Studien verstehen (2): Prüfen der Signifikanz

Woher kommt das kleine „p“?

Häufig werden Ergebnisse von Studien so angepriesen: „Das neue Medikament ist wirksamer als das alte, der Unterschied ist statistisch signifkant.“ Vielen Medizinern ist aber unklar, was „statistisch signifikant“ genau bedeutet – obwohl sich in der Wissenschaft enorm viel darum dreht. Dr. Christel Weiß, Leiterin der Biomathematik am Universitätsklinikum Mannheim, weiht Sie in das Geheimnis der Signifikanzprüfung ein.

Am 12. Februar 2004 erschien in der Wochenzeitung „Die Zeit“ unter dem Titel „Das Wunder von Innsbruck“ ein Bericht über einen ungarischen Fußballspieler, der im Alter von 24 Jahren während eines Fußballspiels einen Herzinfarkt erlitten hatte und kurz danach verstorben war. Der Notarzt hatte ihm – nachdem Herzmassage und Elektroschock nichts bewirkten – einen Adrenalinstoß verabreicht. „Vielleicht würde der Fußballspieler noch leben“, mutmaßte der Journalist, „wenn die Ärzte ihm ein anderes Mittel gegeben hätten.“ Er pries das Hormon Vasopressin (antidiuretisches Hormon, ADH oder Adiuretin) als ein Wundermittel, das geeignet sei, im Falle eines Herzstillstands „ein paar Menschen mehr zu retten“. Dabei berief er sich auf ein Paper, das wenige Wochen zuvor in der renommierten Fachzeitschrift „New England Journal of Medicine“ erschienen war.

In diesem Paper* präsentierten Forscher die Ergebnisse einer randomisierten Studie der Universitätsklinik Innsbruck: Von 257 Patienten, die nach einem Herzstillstand mit Vasopressin behandelt wurden, überlebten zwölf, also 4,67%. In der Vergleichsgruppe, bei den mit Adrenalin behandelten Patienten, kamen dagegen vier von 262 mit dem Leben davon, also nur 1,53%. Die Überlebensquote bei Vasopressin ist demnach dreimal so hoch wie bei Adrenalin. Die statistische Analyse zeigte: Die Vasopressin-Therapie ist signifikant überlegen – der p-Wert ist 0,0384! Kardiologen im In- und Ausland forderten daraufhin, in der Notfallmedizin sofort Vasopressin statt Adrenalin einzusetzen. Sind diese Forderungen gerechtfertigt? Zwingt das Ergebnis der statistischen Analyse dazu, die bisherige Standardtherapie mit Adrenalin über Bord zu werfen? Diese Frage ist von fundamentaler Bedeutung – geht es doch hier um zwei Therapien, die über Leben oder Tod eines Patienten entscheiden können.

Zufall und Signifikanz – zwei Mysterien ▶ Maßgeblich dazu beigetragen, dass die Innsbrucker Studie überhaupt publiziert wurde, hat wahrscheinlich der p-Wert – genauer gesagt: dass der p-Wert kleiner war als 0,05. Bei einem p-Wert unter 0,05 wird das Ergebnis eines statistischen Tests als „signifikant“ bezeichnet – soviel wissen noch viele Mediziner, die wissenschaftliche Studien lesen. Doch wie erhält man diesen Wert? Und was bedeutet er? Diese Fragen lassen sich schon weniger leicht beantworten.

Hinter jedem p-Wert steckt ein Signifikanztest. Zu den bekanntesten Tests, die in der medizinischen Forschung angewandt werden, zählen der Chiquadrat-Test und der t-Test. Sie werden häufig zum Vergleich zweier Häufigkeiten oder zweier Mittelwerte herangezogen. Daneben existiert noch eine ganze Reihe weiterer Testverfahren. Welchen Test man verwendet, hängt von der Fragestellung und den Daten ab. Die Vorgehensweise ist aber bei jedem prinzipiell dieselbe (☎ ***Tabelle***): Passend zu dem, was man herausfinden möchte, formuliert man zunächst zwei Hypothesen. Die erste davon, die Nullhypothese, ist konservativ

» DIE NULLHYPOTHESE IST KONSERVATIV UND KEINESFALLS AUFSEHEN ERREGEND.

und in keiner Weise Aufsehen erregend. Sie besagt, dass es keinen Unterschied gibt. In unserem Beispiel heißt die Nullhypothese: Egal, ob Ärzte Patienten mit Herzstillstand Adrenalin oder Vasopressin spritzen, die Wahrscheinlichkeit zu überleben ist für den Patienten gleich. Die Unterschiede, die von den Innsbrucker Forschern gemessen wurden, sind zufällige Schwankungen. Die zweite, zur Nullhypothese konkurrierende Alternativhypothese, stellt dagegen Althergebrachtes in Frage. Sie ist innovativ und heißt in unserem Fall: Die Überlebenswahrscheinlichkeit mit

SO GEHT EIN SIGNIFIKANZTEST

Einzelne Schritte:		Beispiel:
1.	Formulieren der Nullhypothese	„Die Überlebenswahrscheinlichkeiten der beiden Therapiegruppen sind gleich.“
2.	Formulieren der Alternativhypothese	„Überlebenswahrscheinlichkeiten sind unterschiedlich.“
3.	Wahl des Signifikanzniveaus	$\alpha = 0{,}05$
4.	Berechnen der Prüfgröße	$\chi^2 = 4{,}2882$ (mit Chi²-Vierfeldertest)
5.	Ermitteln des p-Wertes	$p = 0{,}0384$
6.	Entscheidung: falls $p < \alpha \rightarrow$ Alternativhypothese wird angenommen falls $p \geq \alpha \rightarrow$ Nullhypothese wird beibehalten	„Der Unterschied ist statistisch signifikant; die Alternativhypothese wird angenommen.“

* Wenzel V et al: A Comparision of Vasopressin and Epinephrine for Out-of-Hospital Cardiopulmonary Resuscitation, N. Eng. J. Med. 2004; 350: 2206–2209

▲ *Viele Forscher schauen vor allem auf den p-Wert, wenn sie eine Studie lesen. Natürlich ist der p-Wert wichtig – aber er ist nicht das einzige Qulitätsmerkmal!*

Vasopressin unterscheidet sich von der mit Adrenalin. Die Daten wertet man je nach Test mit einer bestimmten Rechenvorschrift aus. Das Resultat ist die sogenannte Prüfgröße, aus der sich der p-Wert ergibt. Natürlich rechnet ein Statistiker das Ganze heutzutage nicht mehr manuell aus, sondern benutzt eine leistungsfähige Software.

Was sagt mir der p-Wert? ▶ Ein Forscher favorisiert in der Regel das Neue und möchte deshalb die Nullhypothese („Es gibt keinen Unterschied zwischen neuer und alter Therapie") am liebsten verwerfen. Ob ihm das gelingt, hängt vom p-Wert ab. Wenn das p kleiner ist als ein vorab festgelegtes Niveau, hat er sein Ziel erreicht: Das Testergebnis ist signifikant.

Aber was beschreibt der p-Wert nun genau? Etwas salopp formuliert, ist der p-Wert die Wahrscheinlichkeit dafür, dass der beobachtete Effekt ein reiner Zufallsbefund ist. Man spricht auch von der Irrtumswahrscheinlichkeit. Denn: Selbst wenn Vasopressin und Adrenalin genau gleich wirksam wären, hätte man zufällig den in unserem Beispiel gefundenen Unterschied beobachten können. Die Wahrscheinlichkeit dafür ist jedoch mit 3,84 % ($p = 0{,}0384$) ziemlich gering. Es ist also unwahrscheinlich, dass das Ergebnis allein auf den Zufall zurückzuführen ist – oder anders formuliert: dass er unter der Nullhypothese zustande gekommen ist. Deshalb lehnt man die Nullhypothese ab und nimmt notgedrungen die Alternativhypothese an. Dass die Alternativhypothese wirklich richtig ist, hat man damit aber keineswegs bewiesen. Das gefundene Ergebnis ist nur schwer mit der Nullhypothese vereinbar. Genauso gilt umgekehrt: Ein nicht-signifikantes Testergebnis ist nicht gleichbedeutend mit der Aussage „Es gibt keinen Unterschied". Man kann lediglich schlussfolgern: „Mit diesen Daten kann kein Unterschied nachgewiesen werden".

Geschichte einer Trennungslinie ▶ Die magische Grenze für den p-Wert, die ein Ergebnis entweder „signifikant" oder „nicht-signifikant" werden lässt, hat einen unspektakulären Ursprung: In früheren Zeiten – als keine Computer existierten – musste man die Prüfgröße manuell bestimmen und diese dann mit einem „kritischen Punkt" vergleichen. Den suchte man mühsam in Tabellen, die heute noch in vielen Statistikbüchern zu finden sind. Für jeden Signifikanztest und für jede Grenze brauchten die Statistiker eine extra Tabelle. Sich auf eine bestimmte Grenze zu einigen erschien daher sinnvoll – obwohl der Anwender eines Tests sie theoretisch nach Belieben festlegen kann. In den Biowissenschaften hat sich eine Grenze von 5% etabliert (maximale Irrtumswahrscheinlichkeit oder Signifikanzniveau $\alpha = 0{,}05$). Das heißt: Ist die Wahrscheinlichkeit, dass ein Ergebnis zufällig zustande gekommen ist, kleiner als 5%, gilt es als „signifikant" ($p < 0{,}05$). So lassen sich statistisch abgesicherte Entscheidungen miteinander vergleichen. Nur wenn es um Fragestellungen von besonderer Tragweite geht, legt man manchmal eine maximale Irrtumswahrscheinlichkeit von 1% oder sogar 0,1% fest ($\alpha = 0{,}01$ oder sogar $\alpha = 0{,}001$).

Was sagt der p-Wert nicht? ▶ Zweifelsohne ist der p-Wert eine fundamentale Größe einer statistischen Analyse, da durch ihn der Zufall kontrolliert werden kann. Allerdings darf seine Bedeutung nicht überschätzt werden. „Statistisch signifikant" ist nicht gleichbedeutend mit „klinisch relevant" oder „wissenschaftlich bedeutsam". Ein p-Wert unter der magischen Grenze von 0,05 besagt lediglich, dass der beschriebene Effekt nicht allein durch den Zufall erklärt werden kann. Wie groß der nachgewiesene Effekt ist, was er für Ursachen hat oder welche Konsequenzen sich daraus ergeben – darüber schweigt der p-Wert sich aus.

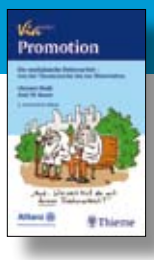
Zumindest bei der Größe des nachgewiesenen Effektes kann aber ein anderer Wert aus der Statistik weiterhelfen: Mit Vasopressin überlebten 4,67% der Patienten, mit Adrenalin 1,53% – die Differenz der Überlebenswahrscheinlichkeiten beträgt in unserer Stichprobe also 3,14%. Eine leistungsfähige Software

▶▶▶

VIA MEDICI SPEZIAL

Lust an der Wissenschaft

Wenn man sich aufmacht, Frau oder Herr Doktor zu werden, ist es ratsam, schon vorab informiert zu sein, was da auf einen zukommt. Das Via medici-Buch „Promotion" hilft Ihnen – beim Planen und beim Literaturstudium der Doktorarbeit genauso wie beim Schreiben und der statistischen Auswertung. Wir verlosen 10 Bücher. Teilnehmen können Sie unter www.thieme.de/viamedici/zeitschrift/spezial mit dem Stichwort „Promotion". Einsendeschluss ist der 6.8.2007.



Ein ungarischer Fußballspieler erleidet auf dem Feld einen Herzinfarkt und stirbt – trotz Reanimationsversuchen und Adrenalin. Hätte man ihm besser Vasopressin gegeben? Der p-Wert allein kann das nicht klären.

kann uns jetzt das 95%-Konfidenzintervall berechnen. Dieses gibt uns mit einer Wahrscheinlichkeit von 95% an, zwischen welchen Grenzen die „wahre" Differenz liegt. Hier ist es: [0,0017 ; 0,0612]. Der „wirkliche" Unterschied zwischen Vasopressin und Adrenalin liegt also mit 95-prozentiger Wahrscheinlichkeit irgendwo zwischen 0,17% und 6,12%. Es bleibt dem Leser des Papers überlassen zu beurteilen, inwieweit ein Therapieunterschied in dieser Größenordnung praktisch relevant ist. Während der p-Wert nur besagt, *ob* ein Unterschied angenommen werden darf, informiert das Konfidenzintervall über die Größe des Unterschiedes. Je kleiner – oder enger – das Intervall ist, desto präziser ist die Schätzung.

Ein p-Wert alleine sagt nichts darüber aus, ob ein Zusammenhang kausal bedingt ist. Waren die Patienten der Adrenalin-Gruppe vielleicht älter? Da die Studie randomisiert durchgeführt wurde, ist davon auszugehen, dass die Therapiegruppen sich nicht wesentlich unterschieden. duie Waren die In jedem Fall muss man sich aber Gedanken über die Ursachen eines möglichen Effekts machen.

Welche Konsequenzen lassen sich aus der Untersuchung ziehen? Auch dazu schweigt der p-Wert. Aus der Studie geht aber klar hervor: Die meisten Patienten, die einen Herzstillstand erleiden, überleben diesen nicht – weder mit Vasopressin noch mit Adrenalin. In beiden Gruppen sterben mehr als 95%. Lediglich 3,14% der behandelten Patienten profitieren von der Behandlung mit Vasopressin – das ist einer von 32! Nun lässt sich argumentieren, dass auch ein einziger Patient von 32, dessen Leben gerettet werden kann, den Einsatz von Vasopressin rechtfertigt. Liest man das Paper jedoch genau, steht darin, dass 40% der Patienten, die mit Vasopressin reanimiert werden konnten, nicht mehr aus dem Koma erwachten. Dieser Preis für das Überleben ist nicht im p-Wert enthalten.

Frage, Test und Stichprobe beeinflussen das p Auch die Fragestellung hat einen wesentlichen Einfluss auf das Ergebnis des Signifikanztests: Bei einer zweiseitigen Fragestellung lässt man offen, ob eine neue Methode besser oder schlechter wirkt, man formuliert die Alternativhypothese einfach als: „Die Methoden unterscheiden sich". Bei einer einseitigen Fragestellung legt man vorher fest, in welche Richtung der Unterschied geht. Die Alternativhypothese hieße beispielsweise: „Die Überlebenswahrscheinlichkeit bei Vasopressin ist höher als bei Adrenalin". Einseitige Fragestellungen haben den Vorteil, dass sich der p-Wert halbiert! Allerdings sollte eine einseitige Fragestellung immer stichhaltig begründet werden.

Eine entscheidende Rolle für den p-Wert spielt außerdem der Stichprobenumfang: Je mehr Patienten die Stichprobe umfasst, desto unwahrscheinlicher werden zufällige Unterschiede zwischen den Gruppen. Mit sehr großen Patientengruppen lassen sich kleinste p-Werte erzeugen und Unterschiede

» MIT GROSSEN GRUPPEN LASSEN SICH KLEINSTE P-WERTE ERZEUGEN.«

nachweisen, die praktisch bedeutungslos sind. Eine sehr kleine Stichprobe gibt dagegen der Alternativhypothese keine Chance!

Publikationen neigen zur Signifikanz Im Hinterkopf behalten sollte man auch, dass sich eine Studie mit signifikantem Ergebnis leichter publizieren lässt als ohne. Bei Übersichtsarbeiten und Meta-Analysen kann daher manchmal ein positiver Effekt vorgetäuscht werden, weil die Studien mit negativen Ergebnissen einfach nicht veröffentlicht wurden. Dieses Problem ist so bedeutsam, dass Wissenschaftler ihm einen Namen gegeben haben: „Publication bias".

Der Leser eines Papers sollte sich also von einem winzigen p-Wert nicht blenden lassen, sondern weitere Informationen einholen, und erst dann beurteilen, wie relevant der beschriebene Effekt für seine Patienten oder sein Labor wirklich ist. Auch das Vasopressin hat Adrenalin letzten Endes nicht aus der Notfallversorgung verdrängen können – dem Innsbrucker p-Wert zum Trotz!

Dr. sc. hum. Christel Weiß

Dr. sc. hum. Christel Weiß ist Leiterin der Abteilung für Statistik, Biomathematik und Informationsverarbeitung am Universitätsklinikum Mannheim.

Fotos: D. Schmid