2.3 Biostatistik

Marcel Zwahlen

Wir lesen in einem Fachartikel, dass bei einer bestimmten Therapieform von 100 Behandelten nur halb so viele versterben wie bei einer anderen Form der Therapie. Ist dieser Unterschied statistisch gut abgesichert (*statistisch signifikant*)? Oder ist es möglich, dass er nur auf Zufall beruht? Es könnte z. B. sein, dass in der ersten Gruppe eine Person verstarb, in der zweiten jedoch zwei. In der ersten Gruppe starben damit tatsächlich nur halb so viele Menschen wie in der zweiten Gruppe. Wie stark unterscheidet sich der Therapieerfolg bei diesen beiden Behandlungsformen nun wirklich? Mit Hilfe der Statistik versuchen wir, über numerische Informationen Antworten auf solche Fragen zu erhalten. Statistik befasst sich mit dem Sammeln, Zusammenfassen, Darstellen und Interpretieren von Daten. *Biostatistik* ist der Zweig der Statistik, der diese Aufgaben in der Biomedizin und in Public Health übernommen hat.

Wir lernen in diesem Kapitel die Grundprinzipien zur Zähmung der Variabilität kennen, d. h. wir erfahren, wie man trotz vorhandener statistischer Unsicherheit möglichst wahrheitsgemäße Schlussfolgerungen über Populationen und Patientengruppen ziehen kann. Statistik kommt dabei nicht ganz ohne mathematische Formeln aus. Sie wird daher von Vielen oftmals als schwierig oder unangenehm angesehen. Wir versuchen hier den mathematischen Formalismus auf das Nötigste zu beschränken.

Schweizerische Lernziele: CPH 13–16

2.3.1 Warum brauchen wir Statistik?

„In God we trust. All others must have data."
W.E. Demming (1900–1993; amerik. Physiker und Statistiker)

„It is easy to lie with statistics. It is hard to tell the truth without statistics."
A. Dunkels (1939–1998; schwed. Mathematiker und Lehrer)

Statistik und statistische Verfahren dienen dazu, aus Situationen, die typischerweise mit einer gewissen Variabilität auftreten, möglichst wahrheitsgemäße Schlüsse zu ziehen. Insbesondere biologische Prozesse zeigen oft eine solche inhärente Variabilität. Dies spiegelt sich dann auch in biomedizinischen Messwerten wider. So variiert beispielsweise der arterielle Blutdruck nicht nur von Mensch zu Mensch, sondern auch bei einem Individuum von Stunde zu Stunde. In einer Population von Individuen äußert sich Variabilität in Form von zufällig auftretenden Ereignissen oder Messwerten. Einerseits können beispielsweise Personen, die gegen eine bestimmte Infektionskrankheit geimpft wurden, trotz Impfung an dieser Infektion erkranken, andererseits können ungeimpfte Personen gesund bleiben. Wenn wir diese Situation aus statistischer Sicht betrachten, stellen sich uns u. a. folgende Fragen:

- Was kann daraus geschlossen werden, wenn bei den geimpften Personen ein größerer Anteil gesund bleibt als bei den ungeimpften?

- Wie wirksam ist der Impfstoff? Ist der Unterschied zwischen Geimpften und Ungeimpften vielleicht zufällig zustande gekommen?
- Gaukelt uns eine Verzerrung bei der Studienpopulation möglicherweise eine Wirkung der Impfung nur vor? So könnte z. B. die Gruppe der Geimpften mehr Interesse an präventiven Maßnahmen gezeigt haben als die der Ungeimpften. Damit wäre denkbar, dass sich beide Gruppen im Gesundheitsverhalten und in den generellen Lebensumständen unterscheiden. Dies alles sind Faktoren, die die Erkrankungswahrscheinlichkeit beeinflussen könnten.
  Statistische Methoden erlauben es, die ersten beiden Fragen zu beantworten. Das in der dritten Frage angesprochene Problem eines Selektionsbias (s. Kap. 2.1.4) kann durch eine sorgfältige Planung der durchzuführenden Studie verhindert werden.

Die *Hauptarbeitsbereiche der Biostatistik* sind

- die Mithilfe bei der Planung von Studien (s. die verschiedenen Studientypen in Kap. 2.1)
- die Beschreibung und Zusammenfassung von erhobenen Daten (z. B. des mittleren Blutdrucks in einer Population, s. „deskriptive Statistik")
- die Quantifizierung von wichtigen Kenngrößen in Populationen oder Patientengruppen (z. B. die Inzidenz einer Infektion, s. „Schätzen von Parametern")
- das Testen von präzisen quantitativen Hypothesen („Impfstoff A ist 20 % wirksamer als Impfstoff B")

2.3.2 Klassifikation von Daten

Um in der Biomedizin und in Public Health Antworten auf Fragen zu bekommen, werden in der Regel Studien durchgeführt, die Messungen beinhalten. Gemessen werden bestimmte Charakteristika (**Variablen**), die Antworten auf die bestehenden Fragen versprechen. Häufig sind dies Untersuchungen bei StudienteilnehmerInnen. Es kann sich aber auch um Messwerte handeln, die an Versuchstieren gewonnen wurden oder um Charakteristika von Krankenhäusern oder Analyseergebnisse aus Urinproben. Jeder Aspekt, der untersucht wird, wie etwa der Blutdruck, der Cholesterinspiegel oder das Geschlecht, entspricht in der Regel einer Variablen. Bevor die Anwendung bestimmter statistischer Verfahren festgelegt und erste Berechnungen durchgeführt werden, lohnt es sich, die vorhandenen Daten anzusehen und sie nach Datentypen zu ordnen. In einem ersten Schritt wird zwischen quantitativer und kategorischer Information unterschieden.

Quantitative Daten sind entweder *kontinuierliche* oder *diskrete Daten*. Als kontinuierliche Variable bezeichnet man einen Messwert, der sich auf einer kontinuierlichen Skala mit einer definierten Maßeinheit abbilden lässt. Kontinuierliche Variablen sind z. B. das Körpergewicht oder ein Cholesterinwert. Sie können jeden beliebigen Wert auf der Skala des Messgerätes einnehmen. Im Gegensatz dazu kann eine diskrete Variable nur eine beschränkte Anzahl, meist ganzzahliger Werte annehmen. Beispiele hierfür sind die Anzahl von Geburten oder von Krankenhausaufenthalten im letzten Jahr.

Kategorische Daten werden auch *nominale* oder *qualitative Daten* genannt. Hierbei handelt es sich um nicht-numerische Daten, wie beispielsweise der Geburtsort, die

Nationalität, die Augenfarbe oder die Art eines Medikaments. Eine wichtige Untergruppe kategorischer Daten sind so genannt *binäre oder dichotome Variablen,* die nur zwei mögliche Werte kennen. So ist das Geschlecht entweder weiblich oder männlich, und der Teilnehmer an einer Studie ist bei Studienende entweder am Leben oder gestorben.

Bei **geordneten kategorischen Daten** gehen wir davon aus, dass den Kategorien – auch wenn sie nicht-numerischer Art sind – eine natürliche Ordnung zukommt. Geordnete kategorische Daten sind z. B. die Antworten auf die folgende Frage:

„Während meines Krankenhausaufenthaltes wurde ich mit Respekt und Würde behandelt."

Bitte beantworten Sie, ob Sie dieser Aussage

a. überhaupt nicht zustimmen
b. ein wenig zustimmen
c. stark zustimmen
d. vollumfänglich zustimmen

Ein weiteres Beispiel für eine solche natürliche Ordnung sind die Stadien einer Krebserkrankung: Stadium I hat eine bessere Prognose als Stadium IV.

2.3.3 Transparentes Zusammenfassen der erhobenen Daten

Die quantitativen Daten, die in einer Studie erhoben wurden, müssen in einem ersten Schritt geeignet zusammengefasst werden, um eine bessere Übersichtlichkeit zu erreichen.

Betrachten Sie z. B. die folgende Situation:

In einer Studie, an der 200 Personen teilnahmen, wurden u. a. Gewicht und Körpergröße gemessen. Anhand dieser Werte wurde anschließend der Body Mass Index (BMI) der TeilnehmerInnen durch Division von Körpermasse (in Kilogramm) durch das Quadrat der Körpergröße (in Metern) berechnet. Die alleinige Auflistung der 200 BMI-Werte wäre nun bei der Beurteilung dieser Daten wenig hilfreich:

BMI-Werte [kg/m^2] der Personen 1–10:
24,0; 27,6; 28,7; 29,0; 25,4; 25,7; 27,8; 25,3; 28,4; 29,0

BMI-Werte [kg/m^2] der Personen 191–200:
28,5; 24,8; 28,6; 21,8; 24,4; 24,4; 21,3; 26,8; 27,7; 22,9

Es ist sinnvoller, eine leicht verständliche Zusammenfassung dieser Werte zu erstellen. Das kann mittels *grafischer Darstellung* oder mit Hilfe geeigneter *Kennzahlen* geschehen. Nützlich ist in diesem Zusammenhang die so genannte **Fünf-Zahlen-Zusammenfassung** (*Five-Number Summary*). Zu diesen fünf Zahlen gehören:

- **Tiefster Wert** (Minimum)
- **Unteres Quartil**: Der Wert, der die vorliegende Reihe von Werten so unterteilt, dass 25 % der Werte kleiner als dieser Wert sind.
- **Median** (m): Der Wert, der die Reihe so unterteilt, dass (höchstens) die Hälfte der Werte kleiner als m und (höchstens) die Hälfte der Werte größer als m sind. Bei

einer geraden Anzahl von Werten (k = Anzahl der vorliegenden Werte) wird die Mitte zwischen dem (k/2)-ten und (k/2 + 1)-ten Wert genommen.

- **Oberes Quartil**: Der Wert, der die Reihe von Werten so unterteilt, dass 75 % der Werte kleiner als das obere Quartil sind.
- **Höchster Wert** (Maximum).

Bei den 200 Personen, für die der BMI berechnet wurde, ergäben sich daraus z. B. die folgenden Werte (in kg/m²) der Fünf-Zahlen-Zusammenfassung:

Minimum: 16,70; Unteres Quartil: 23,50; Median: 25,10; Oberes Quartil: 26,85; Maximum: 39,20.

Diese fünf Kennzahlen lassen sich auch in einem so genannten **Boxplot** (Kastengrafik) oder *Box-Whisker-Plot* darstellen (Abb. 2.12). Im Boxplot sehen wir in der Mitte eine dunkler eingefärbte Box, welche durch die Werte des unteren und oberen Quartils begrenzt ist. 50 % aller Werte liegen innerhalb des Interquartilbereichs zwischen 23,50 und 26,85. In der Mitte dieser Box ist der Median-Wert eingezeichnet. Die beiden Linien, die von den Rändern der Box ausgehen, werden *Whisker* (Antennen, Fühler) genannt. Die Länge dieser Whisker ist auf das 1,5-Fache des Interquartilabstands beschränkt. In unserem Beispiel beträgt der Interquartilabstand 26,85 – 23,5 = 3,35. Die Begrenzung des oberen Whiskers liegt also maximal bei 26,85 + (1,5 × 3,35) = 31,875. Der untere Whisker erstreckt sich bis maximal 23,5 – (1,5 × 3,35) = 18,475. Werte, die weiter als die beiden Whisker vom Median entfernt liegen, werden einzeln dargestellt und als Ausreißerwerte bezeichnet.

Eine andere Form der grafischen Darstellung ist das **Histogramm**. Hierbei werden zuerst Werteintervalle gebildet, anschließend wird gezählt, wie viele der vorliegenden Werte in die jeweiligen Intervalle fallen. Das Ergebnis kann man dann auf zwei verschiedene Arten grafisch darstellen. Entweder wird die Anzahl oder der Prozentsatz der Werte aufgezeigt, die jeweils in die gebildeten Werteintervalle fallen. Abb. 2.13 zeigt eine solche Darstellung. Die gewählten Intervalle haben hier eine Länge von 2 kg/m²

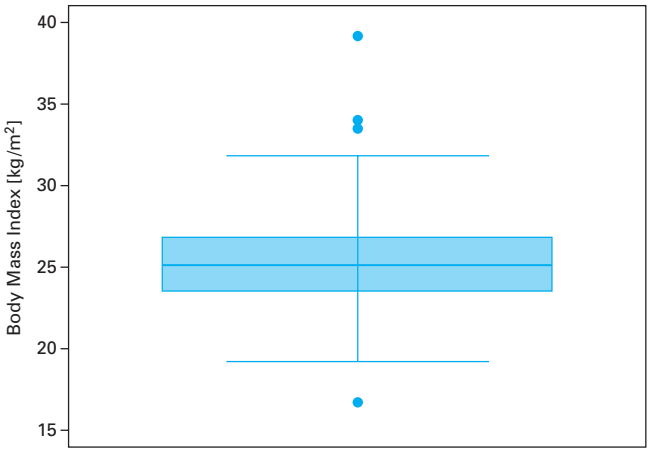

Abb. 2.12: Boxplot-Darstellung der BMI-Werte der 200 Personen aus dem Anwendungsbeispiel. Bei den Punkten außerhalb der Whisker handelt es sich um so genannte Ausreißerwerte.

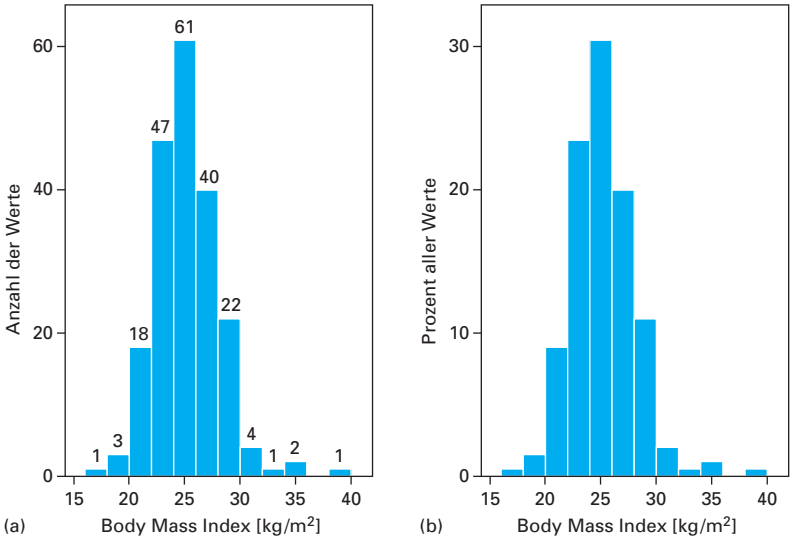

Abb. 2.13: Histogramm der BMI-Werte der 200 Personen aus dem Anwendungsbeispiel. (a) Histogramm, bei dem die Anzahl der Werte in der jeweiligen Wertegruppe angegeben sind. (b) Histogramm, das die jeweiligen Prozentsätze angibt.

(z. B. 16 bis < 18 kg/m², 18 bis < 20 kg/m² etc.). Der Nachteil eines solchen Histogramms ist, dass es von der gewählten Intervalleinteilung abhängt, welches Bild man erhält.

Eine weitere Möglichkeit der grafischen Darstellung ist das **Streudiagramm**. Hierdurch können zwei verschiedene Merkmale gleichzeitig dargestellt werden. In unserem Anwendungsbeispiel ließe sich auf diese Weise etwa der BMI mit dem Alter der 200 untersuchten Personen verknüpfen (Abb. 2.14).

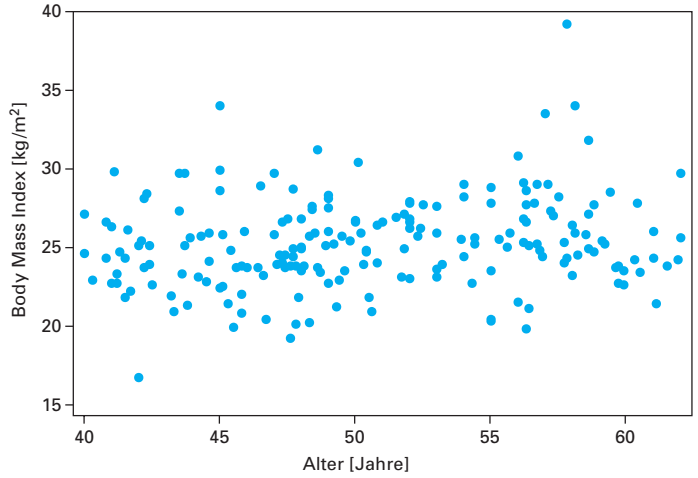

Abb. 2.14: Streudiagramm (*Scatter Plot*), das das Alter der 200 Personen aus dem Anwendungsbeispiel zu ihrem Body Mass Index in Relation setzt.

Oft werden auch der **Mittelwert** und die **Standardabweichung** als Kennzahlen zur Zusammenfassung der vorliegenden Werte verwendet. Die Anzahl der Werte des Datensatzes bezeichnet man hierbei mit N.

Formel 2.1: Formeln für die Berechnung des Mittelwertes und der Standardabweichung einer Wertereihe. SD = Standard Deviation (engl.)

$$\text{Mittelwert} = \frac{X_1 + \ldots + X_N}{N} = \frac{\sum_{j=1}^{j=N} X_j}{N}$$

$$\text{Standardabweichung (SD)} = \sqrt{\frac{\sum_{j=1}^{j=N} (X_j - \text{Mittelwert})^2}{N - 1}}$$

Der Mittelwert ist eine Kennzahl für typische Werte in der Mitte einer Datenreihe, die Standardabweichung kennzeichnet dagegen die Variabilität der betrachteten Werte. Zu beachten ist, dass bei der Berechnung der Standardabweichung die Summe der quadrierten Abstände zum Mittelwert durch die um 1 reduzierte Anzahl der Werte geteilt wird (N–1).

Ständen uns alle 200 BMI-Werte aus unserem Anwendungsbeispiel zur Verfügung, ließe sich daraus ein Mittelwert von 25,3 kg/m² sowie eine Standardabweichung von 2,92 kg/m² berechnen. In unserem Beispiel nimmt die Standardabweichung damit einen ähnlichen Wert wie der Interquartilabstand ein, der 3,35 kg/m² betrug.

Mittelwert und Standardabweichung reagieren empfindlich darauf, wenn einige wenige Werte weit außerhalb des übrigen Wertebereichs liegen. So würde sich die Standardabweichung z. B. von 2,92 auf 4,98 kg/m² erhöhen, wenn unter den Werten unseres Beispiels anstatt der zehn höchsten BMI-Werte zwischen 29,8 kg/m² und 39,2 kg/m² zehn Werte von jeweils 45 kg/m² gewesen wären. Der Mittelwert würde nun 25,9 kg/m² betragen. Median und Interquartilbereich würden sich jedoch nicht ändern. Tab. 2.6 fasst die *Vor- und Nachteile der Kennzahlen quantitativer Daten* zusammen.

Will man dagegen die Resultate von **qualitativen Daten** zusammenfassen, ist es nicht sinnvoll, Median oder Mittelwert zu berechnen. Dies gilt auch dann, wenn Zahlencodes verwendet wurden, wie z. B. die Zahlen 1 bis 5 zur Kodierung des Personenstands (schweizerisch: Zivilstand) in ledig, verheiratet, geschieden, verwitwet, getrennt lebend. Eine nützliche Information bei qualitativen Daten ist die *prozentuale Verteilung* auf die verschiedenen Kategorien. Diese kann dann anhand einer **Tabelle** (Tab 2.7) oder grafisch in Form eines **Kuchendiagramms** (*Pie chart*; Abb. 2.15) oder eines **Häufigkeitsdiagramms** (s. Web-Abb. 2.3.1 auf unserer Lehrbuch-Homepage) dargestellt werden. Die Kuchengrafik bezeichnet man auch als *Kreisdiagramm*, die Häufigkeitsgrafik als *Balkendiagramm* (*Bar chart*).

Tab 2.6: Vor- und Nachteile der Kennzahlen quantitativer Daten.

	Vorteil	Nachteil
Kennzahlen für die Mitte		
Mittelwert	Einfach zu berechnen, gute statistische Eigenschaften	Reagiert empfindlich auf Ausreißerwerte
Median	Einfach zu verstehen, reagiert nicht sensibel auf Ausreißerwerte (= robust gegenüber Ausreißerwerten)	Hat komplexe statistische Eigenschaften
Kennzahlen für die Variabilität		
Standardabweichung	Hat gut verstandene statistische Eigenschaften	Ist kompliziert zu berechnen, reagiert empfindlich auf Ausreißerwerte
Interquartilbereich	Einfach zu verstehen: 50 % aller Werte liegen in diesem zentralen Bereich	Hat komplexe statistische Eigenschaften

Tab. 2.7: Zivilstand (Personenstand) der 30- bis 49-jährigen Männer und Frauen in der Schweiz (Schweizerische Gesundheitsbefragung 2007).

Zivilstand	Männer	Frauen
Ledig	26,5 %	18,5 %
Verheiratet	64,8 %	68,7 %
Verwitwet	0,6 %	1,4 %
Geschieden	6,3 %	9,4 %
Getrennt lebend	1,8 %	2,0 %
Gesamt	100 %	100 %

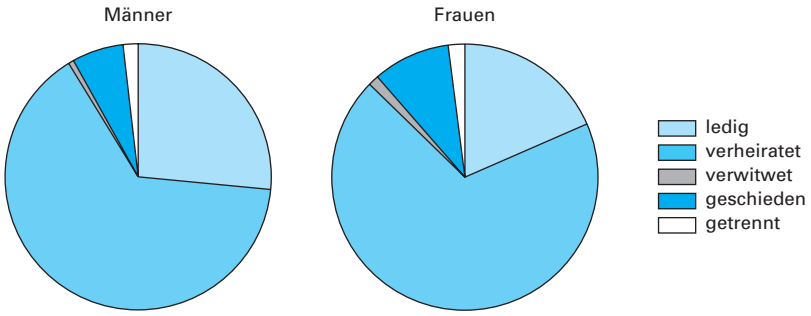

Abb. 2.15: Kuchengrafik, die den Zivilstand (Personenstand) der 30- bis 49-jährigen Männer und Frauen in der Schweiz wiedergibt (Schweizerische Gesundheitsbefragung 2007; die genauen Prozentsätze zeigt Tab. 2.7).

2.3.4 Variabilität des Mittelwertes bei wiederholten Zufalls-Stichproben

Da es nur ausnahmsweise möglich ist, Untersuchungen ganzer **Populationen** durchzuführen, ist man in der Regel dazu gezwungen, sich mit der Analyse einer Teilmenge einer Population zu begnügen. Wenn diese Teilmenge nach dem Zufallsprinzip ausgewählt wurde, wird sie als *zufällig gezogene Stichprobe* bezeichnet. Die Wahrscheinlichkeitslehre erlaubt es nun, anhand einer solchen Stichprobe Aussagen darüber zu machen, wie sich die statistischen Kennzahlen der zufällig gezogenen Stichprobe von den wahren Werten in der Gesamtpopulation unterscheiden. Wenn man anhand der Resultate einer Stichprobe Aussagen über die ganze Population machen will, muss man allerdings berücksichtigen, dass aufgrund des Zufalls mehrere, nach dem gleichen Zufallsprinzip gezogene Stichproben unterschiedliche Kennzahlen liefern. Dieses Phänomen der „Stichprobenvariation" soll nun anhand von Computersimulationen illustriert werden.

Computersimulation der Stichprobenvariabilität

Hierzu stellen wir uns vor, dass wir im Jahr 2007 in der Schweiz alle rund 2,4 Mio. Personen im Alter zwischen 30 und 49 Jahren nach ihrem Personenstand (Zivilstand) befragt hätten. Es zeigte sich, dass exakt 60 % der Befragten verheiratet waren. Wir ziehen nun am Computer eine zufällige Stichprobe von einer bestimmten Größe (z. B. 50 Personen) aus der Gesamtpopulation und berechnen anschließend den Prozentsatz der verheirateten Personen aus dieser Stichprobe. Das Ganze wird 10.000-mal wiederholt, und zum Schluss wird die Verteilung der in den Stichproben berechneten Prozentsätze mittels eines Histogramms beschrieben. Dieses Prozedere wird dann für eine Stichprobengröße von 100, 300 und 500 Personen wiederholt (Abb. 2.16). Es wird bei allen Stichprobengrößen deutlich, dass die Verteilung des berechneten Prozentsatzes jeweils um die Mitte, den wahren Wert von 60 % schwankt. Allerdings variieren die Resultate bei einer Stichprobengröße von 50 Personen relativ stark zwischen 40 % und 80 %. Dagegen kommt es bei einer Stichprobengröße von 300 Personen kaum vor, dass der berechnete Prozentsatz kleiner als 50 % oder größer als 70 % wird. Je mehr Personen eine Stichprobe umfasst, desto weniger variieren also die in der Stichprobe berechneten Resultate. Im Grenzfall einer Vollerhebung entspricht der berechnete Wert exakt dem wahren Wert.

Auch für die Berechnung des Mittelwertes einer Stichprobe gilt, dass der hier berechnete Wert vom wahren Mittelwert umso weniger abweicht, je größer die Stichprobe ist. Bei unserem Beispiel der rund 2,4 Mio. SchweizerInnen im Alter zwischen 30 und 49 Jahren sind die Body-Maß-Index-Werte normalverteilt mit einem Mittelwert von 25 kg/m^2 und einer Standardabweichung von 4 kg/m^2. Werden nun aus der Gesamtpopulation wiederholt Stichproben verschiedener Größe gezogen und wird pro Stichprobe der Mittelwert des BMI berechnet, so ergibt sich hieraus eine Verteilung der für die Stichproben berechneten Mittelwerte um die Mitte, den wahren Wert von 25 kg/m^2 herum. Die Mittelwerte variieren dabei umso mehr, je kleiner die Anzahl an Personen in der Stichprobe ist (s. Web-Abb. 2.3.2 auf unserer Lehrbuch-Homepage).

Aus der Wahrscheinlichkeitslehre ergibt sich auch, dass die Stichprobenvariabilität einer berechneten Proportion oder eines berechneten Mittelwertes annährungsweise durch die so genannte *Normalverteilung* beschrieben werden kann, wenn die Stichpro-

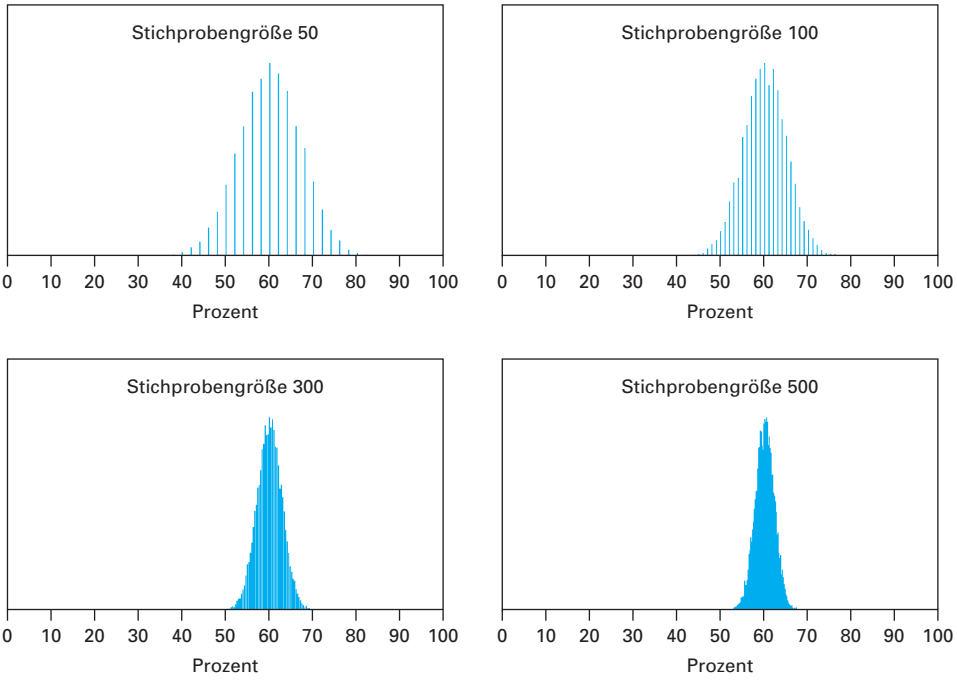

Abb. 2.16: Stichprobenvariabilität in Abhängigkeit von der Stichprobengröße (50, 100, 300 und 500 Personen) für den Prozentsatz an verheirateten Personen bei einem wahren Prozentsatz von 60 % in der Gesamtpopulation. Resultate von Computersimulationen mit jeweils 10.000 Stichproben.

ben nach dem Zufallsprinzip gezogen wurden. Je größer hierbei die Stichprobengröße N ist, desto exakter stimmt diese Annäherung. Die normalverteilten Werte liegen dabei zentriert um den wahren Wert herum. Die „Breite" der Normalverteilung muss allerdings geeignet gewählt werden. Die Web-Abb. 2.3.3 auf unserer Lehrbuch-Homepage zeigt geeignete Normalverteilungskurven für die Berechnung des Prozentsatzes verheirateter Personen (oben) sowie für die Berechnung des mittleren BMI-Wertes (unten), jeweils in Abhängigkeit von der Stichprobengröße. Diese entsprechen annähernd den Simulationsverteilungen, die in den oberen Hälften von Abb. 2.16 und Web-Abb. 2.3.2 zu sehen sind.

2.3.5 Die Normalverteilung in aller Kürze

Es ist sinnvoll, sich eingehender mit der *Normalverteilung* (Gauß-Verteilung) auseinander zu setzen. „Normal" bedeutet hier, dass diese statistische Verteilung in vielen Situationen einer guten Annäherung an die wahren Werte entspricht, sodass sie auch als Grundlage für Berechnungen dienen kann. Abb. 2.17 zeigt die *Dichtefunktion der Standard-Normalverteilung*. Die Dichtefunktion kann man sich als „geglättetes Histogramm" von unendlich vielen Werten vorstellen. Es stellt aber nicht die Anzahl der beobachteten Werte dar, sondern die prozentuale Verteilung dieser Werte. Hierbei beträgt die gesamte Fläche zwischen der Funktionslinie und der X-Achse genau 100 %.

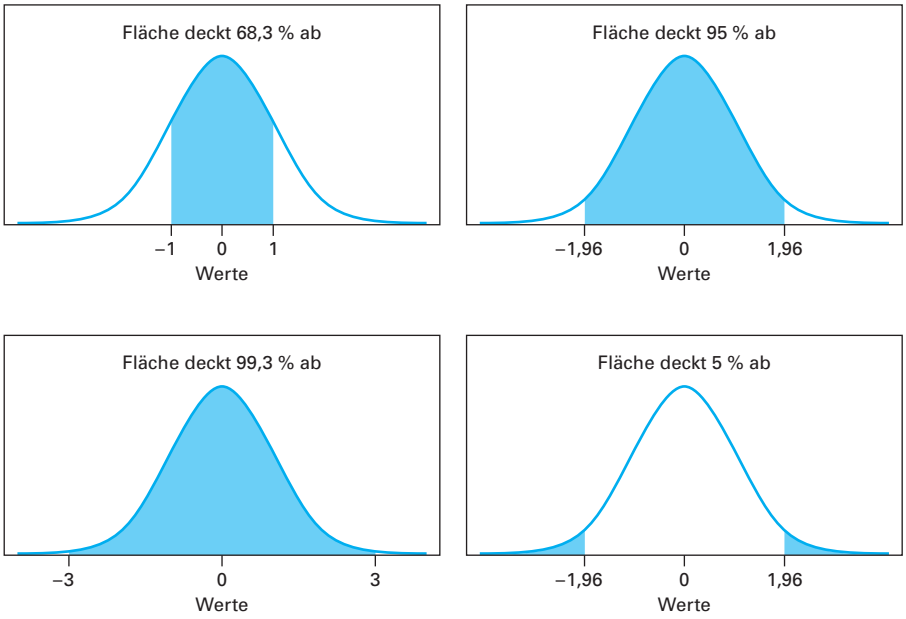

Abb. 2.17: Die Standard-Normalverteilung mit dem Mittelwert (MW) = 0 und der Standardabweichung (SD) = 1. Die gesamte Fläche zwischen der Linie der Dichtefunktion und der X-Achse beträgt 100 %.

Gibt man ein bestimmtes Intervall vor, dann ergibt die *Fläche unter der Kurve* den Prozentsatz der Werte in diesem Intervall. Im Intervall zwischen – 1 und + 1 befinden sich z. B. 68,3 % aller Werte. Da die Kurve symmetrisch zum Wert 0 ist, kann man daraus ableiten, dass 34,15 % der Werte zwischen 0 und + 1 liegen. Dem entsprechend befinden sich 95 % aller Werte zwischen – 1,96 und + 1,96. Fünf Prozent der Werte liegen damit weiter als 1,96 von Null entfernt.

Leider ist es nicht möglich, eine einfache Formel anzugeben, mit der man diese Flächenabschnitte für jedes Intervall selber berechnen kann. Früher gab es daher in Statistikbüchern eine Tabelle, die die Werte für die Flächenabschnitte von minus Unendlich bis zu einem bestimmten Wert („Z-Wert" genannt) angab. Heute kann man diese mit Hilfe verschiedener Computerprogramme berechnen (s. Internet-Ressourcen).

2.3.6 Das 95 %-Vertrauensintervall

Wir wissen nun, dass bei einer großen Stichprobe der hierbei berechnete Mittelwert bzw. ein bestimmter Prozentsatz der Variablen annähernd einer Normalverteilung um den wahren Wert in der Gesamtpopulation folgt. Damit ist es uns möglich, rund um den in der Stichprobe berechneten Wert (Mittelwert oder Prozentsatz) ein Intervall zu berechnen, das mit einer gewünschten Wahrscheinlichkeit den wahren Wert enthält. Was uns noch fehlt, ist eine Formel, die angibt, welche Breite die zu benützende Normalverteilung haben soll. Sie kann durch die Formeln für den so genannten Standardfehler berechnet werden. In Englisch wird der Standardfehler als *Standard Error* bezeichnet und mit „SE" abgekürzt.

Den Standardfehler für einen Mittelwert errechnet man, indem man die Standardabweichung aller Werte in der Population durch die Quadratwurzel der Stichprobengröße dividiert.

Formel 2.2: Formeln für die Berechnung des Standardfehlers eines Mittelwertes (a) und einer Proportion (b). SE = Standard Error (engl.)

$$\text{a. Standardfehler (SE) für Mittelwert} = \frac{\text{Standardabweichung der Werte}}{\sqrt{N}} = \frac{SD}{\sqrt{N}}$$

$$\text{b. Standardfehler (SE) für Proportion} = \sqrt{\frac{\text{Proportion} \times (1 - \text{Proportion})}{N}}$$

In der Praxis ist die Standardabweichung für alle Werte einer Population in der Regel jedoch nicht bekannt. Deshalb wird hierzu die Standardabweichung der vorliegenden Werte benützt. Beide Werte stimmen näherungsweise überein. Auch hier gilt: Die Annäherung ist umso besser, je größer die Stichprobe ist. Eine analoge Formel gibt es für den Standardfehler einer Proportion.

Um nun das **95%-Vertrauensintervall** (**VI**, auch: Konfidenzintervall) zu berechnen, wird anschließend zum erhaltenen Wert noch 1,96-mal der Standardfehler für diesen Wert auf der einen Seite addiert, auf der anderen Seite subtrahiert. Das gewählte Vertrauensintervall umfasst denjenigen Bereich um den geschätzten Wert herum, der mit einer zuvor festgelegten Wahrscheinlichkeit (hier: 95 %) die wahre Lage dieses Wertes angibt.

Formel 2.3: Formeln für die Berechnung des 95%-Vertrauensintervalls eines Mittelwertes.

$$\begin{aligned}\text{95\%-Vertrauensintervall (VI) für Mittelwert} &= [\text{MW} - 1{,}96 \times \text{SE(MW)};\\ &\quad\ \text{MW} + 1{,}96 \times \text{SE(MW)}]\\ &= \text{MW} \pm 1{,}96 \times \text{SE(MW)}\end{aligned}$$

MW = Mittelwert
SE = Standardfehler

Analog hierzu lässt sich auch das 95%-Vertrauensintervall für eine Proportion berechnen.

Wenn nun z. B. bei einer randomisierten Behandlungsstudie zwei Gruppen miteinander verglichen werden, dann interessiert uns in der Regel eine Größe, die den Unterschied zwischen den beiden Gruppen beschreibt. Dies kann z. B. die Differenz zwischen den beiden Mittelwerten der betrachteten Gruppen sein oder die Differenz zwischen zwei Proportionen dieser Gruppen. Auch die Risiken, dass ein bestimmtes Ereignis wie Rückfall, Herzinfarkt oder Tod eintritt, können in beiden Gruppen unterschiedlich verteilt sein. Hier lässt sich ebenfalls ein 95%-Vertrauensintervall analog zum oben beschriebenen Vorgehen konstruieren. Zum beobachteten Wert für die Differenz wird auf der einen Seite der Standardfehler für die interessierende Größe 1,96-mal addiert, auf der anderen Seite subtrahiert.

Entsprechende Formeln kommen bei der Berechnung des 95%-Vertrauensintervalls eines *relativen Risikos* und der *Odds Ratio* zur Anwendung (s. a. Kap. 2.1). Hier muss jedoch beachtet werden, dass die Werte vor der Berechnung logarithmiert und dann zum Schluss auf die gewünschte Skala zurück transformiert werden müssen. Verwendet

wird dabei der natürliche Logarithmus zur Basis der eulerschen Zahl *e*. Die Rücktransformation muss daher mit der Exponentialfunktion e^x geschehen.

Es stellt sich nun natürlich die Frage, wie gut diese Annäherungen (*Approximationen*) in der praktischen Anwendung sind. Anstelle der annähernden Berechnung des wahren Mittelwertes mit Hilfe des Stichprobenmittelwertes und seines 95%-Vertrauensintervalls gibt es noch eine verfeinerte Approximation mittels der *Familie von t-Verteilungen*. Die t-Verteilungen zeigen im Gegensatz zur Normalverteilung für *kleine n-Werte* eine größere Breite und eine Betonung der Flanken. Ein Vergleich zwischen beiden Methoden zeigt, dass z. B. bei einer Wertereihe von nur 50 (bzw. 20, 10 oder 5) Messungen für die Formel des 95%-Vertrauensintervalls anstatt 1,96 besser der Wert 2 (bzw. 2,1; 2,23; 2,57) verwendet werden sollte. Bei der Berechnung von Proportionen, wie z. B. eines Risikos aus einer t-Verteilung (= Anzahl Ereignisse / Anzahl Personen in der Gruppe), wird die approximative Berechnung des 95%-Vertrauensintervalls unzuverlässlich, sobald die Anzahl der Ereignisse bzw. der Nichtereignisse kleiner als 5 wird. In diesen Fällen müssen anstatt der Annäherung über die Normalverteilung andere Methoden verwendet werden.

2.3.7 Der Umgang mit Wahrscheinlichkeiten: Interpretation von Untersuchungen und Tests

Obwohl die Wahrscheinlichkeitsrechnung als Teilgebiet der Mathematik nur wenige Rechenregeln kennt und daher auf den ersten Blick einfach erscheint, macht der Umgang mit Wahrscheinlichkeiten nicht selten Probleme. Die Web-Box 2.3.1 auf unserer Lehrbuch-Homepage fasst wichtige Regeln der Wahrscheinlichkeitsrechnung zusammen.

Probleme entstehen besonders häufig bei der Interpretation der Resultate von Untersuchungen und Tests. Hier sind mehrere, unterschiedlich definierte *bedingte Wahrscheinlichkeiten* von Bedeutung. So ist die **Sensitivität** eines Tests die Wahrscheinlichkeit, dass der Test bei einer tatsächlich erkrankten Person positiv ausfällt. In mathematischer Schreibweise wird dies folgendermaßen ausgedrückt: P(positiver Test | Person hat die Krankheit). Als **Spezifität** eines Tests bezeichnet man dagegen die Wahrscheinlichkeit, dass bei einer nicht erkrankten Person der Test auch tatsächlich negativ ausfällt. Die Notation lautet hier: P(negativer Test | Person hat die Krankheit nicht). Es handelt sich hierbei um bedingte Wahrscheinlichkeiten, da man jeweils das Eintreten eines Ereignisses *A* (positiver/negativer Test) unter der Bedingung anschaut, dass ein anderes Ereignis *B* (Person hat eine Krankheit/keine Krankheit) eingetreten ist.

Was ist nun die Wahrscheinlichkeit für einen falsch positiven Test? Um dies zu beantworten gilt es, die Frage zu präzisieren. Bezieht sich diese Aussage auf nicht erkrankte Personen, dann ist die Wahrscheinlichkeit für einen falsch positiven Test einfach gleich 1 – Spezifität: Die beiden Wahrscheinlichkeiten verhalten sich komplementär (s. Web-Box 2.3.1). In der Praxis bezieht sich die Frage jedoch häufig auf die Personen mit positiven Testresultaten. Um diese beraten zu können, muss man wissen, wie häufig Personen mit einem positiven Test die Krankheit nicht haben: P(Person hat die Krankheit nicht | positiver Test). Wir können dies am Beispiel von systematisch durchgeführten Mammografien bei 50-jährigen Frauen zur Früherkennung von Brustkrebs illustrieren. Um die Frage beantworten zu können, benötigen wir einige zusätzliche Angaben:

- Die *Sensitivität* der Mammografie-Untersuchung liegt bei 85 %, d. h. bei 100 Frauen mit Brustkrebs wird der Test in 85 Fällen positiv ausfallen.
- Die *Spezifität* der Mammografie-Untersuchung liegt bei 97 %, d. h. bei 100 Frauen ohne Brustkrebs wird der Test in 3 Fällen positiv ausfallen.
- Weiter wird angenommen, dass von tausend 50-jährigen Frauen, die sich gesund fühlen, zwei unerkannt an Brustkrebs erkrankt sind (*Prävalenz*, s. a. Kap. 2.1).

Diese realistischen Annahmen liegen den Berechnungen in Tab. 2.8 zugrunde. Hier wurden 10.000 Frauen mammografiert. Da nach unserer Annahme zwei von 1.000 sich gesund fühlenden Frauen an Krebs erkrankt sind, haben in unserer Gruppe 20 Frauen Brustkrebs. Die restlichen 9.980 Frauen sind nicht an Brustkrebs erkrankt. Bei einer Test-Sensitivität von 85 % hätten 17 der 20 Frauen mit Brustkrebs ein positives Test-Resultat. Drei Brustkrebsfälle blieben dagegen unerkannt. Darüber hinaus gäbe es auch bei 3 % (= 299 Frauen) der 9.980 Frauen ohne Brustkrebs ein positives Test-Resultat (Spezifität 97 %). Aus diesen Berechnungen ergibt sich, dass insgesamt 299 der 316 (17 + 299) positiven Tests falsch positiv sind, was 94,6 % entspricht!

Tab. 2.8: Interpretation der Resultate eines Tests am Beispiel eines Mammografie-Screenings bei 10.000 Frauen.

Annahmen:
- Die Sensitivität des Tests ist 85 %
- Die Spezifität des Tests ist 97 %.
- Die Prävalenz der Krankheit beträgt 2 von 1.000 (also 20 von 10.000)

Testresultat	Personen mit der Krankheit	Personen ohne die Krankheit	Gesamt
Positiv	17	299	316
Negativ	3	9.681	9.684
Gesamt	20	9.980	10.000

- Der *positiv prädiktive Wert* des Tests beträgt 17/316 = 5,4 %.
- 94,6 % der positiven Tests sind falsch positiv.
- Der *negativ prädiktive Wert* des Tests beträgt 9.681/9.684 = 99,97 %

Als **positiv prädiktiven Wert** (*PPV = Positive Predictive Value*) bezeichnet man den Anteil der tatsächlich erkrankten Personen unter allen Personen mit positivem Test: P(Person hat die Krankheit | positiver Test). In der geschilderten Situation wären dies 17 von 317 Frauen, d. h. nur 5,4 % der Frauen mit positivem Test wären tatsächlich erkrankt. Führt man dieselbe Berechnung mit einem anderen Prävalenzwert durch, ändert sich auch der PPV. Je höher die Krankheitshäufigkeit ist, desto höher liegt auch die Zahl der tatsächlich Erkrankten unter den positiv getesteten Personen und damit der PPV. Dies ist einer der Gründe, warum das Mammografie-Screening bei Frauen unter 50 Jahren nicht empfohlen wird: Brustkrebs ist in dieser Altersgruppe weniger häufig als bei älteren Frauen.

Der **negative prädiktive Wert** (*NPV = Negative Predictive Value*) ist definiert als der Anteil der tatsächlich gesunden Personen unter allen Personen mit negativem Test:

P(Person hat die Krankheit nicht | negativer Test). In unserem Fall wären das 9.681 von 9.684 Frauen. Dies bedeutet, dass 99,97 % aller Frauen mit einem negativen Testergebnis tatsächlich nicht an Brustkrebs erkrankt waren. Hier gilt: Je niedriger die Prävalenz, desto höher ist der NPV. Um sich diese Zusammenhänge einzuprägen, wiederholen Sie am besten die Berechnungen in Tab. 2.7 mit einer Prävalenz von 20 %.

Eine ausführliche Diskussion über die Vor- und Nachteile von Screening-Untersuchungen finden Sie in Kap. 4.5.

2.3.8 Statistische Signifikanz und p-Wert

Oft liest man in wissenschaftlichen Zeitschriften, dass die Resultate einer Studie „statistisch signifikant" seien. Um zu erläutern, was damit gemeint ist, betrachten wir die Resultate einer im Jahr 2010 im englischen Medizinjournal *The Lancet* veröffentlichten randomisierten Studie. Die Studie untersuchte, ob eine einmalige Sigmoidoskopie (= endoskopische Untersuchung des Enddarms einschließlich der S-förmigen Grimmdarmschlinge) bei klinisch gesunden Personen im Alter von 55 bis 64 Jahren die Darmkrebs-Sterblichkeit in den nächsten 11 Jahre reduziert. Als Vergleichsgruppe dienten Gleichaltrige, bei denen keine solche Untersuchung durchgeführt wurde. Die Studienautoren untersuchten neben der Darmkrebs-Sterblichkeit auch die Gesamtsterblichkeit in beiden Gruppen (Tab. 2.9). Die Berechnungen ergaben, dass die Darmkrebs-Sterblichkeit in der Sigmoidoskopie-Gruppe im Vergleich zu Kontrollgruppe um 31 % gesenkt werden konnte. Das relative Risiko (RR) betrug 0,69 bei einem 95 % VI von 0,59 bis 0,80. Die Gesamtsterblichkeit sank dadurch nach Angaben der Autoren um 3 % (RR: 0,97; 95 % VI: 0,95; 1,00).

Tab. 2.9: Randomisierte Studie zur Wirksamkeit einer einmaligen Sigmoidoskopie als Mittel der Darmkrebs-Früherkennung: Zahl der Todesfälle insgesamt sowie der Darmkrebs-Todesfälle in beiden Gruppen während eines Zeitraums von etwa 11 Jahren (Resultate übernommen aus Lancet 2010; 375: 1624–33, Tab. 1).

	Gruppe mit einmaliger Sigmoidoskopie	Kontrollgruppe ohne Sigmoidoskopie	Relatives Risiko (95 % VI)	p-Wert
Darmkrebs-Todesfälle	221	637	0,69 (0,59; 0,80)	< 0,0001
Alle Todesfälle	6.775	13.768	0,97 (0,95; 1,00)	0,052
Gesamtzahl der Personen	57.099	112.939		

95 %-VI: 95 %-Vertrauensintervall

Das *relative Risiko* (RR) vergleicht die Gruppe, bei deren Mitgliedern jeweils eine einmalige Sigmoidoskopie durchgeführt wurde, mit der Kontrollgruppe (s. Kap. 2.1.3).

Die Autoren veröffentlichten zusätzlich den so genannten **p-Wert**, der in der englischen Terminologie als „p-value" bezeichnet wird. Auch der p-Wert ist eine *bedingte* Wahrscheinlichkeit. Für Studien, die die Wirksamkeit einer bestimmten Intervention untersuchen, wird in der Regel als Bedingung die so genannte „**Null-Hypothese**" gewählt. Bei dieser Hypothese geht man davon aus, dass die Intervention keine Wirkung hat.

Unter dieser Annahme wird nun die Wahrscheinlichkeit berechnet, dass der tatsächlich beobachtete oder ein noch größerer Unterschied rein zufällig zustande gekommen sind.

Bei der Sigmoidoskopie-Studie sagt der p-Wert von < 0,0001 für die Darmkrebs-Sterblichkeit folgendes aus: Unter der Annahme, dass eine einmalige Sigmoidoskopie die Darmkrebs-Sterblichkeit bei den untersuchten Personen nicht reduziert – das relative Risiko also 1 ist –, ist die Wahrscheinlichkeit kleiner als 1 zu Zehntausend, ein relatives Risiko von ≤ 0,69 oder ≥ 1,45 (= 1/0,69) rein zufällig zu beobachten. Bei der Analyse der Gesamtsterblichkeit nennen die Autoren einen p-Wert von 0,052. Dies bedeutet analog, dass unter der Annahme, die einmalige Sigmoidoskopie reduziere die Gesamtsterblichkeit bei den untersuchten Personen nicht, eine Wahrscheinlichkeit von 5,2 % besteht, ein relatives Risiko von ≤ 0,97 oder ≥ 1,03 (1/0,97) zu beobachten, das rein durch Zufall zustande gekommen ist. Diese p-Werte werden „zweiseitig" genannt, weil hier die Entfernung zum Null-Wert, der für „keine Wirksamkeit" steht, sowohl nach oben („Nutzen durch Behandlung") als auch nach unten („Schaden durch Behandlung") betrachtet wird.

Die Berechnung des p-Wertes

Der **p-Wert** lässt sich unter Verwendung der *Standard-Normalverteilung* in drei Schritten berechnen.

- Zuerst berechnet man die Distanz zwischen dem Studien-Wert für die Wirksamkeit einer Methode und der Null-Hypothese, d. h. dem Wert, der „keine Wirksamkeit" beschreibt. Wenn *relative Risiken* (RR) betrachtet werden, müssen die Berechnungen auf der logarithmischen Skala durchgeführt werden. Für die Gesamtsterblichkeit bei der betrachteten Sigmoidoskopie-Studie berechnet sich dies aus ln(0,97332) – ln(1). Als Ergebnis erhalten wir – 0,02704.
- Der Absolutbetrag dieses Resultates wird anschließend durch den Standardfehler dividiert. Auch hier muss bei relativen Risiken die logarithmische Skala verwendet werden. Berechnet man den Logarithmus des Standardfehlers SE(ln(RR)), so ergibt dies 0,01392. Teilt man nun 0,02704 durch 0,01392, erhält man den Wert 1,942529. Dieser Wert wird als **Z-Wert** zur Berechnung des p-Wertes bezeichnet.
- In einem dritten Schritt wird nun berechnet, welcher Prozentsatz der Werte bei der Standard-Normalverteilung weiter als Z von Null entfernt liegt. In unserem Beispiel bedeutet dies: Welcher Prozentsatz ist kleiner/größer als der errechnete Z-Wert, d. h. kleiner als – 1,942529 oder größer als 1,942529? Wir wissen, dass bei der Standard-Normalverteilung genau 5 % aller Werte außerhalb von ±1,96 liegen. Also erwarten wir etwas mehr als 5 %. Die genaue Berechnung ergibt 5,2 %.
  Auch für die Differenz der Mittelwerte aus zwei Behandlungsgruppen lässt sich analog ein p-Wert berechnen. In die Berechnung des p-Wertes fließt also der *Standardfehler* und dadurch auch die *Größe der Studie* mit ein.

Dualität zwischen 95%-Vertrauensintervall und „statistischer Signifikanz"

Es hat sich eingebürgert, dass p-Werte, die kleiner als 0,05 sind, als „statistisch signifikant" bezeichnet werden. Die Wahl der 0,05-Grenze hat den Vorteil, dass eine Du-

alität zwischen dem 95%-Vertrauensintervall und der „statistischen Signifikanz“ besteht. In den Fällen, in denen das 95%-Vertrauensintervall den Wert für „keine Wirksamkeit“ ausschließt, ist der p-Wert kleiner als 0,05 und damit das Resultat „statistisch signifikant“ (und umgekehrt).

Dies sehen wir z. B. bei den Darmkrebstodesfällen in der Sigmoidoskopie-Studie. Das 95%-Vertrauensintervall für das relative Risiko reicht von 0,59 bis 0,80 und schließt damit den Wert 1 (= „keine Wirksamkeit“) klar aus. Entsprechend ist der p-Wert deutlich kleiner als 0,05. Dagegen reicht das 95%-Vertrauensintervall für das relative Risiko bei der Gesamtsterblichkeit von 0,95 bis 1,00. Es berührt also den Wert 1, der für „keine Wirksamkeit“ steht. Aufgrund der Dualität zwischen dem 95%-Vertrauensintervall und „statistischer Signifikanz“ sollte hier der p-Wert bei 0,05 (= 5 %) liegen. Gibt man das Resultat mit mehr als zwei Stellen nach dem Komma an, sieht man, dass das obere Ende des 95%-Vertrauensintervalls 1,00024 beträgt. Es schließt also die 1 noch knapp mit ein. Damit muss der p-Wert etwas größer als 5 % sein.

2.3.9 Statistische Signifikanz und klinische Relevanz

Statistische signifikante Resultate sind nicht zwingend auch klinisch relevant. In Tab. 2.10 sind die hypothetischen Resultate von drei randomisierten plazebokontrollierten Studien zur Senkung des LDL-Cholesterins im Blut dargestellt. In allen drei Studien wurden die TeilnehmerInnen zufällig entweder derjenigen Gruppe zugeteilt, in der sie das neue Medikament (A, B oder C) erhielten oder der Plazebo-Gruppe. Dort wurde ihnen statt des zu testenden Medikaments ein Scheinmedikament (*Plazebo*) verabreicht. Nach einer Behandlungsdauer von 3 Monaten wurde in allen Gruppen der Blutspiegel des LDL-Cholesterins gemessen. Daraus wurden nun die Mittelwerte pro Behandlungsgruppe sowie die Differenzen der Mittelwerte berechnet (Spalte 4 von Tab. 2.10). Anschließend wurde der Standardfehler für die Differenz von Mittelwerten (Spalte 5) und die Grenzen des 95%-Vertrauensintervalls (Spalte 6) ermittelt. Zum Schluss wurden der Z-Wert sowie der p-Wert berechnet (Spalten 7 und 8).

Tab. 2.10: Hypothetische Resultate von drei plazebokontrollierten, randomisierten Studien zur Senkung des LDL-Cholesterins im Blut.

Studie	Medikament	Anzahl der Patienten pro Gruppe	Differenz der Mittelwerte des LDL-Cholesterins (mg/dl) zwischen der Medikamenten-Gruppe und der Placebo-Gruppe	Standardfehler für die Differenz der Mittelwerte des LDL-Cholesterins	Grenzen des 95%-Vertrauensintervalls für die Differenz der Mittelwerte des LDL-Cholesterins	Z-Wert	p-Wert
1	A	40	–20	33	–84,7 bis 44,7	–0,606	0,544
2	B	4000	–2	3,3	–8,47 bis 4,47	–0,606	0,544
3	C	5000	–5	2	–8,92 bis –1,08	–2,5	0,012

Interessanterweise ergaben die Berechnungen sowohl für Medikament A als auch für Medikament B den gleichen p-Wert von 0,544. Die Resultate sind also beide statistisch

nicht signifikant. Das erstaunt nicht. In beiden Studien ist zu sehen, dass der Wert 0 deutlich im 95%-Vertrauensintervall enthalten ist. Betrachtet man die Grenzen des 95%-Vertrauensintervalls von Studie 1, fällt auf, dass der Behandlungseffekt hiermit nicht sinnvoll eingegrenzt wurde. Er liegt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 84,7 mg/dl und einer Erhöhung um 44,7 mg/dl. Da in Studie 1 nur 40 Patienten pro Gruppe untersucht wurden, überrascht dieses unpräzise Resultat nicht. Es sind also größere Studien notwendig, um die Wirksamkeit von Medikament A abzuklären. In Studie 2 umfasste jede Gruppe 4.000 Personen. Die Wirksamkeit von Medikament B konnte recht präzise quantifiziert werden. Sie liegt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 8,47 mg/dl und einer Erhöhung um 4,47 mg/dl. Eine klinisch relevante Senkung um ≥10 mg/dl ist daher sehr unwahrscheinlich. Bei Medikament C liegt mit einen p-Wert von 0,012 ein *statistisch signifikanter* Behandlungseffekt vor. Der Wert 0 ist nicht im 95%-Vertrauensintervall enthalten. Betrachtet man die Grenzen des 95%-Vertrauensintervalls, so liegt der Behandlungseffekt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 8,92 mg/dl und einer Senkung um 1,08 mg/dl. Damit liegt zwar eine „statistisch signifikante" Senkung vor, aber auch hier ist eine Senkung um ≥10 mg/dl eher unwahrscheinlich.

Um alle drei Studien abschließend beurteilen zu können, ist es wichtig zu wissen, welches Ausmaß einer Senkung des LDL-Cholesterinspiegels im Blut klinisch relevant ist. Geht man davon aus, dass dies erst bei einer Senkung um mindestens 10 mg/dl der Fall ist, dann ist auch Medikament C nicht geeignet, da es ja nur mit einer sehr geringen Wahrscheinlichkeit eine solche Senkung erreicht. Es zeigt sich, dass die Information, ob ein Behandlungseffekt statistisch signifikant ist, allein nicht ausreicht, um die klinische Relevanz der Resultate einer Studie beurteilen zu können. Die Information des 95%-Vertrauensintervalls ist hier wesentlich nützlicher. Wir erhalten einen 95 %-Wahrscheinlichkeitsbereich für den Behandlungseffekt und können daraus auch ableiten, ob der Behandlungseffekt in einem Bereich liegt, der klinisch relevant ist.

Internet-Ressourcen

Auf unserer Lehrbuch-Homepage (**www.public-health-kompakt.de**) finden Sie die Formeln für die Berechnungen in Tab. 2.10, Hinweise auf weiterführende Literatur, zusätzliche Tabellen, Abbildungen und Boxen sowie Links zu frei verfügbarer Software zur Datenanalyse.