Multiples Testen

– Artikel Nr. 12 der Statistik-Serie in der DMW –

Multiple testing

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
- Multiple Vergleiche
- Multiples Signifikanzniveau
- Statische Macht
- Multiple Testprozedur

Key words
- Multiple comparisons
- Multiple significance level
- Power
- Multiple testing procedure

Bibliografie
DOI 10.1055/s-2007-959035
Dtsch Med Wochenschr 2007; 132: e26–e29 · © 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

Irrtumswahrscheinlichkeiten bei multiplen Signifikanztests

Häufig werden in der biomedizinischen Forschung zum Nachweis von Effekten oder Zusammenhängen statistische Signifikanztests [9] verwendet und das Ergebnis in Form von *p*-Werten [4] angegeben. Ist dieser ***p***-Wert kleiner als das vorgegebene Signifikanzniveau α (häufig $\alpha = 0{,}05$), so wird die Nullhypothese, dass kein Effekt existiert, verworfen. Das Signifikanzniveau α begrenzt die Wahrscheinlichkeit für den Fehler 1. Art, nämlich beim Ablehnen der Nullhypothese eine falsche Entscheidung zu treffen. Die Kontrolle dieser Irrtumswahrscheinlichkeit ist die wesentliche Eigenschaft eines Signifikanztests zum Niveau α. Die Begrenzung der Irrtumswahrscheinlichkeit auf höchstens α gilt allerdings nur beim Testen ***einer*** Hypothese mit Hilfe ***eines*** Signifikanztests. Werden zur Untersuchung einer Fragestellung mehrere Tests jeweils zum Niveau α durchgeführt (Mehrhypothesenproblem), so wird zwar für jeden einzelnen Test die **individuelle Irrtumswahrscheinlichkeit** (engl.: individual error rate) kontrolliert, die **versuchsbezogene Irrtumswahrscheinlichkeit** (engl.: experimentwise error rate) für das gesamte Mehrhypothesenproblem ist jedoch größer als α.

Zur Beschreibung der Testeigenschaften im Rahmen von Mehrhypothesenproblemen müssen verschiedene Signifikanzniveaus unterschieden werden. Bei der Anwendung multipler Tests hält man das **lokale Signifikanzniveau** (engl.: local significance level) von α ein, wenn jede individuelle Nullhypothese höchstens mit Wahrscheinlichkeit α irrtümlich abgelehnt wird. Die versuchsbezogene Irrtumswahrscheinlichkeit ist jedoch durch die Wahrscheinlichkeit gegeben, mindestens eine der individuellen Nullhypothese irrtümlich abzulehnen; um diese zu kontrollieren, genügt nicht die Einhaltung des lokalen Signifikanzniveaus. Ein multiples Testverfahren hält das **globale Signifikanzniveau** (engl.: global significance level) von α ein, wenn die Wahrscheinlichkeit, mindestens eine der einzelnen Nullhypothesen abzulehnen, unter der Annahme, dass alle Nullhypothesen zutreffen, höchstens α beträgt. Die Annahme der globalen Nullhypothese, nämlich dass alle einzelnen Nullhypothesen gleichzeitig richtig sind, ist jedoch in der Praxis meist unrealistisch. In der Regel möchte man sich vor der Fehlentscheidung schützen, mindestens eine wahre Nullhypothese abzulehnen, und zwar unabhängig davon, welche der anderen Nullhypothesen wahr oder falsch sind. Daher ist das **multiple Signifikanzniveau** (engl.: multiple significance level) definiert als die maximale Wahrscheinlichkeit mindestens eine der einzelnen Nullhypothesen irrtümlich abzulehnen, unabhängig davon, welche der anderen Nullhypothesen richtig sind und welche nicht. Die Einhaltung des multiplen Signifikanzniveaus ist das stärkste Kriterium, um sich bei der Anwendung multipler Signifikanztests vor Fehlentscheidungen zu schützen [2, 8].

Allgemeine Methoden

Die Berechnung der versuchsbezogenen Irrtumswahrscheinlichkeit ist einfach, wenn es sich um ***unabhängige*** Tests handelt (z. B. beim Testen jeweils einer Hypothese in mehreren sich nicht überschneidenden Gruppen). Falls k unabhängige Signifikanztests jeweils zum lokalen Niveau α durchgeführt werden, so ist die Wahrscheinlichkeit für einen einzelnen Test, diesen korrekterweise abzulehnen, $1-\alpha$. Da die Tests unabhängig sind, ist die Wahrscheinlichkeit, alle k Tests korrekterweise abzulehnen, das Produkt der einzelnen Wahrscheinlichkeiten, also $(1-\alpha)^k$. Damit ist die Wahrscheinlichkeit, mindestens eine der k Nullhypothesen fälschlicherweise abzulehnen $1-(1-\alpha)^k$. Mit steigender Zahl der Tests steigt

auch die versuchsbezogene Irrtumswahrscheinlichkeit. Bei $\alpha = 0{,}05$ und $k = 100$ unabhängigen Tests beträgt die versuchsbezogene Irrtumswahrscheinlichkeit

$$1-(1-0{,}05)^{100} = 0{,}994$$

Mit anderen Worten: Beim Testen von 100 unabhängigen, in Wahrheit richtigen Nullhypothesen erhält man fast sicher mindestens ein falsch signifikantes Resultat. Mit Hilfe dieser Berechnung lässt sich auch eine einfache Korrektur für multiples Testen durchführen. Das als Šidák-Methode bekannte Verfahren besagt, dass man das multiple Signifikanzniveau von α einhält, wenn man die k einzelnen Tests jeweils zum Niveau $1-(1-\alpha)^{1/k}$ durchführt [13].

Die Berechnung der versuchsbezogenen Irrtumswahrscheinlichkeit ist weitaus schwieriger, wenn es sich um ***abhängige*** Tests handelt (z. B. Signifikanztests bezüglich mehrerer Zielvariablen der gleichen Stichprobe). Dies ist in der Praxis der häufigste Fall. Da die versuchsbezogene Irrtumswahrscheinlichkeit von der Abhängigkeitsstruktur der Tests untereinander abhängt, kann man keine allgemein gültige Formel herleiten. Man kann aber die versuchsbezogene Irrtumswahrscheinlichkeit nach oben abschätzen: sie kann auf keinen Fall größer sein als die Summe der individuellen Irrtumswahrscheinlichkeiten, d. h. die versuchsbezogene Irrtumswahrscheinlichkeit bei k (möglicherweise abhängigen) Tests jeweils zum Niveau α ist $\leq k \times \alpha$. Aus dieser Ungleichung leitet sich die bekannte Bonferroni-Methode [7] ab, die besagt, dass man das multiple Signifikanzniveau von α einhält, wenn man die einzelnen Tests jeweils zum Niveau α/k durchführt. Alternativ hierzu kann man auch die einzelnen *p*-Werte mit k multiplizieren, um für multiples Testen adjustierte *p*-Werte zu erhalten.

Die Bonferroni-Methode ist sehr einfach durchzuführen und global anwendbar auf alle multiplen Testsituationen. Sie hat allerdings den Nachteil, dass sie – bedingt durch die grobe Abschätzung der versuchsbezogenen Irrtumswahrscheinlichkeit – die Macht (engl.: power) der Tests unnötigerweise stark reduziert, insbesondere wenn die Zahl der Tests hoch ist und die Tests untereinander stark korreliert sind. Daher wurden in den letzten Jahren eine Reihe weiterer multipler Testprozeduren entwickelt. Diese kann man einteilen in allgemein anwendbare Methoden und solche, die für spezielle Testsituationen entwickelt wurden. Bei den allgemein anwendbaren Verfahren sind in erster Linie die Verfahren zu nennen, die sich aus dem Abschlusstest-Prinzip (engl.: closed test principle) herleiten [14], sowie neuere rechenintensive Verfahren, die auf Resampling [16] basieren. Beim Abschlusstest-Prinzip wird die logische Struktur der einzelnen Hypothesen ausgenutzt. Ein vereinfachtes Abschlusstest-Verfahren stellt z. B. die bekannte Methode von Holm dar [1]. Unter Resampling versteht man die Erzeugung einer großen Zahl von Pseudo-Datensätzen durch wiederholte Stichprobenziehung mit Zurücklegen aus dem Ausgangsdatensatz. Auf diese Weise ist es möglich, Informationen über die Abhängigkeiten und Verteilungseigenschaften der einzelnen Teststatistiken zu gewinnen und auszunutzen. Auf diese Verfahren kann im Rahmen dieses Artikels nicht im Detail eingegangen werden. Für spezielle Multiplizitätssituationen gibt es eine Reihe bekannter Verfahren, die im Folgenden kurz zusammengefasst werden. Eine ausführlichere Übersicht findet man in der Literatur [5, 8].

Mehr als 2 Gruppen

Für den Vergleich von mehr als 2 Mittelwerten mit Hilfe der Varianzanalyse [6] existieren die meisten multiplen Testprozeduren. Mit Hilfe des ***F***-Tests kann entschieden werden, ob es überhaupt Unterschiede zwischen den Gruppen gibt; die anschließende Anwendung einer multiplen Testprozedur gibt Aufschluss darüber, zwischen welchen Gruppen Unterschiede bestehen. Die bekanntesten Prozeduren, die auch häufig in statistischen Programmpaketen enthalten sind, sind die simultanen Testprozeduren von Scheffé und Tukey, die Methode von Dunnett, bei der mehrere Gruppen jeweils mit der gleichen Referenzgruppe verglichen werden, und das mehrstufige Verfahren von Ryan, Einot, Gabriel und Welsch (REGW-Prozedur). All diese Verfahren kontrollieren das multiple Signifikanzniveau, zumindest in balancierten Designs (d. h. mit gleichen Stichprobenumfängen pro Gruppe). Für den häufigen Fall von 3 Gruppen gibt es die einfache Methode nach Bonferroni-Holm-Shaffer, die das multiple Signifikanzniveau kontrolliert. Zunächst testet man mit einem globalen Test zum Niveau α (z. B. ***F***-Test oder Kruskal-Wallis-Test), ob überhaupt signifikante Unterschiede zwischen den 3 Gruppen bestehen. Nur wenn der globale Test signifikant ist, kann im nächsten Schritt mit paarweisen Vergleichen (z. B. ***t***-Test oder Wilcoxon Rangsummentest) ebenfalls zum Niveau α getestet werden, zwischen welchen Gruppen die Unterschiede bestehen.

Mehr als 1 Endpunkt

Der Fall multipler Endpunkte ist das häufigste Multiplizitätsproblem in klinischen Studien. Es gibt mehrere mögliche Strategien zum Umgang mit multiplen Endpunkten. Die einfachste Möglichkeit ist, einen einzigen primären Endpunkt zu spezifizieren. Dies macht eine Adjustierung für multiples Testen unnötig; allerdings sind dann Signifikanztests bezüglich sekundärer Endpunkte untergeordnete Analysen und nur einer explorativen Interpretation zugänglich. Zweitens können multiple Endpunkte zu einem einzigen Endpunkt aggregiert (zusammengefasst) werden. In der UK Prospective Diabetes Study (UKPDS) wurden z. B. Ereignisse wie Tod durch Hyperglykämie oder Hypoglykämie, Nierenversagen, Amputation, Blindheit u. a. zur Zielvariable „irgendein diabetesbezogener Endpunkt“ zusammengefasst [15]. Allerdings erhält man dann keine Resultate für die einzelnen Endpunkte. Drittens können multivariate Methoden, z. B. eine multivariate Varianzanalyse (MANOVA) verwendet werden, wobei jedoch auch hier, wie bei aggregierten Endpunkten, keine Interpretation der einzelnen Variablen möglich ist. Falls es mehrere gleichwertige Endpunkte gibt, von denen kein primärer Endpunkt spezifiziert werden kann, oder falls die Resultate der einzelnen Endpunkte interessieren, so ist die Anwendung einer multiplen Testprozedur notwendig. Hierfür können wiederum die o. g. allgemeinen Methoden nach dem Abschlusstest-Prinzip [8, 14] und die Resampling-Verfahren [16] verwendet werden.

Messwertwiederholungen

Obwohl hochentwickelte statistische Methoden zur Analyse von Messwertwiederholungen vorhanden sind, gibt es beim Vorliegen von Messwertwiederholungen nur sehr wenige multiple

Testprozeduren für ganz spezielle Datensituationen. Handelt es sich um Verlaufskurven, so kann in vielen Fällen das Multiplizitätsproblem verringert oder sogar ganz vermieden werden, wenn anstelle der Verlaufskurven geeignete Kurvenkenngrößen ausgewertet werden [4]. Dies führt oftmals zu varianzanalytischen Fragestellungen, die mit den entsprechenden Methoden untersucht werden können (siehe oben).

Subgruppen-Analysen

In aller Regel sind Subgruppenanalysen schwierig zu interpretieren. Prinzipiell gilt, dass Analysen bezüglich a posteriori definierten Subgruppen nur explorativen Charakter haben, egal ob für multiples Testen adjustiert wird oder nicht. Ist die Untersuchung eines Effektunterschieds zwischen a priori definierten Subgruppen das Ziel einer konfirmatorischen Studie, so ist die adäquate Methode im Allgemeinen ein Test auf Signifikanz der entsprechenden Wechselwirkung. Zur Untersuchung der unterschiedlichen Effekte in den Subgruppen können die klassischen Methoden nach dem Abschlusstest-Prinzip [8, 14] und die Resampling-Verfahren [16] verwendet werden.

Zwischenauswertungen

Häufig werden in klinischen Studien Daten über längere Zeiträume gesammelt und bereits vor dem definierten Ende der Studie Zwischenauswertungen durchgeführt. Bei solchen Studien besteht die Möglichkeit, die Studie ggf. vorzeitig abbrechen zu können, so dass möglichst wenige Patienten eingeschlossen werden und die Effektivität neuer Therapiemethoden möglichst früh erkannt wird. Prinzipiell führen Zwischenauswertungen zu einer Erhöhung der Irrtumswahrscheinlichkeit 1. Art. Daher muss eine adäquate Adjustierung für multiples Testen vorgenommen werden. Eine einfache Regel, die in der Praxis häufig ausreicht, ist die folgende: Wenn man bei Vorliegen eines Endpunkts und höchstens 10 Zwischenauswertungen jeweils zum Niveau $\alpha = 0{,}01$ testet, so hält man insgesamt das Signifikanzniveau von $\alpha = 0{,}05$ ein. Die erforderlichen Signifikanzniveaus für andere Anzahlen von Zwischenauswertungen lassen sich mit der Methode von Pocock [11] berechnen. Der Nachteil dieser Methode ist, dass auch am Ende der Studie mit dem gleichen niedrigen Niveau wie bei den Zwischenauswertungen getestet werden muss. Es gibt eine Reihe von Verfahren, die es ermöglichen, am Ende der Studie möglichst nahe am Niveau $\alpha = 0{,}05$ zu testen; dafür werden vorher die Zwischenauswertungen zu einem sehr viel kleineren Signifikanzniveau durchgeführt. Eines der bekanntesten Verfahren dieser Art ist das von O'Brien und Fleming [10].

Wann muss multiples Testen berücksichtigt werden?

Es stellt sich nun die Frage, in welchen Fällen eine Adjustierung für multiples Testen notwendig ist und wann nicht. Prinzipiell gilt, dass keine Berücksichtigung für multiples Testen erforderlich ist, wenn es genügt, das lokale Signifikanzniveau einzuhalten. Soll dagegen das multiple oder zumindest das globale Signifikanzniveau kontrolliert werden, so ist eine Adjustierung für multiples Testen zwingend erforderlich. Schwieriger ist nun die Beurteilung, in welchen Situationen welches Signifikanzniveau eingehalten werden sollte; dies wird in der Literatur kontrovers beurteilt [5, 12]. In explorativen Versuchen, in denen häufig eine Vielzahl von Signifikanztests verwenden werden, um Hypothesen zu generieren, halten wir die Anwendung multipler Testprozeduren nicht für unbedingt erforderlich. „Signifikante" Ergebnisse solcher Studien haben jedoch nur explorativen Charakter und müssen als solche kenntlich gemacht werden. Darüber hinaus ist selbstverständlich eine konfirmatorische Validierung dieser Resultate in späteren Studien erforderlich.

Tab. 1 Übersetzung (deutsch – englisch).

Signifikanztest	significance test
p-Wert	p value
Signifikanzniveau	significance level
Fehler 1. Art	type 1 error
individuelle Irrtumswahrscheinlichkeit	individual error rate = comparisonwise error rate
versuchsbezogene Irrtumswahrscheinlichkeit	experimentwise error rate = familywise error rate
lokales Signifikanzniveau	local significance level
globales Signifikanzniveau	global significance level
multiples Signifikanzniveau	multiple significance level
Abschlusstest-Prinzip	closed test principle
paarweise Vergleiche	pairwise comparisons
Macht	power
Varianzanalyse	analysis of variance (ANOVA)
multivariate Varianzanalyse	multivariate analysis of variance (MANOVA)
balanciertes Design	balanced design
Messwertwiederholungen	repeated measurements
Subgruppen-Analysen	subgroup analyses
Wechselwirkung	interaction
Zwischenauswertungen	interim analyses
konfirmatorischer Versuch	confirmatory trial
explorativer Versuch	exploratory trial

In konfirmatorischen Versuchen sollte sorgfältig überprüft werden, ob die Anwendung einer multiplen Testprozedur erforderlich ist. Hierfür muss zunächst definiert werden, welche Signifikanztests eine inhaltlich zusammenhängende Familie von Tests zur Untersuchung einer Fragestellung darstellen [12]. Ein „blindes" Adjustieren für multiples Testen mit Verfahren wie der Bonferroni-Methode ist nicht sinnvoll, insbesondere dann nicht, wenn aus den einzelnen Tests gar keine gemeinsame Schlussfolgerung gezogen werden soll. Liegt allerdings eine klar definierte Fragestellung in Form eines Mehrhypothesenproblems vor, dann ist für die Familie inhaltlich zusammenhängender Tests zur Einhaltung des multiplen Signifikanzniveaus die Anwendung einer adäquaten multiplen Testprozedur erforderlich. Denn nur dann kann in sinnvoller Weise der Fehler 1. Art kontrolliert werden.

Für viele Standardsituationen sind eine Vielzahl von multiplen Testprozeduren entwickelt worden. Diese werden bisher in der medizinischen Literatur nicht genügend beachtet. In der Praxis ist es schwierig, multiples Testen adäquat zu berücksichtigen, wenn es verschiedene Multiplizitätsebenen gibt (z. B. mehr als 1 Endpunkt ***und*** mehr als 2 Gruppen ***und*** Messwertwiederholungen). In der Studienplanung sollte man daher darauf achten,

dass Multiplizitätsprobleme so gering wie möglich gehalten werden. Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt **Tab. 1**.

kurzgefasst

In konfirmatorischen Studien benötigt man zur Einhaltung des vorgegebenen Signifikanzniveaus bei der Anwendung multipler Signifikanztests zur Untersuchung einer Fragestellung Methoden zur Adjustierung für multiples Testen. Bei der Planung von Studien sollten Multiplizitätsprobleme so gering wie möglich gehalten werden, so dass entweder eine Adjustierung für multiples Testen unnötig wird oder eine der zahlreichen multiplen Testprozeduren angewendet werden kann.

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

Literatur

1 *Aickin M, Gensler H.* Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. Am J Public Health 1996; 86: 726–728
2 *Bauer P.* Multiple testing in clinical trials. Stat Med 1991; 10: 871–890
3 *Bender R, Lange S.* Was ist der *p*-Wert? Dtsch Med Wochenschr 2007; 132: e15–e16
4 *Bender R, Lange S.* Verlaufskurven. Dtsch Med Wochenschr 2007; 132: e22–e23
5 *Bender R, Lange S.* Adjusting for multiple testing – when and how? J Clin Epidemiol 2001; 54: 343–349
6 *Bender R, Ziegler A, Lange S.* Varianzanalyse. Dtsch Med Wochenschr 2007; 132: e57–e60
7 *Bland JM, Altman DG.* Multiple significance tests: The Bonferroni method. Br med J 1995; 310: 170
8 *Horn M, Vollandt R.* Multiple Tests und Auswahlverfahren. Fischer, Stuttgart, 1995
9 *Lange S, Bender R.* Was ist ein Signifikanztest? Dtsch Med Wochenschr 2007; 132: e19–e21
10 *O'Brien PC, Fleming TR.* A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549–556
11 *Pocock SJ.* Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64: 191–199
12 *Proschan MA, Waclawiw MA.* Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials 2000; 21: 527–539
13 *Sachs L.* Angewandte Statistik. Anwendung statistischer Methoden (9. überarbeitete Auflage) (Hrsg). Heidelberg: Springer, 1999
14 *Sonnemann E.* Allgemeine Lösungen multipler Testprobleme. EDV Med Biol 1982; 13: 120–128
15 *The UK Prospective Diabetes Study (UKPDS) Group.* Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. Br med J 1998; 317: 703–713
16 *Westfall PH, Young SS.* Resampling-Based Multiple Testing. New York: Wiley, 1993