Was ist ein Signifikanztest? Allgemeine Aspekte

– Artikel Nr. 9 der Statistik-Serie in der DMW –

What is a significance test? General issues

Autoren S. Lange [1] R. Bender [1]

Institut [1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

Zum statistischen Nachweis von Unterschieden oder Effekten werden häufig Signifikanztests eingesetzt. Das Ergebnis eines solchen Tests wird zumeist als *p*-Wert [1] ausgegeben. Anhand dieses *p*-Werts wird entschieden, ob beobachtete Unterschiede statistisch signifikant sind (wenn der *p*-Wert kleiner ist als das Signifikanzniveau α von zum Beispiel 5%) oder nicht.

Bei der Anwendung von Signifikanztests sind allgemein zwei Aspekte zu beachten: Zum einen kann mit Hilfe von Signifikanztests nichts (statistisch) „abgesichert" werden. Der Begriff „Absichern" impliziert eine 100%ige Sicherheit, also den Ausschluss einer Irrtumsmöglichkeit. Aber diese Irrtumsmöglichkeit ist den Signifikanztests geradezu immanent; sie ist allenfalls unter Einhaltung bestimmter Voraussetzungen quantifizierbar (Signifikanz- bzw. Irrtumsniveau α [1]). Zum anderen sind Signifikanztests im Sinne einer wissenschaftlichen Hypothesenüberprüfung – einer konfirmatorischen Statistik – nur dann einsetzbar, wenn die zu prüfende Hypothese vor Kenntnis der Daten aufgestellt wurde. Gegen diesen Grundsatz wird allerdings in der Praxis häufig verstoßen. Im anderen Fall können Signifikanztests nur noch der weitergehenden Beschreibung der erhobenen Daten dienen. Für solche deskriptiven Zwecke sind aber häufig Konfidenzintervalle [2] besser geeignet.

Ob nun überhaupt ein Signifikanztest eingesetzt werden soll, und wie er dann gegebenenfalls zu interpretieren ist, hängt also von der wissenschaftlichen Vorgehensweise ab. Wir wollen diese Problematik im Folgenden nicht weiter erörtern. Hat man sich für die Durchführung eines Signifikanztests entschieden, ist die Wahl des zu verwendenden Tests von der Fragestellung, dem Studiendesign und dem Messniveau des betrachteten Merkmals abhängig.

Zunächst muss man entscheiden, ob sich die zu testende Hypothese auf Verteilungsparameter **einer** Stichprobe bezieht (zum Beispiel ob ein Mittelwert signifikant von 0 verschieden ist), oder ob **mehrere** Stichproben verglichen werden sollen (zum Beispiel ob sich zwei Mittelwerte signifikant unterscheiden), was den weitaus häufigsten Fall darstellt. Bei der Auswertung von zwei oder mehr Stichproben muss man den Abhängigkeitsstatus der Stichproben berücksichtigen. Handelt es sich um den Vergleich von unabhängigen Gruppen (zum Beispiel im Parallel-Gruppen-Design einer kontrollierten Studie) so müssen Verfahren für **unabhängige Stichproben** verwendet werden. Handelt es sich dagegen um den Vergleich von abhängigen Werten (zum Beispiel bei Messwertwiederholungen [3] an denselben Probanden), so kommen Verfahren für **abhängige Stichproben** in Frage. Der nächste entscheidende Faktor ist das Messniveau der betrachteten Zielgröße. In der Praxis genügt hierbei die Unterscheidung zwischen den Messniveaus **binär** (ja/nein), **nominal** (ungeordnete Kategorien, zum Beispiel Tumorentitäten), **ordinal** (geordnete Kategorien, zum Beispiel Tumorstadien), **stetig** (quantitatives Merkmal mit theoretisch unendlich vielen Merkmalsausprägungen, zum Beispiel Herzfrequenz) und **zensiert** (Überlebenszeiten).

kurzgefasst

Signifikanztests dienen zumeist dem statistischen Nachweis von Unterschieden oder Effekten. Dabei versucht man, die Nullhypothese zu widerlegen. Signifikanztests sind nur dann einsetzbar, wenn die Hypothese vor Kenntnis der Daten aufgestellt wurde. Das Ergebnis des Tests wird häufig als *p*-Wert angegeben. Signifikanz liegt vor, wenn der *p*-Wert kleiner ist, als das zuvor festgelegte Signifikanzniveau. Mit Signifikanztests kann man die Irrtumswahrscheinlichkeit quantifizieren, nicht ausschließen.

Schlüsselwörter
- Signifikanztest
- Nullhypothese
- *p*-Wert
- *t*-Test

Key words
- Significance test
- Null hypothesis
- *p*-value
- *t*-test

Bibliografie
DOI 10.1055/s-2007-959032
Dtsch Med Wochenschr 2007; 132: e19–e21 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. Stefan Lange
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail stefan.lange@iqwig.de

Student *t*-Test

Ein häufig verwendeter Signifikanztest ist der „(Student) ***t***-Test“. Er kommt zum Einsatz, wenn es um die Betrachtung stetiger Zielgrößen geht. Wir wollen das anhand eines Beispiels konkretisieren: Patienten mit einer chronischen venösen Insuffizienz (CVI) leiden ab einem bestimmten Krankheitsstadium unter Ödemen der abhängigen Körperpartien. Ein anerkanntes Zielkriterium im Rahmen von klinischen Therapieprüfungen bei solchen Patienten ist die Differenz des Unterschenkelvolumens im Verlauf als Surrogat für eine Ödemreduktion. Hierbei handelt es sich also um eine stetige Zielgröße. Soll nun zum Beispiel die Wirksamkeit einer medikamentösen Therapie (im Folgenden als „Verum“ bezeichnet) geprüft werden, gilt als Standardverfahren, die Patienten zufällig (randomisiert) zwei Gruppen zuzuordnen, die während der Studie entweder mit Verum oder mit einem Placebo behandelt werden. Es ist somit die Situation zweier unabhängiger Stichproben gegeben.

Das Ziel der Studie wäre es, zu demonstrieren, dass sich unter der Therapie mit Verum eine andere (größere) Volumendifferenz einstellt als unter Placebo. Statistische Tests bedienen sich einer deduktiven Schlussweise. Das bedeutet, es wird zunächst eine Nullhypothese [1] aufgestellt, dass nämlich **kein** Unterschied zwischen den Gruppen besteht, mit dem Ziel, diese Hypothese zu verwerfen, um das Gegenteil, die Alternativhypothese (es besteht ein Unterschied) annehmen zu können. Man prüft also, ob die beiden Gruppen der gleichen Grundgesamtheit entstammen.

Beim ***t***-Test wird die Gleichheit bzw. Unterschiedlichkeit der zu vergleichenden Gruppen an einem Parameter gemessen, und zwar an dem Erwartungswert μ. Deshalb wird der ***t***-Test auch als ein **parametrischer Test** bezeichnet. Der Erwartungswert ist im übertragenen Sinn der Mittelwert der Grundgesamtheit, oder umgekehrt: Der Mittelwert einer Stichprobe ist ein Schätzwert für den Erwartungswert der Grundgesamtheit.

Wenn nun beide Gruppen derselben Grundgesamtheit angehören, besitzen sie den gleichen Erwartungswert. Kennzeichnend für die statistische Schlussweise ist, dass es für die beobachtete Variable zwar einen (theoretischen) Erwartungswert gibt, zum Beispiel eine Abnahme des Unterschenkelvolumens um 50 ml, dass aber beim einzelnen Patienten praktisch nie exakt dieser Erwartungswert beobachtet wird. Die Abweichungen vom Erwartungswert werden dabei als Ergebnis eines Zufallsprozesses betrachtet und durch die Standardabweichung σ [6] quantifiziert. Auch der Mittelwert einer Stichprobe und die Differenz zweier Mittelwerte stellen Zufallsvariablen dar.

Bei der oben skizzierten Studie wird man also auch bei tatsächlicher Gleichheit der beiden Gruppen nicht beobachten, dass die Differenz der beiden Stichprobenmittelwerte genau Null ist, sondern es wird eine (zufallsbedingte) Abweichung von der theoretischen Erwartung geben. Erst wenn diese Abweichung eine bestimmte Größenordnung überschreitet, wird man sich für die Ablehnung der Nullhypothese entscheiden. Diese Größenordnung muss unbedingt vor Testdurchführung durch das Signifikanzniveau und den Stichprobenumfang festgelegt werden. Anderenfalls ist das Ergebnis des statistischen Tests nicht mehr im Sinne einer Entscheidungsregel eindeutig interpretierbar.

Eine Voraussetzung für den Einsatz des ***t***-Tests ist die Annahme einer Normalverteilung der zu betrachtenden Zielvariable. Normalverteilung bedeutet folgendes: Bestimmte man von allen Patienten der Erde mit einer CVI nach einer 12wöchigen Therapie mit Verum oder Placebo die Differenz des Unterschenkelvolumens („Grundgesamtheit“), würde die Verteilung dieser Werte eine bestimmte glockenförmige Gestalt annehmen, die zuerst von dem Mathematiker Gauß beschrieben und formalisiert wurde (Gauß’sche Glockenkurve). Jede Normalverteilung kann durch eine einfache Umrechnung auf die so genannte Standardnormalverteilung mit $\mu = 0$ und $\sigma = 1$ zurückgeführt werden. Dies hat enorme praktische Bedeutung, da damit die Eigenschaften der Standardnormalverteilung auf jegliche Normalverteilung übertragen werden können.

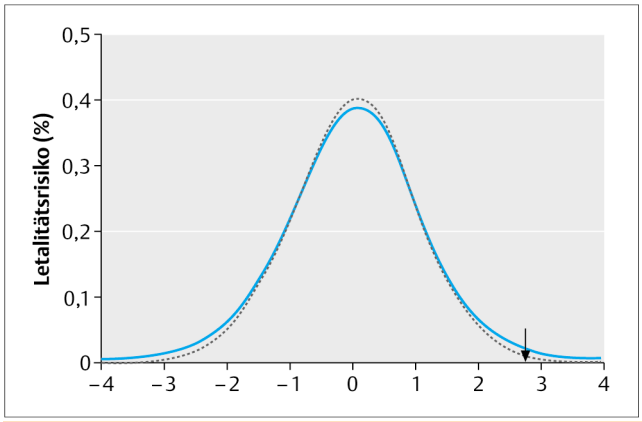

Abb. 1 Wahrscheinlichkeitsdichten der Standardnormalverteilung (gestrichelte Linie) und einer *t*-Verteilung mit 9 Freiheitsgraden (durchgezogene Linie). Eingezeichnet (Pfeil) ist der Wert der Prüfgröße (2,8) für den im Text beschriebenen *t*-Test.

Beim ***t***-Test wird die Differenz zweier Stichprobenmittelwerte dividiert durch den Standardfehler dieser Differenz als Prüfgröße (Teststatistk) T herangezogen. Die Division durch den Standardfehler führt zu einer Normierung [1] ähnlich der oben genannten Umrechnung. Ausgehend von der Annahme einer Normalverteilung der Zielvariable, folgt T einer ähnlichen Wahrscheinlichkeitsverteilung, nämlich der ***t***-Verteilung. Diese ist durch einen Parameter, die so genannten „Freiheitsgrade“ (FG) charakterisiert. Mit zunehmenden Freiheitsgraden nähert sich die ***t***-Verteilung der Normalverteilung an (**Abb. 1**).

Wir wollen auf den Begriff der „Freiheitsgrade“ nicht näher eingehen. Bei der ***t***-Verteilung ergeben sie sich als eine Funktion des Stichprobenumfangs n bei Betrachtung einer Stichprobe ($FG = n-1$) bzw. der Stichprobenumfänge n_1 und n_2 bei Betrachtung von zwei Stichproben ($FG = n_1+n_2-2$).

kurzgefasst

Mit dem *t*-Test kann die Signifikanz beim Vergleich stetiger Zielgrößen geprüft werden, indem die Gleichheit bzw. Verschiedenheit zweier Stichproben anhand der Differenz ihrer Erwartungswerte gemessen wird. Erwartungswerte entsprechen Mittelwerten von (fiktiven) unendlichen Grundgesamtheiten. Die Mittelwerte aus Stichproben sind Schätzwerte für die entsprechenden Erwartungswerte. Vor Durchführung eines Signifikanztests muss festgelegt werden, bei welchem Irrtumsniveau die Nullhypothese abgelehnt werden soll.

Beispiel: Klinische Studie

In der eingangs beschriebenen Studie zum Nachweis der Wirksamkeit von Verum war im Studienprotokoll für das Signifikanzniveau α der allgemein übliche Wert von 0,05 (bzw. 5%) festgelegt. Es wurde folgendes Ergebnis beobachtet: 95 mit Verum behandelte Patienten hatten im Studienverlauf eine mittlere Abnahme des Unterschenkelvolumens von 44 ml, die (empirische) Standardabweichung betrug 111 ml. In die Placebogruppe wurden 46 Patienten aufgenommen, bei denen es im Mittel zu einer Zunahme des Unterschenkelvolumens um 10 ml (Standardabweichung 102 ml) kam [5].

Die Differenz der beiden Stichprobenmittelwerte ergibt also 54 ml, aus den empirischen Standardabweichungen und den Stichprobenumfängen errechnet sich ein Standardfehler von 19 ml. Die Prüfgröße beträgt somit 54/19 = 2,8. Falls die beiden Stichproben tatsächlich der gleichen Grundgesamtheit angehören sollten, wäre es sehr unwahrscheinlich, einen solchen oder noch extremer von der Nullhypothese abweichenden Wert zu beobachten (vergleiche Abbildung 1). Im konkreten Fall beträgt die Wahrscheinlichkeit dafür etwa 0,006. Diese Wahrscheinlichkeit entspricht dem ***p***-Wert [1]. Da der ***p***-Wert kleiner ist als das vorgegebene Signifikanzniveau α, ist der in der Studie beobachtete Unterschied statistisch signifikant.

Die Beschreibung des ***t***-Tests erfolgte für Fragestellungen, bei denen Abweichungen von der Nullhypothese in beide Richtungen (**zweiseitig**) entdeckt werden sollen. Dies entspricht der gängigen biometrischen Praxis. So galt das Interesse der obengenannten Studie hauptsächlich der Überlegenheit von Verum gegenüber Placebo, aber auch eine Unterlegenheit von Verum wäre nicht ohne Konsequenzen geblieben (zum Beispiel Abbruch aller weiteren klinischen Untersuchungen). Es sind natürlich auch Fälle denkbar, in denen tatsächlich nur eine Abweichungsrichtung (**einseitig**) interessant ist. Wir wollen diese Problematik hier allerdings nicht weiter vertiefen.

Als Voraussetzung für die Anwendbarkeit des ***t***-Tests wird im Allgemeinen das Vorliegen einer Normalverteilung gefordert. Dies ist eine theoretische Forderung, die in praxi nie erfüllt werden kann: Der Wertebereich realer Daten ist stets beschränkt und besteht aus nur endlich vielen (diskreten) Werten, was der Normalverteilungsannahme widerspricht. Das entscheidende Kriterium für die Zuverlässigkeit eines statistischen Tests ist die Einhaltung des vorgegebenen Signifikanzniveaus. Tests, die das Niveau eher überschreiten, nennt man antikonservativ, die es eher unterschreiten konservativ.

Die wesentliche Voraussetzung, dass ***t***-Tests in guter Näherung ihr Niveau halten, ist die Symmetrie der Verteilung der Teststatistik T unter der Nullhypothese. Diese Voraussetzung ist im Falle von zwei Stichproben im Allgemeinen unkritisch, da die Differenzen der Mittelwerte unter der Nullhypothese im Prinzip eine symmetrische Verteilung besitzen. Probleme können dann entstehen, wenn die zu vergleichenden Stichproben Grundgesamtheiten mit zwar identischen Erwartungswerten, aber ungleichen Varianzen oder unterschiedlicher Schiefe entstammen. Experimentelle Studien, die randomisiert und doppelblind durchgeführt werden, lassen jedoch zumeist eine strukturelle Gleichheit unter der Nullhypothese erwarten.

Das zweite Kriterium, an dem ein statistischer Test gemessen wird, ist seine Trennschärfe: Vorhandene, relevante Unterschiede sollen bei möglichst geringem Aufwand (Stichprobenumfang) mit hoher Sicherheit erkannt werden (s. u.). Bei normalverteilten Grundgesamtheiten ist der ***t***-Test der trennschärfste Test, und er behält seine guten Eigenschaften auch bei leichten bis mäßigen Abweichungen von der Normalverteilungsannahme. Es gibt jedoch Situationen (besonders schiefe oder ausreißerbehaftete Verteilungen), in denen der ***t***-Test ein schlechter (trennschwacher) Test ist. Hier gibt es prinzipiell zwei Lösungsansätze: Entweder eine Transformation der Daten (z. B. Logarithmus-Transformation) oder die Verwendung von Tests ohne spezielle Verteilungsannahmen, sog. „nicht-parametrische" Tests [4].

Trennschärfe

Formal ausgedrückt, bezeichnet die Trennschärfe (engl.: Power) die Wahrscheinlichkeit, die Nullhypothese abzulehnen, wenn sie tatsächlich falsch ist, also „in Wahrheit" (irgend-)ein Unterschied zwischen den Gruppen – bzw. allgemeiner, eine Abweichung von der Nullhypothese – besteht. Das „Gegenteil" der Power, also die Wahrscheinlichkeit, die Nullhypothese nicht abzulehnen, obwohl sie falsch ist, stellt die zweite Irrtumsmöglichkeit bei der Durchführung eines statistischen Tests dar; sie wird – in Analogie zum Signifikanzniveau α [1] – durch den griechischen Kleinbuchstaben β quantifiziert. Es gilt demnach: Power = $1 - \beta$ (bzw. $100 - \beta$ bei Angaben in Prozent). Neben dem verwendeten statistischen Test und dem tatsächlichen Unterschied zwischen den Gruppen hängt die Trennschärfe noch von der Variabilität und vom Stichprobenumfang ab. Das heißt, große Gruppenunterschiede bei geringer Variabilität können mit einer vergleichsweise kleinen Fallzahl statistisch entdeckt werden, kleine Unterschiede bei hoher Variabilität erfordern dagegen große Fallzahlen. Damit wird deutlich, dass zu einer guten Planung von klinischen Studien (und später dann auch zur Präsentation der Ergebnisse) insbesondere die Spezifizierung von (a) zu entdeckenden Unterschieden, (b) von der zu erwartenden Variabilität und (c) von der gewünschten Power mit der daraus resultierenden Fallzahl gehören (Fallzahlplanung). Die oben beschriebene Studie hatte mit 95 Patienten in der Verum- und 46 Patienten in der Placebogruppe eine Power von etwa 80%, um einen Unterschied in der Ödemreduktion von 50 ml zwischen Verum und Placebo bei einer Standardabweichung von 100 ml mit Hilfe eines Signifikanztests (***t***-Test) entdecken zu können. Ohne Angaben zur gewünschten Power, zum zu entdeckenden Unterschied und zur erwarteten Variabilität kann der negative (nicht signifikante) Ausfall eines statistischen Tests nicht interpretiert werden.

Dieser Beitrag ist eine überarbeitete Fassung aus dem Supplement Statistik aus dem Jahr 2001.

Literatur

1 *Bender R, Lange S.* Was ist der *p*-Wert? Dtsch Med Wochenschr 2007; 132: e15–e16
2 *Bender R, Lange S.* Was ist ein Konfidenzintervall? Dtsch Med Wochenschr 2007; 132: e17–e18
3 *Bender R, Lange S.* Verlaufskurven. Dtsch Med Wochenschr 2007; 132: e22–e23
4 *Bender R, Lange S, Ziegler A.* Wichtige Signifikanztests. Dtsch Med Wochenschr 2007; 132: e24–e25
5 *Diehm C, Trampisch HJ, Lange S, Schmidt C.* Comparsion of leg compression stocking and oral dried horse chestnut seed extract therapy in patients with chronic venous insufficiency. Lancet 1996; 347: 292–294
6 *Lange S, Bender R.* Variabilitätsmaße. Dtsch Med Wochenschr 2007; 132: e5–e6