Wichtige Signifikanztests

– Artikel Nr. 11 der Statistik-Serie in der DMW –

Common significance tests

Autoren **R. Bender** [1] **S. Lange** [1] **A. Ziegler** [2]

Institut
[1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

Schlüsselwörter
- Signifikanztest
- Messniveau
- Abhängige Stichproben
- Nichtparametrische Methoden

Key words
- Significance test
- Measurement scale
- Dependent samples
- Nonparametric methods

Bibliografie
DOI 10.1055/s-2007-959034
Dtsch Med Wochenschr 2007; 132: e24–e25 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

Zum statistischen Nachweis von Unterschieden oder Effekten werden in der medizinischen Forschung häufig Signifikanztests [6] verwendet. Die Grundprinzipien statistischer Tests, die von Lange und Bender [6] am Beispiel des ungepaarten ***t***-Tests erläutert werden, gelten auch für andere Testverfahren. Der ***t***-Test kommt für viele Testsituationen der medizinischen Forschung in Frage, er deckt aber nicht alle Anwendungsgebiete ab. Je nach Fragestellung und Datensituation benötigt man andere statistische Testmethoden. Die für medizinische Anwendungen wichtigsten statistischen Tests für die einfachen Standardsituationen werden in dieser Arbeit kurz und übersichtlich zusammengefasst.

Zur Auswahl eines adäquaten Tests benötigt man im Wesentlichen die folgenden Informationen:
- die Zahl der zu analysierenden Stichproben
- den Abhängigkeitsstatus der Stichproben
- das Messniveau und die Verteilung der zu analysierenden Zielvariable.

Bei der **Stichprobenanzahl** genügt die Unterscheidung in 1, 2 oder mehr als 2 Stichproben. Die korrekte Zahl der vorliegenden Stichproben ergibt sich aus der Fragestellung. Möchte man z. B. untersuchen, ob die Erfolgswahrscheinlichkeit einer Behandlungsmethode signifikant über 50% liegt, so ergibt sich hieraus ein Einstichprobenproblem. Möchte man dagegen die Erfolgswahrscheinlichkeiten zwischen 2 oder mehr als 2 Behandlungsmethoden vergleichen, so liegen 2 bzw. mehr als 2 Stichproben vor.

Der **Abhängigkeitsstatus** der Stichproben ergibt sich aus dem Studiendesign. Beim Vergleich unverbundener Gruppen liegen ***unabhängige*** Stichproben vor. Der Vergleich mehrerer Behandlungsmethoden anhand paralleler Therapiegruppen in randomisierten klinischen Studien und der Vergleich exponierter und nicht exponierter Personen in Kohortenstudien stellen Standardbeispiele für unabhängige Stichproben dar. Beim Vergleich von abhängigen Werten an gleichen oder über gewisse Merkmale verbundenen Untersuchungseinheiten liegen ***abhängige*** Stichproben vor. Häufige Designs, die zu abhängigen Stichproben führen, sind Crossover-Studien, bei denen alle Patienten die zu vergleichenden Therapien in verschiedenen Studienphasen (meist in zufälliger Reihenfolge) erhalten und gematchte Fall-Kontroll-Studien, bei denen jedem Patient mit der interessierenden Erkrankung („Fälle“) ein oder mehrere Patienten ohne diese Erkrankung („Kontrollen“) so zugeordnet werden, dass Fälle und Kontrollen für bestimmte wichtige Merkmale (z. B. Alter und Geschlecht) gleiche oder zumindest ähnliche Ausprägungen besitzen.

Bezüglich des **Messniveaus** der Zielvariablen genügt in der Praxis die Unterscheidung zwischen den Messniveaus ***binär*** (ja/nein, z. B. Dialysenotwendigkeit), ***nominal*** (ungeordnete Kategorien, z. B. Blutgruppe), ***ordinal*** (geordnete Kategorien, z. B. Retinopathiestadien), ***stetig*** (quantitatives Merkmal, z. B. systolischer Blutdruck) und ***zensiert*** (Überlebenszeiten, z. B. Zeit bis zum Tod) [3]. Das Messniveau der Zielvariablen bezieht sich auf die Datenerfassung bei den Untersuchungseinheiten (z. B. Patienten), nicht auf Größen, die aus den Stichprobendaten berechnet werden und diese zusammenfassen. Beispielsweise liegt bei der Untersuchung der Erfolgswahrscheinlichkeit einer Behandlungsmethode in einer definierten Gruppe von Patienten eine binäre Zielvariable vor. Die Erfolgswahrscheinlichkeit ist zwar stetig (zwischen 0 und 1), sie stellt aber eine Größe dar, die aus den einzelnen Stichprobenwerten berechnet wurde; auf der Ebene der Patienten liegt eine binäre Zielvariable vor, nämlich Erfolg ja/nein.

Die üblichen parametrischen Verfahren für stetige Daten setzen strenggenommen auch noch **Normalverteilung** [1] voraus. Allerdings ist diese Voraussetzung bei großen Stichproben im Allgemeinen vernachlässigbar [8]. Bei kleinen Stichproben mit stetigen aber nicht normalverteilten Daten können die entsprechenden **nichtparametrischen Tests**, für die mindestens ordinales Messniveau erforderlich ist, verwendet werden. Das Grundprinzip der nichtpa-

Tab. 1 Übersicht über die wichtigsten statistischen Signifikanztests.

Stichproben		Messniveau der Zielvariable		*t*-Test		
Anzahl	**Status**	**binär**	**nominal**	**ordinal o. stetig (nicht normal-verteilt)**	**stetig und normalverteilt**	**zensiert**
1	–	Binominaltest χ^2-Test	χ^2-Test	(Vor)zeichentest Wilcoxon Vorzeichenrangtest	*t*-Test	–
2	unabhängig	χ^2-Test Fisher's exakter Test	χ^2-Test Fisher's exakter Test	Wilcoxon Rangsummentest = Mann-Whitney U-Test	ungepaarter *t*-Test	Log-Rang-Test
> 2	unabhängig	χ^2-Test Fisher's exakter Test	χ^2-Test Fisher's exakter Test	Kruskal-Wallis-Test	*F*-Test (ANOVA)	Log-Rang-Test
2	abhängig	McNemar Test	–	(Vor)zeichentest Wilcoxon Vorzeichenrangtest	gepaarter *t*-Test	–
> 2	abhängig	Cochran's Q	–	Friedman-Test	ANOVA für Messwert-wiederholungen	–

Beispiel: Für den statistischen Vergleich von 3 unabhängigen Gruppen (z. B. bei einer randomisierten 3-armigen klinischen Studie) bezüglich einer ordinalen Zielvariable (z. B. einem Score zur Lebensqualität) kann man den Kruskal-Wallis-Test verwenden.

rametrischen Tests ist die so genannte Rangbildung. In die Berechnung der entsprechenden Teststatistiken gehen nicht die Daten selbst, sondern deren Ränge, d. h. deren Platzierung in der vom kleinsten zum größten Wert sortierten Stichprobe, ein. Ein häufiges Problem in der Praxis stellen hierbei Bindungen dar, d. h. gleiche Ränge durch identische Stichprobenwerte, vor allem bei ordinalen Zielvariablen mit wenigen Kategorien. Eine geringe Anzahl von Bindungen kann in der Praxis vernachlässigt werden. Ist die Anzahl der Bindungen beträchtlich, so sollte man diese berücksichtigen. Meist wird die Bildung von Durchschnittsrängen empfohlen. Für die meisten nichtparametrischen Tests liegt eine so genannte bindungskorrigierte Version vor, die bei großen Stichproben mit vielen Bindungen verwendet werden sollte. Bei kleinen Stichproben ($n < 10$) mit oder ohne Bindungen sollte in jedem Fall die exakte Verteilung der Teststatistik berechnet werden, anstelle der sonst üblichen Approximationen. Die Verwendung **exakter Tests** ist meist mit einem enormen Rechenaufwand verbunden, allerdings gibt es seit einigen Jahren hierfür spezielle Statistik-Software wie z. B. StatXact [7].

Sowohl die üblichen parametrischen als auch die nichtparametrischen Tests setzen die so genannte **Homoskedastizität**, d. h. identische Varianzen in den verschiedenen Gruppen, voraus. Diese Annahme ist in der Praxis häufig verletzt. Im Bereich der Therapiestudien findet man z. B. oftmals eine größere Streuung in der behandelten Gruppe, die sich möglicherweise durch ein unterschiedliches Ansprechverhalten auf die Therapie erklären lässt. Ist die Annahme der Homoskedastizität deutlich verletzt, so sollten **modifizierte Tests** verwendet werden, die keine identischen Varianzen voraussetzen [9].

In **Tab. 1** sind die wichtigsten Signifikanztests für die häufigsten und einfachsten Standardsituationen aufgeführt. Es gibt eine Reihe von weiteren Verfahren, die sich wegen ihrer Komplexität nicht in einer einfachen Übersicht zusammenfassen lassen. Ein in der Praxis häufiges Problem ist z. B. der Vergleich unabhängiger Gruppen, wobei nach wichtigen Kovariablen adjustiert werden soll. Hierfür kommen dann als Verallgemeinerung von *t*-Test [6] und Varianzanalyse [2,4] die Methoden der multiplen Regressionsanalyse in Frage [5].

kurzgefasst

Zur Auswahl eines geeigneten Signifikanztests für die einfachen Standardsituationen in der medizinischen Statistik kann man die Übersicht in Tab. 1 verwenden. Es werden lediglich die Informationen über die Zahl der zu analysierenden Stichproben, deren Abhängigkeitsstatus sowie das Messniveau und die Verteilung der Zielvariable benötigt.

Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt **Tab. 2**.

Tab. 2 Übersetzung biometrischer Begriffe (deutsch – englisch).

Signifikanztest	significance test
Stichprobe	sample
unabhängig	independent
abhängig	dependent
Messniveau	measurement scale
binär	binary
nominal	nominal
ordinal	ordinal
stetig	continuous
zensiert	censored
Überlebenszeit	survival time
nichtparametrisch	nonparametric
Rang	rank
Bindung	tie
Kovariable	covariable
Vorzeichentest	sign test
Wilcoxon Vorzeichenrangtest	Wilcoxon sign rank test
Wilcoxon Rangsummentest	Wilcoxon rank sum test
gepaarter t-Test	paired *t*-test
Varianzanalyse	analysis of variance (ANOVA)
Messwertwiederholungen	repeated measurements
Homoskedastizität	homoscedasticity

Dieser Beitrag ist eine überarbeitete Fassung aus dem Supplement Statistik aus dem Jahr 2002.

Literatur

1 *Altman DG, Bland JM.* The Normal distribution. BMJ 1995; 310: 298
2 *Altman DG, Bland JM.* Comparing several groups using analysis of variance. BMJ 1996; 312: 1472–1473
3 *Altman DG, Bland JM.* Time to event (survival) data. BMJ 1998; 317: 468–469
4 *Bender R, Ziegler A, Lange S.* Varianzanalyse. Dtsch Med Wochenschr 2007; 132: e57–e60
5 *Bender R, Ziegler A, Lange S.* Multiple Regression. Dtsch Med Wochenschr 2007; 132: e30–e32
6 *Lange S, Bender R.* Was ist ein Signifikanztest? Dtsch Med Wochenschr 2007; 132: e19–e21
7 *Mehta CR, Patel NR.* StatXact 5 for Windows. Statistical Software for Exact Nonparametric Inference. CYTEL Software Corporation, Cambridge, MA, 2001
8 *Sawilowsky SS, Blair RC.* A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychol Bull 1992; 111: 352–360
9 *Skovlund E, Fenstad GU.* Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol 2001; 54: 86–92