

Varianzanalyse

– Artikel Nr. 21 der Statistik-Serie in der DMW –

Analysis of variance

Autoren

R. Bender¹ A. Ziegler² S. Lange¹

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

Ungepaarter *t*-Test

Mit Hilfe des ungepaarten *t*-Tests lässt sich herausfinden, ob sich zwei unabhängige Gruppen bezüglich ihrer Mittelwerte statistisch signifikant unterscheiden [3,9]. Streng genommen setzt der *t*-Test die Normalverteilung des betrachteten Merkmals voraus. Der *t*-Test ist aber sehr robust gegenüber Abweichungen von der Normalverteilung [11]. Bei genügend großen Stichproben ist der *t*-Test auch anwendbar, wenn die Zielvariable nicht normalverteilt ist [10]. Bei kleinen Stichproben mit stetigen aber nicht normalverteilten Daten kann ein entsprechender nicht-parametrischer Test, wie z. B. der Wilcoxon-Rangsummentest verwendet werden [3]. Des Weiteren setzt der *t*-Test gleiche Varianzen in den beiden Gruppen, die so genannte Homoskedastizität, voraus. Ist diese Annahme stark verletzt, so sollte auf modifizierte Tests zurückgegriffen werden, die diese Annahme nicht benötigen [3]. Allerdings beseitigt die Anwendung solcher modifizierter Tests nicht die Notwendigkeit einer inhaltlichen Überprüfung, ob ein Gruppenvergleich in einer Situation, bei der die Gruppenvarianzen stark unterschiedlich sind, überhaupt sinnvoll ist. Bei geringeren Unterschieden in der Varianz ist jedoch in vielen praxisrelevanten Situationen der gewöhnliche ungepaarte *t*-Test anwendbar. Da sehr häufig in der medizinischen Forschung genau zwei Gruppen bezüglich eines stetigen Merkmals verglichen werden, ist der *t*-Test eines der wichtigsten Verfahren in der Medizinischen Statistik.

Ist in einer Studie jedoch das Ziel der Vergleich von Mittelwerten zwischen mehr als zwei Gruppen, so wird eine flexiblere Methode benötigt, die den Vergleich von mehr als zwei Mittelwerten zulässt. Situationen, in denen mehr als zwei Gruppen verglichen werden, sind z. B. klinische Versuche mit mehr als zwei verschiedenen Medikamenten oder epidemiologische Studien, in

denen die Exposition in mehr als zwei Ausprägungen gemessen wird (z. B. Raucher, Ex-Raucher, Nicht-Raucher). Die Verallgemeinerung des *t*-Tests auf solche Situationen ist die Varianzanalyse, die international durch ANOVA (Analysis of variance) abgekürzt wird [1]. Je nach Studiendesign und Datenlage ergeben sich unterschiedliche Varianzanalysemodelle. Die für die Anwendung grundlegenden Modelle der ANOVA werden im Folgenden kurz beschrieben und erklärt.

Modell der Einfachklassifikation

Das einfachste varianzanalytische Modell ist das der Einfachklassifikation. In diesem Modell kann ein einziger Einflussfaktor untersucht werden, der in mehr als zwei Ausprägungen vorliegt. Wir betrachten als Beispiel eine Studie, in der es um die Untersuchung der Effekte von Rauchverhalten und Sport auf die Lungenfunktion geht (zur Illustration wird ein künstlicher Datensatz verwendet). Bei 60 Probanden wurden die Lungenfunktion mit Hilfe des forcierten expiratorischen Volumens in einer Sekunde (FEV₁) sowie das Raucherhalten in drei Kategorien (0 = Nicht-Raucher, 1 = Ex-Raucher, 2 = Raucher) und das Vorhandensein regelmäßiger sportlicher Aktivität in zwei Kategorien (1 = ja, 0 = nein) gemessen. In **Tab. 1** findet man einen deskriptiven Überblick über den verwendeten Datensatz.

Unter Vernachlässigung des Merkmals Sport betrachten wir zunächst die Nullhypothese, dass alle drei Gruppen bezüglich des Rauchverhaltens im Durchschnitt gleiche FEV₁-Werte haben. Da es sich dann um den Vergleich von drei Gruppen handelt, ist der *t*-Test nicht anwendbar. Das richtige Verfahren ist das Varianzanalysemodell der Einfachklassifikation. Dieses Modell wird allgemein formuliert über die Gleichung

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

$i = 1, \dots, m$ (Gruppen) und $j = 1, \dots, n_i$ (Probanden)

Schlüsselwörter

- ▶ Varianzanalyse
- ▶ *t*-Test
- ▶ Wechselwirkung
- ▶ Multiple Vergleiche

Key words

- ▶ Analysis of variance (ANOVA)
- ▶ *t* test
- ▶ Interaction
- ▶ Multiple comparisons

Bibliografie

DOI 10.1055/s-2007-959044
Dtsch Med Wochenschr 2007;
132: e57–e60 · © Georg Thieme
Verlag Stuttgart · New York ·
ISSN 0012-0472

Korrespondenz

Privatdozent Dr. rer. biol. hum.

Ralf Bender

Institut für Qualität und
Wirtschaftlichkeit im Gesund-
heitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

Tab. 1 Deskriptive Statistik bezüglich des Merkmals FEV₁ (Angaben in Liter) in einer Studie zur Untersuchung des Effekts von Rauchverhalten und Sport bei 60 Probanden (künstliche Daten).

	Sportler			Nicht-Sportler			Gesamt		
	n	MW	SD	n	MW	SD	n	MW	SD
Nicht-Raucher	7	4,58	0,33	13	3,57	0,56	20	3,93	0,69
Ex-Raucher	8	3,68	0,85	12	3,53	0,50	20	3,59	0,64
Raucher	10	2,91	0,71	10	2,78	0,50	20	2,84	0,60
Gesamt	25	3,62	0,95	35	3,33	0,62	60	3,45	0,78

MW = Mittelwert; SD = Standardabweichung

wobei y_{ij} der Merkmalswert des j -ten Probanden in der i -ten Gruppe ist, μ der Mittelwert über alle Probanden und α_i der Effekt der i -ten Gruppe. Damit kann man auch schreiben, dass der Mittelwert der i -ten Gruppe durch $\mu_i = \mu + \alpha_i$ darstellbar ist und α_i gerade die Abweichung des i -ten Gruppenmittelwerts vom Gesamtmittel ist. Schließlich ist e_{ij} die zufällige Abweichung des j -ten Probanden vom i -ten Gruppenmittelwert. Im Beispiel ist y_{ij} also der FEV₁-Wert des j -ten Probanden in der i -ten Raucherstatusgruppe, wobei der Index i von 1 bis $m = 3$ läuft und j jeweils von 1 bis $n_i = 20$ für alle i . In einer Beobachtungsstudie ist ein balanciertes Design, also gleich viele Probanden in jeder Gruppe, eher die Ausnahme. Die Datenanalyse ist jedoch einfacher und hat mehr statistische Power, wenn das Design balanciert ist, so dass wir zur Illustration hier zunächst ein balanciertes Design betrachten.

Um die Teststatistik und den p -Wert für die Nullhypothese der Gleichheit aller Gruppenmittelwerte zu berechnen, bedient man sich einer so genannten Varianzanalysetabelle, die in Tab. 2 aufgeführt ist.

Die Nullhypothese der Gleichheit aller Gruppenmittelwerte lässt sich mit Hilfe eines F -Tests testen, dessen Teststatistik sich aus mittleren Quadratsummen zusammensetzt, die mit Hilfe der Varianzanalysetabelle berechnet werden können. Die Summe von Abstandsquadraten gemäß dem Konzept von Gauß bildet dabei das Grundgerüst für die Berechnung von Varianzen bzw. Standardabweichungen [8]; auf die genauen Formeln wollen wir hier nicht näher eingehen. Der F -Test vergleicht die Varianz zwischen den Gruppen mit der Varianz innerhalb der Gruppen; daher kommt auch der Name des Verfahrens: Varianzanalyse. Die Idee des Tests ist dabei, dass die Varianzen innerhalb der Gruppen klein und die zwischen den Gruppen erwartungsgemäß groß sind, wenn sich in Wahrheit die Gruppen unterscheiden. Der p -Wert findet sich in der Varianzanalysetabelle rechts oben. Aufgrund des kleinen p -Werts ($p < 0.0001$) kann man hier also die Gleichheit aller drei Gruppenmittelwerte (Schätzung aus den Daten: Nicht-Raucher 3,93, Ex-Raucher 3,59, Raucher 2,84 Liter, siehe Tab. 1) ablehnen, so dass dieses Modell einen statistisch signifikanten Einfluss des Rauchverhaltens auf die Lungenfunktion zeigt.

Der Nachteil dieses Modells ist jedoch, dass ein möglicher Effekt regelmäßigen Sports nicht berücksichtigt ist. Um auch den möglichen Effekt regelmäßigen Sports zu berücksichtigen, muss das Modell erweitert werden.

Modell der Zweifachklassifikation

Wäre regelmäßiger Sport die einzige Einflussgröße, so könnte man zum Vergleich der Sportler mit den Nicht-Sportlern den t -

Test verwenden. Die beiden rohen Mittelwerte lauten 3,33 Liter für Nicht-Sportler und 3,62 Liter für Sportler (siehe Tab. 1). Mit Hilfe des t -Test lässt sich hier jedoch kein signifikanter Unterschied zwischen Sportlern und Nicht-Sportlern belegen ($p = 0,1605$). Allerdings berücksichtigt diese Analyse nicht den Effekt des Rauchverhaltens und somit besteht die Gefahr einer Verzerrung, wenn nur die rohen Mittelwerte verglichen werden. Es wird daher ein Verfahren benötigt, das gleichzeitig den Effekt des Rauchverhaltens und den Effekt des Sports berücksichtigt. Ein solches Verfahren stellt das Varianzanalysemodell der Zweifachklassifikation dar. Dieses Modell wird allgemein formuliert über die Gleichung

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \quad i = 1, \dots, m; j = 1, \dots, r; k = 1, \dots, n_{ij}$$

wobei y_{ijk} der Merkmalswert des k -ten Probanden in der i -ten Gruppe des 1. Faktors (im Beispiel Rauchverhalten) und in der j -ten Gruppe des 2. Faktors (im Beispiel Sport) ist, μ der Mittelwert über alle Probanden, α_i der Effekt der i -ten Gruppe 1. Faktors, β_j der Effekt der i -ten Gruppe 2. Faktors und e_{ijk} die zufällige Abweichung des k -ten Probanden vom jeweiligen Gruppenmittelwert.

Bei der Berechnung der Quadratsummen und der benötigten Teststatistiken und p -Werte ist – im Gegensatz zur vorherigen Berechnung – jetzt zu beachten, dass das zugrunde liegende Design unbalanciert ist, da nicht alle Zellen der Kreuzklassifikation zwischen Rauchverhalten und Sport mit gleich vielen Probanden besetzt sind. Um mit diesem Problem umzugehen, wurden verschiedene Lösungen entwickelt, die so genannten Quadratsummen des Typs I, II, III und IV. In der Regel sollte man die Quadratsummen des Typs II oder III verwenden, wobei aus inhaltlichen Gründen primär in der Praxis die Quadratsummen des Typs III verwendet werden. Technische Details sind in der Literatur zu finden [7]. Mit Hilfe der Quadratsummen des Typs III erhält man sowohl für den Faktor Rauchverhalten ($p < 0,0001$) als auch für den Faktor Sport ($p = 0,0145$) ein signifikantes Ergebnis. Regelmäßiger Sport hat also – wenn man den Effekt des Rauchverhaltens berücksichtigt – einen signifikanten Einfluss auf die Lungenfunktion. Dieses Ergebnis deutet an, dass der Vergleich der rohen Mittelwerte mit Hilfe des t -Tests (s. o.) keine adäquate Analyse darstellt.

Modell der Zweifachklassifikation mit Wechselwirkung

Das Modell der Zweifachklassifikation besitzt eine wichtige Voraussetzung, nämlich dass sich die Effekte der beiden betrachteten Faktoren additiv verhalten. Das bedeutet, dass die Effekte des einen Faktors in jeder Faktorstufe des anderen Faktors identisch sind. Um dies zu überprüfen, eignen sich z. B. grafische

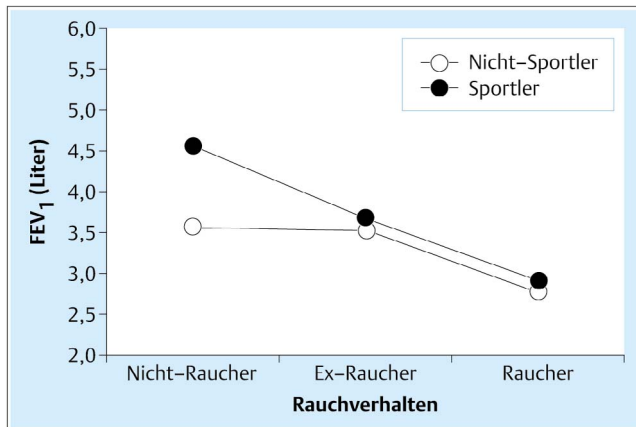


Abb. 1 FEV₁-Mittelwerte in Abhängigkeit von Rauchverhalten und regelmäßiger sportlicher Tätigkeit bei 60 Probanden.

Methoden. In **Abb. 1** sind die mittleren FEV₁-Werte aus **Tab. 1** getrennt nach Rauch- und Sportverhalten grafisch dargestellt.

Wenn die Unterschiede im FEV₁-Wert zwischen Sportlern und Nicht-Sportlern in allen drei Raucherstatusgruppen gleich wären, müssten die Linien parallel verlaufen. Dies ist jedoch bei den betrachteten Daten nicht der Fall. Da der Unterschied zwischen Sportlern und Nicht-Sportlern bei Nicht-Rauchern sehr viel ausgeprägter ist als in den beiden anderen Raucherstatusgruppen, ist die Annahme additiver Effekte hier fragwürdig. Um unterschiedliche Effekte des Sports in den drei Raucherstatusgruppen zu modellieren, muss man so genannte Wechselwirkungen ins Modell aufnehmen. Das Modell der Zweifachklassifikation mit Wechselwirkung wird allgemein formuliert über die Gleichung

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \quad i = 1, \dots, m; j = 1, \dots, r; k = 1, \dots, n_{ij}$$

wobei γ_{ij} die Wechselwirkung zwischen den beiden Faktoren darstellt und die anderen Parameter die gleiche Bedeutung haben wie im letzten Abschnitt.

In diesem Modell erhält man ähnlich wie vorher signifikante Effekte für das Rauchverhalten ($p < 0,0001$) und für den Faktor Sport ($p = 0,0089$). Aber auch die Wechselwirkung zwischen Rauchen und Sport ist zum 5% Niveau statistisch signifikant ($p = 0,0442$). Die entsprechende Varianzanalysetabelle findet man in **Tab. 3**. In dieser Studie hat man also statistisch signifikante Effekte regelmäßigen Sports sowie des Rauchverhaltens auf die Lungenfunktion gefunden, wobei der Effekt des Sports zwischen den Raucherstatusgruppen allerdings unterschiedlich ist.

In ähnlicher Weise lassen sich Modelle mit noch mehr Faktoren bilden. Für jeden betrachteten Faktor und jede bedeutsame Wechselwirkung kommt ein entsprechender Term im Modell hinzu. Da das Grundprinzip der Modellierung gleich bleibt, werden wir hier auf Modelle mit mehr als zwei Faktoren nicht weiter eingehen.

Multiple Vergleiche

Die bisher betrachteten Ergebnisse der Varianzanalysen lassen nur globale Schlussfolgerungen zu. Es konnte u. a. gezeigt werden, dass es einen signifikanten Einfluss des Rauchverhaltens

Tab. 2 Varianzanalysetabelle für das Modell der Einfachklassifikation zur Untersuchung von Unterschieden zwischen drei Raucherstatus-Gruppen bezüglich des Merkmals FEV₁.

	Quadratsumme	Freiheitsgrade	Mittel der Quadrate	F	p-Wert
Zwischen den Gruppen	12,312	2	6,156	14,76	< 0,0001
Innerhalb der Gruppen	23,773	57	0,417		
Gesamt	36,085	59			

Tab. 3 Varianzanalysetabelle für das Modell der Zweifachklassifikation mit Wechselwirkung zur Untersuchung des Effekts von Rauchverhalten und Sport auf das Merkmal FEV₁.

	Quadratsumme	Freiheitsgrade	Mittel der Quadrate	F	p-Wert
Rauchverhalten	14,901	2	7,450	21,16	< 0,0001
Sport	2,595	1	2,595	7,37	0,0089
Rauchen × Sport	2,329	2	1,164	3,31	0,0442
Fehler	19,016	54	0,352		

gibt, d. h. es sind Unterschiede bezüglich des mittleren FEV₁-Werts zwischen Rauchern, Ex-Rauchern und Nicht-Rauchern vorhanden. Man weiß damit allerdings noch nicht, zwischen welchen Gruppen signifikante Unterschiede bestehen. Die Anwendung multipler *t*-Tests ist nicht adäquat, da damit eine Inflation des Typ-1-Fehlers einhergeht [4]. Der Vorteil der Varianzanalyse ist, dass hier adäquate Methoden für multiple Vergleiche zur Verfügung stehen. Ein weiterer Vorteil dieser Modelle ist, dass es möglich ist, adjustierte Mittelwerte zu berechnen, bei denen die Effekte weiterer erklärender Variablen berücksichtigt werden.

Aus dem Modell der Zweifachklassifikation mit Wechselwirkung lassen sich für den Faktor Rauchverhalten folgende adjustierte FEV₁-Mittelwerte berechnen: Nicht-Raucher 4,07, Ex-Raucher 3,61, Raucher 2,84 Liter. Bezüglich des Faktors Sport erhält man die adjustierten FEV₁-Mittelwerte 3,29 für Nicht-Sportler und 3,72 für Sportler. Vergleicht man diese Werte mit den rohen, nicht adjustierten Mittelwerten aus **Tab. 1**, so zeigt sich, dass insbesondere der Unterschied zwischen Sportlern und Nicht-Sportlern bei den adjustierten Mittelwerten größer ist als bei den rohen Mittelwerten. Die Berücksichtigung wichtiger Effekte in einem multifaktoriellen Modell führt zur Verminderung von Verzerrung sowie zur Erhöhung der statistischen Power, was im betrachteten Datenbeispiel dazu führt, dass ein statistisch signifikanter Effekt des Sports gefunden werden kann.

Mit Hilfe der so genannten Tukey-Kramer-Methode, die z. B. in der Software PROC GLM von SAS für multiple Vergleiche im Rahmen von Varianzanalysen zur Verfügung steht [7], ergibt sich die Aussage, dass sich alle drei Raucherstatusgruppen signifikant voneinander unterscheiden (alle $p < 0,05$). Aufgrund der signifikanten Wechselwirkung stellt sich noch die Frage, welche Gruppen der Kreuzklassifikation zwischen Rauchen und Sport

Tab. 4 Ergebnisse der Tukey-Kramer-Methode (p -Werte) für multiple Vergleiche aus dem Modell der Zweifachklassifikation mit Wechselwirkung zur Untersuchung des Effekts von Rauchverhalten und Sport auf das Merkmal FEV_1 .

Nr.	Rauchstatus	Sport	MW	Gruppennummer					
				1	2	3	4	5	6
1	Nicht-Raucher	ja	4,58	-	0,0085	0,0532	0,0064	<0,0001	<0,0001
2		nein	3,57	0,0085	-	0,9988	1,000	0,0973	0,0268
3	Ex-Raucher	ja	3,68	0,0532	0,9988	-	0,9946	0,0843	0,0265
4		nein	3,53	0,0064	1,000	0,9946	-	0,1530	0,0473
5	Raucher	ja	2,91	<0,0001	0,0973	0,0843	0,1530	-	0,9963
6		nein	2,78	<0,0001	0,0268	0,0265	0,0473	0,9963	-

Tab. 5 Übersetzung wichtiger englischer Begriffe (deutsch – englisch).

ungepaarter t -Test	unpaired t test
Homoskedastizität	homoscedasticity
Varianzanalyse	analysis of variance (ANOVA)
Modell der Einfachklassifikation	one-way ANOVA
Quadratsummen	sum of squares
Varianzanalysetabelle	ANOVA table
Modell der Zweifachklassifikation	two-way ANOVA
unbalanciert	unbalanced
Wechselwirkung	interaction
multiple Vergleiche	multiple comparisons
adjustierter Mittelwert	adjusted mean (least squares mean)
Varianzanalyse für Messwertwiederholungen	ANOVA for repeated measurements
Modellgüte	goodness-of-fit

sich signifikant voneinander unterscheiden. Diese Ergebnisse lassen sich am übersichtlichsten in einer (6×6)-Tabelle darstellen (siehe Tab. 4). Die höchsten FEV_1 -Werte hat im Durchschnitt die Gruppe der Nicht-Raucher mit regelmäßigem Sport; diese Gruppe ist signifikant von allen anderen verschieden (alle $p < 0,01$) außer der Gruppe der Ex-Raucher mit regelmäßigem Sport ($p = 0,0532$). Die Ergebnisse der anderen paarweisen Gruppenvergleiche sind Tab. 4 zu entnehmen.

Hinweise zur Modellbildung

Eine sinnvolle Anwendung von Modellen der Varianzanalyse setzt einen ausreichenden Stichprobenumfang voraus. Analog zum allgemeinen multiplen Regressionsmodell sollten für jeden freien Faktorterm im Modell mindestens 10 Beobachtungen zur Verfügung stehen [5]. Die Zahl der freien Faktorterm ergibt sich hierbei aus der Zahl der Faktorstufen abzüglich 1. Die Zahl der freien Wechselwirkungsterme ergibt sich aus der Multiplikation der beteiligten freien Faktorterm. In unserem Datenbeispiel ergibt sich mit dieser Faustregel für das Modell der Zweifachklassifikation ein minimaler Stichprobenumfang von $n = 30$ und für das Modell der Zweifachklassifikation mit Wechselwirkung von $n = 50$.

Die Varianzanalyse besitzt im Prinzip dieselben Voraussetzungen wie der t -Test, nämlich Normalverteilung und Homoskedastizität. In der Praxis ist – wie im ersten Abschnitt geschildert – die Annahme der Normalverteilung eher unkritisch. Bei kleinen Stichproben und deutlicher Abweichung von der Normalverteilung können entsprechende nicht-parametrische Verfahren wie z. B. der Kruskal-Wallis-Test eingesetzt werden [3]. Es existieren auch neuere nicht-parametrische Verfahren für mehrfaktorielle

Versuchspläne [6]. Ist die Annahme der Homoskedastizität stark verletzt, so sollten vor Anwendung von ANOVA-Modellen varianzstabilisierende Transformationen verwendet werden.

Die gewöhnliche Varianzanalyse ist als Verallgemeinerung des ungepaarten t -Tests anwendbar zur Analyse unabhängiger Stichproben. In der Datensituation mehrerer abhängiger Stichproben, also sozusagen in der Verallgemeinerung der Situation für den gepaarten t -Test, müssen andere Verfahren, z. B. die Varianzanalyse für Messwertwiederholungen angewendet werden [2].

Ähnlich wie bei der multiplen Regression spielen bei der Modellbildung von Varianzanalysen Verfahren zur Untersuchung der Modellgüte eine Rolle [5], auf die wir hier nicht eingehen können. Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt Tab. 5.

kurzgefasst

Mit Hilfe der Varianzanalyse lassen sich Unterschiede zwischen mehr als zwei Gruppen bezüglich ihrer Mittelwerte einer stetigen Zielvariablen statistisch untersuchen. Eine Berücksichtigung mehrerer Faktoren ist mit Hilfe von Modellen der Mehrfachklassifikation möglich. Um herauszufinden, zwischen welchen Gruppen signifikante Unterschiede bestehen, benötigt man Methoden für multiple Vergleiche.

Literatur

- Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ* 1996; 312: 1472–1473
- Bender R, Grouven U, Ziegler A. Varianzanalyse für Messwertwiederholungen. *Dtsch med Wochenschr* 2007; 132: e61–e64
- Bender R, Lange S, Ziegler A. Wichtige Signifikanztests. *Dtsch med Wochenschr* 2007; 132: e24–e25
- Bender R, Lange S, Ziegler A. Multiples Testen. *Dtsch Med Wochenschr* 2007; 132: e26–e29
- Bender R, Ziegler A, Lange S. Multiple Regression. *Dtsch med Wochenschr* 2007; 132: e30–e32
- Brunner E, Munzel U. Nicht-parametrische Datenanalyse. Springer, Berlin, Heidelberg, New York, 2002
- Freund RJ, Littell RC, Spector PC. SAS System for Linear Models. SAS Institute Inc., Cary, NC, 1991
- Lange S, Bender R. Variabilitätsmaße. *Dtsch Med Wochenschr* 2007; 132: e5–e6
- Lange S, Bender R. Was ist ein Signifikanztest? *Dtsch med Wochenschr* 2007; 132: e19–e21
- Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* 2002; 23: 151–169
- 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