgelegt und ihre Standardabweichung auf 15 Punkte. Wenn es also um Intelligenz geht, haben wir es mit dem Spezialfall zu tun, dass wir im Prinzip auf die Bestimmung einer Effektgröße verzichten können. Denn da die IQ-Verteilung auf die eben genannten Parameter standardisiert wurde, weiß jeder, was ein IQ von 130 oder eine IQ-Differenz von 20 bedeutet. Ein IQ von 130 liegt beispielsweise zwei Standardabweichungseinheiten über dem Durchschnitt – also schon relativ weit am Rand der Verteilung. Wer auch immer eine Studie zum Thema Intelligenz macht, wird seine Effekte in IQ-Einheiten ausdrücken, und damit weiß jeder mit der Größe dieser Effekte etwas anzufangen. Das gleiche Prinzip verbirgt sich nun hinter der Idee der Effektgrößen. Für ihre Berechnung gibt es drei Möglichkeiten: aus den Rohdaten, aus anderen Effektgrößen und aus Signifikanztestergebnissen.

4.2 Effektgrößen aus Rohdaten

Der anschaulichste Weg zur Bestimmung von Effektgrößen ist deren Berechnung aus den Rohwerten der Studie. Diese liegen je nach untersuchter Fragestellung in der Form eines Mittelwertsunterschiedes aus abhängigen oder unabhängigen Messungen oder eines Zusammenhangsmaßes vor.

Effektgrößen für Unterschiede bei unabhängigen Messungen

Für die Berechnung der Effektgröße bei unabhängigen Messungen teilen wir den gefundenen Mittelwertsunterschied durch die gemeinsame Streuung der beiden Stichproben. Die resultierende Effektgröße ist das *Abstandsmaß d*:

$$d = \frac{\bar{x}_A - \bar{x}_B}{s_{AB}}$$

Wobei sich die gemeinsame Streuung bei gleichen Gruppengrößen aus den Streuungen der einzelnen Gruppen bestimmt:

$$s_{AB} = \sqrt{\frac{s_A^2 + s_B^2}{2}}$$

Effektgrößen für Unterschiede werden Abstandsmaße genannt, weil sie den Abstand der beiden Mittelwerte repräsentieren. Der „Erfinder" der Effektgröße *d* ist Cohen, weshalb sie auch als *Cohen's d* bezeichnet wird. Durch die Standardisierung ist der Mittelwertsunterschied nun an der Streuung der beiden Messungen relativiert und damit indirekt auch an der Stichprobengröße. Die Formel macht deutlich, dass sich die Effektgröße erhöht, wenn die Streuungen der Messungen kleiner werden. Wie Sie wissen, ist ein Mittelwert, der mit einer kleineren Streuung behaftet ist, wesentlich aussagekräftiger. Zwei Mittelwerte mit jeweils kleiner Streuung führen damit auch zu einem bedeutsameren Effekt, was sich in der höheren Effektgröße ausdrückt.

Die Effektgröße *d* drückt einen Mittelwertsunterschied durch die Standardisierung folglich in Standardabweichungseinheiten aus. Ein d von 1 oder -1 entspricht also einer Standardabweichungseinheit und kann auch entsprechend interpretiert werden. Sehen wir uns zur Berechnung ein Beispiel an. Oben hatten wir überlegt, ob ein Unternehmen eine Umgestaltung der Arbeitsplätze seiner Mitarbeiter vornehmen soll, um die Arbeitszufriedenheit zu erhöhen. Dafür wurde an 10 zufällig ausgewählten Arbeitsplätzen die Umstellung probehalber eingeführt. Die 10 Mitarbeiter an diesen Arbeitsplätzen sollten eine Woche lang jeden Tag ihre Arbeitszufriedenheit auf einer Skala von 1 bis 10 Punkten angeben. Diese Angaben werden mit denen einer Kontrollgruppe von 10 anderen Arbeitern verglichen, die noch am alten Arbeitsplatz beschäftigt waren. Die möglichen Ergebnisse sind in Tabelle 4.1 dargestellt.

Tabelle 4.1 Fiktive Ergebnisse einer Studie zur Arbeitszufriedenheit

	Versuchsgruppe	Kontrollgruppe
M	8,8	7,6
SD	1,1	1,5

Wie wir sehen, geht der Unterschied in die richtige Richtung. Berechnen wir die Effektgröße *d*. Die gemeinsame Streuung beträgt:

$$s_{AB} = \sqrt{\frac{s_A^2 + s_B^2}{2}} = \sqrt{\frac{1{,}1^2 + 1{,}5^2}{2}} = 1{,}32$$

Die Effektgröße beträgt damit:

$$d = \frac{\bar{x}_A - \bar{x}_B}{s_{AB}} = \frac{8{,}8 - 7{,}6}{1{,}32} = 0{,}91$$

Der Unterschied beträgt demnach 0,9 Standardabweichungseinheiten, was ein sehr großer Effekt ist (wir kommen später noch zur Interpretation von Effektgrößen).

Eine Alternative zu *d* wurde von Hedges vorgeschlagen, der argumentiert hat, dass die Streuung s_{AB} keine exakte Schätzung der Populationsstreuung liefert. Stattdessen sollte man nicht die Stichprobenstreuung s_{AB} verwenden, sondern die Populationsstreuung $\hat{\sigma}_{AB}$. Die Besonderheit deren Berechnung kennen Sie schon: hier wird nicht durch *n*, sondern durch $n - 1$ geteilt. Die gemeinsame Populationsstreuung berechnet sich dann wie folgt:

$$\hat{\sigma}_{AB} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2}}$$

$$wobei\ \hat{\sigma}_A^2 = \frac{n}{n-1} s_A^2\ und\ \hat{\sigma}_B^2 = \frac{n}{n-1} s_B^2$$

Die Formel für die Effektgröße – sie wird *g* oder *Hedges' g* genannt – sieht dann so aus:

$$g = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{AB}}$$

Hedges' g ist damit stets etwas kleiner als *d*, liefert aber etwas exaktere Schätzungen. Für welche der beiden Effektgrößen man sich entscheidet, bleibt dem Anwender überlassen. Allerdings ist *d* die weitaus gebräuchlichere Variante.

Effektgrößen für Unterschiede bei abhängigen Messungen

Bei abhängigen Messungen vereinfacht sich die Berechnung der Effektgrößen etwas, da hier nur noch die Differenzen der Messwerte zwischen der ersten und der zweiten Messung und deren Streuung eingehen:

$$d = \frac{\bar{x}_{Differenzwerte}}{s_{Differenzwerte}}$$

Die Streuung der Differenzwerte beträgt:

$$s_{Differenzwerte} = \sqrt{\frac{\sum(x_{diff_i} - \bar{x}_{diff})^2}{n}}$$

Alternativ kann auch hier die Effektgröße *g* berechnet werden, indem man die geschätzte Populationsstreuung der Differenzwerte benutzt:

$$g = \frac{\bar{x}_{Differenzwerte}}{\hat{\sigma}_{Differenzwerte}}$$

Die Populationsstreuung beträgt:

$$\hat{\sigma}_{Differenzwerte} = \sqrt{\frac{\sum(x_{diff_i} - \bar{x}_{diff})^2}{n-1}}$$

Effektgrößen für Zusammenhänge

Zusammenhänge von Variablen betrachten wir in aller Regel in der Form von Korrelationen. Für die Angabe einer Effektgröße kommt uns dabei ein sehr glücklicher Umstand zu gute. Erinnern Sie sich an die Formel für die Berechnung des Korrelationskoeffizienten:

$$r_{xy} = \frac{cov}{s_x s_y}$$

Die Kovarianz der beiden Variablen (ihr gleichsinniges Variieren) wird hier bereits durch die gemeinsame Streuung der Variablen geteilt. Das bedeutet, dass der Korrelationskoeffizient bereits eine Effektgröße ist. Da er die Enge des Zusammenhanges relativ zu einer perfekten Geraden beurteilt und bei 1 bzw. -1 seine Grenze hat, wissen wir auch, was er inhaltlich bedeutet. Neben den Abstandsmaßen *d* und *g* stellt das Zusammenhangsmaß *r* damit die dritte der häufig anzutreffenden Effektgrößen dar. Es gibt noch wenige andere Effektgrößen, die für spezielle Fragestellungen relevant sind. Auf diese werden wir bei der Betrachtung dieser Fragestellungen dann jeweils eingehen.

4.3 Effektgrößen aus anderen Effektgrößen

Manchmal möchte man Effektgrößen aus bereits vorhandenen Effektgrößen berechnen, zum Beispiel, wenn in verschiedenen Studien verschiedene Effektgrößen berechnet wurden, die man miteinander vergleichen möchte. Da Unterschieds- und Zusammenhangsfragestellungen prinzipiell ineinander überführbar sind, kann man auch Abstands- und Zusammenhangsmaße ineinander überführen.

Abstandsmaße können folgendermaßen ineinander überführt werden:

$$d = g\sqrt{\frac{n}{df}} \qquad g = d\sqrt{\frac{df}{n}}$$

Die Stichprobengröße bezieht sich dabei auf die gesamte Stichprobe. Bei einem Vergleich von zwei Gruppen mit je 10 Personen, beträgt $n = 20$. Die Anzahl der Freiheitsgrade bestimmt sich bei Gruppenvergleichen immer durch $df = n - k$ (der Stichprobengröße minus der Anzahl von Gruppen *k*, also 2).

Aus Abstandsmaßen können – bei gleichen Stichprobengrößen – Korrelationen wie folgt berechnet werden:

$$r = \frac{d}{\sqrt{d^2 + 4}} \qquad r = \frac{g}{\sqrt{g^2 + 4(\frac{df}{n})}}$$

Und schließlich kann man Abstandsmaße aus Korrelationen berechnen:

$$d = \frac{2r}{\sqrt{1-r^2}} \qquad g = \frac{r}{\sqrt{1-r^2}}\sqrt{\frac{(n_A+n_B)df}{n_A n_B}}$$

Man sollte allerdings beachten, dass es in der Regel wenig Sinn macht, eine Korrelation zweier kontinuierlicher Variablen als Unterschiedsmaß auszudrücken, da es keine Gruppen gibt, die man vergleichen könnte. Die Verwendung von Effektgrößen sollte sich demnach immer an der inhaltlichen Fragestellung orientieren.

4.4 Effektgrößen aus Signifikanztestergebnissen

Das Ergebnis eines Signifikanztests enthält eine indirekte Information über die Größe des Effektes, die man sich zunutze machen kann. Denn dieses Ergebnis ist natürlich vom gefundenen Effekt (zum Beispiel einem Mittelwertsunterschied) abhängig (aus ihm wird ja der *t*-Wert bestimmt). Bemängelt hatten wir am Signifikanztest, dass man aus seinem Ergebnis die Größe des Effektes nicht ablesen kann, weil es von der Stichprobengröße abhängig ist. Diesen Umstand kann man dadurch „umgehen", dass man das Signifikanztestergebnis durch die Stichprobengröße teilt. Als Faustregel kann man sich das so vorstellen:

$$Effektgröße = \frac{Signifikanztestergebnis}{Größe\ der\ Stichprobe}$$

Mit „Signifikanztestergebnis" ist dabei die jeweilige Prüfgröße (zum Beispiel ein *t*-Wert) gemeint, und die „Größe der Stichprobe" wird meist mit Hilfe der Freiheitsgrade angegeben. Für alle Arten von Signifikanztests gibt es entsprechende Formeln, mit denen man aus der Prüfgröße die Effektgröße berechnen kann. Diese Formeln werden wir uns bei den jeweiligen Signifikanztests immer mit ansehen. Die Effektgrößen gelangen im Übrigen stets zum selben Ergebnis – egal, auf welchem der drei beschriebenen Wege sie berechnet werden.

4.5 Interpretation von Effektgrößen

Da wir Effektgrößen als standardisierte Maße verwenden wollen, die über Studien und Themenbereiche hinweg vergleichbar sein sollen, stellt sich noch die Frage, wie sie zu interpretieren sind. Wann ist eine Effektgröße groß und wann klein? Darauf kann man zwei Antworten geben. Erstens ist diese Interpretation von der jeweiligen Fragestellung der Studie abhängig. In manchen Gebieten – wir hatten das Krebsmedikament als Beispiel – kann schon ein sehr kleiner Effekt bedeutsam sein. In anderen Bereichen möchte man eher große Effekte erzielen. Eine gute Orientierungshilfe sind dabei auch die Effekte, die in vorangegangenen Studien zum gleichen Thema gefunden wurden. Wenn etwa Studien zum Einfluss von Hintergrundmusik auf die Gehgeschwindigkeit in Fußgängerzonen durchschnittliche Effekte *d* von ca. 0,4 zeigen, kann man in seiner eigenen Studie zum selben Thema kaum Effekte von 0,8 erwarten. Man sollte dann allerdings auch skeptisch sein, wenn man selbst nur einen Effekt von 0,1 gefunden hat.

Die andere Möglichkeit, die Bedeutsamkeit einer Effektgröße zu beurteilen, besteht in der Anwendung von Konventionen. Solche Konventionen wurden von Cohen (1988) vorgeschlagen und werden in der Psychologie für die Interpretation von Effektgrößen in der Mehrzahl der Fälle verwendet. Sie sind in Tabelle 4.2 zusammengefasst.

Tabelle 4.2 Konventionen für die Interpretation von Effektgrößen

	d und *g*	*r*
klein	ab 0,2 bzw. -0,2	ab 0,1 bzw. -0,1
mittel	ab 0,5 bzw. -0,5	ab 0,3 bzw. -0,3
groß	ab 0,8 bzw. -0,8	ab 0,5 bzw. -0,5

Die Abstandsmaße *d* und *g* sind in ihrer Größe prinzipiell nach oben offen. Allerdings findet man in der Psychologie nur selten Effekte, die über 1 bzw. -1 hinausgehen. Der Korrelationskoeffizient *r* hat seine Grenze bei 1 bzw. -1. Bei der Interpretation ist allerdings zu beachten, dass Unterschiede in der Größe von *r* im oberen Wertebereich viel bedeutsamer sind als Unterschiede im unte-

ren Bereich. So ist der Unterschied zwischen zwei Korrelationskoeffizienten von 0,8 und 0,9 wesentlich bedeutsamer als der Unterschied zwischen 0,2 und 0,3.

4.6 Effektgrößen, Konfidenzintervalle und Signifikanztests im Vergleich

Wie wir gesehen haben, kann man Effektgrößen nicht direkt als inferenzstatistische Aussagen betrachten. Sie liefern lediglich eine standardisierte Schätzung der Größe des in der Stichprobe gefundenen Effektes. Effekt und Effektgröße sind die beiden aussagekräftigsten Angaben über die Ergebnisse einer Studie. Hinzu kommt die Angabe über die Güte der Schätzung des Effektes von der Stichprobe auf die Population, die mit Hilfe der drei möglichen inferenzstatistischen Verfahren gemacht werden kann. Der Standardfehler wird bei Effekten meist nicht angegeben. Allerdings bildet er die Grundlage für die Berechnung von Konfidenzintervallen. Die beiden Grenzen von Konfidenzintervallen – in Rohwerten ausgedrückt – liefern eine leicht verständliche inferenzstatistische Aussage. Diese sollte als echte Alternative zum Signifikanztest gesehen werden, dessen Interpretation immer wieder Schwierigkeiten bereitet. Für welche Möglichkeit man sich letztendlich entscheidet, bleibt dem Anwender überlassen. Die Angabe von Signifikanztestergebnissen ist in wissenschaftlichen Publikationen leider immer noch Standard. Allerdings wird inzwischen die Angabe von Effektgrößen ausdrücklich gefordert (zum Beispiel durch die Publikationsrichtlinien der American Psychological Association, APA). Abbildung 4.1 zeigt die verschiedenen Möglichkeiten noch einmal in der Übersicht am Beispiel eines Mittelwertsunterschiedes.

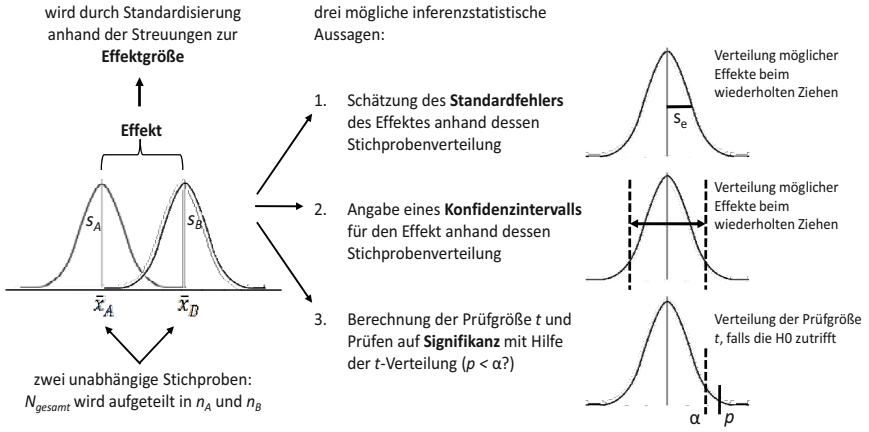

Abbildung 4.1 Effektgrößen und inferenzstatistische Verfahren am Beispiel eines Mittelwertsunterschiedes

Für Zusammenhangsfragestellungen gilt diese Übersicht gleichermaßen. Der Effekt besteht dann im Korrelationskoeffizienten *r* (bzw., bei mehreren Prädiktoren, in den Regressionsgewichten *b*), für den der Standardfehler und das Konfidenzintervall angegeben werden kann. Die Prüfgröße für den Signifikanztest ist hier ebenfalls ein *t*-Wert.

Literaturempfehlung:

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 9)

5 Das Allgemeine Lineare Modell und die Multiple Regression

5.1 Das Allgemeine Lineare Modell (ALM): Alle Fragestellungen sind Zusammenhänge

Signifikanztests berechnen aus einem gefundenen Effekt (zum Beispiel einem Mittelwertsunterschied) eine Prüfgröße (zum Beispiel einen *t*-Wert), die dann mit Hilfe einer Prüfverteilung auf Signifikanz geprüft wird. Unterschiedliche Fragestellungen führen zu unterschiedlichen Effekten und schließlich zu unterschiedlichen Prüfgrößen und Prüfverteilungen. Daher gibt es eine Reihe verschiedener Signifikanztests, die jeweils für spezielle Fragestellungen zu verwenden sind. Das Prinzip dieser Tests ist dabei aber stets das gleiche.

Generell lassen sich zwei größere Gruppen von Tests unterscheiden: Tests für Mittelwertsunterschiede (also für den Vergleich verschiedener Gruppen) und Tests für Zusammenhänge. An dieser Stelle könnten wir nun beginnen, diese Tests nacheinander vorzustellen und zu diskutieren, für welche Fragestellung sie gelten. Wir werden allerdings etwas anders verfahren und ein wichtiges Prinzip an den Anfang stellen, dessen Verständnis die Grundlage für alle Signifikanztests bildet. Diesem Prinzip sind Sie schon mehrmals „am Rande" begegnet. Sicher erinnern Sie sich an die Behauptung, dass Unterschieds- und Zusammenhangsfragestellungen ineinander überführbar sind und dass die Frage nach Zusammenhängen die grundlegendste aller statistischen Fragen darstellt. Das heißt, dass jede Fragestellung im Grunde als eine Zusammenhangsfragestellung betrachtet werden sollte. Zusammenhänge fragen immer danach, wie Variablen miteinander in Beziehung stehen und wie sich Variablen aus anderen Variablen vorhersagen lassen. Diese Beziehung zwischen Variablen ist eigentlich eine relativ einfache Sache. Und tatsächlich lässt sie sich in einem einfachen mathema-

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_6,

tischen Prinzip formulieren: dem sogenannten *Allgemeinen Linearen Modell*, oder kurz ALM. Das ALM spannt sich wie eine Art mathematischer Schirm über fast alle Arten von Signifikanztests und vereint die in verschiedenen Tests auftauchenden Berechnungen. Das erklärt schon den Begriff *allgemein* im ALM. Aber was bedeutet *linear*? Diese Frage ist schnell beantwortet: damit ist nichts anderes gemeint als der lineare Zusammenhang von Variablen. Das heißt, die Variablen stehen in einer Beziehung, die sich durch eine lineare Regressionsgerade beschreiben lässt (nicht-lineare oder auch rückgekoppelte Beziehungen werden also vernachlässigt). Das ALM verallgemeinert das Prinzip der Regression noch ein wenig. Wie genau, das sehen wir uns gleich im Detail an. Als wichtigsten Punkt können wir aber zunächst festhalten, dass das ALM die mathematische Grundlage für die meisten Signifikanztests bildet, dass es alle Arten von Fragestellungen als Zusammenhänge definiert und dass die Regression eine direkte Ableitung aus dem ALM darstellt.

Bleibt noch der Begriff *Modell* zu klären. Und damit sind wir an einem sensiblen Punkt in der Methodenlehre, an dem sich Statistik und Wirklichkeit berühren. Ein Modell ist eine vereinfachte Darstellung der Wirklichkeit. Der Sinn dieser Vereinfachung ist, dass man leichter damit arbeiten kann. Für die Psychologie bedeutet das, dass wir die „Wirklichkeit" des Erlebens und Verhaltens sowohl in handhabbare Einzelheiten zerlegen als auch in bestimmte mathematische Vorstellungen pressen. Zumindest ist das das Vorgehen der quantitativen Methoden. Erinnern Sie sich an das Problem des *Messens* – es bestand in der Schwierigkeit, die psychologische Wirklichkeit empirisch einzufangen und in Zahlen zu übersetzen. Was dabei übrig bleibt, ist ein Modell dieser Wirklichkeit. Für alle Analyseverfahren, die wir jetzt noch besprechen werden, gilt dasselbe. Sie liefern im Prinzip keine Ergebnisse über die Wirklichkeit, sondern lediglich über unsere Modelle, die wir uns von der Wirklichkeit machen. Damit wäre es streng genommen sinnvoller, nicht von Analyseverfahren, sondern immer von Analyse*modellen* zu sprechen. Während die Bezeichnung Allgemeines Lineares Modell dem Rechnung trägt, tun das alle anderen Verfahren nicht. Es wäre also schön, wenn Sie diesen wichtigen Punkts stets im Hinterkopf behalten.

Die Mathematik hinter dem ALM

Hinter dem ALM verbirgt sich nur eine einzige Formel, deren Aussage sehr leicht zu verstehen ist. Sie beschreibt, wie ein konkreter Messwert zustande kommt. Egal, ob wir den IQ, die soziale Kompetenz, das Selbstwertempfinden, die Extraversion oder was auch immer messen – diese Messungen werden immer an Personen vorgenommen. Das heißt, dass für jede Person ein konkreter Messwert Y existiert. Und das grundlegende Ziel der Psychologie ist es, Variablen zu finden, die diese Messwerte vorhersagen können. Diese Vorhersage beruht also auf Prädiktoren oder Prädiktorvariablen (X_1, X_2, X_3 usw.). Diese Prädiktoren finden sich in der Formel des ALM wieder:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + e$$

Y ist der konkrete Wert, den eine Person auf einer Variable tatsächlich hat, also beispielsweise ihr wahrer Wert für die Variable Empathievermögen. In der Regression wird diese Variable als *Kriterium* (manchmal auch einfach als *AV*) bezeichnet. Wie kann dieser Wert vorhergesagt werden? Die beste Vorhersage in dem Fall, dass man keine weiteren Variablen untersucht hat, ist der Mittelwert der Empathie einer Vielzahl von Personen, für die bereits Messungen vorliegen. Nehmen wir an, wir hätten 100 Personen untersucht und ihr Mittelwert vom Empathievermögen beträgt 6,8 auf einer Skala von 1 bis 10. Wenn wir keine weiteren Informationen haben, müssen wir diesen Mittelwert als den besten Schätzwert auch für die eine Person benutzen, um die es geht. Diese Schätzung wird durch die Regressionskonstante *a* repräsentiert. Diese ist für alle Personen gleich und liefert meist schon eine recht brauchbare Vorhersage für Y (da der Mittelwert immer eine recht zuverlässige Schätzung ist). Mit der Regressionskonstante allein können wir aber den Wert von Y nie exakt vorhersagen. Es sei denn, der Wert dieser Person ist tatsächlich genau mit dem Mittelwert identisch, aber in den meisten Fällen wird das nicht so sein. Wie können wir die Vorhersage verbessern? Diese Verbesserung steckt im Rest der Formel. Wir erheben weitere Variablen, von denen wir glauben, dass sie in der Lage sind, Y vorherzusagen. Das sind also Variablen, die – abgesehen von *a* – die konkrete Vorhersage von Y noch verbessern können. Eine solche Variable könnte in unserem Fall das Alter sein. Alter wäre dann die Variable X_1, also ein erster Prädiktor, der den vorhergesagten Wert von Y nun vom Mittelwert ausgehend noch etwas verändert, um die Vorhersage zu verbessern. Damit sollte der wahre Wert von Y ein Stück besser getroffen

sein. Wie groß der Einfluss des Alters auf die Vorhersage von Y ist, wird durch das Regressionsgewicht b_1 festgelegt. Es beschreibt lediglich, wie stark das Alter überhaupt mit Empathievermögen zusammenhängt und wie gut es daher zur Vorhersage geeignet ist. Das Produkt b_1X_1 verbessert also die Vorhersage ein Stück. Diese Prozedur können wir nun für weitere Prädiktoren wiederholen. Zum Beispiel könnten wir noch den IQ erheben, in der Hoffnung, dass auch er einen Einfluss auf Empathievermögen ausübt. Wenn ja, würde das Produkt b_2X_2 die Vorhersage von Y weiter verbessern. Das können wir nun solange fortsetzen, bis wir alle möglichen Prädiktoren in die Gleichung aufgenommen haben. Die Vorhersage von Y würde sich dabei immer mehr verbessern. Allerdings werden wir nie alle möglichen Prädiktoren finden bzw. messen können, die irgendwie mit Y zusammenhängen. Denn psychologische Merkmale sind in der Regel von einer Vielzahl von anderen Merkmalen abhängig, die man nie alle berücksichtigen kann.

Und es kommt ein zweites Problem hinzu. Beim Messen all dieser Variablen machen wir Fehler. Solche *Messfehler* sind unvermeidbar. Die unbekannten bzw. nicht untersuchten Variablen und die Messfehler führen dazu, dass wir den Wert von Y nie exakt vorhersagen können. Um dem Rechnung zu tragen, endet die Gleichung des ALM mit einem Fehlerterm e (der alle konkreten Fehler enthält). Erst wenn wir diesen Fehler mit einbeziehen, wäre eine exakte Vorhersage von Y möglich. Das Problem dabei liegt natürlich auf der Hand: wir kennen den Fehler nicht. Der Fehler ist eine unbekannte Größe für uns. Daher müssen wir uns immer damit begnügen, dass wir den wahren Wert von Y nicht exakt bestimmen können. Stattdessen können wir ihn lediglich schätzen.

Die Gleichung des ALM sagt also nichts weiter aus, als dass ein konkreter Wert einer Person aus der Regressionskonstante, einer Reihe von Prädiktoren und einem Fehler vorhergesagt werden kann. Die Prädiktoren können dabei natürlich auch als unabhängige Variablen in Experimenten aufgefasst werden. Besteht ein Experiment beispielsweise im Vergleich einer Kontrollgruppe und einer Experimentalgruppe, so besagt das ALM, dass der Wert einer Person auf der AV durch die Gruppenzugehörigkeit vorhergesagt werden kann. Der Gruppenunterschied wird also als Zusammenhangsfragestellung aufgefasst. Und natürlich ist die Messung in einem Experiment immer mit einem Fehler behaftet. Das ALM beschreibt daher das grundlegende Prinzip des Zusammenhangs und der Vorhersage von Variablen, auf das sich jede Fragestellung reduzieren lässt.

Sehen wir uns die Formel des ALM noch einmal an, erkennen wir das eben genannte Problem, dass wir offenbar für den Fehler e keinen Wert einsetzen kön-

nen, da dieser immer unbekannt ist. Um das ALM für statistische Berechnungen nutzbar zu machen, muss der Fehler aus der Gleichung entfernt werden. Das führt dazu, dass der Wert von Y nicht mehr exakt vorhergesagt, sondern nur noch geschätzt werden kann. Genau diese kleine Abwandlung der Formel führt damit zum grundlegendsten Verfahren der Statistik: der Multiplen Regression.

Literaturempfehlung

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall. (Chapter 15)

5.2 Die Multiple Regression

Das Rechenverfahren, das direkt aus dem ALM folgt, ist die Multiple Regression. Sie beschreibt, wie man den Wert einer Person auf einer Variable aufgrund mehrerer Prädiktoren schätzen kann. Da der Messfehler bei einer solchen Vorhersage nicht bekannt ist, resultiert aus dieser Vorhersage immer nur ein Schätzwert für Y. Daher erhält das Y ein Dach, das diese Schätzung anzeigt: $\hat{Y}$. Die Formel für die Multiple Regression besteht damit nur noch aus der Regressionskonstante und den Prädiktoren mit ihren Regressionskoeffizienten:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots$$

Auch an dieser Formel wird deutlich, dass sich die Schätzung von Y immer mehr verbessert, je mehr relevante Prädiktoren gefunden werden, die Y vorhersagen können. Unbekannte oder nicht berücksichtigte Prädiktoren würden mit zum Messfehler gehören und damit die Schätzung ungenauer machen. Ungenaue Schätzungen liegen dann vor, wenn der vorhergesagte Wert für Y, also $\hat{Y}$, nicht mit dem tatsächlichen Wert für Y übereinstimmt. Diese Differenz kann man nur bestimmen, wenn man den wahren Wert von Y kennt. Das ist aber nur dann der Fall, wenn man die wahren Werte der Personen gemessen hat. Mit diesem Sonderfall haben wir es meist zu tun, denn wir machen ja entsprechende Studien und kennen daher die wahren Werte der Studienteilnehmer. Wir können für die erhobene Stichprobe also diese Differenz ausrechnen. Und Sie wissen auch schon, wie sie genannt wird: sie heißt *Vorhersagefehler* oder *Residuum*.

Erinnern Sie sich aber daran, dass wir in der Inferenzstatistik versuchen, gefundene Zusammenhänge auf die Population zu verallgemeinern. Das heißt,

dass wir dann natürlich Vorhersagen für *unbekannte* Personen machen, bei denen wir den wahren Wert *nicht* kennen. Daher wissen wir bei einer konkreten Person nie, wie nahe die Vorhersage an ihrem wahren Wert liegt. Wir können nur den durchschnittlichen Fehler angeben, den wir bei dieser Schätzung machen. Das sehen wir uns später noch genauer an.

Varianzaufklärung: Schlüsselkonzept von Methodenlehre und Statistik

Die Multiple Regression beschreibt – genau wie das ALM – die Tatsache, dass konkrete Werte von Personen auf bestimmten Merkmalen (Variablen) durch deren Ausprägung auf anderen Variablen vorhersagbar sind. In dieser Tatsache ist die grundlegende Idee der Psychologie im Allgemeinen und der Methodenlehre und Statistik im Speziellen verkörpert: die Idee der Varianzaufklärung. Menschen haben auf bestimmten uns interessierenden Merkmalen unterschiedliche Ausprägungen. Wenn das nicht so wäre, dann gäbe es sozusagen nichts, was wir erklären könnten. Denn psychologische Erklärungen beziehen sich immer auf Variablen. Etwas, was sich nicht verändert (was nicht variabel ist), ist für die Psychologie nicht von Interesse. Wo immer es Veränderungen – also Varianz – gibt, ist die Psychologie auf der Suche nach Erklärungen. Und im Prinzip gibt es zwei Wege, die die psychologische Forschung geht, um solche Erklärungen für menschliches Erleben und Verhalten zu finden. Der erste Weg besteht im Beobachten und Befragen, das heißt, im bloßen Sammeln von Daten, in denen anschließend nach Zusammenhängen gesucht wird. Beispielsweise könnten wir von Schülern den Schulerfolg, ihren IQ und ihre Sozialkompetenz erheben und später untersuchen, wie diese Variablen statistisch zusammenhängen. Und schließlich werden wir bestimmte Vorstellungen darüber haben, welche Variable eine andere Variable kausal beeinflusst, also als Erklärung in Frage kommt. Mit Hilfe der Regression können wir prüfen, wie stark die Vorhersagekraft der einen für die andere Variable ist.

Der zweite Weg, um Erklärungen zu finden, besteht im Durchführen von Experimenten. Hier wird Erleben und Verhalten nicht einfach beobachtet oder erfragt, sondern hier wird zunächst ein Stück Realität künstlich erschaffen. Das geschieht durch Versuchsbedingungen, die einer systematischen Variation folgen. Bei Experimenten stehen die zu untersuchenden Hypothesen immer schon vorher fest, und das Experiment soll zeigen, welche Hypothese zutrifft. Wenn

wir beispielsweise wissen wollen, ob ein Training der sozialen Kompetenz den Schulerfolg erhöht, würden wir zufällig ausgewählte Schüler zufällig in zwei Gruppen teilen, von denen die eine ein solches Training erhält, die andere nicht. Danach würden wir die Veränderung im Schulerfolg messen und daraufhin entscheiden, ob die Daten für oder gegen die Wirksamkeit des Trainings sprechen. Wenn das Training wirkt, dann wissen wir, dass es als Erklärung für höheren Schulerfolg anzusehen ist. Das hört sich zunächst sehr praktisch an, aber im Grunde steckt in einem solchen Ergebnis immer auch ein wichtiger psychologischer Befund. Wir haben etwas über das Funktionieren des Erlebens und Verhaltens dazu gelernt. In der Grundlagenforschung ist es oft so, dass aus Studien kein direkter praktischer „Nutzen" folgt, sondern sie dienen eher dazu, psychologische Mechanismen aufzudecken. Das Prinzip dahinter ist aber immer dasselbe. Der Forscher versucht Varianz in einer AV herzustellen, indem er Versuchsbedingungen (die UV) variiert. Wenn ihm das nicht gelingt, war seine Variation kein kausaler Faktor für die AV. Wenn es ihm aber gelingt, dann hat er eine Ursache für Veränderungen in der AV gefunden. Psychologische Forschung und psychologischer Erkenntnisgewinn drehen sich also immer um die Aufklärung von Varianz – ganz egal, ob diese einfach nur beobachtet oder künstlich erzeugt wurde.

Und diese Überlegungen bringen uns zurück zur Idee des ALM und der Multiplen Regression: diese beschreiben, wie die Varianzaufklärung rechnerisch dingfest gemacht werden kann. Im einfachsten Fall haben wir es mit einer *einfachen* linearen Regression zu tun. Diese beachtet nur einen Prädiktor. Bei Experimenten, in denen nur eine UV variiert wurde, ist die einfache lineare Regression ausreichend. Der eine Prädiktor besteht dann in der einen UV. Da man bei Experimenten davon ausgeht, dass man Störvariablen kontrolliert bzw. ausgeschaltet hat, müssen in der Regression keine weiteren Variablen auftauchen. Denn der Wert von Y sollte allein durch die UV (und die Regressionskonstante) vorhersagbar sein. Außer der UV gibt es nämlich nichts, was noch variieren konnte. Störende Einflüsse sollten dann nur noch vom Messfehler herrühren.

Sobald in einem Experiment mehrere UVs untersucht wurden, müssten entsprechend mehr Prädiktoren in die Regression aufgenommen werden. Das Gleiche gilt, wenn man durch Beobachtungen und Befragungen eine Vielzahl von Variablen erhoben hat und anschließend eine Variable durch viele andere Variablen vorhersagen möchte. In einer solchen Multiplen Regression können im Prinzip beliebig viele Prädiktoren aufgenommen werden. Im schlimmsten Fall

würde man Prädiktoren in die Regression aufnehmen, die keine Vorhersagekraft für Y haben. Das hätte aber keinen nachteiligen Effekt auf die Vorhersage. Denn das würde lediglich heißen, dass der Prädiktor nicht mit dem Kriterium korreliert. Das Regressionsgewicht b dieses Prädiktors wäre damit 0 und das entsprechende Produkt bX würde einfach aus der Gleichung entfallen. (Und natürlich könnte man dann auch nicht mehr von einem „Prädiktor" sprechen.)

Regressionsgewichte in der Multiplen Regression

Bei der einfachen linearen Regression ist das Regressionsgewicht sozusagen das Hauptergebnis. Bei der Multiplen Regression sieht das etwas anders aus. Hier gibt es mehrere Prädiktoren und jeder Prädiktor erhält ein eigenes Regressionsgewicht. Der Einfachheit halber werden wir nur noch vom standardisierten Regressionsgewicht β sprechen, da dieses am häufigsten benutzt wird und einfacher zu interpretieren ist. Jedes Regressionsgewicht beschreibt, wie stark der Prädiktor mit dem Kriterium zusammenhängt – allerdings in einem *relativen* Sinn. Das bedeutet, dass die Vorhersagekraft eines Prädiktors *abhängig von allen anderen Prädiktoren* beurteilt wird. Anders formuliert soll der *isolierte* bzw. *systematische* Einfluss eines Prädiktors auf das Kriterium beurteilt werden, der nur durch diesen Prädiktor zustande kommt. Was ist damit gemeint? Im Normalfall korrelieren die Prädiktoren untereinander. Wenn wir beispielsweise Schulerfolg durch die beiden Prädiktoren IQ und Sozialkompetenz (SK) vorhersagen wollen, dann werden die beiden Prädiktoren wahrscheinlich auch untereinander korrelieren. Was bedeutet das für die Vorhersage von Y (dem Schulerfolg)? Sehen wir uns die Formel noch einmal an:

$$\hat{Y} = a + \beta_{IQ} X_{IQ} + \beta_{SK} X_{SK}$$

Wenn wir von links nach rechts vorgehen, würden wir zunächst danach fragen, wie groß die Vorhersagekraft von IQ für Y ist. Im zweiten Schritt würden wir fragen, wie stark die Vorhersagekraft von SK für Y ist. Wenn aber IQ und SK korrelieren, heißt das, dass die Vorhersagekraft von SK für Y bereits im IQ teilweise enthalten war. Was damit gemeint ist, kann man sich durch ein sogenanntes *Venn-Diagramm* veranschaulichen. In einem Venn-Diagramm wird für eine Variable ein Kreis gezeichnet, der für die Varianz der Variable steht. Wenn Kreise sich überdecken, heißt das, dass sie einen gemeinsamen Varianzanteil besitzen. Je

mehr sich die Kreise überdecken, desto stärker ist die Korrelation der Variablen. Für unser Beispiel könnte ein Venn-Diagramm etwa so aussehen.

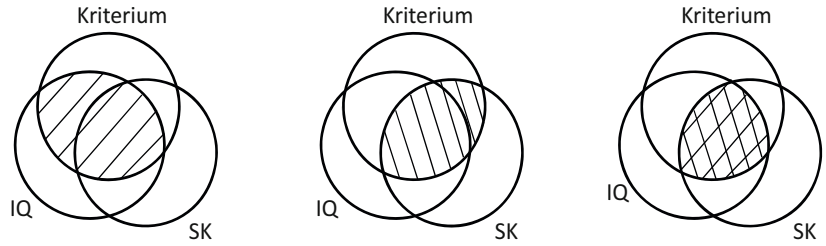

Abbildung 5.1 Venn-Diagramm für drei Variablen mit gemeinsamen Varianzanteilen

Im linken Diagramm ist die Korrelation zwischen IQ und Schulerfolg anhand der schraffierten Fläche dargestellt. Wie man sehen kann, werden ungefähr zwei Drittel der Fläche von Schulerfolg auch vom IQ überdeckt. Das heißt, dass der IQ zwei Drittel der Varianz von Schulerfolg aufklären kann. Ähnlich verhält es sich bei SK und Schulerfolg (mittleres Diagramm). SK kann ungefähr die Hälfte der Varianz von Schulerfolg aufklären. Spannend ist nun aber das ganz rechte Diagramm. Hier sehen wir, dass auch IQ und SK untereinander korrelieren. Damit gibt es einen Flächenanteil von Schulerfolg (der schraffierte), der von *beiden* Prädiktoren gleichzeitig erklärt wird. Für diesen Flächenanteil wäre es also egal, ob er durch IQ oder durch SK erklärt wird. Für die Multiple Regression heißt das, dass für den isolierten Einfluss von SK nun nicht mehr die gesamte Korrelation mit dem Kriterium relevant ist, sondern nur noch der Anteil, der *nicht ebenfalls* durch IQ erklärt wird. In diesem Fall – wenn die Prädiktoren korreliert sind – ist β daher nicht mehr mit der Korrelation zwischen Prädiktor und Kriterium identisch, sondern beschreibt den isolierten Einfluss des Prädiktors, der von allen anderen Einflüssen *bereinigt* ist. Anders ausgedrückt: β gibt die spezifische Vorhersagekraft wieder, die nur durch diesen Prädiktor und durch keinen anderen gegeben ist.

In der Multiplen Regression beschreiben die Regressionsgewichte β_i den spezifischen bzw. isolierten Einfluss eines Prädiktors auf die Vorhersage des Kriteriums. Die Regressionsgewichte sind dabei um den Einfluss anderer Prädiktoren bereinigt.

Wenn die Prädiktoren untereinander nicht korreliert sind, dann ist ihr isolierter Einfluss genauso groß wie ihre jeweilige Korrelation mit dem Kriterium. Dann sind die Regressionsgewichte also wieder mit den einzelnen Korrelationskoeffizienten identisch. Unkorrelierte Prädiktoren kommen aber nur äußerst selten vor.

Man kann sich die Vorhersage eines Kriteriums durch mehrere Prädiktoren in einem 3D-Diagramm vorstellen (siehe Abbildung 5.2). Hier wird der Einfluss unserer beiden Prädiktoren gleichzeitig sichtbar. Wenn es bei einem Prädiktor eine Regressions*gerade* gibt, müssten Sie sich bei zwei Prädiktoren entsprechend eine Regressions*ebene* vorstellen, deren Zentrum genau durch den Mittelwert aller Daten verläuft und die so geneigt ist, dass sie wiederum alle Datenpunkte so gut wie möglich repräsentiert. Je nachdem, wie sich die Werte von Personen auf beiden Prädiktoren verändern, würde die Ebene anzeigen, wie sich der Wert des Kriteriums verändert.

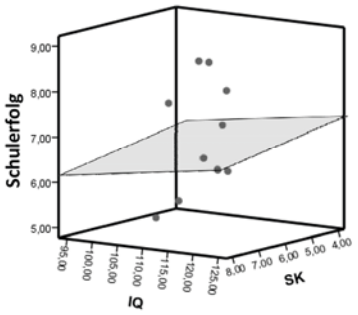

Abbildung 5.2 Regressionsebene bei zwei Prädiktoren

Das Problem an dieser Art Darstellung ist, dass sie nur für zwei Prädiktoren geeignet ist. Ab drei Prädiktoren ist keine grafische Darstellung mehr möglich. Dennoch kann diese Abbildung helfen, sich die Wirkung mehrerer Prädiktoren besser vorzustellen.

Die globale Güte der Vorhersage

Bei der einfachen linearen Regression steckt die Güte der Vorhersage bereits im Regressionsgewicht β, welches mit der Korrelation r von Prädiktor und Kriterium identisch ist. Der Determinationskoeffizient ergibt sich aus dem Quadrat der Korrelation (r^2) und gibt den Anteil der aufgeklärten Varianz wieder. Bei der Multiplen Regression sieht das etwas anders aus. Hier können die Regressionsgewichte nur noch als relative Einflussgrößen interpretiert werden. Das heißt, dass zwar immer noch größere Betas einen größeren Einfluss des Prädiktors auf das Kriterium anzeigen. Allerdings kann man aus den Betas die absolute Größe des Zusammenhangs nicht mehr ablesen. Mit anderen Worten: wir können ein Beta nicht einfach quadrieren und daraus den Anteil aufgeklärter Varianz ablesen, weil es eben nicht mehr der simplen Korrelation entspricht.

Allerdings kommt uns hier der Umstand zu Hilfe, dass wir in der Regel gar nicht an den einzelnen Varianzaufklärungen interessiert sind. Denn was uns eigentlich interessiert, ist die Frage, wie gut *alle Prädiktoren zusammen* das Kriterium vorhersagen bzw. erklären können. Die Auflistung mehrerer Prädiktoren in der Regression wird übrigens *Modell* genannt. In unserem Beispiel besteht das Modell aus zwei Prädiktoren: IQ und SK. Wir wollen nun wissen, wie gut das gesamte Modell für die Vorhersage von Schulerfolg geeignet ist. Dafür gibt es zwei verschiedene Möglichkeiten: den Standardschätzfehler und den multiplen Determinationskoeffizient R^2, die wir schon aus der einfachen linearen Regression kennen.

Beginnen wir mit dem multiplen Determinationskoeffizient R^2, der das Ausmaß der aufgeklärten Varianz angibt (in der Multiplen Regression wird das R groß geschrieben). Dessen Bestimmung ist relativ einfach, da wir die Varianz des Kriteriums Y kennen (also die Varianz der tatsächlichen Werte) und auch die Varianz der vorhergesagten Werte $\hat{Y}$. Diese beiden Varianzen müssen wir nur ins Verhältnis setzen:

$$R^2 = \frac{Varianz\ der\ vorhergesagten\ Werte}{Varianz\ der\ tatsächlichen\ Werte} = \frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2}$$

Diese Formel gilt für standardisierte Werte von Y und $\hat{Y}$. Auf die Berechnung der Varianzen müssen wir an dieser Stelle nicht genauer eingehen. Entscheidend ist das Prinzip der Varianzaufklärung. Der Determinationskoeffizient R^2

wird von allen Statistikprogrammen als Gütemaß ausgegeben und wird meist als Hauptergebnis der Multiplen Regression benutzt.

> Der multiple Determinationskoeffizient R^2 gibt den Anteil von Varianz des Kriteriums wieder, der durch alle Prädiktoren gemeinsam erklärt wird. Er kann maximal 1 sein, was einer Varianzaufklärung von 100% entspricht.

Alternativ kann man sich den Determinationskoeffizienten auch als das Quadrat des sogenannten *multiplen Korrelationskoeffizienten R* vorstellen. *R* ist nichts anderes als die gemeinsame Korrelation des Kriteriums mit all seinen Prädiktoren. Wenn alle Prädiktoren unkorreliert sind, ergibt sich *R* aus der Summe aller *r*s zwischen dem jeweiligen Prädiktor und dem Kriterium. Wenn die Prädiktoren korreliert sind (so wie IQ und SK in Abbildung 5.1), ist *R* immer kleiner als die Summe aller *r*s. Entscheidend ist, welche Fläche des Kriteriums von den Prädiktoren insgesamt überdeckt und damit erklärt wird.

Der Determinationskoeffizient steht in direkter Beziehung zum Standardschätzfehler, der eher selten als Gütemaß angegeben wird. Er beschreibt die durchschnittliche Differenz von vorhergesagten und tatsächlichen Werten und kann direkt aus R^2 bestimmt werden:

$$s_e = s_y\sqrt{1 - R^2}$$

An der Formel erkennt man, dass ein Determinationskoeffizient nahe 1 zu einem Standardschätzfehler führt, der gegen 0 geht.

> Der Standardschätzfehler s_e in der Multiplen Regression gibt an, wie stark die vorhergesagten Werte von den tatsächlichen Werten des Kriteriums im Durchschnitt abweichen.

Standardschätzfehler und Determinationskoeffizient geben das Ausmaß der Varianzaufklärung an und verkörpern damit die Güte, mit der es uns gelingt, Variablen als Erklärungen oder gar Ursachen für andere Variablen zu finden. Neben der Höhe der Varianzaufklärung stellt sich aber auch hier die Frage, wie gut ein gefundenes Modell auf die Population verallgemeinert werden kann. Dafür benötigen wir inferenzstatistische Aussagen, die wir uns im Folgenden anschauen.

Der Signifikanztest bei der Multiplen Regression

Inferenzstatistische Aussagen für die einfache lineare Regression haben wir in den vorangegangenen Kapiteln bereits besprochen. Das zentrale Ergebnis der einfachen linearen Regression war das Regressionsgewicht *b*. Es gibt die Stärke des Zusammenhangs zwischen Prädiktor und Kriterium an. Seine standardisierte Variante β ist in der einfachen linearen Regression mit dem Korrelationskoeffizienten *r* identisch. Für das Regressionsgewicht können wir einen Standardfehler oder ein Konfidenzintervall angeben sowie einen Signifikanztest berechnen, um zu erfahren, ob der gefundene Zusammenhang nur für unsere Stichprobe gilt oder auf die Population verallgemeinert werden kann. Die Prüfverteilung ist dabei die bereits bekannte *t*-Verteilung, die für Korrelationskoeffizienten und für Regressionsgewichte ebenso gilt wie für Mittelwertsunterschiede. Der gefundene Korrelationskoeffizient muss nur in einen *t*-Wert umgerechnet werden:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

n steht hier für die Anzahl der Wertepaare, die in der Korrelation benutzt wurden. Die Anzahl der Freiheitsgrade beträgt immer $df = n - 2$. Hätten wir etwa eine Korrelation von $r = .40$ bei 20 Personen gefunden, entspräche das einem *t*-Wert von: $t = \frac{0{,}4\sqrt{20-2}}{\sqrt{1-0{,}4^2}} = 1{,}85$. Laut *t*-Tabelle wäre dieser Wert bei einem Alpha-Niveau von 5% signifikant. Ein standardisiertes Regressionsgewicht β kann genau wie *r* behandelt und in die Formel eingesetzt werden. Das unstandardisierte *b* kann durch eine Standardisierung in *r* überführt werden:

$$r = \beta = b\frac{s_x}{s_y}$$

Wie wir gesehen haben, sind die Regressionsgewichte in der Multiplen Regression nicht mehr mit den Korrelationskoeffizienten identisch. Auf die Berechnung der Regressionsgewichte können wir aus Platzgründen nicht weiter eingehen. Und im Prinzip ist das auch nicht nötig, solange uns die Aussage der Regressionsgewichte klar ist. Wie oben beschrieben, geben sie den relativen bzw. isolierten Einfluss eines Prädiktors auf das Kriterium an. Die Regressionsgewichte werden von allen Statistikprogrammen ausgegeben. Für jedes einzelne Regressionsgewicht β kann nun ein eigener Standardfehler s_e und ein Konfidenzinter-

vall angegeben werden oder ein Signifikanztest berechnet werden. Dafür wird auch hier aus dem Regressionsgewicht ein *t*-Wert berechnet:

$$t = \frac{\beta}{s_e}$$

Dieser *t*-Wert kann mit $n - 2$ Freiheitsgraden auf Signifikanz geprüft werden. Alle inferenzstatistischen Aussagen würden für jeden einzelnen Prädiktor prüfen, ob er nur in der aktuellen Stichprobe zur Vorhersage des Kriteriums in der Lage war, oder ob er auch in der Population ein verlässlicher Prädiktor sein würde.

Allerdings wird den einzelnen Prädiktoren meist wenig Beachtung geschenkt. Denn ähnlich wie bei den globalen Gütemaßen wollen wir im Prinzip wissen, ob unser *Modell als Ganzes* eine signifikante Varianzsaufklärung leisten kann, die sich auf die Population verallgemeinern lässt. Das Hauptergebnis für dieses gesamte Modell ist der Determinationskoeffizient. Meist wird bei der Multiplen Regression – und das ist gut so – auf die Betrachtung der Signifikanz verzichtet. Stattdessen schaut man sich den Determinationskoeffizienten R^2 an, der eine sehr gut interpretierbare Information darstellt. Dennoch kann man auch für das gesamte Modell einen Signifikanztest berechnen, einen sogenannten *F*-Test. Der *F*-Test ist der Signifikanztest für die Varianzanalyse, die wir uns im übernächsten Kapitel ansehen werden. Wir werden dort diskutieren, wie man ihn zur Prüfung der Signifikanz bei der Multiplen Regression verwendet. (Auch Standardfehler und Konfidenzintervalle kann man für R^2 berechnen; das wird aber so gut wie nie gemacht, da R^2 allein schon eine sehr gut interpretierbare Größe darstellt.)

Literaturempfehlung

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall. (Chapter 12)
Backhaus, K.; Erichson, B.; Plinke, W. und Weiber, R. (2006). *Multivariate Analysemethoden*. Berlin: Springer. (Kapitel 1)

5.3 ALM und Multiple Regression als Grundlage aller Testverfahren

> „…if you were going to a desert island to do psychology research and could take only one computer program with you to do statistical tests, you would want to choose multiple regression" (Aron, Aron & Coups, 2009, S. 612)

Prinzipiell lassen sich alle Fragestellungen mit Hilfe der Regressionsrechnung behandeln. Das einzige, was man dabei tun muss, ist, alle unabhängigen Variablen so zu codieren, dass sie als Prädiktoren in der Regression verwendet werden können. Denn die Regressionsanalyse fordert, dass alle Variablen intervallskaliert sein müssen. Bei vielen Variablen ist das kein Problem, da sie auf Intervallskalenniveau gemessen werden. Schwieriger ist es bei Variablen, die zum Beispiel die unterschiedlichen Gruppen in einem Experiment repräsentieren, etwa die Zahlen 1, 2 und 3 für drei verschiedene Interventionsmethoden, die in einer Studie verglichen werden sollen. Diese Zahlen bilden keine Intervallskala, da es keinen Sinn macht, bei drei Interventionsmethoden von einem inhaltlich gleichen Abstand zu sprechen. Der Unterschied zwischen den drei Methoden besteht nur nominal, nicht aber quantitativ. Man könnte ihnen genauso gut die Zahlen 2, 50 und 900 zuordnen, das würde keinerlei Rolle spielen.

Nominalskalierte Variablen können dennoch mit der Regressionsrechnung behandelt werden, wenn man sie entsprechend codiert. Man spricht dann von einer sogenannten *Dummy-Codierung*, die dafür sorgt, dass die Variable einer Intervallskala entspricht. Einen vereinfachten Fall von Dummy-Codierung haben wir schon einmal erwähnt: den Vergleich von nur zwei Gruppen. Bei einer Variable, die nur zwei Ausprägungen hat – zum Beispiel eine UV, die Kontrollgruppe und Versuchsgruppe codiert – ist es egal, welche beiden Zahlen man zur Codierung verwendet. Denn zwischen zwei Zahlen gibt es immer nur einen Abstand, und damit liegt automatisch Intervallskalenniveau vor (die Forderung der Intervallskala war, dass alle Abstände zwischen Zahlen identisch sind – bei nur einem Abstand liegt daher immer Intervallskalenniveau vor). Eine Variable mit zwei Ausprägungen kann daher problemlos in eine Korrelation oder Regression gegeben werden. Anhand dieses einfachen Beispiels können wir uns auch sehr anschaulich noch einmal die Hauptaussage des ALM verdeutlichen: dass nämlich alle Arten von Fragestellungen als Zusammenhänge aufgefasst werden können. Betrachten wir einen Mittelwertsunterschied zwischen zwei Gruppen A und B (siehe Abbildung 5.2). Hier können wir entweder danach fragen, ob der Mittel-

wertsunterschied signifikant ist oder ob die Steigung der Regressionsgerade, die man durch beide Mittelwerte legt, signifikant ist.

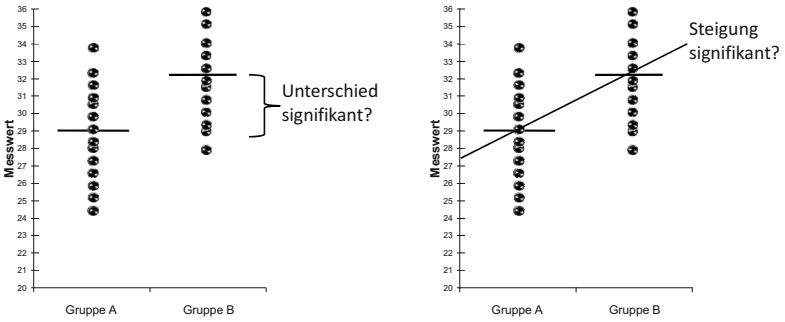

Abbildung 5.2 Äquivalenz von Unterschieden und Zusammenhängen

In dem hier gezeigten Fall wäre es damit egal, ob man einen Signifikanztest für den Mittelwertsunterschied oder einen Signifikanztest für die Steigung der Regressionsgerade (das β-Gewicht) berechnet. Man würde beide Male dasselbe Ergebnis erhalten, das heißt, denselben *p*-Wert. Dieses Prinzip der Äquivalenz von Unterschieden und Zusammenhängen steckt in allen folgenden Testverfahren.

> Das Allgemeine Lineare Modell (ALM) führt alle Testverfahren auf lineare Zusammenhänge zwischen Variablen zurück. Ziel jedes Testverfahrens ist die Prüfung der Varianzaufklärung von Kriteriumsvariablen (abhängigen Variablen).

Nun fragen Sie sich eventuell, warum es neben der Multiplen Regression überhaupt noch andere Testverfahren gibt, wenn doch alle Fragestellungen mit der Regression behandelt werden können. Diese Frage ist schnell beantwortet. Dummy-Codierungen einerseits und die Regressionsrechnung andererseits sind relativ rechenaufwändige Verfahren. Für einige Fragestellungen – wie Mittelwertsunterschiede – ist es schlichtweg nicht nötig, eine Regression zu berechnen. Stattdessen kommt man bei solchen Fragestellungen viel einfacher zu einem Signifikanztestergebnis. Während die Regression das älteste Testverfahren ist, wurden im Lauf der Zeit Testverfahren entwickelt, die einfacher zu berechnen waren, wenn man

es nicht mit intervallskalierten Variablen zu tun hatte. Ronald Fisher, dem wir schon begegnet sind, hat beispielsweise die Varianzanalyse entwickelt – ein Testverfahren, mit dem man mehrere Mittelwerte auf signifikante Unterschiede prüfen kann. Die Varianzanalyse ist – wenn man nur Zettel und Stift zur Verfügung hat – einfacher zu berechnen als eine entsprechende Regression.

Auf diese Weise ist eine Vielzahl von Verfahren für die verschiedensten Fragestellungen entwickelt worden, mit denen wir uns in den nächsten Kapiteln beschäftigen werden. Diese stellen aber alle nichts weiter als Spezialfälle des ALM dar. Da wir heute im Computerzeitalter leben und kaum einen Test per Hand berechnen, kann man argumentieren, dass all diese Testverfahren eigentlich überholt und nicht mehr nötig sind und man stattdessen immer eine Regression berechnen könnte. Das ist zwar richtig, aber die Testverfahren (vor allem *t*-Tests und Varianzanalysen) sind ein so fester Bestandteil des Methodenrepertoires in den Sozialwissenschaften, dass sie so schnell nicht aus den Publikationen verschwinden werden. Behalten Sie aber immer im Hinterkopf, dass die verschiedenen Verfahren nicht unvermittelt nebeneinander stehen, sondern dass die meisten von ihnen aus der Idee des ALM abgeleitet sind – auch wenn man das nicht immer auf den ersten Blick sieht.

Die Mehrzahl aller Testverfahren zur Signifikanzprüfung sind Spezialfälle des ALM.

6 Unterschiede zwischen zwei Gruppen: der *t*-Test

6.1 Das Prinzip des t-Tests

Der *t*-Test geht auf den Engländer William Gosset zurück, der – weil er in einer Brauerei beschäftigt war, deren Mitarbeiter keine Studienergebnisse veröffentlichen durften – unter dem Pseudonym *Student* publizierte. Daher wird die auf ihn zurück gehende *t*-Verteilung auch manchmal Student's *t*-Verteilung genannt. Der *t*-Test ist im Prinzip kein einzelner Test, sondern eine Gruppe von Tests, die für verschiedene Fragestellungen verwendet werden können. Das Prinzip des *t*-Test ist aber immer der Vergleich zwischen zwei Mittelwerten. Dabei kann es sich um Mittelwerte aus unabhängigen oder abhängigen Stichproben handeln oder um einen Mittelwert, der gegen einen theoretisch zu erwartenden Mittelwert getestet wird. Diese Möglichkeiten werden wir uns jetzt anschauen. Außerdem kann der *t*-Test als Testverfahren für andere Kennwerte, die auch einer *t*-Verteilung folgen, verwendet werden. Die beiden wichtigsten haben wir schon kennengelernt: Korrelationskoeffizienten und Regressionsgewichte.

> Der *t*-Test prüft, ob sich zwei Mittelwerte signifikant voneinander unterscheiden.

6.2 t-Test bei zwei unabhängigen Stichproben

Der häufigste Fall für die Verwendung eines *t*-Tests ist der Vergleich zweier Mittelwerte aus unabhängigen Stichproben. Betrachten wir dazu eine Studie, in der untersucht wurde, wie sich der Anteil von Substantiven auf die Verständ-

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_7,

lichkeit von Texten auswirkt. Verglichen werden zwei Texte mit einem Anteil von Substantiven von 30% und 40%. Die Texte werden zwei zufällig gezogenen Gruppen von je fünf Probanden vorgelegt, die die Verständlichkeit auf einer Skala von 1 (gar nicht verständlich) bis 10 (sehr gut verständlich) einschätzen sollen. Wir wollen prüfen, ob die Verständlichkeit bei Texten mit mehr Substantiven höher ausfällt. Die Ergebnisse könnten wie folgt aussehen:

Tabelle 6.1 Beispieldaten für das Textexperiment

	Gruppe 1 (30%)	Gruppe 2 (40%)
	5	5
	6	8
	2	7
	3	9
	7	6
M	4,6	7,0
$\hat{\sigma}^2$	4,3	2,5

Diese Daten könnte man sich zunächst mit einem Diagramm veranschaulichen. Die Mittelwerte zeigen schon einen Unterschied zwischen beiden Gruppen. Der *t*-Test soll nun prüfen, ob dieser Unterschied signifikant ist. Bei der allgemeinen Betrachtung des Signifikanztests haben wir gesehen, dass dabei immer *gegen die Nullhypothese* getestet wird. Wir müssen also den Mittelwertsunterschied unserer Stichprobe gegen den Mittelwertsunterschied testen, den die Nullhypothese für die Population unterstellt. Rechnerisch sieht das folgendermaßen aus:

$$(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)$$

Was ist die Nullhypothese in diesem Fall? Sie behauptet nichts anderes, als dass sich die Mittelwerte in der Population *nicht* unterscheiden sollten. Das heißt, dass

μ_A und μ_B denselben Wert liefern sollten. Damit wäre die Differenz $(\mu_A - \mu_B)$ aber 0 und wir können sie aus der Formel entfernen, die sich dann vereinfacht zu:

$$\bar{x}_A - \bar{x}_B$$

Wie wir schon wissen, ist der Unterschied zweier Mittelwerte umso bedeutsamer, je kleiner die Streuungen sind, mit denen beide Mittelwerte behaftet sind. Das heißt, dass dieser Mittelwertsunterschied nun an den Streuungen beider Mittelwerte relativiert werden muss. Man benutzt dafür den Standardfehler des Mittelwertsunterschiedes. Wenn wir das tun, erhalten wir die Formel für unseren *t*-Test:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{\bar{x}_A - \bar{x}_B}}$$

Der Standardfehler des Mittelwertsunterschiedes im Nenner kann aus den einzelnen geschätzten Populationsvarianzen der beiden Gruppen bestimmt werden, die in Tabelle 6.1 mit angegeben sind:

$$\hat{\sigma}_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{\hat{\sigma}_A^2}{n} + \frac{\hat{\sigma}_B^2}{n}}$$

Diese Formel gilt für gleiche Stichprobengrößen. In unserem Fall ist *n* in beiden Gruppen 5. Der *t*-Wert beträgt für unser Beispiel:

$$t = \frac{4{,}6 - 7{,}0}{\sqrt{\frac{4{,}3}{5} + \frac{2{,}5}{5}}} = \frac{-2{,}4}{1{,}166} = -2{,}06$$

Der so bestimmte *t*-Wert heißt *empirischer t*-Wert. Er wird mit dem *kritischen t*-Wert verglichen, der sich aus der *t*-Verteilung bei einem bestimmten Signifikanzniveau ergibt. Die Freiheitsgrade für den *t*-Test bei unabhängigen Stichproben bestimmen sich wie folgt:

$$df = (n_A - 1) + (n_B - 1)$$

Wir entscheiden uns für ein Signifikanzniveau von 5% und müssen daher bei einer Fläche von 0,95 und bei $df = 8$ in der Tabelle der *t*-Verteilung nachschauen. Wir gehen hier von einem einseitigen Test aus, da wir die Hypothese hatten, dass ein höherer Anteil an Substantiven zu einer besseren Verständlichkeit führt. Der kritische *t*-Wert bei Alpha = 5% beträgt 1,86. Unser empirischer *t*-Wert muss nun

extremer sein als der kritische. Mit extremer ist gemeint, dass er *absolut* größer sein muss, unabhängig vom Vorzeichen, da in der Tabelle nur die positiven *t*-Werte abgetragen sind. Das negative Vorzeichen in unserem berechneten *t*-Wert zeigt ja nur an, dass wir den größeren vom kleineren Mittelwert abgezogen haben. Das hätten wir genauso gut umgekehrt machen können – dafür gibt es keine feste Regel. Da unser empirischer *t*-Wert extremer ist als der kritische, haben wir es mit einem Ergebnis zu tun, das auf dem 5%-Niveau signifikant ist. Ein weiterer Blick in die Tabelle verrät uns übrigens, dass unser Wert auf dem 1%-Niveau nicht signifikant wäre: der kritische *t*-Wert läge dann nämlich bei 2,896.

Hätten wir in unserem Experiment keine gerichtete Hypothese gehabt, hätten wir entsprechend zweiseitig testen müssen und – bei gleichem Signifikanzniveau von 5% – bei einer Fläche von 0,975 nachschauen müssen. Der kritische *t*-Wert wäre dann 2,306 und unser Ergebnis wäre nicht signifikant gewesen. Daran sieht man deutlich, dass gerichtete Hypothesen immer von Vorteil sind.

Während wir hier der Anschaulichkeit halber die Signifikanzprüfung mit Hilfe der Tabelle durchführen, wird diese Aufgabe normalerweise von einem Statistikprogramm übernommen. Wir erhalten dann den genauen *p*-Wert, der dem gefundenen empirischen *t*-Wert entspricht. Dieser *p*-Wert liefert eine genauere Information als die bloße Aussage, dass ein Ergebnis auf einem Niveau von 5% oder 1% signifikant ist. Daher sollte als Ergebnis neben dem *t*-Wert immer der genaue *p*-Wert angegeben werden.

An diesem Beispiel haben wir nun die Prozedur eines Signifikanztests durchlaufen, die sich bei allen Testverfahren wiederholen wird und die die allgemeine Durchführung eines Signifikanztests widerspiegelt. Wir haben eine Nullhypothese formuliert und unser empirisches Ergebnis (den Mittelwertsunterschied) dahingehend getestet, ob er unter der Annahme der Gültigkeit der Nullhypothese wahrscheinlich war oder nicht. Bei einem festgelegten Alpha-Niveau von 5% war unser Ergebnis signifikant und wir verhalten uns nun so, als ob die Alternativhypothese zuträfe. Das heißt, wir schlussfolgern, dass der in unserer Studie gefundene Effekt auf die Population verallgemeinert werden kann. Wir sehen hier auch, dass es sich bei dieser Vorgehensweise um einen *Hybrid* zwischen den Ansätzen von Fisher und Neyman/Pearson handelt, denn wir haben auf die explizite Formulierung einer Alternativhypothese verzichtet. Diese hätten wir allerdings machen können, um den Alpha- und den Beta-Fehler gegeneinander abzuwägen.

Anhand der Formel für den *t*-Test können wir nun auch nachvollziehen, dass diese relativ schnell per Hand ausgerechnet werden kann. Das Ergebnis wäre aber identisch, wenn wir die AV (die Textverständlichkeit) mit der UV (der Gruppenzugehörigkeit, die mit 0 und 1 codiert sein könnte) korreliert hätten und die Korrelation auf Signifikanz getestet hätten. Wir werden gleich noch sehen, dass man aus dem *t*-Wert sehr leicht eine Effektgröße, nämlich einen Korrelationskoeffizienten *r* berechnen kann.

Betrachten wir an dieser Stelle einmal kurz, wie sich das Prinzip der Varianzaufklärung – das wir aus dem ALM abgeleitet hatten – eigentlich im *t*-Test widerspiegelt. Was ist die Varianz, die aufgeklärt werden soll? Das ist die Varianz aller Messwerte aller Personen, unabhängig von ihrer Gruppenzugehörigkeit. Diese Varianz soll erklärt werden. Und es gibt prinzipiell zwei Ursachen für diese Varianz. Die erste – und für uns interessante – liegt darin, dass sich ein Teil der Personen in der einen und ein anderer Teil der Personen in der anderen Gruppe befindet und beide Gruppen ein unterschiedliches Treatment erhalten haben, nämlich die verschiedenen Texte. Optimal wäre es, wenn die Varianz aller Messwerte nur auf dieses Treatment zurückgeht. Die UV (Gruppenzugehörigkeit) würde dann die gesamte Varianz erklären. Es gibt aber eine zweite Ursache für die Varianz der Messwerte, und zwar die Varianz, die auf zufällige Unterschiede zwischen den Personen und auf Messfehler zurückgeht: die Fehlervarianz. Diesen Teil der Varianz können wir nicht erklären. Was wir wissen wollen, ist also, wie das Verhältnis der systematischen erklärten Varianz zur Fehlervarianz ist. Genau das ist das Verhältnis, das wir beim *t*-Test berechnen: wir teilen die systematische Varianz (nämlich den Mittelwertsunterschied) durch die Fehlervarianz (verkörpert durch den Standardfehler). Diesem Prinzip folgen alle Signifikanztests.

6.3 t-Test für abhängige Stichproben

Bei Mittelwertsunterschieden aus abhängigen Stichproben interessieren wir uns rein rechnerisch nicht für die Mittelwerte, die für beide Messzeitpunkte vorliegen, sondern für die Mittelwertsunterschiede, die sich *pro Person* ergeben. Die absolute Größe der Messwerte (also die Rohwerte) und damit auch Unterschiede zwischen den Personen sind also nicht von Bedeutung. Dieses Prinzip ist in Abbildung 6.1 noch einmal dargestellt.

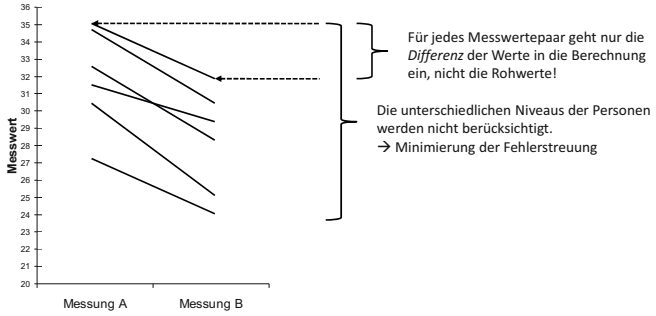

Abbildung 6.1 Mittelwertsunterschiede bei abhängigen Messungen

Wir müssen daher für jede Person ihren Unterschied zwischen der ersten und zweiten Messung berechnen und den Durchschnitt dieser Unterschiede für alle Personen bestimmen. Als Beispiel benutzen wir wieder die Daten aus dem Experiment zur Textverständlichkeit, gehen aber jetzt davon aus, dass dieselben Personen an beiden Bedingungen teilgenommen haben. Es gibt also nur noch fünf Versuchsteilnehmer, die erst den einen und danach den anderen Text bewertet haben (siehe Tabelle 6.2).

Tabelle 6.2 Beispieldaten für das Textexperiment mit abhängigen Messungen

	Messung 1 (30%)	Messung 2 (40%)	Differenz
	5	5	0
	6	8	-2
	2	7	-5
	3	9	-6
	7	6	1
$\bar{X}_{diff}$			-12
$\hat{\sigma}_{diff}$			11,16

Der Mittelwert aller Differenzen beträgt -12 und zeigt einen Unterschied in die richtige Richtung an, wenn wir von der Hypothese ausgehen, dass mehr Substantive die Textverständlichkeit erhöhen. Auch hier müssen wir nun wieder diesen Effekt gegen den Effekt testen, den die Nullhypothese unterstellt:

$$\bar{X}_{diff} - \mu_{diff}$$

Die Nullhypothese sagt aber auch hier, dass es in der Population keine Differenzen gibt, die systematisch von 0 abweichen: $\mu_{diff} = 0$.

Wir können damit diesen Wert wieder aus der Formel streichen, sodass nur $\bar{X}_{diff}$ übrig bleibt. Diese durchschnittlichen Differenzen müssen wieder an der Streuung der Differenzen relativiert werden, und das wird auch hier mit dem Standardfehler getan:

$$\hat{\sigma}_{\bar{X}_{diff}} = \frac{\hat{\sigma}_{diff}}{\sqrt{n}}$$

Die Streuung der Differenzen $\hat{\sigma}_{diff}$ ist in Tabelle 6.2 mit angegeben. Der *t*-Wert berechnet sich nun wie folgt:

$$t = \frac{\bar{X}_{diff}}{\hat{\sigma}_{\bar{X}_{diff}}}$$

Für unser Beispiel ergibt sich damit ein *t*-Wert von:

$$t = \frac{-12}{\frac{11{,}6}{\sqrt{5}}} = -2{,}31$$

Diesen empirischen *t*-Wert vergleichen wir wieder mit dem kritischen *t*-Wert aus der Tabelle, den wir bei abhängigen Stichproben bei

$$df = n - 1$$

Freiheitsgraden nachsehen müssen. Die Stichprobengröße beträgt hier nur noch 5, damit wird es auch schwieriger, ein signifikantes Ergebnis zu erhalten. Bei 4 Freiheitsgraden und einem Signifikanzniveau von 5% liefert die Tabelle einen kritischen *t*-Wert von 2,132. Damit ist unser Ergebnis signifikant und wir können die Nullhypothese verwerfen.

6.4 t-Test bei einer Stichprobe

Den *t*-Test kann man auch verwenden, wenn man eigentlich nur eine Gruppe von Personen untersucht hat und deren Mittelwert gegen einen theoretischen Mittelwert testen möchte. Da es nur eine Stichprobe gibt, die man untersucht, spricht man hierbei meist von einem sogenannten *Einstichprobenfall*. Die zweite „Gruppe" besteht dann gewissermaßen nicht in einer echten Gruppe, sondern in einem Mittelwert, den man als gegeben voraussetzt. Dabei kann es sich um bereits bekannte Mittelwerte handeln, wie den Mittelwert des Intelligenzquotienten, der in der Population immer 100 beträgt. Nun könnten wir beispielsweise herausfinden wollen, ob Psychologiestudierende signifikant intelligenter sind als der Durchschnitt mit eben diesem Mittelwert von 100. Wir erheben den IQ von 50 Psychologiestudierenden und finden einen Mittelwert von $\bar{x} = 112{,}0$ und eine Populationsvarianz der Messwerte von $\hat{\sigma} = 17{,}8$. Der *t*-Test bei einer Stichprobe ergibt sich aus der Differenz des empirischen Wertes $\bar{x}$ und des theoretisch zu erwartenden Wertes für die Population μ, die wiederum anhand des Standardfehlers des Mittelwertes relativiert wird:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

Der Standardfehler berechnet sich – wie Sie sich erinnern – wie folgt:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Für unser Beispiel ergibt sich damit ein *t*-Wert von:

$$t = \frac{112{,}0 - 100}{\frac{17{,}8}{\sqrt{50}}} = 4{,}77$$

Dieser *t*-Wert wird wieder auf Signifikanz geprüft, und zwar mit

$$df = n - 1$$

Freiheitsgraden. 49 Freiheitsgrade finden wir nicht direkt in der Tabelle; das müssen wir aber auch gar nicht, denn wir sehen, dass der Wert 4,77 jeden vorhandenen Wert in der Tabelle überschreitet. Unser Ergebnis ist damit hochsignifikant. Den genauen *p*-Wert würde wieder ein Statistikprogramm liefern.

Auf diese Weise können Mittelwerte aus einer Stichprobe gegen beliebige theoretisch zu erwartende Werte getestet werden. In vielen Fällen beträgt der erwartete Wert einfach 0. Die Formel vereinfacht sich dann zu:

$$t = \frac{\bar{x}}{\hat{\sigma}_{\bar{x}}}$$

6.5 Effektgrößen beim t-Test

Für die Bestimmung von Effektgrößen bei Mittelwertsunterschieden gibt es zwei Möglichkeiten. Die erste besteht in der Berechnung von Abstandsmaßen direkt aus den Rohdaten. Diese Berechnungen haben wir bei der Betrachtung der Effektgrößen bereits behandelt. Die zweite Möglichkeit besteht darin, Effektgrößen aus dem Ergebnis eines Signifikanztests zu bestimmen. Dafür gelten einfache Berechnungsvorschriften, die das Signifikanztestsergebnis stets an der Größe der Stichprobe relativieren. Diese wird manchmal durch die Freiheitsgrade ausgedrückt, die dann jeweils den Freiheitsgraden entsprechen, mit denen auch der *t*-Test berechnet wurde, für den eine Effektgröße bestimmt werden soll.

Zwei unabhängige Stichproben

Für den Unterschied zweier unabhängiger Mittelwerte bieten sich meist die Abstandsmaße *d* und *g* als Effektgrößen an:

$$d = \frac{2t}{\sqrt{df}} \qquad und \qquad g = t\sqrt{\frac{n_A + n_B}{n_A \cdot n_B}}$$

Für unser oben betrachtetes Beispiel des Textexperimentes ergeben sich damit die folgenden Effektgrößen:

$$d = \frac{2 \cdot (-2{,}06)}{\sqrt{8}} = -1{,}46 \quad und \quad g = -2{,}06\sqrt{\frac{5+5}{5 \cdot 5}} = -1{,}30$$

Beide Effektgrößen sind nach den Konventionen sehr groß. Das negative Vorzeichen gibt genau wie beim *t*-Wert lediglich an, dass der größere Mittelwert vom kleineren abgezogen wurde.

Da wir argumentiert hatten, dass man sich den Vergleich zweier unabhängiger Stichproben auch als Korrelation zwischen der AV und der Gruppenzugehörigkeit vorstellen kann, bietet sich hier auch die Berechnung der korrelativen Effektgröße *r* an:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Für unser Beispiel ergibt sich damit eine Effektgröße von:

$$r = \sqrt{\frac{-2{,}06^2}{-2{,}06^2 + 8}} = 0{,}59$$

Dieses *r* ist identisch mit dem Korrelationskoeffizienten, den man auch bei der Berechnung einer Korrelation zwischen AV und Gruppenzugehörigkeit erhalten würde.

Zwei abhängige Stichproben

Bei zwei abhängigen Stichproben erhält man die Abstandsmaße, indem man den *t*-Wert, den man für diese abhängigen Stichproben bestimmt hat, in die folgenden Formeln einsetzt:

$$d = \frac{t}{\sqrt{df}} \qquad und \qquad g = \frac{t}{\sqrt{n}}$$

In unserem Textexperiment mit abhängigen Messungen ergeben sich damit:

$$d = \frac{-2{,}31}{\sqrt{4}} = -1{,}16 \quad und \quad g = \frac{-2{,}31}{\sqrt{5}} = -1{,}03$$

Auch bei abhängigen Stichproben kann man prinzipiell die korrelative Effektgröße *r* berechnen. Allerdings hat sich gezeigt, dass diese Berechnung zu verzerrten (meist zu großen) Ergebnissen führt. Daher sollte man auf diese Art der Berechnung wenn möglich verzichten.

Eine Stichprobe

Wenn ein *t*-Test für eine Stichprobe berechnet wurde – der Einstichprobenfall, bei dem ein empirischer Wert mit einem theoretischen Wert verglichen wird – sind die Formeln zur Bestimmung von Effektgrößen identisch mit denen bei abhängigen Stichproben. Für unseren Vergleich zwischen Psychologiestudierenden und dem Durchschnitt der Bevölkerung hinsichtlich des IQ ergibt sich damit:

$$d = \frac{4{,}77}{\sqrt{49}} = 0{,}68 \quad und \quad g = \frac{4{,}77}{\sqrt{50}} = 0{,}67$$

6.6 Voraussetzungen beim t-Test

Signifikanztests sind je nach der Art und Weise ihrer Berechnung an einige Voraussetzungen geknüpft, die sich direkt aus der allgemeinen Idee von Signifikanztests ergeben und sich meist auf die hypothetische Population beziehen, für die man eine Aussage treffen möchte. Auch wenn man einen *t*-Test berechnet, unterstellt man, dass einige Voraussetzungen erfüllt sind. Bei einem *t*-Test für unabhängige Stichproben geht man davon aus, dass die beiden Stichproben, die man gezogen hat, tatsächlich unabhängig sind, das heißt, dass sich die verschiedenen Personen in den beiden Gruppen nicht systematisch gegenseitig beeinflussen.

Weiterhin unterstellen wir beim *t*-Test, dass die AV immer intervallskaliert ist. Eine weitere Voraussetzung betrifft die Verteilung der Werte der AV. Diese sollte in der Population einer Normalverteilung folgen. Diese Forderung ist leicht nachvollziehbar, denn der gesamte Signifikanztest basiert ja auf Normalverteilungen. Ist die Verteilung der Messwerte schief, würde es kaum Sinn machen, die symmetrische *t*-Verteilung als Prüfverteilung zu verwenden. Und schließlich gibt es noch eine Voraussetzung, die die Varianzen der beiden Stichproben betrifft. Diese sollten möglichst gleich groß sein. Auch diese Forderung sollte uns einleuchten. Denn Mittelwerte mit großen Streuungen sind weniger aussagekräftig als solche mit kleinen Streuungen. Daher wäre es wenig sinnvoll, zwei Mittelwerte miteinander zu vergleichen, die mit völlig unterschiedlichen Streuungen behaftet sind.

Die letztgenannten Forderungen sind schwer nachzuprüfen und werden in aller Regel einfach vorausgesetzt. Das ist deswegen legitim, weil der *t*-Test

ein sogenanntes *robustes* Verfahren ist. Das heißt, er ist gegen Verletzungen dieser Voraussetzungen so unempfindlich, dass er trotzdem sehr gute Ergebnisse liefert. Das ist einer der Gründe, warum der *t*-Test ein sehr häufig verwendetes Verfahren ist. Erst wenn die Voraussetzungen sehr deutlich verletzt sind – vor allem, wenn die Verteilung der Werte der AV deutlich von einer Normalverteilung abweicht – sollte man auf die Berechnung eines *t*-Tests verzichten. Als Alternative muss man dann auf sogenannte non-parametrische Testverfahren zurückgreifen, denen wir uns später noch zuwenden werden.

Literaturempfehlung

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall. (Chapters 7, 8)
Bühner, M. und Ziegler, M. (2009). S*tatistik für Psychologen und Sozialwissenschaftler*. München: Pearson. (Kapitel 5)
Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 13)

7 Unterschiede zwischen mehr als zwei Gruppen: die Varianzanalyse

7.1 Das Prinzip der Varianzanalyse

Der *t*-Test ist ein relativ leicht nachvollziehbares Verfahren, das die Bedeutsamkeit der Differenz zweier Mittelwerte untersucht: der Mittelwertsunterschied wird anhand der Streuungen der Mittelwerte relativiert. Der Anwendungsbereich des *t*-Tests ist aber auf den Vergleich von zwei Mittelwerten beschränkt. Da sich psychologische Fragestellungen oft aber auf mehr als zwei Mittelwerte beziehen, benötigen wir hier ein anderes Verfahren. Dieses Verfahren untersucht nicht mehr nur eine Differenz zwischen zwei Mittelwerten, sondern die Variation mehrerer Mittelwerte. Das Verfahren versucht also, die Varianz von Mittelwerten zu erklären und wird daher als *Varianzanalyse* bezeichnet. Als Abkürzung wird der Begriff *ANOVA* (für Analysis of Variance) verwendet. Die Varianzanalyse wurde von dem viel zitierten Statistiker Ronald Fisher entwickelt und ist das zweifellos bekannteste Signifikanztestverfahren überhaupt. Genau wie der *t*-Test ist es eine Sonderform der Multiplen Regression. Die Besonderheit besteht darin, dass bei der Varianzanalyse die UV nominalskaliert sein darf. Damit stellt sie sozusagen den Königsweg für die Auswertung von Experimenten dar. Denn das Prinzip eines Experimentes ist in aller Regel der kontrollierte Vergleich mehrerer Versuchsgruppen hinsichtlich einer AV. Die UV besteht hier ebenfalls in der Gruppenzugehörigkeit, die jetzt aber – anders als beim *t*-Test – auch drei oder noch mehr Gruppen umfassen kann. Die Varianzanalyse stellt damit auch ein viel umfassenderes Verfahren als der *t*-Test dar. Außerdem – und vielleicht ahnen Sie es schon – ist der *t*-Test selbst wiederum nur eine Sonderform der Varianzanalyse. Hier besteht die Besonderheit in der eben erwähnten Einschränkung, dass der *t*-Test nur zwei Mittelwerte vergleichen kann, während die Varianzana-

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_8,

lyse zwei oder mehr Mittelwerte vergleichen kann. Dabei sind die Berechnungen bei der Varianzanalyse nur wenig komplexer als die beim *t*-Test. Sie folgen allerdings einer etwas anderen Logik, die wir uns gleich anschauen wollen.

Die Varianzanalyse ist im Prinzip dasjenige Verfahren, welches das Anliegen der Psychologie am anschaulichsten verkörpert. Wir sind daran interessiert, menschliches Erleben und Verhalten zu verstehen und zu erklären. Dieser Drang zum Verstehen erwächst direkt aus unserer Alltagspsychologie, aus unseren Fragen über den Menschen. Warum verhält sich der eine so, der andere so? Warum verhält sich jemand in einer bestimmten Situation anders als in einer anderen? Warum ging es mir gestern anders als heute? Wenn wir nach Antworten auf solche Fragen suchen, dann tun wir das – ohne es zu wissen – nach dem Prinzip der Varianzanalyse. Wir versuchen nämlich Ursachen zu finden, die für die Varianz, die wir im Erleben und Verhalten von uns und anderen Personen beobachten, verantwortlich sein könnten. Wenn jemand die Augen zusammenkneift, während wir einen Vortrag halten, überlegen wir, woran das liegen könnte. Und wir ziehen verschiedene Variablen als Ursachen in Betracht. Sprechen wir zu leise? Drücken wir uns unverständlich aus? Oder ist die Schriftgröße unserer Präsentation zu klein? Nun beginnen wir zu experimentieren und sprechen lauter, drücken uns verständlicher aus oder vergrößern die Schrift auf unseren Folien. Nach jeder Änderung würden wir schauen, ob die Zuhörer jetzt einen zufriedeneren Eindruck machen. Wenn ja, dann haben wir die richtige Variable als Ursache identifiziert. Das ist Varianzaufklärung!

Auf diese Weise kann man sich viele alltagspsychologische Fragestellungen als varianzanalytische Fragestellungen veranschaulichen. Es gibt immer eine Vielzahl von Variablen, die mit ihren verschiedenen Ausprägungen als Ursachen und Erklärungen in Frage kommen. Das Ziel ist, die richtige(n) zu finden. Statistisch versuchen wir bei der Varianzanalyse das Gleiche zu tun: die wichtigste Quelle (verursachende Variable) für das Zustandekommen von Varianz zu finden.

Wir werden hier mit einem Beispiel arbeiten, in dem drei Mittelwerte verglichen werden sollen. Prinzipiell ist die Anzahl der Mittelwerte aber nicht begrenzt. Das Prinzip bleibt stets das gleiche. Die drei Mittelwerte stellen die drei Ausprägungen einer UV dar. Und wir gehen davon aus, dass es sich um drei experimentelle Gruppen handelt. Wir wollen die Hypothese untersuchen, dass der Stress (gemessen am Adrenalingehalt im Blut mit einem Index von 1 bis 100) bei der Nutzung verschiedener Verkehrsmittel verschieden stark ist: beim Straßenbahnfahren, beim Autofahren und beim Radfahren. Da wir hier nur eine un-

abhängige Variable (einen Faktor) untersuchen, haben wir es mit einer *einfaktoriellen* Varianzanalyse zu tun. Das ist der einfachste Fall.

7.2 Eine UV: die einfaktorielle ANOVA

Um uns das Prinzip der Varianzanalyse zu veranschaulichen, sehen wir uns an, um welche Varianzen es überhaupt geht. Abbildung 7.1 zeigt, wie sich die Messwerte von 45 Personen verteilen, die zufällig auf die drei Gruppen aufgeteilt wurden.

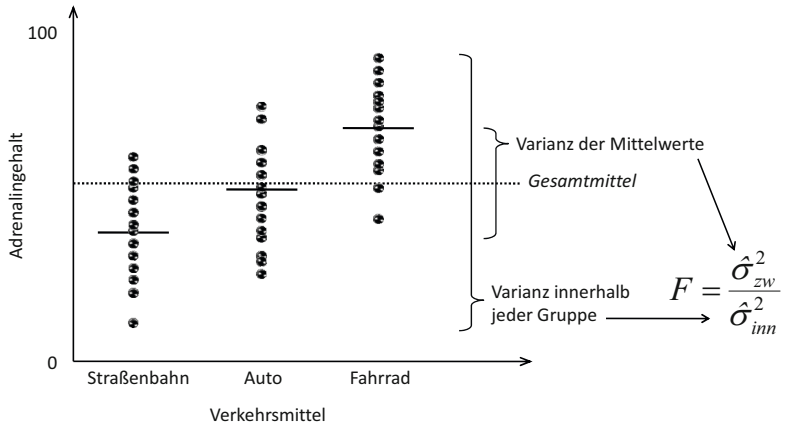

Abbildung 7.1 Das Prinzip der Varianzanalyse

Genau wie beim *t*-Test sind es drei Arten von Varianz, die hier von Interesse sind. Die erste ist die Gesamtvarianz aller 45 Messwerte, also die Varianz der AV. Diese zeigt einfach an, dass sich die Messwerte über alle Personen hinweg unterscheiden. Diese Varianz ist es, die wir aufklären wollen. Welche Erklärungen stehen uns dabei zur Verfügung? Die erste Erklärung steckt in unserer „Manipulation“: wir haben es ja mit Personen zu tun, die verschiedene Verkehrsmittel benutzt haben. Das ist unsere UV, und die sollte den größten Teil der Varianz aufklären. Wo finden wir die Varianz, die auf die UV zurückgeht? Die steckt natürlich in der Differenz der drei Mittelwerte. Die Tatsache, dass die drei Mittelwerte voneinander verschieden sind, zeigt an, dass unsere Manipulation

einen Effekt hatte. Allerdings gibt es hier nun nicht mehr nur *eine* Differenz zwischen zwei Mittelwerten, sondern gleich drei Differenzen: Gruppe 1 und Gruppe 2, Gruppe 1 und Gruppe 3, sowie Gruppe 2 und Gruppe 3. Wir sprechen daher nicht mehr von Mittelwertsdifferenzen, sondern von der Varianz der Mittelwerte: sie unterscheiden sich. Weil diese Varianz durch die unterschiedliche Manipulation zwischen den Gruppen hervorgerufen wird, heißt sie *Varianz zwischen den Gruppen* oder *between*-Varianz. Und weil sie den Teil der Varianz repräsentiert, den wir durch unsere Manipulation erklären können, wird sie auch manchmal *systematische* Varianz genannt.

Die zweite Erklärung für die Gesamtvarianz steckt in der Tatsache, dass die Messwerte auch *innerhalb einer jeden Gruppe* variieren. Diese *Varianz innerhalb der Gruppen* oder *within*-Varianz geht einfach darauf zurück, dass Menschen sich in der Ausprägung von Merkmalen nun mal unterscheiden und daher nicht alle denselben Messwert liefern. Außerdem fallen Messfehler in die within-Varianz. Dieser Teil der Varianz ist zufällig und kann nicht erklärt werden. Er stellt einen Fehler dar, der die Aussagekraft unserer Mittelwerte einschränkt. Diese Varianz wird daher auch *Fehlervarianz* oder *unsystematische Varianz* genannt. Tabelle 7.1 fasst diese Streuungszerlegung der Gesamtstreuung zusammen.

Tabelle 7.1 Streuungszerlegung bei der Varianzanalyse

Gesamtvarianz =	erklärte Varianz	+ nicht erklärte Varianz
	= systematische Varianz = Varianz zwischen den Gruppen = between-Varianz	= unsystematische Varianz = Varianz innerhalb der Gruppen = within-Varianz = Fehlervarianz

Die Fehlervarianz stellt damit – genau wie beim *t*-Test – eine Art Rauschen dar, welches die Bedeutsamkeit der Mittelwertsdifferenzen etwas einschränkt. Daher versucht die Varianzanalyse zu prüfen, ob die gefundenen Mittelwertsdifferenzen deutlich genug sind, dass wir sie auf die Population verallgemeinern können. Mit anderen Worten, sie prüft das Verhältnis von erklärter und nicht erklär-

ter Varianz. Dieses Verhältnis wird durch die Prüfgröße *F* ausgedrückt, deren Formel auch in Abbildung 7.1 dargestellt ist:

$$F = \frac{\hat{\sigma}^2_{zw}}{\hat{\sigma}^2_{inn}}$$

Je kleiner die Fehlervarianz (inn = Varianz innerhalb der Gruppen) bzw. je größer die Varianz der Mittelwerte (zw = Varianz zwischen den Gruppen), desto größer wird der Wert für *F*. Der *F*-Wert ist eine Prüfgröße wie der *t*-Wert. Sie repräsentiert ein Varianz-Verhältnis. Und auch für diese Prüfgröße gibt es eine Prüfverteilung – die *F*-Verteilung – in der alle möglichen solcher Varianz-Verhältnisse verteilt sind, die in der Population unter Annahme der Nullhypothese vorkommen können. Auch hier wird also wieder geprüft, wie wahrscheinlich ein gefundener *F*-Wert unter der Annahme der Nullhypothese war. Bei einem signifikanten Ergebnis kann man die Nullhypothese ablehnen.

> Die Varianzanalyse ist ein Verfahren zum Vergleich von Unterschieden zwischen Gruppen. Dabei wird das Verhältnis zwischen erklärter Varianz (zwischen den Gruppen) und nicht erklärter Varianz (innerhalb der Gruppen) gebildet. Ist die erklärte Varianz in diesem Verhältnis groß genug, führt das zu einem signifikanten Ergebnis.

Aber wie werden die einzelnen Varianzen berechnet? Prinzipiell genauso wie Varianzen immer berechnet werden: man zieht von einem Wert den Mittelwert ab, quadriert diese Differenz, addiert alle quadrierten Differenzen und teilt die Summe durch die Freiheitsgrade. Hier haben wir es aber mit drei Varianzen zu tun, und wir wollen uns anschauen, wie diese im Einzelnen berechnet werden. Bevor wir das tun, sehen wir uns eine Besonderheit bei der Varianzanalyse an. Zur Veranschaulichung der einzelnen Varianzen wird hier meist die Summe der quadrierten Differenzen gar nicht erst durch die Freiheitsgrade geteilt. Man spricht daher von sogenannten *Quadratsummen*. Diese geben ebenfalls die Streuung an, nur eben eine, die noch nicht an der Größe der Stichprobe relativiert wurde. Die Quadratsummen sind uns schon aus der Regressionsrechnung bekannt. Dort wurden jeweils die vertikalen Abstände aller Punkte zur Regressionsgerade quadriert und aufsummiert. Bei der Varianzanalyse wird der Abstand der Punkte zum Mittelwert benutzt. Die Frage bleibt aber die gleiche: wie

gut kann der Mittelwert die Daten repräsentieren, also wie groß sind die Abweichungen der Punkte?

> In der Varianzanalyse werden als Maß für die Streuung meist die Quadratsummen *QS* verwendet. Diese sind noch nicht an der Stichprobengröße relativiert.

Beginnen wir mit der Gesamtvarianz aller Daten. Die Quadratsumme QS_{ges} berechnet sich wie folgt:

$$QS_{ges} = \sum (x - \bar{\bar{x}})^2$$

Wie gewohnt wird hier die Differenz der einzelnen Messwerte x zum gemeinsamen Mittelwert gebildet, quadriert und aufsummiert. Der gemeinsame Mittelwert aller Daten ist in Abbildung 7.1 durch die gestrichelte Linie dargestellt. Um ihn von den einzelnen drei Mittelwerten der Gruppen abzugrenzen, wird er mit zwei Querstrichen versehen ($\bar{\bar{x}}$). Insgesamt gibt es genauso viele Quadrate wie es Messwerte gibt. Die Quadratsumme besteht daher aus n Summanden.

Wie wir eben behauptet haben, kann diese Gesamtvariation in zwei Teile aufgeteilt werden. Der systematische Teil QS_{zw} geht auf die Variation der Mittelwerte zurück, für die wir nun ebenfalls eine Quadratsumme berechnen können:

$$QS_{zw} = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$$

Wie wir hier sehen, wird der Gesamtmittelwert $\bar{\bar{x}}$ von jedem Gruppenmittelwert $\bar{x}_i$ abgezogen. In unserem Fall gibt es drei solcher Gruppenmittelwerte. (Für den Index i kann man sich also die Zahlen 1, 2 oder 3 vorstellen. Demnach gäbe es hier drei Quadrate, die man aufsummieren muss.) Die Quadrate werden jeweils noch mit der Stichprobengröße der jeweiligen Gruppe (n_i) multipliziert.

Wir berechnen hier sozusagen die Streuung einer „Stichprobe“, die nur aus drei Werten besteht, nämlich unseren drei Mittelwerten. Diese Streuung sollte möglichst groß sein, denn wir hoffen ja, dass sich unsere Mittelwerte unterscheiden. Bleibt schließlich noch die Variation der Daten innerhalb einer jeden Gruppe, die die Fehlervarianz darstellt, weil die Verschiedenheit der Messwerte innerhalb der Gruppen für uns nicht systematisch erklärbar ist. Diese Fehlerstreuung berechnet sich schließlich wie folgt:

$$QS_{inn} = \sum (x - \bar{x}_i)^2$$

Das sieht so ähnlich aus wie bei der Gesamtstreuung, nur dass wir hier nicht den Gesamtmittelwert von jedem Messwert abziehen, sondern den Mittelwert der jeweiligen Gruppe, aus der der Datenpunkt stammt. Denn nur so erfahren wir, wie stark die Werte innerhalb einer jeden Gruppe variieren. (Je nach Messwert, um den es gerade geht, nimmt $\bar{x}_i$ also drei verschiedene Werte an. Und Summanden gibt es wieder soviele, wie es einzelne Werte gibt.) Wir können nun die Zerlegung der Gesamtstreuung in ihre beiden Bestandteile folgendermaßen zusammenfassen:

$$QS_{ges} = QS_{zw} + QS_{inn}$$

$$\sum (x - \bar{\bar{x}})^2 = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2 + \sum (x - \bar{x}_i)^2$$

Mit Hilfe der Quadratsummen ist das Zustandekommen der Variationen gut verständlich. Zum Schluss müssen wir aber wieder auf die Varianzen kommen. Dafür teilen wir wie gehabt die Quadratsummen durch entsprechende Freiheitsgrade:

$$F = \frac{\hat{\sigma}^2{}_{zw}}{\hat{\sigma}^2{}_{inn}} = \frac{\frac{QS_{zw}}{df_{zw}}}{\frac{QS_{inn}}{df_{inn}}}$$

Die Freiheitsgrade kann man folgendermaßen bestimmen:

$$df_{zw} = k - 1 \quad und \quad df_{inn} = \sum (n_i - 1) \quad oder\ auch \quad N - k$$

Dabei steht *k* für die Anzahl von Gruppen und *N* für die Gesamtstichprobengröße. Bei df_{inn} wird von jeder Gruppengröße n_i 1 abgezogen. Von diesen Differenzen gibt es dann *k* Stück, die aufsummiert werden. In unserem Beispiel gab es drei mal 15 Personen. Die Freiheitsgrade wären dann:

$$df_{zw} = 3 - 1 = 2 \quad und \quad df_{inn} = (15 - 1) + (15 - 1) + (15 - 1) = 42$$

Die Bestimmung der verschiedenen Varianzen wird man – vor allem weil die Varianzanalyse sehr komplex sein und sehr große Stichproben umfassen kann – immer dem Computer überlassen. Was aber anhand der Formeln deutlich werden sollte, ist das Prinzip, dass die Differenz von Mittelwerten groß und die Streu-

ung der Daten um ihre Mittelwerte herum klein sein sollte. Demnach ergeben sich größere Werte für *F*, wenn die Mittelwerte weiter auseinanderliegen oder aber die Streuungen um die Mittelwerte kleiner sind.

Zur Prüfung des *F*-Wertes auf Signifikanz mit Hilfe der *F*-Tabelle benötigt man die Zähler- und Nennerfreiheitsgrade. In der Tabelle sind die kritischen *F*-Werte für verschiedene Signifikanzniveaus aufgeführt. Der kritische Wert für ein Signifikanzniveau von 5% ist für unser Beispiel mit 2 Zähler- und 42 Nennerfreiheitsgraden 3,23 (diesen Wert müssen wir bei 40 Nennerfreiheitsgraden ablesen, da die Tabelle keinen Wert für 42 Freiheitsgrade enthält – wir können ihn aber annäherungsweise benutzen). Allerdings gilt auch hier, dass diese Art der Signifikanzprüfung mit Hilfe der Tabelle eher Übungszwecken dient. Den genauen *p*-Wert für einen bestimmten *F*-Wert erfahren wir nur von einem Statistikprogramm, und man sollte stets diesen *p*-Wert angeben. Tabelle 7.2 zeigt den typischen Aufbau einer Tabelle, wie sie Statistikprogramme als Ergebnis der Varianzanalyse liefern.

Tabelle 7.2 Aufbau einer Ergebnistabelle bei der Varianzanalyse

Ursprung der Varianz	Quadratsummen (sum of squares, *SS*)	Freiheitsgrade df	geschätzte Varianz (mean squares, *MS*)	*F*-Wert	*p*-Wert
zwischen den Gruppen	…	…	…	…	…
innerhalb der Gruppen	…	…	…		
gesamt	…	…			

F-Test und t-Test

Natürlich kann man eine Varianzanalyse auch dann berechnen, wenn man nur zwei Gruppen untersucht hat (also anstelle des *t*-Tests). Denn auch dann haben

die Mittelwerte eine Varianz (der einfache Mittelwertsunterschied). Beide Analysen kommen daher in diesem Fall zum selben Ergebnis. *F*-Werte und *t*-Werte sind problemlos ineinander überführbar:

$$t = \sqrt{F} \quad bzw. \quad F = t^2$$

Im Unterschied zum *t*-Wert kann *F* aber nie negativ werden, denn es gibt keine negativen Varianzen. Der *F*-Test testet folglich immer einseitig. Die Varianzanalyse kann damit auch die Richtung des Unterschiedes nicht identifizieren. Es spielt nämlich keine Rolle, in welcher Reihenfolge man die Gruppen in ein Diagramm wie in Abbildung 7.1 aufnimmt. (Man hätte die Autofahrer auch in die Mitte setzen können.) Für die Berechnung der Quadratsummen ist diese Reihenfolge unerheblich. Und durch die Quadrierung werden alle Differenzen immer positiv.

An der Beziehung zwischen *F*-Test und *t*-Test wird außerdem deutlich, dass der *t*-Test lediglich ein Spezialfall der Varianzanalyse ist. Nun fragen Sie sich aber eventuell, ob man anstelle eines *F*-Tests auch mehrere *t*-Tests berechnen könnte, wenn man mehr als zwei Gruppen vergleicht. In unserem Beispiel könnten wir ja auch drei *t*-Tests berechnen. Das ist in der Regel aber *nicht* möglich. (Auf eine Alternative kommen wir gleich noch zu sprechen.) Die Erklärung dafür liegt in der Logik des Signifikanztests. Erinnern Sie sich daran, dass eine Irrtumswahrscheinlichkeit von 5% bedeutet, dass Sie nur in 5 von 100 Fällen einen Fehler machen, wenn Sie die Nullhypothese ablehnen. Anders ausgedrückt: Wenn die Nullhypothese stimmt und Sie 100 Signifikanztests berechnen, dann werden im Durchschnitt 5 davon fälschlicherweise signifikant. Wenn Sie also mit denselben Daten einen zweiten Signifikanztest rechnen, dann verdoppelt sich die Wahrscheinlichkeit eines Alpha-Fehlers (denn es kommt erneut zu einer 5%-igen Fehlerwahrscheinlichkeit). Das heißt, die Wahrscheinlichkeit, eines der 5 von 100 falschen Testergebnisse zu erwischen, wird immer größer. Man sagt daher, dass sich der Alpha-Fehler *kumuliert*. Es ist daher nicht zulässig, mehrere Signifikanztests mit denselben Daten zu berechnen.

Einzelvergleiche (Post-hoc Tests)

Kommen wir zurück zum Ergebnis der Varianzanalyse. Bei einem signifikanten Ergebnis wissen wir, dass sich die Mittelwerte der Versuchsgruppen unterschei-

den. Eine Frage bleibt allerdings offen: Wir wissen nicht, *wie* sie sich unterscheiden. Die ANOVA liefert ein sogenanntes *overall*-Ergebnis, das heißt, das Ergebnis gilt für alle Mittelwerte insgesamt. Wir wissen nur, dass diese sich *irgendwie* unterscheiden. In der Regel sind wir aber daran interessiert zu erfahren, zwischen welchen Mittelwerten ein bedeutsamer Unterschied besteht. Wenn in unserem Beispiel aus Abbildung 7.1 die Gruppen 1 und 2 den gleichen Mittelwert gehabt hätten, die Gruppe 3 aber einen sehr viel höheren Mittelwert, hätte das ebenfalls zu einem signifikanten Ergebnis führen können.

Einen ersten Hinweis liefert natürlich zunächst ein Diagramm, das immer *vor* einer jeden Berechnung angeschaut werden sollte. Im Diagramm sind die Mittelwertsunterschiede schon entsprechend zu erkennen. Um im Nachhinein (post hoc) die einzelnen Mittelwertsunterschiede auf Signifikanz zu prüfen, kann man sogenannte *Einzelvergleiche* berechnen. Diese funktionieren im Prinzip wie einzelne *t*-Tests. Allerdings wird hier die eben angesprochene Alpha-Fehler-Kumulation berücksichtigt, indem diese Einzelvergleiche eine sogenannte *Alpha-Korrektur* erhalten. Für eine solche (mathematische) Prozedur gibt es verschiedene Möglichkeiten, und entsprechend gibt es eine Vielzahl von Einzelvergleichstests. Die bekanntesten beiden sind der *Scheffé*-Test und der *Bonferroni*-Test. Anschließend an eine Varianzanalyse gibt es immer so viele Einzelvergleichstests wie es mögliche Vergleiche gibt (in unserem Beispiel waren es drei). Diese Tests liefern als Ergebnis lediglich einen *p*-Wert für jeden einzelnen Mittelwertsunterschied, den man wie jeden anderen *p*-Wert interpretieren kann. So könnten wir beispielsweise mit Hilfe der Einzelvergleiche erfahren, dass sich die Straßenbahnfahrer von den Radfahrern unterscheiden (deren Mittelwerte liegen am weitesten auseinander), dass es aber zwischen Straßenbahnfahrern und Autofahrern sowie zwischen Autofahrern und Radfahrern keine signifikanten Unterschiede gibt. Der signifikante overall-Test der ANOVA wäre damit nur durch diesen einen signifikanten Mittelwertsunterschied zustande gekommen. Es können aber auch zwei oder alle drei Mittelwertsunterschiede signifikant sein.

Die Betrachtung eines Diagramms mit Mittelwerten oder eines Boxplots sollte immer am Anfang der Analyse stehen. Wenn man dort sieht, dass Mittelwertsunterschiede in eine Richtung gehen, die der Hypothese widerspricht, erübrigt sich natürlich die Berechnung von ANOVA bzw. Einzelvergleichen. Nur wenn es keine gerichteten Hypothesen gab und man einfach schauen will, ob es überhaupt irgendwo Unterschiede gibt, kann man in jedem Fall diese Analysen durchführen.

Einzelvergleiche kann man sich als Zusatz bei der Berechnung einer ANOVA von jedem Statistikprogramm ausgeben lassen. Zu bedenken ist, dass mit der Anzahl von Versuchsgruppen auch die Anzahl möglicher Einzelvergleiche immer größer wird. Bei vier Versuchsgruppen gibt es schon sechs mögliche Einzelvergleiche. Diese haben es aufgrund der Alpha-Fehler-Korrektur immer schwerer, signifikant zu werden. Einfach ausgedrückt würde die Korrektur hier dafür sorgen, dass der ursprüngliche Alpha-Fehler durch 6 geteilt wird (zum Beispiel 5% : 6 = 0,8%). Es wird dabei also immer schwieriger, ein signifikantes Ergebnis zu finden. Als Alternative kann man – wenn man gerichtete Hypothesen über die einzelnen Mittelwertsunterschiede hat – sogenannte *Kontraste* bzw. *Kontrastanalysen* berechnen, auf die wir hier nicht näher eingehen wollen.

Literaturempfehlung zu Kontrastanalysen:

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie.* München: Pearson. (Kapitel 16)

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

7.3 Mehr als eine UV: die mehrfaktorielle Varianzanalyse

Im eben diskutierten Beispiel haben wir eine UV mit drei Ausprägungen untersucht. Wir hatten es also mit einem einfaktoriellen Design zu tun und können daher die durchgeführte Varianzanalyse als *einfaktorielle* ANOVA bezeichnen. Wie wir aus den Überlegungen zu den verschiedenen experimentellen Designs wissen, können wir aber auch gleichzeitig eine zweite UV untersuchen, die ebenfalls zwei oder mehr Ausprägungen haben kann. Wir haben es dann mit einem *zweifaktoriellen* Design zu tun. Nehmen wir an, wir hätten bei dem Vergleich der verschiedenen Verkehrsteilnehmer zusätzlich erhoben, ob es Unterschiede im Geschlecht gibt. Wir könnten dabei die Hypothese haben, dass bei Männern die Unterschiede im Stressniveau nicht so groß sind wie bei Frauen. Das Geschlecht ist nun unsere zweite UV mit zwei Ausprägungen. Das bedeutet, dass wir es mit einem 3x2-Design zu tun haben (3 Ausprägungen für die erste UV und 2 Ausprägungen für die zweite UV). Tabelle 7.3 zeigt beispielhafte Daten für dieses Design, und in Abbildung 7.2 sind diese durch Liniendiagramme dargestellt.

Tabelle 7.3 Faktorielles Design in einer zweifaktoriellen ANOVA mit Beispieldaten

		UV A: Verkehrsmittel			
		Straßenbahn	Auto	Fahrrad	
UV B: Geschlecht	Frauen	35	50	65	$\bar{x}_{Frauen}$
	Männer	48	50	49	$\bar{x}_{Männer}$
		$\bar{x}_{Straßenbahn}$	$\bar{x}_{Auto}$	$\bar{x}_{Fahrrad}$	$\bar{\bar{x}}$

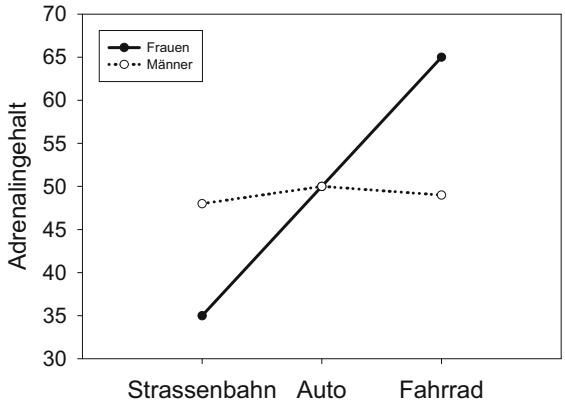

Abbildung 7.2 Liniendiagramm für die Beispieldaten

Der Vorteil der Varianzanalyse ist nun, dass sie beide UVs gemeinsam untersuchen kann. Wie kann man sich das vorstellen? Zunächst wollen wir natürlich für jede der beiden UVs wissen, ob sie zu einem signifikanten Effekt geführt hat oder nicht. An den Berechnungen, die wir hierfür durchführen müssen, ändert sich dabei gar nichts. Wir würden für jede UV einen *F*-Wert berechnen:

$$F_{UVA} = \frac{\hat{\sigma}^2_{zw\,A}}{\hat{\sigma}^2_{inn}} \quad und \quad F_{UVB} = \frac{\hat{\sigma}^2_{zw\,B}}{\hat{\sigma}^2_{inn}}$$

Die beiden UVs sind hier mit A und B bezeichnet. Für die erste UV ergibt sich die Varianz zwischen den Gruppen aus den Mittelwertsunterschieden, die durch die verschiedenen Verkehrsmittel hervorgerufen werden. Diese Varianz wird genauso berechnet wie in der einfaktoriellen ANOVA. Das heißt, für diese Berechnung bleibt die zweite UV unberücksichtigt! Für die zweite UV ergibt sich die Varianz zwischen den Gruppen entsprechend durch die unterschiedlichen Mittelwerte, die durch das Geschlecht zustande kommen. Hier wird wiederum die erste UV außer Acht gelassen. Die Varianz innerhalb der Gruppen wird wie gehabt über *alle* Messwerte berechnet. Diese ist also für beide Analysen dieselbe. Welche *F*-Werte könnten wir erwarten, wenn die Daten so aussehen wie in Abbildung 7.2? Für die verschiedenen Verkehrsmittel ergibt sich wahrscheinlich ein großer Wert für *F*, denn die drei Mittelwerte liegen – wenn man das Geschlecht außer Acht lässt – relativ weit voneinander entfernt. Beim Geschlecht sieht das allerdings anders aus. Um zu erkennen, ob das Geschlecht auch einen Effekt hatte, müssen wir das Verkehrsmittel außer Acht lassen. Das heißt, wir müssen uns vorstellen, wo der Mittelpunkt der gestrichelten und der Mittelpunkt der durchgezogenen Linie liegen. Das wären die Mittelwerte für Frauen und Männer. Offenbar liegen die aber auf der gleichen Höhe. Sie unterscheiden sich also nicht. Das Geschlecht hat damit keinen Effekt auf das Stressniveau: Männer und Frauen empfinden – im Durchschnitt – denselben Stress.

Damit haben wir für jede der beiden UVs den Effekt bestimmt. In der Varianzanalyse werden diese Effekte als *Haupteffekte* bezeichnet. Bei einer zweifaktoriellen ANOVA gibt es zwei mögliche Haupteffekte. Bei mehr als zwei UVs würde man einfach von einer mehrfaktoriellen ANOVA sprechen. Bei einer ANOVA mit drei UVs gäbe es entsprechend drei mögliche Haupteffekte usw. In unserem Beispiel gibt es wahrscheinlich nur einen signifikanten Haupteffekt, nämlich für die UV Verkehrsmittel. Der andere Haupteffekt wäre sicher nicht signifikant.

Welchen Vorteil bringt uns nun diese Varianzanalyse? Die Antwort darauf versteckt sich in dem Diagramm in Abbildung 7.2. Wie wir sehen, sind die Verläufe der Linien für Frauen und Männer völlig verschieden. Bei den Frauen zeigt sich ein Effekt für die Variable Verkehrsmittel, bei den Männern aber nicht. Anders ausgedrückt: der Effekt, den wir für das Verkehrsmittel gefunden haben, kam nur durch die weiblichen Versuchsteilnehmer zustande. Bei Männern spielt das Verkehrsmittel keine Rolle. Das bedeutet also, dass die eine UV (das Geschlecht) darüber entscheidet, ob die andere UV (das Verkehrsmittel) zu einem

Effekt führt oder nicht. Diese gegenseitige Beeinflussung wird *Interaktion* genannt und sie ist der eigentliche Vorteil der mehrfaktoriellen Varianzanalyse. Auch für die Interaktion kann man einen *F*-Wert berechnen, der darüber Auskunft gibt, ob die UVs in ihrer Wirkung signifikant voneinander abhängen:

$$F_{AxB} = \frac{\hat{\sigma}^2_{AxB}}{\hat{\sigma}^2_{inn}}$$

Die Interaktion zweier Variablen wird hier mit AxB gekennzeichnet. Auch hier wird die Varianz, die durch die Interaktion aufgeklärt wird, durch die Fehlervarianz geteilt. Die Fehlervarianz ist wieder dieselbe wie in den obigen Formeln. Die Bestimmung der Varianz für die Interaktion ist denkbar einfach: sie ist diejenige Varianz, die übrig bleibt, wenn man von der Gesamtvarianz alle bereits bekannten Varianzen abzieht:

$$\hat{\sigma}^2_{AxB} = \hat{\sigma}^2_{gesamt} - \hat{\sigma}^2_{zw\,A} - \hat{\sigma}^2_{zw\,B} - \hat{\sigma}^2_{inn}$$

Anders ausgedrückt: Wenn man die aufgeklärten Varianzen aus den beiden Haupteffekten sowie die Fehlervarianz von der Gesamtvarianz abzieht und danach immer noch ein Rest bleibt, dann kann dieser Rest nur noch auf eine Interaktion der UVs zurückgehen. Wir haben es also mit einem Varianzanteil zu tun, der sich nicht durch die einzelnen Haupteffekte allein erklären lässt. Die Interaktion in unserem Beispiel wäre wahrscheinlich signifikant. (Den *p*-Wert für die Interaktion kann man auf dem herkömmlichen Weg in der Tabelle der *F*-Verteilung nachschauen.)

> Bei der mehrfaktoriellen Varianzanalyse kann man Haupteffekte und Interaktionen unterscheiden. Die Haupteffekte prüfen, ob die einzelnen unabhängigen Variablen einen signifikanten Effekt auf die AV ausüben. Die Interaktion prüft, ob sich die unabhängigen Variablen gegenseitig in ihrer Wirkung beeinflussen.

Wie ist eine solche Interaktion zu interpretieren? Allgemein formuliert, bedeutet eine Interaktion, dass sich der Effekt der einen UV in Abhängigkeit der anderen UV verändert. Im Speziellen muss man für eine genaue Interpretation immer das Ergebnisdiagramm betrachten. Die Interpretation für unser Beispiel würde etwa lauten: Wenn jemand eine Frau ist, dann wird sich das Stressniveau bei der Benutzung verschiedener Verkehrsmittel unterscheiden; wenn jemand ein Mann

ist, dann gibt es keine Unterschiede. Da man das Ergebnis in einer solchen Wenn-dann-Bedingung ausdrücken kann, spricht man bei Interaktionen auch manchmal von *bedingten Mittelwertsunterschieden*. Noch einmal: die Interaktion ist nicht durch die Wirkung der einzelnen Faktoren erklärbar, nach denen die Gruppen unterschieden wurden, sondern allein aus deren Kombination.

Wie wir in Abbildung 7.2 sehen können, ist die Interaktion dadurch charakterisiert, dass die beiden Linien für Männer und Frauen *nicht parallel* verlaufen. Das ist immer ein Hinweis auf Interaktionen. Es bedeutet ja nichts anderes, als dass die Mittelwerte in beiden Gruppen einen anderen Verlauf nehmen. Die Linien können sich dabei auch kreuzen. Entsprechend kann es weitere mögliche Linienmuster geben, die auf Interaktionen hindeuten. Es gilt aber immer, dass nicht-parallele Linien eine Interaktion anzeigen. Ob diese signifikant ist, kann natürlich nur der *F*-Test klären. Verschiedene Arten von Interaktionen sind in Abbildung 7.3 dargestellt.

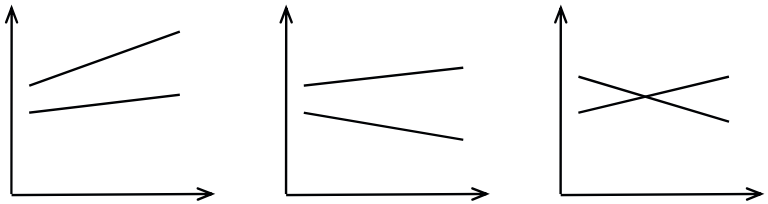

Abbildung 7.3 Schematische Darstellung von Interaktionen in einem 2x2-Design

Das ganz rechte Diagramm zeigt übrigens den Sonderfall, dass es nur eine Interaktion aber keine Haupteffekte gibt. Die Variable, die auf der X-Achse abgetragen ist, führt in beiden Ausprägungen zum selben Mittelwert. Die Variable, die durch die beiden Linien repräsentiert wird, zeigt ebenfalls für beide Ausprägungen denselben Mittelwert. Nur die Kombination beider Variablen führt zu einem Effekt. (Übrigens ist es dem Anwender überlassen, welche UV auf der X-Achse abgetragen und welche durch die verschiedenen Linien dargestellt werden soll.) Solche Haupteffekte und Interaktionen sind tatsächlich gar nicht so leicht zu erkennen und zu durchschauen, wenn man darin wenig Übung hat. Deshalb versuchen Sie jedes Mal, wenn Sie irgendwo Diagramme finden – und das passiert ja häufig in der Psychologie – herauszufinden, was es für Haupteffekte und Interaktionen geben könnte. Üben Sie auch jeweils, sich

die korrekte Interpretation für solche Interaktionen zu überlegen. Übrigens kann man auch nach einer mehrfaktoriellen ANOVA Einzelvergleiche für alle UVs berechnen, die mehr als zwei Ausprägungen haben. Das Vorgehen ändert sich dabei nicht.

Das Aufdecken von Interaktionen ist ein großer Vorteil der Varianzanalyse. Mit Hilfe einer multiplen Regression kann man Interaktionen prinzipiell auch untersuchen; das ist jedoch wesentlich schwieriger, weil man dafür wieder eine Dummy-Codierung für die Interaktion durchführen muss. Bei der Untersuchung von Interaktionen ist allerdings darauf zu achten, dass ihre Anzahl bei mehr als zwei UVs sehr schnell zunimmt. Bei drei UVs gibt es bereits vier mögliche Interaktionen: zwischen Faktor A und B (AxB), zwischen Faktor A und C (AxC), zwischen Faktor B und C (BxC) sowie eine weitere Interaktion zwischen allen drei Faktoren (AxBxC). Eine solche Vielzahl von Interaktionen ist in aller Regel kaum noch sinnvoll zu interpretieren. Tabelle 7.4 zeigt auch für eine zweifaktorielle ANOVA den schematischen Aufbau einer Ergebnistabelle, wie sie Statistikprogramme liefern.

Tabelle 7.4 Ergebnistabelle einer zweifaktoriellen ANOVA

Ursprung der Varianz	Quadratsummen (sum of squares, *SS*)	Freiheitsgrade df	geschätzte Varianz (mean squares, *MS*)	*F*-Wert	*p*-Wert
Faktor A (zwischen)	…	…	…	…	…
Faktor B (zwischen)	…	…	…		
Interaktion (AxB)	…	…	…	…	…
Fehler (innerhalb)	…	…	…		
gesamt	…	…			

7.4 Varianzanalyse mit Messwiederholung

Wir haben uns bisher Designs für unabhängige Messungen angesehen. Aber es kann natürlich – genau wie beim *t*-Test – vorkommen, dass die verschiedenen Versuchsgruppen mit denselben Personen besetzt sind. Wir haben es dann mit abhängigen Messungen bzw. mit echten Messwiederholungen zu tun. Nehmen wir als Beispiel die gleiche Studie wie oben, in der der Effekt verschiedener Verkehrsmittel auf das Stressniveau untersucht wurde. Wir gehen aber jetzt davon aus, dass dieselben Personen alle drei Bedingungen durchlaufen haben. Wenn sich die Mittelwerte der Personen unterscheiden je nachdem, welches Verkehrsmittel sie benutzt haben, können wir auch hier mit Hilfe eines *F*-Tests prüfen, ob diese Unterschiede signifikant sind. Das Prinzip des *F*-Tests ändert sich dabei nicht. Allerdings werden die einzelnen Varianzen anders bezeichnet. Schauen wir uns das genauer an und benutzen dafür Abbildung 7.4.

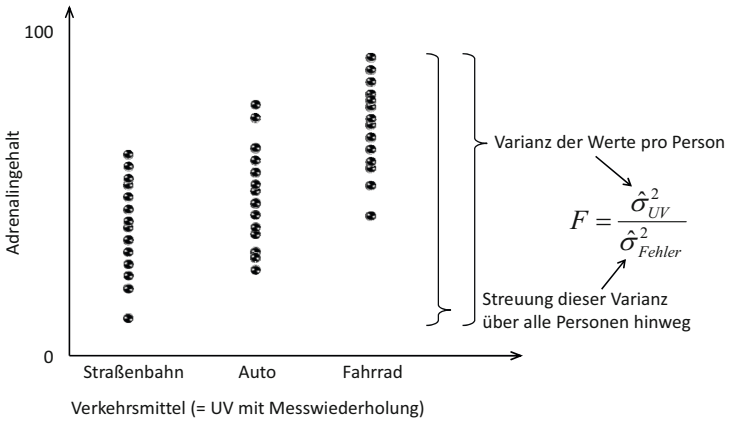

Abbildung 7.4 Beispiel für ein Messwiederholungsdesign

Die aufgeklärte Varianz besteht bei Messwiederholungen darin, dass sich über die Messzeitpunkte hinweg (hier sind das die Verkehrsmittel) für alle Personen Unterschiede ergeben und dass diese Unterschiede im Mittel von 0 verschieden sind. Man kann hier also nicht mehr von einer *Varianz zwischen den Gruppen* sprechen, weil es jetzt nur noch eine Gruppe von Personen gibt. Stattdessen sprechen wir hier einfach von der Varianz, die durch die UV hervorgerufen wird und be-

zeichnen sie mit $\hat{\sigma}^2_{UV}$. Diese systematische Varianz muss nun wieder durch die Fehlervarianz geteilt werden. Worin besteht die Fehlervarianz hier? Wenn die UV (also die Messwiederholung) auf alle Personen den gleichen Effekt ausüben würde, dann müsste für jede Person die gleiche Differenz ihrer Messwerte über die drei Messzeitpunkte hinweg entstehen. Das wird aber in der Regel nicht der Fall sein, weil wir einerseits Messfehler machen und andererseits jede Person individuell anders auf die drei Bedingungen reagiert. Diese beiden Einflüsse können wir aber nicht erklären, daher landen sie in der Fehlervarianz. (Diese wird daher manchmal auch als *Residualvarianz* bezeichnet.) Wir bezeichnen sie hier einfach mit $\hat{\sigma}^2_{Fehler}$. Der *F*-Wert berechnet sich dann wie gehabt:

$$F = \frac{\hat{\sigma}^2{}_{UV}}{\hat{\sigma}^2{}_{Fehler}}$$

Auch dieser *F*-Wert kann mit Hilfe der *F*-Tabelle geprüft werden. Die Freiheitsgrade sind:

$$Zählerfreiheitsgrade\colon df_{UV} = k - 1$$

$$Nennerfreiheitsgrade\colon df_{Fehler} = (k-1)(n-1)$$

Dabei steht k für die Anzahl von Messwiederholungen und n für die Anzahl der Personen.

Auch für Varianzanalysen, die bei Messwiederholungen durchgeführt wurden, kann man Einzelvergleiche berechnen. Das Vorgehen ist dabei identisch zur normalen ANOVA. Auch hier werden die entsprechenden Ergebnisse von jedem Statistikprogramm ausgegeben. Tabelle 7.5 zeigt wieder den schematischen Aufbau einer Ergebnistabelle bei einer ANOVA mit Messwiederholung. (Bei „Personen“ wird hier zusätzlich die Varianz zwischen den Personen angegeben. Diese entspricht der Varianz innerhalb der Gruppen bei der normalen ANOVA. Allerdings spielt diese hier keine Rolle, da das Niveau, auf dem sich die Personen befinden, für die Messwiederholung völlig irrelevant ist.)

Tabelle 7.5 Aufbau einer Ergebnistabelle bei einer ANOVA mit Messwiederholung

Ursprung der Varianz	Quadratsummen (sum of squares, *SS*)	Freiheitsgrade *df*	geschätzte Varianz (mean squares, *MS*)	*F*-Wert	*p*-Wert
Personen	…	…	…		
UV	…	…	…	…	…
Fehler	…	…	…		
gesamt	…	…			

Gemischte Designs

Manchmal kann es vorkommen, dass man in einer mehrfaktoriellen Varianzanalyse abhängige und unabhängige Messungen vermischt hat. Man spricht dann von *gemischten Designs* oder *mixed models*. Beispielsweise hätten wir auch für unser Messwiederholungsdesign zusätzlich untersuchen können, ob sich die Effekte bei Männern und Frauen unterscheiden. Dann hätten wir eine abhängige Messung (Verkehrsmittel) mit einer unabhängigen Messung (Geschlecht) vermischt. Für die Varianzanalyse stellt das kein Problem dar. Auch hier können für beide Haupteffekte und für die Interaktion die *F*-Werte bestimmt werden.

7.5 Effektgrößen bei der Varianzanalyse

Die Effektgröße, die man für Varianzanalysen berechnen kann, nennt sich Eta-Quadrat (η^2). Die Idee hinter dieser Effektgröße ist relativ einfach: sie fragt danach, wie groß der Anteil der durch die UV aufgeklärten Varianz an der Gesamtvarianz ist. Bei der einfaktoriellen ANOVA sieht das folgendermaßen aus:

$$\eta^2 = \frac{QS_{zw}}{QS_{gesamt}}$$

η^2 kann man sich damit in etwa so vorstellen wie den Determinationskoeffizient in der Regressionsanalyse. Multipliziert mit 100 gibt er an, wie viel Prozent an Gesamtvarianz die UV aufklären kann. Einfacher lässt sich η^2 direkt aus dem *F*-Wert berechnen:

$$\eta^2 = \frac{F \cdot df_{zw}}{F \cdot df_{zw} + df_{inn}}$$

Wir können diese Formel auch verallgemeinern, um für alle Arten von Effekten das η^2 zu bestimmen:

$$\eta^2 = \frac{F_{Effekt} \cdot df_{Effekt}}{F_{Effekt} \cdot df_{Effekt} + df_{inn}}$$

Für F_{Effekt} können wir nun jeden beliebigen *F*-Wert (und die dazugehörigen Freiheitsgrade) einsetzen, den wir in den oben genannten Berechnungen erhalten haben – egal, ob es sich um *F*-Werte für Haupteffekte, Interaktionen oder Messwiederholungen handelt. Bei Varianzanalysen mit Messwiederholung würde man dabei für die Fehlerstreuung nicht df_{inn}, sondern entsprechend df_{Fehler} einsetzen.

Bei mehrfaktoriellen Varianzanalysen wird das η^2 auch *partielles* η^2 genannt. Das bedeutet, dass sich das jeweilige η^2 nur auf einen Teil (part) der Varianzaufklärung bezieht. Das η^2 für einen Haupteffekt beschreibt zum Beispiel nur den Varianzanteil, der auf die eine UV zurückgeht. Wenn es in der Analyse aber noch einen anderen Haupteffekt und eine Interaktion gab, dann haben auch diese jeweils ein partielles η^2. Diese Unterscheidung wird deswegen gemacht, weil manchmal auch für die gesamte Analyse ein eigenes η^2 berechnet wird. Dieses würde dann angeben, wie groß der Anteil der aufgeklärten Varianz *durch alle Haupteffekte und Interaktionen zusammen* ist. Statistikprogramme geben in der Regel ein solches globales η^2 für die gesamte Analyse sowie alle partiellen η^2 für die einzelnen Haupteffekte und Interaktionen aus. Auch für die Interpretation von η^2 gibt es Konventionen, die in Tabelle 7.6 dargestellt sind. Wir sehen dabei, dass bereits eine Varianzaufklärung von 14% als großer Effekt angesehen wird.

Tabelle 7.6 Konventionen für die Interpretation von Effektgrößen bei der Varianzanalyse (nach Cohen, 1988)

kleiner Effekt	ab $\eta^2 = 0{,}01$
mittlerer Effekt	ab $\eta^2 = 0{,}06$
großer Effekt	ab $\eta^2 = 0{,}14$

7.6 Voraussetzungen bei der Varianzanalyse

Die Voraussetzungen, die für die Durchführung einer Varianzanalyse erfüllt sein müssen, sind identisch mit denen beim *t*-Test. Zunächst muss die abhängige Variable intervallskaliert sein. Des Weiteren müssen die Messwerte in allen untersuchten Gruppen einer Normalverteilung folgen, um sicherzustellen, dass auch die Werte in der Population normalverteilt sind. Und schließlich sollten sich die Varianzen der Messwerte in allen Gruppen nicht zu stark unterscheiden.

Nur unter diesen Bedingungen ist es gerechtfertigt, die *F*-Verteilung als Prüfverteilung zu benutzen. Allerdings gilt auch für die Varianzanalyse, dass sie relativ robust gegenüber Verletzungen dieser Voraussetzungen ist. Bei sehr starken Verzerrungen – zum Beispiel bei Verteilungen, die eindeutig nicht einer Normalverteilung folgen – sollte man aber auf die Berechnung eines *F*-Wertes verzichten und auf ein nonparametrisches Verfahren zurückgreifen (siehe nächstes Kapitel).

Literaturempfehlung

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall. (Chapters 9, 10)

Backhaus, K.; Erichson, B.; Plinke, W. und Weiber, R. (2006). *Multivariate Analysemethoden*. Berlin: Springer. (Kapitel 2)

Bühner, M. und Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson. (Kapitel 6)

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 14)

7.7 Der F-Test als Signifikanztest bei der Regressionsrechnung

Bevor wir dieses Kapitel abschließen, sehen wir uns noch an, wie man den *F*-Test einsetzen kann, um das Ergebnis einer (Multiplen) Regression auf Signifikanz zu prüfen. Der *F*-Wert wird manchmal als Hauptergebnis für die Güte der Vorhersage für das gesamte Regressionsmodell neben dem Determinationskoeffizienten R^2 angegeben. Während der Determinationskoeffizient das Ausmaß der aufgeklärten Varianz angibt, soll der *F*-Wert Auskunft darüber geben, ob das Ergebnis der Regressionsrechnung auf die Population übertragen werden kann. Das Prinzip des *F*-Tests ändert sich dabei nicht – auch hier wird wieder die durch den Prädiktor bzw. die Prädiktoren aufgeklärte Varianz ($\hat{\sigma}^2_{Regression}$) durch die Fehlervarianz geteilt:

$$F = \frac{\hat{\sigma}^2_{Regression}}{\hat{\sigma}^2_{Fehler}}$$

Die erklärte Varianz ergibt sich dabei aus der Vorhersage der y-Werte. Die Vorhersage schätzt y-Werte ($\hat{y}_i$), die entsprechend der jeweiligen Ausprägung von x vom Mittelwert aller Daten ($\bar{y}$) verschieden sind:

$$\hat{\sigma}^2_{Regression} = \frac{\sum(\hat{y}_i - \bar{y})^2}{df_{Regression}}$$

Die Fehlervarianz entspricht den mittleren Residuen, also dem nicht erklärten Varianzanteil, der sich aus der Differenz von vorhergesagten ($\hat{y}_i$) und tatsächlichen Werten (y_i) ergibt:

$$\hat{\sigma}^2_{Fehler} = \frac{\sum(y_i - \hat{y}_i)^2}{df_{Fehler}}$$

Auch dieser *F*-Wert kann wie jeder andere *F*-Wert auf Signifikanz geprüft werden. Ein signifikanter Wert gibt dabei an, dass sich die Varianzaufklärung, die mit der Regression erreicht werden kann, auf die Population verallgemeinern lässt und nicht etwa ein zufälliges Ergebnis in der Stichprobe war. Den *F*-Wert kann man übrigens auch direkt aus R^2 berechnen:

$$F = \frac{R^2(n-k-1)}{(1-R^2)k}$$

Dabei steht *k* für die Anzahl von Prädiktoren und *n* für die Anzahl der untersuchten Personen.

Der *F*-Wert als Ergebnis einer Regressionsanalyse ist allerdings insgesamt wenig aussagekräftig. Ein signifikantes Ergebnis heißt lediglich, dass die Varianzaufklärung durch das Regressionsmodell größer als 0 ist. Wie groß sie tatsächlich ist, lässt sich aus dem *F*-Wert nicht ablesen. Letztlich ist das aber die interessante Information, die eine Regression liefern soll. Daher ist der Determinationskoeffizient immer das sinnvollere Ergebnis.

Literaturempfehlung:

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 15.3)

8 Testverfahren für nominalskalierte und ordinalskalierte Daten

8.1 Parametrische und nonparametrische Testverfahren

Wir haben bisher über eine Reihe von Analyseverfahren gesprochen, die Zusammenhänge von Variablen untersuchen (wie Korrelation und Regression) oder speziell bei der Untersuchung verschiedener Gruppen angewendet werden (*t*-Tests und Varianzanalyse). All diese Verfahren haben die Besonderheit, dass sie auf bestimmten Annahmen beruhen, die sich auf die Verteilung der Messwerte in der Population beziehen: diese sollten immer einer Normalverteilung folgen. Das ist deswegen notwendig, weil diese Verfahren mit Prüfverteilungen arbeiten, die nur dann exakte Werte liefern, wenn die zugrunde liegende Populationsverteilung normalverteilt ist.

Es gibt nun aber zwei denkbare Fälle, in denen diese Voraussetzung verletzt ist. Der eine Fall tritt dann ein, wenn wir abhängige Variablen erheben, die nicht intervallskaliert sind. Das trifft auf alle Variablen auf Nominal- oder Ordinalskalenniveau zu. Bei solchen Variablen stehen die Zahlen, die man für die Messungen vergibt (zum Beispiel „Rang 1", „Rang 2" usw.) lediglich in einer größer/kleiner-Relation zum Inhalt (wir wissen nicht, in welcher absoluten Größe sich Rang 1 und Rang 2 unterscheiden). Ähnlich bei einer Variable wie Autofarbe: diese können wir zum Beispiel mit den Zahlen 1, 2 und 3 für rot, schwarz und blau codieren, aber wir könnten auch beliebige andere Zahlen verwenden. Die Zahlen können hier nicht als quantitative Unterschiede aufgefasst werden, sondern sie sind nur Symbole zur Unterscheidung der Messwerte.

Der zweite Fall, bei dem die Voraussetzung normalverteilter Messwerte verletzt ist, tritt dann ein, wenn intervallskalierte Messwerte eine schiefe Verteilung ergeben. Wir könnten also beispielsweise eine Fragestellung mit Hilfe

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4_9,

eines *t*-Tests untersuchen wollen und stellen bei der Betrachtung der Verteilung fest, dass diese keiner Normalverteilung folgt. Decken- und Bodeneffekte können etwa dazu führen, dass die Verteilung an den Rand des Wertebereiches gedrückt wird und daher linksschief oder rechtsschief ist. Wenn solche schiefen Verteilungen dann zu stark von einer Normalverteilung abweichen, sollte man auf die Berechnung von *t*-Tests oder Varianzanalysen verzichten. Sehen wir uns ein Beispiel für eine schiefe Verteilung an. In einer Studie wurde gefragt, wie stark sich ein Seitensprung des Partners auf die Entscheidung auswirken würde, sich zu trennen. Die Befragten sollten ihre Antwort auf einer Skala von 1 (geringer Einfluss) bis 5 (starker Einfluss) angeben. Wie zu erwarten, werden hier die meisten Personen einen Wert nahe 5 angeben und die Verteilung wäre gegen den rechten Rand des Wertebereiches gedrückt (siehe Abbildung 8.1). Diese Verteilung ist linksschief bzw. rechtssteil.

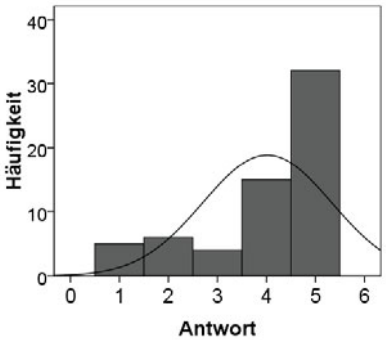

Abbildung 8.1 Linksschiefe Verteilung mit eingezeichneter Normalverteilung

In das Diagramm ist gleichzeitig eine Normalverteilungskurve gelegt, die ihren Gipfel genau über dem Mittelwert der Daten hat. Diese Kurve zeigt an, wie die Daten eigentlich verteilt sein müssten. Und wie wir sehen, kann man bei diesen Daten nicht mehr von einer Normalverteilung sprechen. Stellen wir uns vor, wir wollten für diese Variable zwei Gruppen vergleichen – Männer und Frauen – und die Verteilung würde in beiden Gruppen so schief aussehen. Die Mittelwerte könnten sich nun unterscheiden und wir wollten einen *t*-Test berechnen. Dass das hier keinen Sinn macht, liegt nun auf der Hand, denn der *t*-Test vergleicht die beiden Mittelwerte unter der Voraussetzung, dass die Messwerte in

beiden Gruppen normalverteilt sind. Das Diagramm zeigt aber, dass die Betrachtung des Mittelwertes hier überhaupt keinen Sinn macht. Der Mittelwert liegt bei ca. 4, aber er repräsentiert die Daten der Verteilung alles andere als gut. (Der Mittelwert repräsentiert die Daten nur dann gut, wenn die Verteilung zu beiden Seiten des Mittelwertes symmetrisch abflacht!) Solche Mittelwerte mit *t*-Tests oder Varianzanalysen zu vergleichen, würde zu völlig verzerrten Ergebnissen führen.

Nun ist es nicht immer so leicht zu erkennen, ob eine Verteilung „zu stark" von einer Normalverteilung abweicht. Daher kann man hierfür einen eigenen Signifikanztest berechnen, der Verteilungen dahingehend prüfen kann, ob sie einer Normalverteilung folgen. Dieser Test heißt *Kolmogorov-Smirnov-Test* und sollte ein nicht-signifikantes Ergebnis liefern, da sich die gefundene Verteilung nicht signifikant von einer Normalverteilung unterscheiden sollte. Ein signifikantes Ergebnis in diesem Test würde also anzeigen, dass man die Daten nicht mehr mit Hilfe von *t*-Tests oder Varianzanalysen untersuchen kann.

Was ist nun zu tun, wenn man es entweder mit nominal- bzw. ordinalskalierten Daten oder mit nicht-normalverteilten Intervall-Daten zu tun hat? Die Lösung besteht in der Anwendung von Testverfahren, die keine Annahmen über die Verteilungen der Werte in der Population machen. Solche Verfahren werden daher *verteilungsfreie* Verfahren genannt. Bei diesen Verfahren ist es egal, wie bestimmte Werte in der Population verteilt sind. Mit solchen Werten können dabei normale Messwerte aber auch Mittelwerte und Mittelwertsunterschiede gemeint sein. Grob gesagt kann man einfach von *Parametern* sprechen. Da die verteilungsfreien Verfahren keine Annahmen über die Verteilung dieser Parameter machen, werden sie auch als *non-parametrische* Tests bezeichnet. Damit grenzen sie sich von den parametrischen Tests ab, die wir bisher kennengelernt haben und die alle von einer Normalverteilung der Parameter ausgegangen sind.

> Parametrische Testverfahren setzen eine bestimmte Verteilung – meist eine Normalverteilung – von Populationswerten oder -parametern voraus. Nonparametrische oder verteilungsfreie Verfahren machen keine Annahmen über die Verteilung dieser Werte oder Parameter.

Der Vorteil der non-parametrischen Verfahren liegt neben der Verteilungsfreiheit auch darin, dass man mit ihnen sehr kleine Stichproben untersuchen kann. Prinzi-

piell kann man alle Fragestellungen, die man mit parametrischen Testverfahren untersuchen könnte, auch mit non-parametrischen Verfahren untersuchen. Das macht allerdings wegen eines großen Nachteils der non-parametrischen Verfahren keinen Sinn. Dieser Nachteil besteht darin, dass diese Verfahren in der Regel viel schwerer signifikant werden. Das bedeutet, dass Effekte, die es in der Population möglicherweise gibt, viel schwerer entdeckt werden können (wobei mit entdecken das Finden eines signifikanten Ergebnisses gemeint ist). Man spricht hier auch von einer geringeren Teststärke (oder Power). Die geringere Teststärke ist der wichtigste Grund, warum man wann immer möglich versucht, parametrische Verfahren zu verwenden.

Wir werden uns in diesem Kapitel mit den verschiedenen non-parametrischen Verfahren im Überblick beschäftigen. Da diese Verfahren im Vergleich zu den parametrischen Verfahren sehr selten zum Einsatz kommen, werden wir hier nicht so sehr in die Tiefe gehen und uns stattdessen eher auf die grundlegenden Prinzipien dieser Tests konzentrieren. Beginnen wir mit den Testverfahren für ordinalskalierte Daten.

8.2 Testverfahren zur Analyse ordinalskalierter Daten

Wie Sie aus der Betrachtung der verschiedenen Skalenniveaus wissen, liegen ordinalskalierte Daten in der Regel als Ränge vor. Das heißt, unterschiedliche Messergebnisse – also unterschiedliche Ränge – repräsentieren lediglich eine Reihenfolge. Sie lassen größer-kleiner-Vergleiche oder besser-schlechter-Vergleiche zu, aber sie sagen nichts über die absolute inhaltliche Differenz zwischen zwei Rängen. Wenn wir – wie oben beschrieben – eigentlich intervallskalierte Messwerte haben, aber feststellen, dass diese nicht normalverteilt sind, dann werden diese Messwerte ebenfalls als Ränge behandelt. Damit umgeht man praktisch die Normalverteilungsannahme und testet zum Beispiel Mittelwertsunterschiede mit Hilfe von Tests für Rangdaten. Für Messwerte wie in Abbildung 8.1 würde das bedeuten, dass sie nur noch eine Rangfolge wiedergeben, dass aber die Abstände zwischen den Zahlen (1, 2, 3, 4 und 5) nicht mehr als inhaltlich gleich groß angenommen werden.

Egal ob wir tatsächliche Rangdaten erhoben haben oder aber eigentlich intervallskalierte Daten wie Rangdaten behandeln wollen – die Testverfahren sind hier die gleichen. Zu jedem Test, den wir bisher kennengelernt haben, gibt

es dabei eine non-parametrische Entsprechung. Alle Verfahren für ordinalskalierte Daten sind in Abbildung 8.2 dargestellt.

Non-parametrische Testverfahren für Ordinaldaten können für Zusammenhangs- und Unterschiedsfragestellungen durchgeführt werden. Sie haben alle eine parametrische Entsprechung. Der Unterschied besteht darin, dass die Messwerte als Ränge behandelt werden.

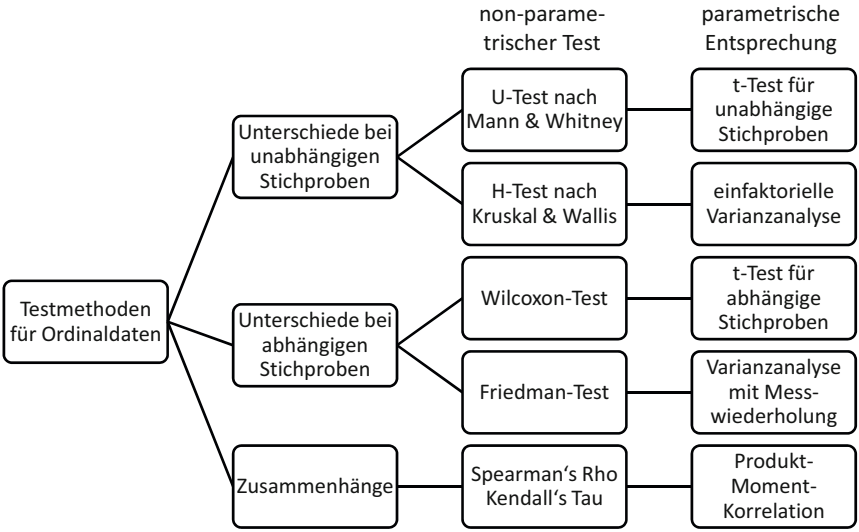

Abbildung 8.2 Überblick über non-parametrische Verfahren für Ordinaldaten

Test für Unterschiede bei zwei unabhängigen Stichproben

Im einfachsten Fall haben wir es mit zwei unabhängigen Messungen zu tun. Die parametrische Entsprechung wäre ein *t*-Test, der die beiden Mittelwerte vergleichen würde. Der non-parametrische Test kann nun nicht mehr mit den Mittelwerten arbeiten. Wie Sie aus der Betrachtung der verschiedenen Skalenniveaus wissen, sind Mittelwerte erst ab Intervallskalenniveau berechenbar bzw. interpretierbar. Tests für Ordinaldaten benutzen daher anstelle des Mittelwertes ein anderes Lagemaß: den Median. Der Median ist hier nichts weiter als der

mittlere Rang in einer Reihe von Rängen. Das Ziel des Tests ist schließlich ein Vergleich der mittleren Ränge für beide Gruppen.

Der Test, der zwei unabhängige Stichproben vergleicht, heißt *U-Test* nach Mann und Whitney. Die Prüfgröße *U*, die für diesen Test bestimmt werden muss, beschreibt den Abstand zwischen den beiden mittleren Rängen. Der *U*-Wert kann wieder mit Hilfe einer Tabelle oder mit Hilfe eines Statistikprogramms auf Signifikanz geprüft werden. Bei großen Stichproben hat sich gezeigt, dass der *U*-Wert annähernd normalverteilt ist. Daher kann man ihn in einen *z*-Wert umrechnen und diesen auf dem herkömmlichen Weg mit Hilfe der *z*-Verteilung auf Signifikanz prüfen. In jedem Fall liefert das Programm auch hier einen genauen *p*-Wert, den man als Ergebnis angeben sollte.

Test für Unterschiede bei mehr als zwei unabhängigen Stichproben

Das Vorgehen zur Untersuchung von mehr als zwei unabhängigen Stichproben ist prinzipiell identisch mit dem eben beschriebenen Vorgehen bei zwei Stichproben. Auch hier wird für jede Gruppe der mittlere Rang bestimmt und anschließend geprüft, ob sich diese mittleren Ränge signifikant voneinander unterscheiden. Der Test, der hier berechnet wird, heißt *H-Test* nach Kruskal und Wallis. Auch hier gilt, dass die Prüfgröße *H* bei großen Stichproben einer bestimmten Verteilung folgt, die sich *Chi-Quadrat-Verteilung* nennt. Diese Verteilung und den dazugehörigen *Chi-Quadrat-Test* werden wir uns später noch genauer ansehen. Dieser Test wird in der Regel als Signifikanztest für den *H*-Test verwendet. Der *H*-Test stellt das non-parametrische Äquivalent zur einfaktoriellen Varianzanalyse dar.

Test für Unterschiede bei zwei abhängigen Stichproben

Beim *t*-Test für abhängige Stichproben hatten wir gesehen, dass hier nicht die Unterschiede zwischen den einzelnen Personen von Interesse sind, sondern die Differenzen der Messwerte, die sich innerhalb von Personen ergeben. Das gleiche Prinzip wird auch für abhängige Stichproben mit Ordinaldaten verwendet. Nur dass hier wiederum nicht die absoluten Differenzen verwendet werden, sondern die Ränge dieser Differenzen. Der Test zur Untersuchung zweier ab-

hängiger Messungen heißt *Wilcoxon-Test* oder auch *Vorzeichenrangtest*. Seine Prüfgröße heißt *T*. Statistikprogramme berechnen zur Signifikanzprüfung beim Wilcoxon-Test ebenfalls einen *z*-Wert, und die Analyse liefert einen *p*-Wert, der als Ergebnis berichtet werden sollte.

Test für Unterschiede bei mehr als zwei abhängigen Stichproben

Zum Vergleich von mehr als zwei abhängigen Messungen würden wir im Normalfall eine Varianzanalyse mit Messwiederholung berechnen. Die non-parametrische Entsprechung ist der *Friedman-Test* bzw. die *Rangvarianzanalyse*. Das Vorgehen ist hier analog zum eben dargestellten Vorgehen: für alle Messwertdifferenzen über die verschiedenen Messzeitpunkte hinweg werden Ränge vergeben und diese anschließend auf Signifikanz geprüft. Die Prüfgröße des Friedman-Tests ist wiederum das Chi-Quadrat, das von Statistikprogrammen mit einem entsprechenden *p*-Wert ausgegeben wird.

Tests für Zusammenhänge: Rangkorrelationen

Wenn Messwerte nicht intervallskaliert sind oder keiner Normalverteilung folgen, lassen sich natürlich auch keine herkömmlichen Korrelationen zwischen Variablen berechnen. Daher gibt es auch für diesen Fall non-parametrische Entsprechungen – die sogenannten *Rangkorrelationen*. Die Rangkorrelationen korrelieren nicht die Rohwerte, sondern wiederum die den Rohwerten zugewiesenen Ränge. Die einzige Bedingung ist dabei, dass die Beziehung beider Variablen *monoton steigt*. Das bedeutet, dass die Linie, die den Zusammenhang beschreibt, zwar keine Gerade sein muss, dass sie aber nicht ihre Richtung ändern darf. In Abbildung 8.3 sind einige Verläufe solcher Linien dargestellt. Das ganz rechte Diagramm zeigt einen Zusammenhang, der erst steigt und dann wieder fällt. Dieser Richtungswechsel stellt einen nicht-monotonen Zusammenhang dar. Alle anderen Zusammenhänge folgen einer monotonen Steigung.

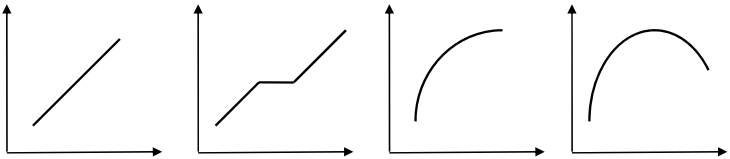

Abbildung 8.3 Beispiele für monotone und nicht-monotone Zusammenhänge

Wenn man die Messwerte, die korreliert werden sollen, nicht bereits als Ränge vorliegen hat, so müssen sie erst in Ränge umgewandelt werden. Dabei wird jedem Messwert abhängig von seiner Größe ein Rang zugewiesen. Bei der Korrelation ist allerdings darauf zu achten, dass die Messwerte beider Variablen einzeln in eine Rangreihe gebracht werden müssen. Man erstellt also eine Rangreihe für die Messwerte von X und eine Rangreihe für die Messwerte von Y. Beide Rangreihen werden anschließend miteinander korreliert. Allerdings kann man sich das Erstellen dieser Rangreihen sparen, wenn man ein Statistikprogramm benutzt, da das diese Umwandlung selbst vornimmt.

Es gibt zwei mögliche Verfahren für die Berechnung einer Rangkorrelation. Die erste Möglichkeit ist die Berechnung einer Prüfgröße, die sich *Spearman's Rho* (ρ) nennt. Rho kann man anwenden, wenn man es eigentlich mit intervallskalierten Daten zu tun hatte, diese aber aufgrund verletzter Voraussetzungen nicht mit einer Pearson-Korrelation untersuchen kann. Wenn man echte Rangdaten erhoben hat (zum Beispiel bei Ergebnisses aus Rankings oder beim Sport), kann man Rho nicht berechnen. In diesem Fall kann man als Alternative eine andere Prüfgröße berechnen, die sich *Kendall's Tau* (τ) nennt.

Rho und Tau können Werte von -1 bis 1 annehmen und sind so zu interpretieren wie die herkömmliche Pearson-Korrelation *r*. Rho folgt dabei einer *t*-Verteilung und wird daher mit einem normalen *t*-Test auf Signifikanz geprüft. Tau folgt einer *z*-Verteilung und kann entsprechend mit einem *z*-Test geprüft werden. Man sollte allerdings beachten, dass man mit den Rangkorrelationen immer weniger Informationen in den Daten berücksichtigen kann als mit einer herkömmlichen Korrelation. Die Rangkorrelationen liefern daher immer kleinere Werte.

Effektgrößen für Ordinaldaten

Für alle parametrischen Testverfahren haben wir Effektgrößen berechnen können, die den gefundenen Effekt an der Streuung relativieren und somit vergleichbare Maße für diesen Effekt darstellen. Effektgrößen für Unterschiedsfragestellungen sind bei Ordinaldaten leider nicht mehr bestimmbar, da wir hier keine Informationen über Mittelwerte erhalten, aus denen sich Effektgrößen ableiten ließen. Wie wir wissen, stecken in diesen Effektgrößen immer Informationen über Standardabweichungseinheiten. Wenn wir es aber mit Rängen zu tun haben, können wir offensichtlich keine Angaben über Standardabweichungseinheiten machen. Bei Unterschiedsfragestellungen muss man sich also – neben der deskriptiven Betrachtung der Rang-Differenzen – auf das Ergebnis der Signifikanztests berufen.

Bei Zusammenhängen kann man die berechneten Rangkorrelationen wie Effektgrößen betrachten und sie im Prinzip so wie die herkömmliche Pearson-Korrelation interpretieren. Wobei man immer beachten sollte, dass die Rangkorrelationen – wie wir eben gesehen haben – in der Regel zu kleineren Werten führen.

8.3 Testverfahren zur Analyse nominalskalierter Daten

Die non-parametrischen Verfahren, die wir zur Analyse ordinalskalierter Daten kennengelernt haben, waren alle direkt vergleichbar mit parametrischen Verfahren zur Analyse von Unterschieds- oder Zusammenhangsfragestellungen. Das liegt daran, dass sich ordinalskalierte Daten und intervallskalierte Daten relativ ähnlich sind. Ihr Unterschied besteht hauptsächlich darin, dass die ordinalskalierten Messwerte zwar immer noch der Größe nach geordnet werden können, dass man aber nicht mehr von gleichen inhaltlichen Abständen zwischen den einzelnen Ausprägungen ausgehen kann.

Bei Daten auf Nominalskalenniveau sieht das ganz anders aus. Hier kann man die Messwerte in keine sinnvolle Rangreihe mehr bringen, da sie mit Zahlen im herkömmlichen Sinne gar nichts zu tun haben. Wenn wir für Messwerte auf Nominalskalenniveau Zahlen verwenden, dann nur, um sie inhaltlich unterscheidbar zu machen. Wir können mit diesen Zahlen aber keinerlei Berechnungen anstellen. Damit lassen sich für Nominaldaten auch keine Aussagen über Lagemaße machen. Wenn wir Personen beispielsweise nach ihrer Lieblings-

musik fragen, dann können wir Messwerte wie „Rock“, „Rap“ oder „Klassik“ erhalten. Diese sind offensichtlich nicht in eine Rangreihe zu bringen. Es wäre uns freigestellt, in einem Diagramm eine beliebige Reihenfolge dieser Musikstile zu wählen, um die Ergebnisse darzustellen.

Bei einer Variable auf Intervallskalenniveau wären wir gezwungen, die möglichen Messwerte auf der X-Achse der Größe nach zu ordnen. Nur so können wir für die entstehende Verteilung einen Mittelwert und eine Streuung bestimmen. Bei den Nominaldaten können wir das nun nicht mehr tun. Wir können lediglich prüfen, ob die Werte, die in der Verteilung auftauchen, bestimmten Erwartungen entsprechen. Dafür gibt es mehrere Möglichkeiten, mit denen wir uns jetzt beschäftigen wollen. Dabei können wir entweder eine Variable untersuchen – wie die Variable Musikstil von eben – oder zwei Variablen in ihrer Kombination. Die Variablen können dabei wieder zwei oder mehr Ausprägungen haben.

Anpassungstest bei einer nominalskalierten Variable

Wie eben angedeutet, können wir bei Nominaldaten lediglich prüfen, ob eine empirische Verteilung, die wir in einer Studie gefunden haben, einer theoretisch zu erwartenden Verteilung entspricht. Der Begriff Verteilung bezieht sich hier allerdings nicht auf Stichprobenverteilungen, sondern auf Häufigkeitsverteilungen. Diese kennen wir aus der deskriptiven Statistik, und sie geben auf der Y-Achse jeweils die Anzahl der Personen an, die einen bestimmten Messwert erzielt haben. Und tatsächlich ist das Einzige, was wir mit Nominaldaten tun können, die Häufigkeiten bestimmter Messwerte zu untersuchen. Für Häufigkeitsverteilungen kann man eine ganz bestimmte Form erwarten, und diese wird mit der Form der gefundenen empirischen Verteilung verglichen.

Weil man hierbei untersucht, wie gut eine empirische und eine theoretische Verteilung zusammenpassen, wird der entsprechende Test *Anpassungstest* genannt. Sein alternativer und gebräuchlicherer Name ist *Chi-Quadrat-Test (χ^2-Test).* χ^2 ist eine Prüfgröße, die wiederum einer bestimmten Verteilung folgt, der χ^2-Verteilung. Um eine empirische Verteilung mit einer theoretisch zu erwartenden Verteilung zu vergleichen, werden die beobachteten mit den erwarteten Häufigkeiten verglichen:

$$\chi^2 = \sum \frac{(f_b - f_e)^2}{f_e}$$

f_b steht dabei für die *b*eobachteten Häufigkeiten, also die Häufigkeiten, mit der bestimmte Merkmalsausprägungen vorgekommen sind. f_e steht für die *e*rwarteten Häufigkeiten, die sich aus theoretischen Überlegungen ergeben. Das Summenzeichen deutet an, dass die Differenz von beobachteten und erwarteten Häufigkeiten so oft berechnet werden muss, wie es Merkmalsausprägungen gibt.

Der χ^2-Anpassungstest prüft, ob eine empirische Häufigkeitsverteilung mit einer theoretisch zu erwartenden Häufigkeitsverteilung übereinstimmt.

Die erwarteten Häufigkeiten stellen beim χ^2-Test die Nullhypothese dar, und es wird auch hier gegen die Nullhypothese getestet, indem die Differenz der beobachteten und erwarteten Werte gebildet wird. Je größer diese Abweichung, desto größer der Wert für χ^2. Dieser empirische χ^2-Wert muss extremer sein als der kritische Wert.

Die zu erwartende Verteilung ist in vielen Fällen eine Gleichverteilung. So würde man beispielsweise beim Würfeln erwarten, dass die Zahlen von 1 bis 6 bei einer großen Zahl von Würfen in etwa gleich häufig auftreten. Die zu erwartende Verteilung kann sich aber auch aus theoretischen Überlegungen oder praktischen Erfahrungen ergeben. Wollen wir etwa untersuchen, ob ein bestimmtes Buch – sagen wir ein Liebesroman – signifikant häufiger von Frauen als von Männern gekauft wurde als ein herkömmlicher Liebesroman, dann können wir für die erwarteten Häufigkeiten keine Gleichverteilung voraussetzen. Denn Liebesromane werden generell eher von Frauen als von Männern gekauft. Diese generelle Häufigkeit – etwa 70% zu 30% – müssten wir hier als erwartete Häufigkeit benutzen.

Unabhängigkeitstest bei zwei nominalskalierten Variablen

Der Anpassungstest prüft, wie sich die Verteilung einer einzigen Variable von einer theoretischen Verteilung unterscheidet, die man für diese Variable erwartet. Natürlich kann man aber auch mehr als eine nominalskalierte Variable für eine Stichprobe erheben. Zur Auswertung kann man nun zunächst für jede dieser Variablen einen Anpassungstest machen – nach dem eben vorgestellten Prinzip. Des Weiteren kann man sich aber fragen, ob die Verteilungen der beiden

Variablen irgendeine Beziehung zueinander aufweisen, ob also die Form der einen Häufigkeitsverteilung von der Form der anderen Häufigkeitsverteilung abhängt. Der Test, der das prüfen soll, heißt entsprechend *Unabhängigkeitstest*. Dieser Test ist ebenfalls ein χ^2-Test, da er sich auch auf beobachtete und erwartete Häufigkeiten bezieht. Die beiden Variablen, die man mit dem Unabhängigkeitstest untersuchen möchte, können dabei im Prinzip beliebig viele Ausprägungen haben.

> Der Unabhängigkeitstest prüft, ob die Verteilung der Ausprägungen einer Variable unabhängig von der Verteilung der Ausprägungen einer anderen Variable ist.

Die Kombinationen der Ausprägungen solcher Variablen kann man in einer sogenannten *Kreuztabelle* oder *Kontingenztafel* darstellen.

> Kreuztabellen oder Kontingenztafeln bilden die verschiedenen Kombinationen der Ausprägungen nominalskalierter Variablen ab.

Im häufigsten Fall wird die Abhängigkeit von nur 2 Variablen untersucht. Dabei kann die eine Variable k und die andere Variable l Ausprägungen haben. Der χ^2-Unabhängigkeitstest wird daher auch *k x l - χ^2* (sprich *k mal l Chi-Quadrat*) genannt. Um ihn zu berechnen, werden wieder für jede der k x l Merkmalskombinationen die beobachteten und die erwarteten Häufigkeiten benötigt. Wenn der Test schließlich signifikant ist, dann heißt das, dass die Häufigkeitsverteilung der einen Variable *nicht unabhängig* von der Verteilung der anderen Variable ist. Wo genau die Unterschiede liegen, kann man allerdings nur anhand der Kreuztabelle im Einzelnen erkennen.

Auch beim Unabhängigkeitstest können die erwarteten Häufigkeiten einer Gleichverteilung entsprechen oder aber aus theoretischen Erwartungen resultieren. Sehen wir uns ein Beispiel an, in dem die erwarteten Häufigkeiten zunächst bestimmt werden müssen. Wir wollen gern wissen, ob Frauen mehr Klassik-CDs kaufen als Männer. Dafür machen wir eine kleine Studie und beobachten die Käufe von 60 Personen in einem Musikgeschäft. Die erste Variable wäre demnach das Geschlecht und die zweite Variable die Frage, ob die Person eine Klassik-CD gekauft hat oder nicht (ja/nein). Wir finden die folgenden Ergebnisse:

Tabelle 8.1 Ergebnisse beim CD-Kauf (beobachtete Häufigkeiten)

		Geschlecht		
		männlich	***weiblich***	*Zeilensumme*
Klassik-CD gekauft?	***ja***	9	35	44
	nein	1	15	16
Spaltensumme		10	50	$N = 60$

Was sind nun die erwarteten Häufigkeiten? Wie wir sehen, haben wir in unserer Stichprobe weder gleich viele Männer und Frauen erwischt, noch haben wir gleich viele Klassik-CD-Käufer wie nicht Klassik-CD-Käufer erwischt. (Wir könnten mit Anpassungstests nun erst einmal prüfen, ob diese Verteilungen für das Geschlecht und für den CD-Kauf von einer Gleichverteilung signifikant abweichen. Aber das ist hier nicht unsere Fragestellung.) Wir können also nicht einfach die 60 Personen durch 4 teilen und das Ergebnis in jede Zelle eintragen. Vielmehr müssen wir hier die Häufigkeiten der Gesamtstichprobe berücksichtigen und in die Bestimmung der erwarteten Häufigkeiten einfließen lassen. Anders ausgedrückt: Wir können nicht erwarten, dass von den 44 verkauften Klassik-CDs 22 von Männern und 22 von Frauen gekauft wurden, denn die Stichprobe enthält ja viel weniger Männer als Frauen. Die erwarteten Häufigkeiten können wir bestimmen, indem wir die sogenannten *Zeilensummen* Z und *Spaltensummen* S multiplizieren und durch N teilen:

$$f_e = \frac{Z \cdot S}{N}$$

Diese Summen geben einfach an, wie viele Männer und Frauen und wie viele Klassik-CD- und nicht-Klassik-CD-Käufe es insgesamt gab. Die erwartete Häufigkeit für die Kombination „Mann und Klassik-CD“ wäre damit $\frac{44 \cdot 10}{60} = 7{,}3$. Auf diese Weise können alle erwarteten Häufigkeiten bestimmt werden (siehe Tabelle 8.2).

Tabelle 8.2 Erwartete Häufigkeiten für das CD-Beispiel

		Geschlecht		
		männlich	***weiblich***	Zeilensumme
Klassik-CD gekauft?	***ja***	7,3	36,7	44
	nein	2,7	13,3	16
	Spaltensumme	10	50	N = 60

Wie man erkennen kann, bleiben die Zeilen- und Spaltensummen natürlich gleich. Die beobachteten und erwarteten Häufigkeiten können nun in die Formel des χ^2-Tests eingesetzt werden:

$$\chi^2 = \frac{(9-7{,}3)^2}{7{,}3} + \frac{(35-36{,}7)^2}{36{,}7} + \frac{(1-2{,}7)^2}{2{,}7} + \frac{(15-13{,}3)^2}{13{,}3} = 1{,}76$$

Die Freiheitsgrade betragen $(k-1)(l-1) = (2-1)(2-1) = 1$. Die χ^2-Tabelle liefert bei einem Signifikanzniveau von 5% einen kritischen Wert von 3,84. Unser Ergebnis ist also nicht signifikant und wir können nicht behaupten, dass der Kauf von Klassik-CDs vom Geschlecht abhängt (die Variablen sind also unabhängig).

Dieses gerade vorgestellte Beispiel stellt übrigens den Sonderfall dar, dass beide Variablen nur zwei Ausprägungen haben. Die Kreuztabelle, die dabei entsteht (Tabelle 8.1), besitzt nur vier Felder und wird daher oft als *Vierfeldertafel* bezeichnet.

Unabhängigkeitstest bei Messwiederholungen

Neben den beiden bisher diskutierten Möglichkeiten, dass man eine Variable oder zwei Variablen untersucht, kann der Fall auftreten, dass man *eine Variable zweimal* gemessen hat. Das würde einer klassischen Messwiederholung entsprechen. Hier wäre man daran interessiert zu prüfen, ob sich die Verteilung der Messwerte bei der ersten Messung von der Verteilung der Messwerte bei der

zweiten Messung unterscheidet. Nehmen wir an, wir würden Raucher und eine Kontrollgruppe von Nichtrauchern einem Nichtraucherseminar unterziehen. Den Anteil von Rauchern und Nichtrauchern erheben wir vor und nach dem Seminar. Unsere Hoffnung ist natürlich, dass das Seminar dazu führt, dass es hinterher weniger Raucher als vorher gibt.

Die erhobenen Daten können wir wieder in einer Tabelle darstellen:

Tabelle 8.3 Tabelle für ein Messwiederholungsdesign bei einer nominalskalierten Variable

		Messung 2 (hinterher)	
		Raucher	***Nichtraucher***
Messung 1 (vorher)	***Raucher***	a	b
	Nichtraucher	c	d

Entscheidend an dieser Tabelle sind die Zellen b und c. Diese würden angeben, dass jemand entweder vorher geraucht hat und hinterher nicht mehr (b) oder vorher nicht geraucht hat aber hinterher (c). Bei Personen in den Zellen a und d hätte sich nichts verändert. Der entsprechende χ^2-Test heißt *Mc-Nemar-*χ^2. Der Mc-Nemar-Test hat eine sehr einfache Formel, in die tatsächlich nur die Zellen b und c eingehen:

$$\boldsymbol{\chi^2 = \frac{(b-c)^2}{b+c}}$$

Ein signifikantes Ergebnis würde hier anzeigen, dass das Verhältnis von Rauchern zu Nichtrauchern nach dem Seminar deutlich anders ist als vor dem Seminar. Ob dabei tatsächlich Raucher zu Nichtrauchern geworden sind und nicht etwa umgekehrt, muss man wieder anhand der Tabelle prüfen.

Effektgrößen für Nominaldaten

Die Bestimmung von Effektgrößen für Nominaldaten wirft die gleichen Probleme auf, wie wir sie bei den Effektgrößen für Ordinaldaten schon diskutiert haben. Da wir auch bei Nominaldaten keine Mittelwerte zur Verfügung haben, können wir keinerlei Abstandsmaße berechnen. Wir können aber für alle Arten von χ^2-Tests eine korrelative Effektgröße bestimmen, denn prinzipiell prüfen alle χ^2-Tests den Zusammenhang von Verteilungen. Die Effektgröße, die hier berechnet werden kann, heißt w und kann für alle χ^2-Tests nach der folgenden Formel berechnet werden:

$$w = \sqrt{\frac{\chi^2}{N}}$$

N bezieht sich dabei immer auf die Größe der Gesamtstichprobe. w kann genauso wie die herkömmliche Pearson-Korrelation interpretiert werden. Beim Spezialfall der Vierfeldertafel wird auch manchmal der sogenannte *Phi-Koeffizient* (ϕ) als Effektgröße verwendet, dessen Ergebnis aber mit w identisch ist.

Literaturempfehlung zu allen Unterschieds-Fragestellungen bei parametrischen und nonparametrischen Tests:

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall. (Chapter 14)

Bühner, M. und Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson. (Kapitel 5 und 6)

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson. (Kapitel 17 und 18)

Literatur

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*. Upper Saddle River: Prentice Hall.

Backhaus, K.; Erichson, B.; Plinke, W. und Weiber, R. (2006). *Multivariate Analysemethoden*. Berlin: Springer.

Bortz, J. und Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bühner, M. und Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pospeschill, M. (2006). *Statistische Methoden*. München: Spektrum.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

Sedlmeier, P. und Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson.

Sedlmeier, P. & Köhlers, D. (2001). *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln*. Braunschweig: Westermann.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research – online, 1*. [Internet: http://www.mpr-online.de/].

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4,

Glossar

abhängige Messungen; 39-41, 45 f., 78, 108, 112, 145: Abhängige Messungen entstehen durch Messwiederholung an derselben Stichprobe oder durch gepaarte (gematchte) Stichproben. Sie zeichnen sich durch eine kleinere Fehlervarianz auf, da Störvariablen wesentlich besser kontrolliert werden können.

Allgemeines Lineares Modell (ALM); 86-89, 99-101: Das ALM spannt sich wie eine Art mathematischer Schirm über fast alle Arten von Signifikanztests und vereint die in verschiedenen Tests auftauchenden Berechnungen. Es führt alle Testverfahren auf lineare Zusammenhänge zwischen Variablen zurück, welche sich durch eine lineare Regressionsgerade beschreiben lassen. Die Gleichung des ALM sagt aus, dass ein konkreter Wert einer Person aus einer Regressionskonstante, einer Reihe von Prädiktoren und einem Fehler vorhersagbar ist.

Alpha-Fehler (Fehler erster Art); 59, 62, 65 ff.: Der Alpha-Fehler ist die Wahrscheinlichkeit, mit der wir beim Signifikanztest aufgrund eines Stichprobenergebnisses fälschlicherweise die Alternativhypothese annehmen und die Nullhypothese verwerfen (obwohl diese eigentlich in der Population gilt).

Alternativhypothese; 56 f., 63 ff., 68f., 72: Die Alternativhypothese (auch als H1 bezeichnet) als Teil des Signifikanztests beschreibt den Effekt (Unterschied, Zusammenhang), den man als mindesten oder interessanten Effekt für die Population annimmt.

Anpassungstest; 148 f., 151: Der Anpassungstest prüft, ob eine empirische Häufigkeitsverteilung mit einer theoretisch zu erwartenden Häufigkeitsverteilung übereinstimmt. Die zu erwartende Verteilung kann einer Gleichverteilung entsprechen bzw. einer anderen Form der Verteilung, welche sich aus theoretischen Überlegungen oder praktischen Erfahrungen ergeben kann.

Beta-Fehler (Fehler zweiter Art); 63, 65 ff.: Der Beta-Fehler ist die Wahrscheinlichkeit, mit der wir beim Signifikanztest aufgrund eines Stichprobenergebnisses fälschlicherweise die Nullhypothese annehmen und die Alternativhypothese verwerfen (obwohl diese eigentlich in der Population gilt).

Chi-Quadrat-Tests; 144, 148: Chi-Quadrat-Tests (χ^2-Tests) sind non-parametrische Tests zur Analyse von Häufigkeiten.

Effekt; 13, 38 f., 41, 44, 46, 48, 50 ff., 55-58, 62, 64 f., 68, 70 f., 73-83, 85, 109, 127 ff., 132 , 134 f., 142, 147: Als Effekt bezeichnet man die Wirkung einer unabhängigen Variable auf eine abhängige Variable. Effekte lassen

T. Schäfer, *Statistik II*, DOI 10.1007/978-3-531-92089-4,

sich in Form von Unterschieden oder Zusammenhängen beschreiben.

Effektgröße (Effektstärke); 71, 73-83, 107, 111 ff., 133 ff., 147, 154: Effektgrößen sind standardisierte Maße für die Größe eines Effektes. Sie sind über Stichproben und Themenbereiche hinweg vergleichbar. Man kann Abstandmaße (z. B. *d* und *g*) und Zusammenhangsmaße (z. B. *r*) unterscheiden.

einseitige Tests; 61 ff., 105, 123: Von einseitigem Testen spricht man, wenn man eine Hypothese testet, die eine Annahme über die Richtung des Effektes beinhaltet.

Einzelvergleiche (Post-hoc-Tests); 123 f., 130, 132: Im Anschluss an Varianzanalysen prüfen Einzelvergleiche, welche Faktorstufen sich signifikant voneinander unterscheiden. Sie funktionieren im Prinzip wie einzelne *t*-Tests. Allerdings wird hier die Alpha-Fehler-Kumulation berücksichtigt, indem diese Einzelvergleiche eine sogenannte Alpha-Korrektur erhalten.

Erwartungswert; 25, 27 ff., 34, 42, 47: Der Erwartungswert einer bestimmten Kenngröße ist der Wert, den wir in der Population erwarten würden, also eine Schätzung des Populationsparameters. Beispielsweise wird der Mittelwert einer Stichprobe als Erwartungswert für den Mittelwert in der Population benutzt.

***F*-Test; 98, 122 f., 129, 131, 136 f.:** Der *F*-Test ist der Signifikanztest der Varianzanalyse. Er prüft das Verhältnis von aufgeklärter zu nicht aufgeklärter Varianz. Er wird auch dafür verwendet Regressionsmodelle auf Signifikanz zu prüfen.

Haupteffekt; 127 ff., 133 f.: Im Zuge der mehrfaktoriellen Varianzanalyse beschreiben Haupteffekte neben der Interaktion zwischen verschiedenen UVs den isolierten Effekt der einzelnen UVs auf die AV.

Inferenzstatistik; 9-14, 24, 38, 50, 73 f., 89: Ziel der Inferenzstatistik sind Schlüsse von einer Stichprobe auf eine Population sowie Aussagen über die Güte dieser Schlüsse.

Interaktion; 128 ff., 133 f.: Im Zuge der mehrfaktoriellen Varianzanalyse beschreibt die Interaktion neben den Haupteffekten der einzelnen UVs die Wechselwirkung zwischen verschiedenen UVs auf die AV.

Irrtumswahrscheinlichkeit (Alpha, Signifikanzniveau); 52, 58 f., 63, 69, 123: Die Irrtumswahrscheinlichkeit ist die Wahrscheinlichkeit eines Ergebnisses (unter Annahme der Nullhypothese), ab der man nicht mehr bereit ist, die Nullhypothese zu akzeptieren. Empirisch gefundene Ergebnisse, deren Wahrscheinlichkeiten kleiner als diese festgelegte Irrtumswahrscheinlichkeit sind ($p < \alpha$), werden als signifikant bezeichnet und führen zur Ablehnung der Nullhypothese. Die Irrtumswahrscheinlichkeit entspricht damit auch der Wahrscheinlichkeit, mit der man beim Ablehnen der Nullhypothese einen Fehler macht.

Konfidenzintervall, 21, 26-35, 38, 43, 48-55, 71, 73 f., 82 f., 97 f.: Ein Konfidenzintervall im Rahmen der Inferenzstatistik ist ein Wertebereich, bei dem wir darauf vertrauen (konfident sein) können, dass er den wahren Wert in der Population mit einer gewissen Wahrscheinlichkeit (der Vertrauenswahrscheinlichkeit) überdeckt.

Kreuztabelle (Kontingenztafel); 150, 152: Kreuztabellen oder Kontingenztafeln bilden die verschiedenen Kombinationen der Ausprägungen nominalskalierter Variablen ab. Die Zellen enthalten die Häufigkeiten, mit denen die Merkmalskombinationen auftreten.

multiple Regression 89-101, 115, 130: Die multiple Regression ist ein Analyseverfahren, welches direkt aus dem ALM folgt. Sie schätzt mithilfe der Ausprägungen auf mehreren Prädiktorvariablen den Wert einer Person auf einer Kriteriumsvariable. Die Formel für die multiple Regression (Regressionsgerade) besteht aus der Regressionskonstante und den Prädiktoren mit ihren Regressionskoeffizienten.

Multipler Determinationskoeffizient; 47, 95 f., 98, 136 f.: Der multiple Determinationskoeffizient R^2 gibt den Anteil von Varianz des Kriteriums wieder, der im Zuge der multiplen Regression durch alle Prädiktoren gemeinsam erklärt wird. Er kann maximal 1 sein, was einer Varianzaufklärung von 100 Prozent entspricht.

nonparametrische (verteilungsfreie) Testverfahren; 114, 139-154: Nonparametrische oder verteilungsfreie Testverfahren testen Zusammenhänge von Variablen oder Unterschiede zwischen Gruppen. Sie machen jedoch im Gegensatz zu parametrischen Testverfahren keine Annahmen, die sich auf die Verteilung der Messwerte in der Population beziehen, und eignen sich daher auch für Daten auf Nominal- und Ordinalskalenniveau.

Nullhypothese; 55-70, 72 f., 104, 106, 109, 119, 123, 149: Die Nullhypothese (auch als H0 bezeichnet) als zentrale Idee des Signifikanztests behauptet, dass es in der Population keinen Effekt (Unterschied, Zusammenhang) gibt.

parametrische Testverfahren; 139-154: Parametrische Testverfahren testen Zusammenhänge von Variablen oder Unterschiede zwischen Gruppen. Sie setzen im Gegensatz zu nonparametrischen Testverfahren eine bestimmte Verteilung – meist eine Normalverteilung – der Messwerte in der Population voraus.

Population (Grundgesamtheit); 10-14, 18-29, 34 f., 37-43, 47, 49 ff., 55 ff., 64 ff., 68, 71, 73 f., 82, 89, 96 ff., 104, 109, 113, 118 f., 135 f., 141: Die Begriffe Grundgesamtheit und Population beziehen sich auf die Gruppe von Menschen, für die eine bestimmte Aussage zutreffen soll (also entweder auf alle Menschen oder auf eine spezifische Subgruppe).

***p*-Wert; 58-61, 63, 68-71, 100, 106, 122, 124, 130, 133, 144 f.:** Der *p*-Wert ist die Wahrscheinlichkeit dafür, dass in einer Stichprobe der gefundene oder ein noch größerer Effekt auftritt unter der Annahme, dass die Nullhypothese gilt.

Quadratsumme; 119-123, 130, 133: Die Quadratsumme QS ist die Bezeichnung für die Summe quadrierter Differenzen. Sie ist ein Maß für die Streuung von Messwerten, welches noch nicht an der Größe der Stichprobe relativiert ist.

Rangkorrelationen; 145 ff.: Rangkorrelationen sind die non-parametrischen Entsprechungen zur herkömmlichen Pearson-Korrelation. Sie korrelieren nicht die Rohwerte, sondern den Rohwerten zugewiesene oder echte Ränge. Die einzige

Bedingung ist dabei, dass die Beziehung der beiden zu korrelierenden Variablen einer monotonen Steigung folgt.

Regressionsgewichte (Regressionskoeffizienten); 47 f., 53 f., 61, 83, 88, 92, 94 f., 97 f., 103: In der Regression beschreiben die Regressionsgewichte den Einfluss eines Prädiktors auf die Vorhersage des Kriteriums. Bei der multiplen Regression gibt es mehrere Prädiktoren, und jeder Prädiktor erhält ein eigenes Regressionsgewicht, welches um den Einfluss anderer Prädiktoren bereinigt ist.

robustes Testverfahren; 114, 135: Ein robustes Testverfahren ist zwar an bestimmte Voraussetzungen geknüpft, ist gegen Verletzungen dieser Voraussetzungen jedoch so unempfindlich, dass es trotzdem sehr gute Ergebnisse liefert.

Signifikanz; 55, 83, 98, 107, 110, 122, 124, 136, 144 ff.: Signifikanz heißt statistische Bedeutsamkeit und bezieht sich immer auf einen Effekt, den man in einer Stichprobe gefunden hat. Signifikante Ergebnisse werden als systematisch und nicht als zufällig angenommen.

Signifikanztest; 38, 43, 55-75, 80, 82 f., 85 f., 97 f., 100, 104, 106 f., 111, 113, 115, 123, 136, 141, 144, 147: Der Signifikanztest als Methode der Inferenzstatistik liefert die Grundlage für eine Entscheidung zwischen gegensätzlichen Hypothesen. Auf der Grundlage von theoretischen Stichprobenverteilungen gibt er Auskunft darüber, ob die statistische Bedeutsamkeit eines Effektes groß genug ist, um ihn auf die Population zu verallgemeinern.

Standardabweichung; 22-25, 31, 42, 76: Die Standardabweichung ist die durchschnittliche Abweichung einer Reihe von Werten von deren gemeinsamen Mittelwert.

Standardfehler; 21-27, 29, 31-35, 38, 42-49, 52 ff., 68, 73 f., 82 f., 97 f., 105, 107, 109 f.: Der Standardfehler eines Effektes entspricht der Standardabweichung der entsprechenden Stichprobenverteilung. Er bezeichnet die Ungenauigkeit, wenn man den Populationseffekt mithilfe des Stichprobeneffektes schätzt.

Stichprobenverteilung; 16-20, 22 ff., 27-30, 33 ff., 41 ff., 48, 50 ff., 56 ff., 60, 68, 70, 83, 148: In einer Stichprobenverteilung als wichtigste Grundlage der Inferenzstatistik sind die Ergebnisse (Mittelwerte, Anteile, Mittelwertsunterschiede, Korrelationen) vieler Stichproben abgetragen. Sie bildet ab, wie sich die einzelnen Ergebnisse verteilen und wie oft bzw. mit welcher Wahrscheinlichkeit ein bestimmtes Ergebnis zu erwarten wäre. Während eine *empirische* Stichprobenverteilung die Ergebnisse einer endlichen Anzahl realer Studien abbildet, zeigt eine *theoretische* Stichprobenverteilung, wie sich die Ergebnisse verteilen würden, wenn wir unendlich viele Stichproben ziehen würden.

Teststärke (Power); 142: Die Teststärke oder Power gibt die Wahrscheinlichkeit an, mit der ein in der Population tatsächlich vorhandener Effekt mithilfe eines Testverfahrens identifiziert werden kann.

***t*-Test; 101, 103-116, 122 f., 139 ff., 143, 146**: Der *t*-Test ist eine Gruppe von Tests, die für verschiedene Fragestellungen verwendet werden können. Das Prinzip des t-

Tests ist immer der Vergleich zweier Mittelwerte (aus unabhängigen oder abhängigen Stichproben oder ein Mittelwert, der gegen einen theoretisch zu erwartenden Mittelwert getestet wird).

unabhängige Messungen; 39 ff., 44 f., 48, 75-78, 83, 131, 133, 143: Unabhängig sind Messungen dann, wenn jede Messung an einer eigenen Stichprobe bzw. in einer eigenen Gruppe vorgenommen wurde, denen die Versuchsteilnehmer rein zufällig zugeordnet wurden.

Unabhängigkeitstest; 149 f., 152: Der Unabhängigkeitstest prüft für nominalskalierte Daten, ob die Ausprägungen einer Variable unabhängig von den Ausprägungen einer anderen Variable sind.

Varianzanalyse; 98, 101, 115-137, 139 ff., 145: Die Varianzanalyse (ANOVA = Analysis of Variance) untersucht die Unterschiede (Variation) der Mittelwerte von zwei oder mehr Gruppen. Sie prüft das Verhältnis zwischen erklärter Varianz (zwischen den Gruppen) und nicht erklärter Varianz (innerhalb der Gruppen) in den Daten. Ist die erklärte Varianz in diesem Verhältnis groß genug, führt das zu einem signifikanten Gruppenunterschied.

Vertrauenswahrscheinlichkeit; 26 ff., 31 ff., 35, 51-54: Die Vertrauenswahrscheinlichkeit (Konfidenz) ist die Wahrscheinlichkeit, mit der wir darauf vertrauen können, dass ein bestimmtes Konfidenzintervall den wahren Wert in der Population überdeckt.

zentraler Grenzwertsatz; 19: Der zentrale Grenzwertsatz besagt, dass die Verteilung einer großen Anzahl von Stichprobenergebnissen immer einer Normalverteilung folgt.

zweiseitige Tests; 61 ff., 106: Von zweiseitigem Testen spricht man beim Testen von Hypothesen, die keine Annahme über die Richtung des Effektes enthalten.