J.B. du Prel[1], B. Röhrig[2], G. Hommel[3], M. Blettner[3]

Auswahl statistischer Testverfahren – Teil 12 der Serie zur Bewertung wissenschaftlicher Publikationen

Choosing statistical tests – part 12 of a series on evaluation of scientific publications

Hintergrund: Zur Interpretation wissenschaftlicher Artikel sind oft Kenntnisse über Verfahren der schließenden Statistik notwendig. Dieser Artikel will über häufig verwendete statistische Tests und deren richtige Anwendung informieren.

Methode: Auf der Grundlage einer selektiven Literaturrecherche zur Methodik in medizinisch-wissenschaftlichen Publikationen werden die am häufigsten verwendeten statistischen Tests identifiziert. Diese und eine Auswahl anderer Standardverfahren der schließenden Statistik werden präsentiert.

Ergebnisse/Schlussfolgerung: Leserinnen und Leser, denen neben deskriptiven Verfahren zusätzlich Pearson's Chi-Quadrat- beziehungsweise der exakte Test nach Fisher sowie der t-Test vertraut sind, können einen großen Teil der wissenschaftlichen Publikationen interpretieren, die im Bereich Humanmedizin veröffentlicht werden. Anhand häufig verwendeter Testformen werden Auswahlkriterien für statistische Tests vermittelt. Algorithmen und eine Tabelle sollen die Entscheidung für einen angemessenen statistischen Test erleichtern.

(Dtsch Zahnärztl Z 2011, 66: 510–516)

Background: The interpretation of scientific articles often requires an understanding of the methods of inferential statistics. This article informs the reader about frequently used statistical tests and their correct application.

Methods: The most commonly used statistical tests were identified through a selective literature search on the methodology of medical research publications. These tests are discussed in this article, along with a selection of other standard methods of inferential statistics.

Results and conclusions: Readers who are acquainted not just with descriptive methods, but also with Pearson's chi-square test, Fisher's exact test, and Student's t test will be able to interpret most medical research articles. Criteria are presented for choosing the proper statistical test to be used out of the more frequently applied tests. An algorithm and a table are provided to facilitate the selection of the appropriate test.

* Nachdruck aus: Dtsch Arztebl Int 2010; 107(19): 343–348; DOI: 10.3238/arztebl.2010.0343 © Deutscher Ärzte-Verlag GmbH Köln

[1] Institut für Epidemiologie, Universität Ulm
[2] MDK Rheinland-Pfalz, Referat Rehabilitation/Biometrie und Epidemiologie, Alzey
[3] Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsklinikum Mainz

Peer reviewed article: eingereicht: 14.10.209, revidierte Fassung angenommen: 22.2.2010
DOI: 10.3238/dzz.2011.0510

Einleitung

Medizinisches Wissen basiert zunehmend auf empirischen Studien, deren Ergebnisse mit statistischen Methoden dargestellt und analysiert werden. Kenntnisse über häufig verwendete statistische Tests sind daher für jeden Arzt vorteilhaft. Nur so kann er/sie die statistische Methodik in wissenschaftlichen Publikationen beurteilen und damit die Studienergebnisse richtig interpretieren. Im Folgenden werden häufig verwendete statistische Tests für unterschiedliche Skalenniveaus und Stichprobenarten vorgestellt. Ausgehend vom einfachsten Fall werden Entscheidungshilfen zur Auswahl statistischer Tests präsentiert.

Häufig verwendete statistische Tests in medizinischen Studien

Die Analyse von 1.828 Publikationen aus sechs Fachjournalen (Allgemeinmedizin, Gynäkologie und Geburtshilfe, Notfallmedizin) ging der Frage nach, welche statistischen Tests in medizinischen Zeitschriften oft angewandt werden. Das Resultat ergab, dass Leser, die neben deskriptiven Verfahren zusätzlich mit Pearson's Chi-Quadrat- beziehungsweise dem exakten Test nach Fisher sowie dem t-Test vertraut sind, zumindest 70 % der Artikel statistisch richtig interpretieren können [9]. Damit wurden frühere Ergebnisse zu häufig verwendeten statistischen Tests in der medizinisch-wissenschaftlichen Literatur bestätigt [5, 6]. Das Spektrum der verwendeten statistischen Tests unterliegt jedoch zeitlichen Veränderungen. Nach einer Auswertung von verwendeten statistischen Analyseverfahren in Publikationen des ersten Halbjahres 2005 der Zeitschrift Pediatrics nahm der Anteil von Methoden der statistischen Inferenz zwischen 1982 und 2005 von 48 % auf 89 % zu [8]. Daneben zeigte sich ein Trend hin zu komplexeren statistischen Testverfahren. Am häufigsten waren allerdings auch hier der t-Test und der Chi-Quadrat-Test beziehungsweise der exakte Test nach Fisher. Daher werden diese und weitere grundlegende statistische Tests einschließlich ihrer Anwendung in diesem Artikel vorgestellt. Mit Kenntnis dieser überschaubaren Testauswahl sollte der Leser einen großen Teil medizinischer Publikationen interpretieren können. Für seltenere statistische Tests wird auf die jeweilige Artikelbeschreibung, auf weiterführende Literatur [1, 3, 12] und auf die Konsultation eines erfahrenen Statistikers verwiesen.

Sinn und Zweck statistischer Tests

In klinischen Studien (zum Beispiel [1, 10]) werden beispielsweise oft Vergleiche zwischen einer Studiengruppe, die ein neues Präparat erhält, und einer Kontrollgruppe, die ein schon etabliertes oder ein Placebo bekommt, hinsichtlich der Wirksamkeit gezogen. Neben der reinen Deskription [13] möchte man wissen, ob die beobachteten Unterschiede zwischen den Behandlungsgruppen lediglich zufällig oder tatsächlich vorhanden sind. Unterschiede könnten ja durch zufällige Variabilität (= Streuung) des Merkmales, also zum Beispiel des Therapieerfolges innerhalb der Studienpersonen, hervorgerufen werden.

Definition

Soll bei einer wissenschaftlichen Fragestellung ein Vergleich zwischen zwei oder mehr Gruppen untersucht werden, so kann man einen statistischen Test durchführen. Dazu muss eine geeignete Nullhypothese, die es zu widerlegen gilt, formuliert und eine geeignete Prüfgröße aufgestellt werden [4, 14].

Wird beispielsweise in einer klinischen Studie untersucht, ob ein Blutdrucksenker besser wirkt als ein Placebo, ist der zu untersuchende Effekt die Reduktion des diastolischen Blutdruckes gemessen anhand der mittleren Blutdruckdifferenz in der Verum- und Placebogruppe. Entsprechend formuliert man dann als Nullhypothese: „Verum und Placebo unterscheiden sich hinsichtlich ihrer blutdrucksenkenden Wirkung nicht" (Effekt = 0).

Ein statistischer Test berechnet dann die Wahrscheinlichkeit, die beobachteten Daten (oder noch extremere) zu erhalten, falls die Nullhypothese zutrifft. Ein kleiner p-Wert besagt, dass diese Wahrscheinlichkeit gering ist. Unterschreitet der p-Wert eine vorab definierte Signifikanzschranke, wird die Nullhypothese abgelehnt. Aus den beobachteten Daten wird eine Prüfgröße (Teststatistik) berechnet, die die Grundlage für den statistischen Test bildet (zum Beispiel Differenz des mittleren Blutdrucks nach sechs Monaten). Mit bestimmten Annahmen über die Verteilung der Daten (zum Beispiel Normalverteilung) kann die theoretische (erwartete) Verteilung der Prüfgröße berechnet werden.

Ablauf eines statistischen Tests

- Aufstellung der Forschungsfrage
- Formulierung von Null- und Alternativhypothese
- Entscheidung für einen geeigneten statistischen Test
- Festlegen des Signifikanzniveaus (zum Beispiel 0,05)
- Durchführen der statistischen Testanalyse: Berechnung des p-Wertes
- Statistische Entscheidung, zum Beispiel
  - $p < 0{,}05$ => Verwerfen der Nullhypothese und Annehmen der Alternativhypothese
  - $p \geq 0{,}05$ Beibehalten der Nullhypothese
- Interpretation des Testergebnisses

Der aus den Beobachtungen berechnete Wert der Prüfgröße wird mit der Verteilung, die man erwarten würde, wenn die Nullhypothese zutrifft, verglichen [1]. Übersteigt oder unterschreitet sie eine bestimmte Größe, die bei Gültigkeit der Nullhypothese wenig wahrscheinlich ist, so wird die Nullhypothese verworfen: das Ergebnis ist „statistisch signifikant zum Niveau α". Der statistische Test ist also eine Entscheidung, ob die beobachtete Größe noch mit Zufall zu erklären ist oder ob sie überzufällig ist (statistisch signifikant). Die Begriffe „Signifikanzniveau" und das Prinzip der Interpretation von p-Werten wurden bereits erörtert [4, 14]. Der grundlegende Ablauf eines statistischen Testes ist im Kasten noch einmal dargestellt.

Sowohl bei Ablehnung als auch bei Beibehaltung der Nullhypothese kann man einen Fehler machen. Das liegt daran, dass die Werte eine gewisse Streuung aufweisen, da zum Beispiel nicht alle Patienten gleich auf ein Medikament reagieren. Für den Fehler erster Art, also die Nullhypothese irrtümlich abzulehnen, entspricht die maximale Irrtumswahrscheinlichkeit dem Signifikanzniveau α. Häufig wird dafür 5 % gewählt [4, 14]. Die Wahrscheinlichkeit für den Fehler

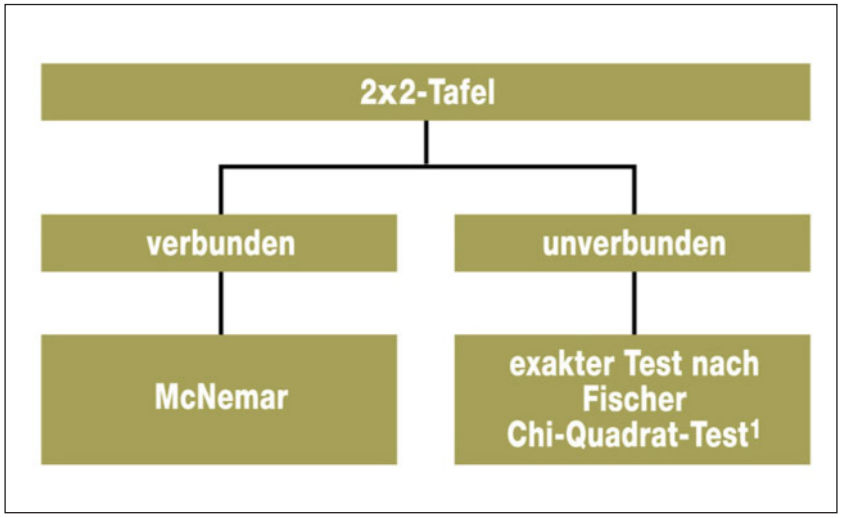
Statistischer Test	Beschreibung
Exakter Test nach Fisher	Geeignet für binäre Daten in unverbundenen Stichproben (2 x 2-Tafel) zum Vergleich der Behandlungseffekte oder der Nebenwirkungshäufigkeiten in zwei Behandlungsgruppen.
Chi-Quadrat-Test	Ähnlich dem exakten Test nach Fisher (allerdings ungenauer), kann man auch mehr als zwei Gruppen sowie mehr als zwei Kategorien der Zielgröße miteinander vergleichen (Voraussetzungen: Fallzahl etwa > 60, erwartete Anzahl in jedem Feld ≥ 5).
Mc-Nemar-Test	Voraussetzungen vergleichbar dem exakten Test nach Fisher, allerdings für verbundene Stichproben.
Student's t-Test	Test für kontinuierliche Daten, untersucht, ob die Erwartungswerte zweier Gruppen gleich sind unter Annahme der Normalverteilung der Daten. Es gibt den Test für gepaarte und ungepaarte Gruppen.
Varianzanalyse	Testvoraussetzungen wie unverbundener t-Test für den Vergleich von mehr als zwei Gruppen. Methoden der Varianzanalyse werden auch beim Vergleich von mehr als zwei verbundenen Gruppen angewendet.
Wilcoxon-Rangsummentest (Anm.: synonym für den unverbundenen Wilcoxon-Rangsummentest wird auch die Bezeichnung Mann-Whitney U-Test verwendet)	Test für ordinale oder kontinuierliche Daten, erfordert im Unterschied zum Student's t-Test keine Normalverteilung der Daten. Auch hier existiert eine Form für gepaarte oder ungepaarte Gruppen.
Kruskal-Wallis Test	Testvoraussetzungen wie unverbundener Wilcoxon-Rangsummentest für den Vergleich von mehr als zwei Gruppen.
Friedman-Test	Vergleich von mehr als zwei verbundenen, mindestens ordinalskalierten Stichproben.
Logrank-Test	Test zur Überlebenszeitanalyse zum Vergleich von zwei und mehr unabhängigen Gruppen.
Korrelationstest nach Pearson	Untersucht, ob zwischen zwei stetigen normal-verteilten Variablen ein linearer Zusammenhang besteht
Korrelationstest nach Spearman	Untersucht, ob zwischen zwei stetigen oder mindestens ordinalen Variablen ein monotoner Zusammenhang besteht.

Tabelle 1 Häufig verwendete statistische Tests (modifiziert nach [6]).
Table 1 Frequently used statistical tests (modified from [6]).

zweiter Art (β), also die Nullhypothese irrtümlich beizubehalten, ist 1 minus der Power der Studie. Die Power der Studie wird vor Studienbeginn festgelegt und hängt unter anderem von der Fallzahl ab. Häufig wird eine Teststärke von 80 % gewählt [4, 14].

Wichtige Schritte bei der Entscheidung für einen statistischen Test

Die Entscheidung für einen statistischen Test erfolgt auf Grundlage der wissenschaftlichen Fragestellung, der Datenstruktur und des Studiendesigns. Vor der Datenerhebung – und damit natürlich auch vor der Wahl des statistischen Tests – müssen die Fragestellung und die Nullhypothese formuliert werden. Test und Signifikanzniveau sind vor Studiendurchführung im Studienprotokoll fest-



Grafik 1 Testauswahl beim Gruppenvergleich von zwei Kategorialen Zielgrößen: *1Voraussetzungen: Fallzahl > 60, erwartete Anzahl pro Feld ≥ 5.

Figure 1 Test selection for group comparison with two categorial endpoints. *1Preconditions: sample size > ca. 60. expected number in early field ≥ 5.

zuhalten. Dabei muss entschieden werden, ob ein- oder zweiseitig getestet werden soll. Zweiseitig bedeutet, dass die Richtung des erwarteten Unterschiedes unklar ist. Man weiß also nicht, ob ein Wirksamkeitsunterschied zwischen Verum und Placebo besteht und lässt offen, in welche Richtung dieser Unterschied gehen könnte (Verum könnte sogar schlechter wirken als Placebo). Ein einseitiger Test sollte nur dann durchgeführt werden, wenn es eine klare Evidenz dafür gibt, dass eine Intervention nur in eine Richtung wirken kann.

Mit der Formulierung der Fragestellung wird auch die Zielgröße (Endpunkt) festgelegt. Für die Wahl des geeigneten statistischen Tests sind zwei Kriterien entscheidend:

– das Skalenniveau der Zielgröße (stetig, binär, kategorial)
– die Art des Studiendesigns (verbunden oder unverbunden).

Skalenniveau: stetig, kategorial oder binär

Die unterschiedlichen Skalenniveaus wurden bereits bei der Wahl der geeigneten Maßzahlen beziehungsweise bei der Wahl grafischer Darstellungsformen in dem Artikel zur deskriptiven Statistik erörtert [11, 13].

Beim Vergleich zweier Antihypertensiva kann der Endpunkt beispielsweise die blutdrucksenkende Wirkung in beiden Behandlungsgruppen sein. Blutdrucksenkung ist eine stetige Zielgröße. Bei einer stetigen Zielgröße ist weiterhin zu unterscheiden, ob sie (angenähert) normalverteilt ist oder nicht.

Würde man beispielsweise nur berücksichtigen, ob der diastolische Blutdruck unter 90 mm Hg gefallen ist oder nicht, so wäre die Zielgröße kategorial (sie wäre sogar binär, da es nur zwei mögliche Ergebnisse gibt). Wenn sich der Wertebereich eines kategorialen Endpunkts sinnvoll ordnen lässt, so spricht man in diesem Fall auch von einem ordinalen Endpunkt.

Unverbundene und verbundene Studiendesigns

Mittels eines statistischen Tests werden die Ergebnisse der Zielgröße für verschiedene Versuchsbedingungen (zum Beispiel Behandlungen) miteinander verglichen; oft dreht es sich dabei um zwei Therapien.

Ist es möglich, für jeden Patienten Ergebnisse unter allen Versuchsbedingungen zu erhalten, so handelt es sich um ein verbundenes (abhängiges) Design. Ein verbundenes Studiendesign läge beim Vergleich von zwei Messzeitpunkten vor, aber auch dann, wenn es sich in zwei Gruppen hinsichtlich bestimmter Merkmale um „Paare“ handelt.

Typisches Beispiel für „Paare“ sind Untersuchungen, die jeweils an einem Auge oder einem Arm derselben Person durchgeführt werden. Typisch für verbundene Designs sind auch Vergleich vor und nach der Behandlung. Eine Besonderheit bilden die „matched pairs“, zum Beispiel in Fall-Kontroll-Studien. Hierbei werden für Probanden aus einer Gruppe hinsichtlich bestimmter Merkmale gleiche Personen aus anderen Gruppen gewählt. Damit sind die Daten nicht mehr unabhängig und sollten so behandelt werden als wären es gepaarte Beobachtungen aus einer Gruppe [1].

Bei einem unverbundenen (unabhängigen) Studiendesign liegen für jeden Patienten nur die Ergebnisse unter einer einzigen Versuchsbedingung vor. Verglichen werden dann die Resultate zweier (oder mehrerer) Gruppen. Hier kann sich die Größe der zu untersuchenden Gruppen unterscheiden.

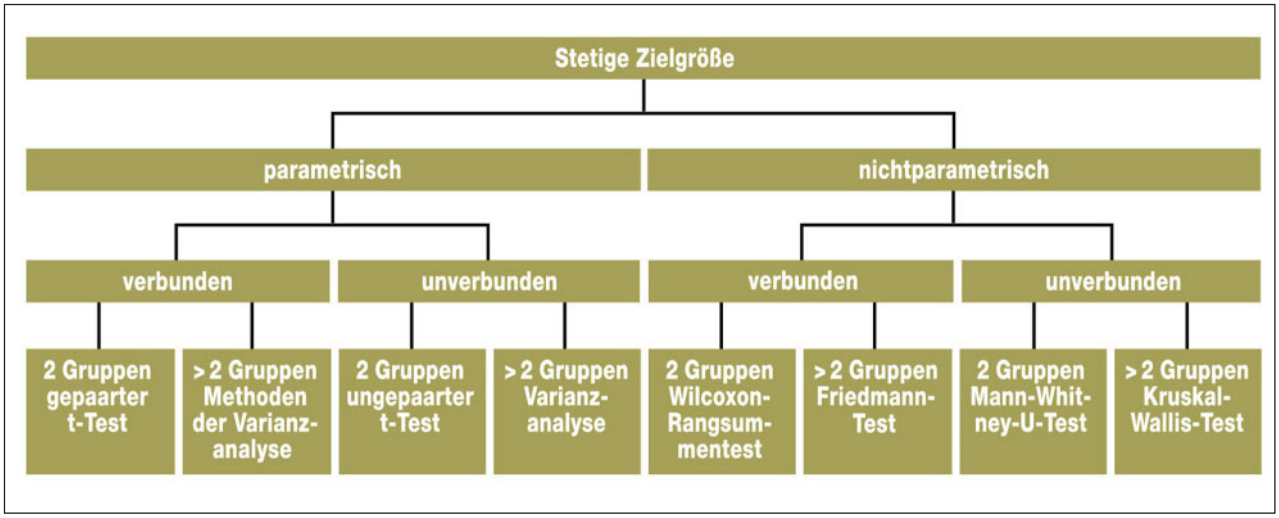
Vorstellung häufiger statistischer Tests

Die wichtigsten statistischen Tests werden in der Tabelle aufgezeigt. Dabei wird immer unterschieden zwischen „kategorial oder stetig“ und „verbunden oder unverbunden“. Ist die Zielgröße stetig, so wird noch unterteilt in normalverteilte und nichtnormalverteilte Größen (Tabelle 1).

Gruppenvergleich von zwei kategorialen Zielgrößen

Der Gruppenvergleich zweier kategorialer Zielgrößen wird hier anhand des einfachsten Falles einer 2 × 2-Tafel (Vierfeldertafel) dargestellt (Grafik 1). Ähnlich wird auch beim Gruppenvergleich mehrstufiger kategorialer Zielgrößen verfahren (Tabelle 1).

– Unverbundene Stichproben:
  Soll die Häufigkeit des Erfolges in zwei Behandlungsgruppen verglichen werden, ist der richtige statistische Test, insbesondere bei kleiner Stichprobengröße, der exakte Test nach Fisher [3]. Bei großem Stichprobenumfang (etwa n > 60) kann auch der Chi-Quadrat-Test durchgeführt werden (Tabelle 1).
– Verbundene Stichproben:
  Ein Beispiel für die Anwendbarkeit dieser Testform ist eine Intervention innerhalb einer Gruppe an zwei Stellen, also zum Beispiel die Implantati-



Grafik 2 Algorithmus zur Testauswahl beim Gruppenvergleich einer stetigen Zielgröße.
Figure 2 Algorithm for test selection for group comparison of a continuous endpoint.

on zweier verschiedener Arten von IOL-Linsen in das rechte und linke Auge mit der Zielgröße „Operationserfolg Ja oder Nein". Die zu vergleichenden Stichproben sind verbunden. In diesem Fall muss man den McNemar-Test durchführen [3].

Stetige und mindestens ordinal skalierte Variablen

Ein Entscheidungsalgorithmus für die Testauswahl findet sich in Grafik 2.

Normalverteilte Variablen – Parametrische Tests:
Wenn die Zielgröße normalverteilt ist, dann können zum statistischen Testen sogenannte parametrische Testverfahren eingesetzt werden.

– Unverbundene Stichproben:
Falls die Probanden beider Gruppen unverbunden voneinander sind (das heißt, die Personen der ersten Gruppe sind andere im Vergleich zur zweiten Gruppe), wird bei normalverteilten, stetigen Merkmalen der unverbundene t-Test angewendet. Werden mehr als zwei unabhängige (unverbundene) Gruppen hinsichtlich eines normalverteilten, stetigen Merkmals miteinander verglichen, ist die Varianzanalyse (ANOVA, „analysis of variance") geeignet (zum Beispiel Studie mit drei oder mehr Therapiearmen). Die ANOVA stellt eine Verallgemeinerung des unverbundenen t-Tests dar. Die ANOVA gibt nur Auskunft darüber, ob sich die Gruppen unterscheiden, aber nicht darüber welche. Hierzu sind Methoden des multiplen Testens erforderlich [14].

– Verbundene Stichproben:
Im Fall eines normalverteilten, stetigen Merkmals bei zwei verbundenen Gruppen wird der verbundene t-Test verwendet. Werden mehr als zwei verbundene Gruppen hinsichtlich eines normalverteilten, stetigen Merkmals miteinander verglichen, sind ebenfalls auf der Varianzanalyse basierende Methoden geeignet. Der Faktor beschreibt die verbundenen Gruppen, zum Beispiel mehr als zwei Erhebungspunkte bei einer Therapieanwendung.

Nichtnormalverteilte Variablen – nichtparametrische Tests: Ist das interessierende Merkmal nicht normalverteilt, aber mindestens ordinalskaliert, dann werden zum statistischen Testen nichtparametrische Testverfahren eingesetzt. Ein solcher Test („Rangtest") basiert nicht direkt auf den beobachteten Werten, sondern auf daraus abgeleiteten Rangzahlen (die Werte werden dazu ihrer Größe nach geordnet und fortlaufend nummeriert). Aus diesen Rangzahlen wird dann die Prüfgröße des statistischen Tests berechnet. Wenn die Voraussetzungen erfüllt sind, sind parametrische Tests trennschärfer als nichtparametrische. Sind sie nicht erfüllt, kann die Trennschärfe der parametrischen Tests jedoch drastisch sinken.

– Unverbundene Stichproben:
Beim Vergleich zweier unverbundener Stichproben bezüglich eines nichtnormalverteilten, jedoch mindestens ordinalskalierten Merkmals kann der Mann-Whitney U-Test (= Wilcoxon-Rangsummentest) eingesetzt werden [1]. Sind mehr als zwei unverbundene Stichproben zu vergleichen, so kann der Kruskal-Wallis-Test als Verallgemeinerung des Mann-Whitney U-Tests eingesetzt werden [7]

– Verbundene Stichproben:
Beim Vergleich zweier verbundener Stichproben bezüglich eines nichtnormalverteilten, jedoch mindestens ordinalskalierten Merkmals kann der Wilcoxon-Vorzeichenrangtest eingesetzt werden [7]. Alternativ, wenn die Differenz der beiden Werte nur eine binäre Unterscheidung ermöglicht (zum Beispiel Verbesserung versus Verschlechterung), ist der Vorzeichentest anzuwenden [3]. Beim Vergleich von mehr als zwei verbundenen Stichproben kann der Friedman-Test als Verallgemeinerung des Vorzeichentests eingesetzt werden.

Andere Testverfahren

Überlebenszeitanalyse

Interessiert nicht der Endpunkt selbst, sondern die Zeit bis zum Erreichen desselben, ist die Überlebenszeitanalyse das

geeignete Verfahren. Dabei werden zwei oder mehrere Gruppen bezüglich der Zeiten bis zum Erreichen eines Endpunktes innerhalb eines Beobachtungszeitraumes miteinander verglichen [7]. Ein Beispiel ist der Vergleich der Überlebenszeit von Patienten aus zwei Gruppen mit einer onkologischen Erkrankung und zwei unterschiedlichen Chemotherapien. Zielkriterium ist hier der Tod, könnte aber auch das Auftreten von Metastasen sein. Im Unterschied zu den vorangegangenen Tests kann bei der Überlebenszeitanalyse aufgrund der begrenzten Beobachtungszeit fast nie bei allen Subjekten das Erreichen des Endpunktes vernommen werden. Deshalb werden die Daten auch als (rechts)zensiert bezeichnet, da man zum Beobachtungsende nicht bei allen Probanden weiß, wann sie den Endpunkt erreichen werden. Der übliche statistische Test für den Vergleich der Überlebensfunktionen zwischen zwei oder mehreren Gruppen ist der Log-rank-Test. Aus den beobachteten und den erwarteten Zahlen an Ereignissen wird anhand einer Formel ein bestimmter Wert, die Prüfgröße, berechnet. Dieser Wert kann dann mit einer bekannten Verteilung, die man erwarten würde falls die Nullhypothese zutrifft, hier die Chi^2-Verteilung, verglichen und ein p-Wert ermittelt werden. Damit kann eine Entscheidungsregel für oder gegen die Nullhypothese angegeben werden.

Korrelationsanalyse

Die Korrelationsanalyse untersucht die Stärke des Zusammenhangs zwischen zwei Zielgrößen, zum Beispiel wie stark das Körpergewicht von Neugeborenen mit ihrer Körpergröße korreliert. Die Wahl eines geeigneten Assoziationsmaßes hängt vom Skalenniveau und der Verteilung beider Größen ab. Während die parametrische Variante, der Korrelationskoeffizient nach *Pearson*, ausschließlich lineare Zusammenhänge zwischen stetigen Merkmalen prüft, untersucht die nichtparametrische Alternative, der Rangkorrelationskoeffizient nach *Spearman*, lediglich monotone Beziehungen bei mindestens ordinal-skalierten Merkmalen. Vorteil des Letzteren ist seine Robustheit gegenüber Ausreißern und schiefen Verteilungen. Korrelationskoeffizienten messen die Assoziationsstärke und können Werte zwischen –1 und +1 annehmen. Je näher sie an 1 liegen, desto stärker ist der Zusammenhang. Aus dem Korrelationskoeffizient kann wiederum eine Prüfgröße und damit ein statistischer Test konstruiert werden. Die Nullhypothese, die geprüft werden soll, heißt hier: Es liegt kein linearer (beziehungsweise monotoner) Zusammenhang vor.

Diskussion

Neben den vorgestellten statistischen Tests, bei denen in der Nullhypothese Gleichheit der Gruppen formuliert ist, gibt es noch andere Testverfahren. Trendtests prüfen, ob es bei mindestens drei Gruppen eine Tendenz zu steigenden oder fallenden Werten gibt.

Zu den häufig vorkommenden Ungleichheitstests („inequality tests“), bei denen die Nullhypothese von Gleichheit zwischen den Gruppen ausgeht, existieren Überlegenheitstests („superiority tests“), Nichtunterlegenheitstests („non-inferiority tests“) und Äquivalenztests („equivalence tests“). Beim Überlegenheitstest wird zum Beispiel von einer neuen, teureren Medikation gefordert, dass sie um eine bestimmte, medizinisch sinnvolle Differenz besser als eine gängige Standardmedikation ist. Beim Nichtunterlegenheitstest wird zum Beispiel von einer neuen, kostengünstigeren Medikation verlangt, dass sie nicht viel schlechter als eine gängige ist. Welche Wirkung noch akzeptabel ist, wird aufgrund medizinischen Sachverstands vor Studienbeginn festgelegt. Bei Äquivalenztests soll gezeigt werden, dass die Medikation eine annähernd gleich große Wirkung wie eine gängige Standardmedikation hat. Vorteile der neuen Medikation können vereinfachte Applikation, weniger Nebenwirkungen oder Kostensenkung sein.

Auf die Methoden der Regressionsanalyse und statistische Tests im Zusammenhang damit wird im Rahmen der Serie zur Bewertung wissenschaftlicher Publikationen noch näher eingegangen.

Die vorliegende Auswahl an statistischen Tests muss unvollständig sein. Es sollte herausgestellt werden, dass die Wahl eines geeigneten Testverfahrens von Kriterien wie dem Skalenniveau der Zielgröße und der zugrunde liegenden Verteilung abhängt. Dem interessierten Leser sei das Buch von *Altman* [1] als praxisnahe Darstellung empfohlen. Für nichtparametrische Tests bietet *Bortz* et al. [3] eine umfangreiche Übersicht.

Mit der Entscheidung für einen statistischen Test vor Studienbeginn wird ausgeschlossen, dass die Studienergebnisse die Testauswahl beeinflussen. Von der Wahl des Testverfahrens hängt zudem die benötigte Fallzahl ab. Auf die Problematik der Fallzahlplanung wird im Rahmen dieser Serie noch näher eingegangen.

Abschließend ist es wichtig festzustellen, dass nicht in jeder Studie ein statistischer Test erforderlich ist. In rein deskriptiven Studien [12] oder wenn sich Zusammenhänge durch naturwisssenschaftliche Plausibilität oder sachlogische Überlegungen ergeben, kann auf die Anwendung eines statistischen Testes verzichtet werden. Bei Untersuchung der Güte eines diagnostischen Testverfahrens oder der Raterübereinstimmmung (zum Beispiel in Form von Bland-Altman-Diagrammen) ist normalerweise ein statistischer Test ebenfalls nicht sinnvoll [2]. Wegen den genannten Irrtumswahrscheinlichkeiten gilt bei Anwendung von statistischen Tests: „So viel wie nötig, so wenig wie möglich“. Die Wahrscheinlichkeit rein zufälliger Ergebnisse ist besonders beim multiplen Testen hoch [14].

Interessenkonflikt: Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Korrespondenzadressen

Dr. med. Jean-Baptist du Prel, MPH
Institut für Epidemiologie
Universität Ulm
Helmholtzstr. 22
89081 Ulm
Tel.: 07 31 / 5 03 10 60
Fax: 07 31 / 5 03 10 69
E-Mail: jean-baptist.du-prel@uni-ulm.de

Prof. Dr. Maria Blettner
Universitätsmedizin der Johannes Gutenberg-Universität Mainz
Institut für Medizinische Biometrie, Epidemiologie und Informatik
Obere Zahlbacher Straße 69
55131 Mainz
Tel.: 0 61 31 / 17 – 32 52
Fax: 0 61 31 / 17 – 29 68
E-Mail: maria.blettner@unimedizin-mainz.de
www.imbei.uni-mainz.de

Literatur

1. Altman DG: Practical statistics for medical research. Chapman and Hall, London 1991
2. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310 (1986)
3. Bortz J, Lienert GA, Boehnke K: Verteilungsfreie Methoden in der Biostatistik. 2. Auflage. Springer, Berlin, Heidelberg, New York 2000
4. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications [Konfidenzintervall oder p-Wert? Teil 4 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 106, 335–339 (2009)
5. Emerson JD, Colditz GA: Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 309, 709–713 (1983)
6. Goldin J, Zhu W, Sayre JW: A review of the statistical analysis used in papers published in Clinical Radiology and British Journal of Radiology. Clin Radiol 51, 47–50 (1996)
7. Harms V: Biomathematik, Statistik und Dokumentation: Eine leichtverständliche Einführung. 7th edition revised. Lindhöft, Harms 1998
8. Hellems MA, Gurka MJ, Hayden GF: Statistical literacy for readers of Pediatrics: a moving target. Pediatrics 119, 1083–1088 (2007)
9. Reed JF 3rd, Salen P, Bagher P: Methodological and statistical techniques: what do residents really need to know about statistics? J Med Syst 27, 233–238 (2003)
10. Röhrig B, du Prel JB, Wachtlin D, Blettner M: Types of study in medical research – part 3 of a series on evaluation of scientific publications [Studientypen in der medizinischen Forschung: Teil 3 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 106, 262–268 (2009)
11. Röhrig B, du Prel JB, Blettner M: Study design in medical research – part 2 of a series on evaluation of scientific publications [Studiendesign in der medizinischen Forschung: Teil 2 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 106, 184–189 (2009)
12. Sachs L: Angewandte Statistik: Anwendung statistischer Methoden. 11. Auflage. Springer, Berlin, Heidelberg, New York 2004
13. Spriestersbach A, Röhrig B, du Prel JB, Gerhold-Ay A, Blettner M: Descriptive statistics: the specification of statistical measures and their presentation in tables and graphspart – part 7 of a series on evaluation of scientific publications [Deskriptive Statistik: Angabe statistischer Maßzahlen und ihre Darstellung in Tabellen und Grafiken: Teil 7 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 106, 578–583 (2009)
14. Victor A, Elsässer A, Hommel G, Blettner M: Judging a plethora of p-values: how to contend with the problem of multiple testing – part 10 of a series on evaluation of scientific publications [Wie bewertet man die p-Wert-Flut – Hinweise zum Umgang mit dem multiplen Testen – Teil 10 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 107, 50–56 (2010)