Studien richtig lesen

Teil 2: Hypothesen, p-Wert, KI und Co.

Unsicherheiten, Variabilität und Fehler sind unvermeidlicher Bestandteil der medizinischen Wissenschaft. Um trotzdem zu einigermaßen verlässlichen Ergebnissen zu kommen, muss man dies in der Auswertung von Studiendaten berücksichtigen.

Messungenauigkeiten sind erst der Anfang Mathematik gilt als überaus exakte Wissenschaft – und ist es auch, solange sie in ihren eigenen Gefilden bleibt und z. B. den Satz des Pythagoras beweist. In der Medizin dagegen können mathematisch komplizierte Berechnungen und formal sehr genaue Zahlenangaben eine Exaktheit suggerieren, die es hier nie geben wird. In medizinischen Studien berechnet man nämlich nicht abstrakte Größen oder geometrische Figuren, sondern untersucht echte Menschen mit konkreten Messgeräten unter realen Bedingungen.

Schon bei so etwas Einfachem wie der Bestimmung der Körpergröße kommt jede Menge Unsicherheit ins Spiel: Steht der Proband gerade? Wie gut ist das Maßband geeicht, und wie genau ist es ablesbar? Wie konstant ist die Körpergröße im Tagesverlauf und je nach vorhergehender Tätigkeit? Da überrascht es nicht, dass man in der Praxis meist auf Zentimeter rundet und die Größe nicht z. B. mit 1,7265 m angibt. Einen Unterschied von 0,3 mm zwischen 2 Personen würde man vermutlich nicht als relevant ansehen, sondern im Rahmen der Messungenauigkeit erwarten. Misst man dagegen 1 cm Unterschied oder bleiben die 0,3 mm über mehrere Messungen erhalten, ist man irgendwann geneigt, einen „wahren" Unterschied anzunehmen. Aber wann wird aus „vermutlich gleich" „vermutlich unterschiedlich"? Die medizinische Statistik versucht, derartige Variabilität, Unsicherheiten und Messfehler mathematisch zu beschreiben und mit gewissen Parametern fassbar zu machen (Mittelwert, Standardabweichung etc.). Abschaffen lässt sich die Variabilität und damit die Unsicherheit aber nicht, weshalb die Medizin nie so exakte Beweise führen kann wie die Mathematik – auch wenn die Ergebnisse mit noch so vielen Nachkommastellen aufwarten. Betrachten wir nacheinander die wichtigsten „Probleme" und wie die Statistik sie angeht.

Die (Standard-)Normalverteilung

Die Normalverteilung wird häufig verwendet für die Beschreibung von Messwerten in einer Stichprobe, die zufällig um einen Mittelwert schwanken. Der Graph ihrer Dichtefunktion ist als Gauß'sche Glockenkurve bekannt (► Abb. 1). Ihre Lage ist bestimmt durch den Mittelwert, ihre Form durch die Standardabweichung. Auch wenn sich bei realen Daten selten eine ideale Glockenform ergibt, nimmt man oft eine Normalverteilung an (nämlich dann, wenn die Stichprobendaten symmetrisch um den Mittelwert verteilt sind und die Verteilung eingipfelig ist). Um einen Unterschied zwischen den Mittelwerten zweier Stichproben / Gruppen feststellen zu können, hilft es, wenn

- der „wahre" Unterschied möglichst groß ist (d. h. die Glockenkurven weiter auseinander liegen) und
- die Streuung der Werte innerhalb der Gruppen möglichst gering ist (die Glockenkurven möglichst schmal sind).

Neben der Normalverteilung verwendet man zur Stichproben-Beschreibung u. a. die Binomialverteilung (z. B. für Komplikationsraten) oder die Exponentialverteilung (z. B. für atomaren Zerfall).

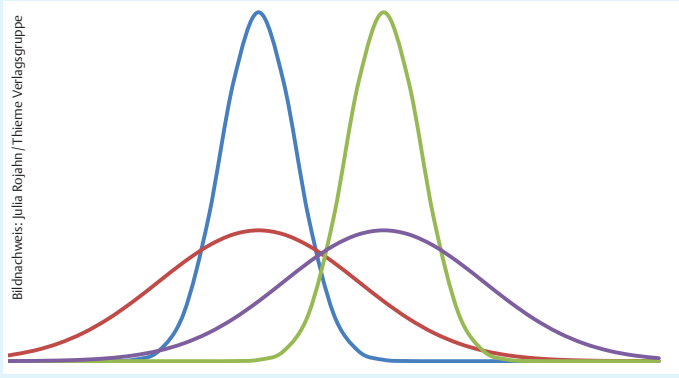

Abb. 1 Fiktive Normalverteilungen von 4 verschiedenen Stichproben. Blau und rot sowie grün und lila haben jeweils den gleichen Mittelwert, blau und grün sowie rot und lila jeweils die gleiche Standardabweichung. Blau und grün sind gut unterscheidbar, rot und lila dagegen kaum: Ihre Mittelwerte liegen genauso weit auseinander, die Streuung ist aber größer. Entsprechend wären auch rechnerisch die Unterschiede zwischen blau und grün besser nachzuweisen als zwischen rot und lila.

1. Problem: Die Menschen sind verschieden

Skalen von Messwerten, Maßzahlen für Lage und Streuung Bevor man anfängt zu rechnen, sollte man sich klar werden, welche Art von Daten man vor sich hat:

- Sind sie binär, nominal, ordinal oder kardinal skaliert?
- Kann ich sie z. B. anhand des Mittelwerts beschreiben, oder ist der Median besser geeignet?

Aus Platzgründen finden Sie eine Übersicht hierzu im Online-Zusatzmaterial zu diesem Beitrag in ► Tab. e1 und e2.

Beispiel: Körpergröße in Schulklassen Wären alle Menschen exakte Klone mit gleichen Umweltbedingungen, würde es evtl. reichen, den Effekt einer Behandlung oder die Güte eines diagnostischen Tests an nur einer Person zu testen. In der Realität muss man stets mehrere Probanden einbeziehen, um einigermaßen verlässliche („repräsentative") Ergebnisse zu erhalten. Am Beispiel der Körpergröße könnte man etwa alle Schüler einer Klasse messen und den Mittelwert („Durchschnitt", arithmetisches Mittel) bilden. Mit diesem kann man weiterrechnen und ihn z. B. mit dem Mittelwert in anderen Klassen vergleichen.

Oft will man auch beschreiben, wie weit die Einzelwerte um den jeweiligen Mittelwert streuen. Das wäre wichtig, falls Sie z. B. Kinderkleidung herstellen und auch die außergewöhnlich großen und kleinen berücksichtigen wollen. Nimmt man an, dass die Körpergröße normalverteilt ist (► Infokasten oben), ist das übliche Streuungsmaß die Standardabweichung (SD, ► Tab. e2 online), d. h. die mittlere Streuung der Messwerte um den Mittelwert. Es gilt:

- 95 % der Messwerte liegen im Bereich Mittelwert ± 1,96 SD
- 99 % der Messwerte liegen im Bereich Mittelwert ± 2,58 SD

Tab. 1 Mittelwerte und Standardabweichung der Körpergrößen in 3 Schulklassen (Normalverteilung unterstellt)

	Klasse 3a in Bremen	Klasse 3a in Leipzig	Klasse 3a in Oslo
Stichprobengröße (n)	25	19	31
Mittelwert	1,34 m	1,41 m	1,43 m
Standardabweichung (SD)	0,15 m	0,21 m	0,09 m
Bereich, der geschätzt 95 % der Werte enthält (Mittelwert ± 1,96 SD)	1,05–1,63 m	1,00–1,82 m	1,25–1,61 m

Zur Beschreibung einer Gruppe von Probanden verwendet man meist ein Maß für die „mittlere Lage" des Parameters (z. B. Mittelwert) und die „Streuung" der Einzelwerte um ihn (z. B. Standardabweichung).

Der Einfachheit halber betrachten wir hier im Weiteren nur Mittelwert und Standardabweichung, außerdem gehen wir von einem kontinuierlichen Merkmal (Körpergröße) aus und unterstellen, dass es normalverteilt ist.

► Tab. 1 zeigt exemplarisch die Mittelwerte und Standardabweichungen der Körpergrößen verschiedener 3. Klassen. Daraus können wir ablesen,

- wie groß die Kinder in jeder Klasse im Durchschnitt waren (Bremen < Leipzig < Oslo) und
- wie stark die einzelnen gemessenen Größen um diesen Mittelwert streuen (Oslo < Bremen < Leipzig).
- Es fällt außerdem auf, dass der Unterschied der mittleren Körpergröße zwischen Bremen und Leipzig deutlich größer ist als zwischen Leipzig und Oslo.

Haben wir nun gezeigt, dass Drittklässler in Bremen generell kleiner sind als in Leipzig, und diese wiederum etwas kleiner als in Oslo? Natürlich nicht.

Tab. 2 Konfidenzintervalle für die Körpergrößen in 3 Schulklassen (Normalverteilung unterstellt, t-Werte aus [1])

	Klasse 3a in Bremen	Klasse 3a in Leipzig	Klasse 3a in Oslo
Stichprobengröße (n)	25	19	31
Mittelwert Körpergröße	1,34 m	1,41 m	1,43 m
Standardabweichung (SD)	0,15 m	0,21 m	0,09 m
Bereich, in dem geschätzt 95 % der Einzelwerte liegen (Mittelwert ± 1,96 SD)	1,05–1,63 m	1,00–1,82 m	1,25–1,61 m
95 %-KI für Mittelwert (gemäß T-Verteilung)	1,34 ± (2,0641 × 0,15/√25) = 1,34 ± 0,06, also KI: 1,28–1,40 m (t-Wert für n = 25 laut Tabelle: 2,0641)	1,41 ± (2,101 × 0,21/√19) = 1,41 ± 0,10, also KI: 1,31–1,51 m (t-Wert für n = 19 laut Tabelle: 2,101)	1,43 ± (2,042 × 0,09/√31) = 1,43 ± 0,03, also KI: 1,40–1,46 m (t-Wert für n = 31 laut Tabelle: 2,042)
99 %-KI für Mittelwert (gemäß T-Verteilung)	1,34 ± (2,797 × 0,15/5) = 1,34 ± 0,08, also KI: 1,26–1,42 m (t-Wert für n = 25 laut Tabelle: 2,797)	1,41 ± (2,878 × 0,21/√19) = 1,41 ± 0,14, also KI: 1,27–1,55 m (t-Wert für n = 19 laut Tabelle: 2,878)	1,43 ± (2,750 × 0,09/√31) = 1,43 ± 0,04, also KI: 1,39–1,47 (t-Wert für n = 31 laut Tabelle: 2,750)

2. Problem: Man betrachtet nur eine Stichprobe

Risiko der Verzerrung Die Ergebnisse, die man aus einer Messung gewinnt, sollen natürlich nicht nur für die jeweils untersuchten Probanden gelten: Sie sind eine Stichprobe, die man stellvertretend für die sog. „Grundgesamtheit" untersucht – also z. B. alle Patienten mit der gleichen Erkrankung, der gleichen Behandlung oder dem gleichen Risikofaktor. In unserem Beispiel betrachten wir die Klasse 3a einer Schule jeweils als Stichprobe für alle Drittklässler in der jeweiligen Stadt. Kann man nun einfach den Mittelwert aus der Stichprobe als Mittelwert aller Drittklässler setzen? Würden Sie sich als Kinderkleidungs-Hersteller blind auf diese Schätzung verlassen? Vermutlich nicht – wir könnten zufällig Klassen mit ungewöhnlich großen oder kleinen Kindern erwischt haben, die unsere Mittelwerte „verzerren". Wie kommen wir der „wahren" Körpergröße der Drittklässler in Bremen, Leipzig und Oslo näher? Sind die Unterschiede im Mittelwert nur „zufällig", verursacht durch Messungenauigkeiten und natürliche Variabilität?

Schon ohne zu rechnen kommt man zu dem Schluss, dass die Schätzung aus der Stichprobe umso verlässlicher / präziser ist,

- je größer die Stichprobe (mehr Schüler gemessen) und
- je homogener die Stichprobe (geringere Streuung der Messwerte).

Das Konfidenzintervall Rechnerisch wird diese Überlegung im sog. Konfidenzintervall (KI) berücksichtigt: Es beschreibt

- den Bereich, der den gesuchten, nicht genau bekannten „wahren" Parameter (hier: die wahre Körpergröße der Drittklässler) mit einer gewissen Wahrscheinlichkeit einschließt.

Anders formuliert: Das KI ist

- der Unsicherheitsbereich für die Schätzung eines (fixen, aber unbekannten) Parameters aus einer Stichprobe [2] bzw.
- der „Intervallschätzer" für einen Parameter der Grundgesamtheit (als Ergänzung zum „Punktschätzer", z. B. dem geschätzten Mittelwert) [3].
- Es gilt: Je größer und / oder homogener die Stichprobe, desto kleiner / schmaler das KI.

Allerdings muss man noch angeben, wie verlässlich das KI sein soll, also mit welcher Wahrscheinlichkeit es den wahren Wert enthalten soll. Meist wird hier 95 % als „Sicherheitsgrad" gesetzt, d. h. man lässt eine Irrtumswahrscheinlichkeit von 5 % zu. In Worten lässt sich das so ausdrücken:

- Man erhält mit einer Wahrscheinlichkeit von 95 % ein KI, das den gesuchten Parameter enthält [2].
- Oder anders formuliert: Wenn man unter identischen Bedingungen 100 verschiedene Stichproben des gleichen Umfangs aus einer Grundgesamtheit zieht und daraus das KI ermittelt, würde der zu schätzende Parameter 95-mal innerhalb des KI liegen [4, 5].
- 95 % aller auf Grundlage von gemessenen Daten berechneten KIs beinhalten den wahren Wert der zu untersuchenden Population.

Die 95 % Sicherheitsgrad (auch: Überdeckungswahrscheinlichkeit oder Vertrauensbereich) sind aber reine Konvention – man kann auch 90 oder 99 % festsetzen.

Das Konfidenzintervall (Konfidenzbereich, Vertrauensbereich, engl.: confidence interval) ist ein Unsicherheitsbereich für die Schätzung eines Parameters aus einer Stichprobe. Es ist umso enger, je größer und homogener die Stichprobe und je kleiner der gewünschte Sicherheitsgrad (üblich sind 95 %). Konfidenzintervalle mit einem Sicherheitsgrad von 99 % sind breiter, solche mit einem Sicherheitsgrad von 90 % schmaler als bei einem Sicherheitsgrad von 95 %.

Die (Student) t-Verteilung

Von William Sealy Gosset unter dem Pseudonym „Student“ eingeführt, beschreibt die t-Verteilung die standardisierte Schätzfunktion des Stichproben-Mittelwerts normalverteilter Daten, wenn die Varianz des Merkmals unbekannt ist und mit der Stichprobenvarianz geschätzt werden muss [6]. Der Graph ähnelt dem der Normalverteilung, allerdings ist die t-Verteilung zusätzlich abhängig von der Stichprobengröße:

- Bei kleinen Stichproben ist die t-Verteilung breiter und flacher als die Normalverteilung,
- bei größeren Stichproben nähert sie sich der Normalverteilung an bzw. geht in sie über.

Hypothesentests mithilfe der t-Verteilung bezeichnet man als t-Tests (s. unten).

Beispiele für Konfidenzintervalle (KIs)

Kaufkraft einer Stichprobe von 30 Haushalten [7]:
- mittlere Ausgaben monatlich 850 €, SD 400 €
- 95 %-KI für Kaufkraft der Grundgesamtheit: 700–1000 €

Unterschied im Gewichtsverlust bei 5,5 vs. 8,5 h Schlaf (dieselben Personen, d. h. abhängige Stichproben) [8]:
- geschätzter Unterschied der Mittelwerte: 0,47 kg
- 95 %-KI: -0,31–1,28 (da die Null eingeschlossen ist, ist es unsicher, ob wirklich ein Unterschied besteht)

Korrelation (Zusammenhang) zwischen Zahl der Störche und der Neugeborenen über 10 Jahre in Chemnitz [8]:
- geschätzter Korrelationskoeffizient r = 0,38
- 95 %-KI für Korrelation: -0,3–0,8 (da die Null eingeschlossen ist, ist Zusammenhang unsicher)

Körpergewicht von 10 Personen vor und nach Diät [6]:
- Mittelwert sinkt von 93,9 auf 91,2 kg
- 95 %-KI für Gewichtsverlust: 0,3–5,1 (knapp signifikant, Ausmaß des Gewichtsverlusts aber sehr unklar)

Metaanalyse zur Prostatakrebs-Früherkennung, Gesamtmortalität mit vs. ohne Screening [9]:
- geschätztes relatives Risiko: 0,99
- 95 %-KI: 0,96–1,03 (Da die 1 eingeschlossen ist, kann man weder eine Senkung der Gesamtmortalität um 4 % noch eine Erhöhung um 3 % durch das Screening ausschließen.)

Berechnung des KI In die Berechnung des KI fließen ein:
- der aus der Stichprobe geschätzte Wert für den gesuchten Parameter (hier: Mittelwert der Körpergrößen)
- die Streuung der Einzelwerte (hier: Standardabweichung)
- der Stichprobenumfang n
- ein Faktor zur Berücksichtigung der festgesetzten Wahrscheinlichkeit

Für unser Beispiel der Körpergrößen lässt sich das KI recht einfach berechnen. Wir können von annähernd normalverteilten Werten ausgehen, für deren Mittelwert das KI berechnet werden soll. Dafür benötigen wir neben Stichprobengröße, Mittelwert und SD noch einen Faktor, der die Verteilung und den gewünschten Sicherheitsgrad berücksichtigt. Dies ist hier der t-Wert aus der sog. Student t-Verteilung (▶ Infokasten oben). Dann gilt für das KI (für den Mittelwert eines normalverteilten Parameters):
- untere Grenze des KI = Mittelwert minus (t-Wert * SD/√n)
- obere Grenze des KI = Mittelwert plus (t-Wert * SD/√n)

Beispiel Schulklassen ▶ Tab. 2 zeigt die Fortführung von ▶ Tab. 1 mit Berechnung der 95 %- und 99 %-KIs. Es zeigt sich:
- Die wahre mittlere Körpergröße liegt mit einer Wahrscheinlichkeit von 95 bzw. 99 %
  - in Bremen im Bereich 1,28–1,40 m bzw. 1,26–1,42 m,
  - in Leipzig im Bereich 1,31–1,51 m bzw. 1,27–1,55 m und
  - in Oslo im Bereich 1,40–1,46 m bzw. 1,39–1,47.
- Die KIs sind in Oslo am schmalsten, in Leipzig am breitesten – wie schon anhand der Standardabweichungen zu vermuten.
- Die 99 %-KIs sind wesentlich breiter als die 95 %-KIs.
- Die 95 %-KIs sind deutlich enger als der Bereich, in dem 95 % der Einzelwerte liegen.

Neu im Vergleich zu ▶ Tab. 1 ist:
- Die 95 %-KIs in Bremen und Leipzig überschneiden sich. Man kann daher (bei diesem Sicherheitsgrad) nicht schließen, dass die Drittklässler unterschiedlich groß sind (d. h. verschiedenen Grundgesamtheiten entstammen) – aber auch nicht, dass sie gleich groß sind (d. h. der gleichen Grundgesamtheit entstammen). Entsprechendes gilt für Leipzig und Oslo.
- Wir können also (bei diesem Sicherheitsgrad) lediglich schließen, dass die Drittklässler in Oslo größer sind als in Bremen.
- Bei einem Sicherheitsgrad von 99 % überschneiden sich sogar alle KIs, sodass wir nicht sicher sagen können, dass sich die „wahren“ Körpergrößen unterscheiden.

Weitere Merkmale des KI/KIs für andere Parameter Bei kontinuierlichen Merkmalen (▶ Tab. e1) wie der Körpergröße liegt das KI für den Mittelwert symmetrisch um den zentralen Schätzwert. Man kann KIs aber auch für andere Parameter berechnen (▶ Infokasten oben), wo sie z. T. asymmetrisch sind [3].

KI und statistische Signifikanz Weicht eine Messgröße so stark von einer anderen oder vom erwarteten Wert ab, dass dieser Unterschied nur sehr unwahrscheinlich rein zufällig bedingt ist, spricht man von einem statistisch signifikanten („überzufälligen“) Unterschied. Als Grenze für „sehr unwahrscheinlich" wird auch hier meist 5 % angesetzt („Signifikanzniveau“). Ob das Ergebnis eines statistischen Tests signifikant ist, entscheidet sich am p-Wert (s. unten).

Statistisch signifikante Unterschiede sind solche, die nur sehr unwahrscheinlich durch Zufall zustande kämen. Wie beim KI gilt meist 5 % als „Unwahrscheinlichkeits“-Grenze.

Was KI und Signifikanz nicht aussagen Grundsätzlich wird die Breite des KIs nur durch Zufallsfehler der Stichprobe, nicht aber durch systematische Erhebungs-, Messfehler etc. bestimmt [3].

Außerdem liefert auch eine signifikante Abweichung keine kausale Erklärung für die beobachteten Unterschiede: Wir können lange spekulieren, warum Drittklässler in Oslo größer sein sollten als in Bremen, wie die 95 %-KIs andeuten:

- Liegt es am geografischen Breitengrad,
- an der Luftverschmutzung,
- am Durchschnittseinkommen,
- am Alkoholkonsum der Eltern oder
- an der Qualität der Kleinkindbetreuung?

Mit viel Rechenaufwand findet man hier vermutlich den einen oder anderen Zusammenhang – seriös und aussagekräftig wären die Schlussfolgerungen trotzdem nicht. Vielleicht gibt es in Wahrheit gar keinen Unterschied, und unsere Stichproben gehören nur zu den 5 %, die per vorab festgelegter Irrtumswahrscheinlichkeit zu einem KI führen, das den „wahren" Wert nicht enthält.

3. Problem: Man muss entscheiden, ob beobachtete Unterschiede „echt" sind

Grundidee: Testen von Hypothesen mit statistischen Mitteln Um Ursache-Wirkungs-Beziehungen mit ausreichender Sicherheit nachzuweisen, sind – neben biologisch halbwegs plausiblen Erklärungen – sog. randomisiert-kontrollierte Studien (RCTs) sinnvoll, in denen man gezielt nur einen Parameter verändert (▶ Studien richtig lesen Teil 1 [10]). Um die Frage der Körpergrößen zu lösen, könnte man daher z. B. 1000 in Bayern geborene Babies im Rahmen einer RCT zufällig auf Bremen und Oslo verteilen, alle sonstigen Einflussgrößen möglichst gleich halten und Jahre später die Körpergrößen messen. Würde man dann signifikante Unterschiede messen, könnte man annehmen, dass die Unterschiede „echt" sind und der Breitengrad tatsächlich die Körpergröße beeinflusst – andernfalls nicht. Dieses Vorgehen nennt man in der medizinischen Statistik

- Testen von Hypothesen oder
- Signifikanztests.

Kurz gefasst geht es darum, eine Hypothese zur Zufälligkeit beobachteter Abweichungen zu formulieren, die man mithilfe der erhobenen Daten statistisch überprüft. Da dies nur an Stichproben möglich ist, gibt es immer eine gewisse Irrtumswahrscheinlichkeit [7].

Bei jedem statistischen Test sind folgende Schritte notwendig:

Schritt 1: Hypothesen formulieren Um näher an den üblichen Fragestellungen medizinischer Studien zu bleiben, verabschieden wir uns an dieser Stelle vom Beispiel der Körpergrößen und betrachten einen Standardfall: Patienten mit Bluthochdruck erhalten in einer RCT entweder ein Plazebo oder ein Medikament, und wir fragen, ob sich die beiden Substanzen bezüglich ihrer Wirkung unterscheiden. Dabei wissen wir: Die beiden Stichproben (Plazebo bzw. Medikament) werden sich schon allein aufgrund von individueller Variabilität, Messungenauigkeiten etc. in ihren Mittelwerten und Standardabweichungen unterscheiden. Trotzdem kann es sein, dass das Medikament in Wahrheit (wenn man nicht nur eine Stichprobe, sondern alle potenziellen Patienten untersuchen würde) nicht besser wirkt als Plazebo – mathematisch ausgedrückt: dass beide Stichproben aus der gleichen Grundgesamtheit stammen. Wir formulieren

- die Nullhypothese (H_0): es gibt keinen Unterschied zwischen Plazebo und Medikament, Abweichungen sind zufällig und
- die Alternativhypothese (H_A): das Medikament senkt den Blutdruck stärker als Plazebo.

H_0 vermutet normalerweise Gleichheit, H_A Ungleichheit. Die Studie zielt dann darauf ab, H_0 zu widerlegen, also z. B. den Vorteil eines neuen Medikaments zu zeigen. Ähnlich könnte man auch Fragen behandeln, ob z. B. Asthma in Peking häufiger vorkommt als in New York oder ob Asthmakranke häufiger per Sectio geboren wurden als Menschen ohne Asthma.

Die Hypothesen können:

- zweiseitig formuliert sein (H_A: Unterschied vorhanden)
- einseitig formuliert sein (H_A: in einer Gruppe ist das Merkmal kleiner / größer als in der anderen)

Entsprechend gibt es ein- und zweiseitige Testverfahren. In unserem Beispiel wäre beides möglich, hier wird zum besseren Verständnis die einseitige Fragestellung bevorzugt.

Die Hypothesen sollten vor Versuchsbeginn formuliert werden – spätestens aber vor Kenntnis der Daten. Im Zweifelsfall sollte man zweiseitig formulieren.

Schritt 2: Signifikanzniveau wählen Ähnlich wie oben für KIs beschrieben, muss man auch für Signifikanztests ein Signifikanzniveau (= Irrtumswahrscheinlichkeit) festlegen. Üblich sind auch hier 5 % bzw. 0,05. Für unser Beispiel bedeutet das: Falls das Medikament in Wahrheit gleichwertig zum Plazebo ist, soll der statistische Test nur dann einen Unterschied zeigen, wenn die beobachtete (oder eine größere) Differenz zwischen den Gruppen mit lediglich 5 % Wahrscheinlichkeit zufällig auftreten kann. Die-

Tab. 3 Mögliche Fehler bei Hypothesentests

	Nullhypothese ist wahr (z. B. kein Unterschied)	Alternativhypothese ist wahr (z. B. es gibt Unterschied)
Test fällt Entscheidung für Nullhypothese	richtige Entscheidung für Nullhypothese > Wahrscheinlichkeit, einen tatsächlich fehlenden Unterschied auch zu zeigen > Wahrscheinlichkeit, einen Fehler 1. Art zu vermeiden > Wahrscheinlichkeit: 1–α, also bestimmt durch definiertes Signifikanzniveau	Fehler 2. Art / Ordnung, Typ-II-Fehler, β-Fehler > Risiko, einen tatsächlich vorhandenen Effekt zu übersehen > Wahrscheinlichkeit: β, meist unbekannt, wird größer bei kleinerem α > angestrebt sind häufig 10–20 % > wird kleiner durch große, homogene Stichprobe und deutliche Unterschiede z. B. zwischen den Mittelwerten > vergleichbar bei diagnostischen Tests: falsch-negatives Ergebnis
Test fällt Entscheidung für Alternativhypothese	Fehler 1. Art / Ordnung, Typ-I-Fehler, α-Fehler > Risiko, einen Effekt festzustellen, wo keiner ist > Wahrscheinlichkeit: maximal α, also bestimmt durch definiertes Signifikanzniveau, häufig 5 % > „Fehlalarmwahrscheinlichkeit“ > Bsp.: Behandlung als wirksam beurteilt, obwohl sie unwirksam ist > vergleichbar bei diagnostischen Tests: falsch-positives Ergebnis	richtige Entscheidung gegen Nullhypothese, entspricht der Power > Wahrscheinlichkeit, einen tatsächlich vorhandenen Unterschied zu zeigen > Wahrscheinlichkeit, einen Fehler 2. Art zu vermeiden > Wahrscheinlichkeit: 1–β, meist unbekannt > Güte, Aussagekraft, „Ansprechwahrscheinlichkeit“, „Trennschärfe“ oder „Teststärke“ eines statistischen Tests > Wahrscheinlichkeit ist größer bei höherem α, größeren und homogeneren Stichproben sowie größerem Unterschied zwischen den Stichproben

ses wäre dann der sog. Ablehnungsbereich der Nullhypothese – im Gegensatz zum Annahmebereich mit 95 %.

Die – mehr oder weniger willkürlich festgelegte – bewusst in Kauf genommene Irrtumswahrscheinlichkeit bezeichnet man auch als Signifikanzniveau.

Eine Irrtumswahrscheinlichkeit von 5 % bedeutet allerdings auch:
> Im Schnitt wird 1 von 20 Untersuchungen, bei denen die Nullhypothese richtig ist (z. B. Medikament = Plazebo), zu dem Schluss kommen, sie sei falsch (und z. B. schließen, das Medikament ist besser als Plazebo).
> In 5 % der Fälle ist man bereit, H_0 abzulehnen, obwohl sie korrekt ist [7].
> Oder: Untersucht man gleichzeitig 20 verschiedene Endpunkte (z. B. Blutdruck in 2 Geschlechtern und je 10 Altersgruppen), wird durch Zufall einer davon einen Unterschied zeigen [5].

Teilweise wird daher auch 1 % oder sogar 0,1 % als Signifikanzniveau gefordert [11] – womit allerdings das Risiko für den komplementären β-Fehler steigt: nämlich bei der Nullhypothese zu bleiben, auch wenn in Wirklichkeit die Alternativhypothese zutrifft.

Fehler 1. und 2. Art, Power Mit jedem Hypothesentest sind 2 potenzielle Fehler verbunden:
> Fehler 1. Art / Ordnung, Typ-I-Fehler, α-Fehler: Man lehnt fälschlicherweise H_0 ab und nimmt H_A an.
> Fehler 2. Art / Ordnung, Typ-II-Fehler, β-Fehler: Man behält fälschlicherweise H_0 bei und nimmt H_A nicht an.

▶ Tab. 3 stellt diese Fehler und ihre Wahrscheinlichkeiten im Überblick dar. Wichtig ist:
> Während der Fehler 1. Art durch das Signifikanzniveau kontrolliert wird,
> ist der Fehler 2. Art meist nicht genau bekannt, da man den „wahren“ Wert (z. B. Unterschied der Mittelwerte) nicht kennt!

Grundsätzlich ist der β-Fehler umso kleiner – und damit die „Power“ der Studie umso größer –,
> je größer und homogener die Stichproben sind und
> je größer der Effekt (z. B. Unterschied) ist.

Eine zu geringe Power führt dazu, dass man einen echten Effekt nicht nachweisen kann. Meist strebt man für die Power 80 %, für den β-Fehler 20 % an. Der angenommene β-Fehler ist also größer als der α-Fehler – u. a., damit man nicht versehentlich eine unwirksame Behandlung etabliert [5]. Das heißt aber auch: Mit 20 % Wahrscheinlichkeit wird man einen vorhandenen Unterschied übersehen – und in der Praxis ist die Power oft noch geringer, weil z. B. nicht genug Patienten rekrutiert werden können [11].

Maximale Werte für die Fehler 1. und 2. Art (α- und β-Fehler) stehen somit vor der Datenanalyse fest: Den einen legt man fest (α), der andere ist durch Studiendesign und die Höhe des „wahren“ Effekts gegeben – seine Höhe aber nur grob abschätzbar (β).

Die Wahl des geeigneten α- und β-Fehlers in der Medizin heißt, zwischen Sicherheit und Machbarkeit abzuwägen:
> Ein niedriger α-Fehler bedeutet hohe Sicherheit, dass bei Annahme der Alternativhypothese dies auch zutrifft – andererseits ist es schwieriger, einen bestehenden Unterschied als solchen zu erkennen.
> Bei einem niedrigen β-Fehler ist die Power höher, sodass man auch kleine Unterschiede findet. Dazu braucht man aber große Stichproben [5].

Schritt 3: Testverfahren wählen und Prüfgröße berechnen Nun folgt der eigentliche Signifikanztest. Er ermittelt für die konkrete Fragestellung, ob die gemessenen Werte im Annahme- oder Ablehnungsbereich von H_0 liegen. Für unseren Medikamentenvergleich – wo wir von normalverteilten Daten ausgehen – wäre der sog. (Student) t-Test geeignet, mit der t-Verteilung (s. oben) als Prüfverteilung und einseitiger Fragestellung.

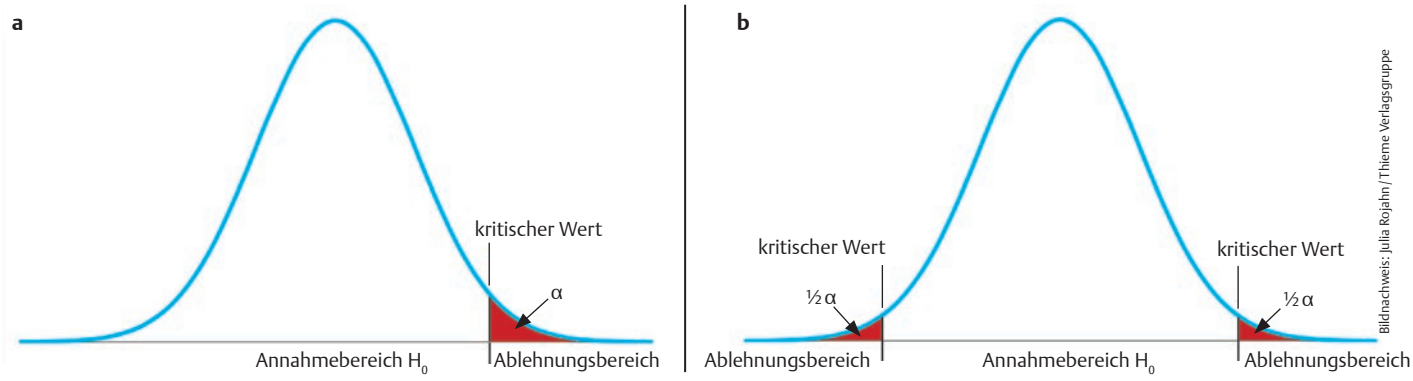

Abb. 2 Beispiel für einen Signifikanztest: t-Test mit Annahme- und Ablehnungsbereich der Nullhypothese H_0. a) einseitige Fragestellung; b) zweiseitige Fragestellung. α = Irrtumswahrscheinlichkeit / Signifikanzniveau des Tests, das im Voraus festzulegen ist (meist 5 %).

Angenommen, wir messen folgende Werte (nach [6], vgl. ▶ Tab. 4):
- Plazebo-Gruppe: n = 40, Mittelwert Blutdruck 181,63 mmHG, SD 6,41 mmHG
- Medikamenten-Gruppe: n = 35, Mittelwert Blutdruck 170,09 mmHG, SD = 5,24 mmHG

Aus diesen empirischen Daten berechnet man die sog. Prüfgröße:
$t = (\text{Mittelwert 1} - \text{Mittelwert 2}) / [\text{„mittlere SD“} * \sqrt{(1/n_1 + 1/n_2)}]$
Für unser Beispiel ergibt sich [6]: t = 8,46. Dies vergleicht man mit dem sog. „kritischen Wert“ („kritischen Schranke“) der t-Verteilung für die jeweilige Stichprobengröße und den gewählten α-Fehler. Die kritischen Werte trennen den Annahme- vom Ablehnungsbereich der Nullhypothese (▶ Abb. 2). Tabellen mit kritischen t-Werten findet man in Lehrbüchern oder online [1]. Im Beispiel wäre der kritische Wert 1,67 bei einem α-Fehler von 5 % und einseitiger Fragstellung. Damit kann eine Testentscheidung getroffen werden.

Schritt 4: Entscheiden Alle statistischen Tests sollen eine Frage beantworten: Liegt die Prüfgröße diesseits oder jenseits der kritischen Werte? Daraus leitet sich die Entscheidung für H_0 oder H_A ab:
- Liegt die (aus den Stichproben ermittelte) Prüfgröße innerhalb des (durch die Prüfverteilung und den gewählten α-Fehler definierten) Annahmebereichs, behält man H_0 bei (kein signifikanter Unterschied). Anders formuliert: Wenn die Prüfgröße in ihrem Betrag kleiner ist als der kritische Wert, ist das Testergebnis nicht signifikant.
- Liegt die Prüfgröße dagegen im kritischen Bereich, lehnt man H_0 ab und entscheidet sich für H_A (es gibt einen signifikanten Unterschied). Anders formuliert: Wenn die Prüfgröße betragsmäßig größer ist als der kritische Wert, ist das Testergebnis signifikant.

Das ist hier der Fall: Die Prüfgröße ist mit 8,46 deutlich größer als der kritische Wert. Damit liegt die Prüfgröße deutlich im Ablehnungsbereich von H_0, der Unterschied ist signifikant, wir würden H_A annehmen: Das Medikament wirkt besser als Plazebo.
▶ Tab. 4 zeigt weitere Beispiele für Hypothesentests.

Der eigentliche statistische Test vergleicht eine Prüfgröße (errechnet aus Versuchsdaten) mit einem kritischen Wert. Dieser ist gegeben durch den verwendeten Test und gewählten α-Fehler und definiert den Annahme- und Ablehnungsbereich der Nullhypothese. Bei zweiseitigen t-Tests gibt es 2 kritische Werte, die sich nur bezüglich ihres Vorzeichens unterscheiden.

Cave: Ein nicht signifikantes Ergebnis heißt aber nicht unbedingt, dass in Wahrheit kein Unterschied besteht! Vielleicht ist nur die Stichprobe zu klein oder heterogen, um ihn nachzuweisen.

Verschiedene Testverfahren möglich Neben dem t-Test gibt es eine ganze Reihe weiterer Tests. Welcher im konkreten Fall geeignet ist, hängt dabei vom Studiendesign ab, z. B. von [2]:
- der Zahl der Stichproben (1, 2 oder mehr als 2)
- der Abhängigkeit der Stichproben untereinander (unabhängig sind z. B. parallele Gruppen einer kontrollierten Studie, abhängig sind z. B. wiederholte Messungen an denselben Probanden)
- der Skala der Daten (binär, nominal, ordinal, kardinal; ▶ Tab. e1)
- der Verteilung des relevanten Parameters (Normal-, Exponential-, Binomial-, Poissonverteilung o. a.)

Häufig verwendete Tests sind [2, 6, 14]:
- (Student) t-Test (für in etwa normalverteilte Daten)
- Chi-Quadrat-Test (X^2-Test) für Häufigkeiten oder Wahrscheinlichkeiten, z. B. in Fall-Kontroll-Studien (siehe [10])
- Fisher's exakter Test (Fisher-Yates-Test, exakter X^2-Test für kleine Stichproben)
- Wilcoxon-Test als Alternative zu t-Test bei fehlenden Voraussetzungen (z. B. nicht stetige oder nicht normalverteilte Merkmale)
- Binomialtest (für dichotome Merkmale)
- F-Test (für Vergleich von Varianzen)

Cave: Jeder Test stellt gewisse Anforderungen an die erhobenen Daten. Sind diese nicht erfüllt, ist der Test streng genommen nicht korrekt und kann falsche Ergebnisse liefern. Probleme können z. B. entstehen, wenn die Stichproben aus Grundgesamtheiten mit

Tab. 4 Beispiele für Hypothesentests

	Schwangerschaftsdauer nach Medikament (verändert nach [15])	Geburtsgewicht von Babies nach mütterlichem Alkoholkonsum [6]	Blutdruck Medikament vs. Plazebo (verändert nach [6])
Fragestellung	Ändert ein Medikament die Schwangerschaftsdauer?	Haben Kinder nach mütterlichem Alkoholkonsum ein anderes Geburtsgewicht als der allgemeine Durchschnitt von 3500 g?	Senkt ein Medikament den Blutdruck gegenüber Plazebo?
Nullhypothese H_0	Abweichungen wären zufällig bedingt.	Abweichungen wären zufällig bedingt.	Der Unterschied wäre zufällig bedingt.
Alternativhypothese H_A	Das Medikament verkürzt oder verlängert die Schwangerschaft.	Die „Risikokinder" sind leichter oder schwerer als der Durchschnitt.	Das Medikament senkt den Blutdruck stärker als Plazebo.
Datenerhebung	16 Mütter mit Medikament entbinden nach durchschnittlich 280 d (SD: 8 d). Der allgemeine Durchschnitt beträgt 281,5 d.	Bei 20 „Risikokindern" messen Sie ein mittleres Geburtsgewicht von 3311,5 g (SD = 410,5 g).	Gruppe A (n = 40) unter Plazebo zeigt Mittelwert 181,63 mmHG (SD 6,41), Gruppe B (n = 35) unter Medikament 170,09 mmHG (SD = 5,24 mmHG).
Berechnung der Prüfgröße und Vergleich mit kritischen Werten für $\alpha = 5\,\%$	t-Test mit zweiseitiger Fragestellung: Die kritischen Werte sind -2,13 und + 2,13. Die Prüfgröße ist 0,75 und liegt damit (deutlich) im Annahmebereich von H_0.	t-Test mit zweiseitiger Fragestellung: Die kritischen Werte sind -2,093 und + 2,093. Die Prüfgröße ist -2,05 und liegt damit (knapp) im Annahmebereich von H_0.	t-Test mit einseitiger Fragestellung: Die Prüfgröße ist 8,46 und damit deutlich größer als der kritische Wert 1,67, d. h. sie liegt (deutlich) im Ablehnungsbereich.
Entscheidung	H_0 annehmen / beibehalten	H_0 annehmen / beibehalten	H_0 ablehnen / verwerfen, H_A annehmen
zusätzliche Angaben	95 %-KI: 276–284 d	p = 0,0541	p-Wert: p < 0,0001 95 %-KI für Unterschied: 8,82–14,26 mmHG

identischen Erwartungswerten, aber unterschiedlichen Varianzen stammen [2]. Mit RCTs kann man zwar meist eine strukturelle Gleichheit unter der Nullhypothese erwarten, bei anderen Studiendesigns kann das aber anders aussehen.

Und wo kommt der p-Wert ins Spiel? Erstaunlicherweise haben wir bisher den vermeintlich wichtigsten Begriff im Zusammenhang mit Signifikanztests noch gar nicht gebraucht: den p-Wert! Er kommt ins Spiel, seitdem man statistische Tests nicht mehr mit Taschenrechner und Tabellenwerken macht, sondern per Computer. Seitdem vergleicht (außer in Statistik-Klausuren) niemand mehr per Hand Prüfgrößen mit kritischen Werten, sondern man füttert ein Statistikprogramm mit allen nötigen Daten und bekommt als Ergebnis des Signifikanztests den p-Wert. Er gibt eine Wahrscheinlichkeit an (p für „probability") und liegt daher zwischen 0 und 1 [16]. Ein Vorteil des p-Werts: Man kann ihn nicht nur für Unterschiede zwischen Mittelwerten zweier Stichproben berechnen (wie hier), sondern auch für relative Risiken, Odds Ratios, Korrelationen etc. – obwohl in diesen Fällen ganz andere Teststatistiken nötig sind [13].
In Worten ausgedrückt ist der p-Wert:

- die Wahrscheinlichkeit, das gefundene Stichprobenergebnis (oder ein noch „extremeres") zu erhalten, wenn H_0 wahr ist
- die Wahrscheinlichkeit, dass ein zufälliger Versuch bei gültiger H_0 mindestens so „extrem" ausgeht wie der beobachtete Versuch
- die Wahrscheinlichkeit, dass ein gefundener Unterschied lediglich zufällig zustande gekommen ist [5]
- die Irrtumswahrscheinlichkeit, mit der man gerade noch H_0 widerlegen kann
- Für z. B. p = 3 %: Wenn in Wahrheit kein Effekt (z. B. kein Unterschied) vorliegt, kann die beobachtete oder eine größere Differenz mit 3 % Wahrscheinlichkeit zufällig auftreten.

Das heißt: Je kleiner der p-Wert,

- desto unwahrscheinlicher ist es, dass der gefundene Effekt zufällig zustande gekommen ist [5],
- desto mehr spricht das Ergebnis gegen H_0,
- desto „signifikanter" ist das Ergebnis bzw. der Unterschied.

Üblicherweise setzt man auch hier wieder 5 % Irrtumswahrscheinlichkeit (α-Fehler) als Grenze für „Signifikanz", d. h.

- wenn p < 5 %, gilt das Ergebnis als signifikant (zum Niveau 5 %),
- bei p ≥ 5 % gilt das Ergebnis als nicht signifikant [17].

Die enthaltene Information ist eigentlich die gleiche wie beim oben geschilderten Vorgehen für Hypothesentests: Ist $p < \alpha$, liegt das beobachtete Ergebnis im kritischen Bereich bzw. im Ablehnungsbereich von H_0. Auch für den p-Wert muss man vorher den korrekten Test wählen, den α-Fehler festsetzen etc. Der p-Wert zeigt aber etwas anschaulicher, wie weit z. B. die Prüfgröße im Ablehnungsbereich von H_0 liegt: In unserem Beispiel Medikament vs. Plazebo lag die Prüfgröße mit 8,46 offensichtlich „deutlich" über dem kritischen Wert, wir haben H_A angenommen. Aber was heißt „deutlich"? Als p-Wert ergibt sich: p < 0,0001 – hier kann man also genauer sagen: Der Unterschied Medikament vs. Plazebo ist so deutlich, dass er nur mit < 0,01 % Wahrscheinlichkeit zufällig auftreten würde. Ergäbe sich dagegen z. B. ein p-Wert von 0,03, könnte man die Nullhypothese auf einem Signifikanzniveau von 0,05 ablehnen, nicht aber von 0,01.
Cave: Die computerbasierte Berechnung des p-Werts erlaubt es im Prinzip, zuerst den p-Wert zu berechnen und dann das Signifikanzniveau so anzupassen, dass man ein „erwünschtes" Ergebnis erhält. Korrekt ist dann lediglich eine Formulierung wie: „Das Ergebnis ist signifikant zum Niveau 5 %, aber nicht zum Niveau 1 %."

Beispiel: p-Wert beim Würfeln

Vergleicht man ein Medikament gegen Plazebo, liegt es nah, einen positiven Effekt zu vermuten. Wie leicht man sich dabei durch einen passenden p-Wert täuschen lässt, zeigt folgendes Beispiel mit 1 scharzen und 1 weißen Würfel [11]: Sie nehmen an, der schwarze Würfel ist „besser" als der weiße, d. h. er würfelt höhere Zahlen. Sie würfeln je einmal (Stichprobe) und erhalten: schwarz 5, weiß 3, d. h. schwarz ist 2 Punkte besser als weiß. Listet man alle 36 möglichen Ergebnisse des Wurfs auf (von schwarz 1 weiß 6 bis schwarz 6 weiß 1), kann man ablesen, mit welcher Wahrscheinlichkeit man dieses oder ein „extremeres" Ergebnis erhält, falls beide Würfel gleich sind (p-Wert):

> In 10/36 (= 0,28) Fällen ist schwarz ≥ 2 Punkte besser als weiß, der p-Wert beträgt also 28 %. Mit dem üblichen Signifikanzniveau von 5 % würde man nach diesem Wurf verneinen, dass schwarz besser ist als weiß (Unterschied nicht signifikant).

Und wenn Sie zufällig schwarz 6 und weiß 1 gewürfelt hätten?

> Der p-Wert für dieses Ereignis (schwarz ≥ 5 Punkte besser als weiß) ist 1/36 = 0,028 = 2,8 %. Das ist weniger als die übliche Grenze von 5 %. Sie würden also nach dieser einen Stichprobe annehmen, dass der schwarze Würfel tatsächlich besser ist als der weiße (signifikanter Unterschied).

Sie könnten damit aber auch falsch liegen, die Würfel sind exakt gleich und das Ergebnis in diesem Wurf rein zufällig zustande gekommen! Beim Würfeln erscheint uns diese Einschränkung allerdings viel einsichtiger als bei medizinischen Studien.

Der p-Wert ist heutzutage die übliche Form, das Ergebnis eines Signifikanztests darzustellen. Er gibt an, wie „extrem" ein beobachtetes Ergebnis ist. Ist p kleiner als ein vorab gewähltes Signifikanzniveau (meist 5 %), gilt der gefundene Effekt als statistisch signifikant, man lehnt H_0 ab.

4. Problem: Man muss die Unterschiede interpretieren

Was der p-Wert nicht leistet Der p-Wert wird häufig falsch interpretiert. Man findet z. B. folgende Aussagen:

> Der p-Wert gibt an, wie wahrscheinlich H_0 oder H_A bei diesem Stichprobenergebnis ist. (FALSCH!)

> Für z. B. p = 3 %: Wenn ich H_A annehme, beträgt die Irrtumswahrscheinlichkeit 3 %. (FALSCH)

> Der p-Wert ist die Wahrscheinlichkeit, dass H_0 oder H_A wahr sind. (FALSCH!)

Denn: Es ist ja immer jeweils entweder H_0 oder H_A wahr, d. h. der im Experiment untersuchte Effekt ist entweder vorhanden oder nicht – man kann nur nicht immer anhand der Stichprobe entscheiden [18] (▶ Infokasten oben)! Eine Hypothese ist entweder richtig oder falsch – zufällig sind lediglich die erhobenen Daten und damit die Testgröße sowie die resultierende Entscheidung [6]. Die Größe des p-Werts sagt auch nichts über die Größe des wahren Effekts: Bei großen Stichproben kann man kleine p-Werte erhalten, obwohl der Effekt sehr klein und medizinisch gar nicht relevant ist, z. B. eine Blutdrucksenkung von 0,5 mmHg – der Effekt ist sehr gering, wurde aber ggf. eindeutig nachgewiesen [18]. Außerdem sollte man sich als Leser fragen, ob der betrachtete „signifikante" Parameter klinisch überhaupt eine Rolle spielt. Für weitere Fallstricke der Interpretation s. auch [10]. Um die Größe des Effekts und die Präzision der Schätzung zu beurteilen, zieht man am besten wieder das KI zurate.

Bildnachweis: MEV

Ein niedriger p-Wert zeigt nicht automatisch klinische Relevanz – und ist auch keine Garantie für wissenschaftlich korrekte Datenanalyse.

Wie hängen KI und p-Wert zusammen? Bei einem niedrigen p-Wert wird auch das KI tendenziell eng sein. Vergleicht man z. B. 2 Mittelwerte, wird das KI für die mittlere Differenz bei sehr kleinem p-Wert eng sein bzw. die Null nicht enthalten. Umgekehrt kann man vom KI auf Signifikanz schließen:

> Enthält das 95 %-KI nicht den Wert des „Null-Effekts" (z. B. 0 bei Differenz der Mittelwerte oder 1 beim relativen Risiko oder einer Korrelation), bedeutet dies ein signifikantes Ergebnis zum Niveau 5 %, der p-Wert wird < 5 % liegen [2].

Das Schöne am KI: Da es Informationen in der gleichen Skala wie die untersuchte Variable liefert, lassen sich die gefundenen Effekte meist besser klinisch interpretieren [2]. Um wie viel der p-Wert kleiner ist als α, kann man am KI aber nicht genau ablesen. Am besten gibt man daher beides an:

> Den p-Wert als genaue Angabe der statistischen Evidenz,

> das KI als klinisch interpretierbare Information über die Unsicherheit des beobachteten Effekts [2].

In unserem Beispiel Medikament vs. Plazebo [6] sähe das so aus:

> p < 0,0001 („Der Unterschied zwischen Medikament und Plazebo ist signifikant zum Niveau 0,0001 bzw. 0,01 %") und

> 95 %-KI: 8,82–14,26 („Mit einer Wahrscheinlichkeit von 95 % senkt das Medikament den Blutdruck um 8,82–14,26 mmHG stärker als Plazebo.")

Das KI ist insgesamt etwas flexibler einsetzbar: Während Signifikanztests nur zulässig sind, wenn die Hypothesen vor Kenntnis der Daten aufgestellt werden (s. oben), kann man KIs auch noch später berechnen – nicht zum Nachweis von Signifikanz, aber zur genaueren Beschreibung der Daten [2].

Beispiel: multiples Testen beim Würfeln

Auch die Fehler beim multiplen Testen sind beim Würfelversuch anschaulicher: Wenn Sie 1 schwarzen und 1 weißen Würfel jedes Mal unter andersfarbigem Licht oft genug werfen, werden Sie irgendwann schwarz 6 und weiß 1 würfeln. Es wäre aber unseriös, diesen einen Wurf herauszugreifen und zu sagen:

> „Die Wahrscheinlichkeit dafür betrug nur 1/36 = 0,028 = 2,8 % – das Ergebnis ist also signifikant.
> Unter Licht mit dieser Wellenlänge ist der schwarze Würfel daher besser als der weiße."

Erscheint ein solcher Unterschied aus physikalischen Gründen irgendwie plausibel, könnten Sie den Befund allenfalls als Anlass nehmen, eine neue Versuchsreihe nur unter dieser Wellenlänge durchzuführen.

Fazit

Um aus den typischen Stichprobendaten belastbare Erkenntnisse zu gewinnen, sind statistische Signifikanztests nötig. Man sollte sich allerdings nicht zu sehr auf den p-Wert als das vermeintlich wichtigste Ergebnis verlassen: Von der Datenerhebung über die gewählte Testmethodik bis zur Interpretation der Ergebnisse gibt es zahlreiche Fehlerquellen, die eine Studie letztendlich wertlos machen können – trotz eines „hochsignifikanten" Ergebnisses.

Statistische Signifikanz (ja oder nein) kann man anhand kritischer Werte + Prüfgröße, anhand des KI oder anhand des p-Werts ausdrücken. Für die praktische Interpretation der Daten ist das KI am anschaulichsten.

Multiples Testen Spätestens seit Computer die p-Werte ausspucken, kann man sehr viele Daten erheben und entsprechend viele Signifikanztests durchführen – im Zweifelsfall solange, bis man signifikante Effekte findet (► Infokasten oben). Man prüft quasi mehrere Hypothesen an derselben Stichprobe. So werden z. B.

> Subgruppen von Patienten gebildet,
> der Behandlungserfolg zu verschiedenen Zeitpunkten geprüft,
> verschiedenste Einflussfaktoren (Umweltparameter, Risikofaktoren, Genvarianten etc.) getestet oder
> neben dem primären auch mehrere sekundäre Endpunkte (z. B. Laborwerte) untersucht.

Dabei steigt allerdings das Risiko für „falsch-positive" Ergebnisse drastisch: Bleibt man bei $\alpha = 5\,\%$, beträgt die Wahrscheinlichkeit für eine korrekte Entscheidung für H_0 dann nicht mehr 95 % (wie beim einfachen Test), sondern bei 10 verschiedenen Parametern nur noch $0{,}95^{10} = 60\,\%$, d. h. der α-Fehler steigt auf 40 %. Dadurch könnte man z. B. fälschlicherweise auf die Wirksamkeit eines Medikaments schließen (zumindest in z. B. der einen Subgruppe mit $p < \alpha$). Oder man macht z. B. eine genetische Assoziationsstudie mit 100 Tests und jeweils $\alpha = 0{,}05$ – und bekommt schon rein zufällig 5 falsch-positive Ergebnisse.

Um nicht auf zufällige Korrelationen hereinzufallen, gibt es mehrere Möglichkeiten, die Irrtumswahrscheinlichkeit anzupassen:

> Bei der Bonferroni-Korrektur teil man α durch die Zahl der Tests und vergleicht die einzelnen p-Werte dann mit diesem Wert (bei 10 Tests wird aus $\alpha = 0{,}05$ dann 0,005 als Signifikanzniveau).
> Man kann auch umgekehrt jeden p-Wert mit der Zahl der untersuchten Hypothesen multiplizieren und das Ergebnis mit dem Gesamtniveau α vergleichen.

Diese Verfahren sind recht „konservativ", d. h. man wird Probleme haben, „wahre" Unterschiede nachzuweisen. Andere Verfahren sind komplizierter, liefern aber mehr Power [16].

Wurden in einer Studie mehrere statistische Tests gemacht (d. h. mehrere p-Werte berechnet), achten Sie auf Folgendes [6, 12]:

> Korrigieren die Autoren das Signifikanzniveau?
> Waren eventuelle Subgruppenanalysen von vornherein eingeplant – oder falls nicht: Werden sie entsprechend zurückhaltend (deskriptiv) interpretiert?
> Diskutieren die Autoren auch die nicht signifikanten Endpunkte, statt sich nur auf die signifikanten zu stützen?
> Verwenden sie statt mehrerer paralleler Tests ein komplexeres Verfahren (z. B. Varianzanalyse statt mehrerer t-Tests)?

Oft können aber nur Experten wirklich beurteilen, ob multiple Tests korrekt berechnet und alle Probleme berücksichtigt sind. Führt man selbst eine Studie durch, sollte man versuchen, mit wenigen Haupthypothesen auszukommen. Niedrige p-Werte in zusätzlichen Tests sollte man vorsichtig interpretieren bzw. zur Hypothesengenerierung nutzen [16].

Testet man mehrere Hypothesen gleichzeitig, muss man die Irrtumswahrscheinlichkeit anpassen. Subgruppenanalysen im Nachhinein sind v. a. beschreibend zu verstehen.

Zu Risiken und Nebenwirkungen fragen Sie Ihren Biometriker oder Statistiker Natürlich kann dieser Artikel nur eine grobe Übersicht geben, was hinter den Studienergebnissen steckt, die wir tagtäglich lesen. Müssen Sie selbst eine Studie planen oder auswerten, sollten Sie unbedingt einen Fachmann zurate ziehen!

Julia Rojahn

Danksagung Die Autorin bedankt sich bei Frau Prof. Dr. Christel Weiß für die gründliche Durchsicht und Korrektur des Manuskripts.

Literatur und Zusatzmaterial online Das Literaturverzeichnis zu diesem Beitrag sowie Tab. e1 und e2 finden Sie im Internet: Unter „www.thieme-connect.de/products" können Sie die Seite der Lege artis aufrufen und beim jeweiligen Artikel auf „Zusatzmaterial" klicken – es frei zugänglich.

Beitrag online zu finden unter http://dx.doi.org/10.1055/s-0041-107449