Parametrische und nichtparametrische simultane Konfidenzintervalle für multiple Kontraste

Edgar Brunner
Abt. Medizinische Statistik
Humboldtallee 32
37073 Göttingen
brunner@ams.med.uni-goettingen.de

Frank Konietschke
Abt. Medizinische Statistik
Humboldtallee 32
37073 Göttingen
Frank.Konietschke@medizin.uni-goettingen.de

Zusammenfassung

Wenn in Versuchen oder Studien in den Biowissenschaften mehr als zwei Stichproben erhoben werden, dann erfolgt die statistische Auswertung in der Regel in drei klassischen Stufen:

1. Testen der Globalhypothese: Haben die Faktoren einen Einfluss auf die Messgröße?
2. Bei Ablehnung der Globalhypothese werden multiple Vergleiche durchgeführt: Welche Komponenten führen zur Ablehnung der Hypothese? Dabei ist eine multiple Fehlerrate einzuhalten.
3. Um die Stärke der Effekte und die Variabilität zu veranschaulichen, werden Konfidenzintervalle für die Effekte berechnet. Diese sollten simultan das gewählte Niveau einhalten und zu den multiplen Vergleichen kompatibel sein.

In der Arbeit von Bretz, Genz und Hothorn (2001) werden Verfahren für normalverteilte Daten vorgestellt, welche diese drei klassischen Schritte der Auswertung solcher Versuche in einem erledigen und gleichzeitig auf die speziellen Fragen von Anwendern zugeschnitten werden können.

Vorgestellt und diskutiert werden neue parametrische und nichtparametrische simultane Konfidenzintervalle und multiple Kontrasttests, welche genau diese Problematik einer ANOVA umgehen.

Schlüsselwörter: Simultane Konfidenzintervalle, multiple Kontrasttests, multivariate t Verteilung

1 Parametrische und nichtparametrische simultane Konfidenzintervalle und multiple Kontrasttests

Die meisten Versuche und Studien in den Biowissenschaften haben eine faktorielle Struktur, d.h. mehrere Faktoren beeinflussen gleichzeitig den interessierenden Messwert. In der Medizin sind dies z.B. multizentrische Studien oder Verlaufskurven für Individuen, die verschiedene Therapien erhalten. Beispiele hierfür findet man in vielen Lehrbüchern, wie z.B. in Brunner und Munzel (2002), Brunner, Domhof und Langer

(2002), Verbeke und Molenberghs (2000), Davis (2002), Diggle, Liang und Zeger (1994).

Die Analyse solcher Daten geschieht klassischerweise unter Annahme der Normalverteilung der Messwerte mithilfe einer Varianzanalyse (ANOVA). Dabei werden die eigentlichen Fragen des Anwenders in drei Schritten beantwortet:

1. Zunächst wird eine ANOVA durchgeführt, welche die Frage beantwortet, ob die im Versuch berücksichtigten Faktoren einen Einfluss auf die Messgröße haben oder nicht. Hat ein als signifikant identifizierter Faktor mehr als zwei Stufen (z.B. mehrere Behandlungen oder Zentren), dann reicht die einfache Antwort der ANOVA „signifikanter Faktoreinfluss“ nicht aus, da der Anwender normalerweise wissen möchte, welche Faktorstufen (Behandlungen) für das signifikante Ergebnis verantwortlich sind. Dazu müssen dann so genannte „multiple Vergleiche“ für die einzelnen Faktorstufen durchgeführt werden.

2. Die sachgerechte Durchführung von multiplen Vergleichen erfordert die Kontrolle des gewählten Niveaus, um eine Aufblähung des Fehlers 1. Art (i.Allg. 5%) infolge der Multiplizität zu vermeiden. Hierzu steht eine Fülle von Verfahren zur Verfügung, die unter teilweise restriktiven Voraussetzungen alle ihre Vor- und Nachteile haben. Ein gleichmäßig bestes Verfahren ist bisher nicht bekannt. Der Nachteil dieser Verfahren ist, dass sie mehr oder weniger konservativ sind, d.h. sie unterschreiten die gewählte multiple Irrtumswahrscheinlichkeit. Dies geht natürlich mit einem mehr oder weniger großen Verlust an Power einher. Für die Praxis bedeutet dies, dass entweder relevante Unterschiede mit zu geringer Wahrscheinlichkeit aufgedeckt werden oder dass der benötigte Stichprobenumfang zu hoch angesetzt werden muss. Fast alle diese Verfahren können die Information gewisser Abhängigkeiten zwischen den Statistiken einzelner Vergleiche nicht ausnutzen.

3. Meist reicht die einfache Antwort, dass zwischen zwei Faktorstufen (z.B. Kontrolle gegen Behandlung 1) ein signifikanter Unterschied besteht, für den Anwender nicht aus. Über das Ausmaß des Unterschieds und einer möglichen Variabilität gibt nur ein Konfidenzintervall für einen Effekt eine ausreichende Information für den Anwender. Ein solches Konfidenzintervall wird sogar von den Regulierungsbehörden bei der Zulassung eines Arzneimittels gefordert (ICH E9 Guideline). Die einfache Angabe eines p-Wertes für einen Vergleich genügt nicht. Die Schwierigkeit bei der Angabe eines Konfidenzintervalls für einen Behandlungseffekt nach einer simultanen Testprozedur besteht nun darin, dass ein solches Konfidenzintervall ebenfalls das multiple Niveau einhalten und zusätzlich noch mit dem verwendeten multiplen Vergleichsverfahren kompatibel sein sollte. Das heißt, dass es z.B. nicht passieren darf, dass die Hypothese in einem Paarvergleich zum multiplen Niveau α abgelehnt wird, das nachfolgend berechnete Konfidenzintervall für den Effekt aber die Null enthält. Von den allgemein gülti-

gen Verfahren erfüllen z.B. die sehr konservativen Verfahren nach Bonferroni und Scheffé diese Anforderung. Für andere Verfahren gibt es erste Ansätze, die aber sehr umständlich sind und zu schwer interpretierbaren Intervallen führen (siehe z.B. Guilbaud, 2008).

Die gesamte Situation ist schon im Bereich der Normalverteilungsmodelle nicht zufriedenstellend. Für den Fall, dass keine Normalverteilung der Daten angenommen werden kann, besonders für rein ordinale Daten, ist das vorhandene Defizit an praktikablen Verfahren noch größer.

Da aber gerade in den Biowissenschaften sehr oft Daten vorliegen, für die keine Normalverteilung angenommen werden kann, ist hier der Bedarf an Verfahren zur adäquaten Auswertung von Versuchen und Studien von besonderem Interesse. Beispielsweise folgen Scoredaten, ordinale Daten oder stetige, schief verteilte Daten keiner Normalverteilung. Um auch diese in der Praxis häufig auftretenden Datentypen ohne die Annahme einer zugrunde liegenden speziellen Verteilung adäquat auswerten zu können, müssen nichtparametrische Verfahren entwickelt werden. Beispiele zu solchen Datensätzen sind z.B. in den Büchern Brunner und Munzel (2002) und Brunner, Domhof und Langer (2002) beschrieben. Diese Beispiele werden alle in der oben beschriebenen (klassischen) Form in drei Schritten ausgewertet, wobei schon der zweite Schritt zu Problemen führen kann und der dritte meistens nicht ausgeführt wird, da keine adäquaten Verfahren zur Verfügung stehen.

Hier bietet nun die Idee von Bretz, Genz und Hothorn (2001) einen hervorragenden Ansatzpunkt zur Entwicklung solcher Verfahren. In dieser Arbeit werden multiple Kontrasttests für homogene normalverteilte Zufallsvariablen vorgestellt, mit denen die Fragen von Anwendern durch speziell auf die Frage zugeschnittene Verfahren beantwortet werden können. Das Verfahren vereinigt die drei Stufen der „klassischen“ Varianzanalyse (vgl. die Punkte 1,2,3) in einem einzigen Schritt. Die Dualität zwischen den simultanen Konfidenzintervallen und den assoziierten Teilhypothesen ist theoretisch gesichert, so dass keine Widersprüche in Form von Ablehnung / Nichtablehnung entstehen können. In der Arbeit von Hasler und Hothorn (2008) werden die Ergebnisse von Bretz et. al. (2001) auf normalverteilte heteroskedastische Modelle verallgemeinert.

Munzel und Hothorn (2001) stellen nichtparametrische simultane Konfidenzintervalle für relative Effekte in Tukey-Vergleichen (All – Pairs Vergleiche) für a unverbundene Stichproben in einer Einweg-Versuchsanlage vor. Diese Verfahren wurden von Konietschke (2009) auf beliebige Kontraste unter Kontrolle des multiplen Niveaus im starken Sinne erweitert.

Im Rahmen des Tutoriums sollen die parametrischen und nichtparametrischen Resultate von Bretz et al. (2001) sowie deren Erweiterungen von Hasler und Hothorn (2008) und von Konietschke (2009) dargelegt werden. Zur Durchführung der umfangreichen Rechnungen ist insbesondere die numerische Verfügbarkeit der multivariaten t-Vertei-

lung mit beliebiger Korrelationsstruktur notwendig. Bisher stehen Algorithmen zur Berechnung multivariater t-Wahrscheinlichkeiten nur in der freien Software R (www.r-project.org) zur Verfügung, so dass die praktische Benutzung bisher nur anhand dieser Software demonstriert werden kann.

Literatur

[1] Bretz, F., Genz, A. und Hothorn, L.A. (2001). On the Numerical Availability of Multiple Comparison Procedures. *Biometrical Journal* **43,** 645-656.

[2] Brunner, E., Domhof, S. und Langer, F. (2002). Nonparametric Analysis of Longitudinal Data in Factorial Experiments. Wiley.

[3] Brunner, E. und Munzel, U. (2002). *Nichtparametrische Datenanalyse*. Springer, Berlin.

[4] Davis, C. S. (2002). Statistical *Methods for the Analysis of Repeated Measurements*. Springer, New York.

[5] Diggle, P. J., Liang, K.-Y. und Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press

[6] Guilbaud, O. (2008). Simultaneous Confidence Regions Corresponding to Holm's Step Down Procedure and Other Closed-Testing Procedures. Biometrical Journal, DOI: 678-69218932131.

[7] Hasler, M. und Hothorn, L.A. (2008). Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*

[8] Munzel, U., und Hothorn, L.A. (2001). A Unified Approach to Simultaneous Rank Test Procedures in the Unbalanced One-Way Layout. *Biometrical Journal*, **43**, 553-569.

[9] Konietschke, F. (2009). Simultane Konfidenzintervalle für nichtparametrische relative Kontrasteffekte. Dissertation an der Georg-August-Universität Göttingen.

[10] Verbeke, G. und Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York

Anhang

Die im Tutorium verwendeten Folien finden Sie als PDF-Datei im deutschsprachigen SAS-Wiki unter http://de.saswiki.org/wiki/KSFE_2010.