Simultane Konfidenzintervalle für multiple Vergleiche - Ansätze für nicht Gauß-verteilte Daten und Inferenz in linearen Modellen

Vorgelegte Habilitationsschrift
zur Erlangung der Lehrbefugnis (*venia legendi*)
für das Fachgebiet Biostatistik
an der Gottfried Wilhelm Leibniz Universität Hannover

von

Dr. rer. hort. Frank Schaarschmidt

geboren am 27.02.1979 in Jena.

Datum der Einreichung: 14.12.2015

Zusammenfassung

Simultane Konfidenzintervalle

In kontrollierten Experimenten mit mehreren Behandlungen stellt sich oft zuerst die Frage nach allgemeinen Unterschieden, also, ob überhaupt irgendwelche der Behandlungen zu Unterschieden in den gemessenen Variablen führen. Globale statistische Tests, wie z.B. der F-Test der Varianzanalyse (ANOVA) oder der χ^2-Test auf Unabhängigkeit in Kontingenztafeln erlauben es, anhand der experimentellen Daten Interpretationen abzuleiten wie 'Die Behandlungen unterscheiden sich im Mittel', oder die 'Häufigkeiten mehrerer Kategorien unterscheidet sich zwischen den Behandlungen'. Bei korrekter Anwendung stellen diese Tests sicher, dass die so generierten Aussagen nur in einem geringen Anteil (z.B. $\alpha = 0.05$) von Anwendungen falsch positiv sind, d.h. rein zufällige Schwankungen in den Daten dazu führen, dass man auf das Vorhandensein von signifikanten Unterschieden schließt.

In den seltensten Fällen ist das Bedürfnis nach Interpretation in Experimenten mit mehreren Behandlungen mit solchen allgemeinen Aussagen erschöpft. Es schließen sich Fragen an, wie: 'Welche der Behandlungen unterscheiden sich untereinander?', 'Für welche Behandlungen kann man erwarten, dass sie zu höheren oder niedrigeren Werten der Variable führen, als welche der übrigen Behandlungen?', 'Wie groß sind die Unterschiede zwischen den Behandlungen?', 'Sind die Unterschiede nur statistisch signifikant, oder sind sie auch praktisch relevant?' Da manche dieser Detailinterpretationen zu den wesentlichen Aussagen eines Experiments werden können, sollte auch für sie die Wahrscheinlichkeit falscher Aussagen auf niedrigem Niveau (z.B. $\alpha = 0.05$) kontrolliert werden. Das heißt, statistische Verfahren werden benötigt, die die Vielzahl sich ergebender Einzelfragestellungen testen und Richtung und Größe der Unterschiede so schätzen, dass daraus abgeleitete Aussagen nur mit geringer Wahrscheinlichkeit falsch sind.

Häufig werden die einzelnen Behandlungen im Experiment untersucht, um eine Gesamtfragestellung zu beantworten: Solche Situationen liegen vor, wenn in einer Vielzahl von Behandlungen bereits einzelne gefundene Unterschiede eine wichtige Interpretation darstellen. Zum Beispiel, wenn unter einer größeren Zahl Behandlungen (ohne *a priori* Ordnung), einige identifiziert werden sollen, die eine Verbesserung versprechen:

- Kann durch eine Reihe verschieden zusammengesetzter Vermehrungsmedien eine Erhöhung der Regenerationsrate in einem *in-vitro* Vermehrungssystem erreicht werden? Welche unter dieser Vermehrungsmedien führen zu einer höheren Regenerationsrate als welche anderen? Um wieviel kann die Regenerationsrate mit dem empirisch besten Medium erhöht werden, im Vergleich zu den übrigen Medien?

- Gibt es in einer Auswahl Genotypen Unterschiede bezüglich der Gehalte bestimmter Inhaltsstoffe? Welche Genotypen enthalten wieviel mehr als welche anderen?

In ähnlicher Weise werden bei Untersuchungen zur Toxizität oder Karzinogenität von Substanzen in der Regel mehrere Dosierungen über einen weiten Dosis-Bereich im Vergleich zu einer unbehandelten Kontrollbehandlung bewertet: 'Führt mindestens eine der geprüften Dosierungen zu einem Anstieg der Häufigkeit von Gewebs-Läsionen, Mutationen oder Tumoren im Vergleich zur unbehandelten Kontrolle?', 'Um wieviel ist deren Anteil im Vergleich zur Kontrolle erhöht?' Ähnliche Probleme ergeben sich auch bei der Bewertung der Unbedenklichkeit von Substanzen: 'Bis zu welcher Dosis ist der Anteil der Läsionen nicht stärker als eine vorgegebene Grenze (z.B. 125% des Werts in der Kontrolle) erhöht?'

Der Fokus der folgenden Publikationen liegt deshalb auf simultanen Konfidenzintervallen. Einfache Konfidenzintervalle geben einen aus den Daten berechneten Bereich (Unter- und Obergrenze) an, in dem **ein** wahrer zu schätzender Parameter mit hoher Wahrscheinlichkeit (z.B. $1 - \alpha = 0.95$) liegt, d.h. in 95% der Anwendungsfälle der Methodik, bei gegebenen Vorraussetzungen. Sie erlauben damit Hypothesentests für diesen Parameter zum Niveau α, d.h. zum Beispiel Aussagen zur Signifikanz des Unterschieds zwischen zwei Behandlungen. Sie erlauben aber darüber hinaus Interpretationen

zu Richtung und Größenordnung des Parameters, und damit fachliche fundierte Interpretationen zur praktischen Relevanz von beobachteten Unterschieden.

Simultane Konfidenzintervalle sind die Erweiterung dieser Methodik auf Fälle, in denen mehrere Parameter geschätzt werden sollen, wenn gleichzeitig die Wahrscheinlichkeit mindestens einer falschen Aussage auf niedrigem Niveau kontrolliert werden soll. Das heißt, für jeden einzelnen zu schätzenden Parameter wird ein Bereich so angegeben, dass die Wahrscheinlichkeit, dass **alle** wahren Parameter in ihrem zugehörigen Bereich liegen, auf hohem Niveau kontrolliert wird, z.B. $1 - \alpha = 0.95$. Umgekehrt beträgt dann die Wahrscheinlichkeit, dass mindestens ein Parameter nicht im zugehörigen Intervall liegt, $\alpha = 0.05$. Das bedeutet, dass simultane Konfidenzintervalle Interpretationen zu allen Typen der oben gestellten Fragen erlauben: Signifikanz der Gesamtfragestellung, Detailunterschiede bezüglich Richtung und Größenordnung der Unterschiede, sowie, bei gegebenen fachlichem Hintergrundwissen, zur praktischen Relevanz der beobachteten Unterschiede. Im Beispiel der Auswahl aus mehreren Medien zur *in-vitro* Vermehrung könnten simultane 95% Konfidenzintervalle für alle paarweisen Differenzen der mittleren Vermehrungsraten der Medien berechnet werden. Wenn mindestens eins dieser Intervalle den Wert 0 nicht enthält, kann zum Signifikanzniveau 5% geschlossen werden, dass es Unterschiede zwischen den Medien bezüglich der mittleren Vermehrungsraten gibt. Gleichzeitig kann interpretiert werden, welche der Medien sich im Einzelnen von welchen anderen unterscheiden (Welche Intervalle der einzelnen paarweisen Differenzen den Wert 0 nicht enthalten). Die Grenzen der Konfidenzintervalle geben (unter Kontrolle der Wahrscheinlichkeit mindestens einer falschen Aussage) an, wie groß die Unterschiede der mittleren Vermehrungsraten zwischen ausgewählten Medien sind.

Verfügbare statistische Methodik

Statistische Methoden zur Berechnung von simultanen Konfidenzintervallen sind seit langem verfügbar für einfaktorielle Versuchsanlagen und die Annahme Gauß-verteilter Residuen mit homogenen Varianzen: Die Verfahren zum Tukey-Test (Tukey, 1953) erlauben auch die Berechnung simultaner Konfidenzintervalle für alle paarweisen Vergleiche zwischen mehreren Behandlungsmittelwerten. Der Test nach Dunnett (Dunnett, 1955) bietet die Lösung für den wichtigen Spezialfall, dass nur die paarweisen Differenzen mehrerer Behandlungsmittelwerte zum Mittelwert einer Kontrollgruppe von Interesse sind. Natürlich lassen sich nicht alle wissenschaftlichen Fragestellungen in diesen beiden Spezialfällen abbilden. Bretz et al. (2001, 2002) beschreiben unter den obigen Annahmen Methoden zur allgemeinen Definition multipler Vergleiche zwischen mehreren Behandlungsmittelwerten: Die Theorie multiple Kontrasttests erlaubt es, nur für diejenigen Differenzen von Behandlungsmittelwerten simultane Konfidenzintervalle zu schätzen, die in einer bestimmten wissenschaftlicher Fragestellung tatsächlich von Interesse sind.

Die obige Annahme Gauß-verteilter, varianzhomogener Residuen ist in praktischen Versuchen in den Biowissenschaften selten eindeutig gegeben, für eine Reihe häufiger Datentypen aber eindeutig nicht gegeben. Westfall et al. (1993) beschreiben Verfahren für alle paarweisen Mittelwertsdifferenzen, die auf Resampling der Residuen basieren und daher nicht auf einer Gaußverteilungsannahme beruhen. Mit diesen Verfahren können auch simultane Konfidenzintervalle für ein nutzerdefiniertes Set von Mittelwertsdifferenzen berechnet werden. Hothorn et al. (2008) beschreiben schließlich asymptotische Verfahren für eine sehr allgemeine Klasse von statistischen Modellen, die ebenfalls die Berechnung simultaner Konfidenzintervalle für eine nutzerdefinerte Auswahl oder Linearkombinationen der Parameter solcher Modelle erlauben. In dieser Methodik sind eine Reihe von Spezialfällen enthalten, die eine wichtige Rolle in den Biowissenschaften spielen: Generalisierte lineare Modelle (McCullagh and Nelder, 1989) erlauben unter anderem die Modellierung binomialer Daten oder Zähldaten, lineare gemischte Modelle sind für die Analyse von Experimenten mit komplexer Randomisierungstruktur notwendig, sowie Modelle für die Analyse von Überlebenszeiten.

All diese Verfahren berücksichtigen zur Berechnung der Intervallgrenzen nicht nur die Anzahl zu schätzender Parameter, sondern auch die Korrelation der zugehörigen Teststatistiken oder Schätzfunktionen. Im Fall einfaktorieller Versuchsanlagen mit Annahme unabhängiger, Gauß-verteilter, varianzhomogener Residuen (Bretz et al., 2001) hängt die Korrelation der Teststatistiken nur von den Stichprobenumfängen und dem Set ausgewählter Mittelwertsvergleiche ab (Bretz et al., 2001), nicht aber von unbekannten Parametern. Die gemeinsame Verteilung der Teststatistiken ist dann die multivariate t-Verteilung (Bretz et al., 2001), deren Korrelation exakt bekannt, und auch die resultierenden Konfidenzintervalle sind exakt, wenn die zugrunde liegenden Annahmen gelten. Die Verfahren von Westfall et al. (1993) bilden die Korrelation der Teststatistiken implizit ab, indem deren gemeinsame Verteilung durch Resampling der Residuen angenähert wird. In den von Hothorn et al. (2008) betrachteten Modellen kann die Korrelation der Teststatistiken zusätzlich von den Parametern von Interesse abhängen: Für große Stichprobenumfänge kann die aufgrund der Stichprobenschätzer eben-

falls geschätzte Korrelation verwendet werden, um kritische Werte der multivariaten Normalverteilung zur Berechnung simultaner Konfidenzintervalle zu verwenden. In der praktischen Anwendung schließt sich daher die Frage an, welche Stichprobenumfänge in gegebenen Modellen groß genug sind, dass die berechneten Intervalle tatsächlich alle Parameter von Interesse mit Wahrscheinlichkeit $(1 - \alpha)$ enthalten.

Aus dieser Situationsbeschreibung ergibt sich folgender Bedarf an biostatistischer Forschung und Darstellung der Methodik:

- Die verfügbaren statistischen Methoden sollten für die praktische Anwendung verständlich dargestellt werden.
- Asymptotische Methoden sollten validiert werden: Für welche Parameterkonstellationen und Stichprobenumfänge enthalten sie alle wahren Parameter annähernd mit der vorgegebenen Wahrscheinlichkeit $(1 - \alpha)$? Diese simultane Überdeckungswahrscheinlichkeit (simultaneous coverage probability) kann in Monte-Carlo Simulationen untersucht werden.
- Können für kleine Stichprobenumfänge Verfahren bereitgestellt werden, die verbesserte Eigenschaften im Vergleich zu den verfügbaren asymptotischen Verfahren aufweisen? Als wesentliche Ansätze werden dazu im Folgenden Verfahren verwendet, die große Stichproben aus der gemeinsamen (d.h. multivariaten) Verteilung der Parameter von Interesse generieren, oder an diese annähern sollen. Auf Basis dieser Stichproben können mit Hilfe der in Besag et al. (1995) und Mandel and Betensky (2008) beschriebenen Perzentilintervalle simultane Konfidenzintervalle berechnet werden.

Einführung in nutzerdefinierte Kontraste von Behandlungsmittelwerten

Multiple Kontraste (Bretz et al., 2001) erlauben es, simultane Konfidenzintervalle für eine ganz bestimmte Auswahl von Differenzen von Mittelwerten zu schätzen. Die zu schätzenden Parameter können dabei auf die speziellen wissenschaftlichen Fragestellungen eines Versuches zugeschnitten werden, wenn Standardmethoden, wie alle paarweisen Vergleiche (Tukey, 1953) oder Vergleiche zur Kontrolle (Dunnett, 1955) die Versuchsfragestellungen nur unzureichend abbilden. Die Arbeit von **Schaarschmidt and Vaas (2009)** stellt die verfügbare Methodik unter Annahme varianz-homogener, Gauß-verteilter, unabhängiger Residuen für eine agrarwissenschaftliche Leserschaft dar. Im besonderen wird auf die Analyse von Versuchen eingegangen, die zwar zwei Behandlungsfaktoren mit jeweils mehreren Stufen enthalten, aber nicht alle möglichen Kombinationen der Stufen beider Faktoren. Solche Behandlungsstrukturen sind dann nicht mit den einfachsten Ansätzen für vollständig kreuzklassifizierte zweifaktorielle Versuche auswertbar, Piepho et al (2006) zeigen, wie allgemeine lineare Modelle strukturiert werden müssen, wenn solche Versuche mit F-tests der Varianzanalyse ausgewertet werden sollen. In Schaarschmidt and Vaas (2009) wird anhand von Beispielen dargestellt, wie mehrere Mittelwertsvergleiche von Interesse als Matrix von Kontrastkoeffizienten definiert werden können, und durch die Darstellung simultaner Konfidenzintervalle Detailinterpretationen zu Signifikanz und Größe der Effekte gemacht werden können, während gleichzeitig die *family-wise error rate* für diese Interpretationen eingehalten wird.

In **Kitsche and Schaarschmidt (2015)** wird ein zweiter Anwendungsbereich simultaner Konfidenzintervalle für multiple Kontraste dargestellt, also bereits publizierte statistische Methodik an Beispieldatensätzen illustriert. In manchen zweifaktoriellen Versuchen ist die Interaktion beider Faktoren nicht störend bei der Interpretation der Haupteffekte, sondern das primäre Versuchsziel besteht im Nachweis der Interaktion, und deren genaueren Analyse. Zum Beispiel werden in Versuchen zur Trockenstress-toleranz einer kleineren Auswahl von Genotypen häufig Pflanzen jedes Genotyps wiederholt sowohl unter Kontrollbedingungen als auch mehreren Stressbehandlungen angebaut. Versuchsziel ist dann nicht der Nachweis von Unterschieden zwischen den Genotypen, oder der prinzipielle Nachweis, dass die Stressbehandlungen physiologische Veränderungen in den Pflanzen erzeugen. Gesamtfragestellung ist vielmehr, ob einzelne Genotypen anders oder erst bei stärkerem Stress reagieren, als andere Genotypen. Diese Gesamtfragestellung kann im F-Test für den Interaktionsterm der zweifaktoriellen Varianzanalyse getestet werden. Da es sich um eine quadratische Teststatistik handelt, kann nur die allgemeine (globale) Signifikanz des Terms angegeben werden. Der F-Test erlaubt keine Detailinterpretationen wie: 'Welche einzelnen Genotypen unterscheiden sich in ihrer Reaktion auf welche der Stressbehandlungen von der Reaktion welcher anderen Genotypen?' oder gerichtete Fragestellungen wie 'Welche Genotypen zeigen unter Stress **geringere** mittlere Ertragsreduktionen im

Vergleich zu Kontrollbedingungen als welche anderen Genotypen?' Diese letzte Formulierung beinhaltet alle paarweise Vergleiche zwischen den Genotypen, bezüglich deren Mittelwertsdifferenzen zwischen Stress- und Kontrollbehandlungen. Die Parameter von Interesse sind also eine Teilmenge aller möglichen Differenzen von Differenzen in beiden Faktoren. In (Kitsche and Schaarschmidt, 2015) wird dargestellt, wie solche speziellen Teilmengen des allgemeinen Tests auf Interaktion als multiple Kontraste kompakt definiert werden können. Neben Verweisen auf ein Zusatzpaket zur R-Software für die Erzeugung solcher Kontrastmatrizen enthält (Kitsche and Schaarschmidt, 2015) auch eine etwas komplexere Auswertung für einen dreifaktoriellen Versuch.

In Versuchen mit mehreren Behandlungen von Interesse bedeutet das Vorhandensein einer Interaktion mit einer numerischen Kovariable, dass die möglichen Behandlungsunterschiede nicht für alle Werte der Kovariable vorhanden, gleich gerichtet, oder gleich groß sind. Die Behandlungsunterschiede können dann in Abhängigkeit der Kovariablenwerte genauer untersucht werden: 'Für welche Wertebereiche der Kovariable sind zwischen welchen der Behandlungen Unterschiede nachweisbar, und wie groß sind diese?' Simultane Konfidenzbänder für Differenzen zwischen mehreren Regressionsgeraden sind in der statistischen Literatur in einer großen Zahl von Spezialfällen beschrieben, es fehlt aber nutzerfreundliche Software. In **Schaarschmidt (submitted)** wird illustriert, dass die in Software verfügbare Methodik in Hothorn et al. (2008) verwendet werden kann, um solche Vergleiche für ein diskretes Set von Kovariablenwerten durchzuführen, was praktisch zur gleichen Interpretation führt, wie die Anwendung simultaner Konfidenzbänder, wenn das Set von Kovariablenwerten groß genug ist. Neben der numerischen Verfügbarkeit hat dieser Ansatz den Vorteil, dass er sich auf Quotienten von Regressionsgeraden, oder den linearen Prädiktor generalisierter linearer Modelle (McCullagh and Nelder, 1989) erweitern lässt. Zur Erzeugung der sehr umfangreichen Kontrastmatrizen wird auf ein R-Paket verwiesen.

Validierung verfügbarer Verfahren und Verbesserungen für kleine Stichprobenumfänge

Dichotome Daten sind ein häufiger Datentyp in Untersuchungen zur Toxizität von Substanzen in Bioassays, Mortalitäten, aber auch Infektionsraten bei der Untersuchung von Pflanzenkrankheiten oder Parasiten, oder auch Heilungsraten in klinischen Versuchen. Wenn im einfachsten Fall vollständig randomisierter Versuche mit mehreren (k) Behandlungen jede Versuchseinheit in eine von zwei möglichen Kategorien eingeordnet wird, können die Daten als $(k \times 2)$ Kontingenztafeln dargestellt werden. Analog zu Tukey-Test oder Dunnett-Test und multiplen Kontrasten für Erwartungswerte Gauß-verteilter Daten, können dann multiple Kontraste für binomiale Proportionen die detaillierten Versuchsfragen abbilden. Bei der Übertragung der Methoden unter Gauß-Verteilungsannahme ergeben sich eine Reihe neuer Probleme: die Darstellung der Behandlungsunterschiede für Proportionen $\pi_i, \pi_{i'}$ kann als Differenz $\pi_i - \pi_{i'}$, Quotient $\pi_i/\pi_{i'}$ oder Oddsratio $(\pi_i/(1-\pi_i))/(\pi_{i'}/(1-\pi_{i'}))$ erfolgen. Schon für Konfidenzintervalle zum Vergleich von zwei Proportionen werden jeweils verschiedene approximative Verfahren diskutiert (z.B. Agresti and Caffo, 2000; Brown and Li, 2005; Newcombe, 1998). Bei multiplen Vergleichen mehrerer binomialer Proportionen, hängt die Korrelation der Teststatistiken bzw. Schätzfunktionen nicht mehr nur von den Stichprobenumfängen und gewählten Kontrasten ab, sondern von den einzelnen zu schätzenden Proportionen selbst (z.B. Bretz and Hothorn, 2002). Auf dieser Grundlage wurden in **Schaarschmidt, Sill and Hothorn(2008)** asymptotische Intervalle für die Differenz zweier Proportionen auf den Fall multipler Kontraste von Proportionen übertragen, indem die geschätzte Korrelationsmatrix zur Berechnung von Quantilen der multivariaten Normalverteilung verwendet wurde. Um akzeptable simultane Überdeckungswahrscheinlichkeiten bei kleineren Stichprobenumfängen zu erreichen, wurden zuvor empfohlene approximative Methoden (Agresti and Caffo, 2000; Brown and Li, 2005) zur Berechnung der Intervallgrenzen für eine Differenz binomialer Proportionen zusätzlich untersucht. In Monte-Carlo-Simulationen wird dann gezeigt, dass diese approximativen Methoden zumindest für Stichprobenumfänge von $n_i > 40$ je Behandlungsgruppe eine simultane Überdeckungswahrscheinlichkeit zwischen 0.94 und 0.96 für weite Bereiche des Parameterraum einhalten, und für den kritischen Fall sehr kleiner Proportionen, zu große, das heißt konservative Überdeckungswahrscheinlichkeiten nahe 1 zeigen.

Die Verfügbarkeit von validen Verfahren zur Berechnung von Konfidenzintervallen zum Vergleich von zwei Proportionen ist eine wichtige Vorraussetzung für die Erweiterung auf simultane Konfidenzintervalle, aber einfacher zu untersuchen. In (Schaarschmidt, accepted) werden diese Grundlagen zur Analyse überdisperser Binomialdaten aus Bioassays für Quotienten von Proportionen (relative risk) gelegt.

Wesentlich gravierendere Probleme bei der Validität von Verfahren für kleine Stichproben ergeben sich bei multiplen Vergleichen für Biodiversitätsindizes in ökologischen Fragestellungen **(Scherer, Schaarschmidt, Prescher und Priesnitz, 2013)**. Ausgangspunkt dieser Arbeit waren Feldversuche zur Sicherheitsbewertung genetisch veränderter Organismen in Bezug auf deren mögliche Wirkung auf Nichtzielorganismen. Biodiversitätsindizes wie der Shannon-Index oder Simpson-Index (z.B. Magurran, 2004) fassen dabei relative Häufigkeiten einzelner Spezies einer Artengemeinschaft zusammen, so dass große Werte hohen Anzahlen ähnlich häufiger Arten entsprechen. Verarmte Artengemeinschaften, in denen einzelne Spezies dominieren und viele seltene Spezies verschwunden sind, stellen sich in niedrigen Werten dieser Indizes dar. Solche Zusammenfassungen könnten mögliche Effekte auf viele seltene Spezies zeigen, für die separate Auswertung nur zu sehr unpräzisen Aussagen, d.h. zu sehr weiten Konfidenzintervallen führen (Rauschen et al., 2010). Simultane Konfidenzintervalle für diese Biodiversitätsindizes wurden bereits von Fritsch and Hsu (1999) und Rogers and Hsu (2001) beschrieben. Diese Methoden gehen von der vereinfachenden Annahme multinomialer Daten für die gezählten Spezies aus. Das primäre Ziel von Scherer et al. (2013) war es zu zeigen, dass diese Annahme in Feldversuchen und ökologischen Erhebungen grob verletzt sein kann: Wenn die erhobenen Speziesanzahlen deutlich höhere Varianzen aufweisen (Überdispersion), führen auch die Verfahren von Fritsch and Hsu (1999) und Rogers and Hsu (2001) nicht mehr zu validen Konfidenzintervallen (Scherer et al., 2013). Alternativen ohne spezielle Verteilungsannahmen sind das Resamplingverfahren von (Westfall et al., 1993) sowie nicht-parametrischer Bootstrap in Kombination mit simultanen Perzentilintervallen nach Besag et al. (1995) und Mandel and Betensky (2008). In Simulationsstudien zeigt sich, dass all diese Verfahren bei deutlicher Überdispersion den Verfahren von Fritsch and Hsu (1999) und Rogers and Hsu (2001) vorzuziehen sind, aber bei geringen Anzahlen von Wiederholungen teilweise nur Überdeckungswahrscheinlichkeiten von 0.8-0.9 oder darunter aufweisen, wenn 0.95 vorgegeben sind.

Im Fall positiver, kontinuierlicher, rechts-schief verteilter Daten werden die Beobachtungen nach log-Transformation oft mit Verfahren unter der Annahme Gauß-Verteilung analysiert. Die darausfolgenden Vergleiche für Mittelwertsdifferenzen der transformierten Variablen entsprechen Quotienten von Medianen auf der Originalskala. Wenn Interesse hauptsächlich am Vergleich der Erwartungswerte besteht, können simultane Konfidenzintervalle **(Schaarschmidt, 2013)** asymptotisch auf Grundlage von Krishnamoorthy and Mathew (2003), Chen and Zhou (2006) sowie Hothorn et al. (2008) berechnet werden. Alternativ können Sampling-Verfahren zur Annäherung der Verteilung der Schätzfunktion des Erwartungswerts für einzelne Behandlungsgruppen (z.B. Chen and Zhou, 2006, generalized pivotal quantities) verwendet werden, um die gemeinsame Verteilung multipler Kontraste darzustellen und mit den Perzentilverfahren von Besag et al. (1995) bzw. Mandel and Betensky (2008) simultane Konfidenzintervalle berechnet werden (Schaarschmidt, 2013). In Simulationsstudien lassen sich mit diesen Verfahren deutliche Verbesserungen im Vergleich zu den asymptotischen Verfahren erzielen (Schaarschmidt, 2013).

Multiple Vergleiche im allgemeinen linearen Modell oder linearen gemischten Modellen werden in der Regel für Differenzen von Modellparametern ausgedrückt. Die Bewertung der Relevanz von Unterschieden kann aber leichter fallen, wenn Unterschiede als Quotienten von Mittelwerten (Dilba et al., 2006) ausgedrückt sind. In anderen Situationen ergibt sich der Parameter von Interesse als Quotient der Modellparameter, z.B. bei der Berechnung von relative potencies (Djira, 2010). Aufbauend auf Young et al. (1997) und Dilba et al. (2006), beschreibt Djira (2010) asymptotische simultane Konfidenzintervalle für solche Quotienten in linearem gemischten Modellen. Hier bestehen ähnliche Probleme wie bei den asymptotischen Methoden zuvor: 1) wird die aus dem Modell geschätzte Kovarianzmatrix der zu schätzenden Parameter eingesetzt, um die Standardfehler der Parameterschätzer, sowie die Korrelationen der Teststatistiken zu berechnen, 2) hängt die Korrelation der Teststatistiken zusätzlich von den unbekannten Quotienten von Interesse ab, so dass zur Berechnung der Quantile der multivariaten Normalverteilung auch hier Schätzwerte eingesetzt werden. In **(Schaarschmidt and Djira, accepted)** werden, neben einfachen Verbesserungsvorschlägen wie der Verwendung multivariater t-Quantile, Sampling-Verfahren aus dem Bereich der Bayes'schen Statistik untersucht: Aus der gemeinsamen Verteilung der Quotienten von Interesse können mittels Markov-Chain-Monte-Carlo (MCMC) in hierarchischen Modellen (z.B. Gelman and Hill, 2007) große Stichproben gezogen werden, aus denen sich simultane Perzentilintervalle (Besag et al., 1995; Mandel and Betensky, 2008) berechnen lassen. Diese Verfahren haben den Nachteil, dass sie aufwendig in Software zu implementieren sind, die Prüfung der Konvergenz anspruchsvoller ist und von einer Reihe technischer Parameter abhängt. Zusätzlich müssen für eine frequentistische Interpretation der Intervalle (d.h. etwa 'Die wahren Parameter sind in 95% der Anwendungsfälle in den Intervallen enthalten') die *prior*-Verteilungen der Parameter so gewählt werden, dass sie möglichst geringen Einfluss auf die Parameter haben. In abgebildeten Szenarien der Simulationsstudien in Schaarschmidt and Djira (accepted) zeigen sich für diese aufwendigen

Methoden nur geringe Verbesserungen im Vergleich zur einfachen Bonferroni-Adjustierung, die oft auch durch Ersatz der multivariaten Normalquantile durch entsprechende t-Quantile erreicht werden können.

Anleihen aus der Bayes'schen Statistik werden auch in der folgenden Arbeit (Schaarschmidt, Gerhard and Vogel, in preparation) verwendet, um simultane Konfidenzintervalle mit verbesserten Eigenschaften bei kleinen Stichprobenumfängen zu erreichen. Wenn vollständig in randomisierten Versuchen die einzelnen Versuchseinheiten nicht in zwei, sondern mehrere Kategorien eingeteilt werden, entstehen im einfachsten Fall multinomiale Daten. Wenn die Kategorien nicht ordinal sind, braucht man mehrere Verhältnisse zwischen den Häufigkeiten der einzelnen Kategorien, um deren Verteilung in der Stichprobe zu beschreiben. Welche dieser Verhältnisse (odds) bei Vergleichen zwischen mehreren Behandlungen von praktischem Interesse sind, hängt naturgemäß von deren genauer fachlicher Bedeutung ab. In dieser Arbeit werden deshalb Methoden beschreiben, die sowohl einer nutzerdefinierte Auswahl der odds als auch eine nutzerdefinierte Auswahl von Vergleichen zwischen mehreren Behandlungsgruppen zulassen. Simultane Konfidenzintervalle können dann einerseits asymptotisch mit Quantilen der multivariaten Normalverteilung berechnet werden. Andererseits kann man ausnutzen, dass die Dirichlet-Verteilung als *prior*-Verteilung für multinomiale Proportionen auch zu einer *a posteriori* Dirichlet-Verteilung der Parameter führt (z.B. Agresti, 2013). Man kann dann leicht aus der gemeinsamen Verteilung der Parameter von Interesse Stichproben ziehen und simultane Konfidenzintervalle berechnen. Diese zeigen in Simulationsstudien insbesondere dann bessere Überdeckungswahrscheinlichkeiten als die asymptotischen Verfahren, wenn durch kleine Stichprobenumfänge oder relativ seltene Kategorien die Erwartungswerte der Anzahlen zumindest für einzelne Kategorien kleiner als 10 sind (Schaarschmidt, Gerhard and Vogel, in preparation).

Literaturverzeichnis

Agresti, A. (2013). Categorical Data Analysis (3rd ed). John Wiley & Sons, Inc., Hoboken; New Jersey.

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. The American Statistician 54, 280-288.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. Statistical Science 10, 3-66.

Bretz, F., Genz, A. and Hothorn, L.A. (2001). On the numerical availability of multiple comparison procedures. Biometrical Journal 43:645-656.

Bretz, F., Hothorn, T. and Westfall, P.H. (2002). On multiple comparisons in R. R News 2:14-17.

Bretz, F. and Hothorn, L. (2002). Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. Statistics in Medicine 21, 3325-3335.

Brown, L. and Li, X. (2005). Confidence intervals for two sample binomial distribution. Journal of Statistical Planning and Inference 130, 359-375.

Chen, Y.-H., Zhou, X.-H. (2006). Interval estimates for the ratio and difference of two lognormal means. Statistics in Medicine 25, 4099-4113.

Dilba, G., Bretz, F., and Guiard, V. (2006). Simultaneous confidence sets and confidence intervals for multiple ratios. Journal of Statistical Planning and Inference, 136(8), 2640-2658.

Djira, G. D. (2010). Relative potency estimation in parallel-line assays - method comparison and some extensions. Communications in Statistics - Theory and Methods, 39(7), 1180-1189.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal American Statistical Association 50: 1096-1121.

Fritsch, K. S., and Hsu, J. C. (1999). Multiple comparison of entropies with application to dinosaur biodiversity. Biometrics 55, 1300-1305.

Gelman, A. and Hill, J. (2007). Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge: Cambridge University Press.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal 50:346-363.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1994). Continuous Univariate Distributions, second ed. In: Wiley Series in Probability and Statistics, Vol. 1. John Wiley & Sons, New York.

Kitsche, A. and Schaarschmidt, F. (2015). Analysis of statistical interactions in factorial experiments. Journal of Agronomy and Crop Science 201 (1): 69-79, DOI: 10.1111/jac.12076

Krishnamoorthy, K. and Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of Statistical Planning and Inference 115, 103-121.

Magurran, A.E. (2004). Measuring Biological Diversity. Blackwell Publishing, Malden, MA.

Mandel, M. and Betensky, R.A. (2008). Simultaneous confidence intervals based on the percentile bootstrap approach. Computational Statistics & Data Analysis 52, 2158-2165.

McCullagh, P. and J.A. Nelder. (1989). Generalized linear models. Chapman & Hall/CRC, Boca Raton, FL.

Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17, 873-890.

Piepho, H.-P., Williams, E.R., and Fleck, M. (2006). A note on the analysis of designed experiments with complex treatment structure. HortScience 41:446-452.

Rauschen, S., Schaarschmidt, F., Gathmann, A. (2010). Occurrence and field densities of Coleoptera in the maize herb layer: implications for Environmental Risk Assessment of genetically modified Bt-maize. Transgenic Research 19:727-744. DOI: 10.1007/s11248-009-9351-3

Rogers, J. A. and Hsu, J. C. (2001). Multiple comparisons of biodiversity. Biometrical Journal 43, 617-625.

Schaarschmidt, F. (2013). Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. Computational Statistics and Data Analysis 58, 265-275. doi:10.1016/j.csda.2012.08.011

Schaarschmidt, F., Djira, G.D. Simultaneous confidence intervals for ratios of fixed effect parameters in linear mixed models. Accepted for publication in Communications in Statistics - Simulation and Computation. DOI: 10.1080/03610918.2013.849741

Schaarschmidt, F., Sill, M., and Hothorn, L.A. (2008). Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions. Biometrical Journal 50(5):782-792.

Schaarschmidt, F. and Vaas, L. (2009). Analysis of trials with complex treatment structure using multiple contrast tests. HortScience 44(1):188-195.

Schaarschmidt, F. Multiple treatment comparisons in analysis of covariance with interaction. Submitted to Statistical Methods and Application.

Scherer, R., Schaarschmidt, F., Prescher, S., Priesnitz, K.U. (2013). Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data. Biometrical Journal 55 (2), 246-263. DOI: 10.1002/bimj.201200157

Schaarschmidt, F.: Confidence Intervals for the Risk Ratio when Analyzing Bioassays in the Presence of Overdispersion. Accepted for publication in Biometrics and Biostatistics International Journal.

Schaarschmidt, F., Gerhard, D. and Vogel, C.: Simultaneous confidence intervals for comparisons of several multinomial samples.

Tukey, J. (1953). The problem of multiple comparisons, unpublished manuscript, reprinted in: Braun, H.I. (Ed.) 1994. The collected works of John W. Tukey. VIII. Multiple comparisons. Chapman and Hall, New York, NY.

Westfall, P.H. and Young, S.S. (1993). Resampling-Based Multiple Testing. John Wiley & Sons, New York.

Young, D.A., Zerbe, G.O., and Hay, W.W. (1997). Fiellers theorem, Scheffes simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effect models. Biometrics, 53(3), 835-847.

Wissenschaftliche Veröffentlichungen

1. Schaarschmidt, F. and Vaas, L. (2009). Analysis of trials with complex treatment structure using multiple contrast tests. HortScience 44(1):188-195.

2. Kitsche, A., Schaarschmidt, F. (2015): Analysis of statistical interactions in factorial experiments. Journal of Agronomy and Crop Science 201 (1): 69-79, DOI: 10.1111/jac.12076

3. Schaarschmidt, F. Multiple treatment comparisons in analysis of covariance with interaction. Submitted to Statistical Methods and Application.

4. Schaarschmidt, F., Sill, M., and Hothorn, L.A. (2008). Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions. Biometrical Journal 50(5):782-792.

5. Scherer, R., Schaarschmidt, F., Prescher, S., Priesnitz, K.U. (2013) Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data. Biometrical Journal 55 (2), 246-263. DOI: 10.1002/bimj.201200157

6. Schaarschmidt, F. (2013). Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. Computational Statistics and Data Analysis 58, 265-275. doi:10.1016/j.csda.2012.08.011

7. Schaarschmidt, F., Djira, G.D. Simultaneous confidence intervals for ratios of fixed effect parameters in linear mixed models. Accepted for publication in Communications in Statistics - Simulation and Computation. DOI: 10.1080/03610918.2013.849741

8. Schaarschmidt, F., Gerhard, D. and Vogel, C.: Simultaneous confidence intervals for comparisons of several multinomial samples. In preparation.

HortScience 44(1):188–195. 2009.

Analysis of Trials with Complex Treatment Structure Using Multiple Contrast Tests

Frank Schaarschmidt[1] and Lea Vaas
Institute of Biostatistics, Natural Sciences Faculty, Leibniz Universität Hannover, Herrenhäuser Str. 2, Hannover D-30419, Germany

Additional index words. augmented factorial, fixed-dose combination, multiple comparisons, simultaneous confidence intervals, linear model

Abstract. **Experiments with complex treatment structures are not uncommon in horticultural research. For example, in augmented factorial designs, one control treatment is added to a full factorial arrangement, or an experiment might be arranged as a two-factorial design with some groups omitted because they are practically not of interest. Several statistical procedures have been proposed to analyze such designs. Suitable linear models followed by F-tests provide only global inference for main effects and their interactions. Orthogonal contrasts are demanding to formulate and cannot always reflect all experimental questions underlying the design. Finally, simple mean comparisons following global F-tests do not control the overall error rate of the experiment in the strong sense. In this article, we show how multiple contrast tests can be used as a tool to address the experimental questions underlying complex designs while preserving the overall error rate of the conclusions. Using simultaneous confidence intervals allows for displaying the direction, magnitude, and relevance of the mean comparisons of interest. Along with application in statistical software, shown by two examples, we discuss the possibilities and limitations of the proposed approach.**

In agricultural and horticultural research, controlled experiments are set up to evaluate the effect of several treatments and their interactions on physiological or developmental variables. If the levels of a first factor of potential influence are combined with all levels of a second factor, the resulting experiment has a two-factorial, completely cross-classified treatment structure. However, often the experimental questions are manifold and, together with the background knowledge on the practical problem, lead to complex treatment structures. Augmented factorial designs are a common example, in which one or several control treatments are added to a completely crossclassified factorial design. Then, comparisons between the control treatment and the full factorial part are of interest as well as the analysis of the factors arranged in the full factorial part. Another reasonable setting arises when the effect of two factors and their interaction is investigated, but not all combinations between the levels of the two factors are of practical interest. Then, these treatments are reasonably omitted from the experiment, leading to a treatment structure that is crossclassified with a few missing cells.

Among others, Marini (2003) discussed different strategies for analyzing augmented factorial designs. He considered four different linear models followed by comparisons of particular treatments. First, a one-way model was performed followed by all pairwise comparisons of the seven treatment means using the least significant difference (LSD) procedure. However, even when protected by a preceding global F-test, decisions based on LSD do not control the familywise error in the strong sense (Hochberg and Tamhane, 1987). Second, a two-way model was performed with subsequent F-tests for the main effect of formulation, the main effect of the concentration, and their interaction. When this approach is used, least square means cannot be estimated properly. Third, a pseudo one-way model comprising all seven treatment groups was used to compute six orthogonal contrasts defining some hypotheses of interest. The orthogonal contrasts involved the comparison of the control group to the average of all other treatments, two contrasts for the comparisons of the three formulations, one contrast for the comparison of the two formulations, and two contrasts for certain interaction effects. However, this approach has severe restrictions; the number of hypotheses that can be formulated is restricted to *k-1* if *k* is the number of treatment groups (Dean and Voss, 1999; Marini, 2003; Petersen, 1994). Often, not all hypotheses of practical interest can be included in this type of analysis. Moreover, the problem of multiple comparisons is not taken into account. In a last approach, a mixed model was performed including the concentrations as a quantitative variable and the formulations as a qualitative variable, whereas the blocks were included as random effect.

Piepho et al. (2006) propose another strategy. The authors show how to develop linear models by insertion of an additional variable containing the value CON for observations in the control and the value TREAT for observations belonging to treatments in the complete factorial structure; thus, the new variable represents the fragmentation of the control group and the six treatments and one is able to write out a nested model representing the experimental layout correctly. Note that in this approach, the nesting of effects is not used to represent nested random effects in a hierarchical model. Rather, the nesting operator is used to define an appropriate design matrix for fixed effects. Based on this model, least square means can be estimated properly and an analysis of variance can be performed that takes all existing groups into account and provides F-tests for relevant hypotheses in the complex treatment structure. One disadvantage of this approach is that results of F-tests generally provide only global information about main effects and interaction effects, i.e., a significant result gives evidence for a difference in means among any of the considered treatments. Information about the location of the difference(s), the effect size, or comparisons of particular interest is not available from this approach.

If the interesting experimental questions can be expressed best as a set of comparisons among particular treatment means, a multiple comparison problem results. If testing an increasing number of hypotheses with the number of true hypotheses unknown, the probability of at least one wrong testing decision increases. If it is the aim of the statistical analysis to control the probability of at least one false rejection among all the tested null hypotheses, procedures are needed that control the familywise error (FWE). Computationally simple procedures like the Bonferroni or Scheffé adjustment (e.g., in Nelson, 1989) are suitable for any type of comparisons between means but are known to be conservative because they ignore the correlations among the comparisons. More advanced standard approaches to control the FWE are the tests according to Tukey (e.g., in Hochberg and Tamhane, 1987, citing Tukey, 1953) or Dunnett (1955). However, these do usually not reflect the experimental questions underlying designs with complex treatment structures. Tukey's procedure is appropriate for all pairwise comparisons and therefore is often considered as conservative when testing and adjusting for more hypotheses than are actually of interest. Dunnett's procedure performs comparisons of several treatments to a control. With complex treatment structures, usually more than these comparisons are of interest. In the recent years, multiple comparisons procedures have been made available (Bretz et al., 2001, 2002; Hothorn et al., 2008a; Westfall 1997; Westfall et al., 1999), which provide the feature of controlling the FWE for a certain user-defined set of comparisons formulated as multiple contrasts of the treatment means. The number of comparisons as well as the correlation among

Received for publication 4 June 2008. Accepted for publication 23 Oct. 2008.
We thank Prof. L.A. Hothorn, M. Hasler, and the referees for their helpful comments on earlier versions of this paper.
[1]To whom reprint requests should be addressed; e-mail schaarschmidt@biostat.uni-hannover.de.

the comparisons is taken into account specifically for each user-defined set of contrasts. Although these methods are computationally more demanding than well-known standard procedures, they are numerically available in various softwares.

If multiple hypotheses are tested without controlling the FWE rate at α, the probability of finding at least one of the considered differences to be significant when this difference is in fact zero can be markedly higher than α. For motivation of multiple comparison procedures, we present simulated FWE probabilities in Tables 1 and 2.

Assuming that no true effect is present in an experiment with three, five, or 10 groups, Table 1 shows the probabilities to find no significant difference and one, two, or more significant differences when all pairwise comparisons are performed among these treatments without multiplicity adjustment or protection by a global F-test. When there are only three treatments, resulting in three tests, one will conclude for at least one significant difference in ≈12 of 100 experiments. If 10 treatments are compared in such a way, ≈61 of 100 experiments will identify an effect that is not reproducible in a follow-up experiment because a relatively large difference in means simply occurred by chance.

Table 2 shows similar numbers for the situation that two, four, or nine independent two-sided *t* tests are performed. This is a situation close to that of testing two, four, or nine orthogonal (independent) contrasts with a common residual degree of freedom (df). With only two independent tests, the overall chance of a Type I error is 0.0975, i.e., $1 - (1 - 0.05)^2$, whereas when testing nine independent hypotheses, the odds are ≈1:2 to observe an effect that is not present in truth. When overall (familywise) Type I errors of this magnitude are not acceptable, LSD or single df contrasts should not be used, and multiple contrast procedures, as described in the following, are more appropriate.

In the following section, we introduce two example data sets and briefly review the well-known concept of multiple contrast tests. Subsequently, we show the application of multiple contrast tests to the two examples and give an interpretation of the results. Finally, we discuss the advantages and limitations of the proposed approach.

Material and Methods

Two example data sets. Marini (2003) describes the analysis of an experiment with an augmented factorial treatment structure. The effect of three formulations of gibberellic acid (f1, f2, f3), each in two different concentrations, 10 and 20 mg·L^{-1}, on fruit set of apple was investigated. Additional to this two-factorial structure, a control group (application of water) was included in the design. The composition of the treatments is summarized in Table 3.

The seven treatment groups were arranged in a randomized complete block design using six apple trees as blocks. As the response variable, the ratio of fruits (65 d after bloom) per 100 flower clusters was presented (Marini, 2003). In previous analyses, Marini (2003) and Piepho et al. (2006) proposed the following experimental questions; the comparison of all treatments with the untreated control aims to show that the experiment was sensitive to reveal marked effects on the fruit set. The comparison of increasing dosages pooled over the formulations aims to assess whether and until which concentration a dose effect is present. For the two-factorial part of the trial, interest was also in the main effects of the formulations and concentrations taking the possibility of an interaction into account. By the formulation of orthogonal contrasts, Marini (2003) addressed the question for interactions more explicitly; namely, whether the different formulations affect the increase of fruit set from dose 10 to dose 20.

As a second example, we consider a fixed-dose combination experiment originally published by Adeli and Varco (2002). The effects of potassium (K) management on cotton yield (kg·ha^{-1}) were investigated. The objective was "to determine potassium (K) fertilizer rate and placement effects on cotton lint yield" (Adeli and Varco, 2002). The experimental setup contained two different application methods for K: "broadcast" (Bc) and "banded" (Bn), each in four concentrations. The factorial arrangement included 0, 68, and 136 kg·ha^{-1} K broadcast in all possible combinations with 0, 34, and 68 kg·ha^{-1} K banded application. Two additional treatments, with 204 kg·ha^{-1} K broadcast with zero banded and 102 kg·ha^{-1} K banded with zero broadcast application, were included in the design. This treatment structure can be imagined as arisen from a complete two-factorial structure with those cells omitted that lead to inappropriately high total dosages of kg·ha^{-1} K. Table 4 summarizes the treatment combinations. The doses were chosen carefully so that the total amount of K is the same in four pairs of treatments (in Table 4, members of each pair are given the same symbol). The resulting 11 treatment groups were assumed to be arranged in a completely randomized design with replication number n = 12. The data used for the analysis are simulated based on the published summary statistics.

The analysis should give information about supposed beneficial influences of the different placement methods and an expected diverging effect of the two methods. Furthermore, the aim is to find treatment combinations resulting in superior cotton lint yield given that the total K fertilizer application is the same. Note that regression method, and more specifically response surface regression, is a viable option to analyze this example. Compared with the methods discussed subsequently, regression methods have the advantage of using fewer parameters to describe the data, but additionally rely on assumptions concerning the dose–response relationship. Response surface regression is particularly useful when the aim is to estimate the optimum combination of the two quantitative variables. For an introduction to response surface regression, see, for example, Montgomery (2005).

Simultaneous confidence intervals for user-defined multiple contrasts. In this section, we review the concept of multiple contrast tests as, for example, described in Bretz et al. (2001). A simple linear model to explain the observations Y_{ij} is:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

with Y_{ij} denoting the jth observation of the ith treatment group, with $i = 1, \ldots, k$, and

Table 3. Experimental layout of Example 1 with three gibberellic acid formulations (f1, f2, f3), each in two different concentrations (10 and 20 mg·L^{-1}), and a water control (H_2O).

	Formulation			
Concn	H_2O	f1	f2	f3
0	CON	—	—	—
10	—	f1(10)	f2(10)	f3(10)
20	—	f1(20)	f2(20)	f3(20)

Table 1. Probability to reject *x* out of *M* tested hypotheses when there is no difference among three, five, and 10 treatments and all pairwise comparisons are performed using multiple *t* tests (comparisons wise Type I error probability 5%) without adjustment for multiple testing or the protection of a global F-test.[z]

Number of treatments	Number of hypotheses *M*	Probability to reject H_0 for *x* hypotheses			
		x = 0	*x = 1*	*x = 2*	*x> = 3*
3	3	0.881	0.090	0.029	0.000
5	10	0.723	0.135	0.083	0.060
10	45	0.390	0.139	0.121	0.350

[z]The probabilities are estimated by 10,000 simulation runs for each setting.

Table 2. Probability to reject *x* out of *M* tested hypotheses when there is no true difference in two, four, and nine independent two-sided contrasts (comparisons wise Type I error probability 5%) without adjustment for multiple testing.[z]

Number of treatments	Number of hypotheses *M*	Probability to reject H_0 for *x* hypotheses			
		x = 0	*x = 1*	*x = 2*	*x> = 3*
3	2	0.903	0.094	0.003	—
5	4	0.819	0.164	0.016	0.002
10	9	0.644	0.284	0.063	0.009

[z]The probabilities are estimated by 10,000 simulation runs for each setting.

Table 4. Treatment structure of Example 2 (Adeli and Varco, 2002) comprising 11 combinations of broadcast (Bc) and banded (Bn) application of different doses of potassium (K) fertilizer.[z]

	Broadcast application			
Banded application	0 kg·ha^{-1} K	68 kg·ha^{-1} K	136 kg·ha^{-1} K	204 kg·ha^{-1} K
0 kg·ha^{-1} K	Bc0Bn0	Bc68Bn0 *	Bc136Bn0 #	Bc204Bn0 °
34 kg·ha^{-1} K	Bc0Bn34	Bc68Bn34 +	Bc136Bn34	—
68 kg·ha^{-1} K	Bc0Bn68 *	Bc68Bn68 #	Bc136Bn68 °	—
102 kg·ha^{-1} K	Bc0Bn102 +	—	—	—

[z]Treatments resulting in the same total amount of kg·ha^{-1} K are labeled with common symbols, * (68 kg·ha^{-1} K), + (102 kg·ha^{-1} K), # (136 kg·ha^{-1} K), and ° (204 kg·ha^{-1} K).

$$j = 1, \ldots, n_i$$

μ_i denoting the mean of the ith treatment group; and
ε_{ij} denoting the residual error for the jth observation in the ith group.

The errors ε_{ij} are assumed to be independent, i.e., the observations are derived from a completely randomized design and to be Gaussian distributed with equal variances: $\varepsilon_{ij} \sim N(0, \sigma^2)$. From fitting this one-way model, we derive estimates $\hat{\mu}_i$ for the treatment means and $\hat{\sigma}^2$ for the residual error. The questions of interest can then be stated as differences of the treatment means μ_i or, more generally, as contrasts of the treatment means μ_i. A contrast L is a weighted difference of the $\mu_1, \mu_2, \ldots, \mu_k$, where the weights c_i are chosen such that a certain difference of interest is built: $L = \Sigma\, c_i\mu_i = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k$. For example, in a design comprising four treatments, $i = 1, 2, 3, 4$, the difference between Treatments 1 and 2, $\mu_2 - \mu_1$ can be written as $\Sigma\, c_i\mu_i$ with $c_1 = -1$, $c_2 = 1$, $c_3 = 0$, $c_4 = 0$. However, also, more complicated differences, like for example the difference of the first treatment's mean to the average mean of the three remaining treatments, $(\mu_2 + \mu_3 + \mu_4)/3 - \mu_1$ can be formulated: $c_1 = -1$, $c_2 = 1/3$, $c_3 = 1/3$, $c_4 = 1/3$.

For the choice of the c_is, we impose only the restriction that the sum of c_is should be zero, $\Sigma c_i = 0$. This ensures that the contrast has expectation 0 if in fact all μ_i are equal. Moreover, we usually choose the c_i such that the sum of all negative coefficients is –1 and hence the sum of all positive coefficients is 1. Then, we can interpret the confidence intervals for the contrasts as differences of (weighted averages of) treatment means. Usually, several, say M, such contrasts are necessary to represent the experimental questions of interest. Note that there are no further restrictions on the choice of the c_i depending on the remaining set of contrasts. That is, there is no necessity to define the M contrasts orthogonal to each other, and there is no restriction on the number of contrasts M, as in the case of orthogonal single df contrasts. The test statistic for one contrast can be calculated from:

$$T = \frac{\sum_{i=1}^{k} c_i\hat{\mu}_i}{\hat{\sigma}\sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}}.$$

Simultaneous confidence intervals for the M contrasts can be calculated from

$$\left[\sum_{i=1}^{k} c_i\hat{\mu}_i \pm q_{1-\alpha,M,R}^{two-sided}\hat{\sigma}\sqrt{\sum_{i=1}^{k}\frac{c_i^2}{n_i}}\right],$$

using $q_{1-\alpha,M,R}^{two-sided}$ as the critical value calculated from the multivariate t-distribution with dimension M and correlation matrix R. For general contrasts, the correlation matrix R has a complicated structure with elements depending on the sample sizes n_i and the contrast coefficients c_i.

For a particular contrast, the null hypothesis that the difference defined by the contrast has the value zero can be rejected if $|T| > q_{1-\alpha,M,R}^{two-sided}$ or if the confidence interval does not contain the value zero. How such critical values can be obtained is described by Westfall et al. (1999) or Bretz et al. (2001). The presented confidence intervals can be obtained using the LSMESTIMATE statement in the SAS PROC GLIMMIX (SAS Institute, 2006) or the package multcomp (Bretz et al., 2002; Hothorn et al., 2008b) in the free statistical software R (R Development Core Team, 2008). In the remaining part of the article, R-2.6.2 is used with multcomp, Version 0.993-2.

We favor the graphical display of the simultaneous confidence intervals for reporting the results of a statistical analysis. From such plots, the significance of a particular difference at a FWE level α can be inferred if the value zero is not included in the confidence interval. Additionally, the direction (decrease or increase), magnitude, and, possibly, relevance of an effect can be assessed. If interpreting the relevance of the measured effect is of interest, confidence intervals are advantageous compared with P values because of displaying the effect size in the scale of the measured variable rather than in the scale of probability. Finally, the uncertainty concerning the estimated effect, depending on the sample variance and the sample size, is displayed by the width of the confidence interval.

Results

Evaluation of Example 1. From previous discussions of Example 1 (Marini, 2003; Piepho et al., 2006), the following experimental questions can be deduced: Is the experimental setting capable of revealing effects on the response? Do the formulations differ? Do the concentrations differ?

In the following, hypotheses that might be of interest are stated as differences of treatment means using the acronyms introduced in Table 3. The contrast coefficients (c_i) leading to the stated differences are summarized in Table 5. The first difference of interest compares the pooled means of all treatments versus the untreated control group:

$$1.\ \mu_{\text{all treatments}} - \mu_{\text{CON}} = 0.$$

For the detection of main effects of the formulation, three pairwise comparisons could be of interest, which pool over the two concentrations:

$$2.\ \mu_{\text{f2}} - \mu_{\text{f1}} = 0;$$

$$3.\ \mu_{\text{f3}} - \mu_{\text{f1}} = 0;$$

$$4.\ \mu_{\text{f3}} - \mu_{\text{f2}} = 0.$$

Analogously, the comparison of the two concentrations can be done by pooling over the formulations:

$$5.\ \mu_{(20)} - \mu_{(10)} = 0.$$

Interactions can be detected by building differences of differences (Petersen, 1994), i.e., comparing the difference between concentrations 10 and 20 between Formulations 1 and 2, and so on, as is done in Comparisons 6 through 8 in Table 5.

$$6.\ [\mu_{\text{f1}(20)} - \mu_{\text{f1}(10)}] - [\mu_{\text{f2}(20)} - \mu_{\text{f2}(10)}] = 0$$

$$7.\ [\mu_{\text{f1}(20)} - \mu_{\text{f1}(10)}] - [\mu_{\text{f3}(20)} - \mu_{\text{f3}(10)}] = 0$$

$$8.\ [\mu_{\text{f2}(20)} - \mu_{\text{f2}(10)}] - [\mu_{\text{f3}(20)} - \mu_{\text{f3}(10)}] = 0$$

Note that the stated contrasts reflect similar hypotheses as have been tested using four

Table 5. Contrast coefficients (c_i) are summarized for the multiple contrast tests indicated in the above text.[z]

		Treatment:						
		Control	f1(10)	f1(20)	f2(10)	f2(20)	f3(10)	f3(20)
Number	Comparison	Contrast coefficients:						
1	All treatments-CON	1	–1/6	–1/6	–1/6	–1/6	–1/6	–1/6
2	f2 – f1	0	–1/2	–1/2	1/2	1/2	0	0
3	f3 – f1	0	–1/2	–1/2	0	0	1/2	1/2
4	f3 – f2	0	0	0	–1/2	–1/2	1/2	1/2
5	(20) – (10)	0	–1/3	1/3	–1/3	1/3	–1/3	1/3
6	Interaction f1 versus f2	0	–1	1	1	–1	0	0
7	Interaction f1 versus f3	0	–1	1	0	0	1	–1
8	Interaction f2 versus f3	0	0	0	–1	1	1	–1

[z]Calculated are the comparison of control versus the pooled treatments (the first contrast), the comparisons of the formulations pooled over concentrations (Contrasts 2, 3, 4), the comparison of the concentrations pooled over formulations (Contrast 5), and contrasts for the interactions between formulation and concentration (Contrasts 6, 7, 8).

F-tests after a suitable model reformulation in the analysis by Piepho et al. (2006).

Simultaneous confidence intervals for the comparisons formally defined in Table 5 are plotted in Figure 1. The complete analysis has been performed in one simple procedure with all done tests adjusted for multiplicity inherently.

Altogether, the six gibberellin treatments lead to a significant increase in the number of fruits per flower cluster. Hence, the experimental setting is capable of revealing effects of the gibberellin treatments compared with an untreated control. On average, over the six gibberellin treatments, we can expect an increase of at least five fruits per 100 clusters more than in the untreated control with 95% confidence. Practically, the mean increase in the response is not very interesting; it is not possible to decide which treatments mainly contribute to the overall effect. Moreover, controlling the FWE for all eight hypotheses, none of the remaining tests is significant at the 5% level. None of the differences among the three formulations, each pooled over the concentrations, are significantly different from 0. Although comparing the average effect of the three formulations at concentration 20 with their average effect at concentration 10 shows a mean increase, the observed difference is not significant when controlling the overall Type I error probability at 5%. Finally, none of the three interaction contrasts differs significantly from zero, although the mean increase in the response when increasing the concentration from 10 to 20 is somewhat more pronounced in f1 and f2 compared with f3 (Comparisons 7 and 8). That is, given the limited sample size of the trial, we cannot conclude that there are differences among the formulations, between the concentrations, and cannot prove the presence of interactions when controlling the FWE at 5%.

The contrasts in Table 5 were constructed to show that hypotheses similar to those of the analysis of variance F-tests used by Piepho et al. (2006) can be tested using a multiple contrast approach. In practice, other comparisons can be more interesting and are as simple to implement. In the following, we show an analysis, alternative to that in Table 5.

First, it could be of interest whether any of the six gibberellin treatments leads to a change in the number of fruit limbs per number of flowers. Hypotheses 1 to 6 represent these comparisons; the resulting contrast coefficients are presented in Table 6.

$$1.\ \mu_{f1(10)} - \mu_{CON} = 0;$$

$$2.\ \mu_{f1(20)} - \mu_{CON} = 0;$$

$$3.\ \mu_{f2(10)} - \mu_{CON} = 0;$$

$$4.\ \mu_{f2(20)} - \mu_{CON} = 0;$$

$$5.\ \mu_{f3(10)} - \mu_{CON} = 0;$$

$$6.\ \mu_{f3(20)} - \mu_{CON} = 0.$$

Second, it might be of interest whether the formulations differ, taking the possibility of an interaction into account. This may result in comparisons of the different formulations at each of the two concentration levels:

$$7.\ \mu_{f2(10)} - \mu_{f1(10)} = 0;$$

$$8.\ \mu_{f3(10)} - \mu_{f1(10)} = 0;$$

$$9.\ \mu_{f3(10)} - \mu_{f2(10)} = 0;$$

$$10.\ \mu_{f2(20)} - \mu_{f1(20)} = 0;$$

$$11.\ \mu_{f3(20)} - \mu_{f1(20)} = 0;$$

$$12.\ \mu_{f3(20)} - \mu_{f2(20)} = 0.$$

Finally, the concentrations 10 and 20 could be compared separately for each formulation:

$$13.\ \mu_{f1(20)} - \mu_{f1(10)} = 0;$$

$$14.\ \mu_{f2(20)} - \mu_{f2(10)} = 0;$$

$$15.\ \mu_{f3(20)} - \mu_{f3(10)} = 0.$$

These are 15 comparisons in total, which could not have been performed using orthogonal contrasts. The correlation structure among these 15 comparisons is not trivial; however, it is taken into account inherently by the statistical software.

Figure 2 shows simultaneous 95% confidence intervals for the contrasts defined in Table 6.

This evaluation results in a more informative interpretation than the first approach: Two of the six gibberellin treatments result in a significant increase of the number of fruits per 100 flower clusters. With 95% confidence, we can expect an increase in the mean number of fruits per 100 flower clusters of at least two when Formulation 1 with Concentration 20 is applied. Using Formulation 2 with Concentration 20, one can expect at least 18 fruits per 100 clusters more than in the untreated control. The remaining combinations of formulation and concentration led to a mean increase of the number of fruits per cluster, but, controlling the FWE for all comparisons, the observed differences are not significant. Furthermore, the pairwise differences among the formulations are not significant when considered separately for each concentration (Comparisons 7 through 12). Finally, a difference between Concentrations 10 and 20 cannot be shown at the 5% level for any of the three formulations.

Evaluation of Example 2. The experiment presented by Adeli and Varco (2002) shows a more complex treatment structure. First, it could be of interest which K rate or application method increases the yield compared with control. That is, the differences of the 10 different K treatments to the untreated control treatment Bc0Bn0 are of primary practical interest. Furthermore, interest is in the magnitude of increase that can at least be expected with high probability, i.e., in lower confidence limits. Hypotheses 1 to 3 compare the three treatments with only broadcast application with the untreated control group:

$$1.\ \mu_{Bc68Bn0} - \mu_{Bc0Bn0} = 0;$$

$$2.\ \mu_{Bc136Bn0} - \mu_{Bc0Bn0} = 0;$$

$$3.\ \mu_{Bc204Bn0} - \mu_{Bc0Bn0} = 0.$$

Hypotheses 4, 5, and 6 compare the treatments with only banded application with the untreated control group:

$$4.\ \mu_{Bc0Bn34} - \mu_{Bc0Bn0} = 0;$$

$$5.\ \mu_{Bc0Bn68} - \mu_{Bc0Bn0} = 0;$$

$$6.\ \mu_{Bc0Bn102} - \mu_{Bc0Bn0} = 0.$$

Hypotheses 7 to 10 compare the treatments with mixed application of banded and broadcast application with the untreated control:

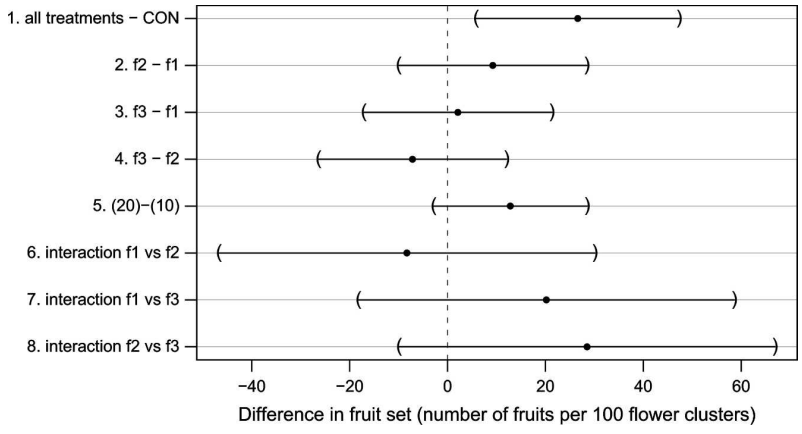


Fig. 1. Simultaneous 95% confidence intervals for the eight contrasts formulated in Table 5. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are the comparison of control versus the pooled effect of all six treatments (the first interval), the comparison of the formulations with pooled concentrations (Contrasts 2, 3, 4), the comparison of the concentration with pooled formulations (Contrast 5), and the interactions between the formulations (Contrasts 6, 7, 8).

Table 6. Contrast coefficients (c_i) of 15 contrasts for the alternative evaluation of Example 1.[z]

		Treatment						
		Control	f1(10)	f1(20)	f2(10)	f2(20)	f3(10)	f3(20)
Number	Comparison	Contrast coefficients						
1	f1(10) – CON	–1	1	0	0	0	0	0
2	f1(20) – CON	–1	0	1	0	0	0	0
3	f2(10) – CON	–1	0	0	1	0	0	0
4	f2(20) – CON	–1	0	0	0	1	0	0
5	f3(10) – CON	–1	0	0	0	0	1	0
6	f3(20) – CON	–1	0	0	0	0	0	1
7	f2(10) – f1(10)	0	–1	0	1	0	0	0
8	f3(10) – f1(10)	0	–1	0	0	0	1	0
9	f3(10) – f2(10)	0	0	0	–1	0	1	0
10	f2(20) – f1(20)	0	0	–1	0	1	0	0
11	f3(20) – f1(20)	0	0	–1	0	0	0	1
12	f3(20) – f2(20)	0	0	0	0	–1	0	1
13	f1(20) – f1(10)	0	–1	1	0	0	0	0
14	f2(20) – f2(10)	0	0	0	–1	1	0	0
15	f3(20) – f3(10)	0	0	0	0	0	–1	1

[z]Defined are the comparisons of the six gibberellin treatments versus the control (Contrasts 1–6), the comparisons of the formulations separate for each concentration (Contrasts 7–12), and the comparisons of Concentrations 10 and 20 separate for each formulation.

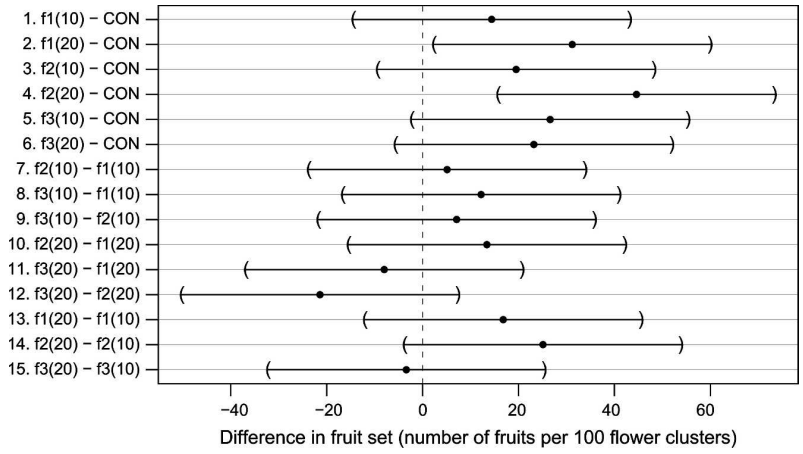


Fig. 2. Simultaneous 95% confidence intervals for the 15 contrasts defined in Table 6. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are differences of the single treatment means to the mean of the control (Contrasts 1–6), the comparisons of the formulations separate for each concentration (Contrasts 7–12), and the comparisons between Concentrations 10 and 20 separate for each formulation (Contrasts 13–15).

$$7.\ \mu_{Bc68Bn34} - \mu_{Bc0Bn0} = 0;$$

$$8.\ \mu_{Bc68Bn68} - \mu_{Bc0Bn0} = 0;$$

$$9.\ \mu_{Bc136Bn34} - \mu_{Bc0Bn0} = 0;$$

$$10.\ \mu_{Bc136Bn68} - \mu_{Bc0Bn0} = 0.$$

Furthermore, the aim could be to analyze whether any treatment that combines broadcast with banded application leads to an increase in yield compared with treatments resulting in the same amount of K with only one application method, broadcast or banded. The corresponding comparisons are:

$$11.\ \mu_{Bc68Bn34} - \mu_{Bc0Bn102} = 0;$$

$$12.\ \mu_{Bc68Bn68} - \mu_{Bc136Bn0} = 0;$$

$$13.\ \mu_{Bc136Bn68} - \mu_{Bc204Bn0} = 0.$$

Finally, it could be of interest whether banded application of 68 kg·ha^{-1} K leads to an increase in yield compared with broadcast application of the same amount:

$$14.\ \mu_{Bc0Bn68} - \mu_{Bc68Bn0} = 0.$$

The contrast coefficients resulting from these 14 contrasts are presented in Table 7.

Simultaneous confidence intervals for the contrasts defined in Table 7 are plotted in Figure 3.

With 95% confidence, one can state that broadcast application of K leads to a significant increase in yield when applied with 136 kg·ha^{-1} or 204 kg·ha^{-1}. For banded application alone, no significant effect can be found. All combinations of banded and broadcast applications lead to a significant increase of cotton yield compared with the untreated control. Applying the combinations Bc68Bn34, BC68Bn68, and Bc136Bn34 leads to a mean increase in yield compared with the untreated control of at least 30 kg·ha^{-1}, 55 kg·ha^{-1}, and 180 kg·ha^{-1}, respectively. None of the treatments combining broadcast and banded application leads to a significant increase in yield compared with treatments applying the same amount of K with only one application method. Finally, there is no significant difference in yield between banded and broadcast application of 68 kg·ha^{-1} K.

Discussion

This article shows that simultaneous confidence intervals for multiple contrasts are a flexible method to evaluate factorial experiments with nonstandard treatment structures as, for example, augmented factorial designs or experiments with two or more factors, which are crossclassified with some factor combinations omitted. The strategy can be summarized as follows; estimators for the treatment means and variance are derived from a simple general linear model with all treatments combined in a single factor (i.e., a pseudo-one-way layout or cell means model). Contrast coefficients are chosen by the user such that the hypotheses of interest are reflected as differences of (weighted averages of) treatment means. Like other procedures following the general linear model (Marini, 2003; Piepho et al., 2006), the described procedure relies on the assumptions that the observations are mutually independent and continuous with normal distributed errors and homogeneous variances. The method is computationally available for the R environment for statistical computing as well as in SAS.

Compared with other methods that have been proposed for evaluation of experiments with complex treatment structures, the described method has a number of advantages. The individual contrasts give more specific information than the global decisions provided by an analysis of variance F-test. When simultaneous confidence intervals are used, the significance, relevance, and direction (increase or decrease) of the effect of interest as well as the uncertainty concerning the estimates can be interpreted in a scale close to that of the measured variable, which is often easier than interpreting *P* values in the scale of probability. Compared with orthogonal single df contrasts, the contrasts formulated for the method described in this article do not need to be mutually orthogonal and are not restricted in their number. Finally, the overall Type I error probability is controlled inherently for a user-defined set of contrasts.

The described method is of limited use if experiments are analyzed that comprise many, say more than 20, treatments. Being still methodologically correct, it then has the drawback that the contrast matrix becomes huge and it is hard to control for typos in the definition of comparisons of interest. Also, when the number of contrasts becomes very high, computations can be very time-consuming or impossible. For some scenarios, the methods described by Piepho et al. (2006) can then be more appropriate.

In this article, we discuss only complex treatment structures and for brevity assume a simple randomization structure

Table 7. Contrast coefficients (c_i) for the 14 comparisons among the 11 treatments of the cotton example.[z]

	K application	Treatment combinations										
	Bc	0	68	136	204	0	0	0	68	68	136	136
	Bn	0	0	0	0	34	68	102	34	68	34	68
	Total K	0	68	136	204	34	68	102	102	136	170	204
Number	Comparison	Contrast coefficients										
1	Bc68Bn0 – Bc0Bn0	–1	1	0	0	0	0	0	0	0	0	0
2	Bc136Bn0 – Bc0Bn0	–1	0	1	0	0	0	0	0	0	0	0
3	Bc204Bn0 – Bc0Bn0	–1	0	0	1	0	0	0	0	0	0	0
4	Bc0Bn34 – Bc0Bn0	–1	0	0	0	1	0	0	0	0	0	0
5	Bc0Bn68 – Bc0Bn0	–1	0	0	0	0	1	0	0	0	0	0
6	Bc0Bn102 – Bc0Bn0	–1	0	0	0	0	0	1	0	0	0	0
7	Bc68Bn34 – Bc0Bn0	–1	0	0	0	0	0	0	1	0	0	0
8	Bc68Bn68 – Bc0Bn0	–1	0	0	0	0	0	0	0	1	0	0
9	Bc136Bn34 – Bc0Bn0	–1	0	0	0	0	0	0	0	0	1	0
10	Bc136Bn68 – Bc0Bn0	–1	0	0	0	0	0	0	0	0	0	1
11	Bc68Bn34 – Bc0Bn102	0	0	0	0	0	0	–1	1	0	0	0
12	Bc68Bn68 – Bc136Bn0	0	0	–1	0	0	0	0	0	1	0	0
13	Bc136Bn68 – Bc204Bn0	0	0	0	–1	0	0	0	0	0	0	1
14	Bc0Bn68 – Bc68Bn0	0	–1	0	0	0	1	0	0	0	0	0

[z]Calculated are the comparisons of the 10 different potassium (K) treatments to the untreated the control (Contrasts 1–10) and the comparisons of groups with equal amount of K but different application methods (Contrasts 11–14).

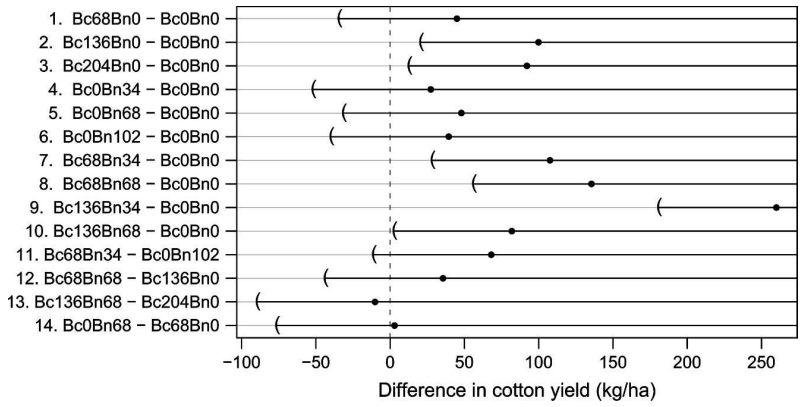


Fig. 3. Simultaneous 95% lower confidence limits for the 14 contrasts defined in Table 7. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are the comparison of the control versus each of the single fertilizer combinations (1–10) and the comparisons of groups with equal amount of potassium but different application methods (11–14).

and homoscedastic Gaussian error distribution for the response variable. Nevertheless, the concept of multiple contrast tests can be extended so that situations with different assumptions or randomization schemes are also covered. Block effects or more complex randomization structures may be included as random effects in a linear mixed effects model, whereas the complex treatment structure remains in the fixed part (for example Piepho et al., 2003). Computationally, approximate simultaneous confidence intervals for multiple contrasts in mixed models are covered in the SAS PROC GLIMMIX as well as the R package multcomp. When the assumption of the Gaussian distribution is not adequate but counts or proportions are considered, generalized linear models (McCullagh and Nelder, 1989; Piepho, 1999) are an alternative. By default, the primary comparisons are then performed on the log scale for count data and on the logit scale for binomial proportions. Using the inverse link function to transform back results in confidence intervals for ratios of means and odds ratios when the log and logit link is used, respectively. Again, these cases are computationally solved in PROC GLIMMIX and multcomp. However, also in the general linear model with Gaussian errors, the comparisons of interest could be formulated in terms of ratios rather than in differences of means (Dilba et al., 2006). When interest is in a combination of one-sided and two-sided hypotheses, Braat et al. (2008) provide a method related to the methods shown in this article.

Literature Cited

Adeli, A. and J.J. Varco. 2002. Potassium management effects on cotton yield, nutrition, and soil potassium level. J. Plant Nutr. 25:2229–2242.

Braat, S., D. Gerhard, and L.A. Hothorn. 2008. Joint one-sided and two-sided simultaneous confidence intervals. J. Biopharm. Stat. 18:293–306.

Bretz, F., A. Genz, and L.A. Hothorn. 2001. On the numerical availability of multiple comparison procedures. Biom. J. 43:645–656.

Bretz, F., T. Hothorn, and P.H. Westfall. 2002. On multiple comparisons in R. R News 2:14–17.

Dean, A. and D. Voss. 1999. Design and analysis of experiments. Springer-Verlag, New York, NY.

Dilba, G., F. Bretz, and V. Guiard. 2006. Simultaneous confidence sets and confidence intervals for multiple ratios. J. Stat. Plan. Infer. 136:2640–2658.

Dunnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. J. Amer. Stat. Assoc. 50:1096–1121.

Hochberg, A.C. and Y. Tamhane. 1987. Multiple comparison procedures. Wiley, New York, NY.

Hothorn, T., F. Bretz, and P. Westfall. 2008a. Simultaneous inference in general parametric models. Biometrical Journal 50:346–363.

Hothorn, T., F. Bretz, P. Westfall, and R.M. Heiberger. 2008b. Multcomp: Simultaneous inference for general linear hypotheses. R package version 0.993-2.

Marini, R.P. 2003. Approaches to analyzing experiments with factorial arrangements of treatments plus other treatments. HortScience 38:117–120.

McCullagh, P. and J.A. Nelder. 1989. Generalized linear models. Chapman & Hall/CRC, Boca Raton, FL.

Montgomery, D.C. 2005. Design and analysis of experiments. 6th Ed. Wiley, Hoboken, NJ.

Nelson, P.R. 1989. Multiple comparisons of means using simultaneous confidence intervals. J. Qual. Technol. 21:232–289.

Petersen, R.G. 1994. Agricultural field experiments—Design and analysis. Marcel Dekker, New York, NY.

Piepho, H.-P. 1999. Analysing disease incidence data from designed experiments by generalized linear mixed models. Plant Pathol. 48:668–674.

Piepho, H.-P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. J. Agron. Crop Sci. 189:310–322.

Piepho, H.-P., E.R. Williams, and M. Fleck. 2006. A note on the analysis of designed experiments with complex treatment structure. HortScience 41:446–452.

Hochberg, A.C. and Y. Tamhane. 1987. Multiple comparison procedures. Wiley, New York, NY.

Hothorn, T., F. Bretz, and P. Westfall. 2008a. Simultaneous inference in general parametric models. Biometrical Journal 50:346–363.

Hothorn, T., F. Bretz, P. Westfall, and R.M. Heiberger. 2008b. Multcomp: Simultaneous inference for general linear hypotheses. R package version 0.993-2.

R Development Core Team. 2008. R: A language and environment for statistical computing. Version 2.6.2. R Foundation for Statistical Computing. Vienna, Austria.

SAS Institute. 2006. The GLIMMIX procedure, June 2006. SAS Institute, Cary, NC.

Tukey, J. 1953. The problem of multiple comparisons, unpublished manuscript, reprinted in: Braun, H.I. (Ed.) 1994. The collected works of John W. Tukey. VIII. Multiple comparisons. Chapman and Hall, New York, NY.

Westfall, P.H. 1997. Multiple testing of general contrasts using logical constraints and correlations. J. Amer. Stat. Assoc. 92:299–306.

Westfall, P.H., R.D. Tobias, D. Rom, R.D. Wolfinger, and Y. Hochberg. 1999. Multiple comparisons and multiple tests using the SAS System. SAS Institute, Cary, NC.

Appendix

The two example data sets are available at http://www.biostat.uni-hannover.de/software/. After loading the data sets into the R workspace under the names ExFruitset (Example 1) and ExKCotton (Example 2), the following R code reproduces the analyses of the two examples shown in this article.

SAS program files, including the data sets and the calculation of the simultaneous intervals, are available at http://www.biostat.uni-hannover.de/software/.

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 1:

```
# Define Block as a factor variable and fit the linear model
ExFruitset$Block <- as.factor(ExFruitset$Block)
model1 <- lm(Fruitset ~ Block + Treatment,data = ExFruitset)
# Load library multcomp
library(multcomp)
# Define the contrast matrix in Table 5
contrast1 <- rbind(
"1. all treatments - CON"=c(-1,1/6,1/6,1/6,1/6,1/6,1/6),
"2. f2 - f1"=          c(0,-1/2,-1/2,1/2,1/2,0,0),
"3. f3 - f1"=          c(0,-1/2,-1/2,0,0,1/2,1/2),
"4. f3 - f2"=          c(0,0,0,-1/2,-1/2,1/2,1/2),
"5. (20)-(10)"=   c(0,-1/3,1/3,-1/3,1/3,-1/3,1/3),
"6. interaction f1 vs f2"=c(0,-1, 1, 1,-1, 0, 0),
"7. interaction f1 vs f3"=c(0,-1, 1, 0, 0, 1,-1),
"8. interaction f2 vs f3"=c(0, 0, 0,-1, 1, 1,-1))
# Multiple comparisons and simultaneous confidence intervals
comps1 <- glht(model1, linfct = mcp(Treatment = contrast1),
 alternative="two.sided")
SCI1<-confint(comps1)
# create the plot in Fig. 1
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI1, main="", xlab="Difference in fruit set
(number of fruits per 100 flower clusters)")
```

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 2:

```
# Define the contrasts in Table 6
contrasts2 <- rbind (
"1. f1(10) - CON" =    c(-1, 1, 0, 0, 0, 0, 0),
"2. f1(20) - CON" =    c(-1, 0, 1, 0, 0, 0, 0),
"3. f2(10) - CON" =    c(-1, 0, 0, 1, 0, 0, 0),
"4. f2(20) - CON" =    c(-1, 0, 0, 0, 1, 0, 0),
"5. f3(10) - CON" =    c(-1, 0, 0, 0, 0, 1, 0),
"6. f3(20) - CON" =    c(-1, 0, 0, 0, 0, 0, 1),
"7. f2(10) - f1(10)"=  c(0,-1, 0, 1, 0, 0, 0),
"8. f3(10) - f1(10)"=  c(0,-1, 0, 0, 0, 1, 0),
"9. f3(10) - f2(10)"=  c(0, 0, 0,-1, 0, 1, 0),
"10. f2(20) - f1(20)"= c(0, 0,-1, 0, 1, 0, 0),
"11. f3(20) - f1(20)"= c(0, 0,-1, 0, 0, 0, 1),
"12. f3(20) - f2(20)"= c(0, 0, 0, 0,-1, 0, 1),
"13. f1(20) - f1(10)"= c(0, -1,1, 0, 0, 0, 0),
"14. f2(20) - f2(10)"= c(0, 0, 0, -1, 1, 0, 0),
"15. f3(20) - f3(10)"= c(0, 0, 0, 0, 0, -1, 1))
# Multiple comparisons and simultaneous confidence intervals
comps2 <- glht(model1, linfct = mcp(Treatment = contrasts2),
 alternative="two.sided")
SCI2<-confint(comps2)
# Create the plot in Fig. 2
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI2, main="", xlab="Difference in fruit set
(number of fruits per 100 flower clusters)")
```

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 3:

```
# Reorder the factor levels in the data set and fit the linear model
ExKCotton$KFertilizer <- factor(ExKCotton$KFertilizer,
levels=c("BC0BN0","BC68BN0","BC136BN0","BC204BN0","BC0BN34",
"BC0BN68","BC0BN102","BC68BN34","BC68BN68","BC136BN34","BC136BN68"))
model2 <- lm(Yield~KFertilizer, data=ExKCotton)
# Define the contrasts in Table 7
contrasts3 <- rbind(
"1.  Bc68Bn0 - Bc0Bn0"    = c(-1,1,0,0,0,0,0,0,0,0),
"2.  Bc136Bn0 - Bc0Bn0"   = c(-1,0,1,0,0,0,0,0,0,0),
"3.  Bc204Bn0 - Bc0Bn0"   = c(-1,0,0,1,0,0,0,0,0,0),
"4.  Bc0Bn34 - Bc0Bn0"    = c(-1,0,0,0,1,0,0,0,0,0),
"5.  Bc0Bn68 - Bc0Bn0"    = c(-1,0,0,0,0,1,0,0,0,0),
"6.  Bc0Bn102 - Bc0Bn0"   = c(-1,0,0,0,0,0,1,0,0,0),
"7.  Bc68Bn34 - Bc0Bn0"   = c(-1,0,0,0,0,0,0,1,0,0),
"8.  Bc68Bn68 - Bc0Bn0"   = c(-1,0,0,0,0,0,0,0,1,0),
"9.  Bc136Bn34 - Bc0Bn0"  = c(-1,0,0,0,0,0,0,0,0,1),
"10. Bc136Bn68 - Bc0Bn0"  = c(-1,0,0,0,0,0,0,0,0,1),
"11. Bc68Bn34 - Bc0Bn102" = c(0,0,0,0,0,0,-1,1,0,0),
"12. Bc68Bn68 - Bc136Bn0" = c(0,0,-1,0,0,0,0,0,1,0),
"13. Bc136Bn68 - Bc204Bn0"= c(0,0,0,-1,0,0,0,0,0,1),
"14. Bc0Bn68 - Bc68Bn0"   = c(0,-1,0,0,0,1,0,0,0,0))
# Multiple comparisons and simultaneous confidence intervals
comps3 <- glht(model2, linfct = mcp(KFertilizer = contrasts3),
 alternative="greater")
SCI3<-confint(comps3)
# Create the plot in Fig. 3
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI3, main="", xlab="Difference in cotton yield (kg/ha)")
```

Journal of Agronomy and Crop Science

J Agro Crop Sci (2015) ISSN 0931-2250

MISCELLANEOUS

Analysis of Statistical Interactions in Factorial Experiments

A. Kitsche & F. Schaarschmidt

Institut für Biostatistik, Leibniz Universität Hannover, Hannover, Germany

Keywords
adjusted P-values; analysis of variance; interaction effect; simultaneous confidence intervals

Correspondence
A. Kitsche
Institut für Biostatistik
Leibniz Universität Hannover
Herrenhäuser Straße 3
30419 Hannover, Germany
Tel.: +49 511 762 3686
Fax: +49 511 762 4966
Email: kitsche@biostat.uni-hannover.de

Accepted April 16, 2014

doi:10.1111/jac.12076

Abstract

Two or higher-order factorial designs are very common in agricultural and horticultural experiments. The evaluation of such trials by analysis of variance (ANOVA) and the corresponding *F*-tests for the interaction effects covers only a global decision concerning the presence of interactions. This study presents a straightforward method, which provides a more detailed analysis of interactions via multiple contrast tests. The presented approach takes both the structure of each factor and the research question into account by building user-defined product-type contrasts. Simultaneous inference for these user-specified interaction contrasts that controls the overall error rate is available. In addition to adjusted P-values, it is recommended to use simultaneous confidence intervals to present the magnitude, direction and the biological relevance of the interaction effects. The proposed method is demonstrated using two horticultural trials. Furthermore, the authors provide a collection of worked examples using the R (A Language and Environment for Statistical Computing, 2013, R Foundation for Statistical Computing, Vienna, Austria) add-on package statint stored on github (https://github.com/AKitsche/statint).

Introduction

Experiments that include two or more treatment factors are frequently set up in agricultural and horticultural research. Such factorial trials permit the experimenter to simultaneously investigate several factors within the same experiment. Some commonly used factors in applied biology are as follows: i) particular varieties or cultivars of a species, ii) different kinds of fertilizers, iii) different concentrations of a fertilizer, iv) different irrigation intensities or v) different seed spacings in a row. As opposed to single-factor experiments, two and higher-order factorial experiments are appropriate to investigate the interaction effects beside the main effects: How does the effect of one factor change, if a second factor is varied? In many two and higher-order factorial experiments, the research hypothesis is at least partially formulated to test for interactions. In some experiments, the detailed investigation of interactions is of primary interest. As examples, see Slauenwhite and Qaderi (2013) who investigated the interactive effects of temperature and light quality on four canola cultivars, or Sahin et al. (2012), who examined the chloride and bromate interaction on oxidative stress in carrot plants.

Using the analysis of variance (ANOVA) and the corresponding *F*-tests for the main and interaction effects covers only an overall decision concerning the presence of an interaction. If a significant interaction is found in the ANOVA, subsequent tests for simple effects are recommended by several statistical textbooks with focus on plant, crop and agricultural science (Gomez and Gomez 1984, Clewer and Scarisbrick 2001 and Hoshmand 2006). That is, the differences in the means of one factor are tested separately on each level of the other factor. This method of post hoc investigation has been criticized in the statistical literature because it does not correspond to the null hypothesis tested by the global *F*-test of interaction in the ANOVA (Marascuilo and Levin 1970). Its focus is on inference for one primary factor, separately for the levels of a secondary factor. Such a distinction between the two factors may not exist, and, more severely, the interactions can be interpreted only indirectly from differing test results among the different levels of the secondary factor. Thus, erroneous conclusions concerning the detailed interaction effects are not directly

controlled via the significance levels of this *post hoc* procedure. Alternatively, graphical methods such as plots of the treatment means or interaction residuals are used for a further descriptive interpretation of the interaction (Harwell 1998), however, lacking any error control for the detailed interpretation of the interaction effects. Therefore, we do not recommend these approaches, particularly if there is no distinction between main effects of primary and secondary interest, and for experiments, where the presence of interactions is not considered as a nuisance in interpreting the main effects, but there is explicit interest in inference on the locations, the directions and the magnitude of interactions.

This study presents a general methodology to construct comparisons of treatment means for an in-depth analysis of interactions. The hypotheses are defined in terms of differences between differences and are related to the hypothesis of the ANOVA *F*-test for interaction. Because in such a detailed analysis several hypotheses are tested, an adequate multiple comparison procedure has to be used. Adjusted P-values for the individual hypotheses are provided, such that the significance of the detailed interpretations can be inferred, while controlling the overall probability of an erroneous decision. Additionally to a statement on rejection/acceptance of the corresponding null hypothesis, the related simultaneous confidence intervals allow interpretations whether the magnitude of the observed interaction effects is biologically relevant. In recent years, several authors applied these methods, see, for example, Hothorn (2003), Hothorn and Bleiholder (2006), Frömke and Bretz (2004), Schaarschmidt and Vaas (2009) and Menke et al. (2011). Nevertheless, there are several reasons why this procedure has hardly been used in practice so far: i) a general description of the statistical methodology to obtain appropriate quantiles for user-defined contrasts was provided not so long ago (Westfall and Young 1993 and Genz and Bretz 1999), ii) the numerical availability in standard statistical software like SAS (SAS Institute Inc. 2013) (LSMESTIMATE statement in PROC GLIMMIX available since 2009 in SAS/STAT 9.2) and R (R Core Team 2013) (add-on package multcomp (Hothorn et al. 2008) first version published on CRAN in 2002) was provided in recent times, iii) the construction and labelling of the interaction contrast matrices can be very cumbersome, and iv) no case studies are available that convincingly present the proposed method. To fill this gap, this manuscript provides a general description of the methodology and its application to two illustrative case studies.

This study is organized as follows: in Section 2, two illustrative examples are introduced. The statistical methodology is presented in Section 3. Subsequently, we demonstrate the application of the proposed method using two examples. Section 5 provides a concluding discussion and hints to straightforward extensions.

Illustrative Examples

Bushy and tall bush bean varieties with different row spacing

The first example was published by Petersen (1985, p. 155). The goal of the experiment was to investigate the effect of row spacing on the yield of different varieties of bush beans. Due to the different growth habits of the considered varieties, it was assumed that the spacing effect differs between the varieties. The selected four varieties differ such that 'NewEra' and 'BigGreen' form low, bushy plants and the two varieties 'LittleGem' and 'RedLake' form erect plants with few branches. The chosen row spacings were of 20, 40 and 60 cm between rows. A randomized complete block design with four blocks and 12 plots per block was used. The yield of dried beans in kilograms per plot was determined after harvest time. Figure 1 displays the mean yield for each variety-by-spacing combination. It is obvious that the mean yield increases for the varieties 'NewEra' and 'BigGreen', which form little, bushy plants, with increasing row spacing. On the other hand, the mean yield decreases for the two varieties 'LittleGem' and 'RedLake', which form erect plants, as the spacing increases. The results of the corresponding ANOVA reveal that this interaction between variety and spacing is highly significant (Table 1). The significant overall interaction may now be further analysed: What is the difference in yield increase for different spacings between the bushy and tall group averages? To what extent do the varieties with similar growth type differ in their reaction to spacing?

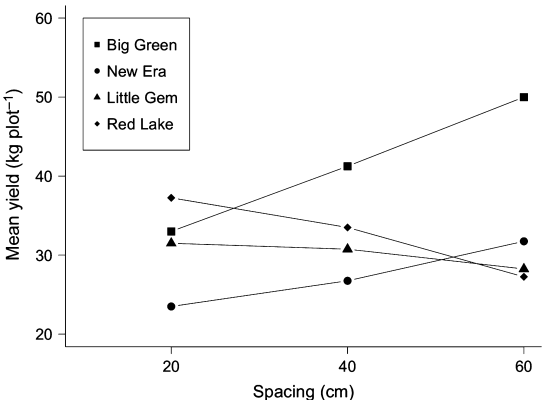


Fig. 1 Interaction plot of cell means which illustrates the relationship between row spacing and yield of four bush bean varieties that form either little, bushy plants (New Era and Big Green) or erect plants with few branches (Little Gem and Red Lake).

Table 1 Two-way ANOVA table of the bush beans data set

	d.f.	Sum Sq	Mean Sq	*F* value	Pr(>*F*)
Block	3	341.90	113.97	8.78	<0.001
Variety	3	1332.56	444.19	34.22	<0.001
Spacing	2	72.67	36.33	2.80	0.075
Variety:Spacing	6	871.00	145.17	11.18	<0.001
Residuals	33	428.35	12.98		

Effect of 1-Methylcyclopropene on *Pelargonium zonale*

The second data set was part of an experiment conducted at the Section of Floriculture within the Institute of Horticultural Production Systems at the Leibniz Universität Hannover in 2010 (data provided by Syariful Mubarok). The goal of the experiment was to investigate the characteristics of the ethylene receptor blocker 1-methylcyclopropene (MCP) for improvement in post-harvest characteristics of *Pelargonium zonale* hybrids. To examine the effect of the ethylene, one group with ethylene treatment and one group without ethylene treatment were considered. To study the effect of the ethylene blocker MCP, three different commercial products using the active agent MCP as ethylene blocker were used. Additionally, a water-sprayed control group was used as reference group. Furthermore, an untreated control group was used to rule out some effect of spray application. Within this experiment, three different cultivars of *Pelargonium zonale* were investigated. The trial was planned in a completely cross-classified treatment structure, laid out as a completely randomized design with three replications for each factorial combination. The primary response to determine the post-harvest characteristics was the chlorophyll content (mg per g fresh weight) after 8 days of treatment.

Figure 2 displays the box plots for the chlorophyll content after 8 days based on each ethylene-by-treatment-by-cultivar combination. From Fig. 2, a remarkable ethylene effect on the chlorophyll content is obvious in the control groups over all cultivars. This ethylene effect is reduced for those treatment groups whose products include the ethylene blocker MCP, whereas this reduction in the ethylene effect is considerably larger for cultivar 3 in contrast to cultivar 1 and cultivar 2. The corresponding three-way ANOVA (Table 2) shows a significant interaction between the factors ethylene, treatment and cultivar. This indicates a different response of the ethylene effect through the treatment groups between the three cultivars under investigation.

In particular, the researcher was interested in investigating a potentially different ethylene response between the control groups and the treatment groups including the ethylene blocker. Furthermore, interest was in the comparison of the ethylene effect between the three products including the ethylene blocker. Note that the research interest was not on a potentially different response of the ethylene effect caused by the different MCP blocking treatments between the three cultivars under investigation. However, to present the flexibility of the proposed approach, the authors provide a detailed analysis of this three-way interaction. Moreover, this analysis is in line with the significant three-way interaction term in Table 2.

Methods

The model

For the sake of simplicity, we assume a completely randomized design with two factors, afterwards denoted as A and B, and their interaction AB. Nevertheless, the presented approach can be extended to designs with more than two factors (see example 2) or with a more complex randomization structure (see section: Interaction contrasts for fixed effects in mixed models). The primary response is a continuous and normally distributed outcome measure. Furthermore, we let I be the number of levels of factor A (with index $i = 1,\ldots,I$) and J be the number of levels of factor B (with index $j = 1,\ldots,J$). The number of experimental units

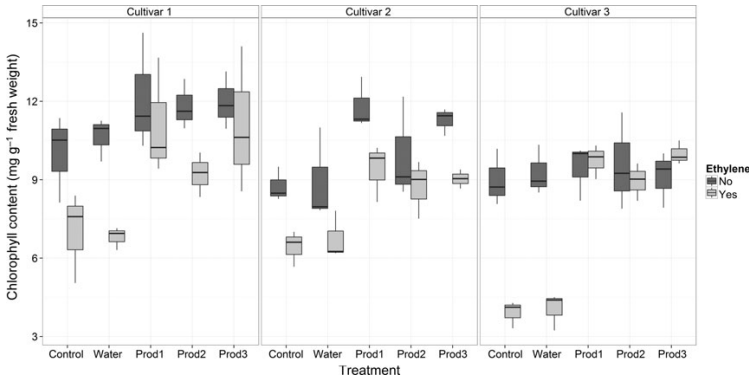


Fig. 2 Box plots of the chlorophyll content after 8 days based on each ethylene-by-treatment-by-cultivar combination.

Table 2 Three-way ANOVA table of the ethylene blocking data set

	d.f.	Sum Sq	Mean Sq	*F* value	Pr(>*F*)
Cultivar	2	53.22	26.61	15.99	<0.001
Ethylene	1	98.85	98.85	59.43	<0.001
Treatment	4	156.05	39.01	23.46	<0.001
Cultivar:Ethylene	2	0.37	0.19	0.11	0.894
Cultivar:Treatment	8	3.05	0.38	0.23	0.984
Ethylene:Treatment	4	35.65	8.91	5.36	<0.001
Cultivar:Ethylene: Treatment	8	28.73	3.59	2.16	0.044
Residuals	60	99.79	1.66		

is permitted to vary between the factor combinations and is denoted by n_{ij}. The corresponding two-way ANOVA model with an interaction term is given by:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad (1)$$

where the parameter μ denotes the overall mean, α_i is the treatment effect for the ith level of factor A, β_j the treatment effect for the jth level of factor B, and $(\alpha\beta)_{ij}$ denotes the joint effect of the ith level of factor A and the jth level of factor B. Furthermore, it is assumed that the error associated with the kth observation for the ijth treatment, with $k = 1,\ldots,N$, is normally distributed with common variance, $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$.

For the purpose of our method, we reformulate the classical ANOVA model as the cell means model as follows:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \qquad (2)$$

where the parameter μ_{ij} denotes the cell mean of the ith level of factor A and the jth level of factor B. Within this study, the dot notation is used to represent the averaging over the respective factor levels, for example $\mu_{i.}$ means the average over the levels of factor B.

Note that it is common practice in agricultural and horticultural research to display the means from a two-way layout in a $I \times J$ table (see e.g. Table 3). Each cell in Table 3 corresponds to a sample mean of a factor combination, which is an unbiased estimator for the parameters μ_{ij} in the cell means model. The lower and right margins in Table 3 represent the marginal means averaged over the levels of the other factor.

Hypotheses for interactions

In this section, we present suitable comparisons for the analysis of interactions, formulated in terms of the cell means model. The null hypotheses for the main effects of A and B test the equality of their marginal means, $H_0^A : \mu_{1.} = \mu_{2.} = \ldots = \mu_{I.}$ and $H_0^B : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.J}$.

As mentioned in Section 1, it is common practice to analyse the single effect of the primary factor after an F-test for interaction. This hypothesis can be formulated using the terms in Eq. (1) as:

$$\begin{aligned} H_0^{A(12)B(2)} : \mu_{12} - \mu_{22} &= (\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}) \\ &\quad - (\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}) \\ &= (\alpha_1 - \alpha_2) + ((\alpha\beta)_{12} \\ &\quad - (\alpha\beta)_{22}) \qquad = 0, \end{aligned} \qquad (3)$$

for $i = 1,2$ and $j = 2$. Obviously, the main effect for factor A and the interaction effect are confounded with one and another, so a clear interpretation concerning the interaction is not possible (as shown in Marascuilo and Levin 1970).

The global null hypothesis for the interaction effect tests the equality of all cell means to the expected additive effect of the two main effects. Expressed in terms of the cell means and marginal means, this can be written as:

$$\begin{aligned} H_0^{AB} : (\mu_{ij} - \overbrace{\mu_{..}}^{\text{overallmean}}) - (\overbrace{\mu_{i.} - \mu_{..}}^{\text{maineffectA}}) - (\overbrace{\mu_{.j} - \mu_{..}}^{\text{maineffectB}}) \\ = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} = 0 \text{ for all } i \text{ and } j. \end{aligned} \qquad (4)$$

The interaction effects are therefore also termed as interaction residuals or corrected cell means (Boik 1993). If the main effects are purely additive and there is no interaction, all pairwise differences between the levels of one factor are the same across all levels of the other factor. Thus, the null hypothesis of no interaction may alternatively be expressed as in terms of the cell means alone:

$$\begin{aligned} H_0^{AB} : \left(\mu_{ij} - \mu_{ij'}\right) - \left(\mu_{i'j} - \mu_{i'j'}\right) \\ = \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \\ = 0 \quad \text{for all } \{i, i'\} \text{ and } \{j, j'\}, \end{aligned} \qquad (5)$$

where $i \neq i'$ and $j \neq j'$ (Kirk 1995). For illustrative purposes, consider the first and third levels of factor A and the first and third levels of factor B (grey-marked cells in

Table 3 Table of cell means for two factorial designs, where factor A has $i = 1,2,3,\ldots,I$ levels and factor B has $j = 1,2,3,\ldots,J$ levels. Grey-marked cells define the tetradic contrasts of cell means $\mu_{11} - \mu_{13} - \mu_{31} + \mu_{33}$

	B_1	B_2	B_3	B_j	B_J	Means
A_1	μ_{11}	μ_{12}	μ_{13}	$\cdots$	μ_{1J}	$\mu_{1.}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\cdots$	μ_{3J}	$\mu_{3.}$
A_i	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	μ_{I1}	μ_{I2}	μ_{I3}	$\cdots$	μ_{IJ}	$\mu_{I.}$
Means	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\cdots$	$\mu_{.J}$	$\mu_{..}$

Table 3). The corresponding local hypothesis to test the interaction effect compares the difference between B_1 and B_3 at A_1 to the difference between B_1 and B_3 at A_3:

$$H_0^{A(13)B(13)} : (\mu_{11} - \mu_{13}) - (\mu_{31} - \mu_{33}) = \mu_{11} - \mu_{13} - \mu_{31} + \mu_{33} = 0 \quad \text{for } i = \{1, 3\} \text{ and } j = \{1, 3\} \tag{6}$$

Conversely, this can be interpreted as comparing (A_1–A_3) at B_1 to (A_1–A_3) at B_3. These comparisons are also known as tetradic contrasts (Bradu and Gabriel 1974), as each compares a set of four cell means. Nevertheless, interest might be only in a subset of these tetradic contrasts. In some experiments, not all pairwise comparisons between the levels of a factor may be of interest, but only comparisons to a control group. Other research hypotheses may require considering a specific set of comparisons, or may consider differences between pooled factor levels, depending on the particular treatment structure. Such user-defined comparisons may be defined in terms of contrasts. Recall that a contrast or comparison among means is a linear combination of means that have *a priori* known weights or coefficients (see e.g. Kirk 1995). The sum of the coefficients in a given contrast should be equal to zero to be interpretable as a difference. Several contrasts can be combined in a contrast matrix, where we will assume here that each row of such a matrix corresponds to one comparison of interest.

For the construction of user-defined interaction contrasts, we assume that the differences of interest with respect to the $i = 1,\ldots, I$ levels of factor A are defined in the contrast matrix $\mathbf{C}_A$ with I columns, whereas the differences of interest among the $j = 1,\ldots,J$ levels of factor B are defined in the contrast matrix $\mathbf{C}_B$ with J columns. The Kronecker product, denoted by $\otimes$, can be used to define the corresponding interaction contrasts. Recall that the Kronecker product of two matrices multiplies each element of the first matrix with the second matrix. Thus, $\mathbf{C}_A \otimes \mathbf{C}_B$ leads to a matrix $\boldsymbol{C}_{AB}$ that defines the M comparisons of interest with each column corresponding to one of the $ij = 11,21,\ldots, IJ$ cell means. Gabriel et al. (1973) presented this general approach to construct comparisons of cell means, which they denoted as (Kronecker) product-type interaction contrasts, to build simultaneous confidence intervals for interactions. All tetradic contrasts in Eq. (5) can be constructed in $\mathbf{C}_{AB} = \mathbf{C}_A \otimes \mathbf{C}_B$ when all pairwise comparisons (Tukey-type contrasts) among the I levels of factor A and among the J levels of factor B are defined in $\mathbf{C}_A$ and $\mathbf{C}_B$, respectively. However, if only a subset of these is of interest *a priori*, statistical inference and corrections for multiple comparisons should be restricted to only this subset in order to avoid too conservative adjustments for multiple testing.

As an example, consider the ethylene blocker data set introduced in Section 2.2. The first factor ethylene consists of two levels, an untreated group and an ethylene-treated group. To investigate the effect of ethylene on the chlorophyll content, the difference between the two groups is investigated. This difference is further formulated in terms of a contrast in the matrix: $\mathbf{C}_A = (1\ -1)$. The second factor, MCP treatment, includes an untreated control group, a water-sprayed control group and three different MCP blocking groups. The objective of including this factor is to estimate the effect of the MCP product in contrast to the control groups on the chlorophyll content. Further interest is in comparing the three different commercial products that include the ethylene blocker MCP. The corresponding contrast matrix $\mathbf{C}_B$ that translates these research hypotheses in terms of contrasts is given in Table 4. To investigate a potential different ethylene effect according to the second factor, a detailed analysis of this ethylene-by-treatment interaction is conducted using the pre-defined contrast matrices $\mathbf{C}_A$ and $\mathbf{C}_B$. The resulting interaction contrast matrix $\mathbf{C}_{AB} = \mathbf{C}_B \otimes \mathbf{C}_A$ is given in Table 5. The first row in $\boldsymbol{C}_{AB}$ compares the ethylene effect between the two control groups and the MCP-treated groups. The remaining three rows compare the ethylene effect between the three different MCP products.

Statistical inference

The objective is now to simultaneously test the M hypotheses represented by the M rows of a given matrix of interaction contrasts $\boldsymbol{C}_{AB}$. The test statistic for one contrast is given by

$$T = \frac{c_{11}\hat{\mu}_{11} + c_{12}\hat{\mu}_{12} + c_{13}\hat{\mu}_{13} + \ldots + c_{IJ}\hat{\mu}_{IJ}}{s\sqrt{\frac{c_{11}^2}{n_{11}} + \frac{c_{12}^2}{n_{12}} + \frac{c_{13}^2}{n_{13}} + \ldots + \frac{c_{IJ}^2}{n_{IJ}}}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\hat{\mu}_{ij}}{s\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{c_{ij}^2}{n_{ij}}}},$$

where s is the square root of the pooled sample variance and $\hat{\mu}_{ij}$ are the estimators for the cell means. To adjust for

Table 4 Matrix of contrasts for the treatment factor for the second example in Section 2. The first row compares the control groups against the ethylene blocking groups. The following three rows conduct all pairwise differences of the ethylene blocking groups

Comparison	Untreated control	Water control	Prod1	Prod2	Prod3
Control – Product	0.5	0.5	−0.33	−0.33	−0.33
Prod1 – Prod2	0	0	1	−1	0
Prod1 – Prod3	0	0	1	0	−1
Prod2 – Prod3	0	0	0	1	−1

Table 5 Interaction contrast matrix obtained from the direct Kronecker product of the matrices $\mathbf{C}_A$ = (1–1) and $\mathbf{C}_B$ from Table 4. The first row compares the ethylene effect between the control groups and the MCP-treated groups. The remaining rows compare the ethylene effect between the three different MCP products

Comparison	Eth.no untr. control	Eth.yes untr. control	Eth.no water control	Eth.yes water control	Eth.no Prod1	Eth.yes Prod1	Eth.no Prod2	Eth.yes Prod2	Eth.no Prod3	Eth.yes Prod3
Ethylene (Control – Product)	0.5	−0.5	0.5	−0.5	−0.33	0.33	−0.33	0.33	−0.33	0.33
Ethylene (Prod1 – Prod2)	0	0	0	0	1	−1	−1	1	0	0
Ethylene (Prod1 – Prod3)	0	0	0	0	1	−1	0	0	−1	1
Ethylene (Prod2 – Prod3)	0	0	0	0	0	0	1	−1	−1	1

multiple comparisons and therefore control the family-wise error rate for the family of M contrasts, we use the framework of multiple contrast tests as described by Hothorn et al. (2008) and Bretz et al. (2010). The null hypothesis for a particular interaction contrast m is rejected if the corresponding $|T| > q_{1-\alpha,M,\mathbf{R},\eta}$ where $q_{1-\alpha,M,\mathbf{R},\eta}$ is the two-sided critical value from a multivariate t-distribution with dimension M, a pre-specified significance level α, degree of freedom $\eta = \sum_{i-1}^{I}\sum_{j-1}^{J}(n_{ij}-1)$ and a correlation matrix R that depends on the sample sizes n_{ij} and the contrast coefficients c_{ij} (for computational details see Bretz et al. (2001)). Corresponding simultaneous two-sided confidence intervals for each of the M interaction contrasts are given by

$$\sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\hat{\mu}_{ij} \pm q_{1-\alpha,M,\mathbf{R},\eta}\cdot s\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{c_{ij}^2}{n_{ij}}},$$

and allow to interpret the magnitude, the direction and the biological relevance of the inferred differences as well as test decisions (if zero is not included in one of the intervals, the contrast is not significantly different from zero). The critical value chosen according to the above methodology accounts for the number of multiple comparisons, M, as well as for the correlations among the M-test statistics under the related null hypothesis that results from repeatedly involving the same cell means in several of the M contrasts. The test decisions provided by this method are thus not conservative in contrast to commonly known adjustments for multiple comparisons, as the Scheffé or Bonferroni method (Nelson 1989). The computational methods are freely available for the software R (R Core Team 2013) using the add-on package multcomp (Hothorn et al. 2008). For users of the SAS software (SAS Institute Inc. 2013), a related method with different computational details is available in the LSMESTIMATE statement of the GLIMMIX procedure (Westfall et al. 2011).

Interaction contrasts for fixed effects in mixed models

Agricultural experiments usually involve randomization structures, for example, are performed as randomized complete block, split plot or more complicated designs. A general way to analyse such data is linear mixed models, where the main effects and their interaction are modelled in the fixed effects part, and various randomization structures may be accounted for in the random effects part (Piepho et al. 2003). Consider the following model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{7}$$

where $\mathbf{X}$ is the design matrix of fixed effects, here containing the dummy-coded treatment levels of the cell means model in Eq. (2), and $\boldsymbol{\beta}$ is the corresponding parameter vector. $\mathbf{Z}$ and $\mathbf{u}$ are the design matrix and the vector for the random effects, driven by the corresponding experimental design. The vector of residuals, $\mathbf{e}$, may assumed to be independent or may model dependencies among observations imposed by repeated measurements. Note that $\mathbf{X}$ may contain further fixed effects like covariates, or a block effect treated as fixed effect. In this situation, parameters corresponding to the ordered vector of cell means introduced above can be obtained by linear combinations of $\boldsymbol{\beta}$, for example, by computing the IJ treatment means pooled over the blocks or by computing treatment means at the average covariate value; see the bush beans example.

For this situation, simultaneous inference is described in Hothorn et al. (2008). Their methods are briefly described in the following: fitting the model yields an estimate of the parameter vector $\hat{\boldsymbol{\beta}}$ and a corresponding covariance matrix $\hat{\sum}$. The vector of the M contrast estimates results from $\hat{\boldsymbol{\delta}} = \mathbf{C}_{AB}\hat{\beta}$, with elements $\hat{\delta}_m$, the corresponding covariance matrix $\hat{\boldsymbol{V}} = \mathbf{C}_{AB}\hat{\sum}\mathbf{C}_{AB}^T$, with the vector of its M diagonal elements denoted $\hat{\boldsymbol{v}}$, and standardizing $\hat{\boldsymbol{V}}$ by its diagonal elements yields the correlation matrix $\hat{\mathbf{R}}$ (Hothorn et al. 2008). The test statistic of the mth contrast is given by $T_m = \hat{\delta}_m/\sqrt{\hat{v}_m}$, and simultaneous confidence intervals for the interaction contrasts can then be obtained by

$$\hat{\delta}_m \pm q\sqrt{\hat{v}_m}, \tag{8}$$

where the appropriate quantile q may be computed from the multivariate normal distribution, $q = Z_{1-\alpha,M,\hat{\mathbf{R}}}$, to yield

asymptotic intervals (Hothorn et al. 2008). This procedure will be on the liberal side in small samples, because error degree of freedom is set to infinity. Alternatively, exact confidence intervals can be obtained, when using quantiles of the multivariate *t*-distribution, $q = q_{1-\alpha,M,\hat{\mathbf{R}},\hat{\eta}}$, where $\hat{\eta}$ is an approximate denominator degree of freedom (e.g. Pinheiro and Bates 2000, p.91) of the related ANOVA *F*-test for the fixed effect corresponding to the cell means.

Analysis of the examples

In this section, the above presented examples are analysed using the previously presented methods.

Bush beans example analysed

To construct the product-type interaction contrast matrix, we first specify the contrast matrices for the main effects for the factors spacing and variety. For the factor spacing, we use all pairwise comparisons between the three row spacings (Table 6). For the factor variety, we first want to compare the average of the two tall varieties with that of the two bushy varieties (see first row in Table 7). Furthermore, we compare the varieties within each of the two growth types, see rows 2 and 3 from Table 7. To get the product-type interaction contrasts, we build the direct Kronecker product of these matrices, which results in the matrix listed in Table 8.

Table 6 Contrast matrix that compares all pairwise differences of the spacing factor for the first example in Section 2

Comparison	20 cm	40 cm	60 cm
20–40	1	−1	0
20–60	1	0	−1
40–60	0	1	−1

Table 7 User-defined contrasts that compare the different varieties from the first example in Section 2

Comparison	BigGreen	NewEra	LittleGem	RedLake
Bushy – Tall	0.5	0.5	−0.5	−0.5
BigGreen – NewEra	1	−1	0	0
LittleGem – RedLake	0	0	1	−1

Figure 3 displays the two-sided simultaneous confidence intervals for the user-defined interaction contrasts in Table 8. The null hypothesis that a single interaction contrast is zero is rejected if the confidence interval does not include the value zero. From the top three confidence intervals in Fig. 3, it is obvious that the spacing effect is different between the two growth types. Considering the second confidence interval in Fig. 3, we conclude that the mean yield of the bushy varieties is at least about 12 kg per plot higher than that of the tall varieties if the row spacing increases from 20 to 60 cm. Furthermore, we can conclude that the difference between the two growth habits increases with increasing row spacings. In contrast to the comparison of the different spacing effects between the two growth habits, the spacing effect is not different between the varieties of the same growth habit.

For the special case of this example, there are two alternative procedures for the evaluation: The first alternative approach takes the hierarchical structure between the growth type and the varieties within the growth type into account. The data may then be analysed by a three-factorial ANOVA with factors growth type and spacing crossed, and the third factor variety nested in growth type (see Table 9). The significant *Type:Spacing* effect indicates that the two growth types, bushy and tall, differ in their mean response when spacing is increased. However, the two varieties within each growth type show no significantly different reaction with increasing spacing (*Type:Spacing:Variety* P = 0.073). These test results coincide with those from the interaction contrast test approach. But, using the simultaneous

Table 8 Interaction contrast matrix obtained from the direct Kronecker product of the matrices $\mathbf{C}_A$ and $\mathbf{C}_B$ from Tables 6 and 7

	BigGreen			NewEra			LittleGem			RedLake		
Comparison	20	40	60	20	40	60	20	40	60	20	40	60
(Bushy – Tall)20 – (Bushy – Tall)40	0.5	−0.5	0	0.5	−0.5	0	−0.5	0.5	0	−0.5	0.5	0
(Bushy – Tall)20 – (Bushy – Tall)60	0.5	0	−0.5	0.5	0	−0.5	−0.5	0	0.5	−0.5	0	0.5
(Bushy – Tall)40 – (Bushy – Tall)60	0	0.5	−0.5	0	0.5	−0.5	0	−0.5	0.5	0	−0.5	0.5
(BigGreen – NewEra)20 – (BigGreen – NewEra)40	1	−1	0	−1	1	0	0	0	0	0	0	0
(BigGreen – NewEra)20 – (BigGreen – NewEra)60	1	0	−1	−1	0	1	0	0	0	0	0	0
(BigGreen – NewEra)40 – (BigGreen – NewEra)60	0	1	−1	0	−1	1	0	0	0	0	0	0
(LittleGem – RedLake)20 – (LittleGem – RedLake)40	0	0	0	0	0	0	1	−1	0	−1	1	0
(LittleGem – RedLake)20 – (LittleGem – RedLake)60	0	0	0	0	0	0	1	0	−1	−1	0	1
(LittleGem – RedLake)40 – (LittleGem – RedLake)60	0	0	0	0	0	0	0	1	−1	0	−1	1

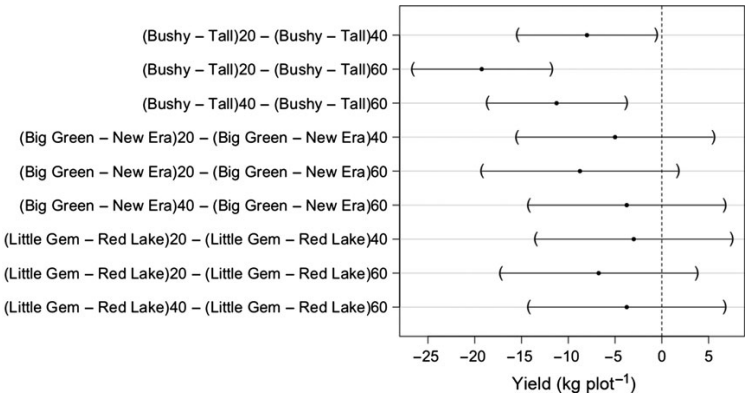


Fig. 3 Simultaneous 95 % confidence intervals for user-defined interaction contrasts as specified in Table 8. Dots denote the estimates for the comparison of interest, and vertical bars the lower and upper limit of the two-sided confidence intervals.

Table 9 ANOVA table of the bush beans data set that takes the hierarchical structure between the growth type and the varieties within the growth type into account

	d.f.	Sum Sq	Mean Sq	*F* value	Pr(>*F*)
Block	3	341.90	113.97	8.78	<0.001
Type	1	105.02	105.02	8.09	0.008
Spacing	2	72.67	36.33	2.8	0.075
Type:Spacing	2	748.17	374.08	28.82	<0.001
Type:Variety	2	1227.54	613.77	47.28	<0.001
Type:Spacing:Variety	4	122.83	30.71	2.37	0.073
Residuals	33	428.35	12.98		

Table 10 Results for the bush beans data set analysed with user-defined contrasts of slopes for the linear regression on spacing. Estimate denotes the estimate for the comparison of interest, the adjusted P-value, Lower and Upper the lower and upper bound of the two-sided 95 % simultaneous confidence interval

Comparison	Estimate	P-Value	Lower	Upper
Bushy – Tall	0.48	<0.001	0.329	0.633
BigGreen – NewEra	0.22	0.044	0.004	0.433
LittleGem – RedLake	−0.17	0.159	−0.383	0.046

confidence intervals of the interaction contrast approach allows the determination of the magnitude and the direction of the interaction effect as described above which is not possible with the ANOVA approach.

The second alternative procedure considers the spacing factor as quantitative. The corresponding general linear model is given by $y_{ijk} = a_i + c_j + b_j x_{ijk} + e_{ijk}$, where a_i are the block effects, c_j are the intercepts of the four varieties and b_j are the variety-specific slopes in regression on spacing, x_{ijk}. In this sparser model, multiple comparisons (Hothorn et al. 2008) among the regression slopes b_j provide a meaningful interpretation of the variety-by-spacing interaction. Table 7 shows user-defined contrasts among the four variety-specific regression slopes that compare the averaged slope of the bushy and tall growth type, as well as the difference of slopes within each growth type. The resulting estimates, multiplicity-adjusted P-values and simultaneous confidence intervals for the differences of regression slopes are shown in Table 10. Again, a significant variety-by-spacing interaction is determined based on the different average slopes between the two growth types. Additionally, a marginally significant variety-by-spacing interaction (P = 0.044) is detected for the comparison of the low, bushy varieties BigGreen and NewEra: the increase in mean yield when increasing spacing is at least slightly bigger for BigGreen than for NewEra.

Ethylene blocking example analysed

For a detailed analysis of the two-way ethylene-by-treatment interaction, the user-defined interaction contrasts derived in Section 3 and displayed in Table 5 are used. These interaction contrasts compare the differences of the ethylene effect between the control groups and the MCP-treated groups, and between all pairs of the MCP-treated groups. Applying the method presented in Section 3, we get the multiplicity-adjusted P-values and simultaneous confidence intervals for the four interaction contrasts of interest in Table 11. The comparison of the ethylene effect between the control groups and the MCP-treated groups results in a significant effect of this interaction term ($P < 0.05$). The corresponding confidence interval additionally provides the information that the ethylene effect is significantly greater for the control groups (lower bound greater than zero). Furthermore, the confidence interval provides some information on the magnitude of the interaction effect: the mean decrease in chlorophyll content is at least 1.11 mg g^{-1} greater in the control groups than in the MCP-treated groups. The scientist can now decide whether the magnitude of this statistically significant interaction

Table 11 Results for the two-way interaction analysis of the ethylene blocking data set. Estimate denotes the estimate for the comparison of interest, the adjusted P-value, Lower and Upper the lower and upper bound of the two-sided 95 % simultaneous confidence interval

Comparison	Estimate	P-value	Lower	Upper
Ethylene (control – Product)	2.52	<0.001	1.11	3.92
Ethylene (Prod1 – Prod2)	−0.44	0.948	−2.61	1.74
Ethylene (Prod1 – Prod3)	0.29	0.983	−1.88	2.47
Ethylene (Prod2 – Prod3)	0.73	0.802	−1.45	2.91

term (1.11 mg g^{-1}) is also of biological relevance, for example, to improve post-harvest cutting quality of *Pelargonium zonale*. From Table 11, we can further conclude that the ethylene effect is not significantly different between the three different MCP-treated groups.

For illustrative purposes, the analysis of the three-way ethylene-by-treatment-by-cultivar interaction is also presented here. For this purpose, we formulate the appropriate contrasts for the factor cultivar in Table 12, which performs all pairwise comparisons between the three cultivars. The Kronecker product $\mathbf{C}_{ABC} = \mathbf{C}_c \otimes \mathbf{C}_{AB}$ results in an interaction contrast matrix for the analysis of the three-way interaction. The first four rows in $\mathbf{C}_{ABC}$ compare the two-way interaction contrasts defined by $\mathbf{C}_{AB}$ (Table 5) between cultivar 1 and cultivar 2, while rows five to eight compare the same two-way contrasts between cultivar 1 and cultivar 3 and so on. Within each of these blocks, the first row compares the differences of the ethylene effect between the control groups and the MCP-treated groups between two cultivars. The second row compares this differing ethylene effect, based on the difference between product 1 and product 2, between two cultivars, and so on.

Using the method presented in Section 3, we get the multiplicity-adjusted P-values and simultaneous confidence intervals for the twelve interaction contrasts of interest defined by $\mathbf{C}_{ABC}$ in Table 13. The comparison of the differing ethylene effect, which is based on the difference of the control groups and the MCP-treated groups, between cultivar 2 and cultivar 3 results in a significant effect of this interaction term ($P < 0.05$). From the corresponding confidence interval, we conclude that the difference of the ethylene effect between the control groups and the MCP-treated groups is smaller in cultivar 2 than in cultivar 3 (upper bound smaller than zero). Additionally, the magnitude of the interaction effect is determined by the confidence interval: the influence of the MCP products on the ethylene effect in contrast to the influence of the control groups on the ethylene effect is at least 1.09 mg g^{-1} (chlorophyll per fresh weight) smaller in cultivar 2 than in cultivar 3. Furthermore, it is concluded that the potentially different influence between the various MCP products on the ethylene effect is not significantly different between the three cultivars (see Table 13).

Table 12 Contrast matrix that compares all pairwise differences of the cultivar factor for the second example in Section 2

Comparison	Cultivar 1	Cultivar 2	Cultivar 3
Cultivar 1 – Cultivar 2	1	−1	0
Cultivar 1 – Cultivar 3	1	0	−1
Cultivar 2 – Cultivar 3	0	1	−1

Table 13 Results for the three-way interaction analysis of the ethylene blocking data set. Estimate denotes the estimate for the comparison of interest, the adjusted P-value, Lower and Upper the lower and upper bound of the two-sided 95 % simultaneous confidence interval

Comparison	Estimate	P-value	Lower	Upper
Ethylene (control – Product) Cult1-Cult2	1.62	0.861	−2.33	5.57
Ethylene (Prod1 – Prod2) Cult1-Cult2	−2.78	0.794	−8.91	3.34
Ethylene (Prod1 – Prod3) Cult1-Cult2	−0.04	1.000	−6.16	6.08
Ethylene (Prod2 – Prod3) Cult1-Cult2	2.75	0.804	−3.38	8.87
Ethylene (control - Product) Cult1-Cult3	−3.42	0.130	−7.37	0.53
Ethylene (Prod1 – Prod2) Cult1-Cult3	−0.66	1.000	−6.78	5.46
Ethylene (Prod1 – Prod3) Cult1-Cult3	−0.45	1.000	−6.58	5.67
Ethylene (Prod2 – Prod3) Cult1-Cult3	0.21	1.000	−5.91	6.33
Ethylene (control – Product) Cult2-Cult3	−5.04	<0.001	−8.99	−1.09
Ethylene (Prod1 – Prod2) Cult2-Cult3	2.12	0.932	−4.00	8.24
Ethylene (Prod1 – Prod3) Cult2-Cult3	−0.42	1.000	−6.54	5.71
Ethylene (Prod2 – Prod3) Cult2-Cult3	−2.54	0.855	−8.65	3.59

The R Code (R Core Team 2013) for all proposed approaches to evaluate the example data sets is given in the Supporting Information.

Discussion

We have presented an approach for the evaluation of a significant interaction effect in factorial designs. Although this study primarily considers the analysis of interactions in two factorial designs, the method is also applicable to higher factorial designs. In those cases, the user has to define appropriate contrast matrices for each factor under consideration and subsequently has to build the Kronecker product of these matrices.

The proposed approach allows the formulation of user-specified comparisons of means that are of main interest in the experiment under consideration. In addition to multiplicity-adjusted P-values, the authors recommend the calculation of simultaneous confidence intervals for the interaction effects. Using these confidence intervals, it is possible to evaluate the amount of the interaction effects. Besides a statement on the statistical significance, this also provides a statement on the biological relevance of the interaction effects.

The presented procedure is easily applicable using the add-on package multcomp (Hothorn et al. 2008) within the statistical software R (R Core Team 2013). In addition, the authors provide the add-on package statint stored on github (https://github.com/AKitsche/statint) with commented R code to reproduce the analysis of the presented examples. To install directly from github, the package devtools is needed:

```
install.packages('devtools')
library(devtools)
install_github(username='AKitsche', repo='statint')
library(statint)
```

The vignette of this package is also given in the online Supporting Information.

In this article, we considered the situation of independently and normally distributed error terms with a common variance. If the assumption of homogeneous variances is not fulfilled, the procedure presented by Hasler and Hothorn (2008) may be used instead. For cases in which the data are not considered to be normally distributed, the ANOVA model can be replaced by a generalized linear model (Hothorn et al. 2008) or the rank-based multiple test procedures presented by Konietschke et al. (2012) can be used.

Acknowledgements

We thank Syariful Mubarok for the permission to show the ethylene blocking data set. The authors are grateful to the editor and two anonymous referees for their numerous helpful comments on earlier drafts, which greatly improved this article. The work was partly supported by the German Science Foundation grant DFG HO1687/9-1.

References

Boik, R. J., 1993: The analysis of two-factor interactions in fixed effects linear models. J. Educ. Stat. 18, 1–40.

Bradu, D., and K. R. Gabriel, 1974: Simultaneous statistical inference on interactions in two-way analysis of variance. J. Am. Stat. Assoc. 69, 428–436.

Bretz, F., A. Genz, and L. A. Hothorn, 2001: On the numerical availability of multiple comparison procedures. Biom. J. 43, 645–656.

Bretz, F., P. H. Westfall, and T. Hothorn, 2010: Multiple comparisons using R. Chapman & Hall/CRC, Boca Raton, FL, USA.

Clewer, A. G., and D. H. Scarisbrick, 2001: Practical statistics and experimental design for plant and crop science. Wiley, Chichester, UK.

R Core Team. 2013: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Frömke, C., and F. Bretz, 2004: Simultaneous tests and confidence intervals for the evaluation of agricultural field trials. Agron. J. 96, 1323–1330.

Gabriel, K. R., J. Putter, and Y. Wax, 1973: Simultaneous confidence intervals for product-type interaction contrasts. J. R. Stat. Soc. Series B Stat. Methodol. 35, 234–244.

Genz, A., and F Bretz, 1999: Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation 63, 361–378.

Gomez, K. A., and A. A. Gomez, 1984: Statistical Procedures for Agricultural Research. An International Rice Research Institute book. 2nd edn, Wiley, New York, NY, USA.

Harwell, M., 1998: Misinterpreting interaction effects in analysis of variance. Meas. Eval. Couns. Dev. 31, 125–136.

Hasler, M., and L. A. Hothorn, 2008: Multiple contrast tests in the presence of heteroscedasticity. Biom. J. 50, 793–800.

Hoshmand, A. R., 2006: Design of experiments for agriculture and the natural sciences. 2nd edn. Chapman & Hall/CRC, Boca Raton, FL, USA.

Hothorn, L. A., 2003: Statistics of interlaboratory in vitro toxicological studies. Altern. Lab. Anim. 31, 43–63.

Hothorn, L. A., and H. Bleiholder, 2006: Statistical aspects of efficacy evaluation of plant protection products in field trials - A comment on EPPO Standard PP1/152. EPPO Bull. 36, 31–45.

Hothorn, T., F. Bretz, and P. Westfall, 2008: Simultaneous inference in general parametric models. Biom. J. 50, 346–363.

Kirk, R. E., 1995: Experimental design: Procedures for the behavioural sciences. 3rd edn. Brooks/Cole, Pacific Grove, CA, USA.

Konietschke, F., L. A. Hothorn, and E. Brunner, 2012: Rank-based multiple test procedures and simultaneous confidence intervals. Electronic J. Stat. 6, 738–759.

Marascuilo, L. A., and J. R. Levin, 1970: Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: the elimination of type IV errors. Am. Educ. Res. J. 7, 397–421.

Menke, S., H.-M. Poehling, and D. Gerhard, 2011: Application of generalized linear model based multiple comparison procedures to greenhouse pest control experiments. Acta Hortic. 893, 1233–1238.

Nelson, P., 1989: Multiple comparisons of means using simultaneous confidence intervals. J. Qual. Technol. 21, 232–241.

Petersen, R. G., 1985: Design and Analysis of Experiments. Volume 66 of Statistics. Dekker, New York, NY, USA.

Piepho, H. P., A. Büchse, and K. Emrich, 2003: A Hitchhiker's Guide to Mixed Models for Randomized Experiments. J. Agron. Crop Sci. 189, 310–322.

Pinheiro, J. C., and D. Bates, 2000: Mixed-Effects Models in S and S-PLUS. Springer-Verlag Inc, New York, NY, USA.

Sahin, O., M. Taskin, Y. Kadioglu, A. Inal, A. Gunes, and D. Pilbeam, 2012: Influence of chloride and bromate interaction on oxidative stress in carrot plants. Sci. Hortic. 137, 81–86.

SAS Institute Inc. 2013: SAS/STAT® 12.3 User's Guide. SAS Institute Inc, Cary, NC, USA.

Schaarschmidt, F., and L. Vaas, 2009: Analysis of trials with complex treatment structure using multiple contrast tests. HortScience 4, 188–195.

Slauenwhite, K. L. I., and M. M. Qaderi, 2013: Single and interactive effects of temperature and light quality on four canola cultivars. J. Agron. Crop Sci. 199, 286–298.

Westfall, P. H., and S. S. Young, 1993: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York, NY, USA.

Westfall, P. H., R. D. Tobias, and R. D. Wolfinger, 2011: Multiple Comparisons and Multiple Tests Using SAS, 2nd edn. SAS Publishing, Cary, NC, USA.

Statistical Methods and Application manuscript No.
(will be inserted by the editor)

Multiple treatment comparisons in analysis of covariance with interaction

Frank Schaarschmidt

2015/09/29

Abstract When multiple treatments are analyzed together with a covariate, a treatment-covariate interaction complicates the interpretation of the treatment effects. The construction of simultaneous confidence bands for differences of the treatment specific regression lines is one option to proceed. The application of these methods is difficult because they are described as a collection of special cases and the implementation requires additional programming or relies on non-standard or proprietary software. A flexible alternative is to compute simultaneous confidence intervals for multiple contrasts of the treatment effects over a grid of covariate values. This approach is available in the R software. Next to treatment differences in the linear model, approximate simultaneous confidence intervals for ratios of expected values and asymptotic extensions to generalized linear models are straightforward. The paper summarizes the available methodology and presents three case studies to illustrate the application to different models, differences and ratios, as well as different types of between treatment comparisons. Simulation studies in the general linear model, for different parameters and different types of comparisons are provided. The R code to reproduce the case studies and a hint to a related R package is provided.

Keywords multiple contrasts · simultaneous confidence intervals · treatment covariate interaction · generalized linear model · multiple ratios · confidence bands

Frank Schaarschmidt
Institute of Biostatistics, Leibniz Universitaet Hannover, Herrenhaeuser Str.2, 30419 Hannover, Germany
Tel.: +49 511 762 5821
Fax: +49 511 762 4966
E-mail: schaarschmidt@biostat.uni-hannover.de

1 Introduction

In linear models, the effects of a classification variable, e.g., the indicator for two or several treatments, can be modeled together with that of covariates. The presence of a significant overall treatment-covariate interaction complicates the interpretation of treatment effects: the significance, magnitude or even direction of the treatment effects depends on the value of the covariate. Nevertheless, the primary objective can be the comparison of the treatments. Often not all possible comparisons but only a special subset of treatment comparisons are of interest. A simplistic approach is to perform these multiple treatment comparisons for one fixed value of the covariate, e.g., the overall mean of the covariate. A more detailed comparison of treatments is provided by methods that yield confidence bands for differences between the treatment-specific regression lines. However, the practical application of such methods is complicated by the fact that they are described as many separate special cases. The focus of this work is on a flexible and user-friendly alternative that is computationally available in free software: simultaneous confidence intervals for multiple contrasts among treatments for a set of pre-specified values of the covariate.

A multitude of publications consider the construction of simultaneous confidence bands for (multiple) differences of regression lines, and it is difficult to review all methodological special cases completely. The methods differ in the number of treatments and the set of contrasts between treatments that can be handled; they differ in whether restrictions are imposed on the treatment specific subsets of the design matrix, the number of covariates and the considered range of the covariate. Most methods have in common that they are based on the assumptions of the general linear model. The very general and easily applicable method by Scheffe (1959) can be used to construct exact simultaneous confidence intervals for all possible contrasts. However, in many applications a restricted set of contrasts among the treatments and a restricted range of the covariates is of interest a priori, for example, all pairwise comparisons, comparisons to a control, special user-defined contrasts or one-sided comparisons. Then, the Scheffe method yields unnecessarily conservative confidence bands. Alternative solutions are provided for all contrasts but a restricted range of covariates (Spurrier 1999; Lu and Chen 2009; Jamshidian et al. 2010). Confidence bands for all pairwise differences and differences to control among several treatments have been proposed (Spurrier, 2002; Bhargava and Spurrier, 2004), however, under restrictive assumptions concerning the equality of the treatment specific design matrices. In a paper addressing various problems (Liu et al. 2004), a number of numerical approaches is described for all pairwise differences and differences to control with three or more treatment groups, several covariates, and, importantly, without severe restrictions on the design matrices. In a recent book (Liu 2010), a number of these approaches is described again. However, all these previous publications concerning exact simultaneous confidence bands have two practical problems: the computational methods are split up in a number of special cases, described in special publications

or book chapters, and, more severely, putting the computation of the critical values into practice usually requires the additional programming of the described algorithms or relies on non-standard or proprietary software packages (Jamshidian et al. 2005).

Alternatively, one may define a set of covariate values and construct simultaneous confidence intervals (SCI) for multiple comparisons among the treatments for this set of values. This approach leads to a more detailed interpretation of the treatment effects in case of an interaction, and treatments can be compared in terms of multiple contrasts which are tailored for the particular experimental question (Bretz et al. 2001). Standard problems as all pairwise comparisons and comparisons to control are contained as special cases. Asymptotically, this approach can be used in generalized linear models (Hothorn et al. 2008), or, treatment effects may be expressed as ratios instead of differences, using approaches of Young et al. (1997) and Dilba et al. (2006). The computational methods to obtain adequate quantiles of multivariate t and multivariate normal distributions are available in the package `mvtnorm` (Genz et al. 2011) in the R software. For the special case of comparing two treatments, this approach has been applied recently by Bretz et al. (2010, p.111-114) and, with different computational details by Westfall et al. (2011).

This manuscript recapitulates the methods to construct simultaneous confidence intervals for multiple contrasts among the treatments for a pre-specified set of covariate values. Approximate extensions to generalized linear models multiple ratios are described. A simulation study is presented to assess the validity of the methods for differences and ratios in the general linear model. Three examples illustrate the application, including all pairwise differences, comparisons to a control in terms of ratios in a model including an interaction to the quadratic term, as well as all pairwise comparisons in log logistic model assuming a binomial response.

2 Material and methods

Consider the general linear model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is an $(N \times 1)$ vector of observations, $\mathbf{X}$ is an $(N \times P)$ design matrix, $\boldsymbol{\theta}$ is a $(P \times 1)$ parameter vector with index $p = 1, ..., P$, and $\boldsymbol{e}$ an $(N \times 1)$ vector of residuals. The residuals are assumed to be identically Gaussian distributed, $e_n \sim N(0, \sigma^2)$, independently for $n = 1, ..., N$. Fitting the model yields the estimate $\hat{\boldsymbol{\theta}}$ and the corresponding $(P \times P)$ covariance matrix $\hat{\boldsymbol{\Sigma}}$.

2.1 SCI for linear combination of parameters

Under the assumption of an independent, homogeneous Gaussian error distribution, the estimates $\hat{\boldsymbol{\theta}}$ follow a multivariate normal distribution. The predictions $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$, linear combinations thereof, or other linear combinations of the

model parameters follow a multivariate normal distribution as well. Simultaneous confidence intervals for M linear combinations of the P model parameters can be constructed using quantiles of the multivariate t distribution with degree of freedom $N-P$, or, asymptotically, using multivariate normal quantiles (Genz et al. 2001; Hothorn et al. 2008). The general methodology according to Hothorn et al. (2008) is:

Let $\mathbf{C}$ be a $(M \times P)$ matrix with elements c_{mp}, $m = 1, ..., M$, which define M linear combinations of the P model parameters, $\boldsymbol{\delta} = \mathbf{C}\boldsymbol{\theta}$. An estimate for $\boldsymbol{\delta}$ is $\hat{\boldsymbol{\delta}} = \mathbf{C}\hat{\boldsymbol{\theta}}$. The $(M \times M)$ covariance matrix of $\hat{\boldsymbol{\delta}}$ can be estimated by $\hat{\mathbf{V}} = \mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^T$, where ^{T} denotes a transposed matrix. Denote the diagonal elements of $\hat{\mathbf{V}}$ by $\hat{\boldsymbol{v}} = (\hat{v}_{11}, \hat{v}_{22}, ..., \hat{v}_{MM})$. Standardizing the covariance matrix $\hat{\mathbf{V}}$ by its diagonal elements yields the correlation matrix $\hat{\mathbf{R}}$ with elements $r_{mm'}$, i.e., $r_{mm'} = \hat{v}_{mm'}\hat{v}_{mm}^{-1/2}\hat{v}_{m'm'}^{-1/2}$.

The lower and upper limits, $\hat{\delta}_m^{(l)}, \hat{\delta}_m^{(u)}$, of simultaneous 95% confidence intervals for the M linear combinations can be constructed by

$$\left[\hat{\delta}_m^{(l)}, \hat{\delta}_m^{(u)}\right] = \left[\hat{\delta}_m \pm t_{0.95,\hat{\mathbf{R}},df=N-P}\hat{v}_{mm}^{-1/2}\right],$$

where $\hat{\delta}_m$ is the mth element of $\hat{\boldsymbol{\delta}}$ and $t_{0.95,\hat{\mathbf{R}},df=N-P}$ is an appropriate two-sided 0.95 quantile of the multivariate t distribution as is computable using the `R`–package `mvtnorm` (Genz et al. 2011): $P\left(|t_m| < t_{0.95,\hat{\mathbf{R}},df=N-P}, \forall m = 1, ..., M\right) = 0.95$, where $\boldsymbol{t} = (t_1, ...t_M)^T$ is a central M-variate t random vector with degree of freedom $N-P$ and correlation $\hat{\mathbf{R}}$. When interest is in one-sided intervals, a quantile $t_{0.95,\hat{\mathbf{R}},df=N-P}$ has to be chosen such that

$$P\left(t_m < t_{0.95,\hat{\mathbf{R}},df=N-P}, \forall m = 1, ..., M\right) = 0.95.$$

The methods implemented in `mvtnorm` can deal with complicated structures of $\hat{\mathbf{R}}$, including the case that $\hat{\mathbf{R}}$ has not full rank. This case is important for the following applications, where confidence sets are constructed for substantially more linear combinations than there are elements in the parameter vector, that is $P < M$. Note that M is bounded at 1000 in this implementation. For the computational details, see Genz and Bretz (2009). An implementation of the complete method relying on a fitted model object and a corresponding contrast matrix $\mathbf{C}$, is available in the `R`–package `multcomp` (Hothorn et al. 2008).

These intervals are simultaneous 95% confidence intervals, i.e., the probability that at least one of the M true parameters δ is not included, is smaller than 5%, $P(\hat{\delta}_m^{(l)} \leq \delta_m \leq \hat{\delta}_m^{(u)}, \forall m = 1, ..., M) = 0.95$. Corresponding hypotheses tests for a hypothetical parameter δ_{m0}, $H_0 : \cap_{m=1}^{M}\delta_m = \delta_{m0}$ vs. $H_1 : \cup_{m=1}^{M}\delta_m \neq \delta_{m0}$ can be rejected if at least one of the hypothesized parameters is excluded by the corresponding lower or upper bounds, $\hat{\delta}_m^{(l)} > \delta_{m0}$ or $\delta_{m0} > \hat{\delta}_m^{(u)}$ for at least one m. For such tests, the familywise error rate (FWER) is controlled in the strong sense (Hothorn et al., 2008), that is, the probability

of erroneously excluding at least one of the true hypothesized parameters is $= 0.05$, irrespective of which of the remaining δ_{m0} are true.

2.2 Differences on the link scale of generalized linear models

Asymptotically, the above methodology can be applied to the scale of the linear predictor in generalized linear models. Consider the systematic part of a generalized linear model,

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{\theta}$ is parameterized as above, and $g()$ is the link function. Relying on the asymptotic normality of $\hat{\boldsymbol{\theta}}$ (McCulloch and Searle 2001; Hothorn et al. 2008), the methods described in Section 2.1 can be applied as well with the exception that a quantile $z_{M,0.95,\hat{\mathbf{R}}}$ will be taken from the multivariate normal distribution with dimension M, correlation matrix $\hat{\mathbf{R}}$. The resulting intervals are constructed for differences on the scale of the linear predictor, $\boldsymbol{\eta}$.

2.3 Multiple ratios in the general linear model

In the general linear model in equation (1), treatment effects may be expressed in terms of ratios instead of differences (Zerbe 1978; Young et al. 1997; Djira 2010). Their methods are briefly reviewed in the following: The parameters of interest are M ratios $\gamma_m = (\boldsymbol{c}_m\boldsymbol{\theta}) / (\boldsymbol{d}_m\boldsymbol{\theta})$, $m = 1, ..., M$. The known coefficients in the vectors $\boldsymbol{c}_m = (c_{m1}, ..., c_{mP})$ and $\boldsymbol{d}_m = (d_{m1}, ..., d_{mP})$ define which linear combinations of $\boldsymbol{\theta}$ are to be compared in the mth ratio. They are summarized in the two $(M \times P)$ matrices $\mathbf{C}$ and $\mathbf{D}$, with elements c_{mp} and d_{mp}, respectively. To construct simultaneous confidence intervals for $\gamma_1, ..., \gamma_M$, consider $W_m = (\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\hat{\boldsymbol{\theta}}$. The joint distribution of W_m is M-variate normal with covariance matrix $\mathbf{U}$, with elements $u_{mm'}$ given in equation (3):

$$u_{mm'} = Cov\,(W_m, W_{m'}) = (\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\boldsymbol{\Sigma}\,(\boldsymbol{c}_{m'} - \gamma_{m'}\boldsymbol{d}_{m'})^T. \tag{3}$$

$\mathbf{U}$ depends on the unknown ratios, an estimate, $\hat{\mathbf{U}}$, can be obtained by evaluating equation (3) at the estimates $\hat{\gamma}_m = \boldsymbol{c}_m\hat{\boldsymbol{\theta}}/\boldsymbol{d}_m\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ (Dilba et al., 2006; Djira, 2010). The corresponding correlation matrix $\hat{\mathbf{R}}$ can be obtained by standardizing $\hat{\mathbf{U}}$ by its diagonal elements. That is, the elements $\hat{\rho}_{mm'}$ of $\hat{\mathbf{R}}$ are then: $\hat{\rho}_{mm'} = \hat{u}_{mm'}\hat{u}_{mm}^{-1/2}\hat{u}_{m'm'}^{-1/2}$. Approximate simultaneous 95% Fieller-type confidence intervals can be obtained by solving the corresponding inequalities

$$\frac{\left[(\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\hat{\boldsymbol{\theta}}\right]^2}{(\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\hat{\boldsymbol{\Sigma}}\,(\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)^T} \leq t^2_{0.95,M,df=N-P,\hat{\mathbf{R}}} \tag{4}$$

for γ_m (Djira, 2010). Note, that the resulting intervals may be unbounded, that is, there might be no solution, or solutions that are not easily interpretable. The method is approximate because the critical value for inverting the test in equation (4), $t^2_{0.95,M,df=N-P,\hat{\mathbf{R}}}$ depends on the unknown parameters of interest via the plug-in of the estimates $\hat{\gamma}_m$ to obtain the correlation matrix, $\hat{\mathbf{R}}$. These methods are implemented in the function `gsci.ratio` in the `R`–package `mratios` (Djira et al. 2011).

2.4 Simultaneous confidence bands over a grid of covariate values: multiple differences between treatments

The above methods can be applied to compare multiple treatments over a grid of covariate values. What remains is to formulate $\mathbf{C}$ for a given model parameter $\boldsymbol{\theta}$ such that $\boldsymbol{\delta}$ defines the comparison of model predictions between treatments for a number of different values of the covariate x. This involves to consider how treatment and treatment-covariate interaction are parameterized in $\boldsymbol{\theta}$, the definition of a set of covariate values, and the definition of the type of treatment comparisons of interest. As a simple introduction, denote the index of I treatments with $i = 1, ..., I$, and denote $j = 1, .., J_i$ as the index of replications of treatment i, such that an experimental unit is identified by ij. The observed values of the covariate and dependent variable in unit ij are denoted x_{ij} and y_{ij}, respectively, and the model (Equation 5) involves treatment specific intercepts α_i and slopes β_i,

$$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \tag{5}$$

where the parameter vector first contains the I intercepts followed by the I slopes, $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_I, \beta_1, ..., \beta_I)^T$. Denote by Q the number of positions of x for which the treatment specific regression lines should be compared, and the actual values by $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_Q)$, with index $q = 1, ..., Q$. Lastly, let $\mathbf{A}$ define a $(K \times I)$ matrix where the rows $k = 1, ..., K$ define the K comparisons of interest between the I treatments. If the parameters in $\boldsymbol{\delta}$ should be interpretable as differences of (weighted arithmetic means) of the treatment specific regression lines for the covariate positions $\tilde{\boldsymbol{x}}$, the coefficients a_{ki} should be defined under the constraints $\sum_{i=1}^{I} a_{ki} = 0$ and $\sum_{i:a_{ki}>0} a_{ki} = 1$ for each row $k = 1, ..., K$. The $M = QK$ comparisons of interest can then shortly be written as

$$\mathbf{C} = \left(\mathbf{1}_Q \ \tilde{\boldsymbol{x}} \right) \otimes \mathbf{A}, \tag{6}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{1}_Q$ denotes a column vector of 1s of length Q. As an illustration, consider a case with $Q = 4$ covariate values of interest, $\tilde{\boldsymbol{x}} = (5, 10, 15, 20)^T$ and the $K = 2$ comparisons to the control group $(i = 1)$ when there are $I = 3$ treatment groups:

$$\mathbf{C} = \begin{pmatrix} 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \end{pmatrix} \otimes \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & -5 & 5 & 0 \\ -1 & 0 & 1 & -5 & 0 & 5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 1 & 0 & -20 & 20 & 0 \\ -1 & 0 & 1 & -20 & 0 & 20 \end{pmatrix}. \tag{7}$$

2.5 Multiple ratios and odds ratios in generalized linear models

In the important case of dichotomous observations, modeled in a generalized linear model with the binomial distribution (or related assumptions) and the canonical logit link, the resulting confidence bounds can be transformed by the exp function and can then be interpreted as intervals for odds ratios between the predicted treatment specific odds at $\tilde{x}_1, ..., \tilde{x}_Q$. Similarly, for count data modeled with the Poisson distribution or related assumptions and the canonical log link, the exp transformation of the confidence bounds leads to confidence bounds for ratios of means between the treatments at $\tilde{x}_1, ..., \tilde{x}_Q$.

2.6 Multiple ratios of model predictions in the general linear model

The matrices of coefficients for the numerator and denominator, $\mathbf{C}$ and $\mathbf{D}$ can be defined in a similar way as described for the difference in Section 2.4. For the model in equation (5) with the parameterization $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_I, \beta_1, ..., \beta_I)^T$, the K ratios among the I treatments can be defined in two $(K \times I)$ matrices $\mathbf{A}$ and $\mathbf{B}$, for the numerator and the denominator, respectively. The $M = QK$ ratios of interest for Q positions of x can then shortly be written as

$$\mathbf{C} = \left(\mathbf{1}_Q \ \tilde{\boldsymbol{x}} \right) \otimes \mathbf{A}, \text{ and } \mathbf{D} = \left(\mathbf{1}_Q \ \tilde{\boldsymbol{x}} \right) \otimes \mathbf{B}. \tag{8}$$

An illustration for a slightly more complicated model is given in Example 4.2.

2.7 Simulation study

Clearly, the above methods will ensure a simultaneous coverage probability only for the pre-specified values $(\tilde{x}_1, ..., \tilde{x}_Q)$ (Figure 1 left panel). For interpolations between adjacent confidence limits of adjacent discrete values $(\tilde{x}_1, ..., \tilde{x}_Q)$ and subsequent interpretations for any possible value of x in the given range, a simultaneous coverage probability of the nominal level 0.95 is not generally ensured. Also, in practice, the number and position of the covariate values may not be clear a priori, raising the question, which and how many values $\tilde{x}_1, ..., \tilde{x}_Q$ to consider. In many practical cases, this problem is no severe restriction, because the number of values Q can be chosen so large, that adding more values within a given range of x does not lead to further increase of the quantile of the multivariate distribution and hence has no effect on the width

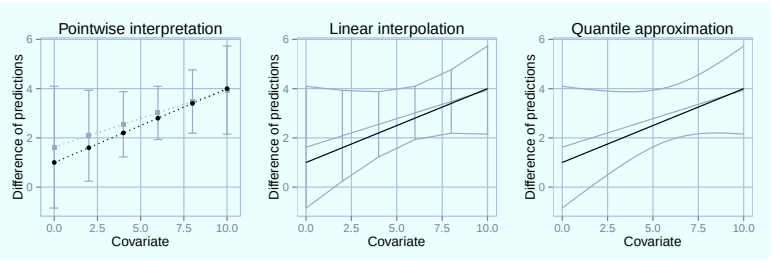


Fig. 1 Point wise interpretation at $Q = 6$ grid points, linear interpolation between adjacent confidence limits in the covariate range $[0; 10]$ and construction of a confidence band over that fixed range, by using the approximate quantile for only six points. Black lines and symbols show the true difference between two predicted lines, gray shows model estimates and corresponding confidence limits

of the confidence limits. For simplicity, consider equidistant values $(\tilde{x}_1, ..., \tilde{x}_Q)$ bounded in a pre-specified range: Increasing the number Q adds estimators for parameters which are highly correlated to estimators already in the set. Adding such values to the set does not change the magnitude of the quantile anymore and the interpretation at many discrete values is very close to an interpretation for any possible value of x in a given range. However, if the number of contrasts between treatments, K, is large, such an approach is limited by the computational limitation $KQ = M < 1000$ in the package `mvtnorm`.

It is therefore tempting to perform the computations for a limited number of covariate values, Q, and then construct simultaneous confidence bands for the whole line in a pre-specified range of a covariate: First, adjacent confidence limits for a given between-treatment comparisons may be joined by lines (Figure 1, middle), this will be referred to as linear interpolation. Second, one may use the quantile of the multivariate t or normal distribution that has been computed for a limited number (Q) of covariate values that span the pre-specified covariate range of interest. This 'approximated' quantile can be used for computing a smooth confidence band over the range of interest (Figure 1, right) this will be referred to as quantile approximation. If the true functions of interest are linear (e.g. differences between treatment specific regression lines in model Equation (5) or example 4.1), it is clear from Figure 1, that the linear interpolation will be a simultaneous confidence band which has slightly too much content and might be slightly conservative, if too few covariate values are used. If one uses the quantile approximation to construct a confidence band over the pre-specified range of the covariate, using too few covariate values when computing the quantile will yield a liberal confidence band. However, for many grid points, Q, that the simultaneous coverage probability of both approaches should be close to the nominal simultaneous confidence, as suggested for discrete confidence bands for the difference of two treatments (Bretz et al., 2010).

A simulation study has been performed to illustrate that with increasing Q, practically valid simultaneous confidence bands can be constructed: The model in Equation (5), has been used to simulate data with x_{ij} sampled from the uniform distribution, with number of treatment groups $I = 3$ or 6, sample sizes of $n_i = 5, 10, 20$ or 100 per treatment group, and parameter configurations involving intercepts and slopes equal, either slopes or intercepts differing between treatments or both intercepts and slope differing between treatments. For each simulated data sets, the methods described above haven been applied for comparisons to control (referred to as Dunnett), all pairwise comparisons (referred to as Tukey) and comparisons of each treatment to the average of treatments (referred to as GrandMean), combined with a set of $Q = 3, 6, 10$, or 20 equidistant grid points spanning the pre-specified covariate range. For each combination of parameter setting and each method, 5 000 data sets have been simulated such that the estimated simultaneous coverage probability for an exact 0.95 simultaneous confidence set can be expected to fall within $[0.944; 0.956]$ with a probability 0.95.

More complications arise if the treatment difference of interest is not a linear function depending on the covariate, for example, when the model involves a treatment-interaction with a quadratic term, as in Example 4.2. In this situation, the point wise interpretation of between treatment differences is still exact, whereas it is obviously unwise to use the linear interpolation with only few covariate values. In this case, the quantile approximation can be supposed to be the better choice to approximate confidence bands. Yet more complications arise when using the ratio approach described in section 2.3: Even for the point wise interpretation, the small sample performance is not clear because the method involves the plug-in of an estimated correlation matrix that depends on the estimated ratios of interest. For this reason, the ratio approach has been simulated for model (5) and the parameter and sample size settings described above. Moreover, a model involving treatment-specific intercepts, slopes, and quadratic terms, $y_{ij} = \alpha_i + \beta_{1i}x_{ij} + \beta_{2i}x_{ij}^2 + e_{ij}$ has been simulated for the sample size settings described above. The parameter settings involved cases without any treatment effect, as well as treatment interactions w.r.t to the linear and/or the quadratic term. For $Q = 3, 6, 10$, and 20, ratios (middle row of Figure 3) and differences (lower row of Figure 3) between model predictions over a covariate grid have been considered. The full details of the simulation settings are provided as supplementary material, part A, which is also available from the GitHub repository.

2.8 Software

The methods can be applied in `R` (R Core Team, 2014) with a few lines of code using basic functionality of `R` and the add-on packages `mvtnorm` (Genz et al. 2011), `multcomp` (Hothorn et al. 2008) and `mratios` (Djira et al. 2011). The code for applying the above methods will involve the model fit, the definition of the treatment contrasts of interest and the grid of covariate values, their

combination by the Kronecker product, and the computation of simultaneous confidence intervals. For the figures, the R package `ggplot2` (Wickham 2009) has been used. The R code for the examples shown below is provided as supplementary material part B.

For even simpler application, the R package `statintcov` is provided on the GitHub repository: The special case of a linear model with one treatment factor and interaction to one covariate (Equation 5) is covered in the functions `scitreatcov`, `sciratiotreatcov`, for differences and ratios, respectively. For slightly more general cases, involving generalized linear models, more than one covariate or interactions with quadratic terms as exemplified in Section 4.2, the functions `cmiacov` and `cmratioiacov` can be used to supply linear combinations of the parameters of a fitted model that are suitable for further use in the function `glht` of package `multcomp` or in function `gsci.ratio` of package `mratios`. The R code for the analysis of the examples using this package is provided as supplementary material part C.

3 Results

For the simple model involving only treatment specific intercepts and slopes and inference in terms of differences, simulated simultaneous coverage probabilities are shown in Figure 2: the linear interpolation provides confidence bands with adequate coverage probabilities already for small numbers of covariate values, such as $Q = 3$ or 6, irrespective of the type of treatment contrast, the number of treatment groups or the sample size settings. When using the multivariate t quantile computed for only $Q = 3$ or 6 equidistant values in the covariate range, constructed confidence bands ('quantile approximation') based on that quantile have too low simultaneous coverage probability. However, the simulation settings used here suggest that already multivariate t quantiles computed for $Q = 10$ or 20 covariate values, lead to confidence bands with actual simultaneous coverage probability very close to the nominal.

The comparison of treatment-specific regression lines in terms of ratios (Figure 3, upper row) shows that, the point wise interpretation for a given set of covariate values yields correct simultaneous coverage probability unless being an approximative approach. The attempt to construct confidence bands using only $Q = 3$ or 6 covariate values with either of the two approaches may yield liberal confidence bands, whereas $Q = 20$ covariate values lead to correct confidence bands in all cases considered here. Note, that for small sample sizes and some simulation settings, up to 13% of the simulated data sets yield unbounded confidence sets and thus the methods appears conservative due to the fact that it yields uninformative confidence bands.

In the quadratic model (with three parameters estimated for each treatment group), the point wise interpretation of differences between treatment-specific model predictions has observed simultaneous coverage close to the nominal level for all settings considered (Figure 3, lower row, left panel). For either approach to construct confidence bands using only $Q = 3, 6$ or 10 lead

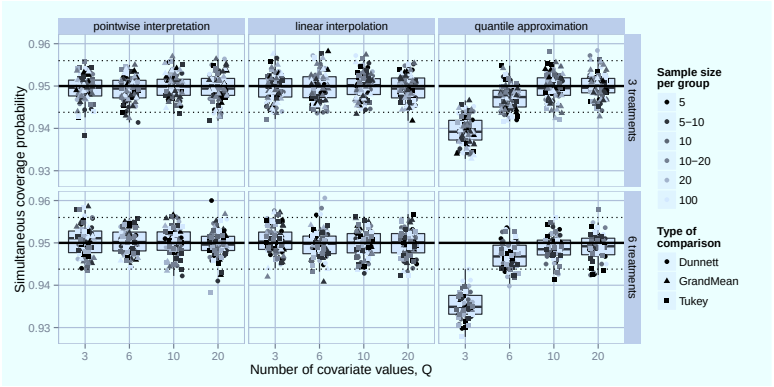


Fig. 2 Simulated simultaneous coverage probabilities (5000 data sets per parameter setting) for the given set of discrete covariate values, confidence bands constructed by linear interpolation, or quantile approximation using a multivariate-t-quantile for approximation using $Q = 3, 6, 10, 20$ equidistant covariate values. Dotted lines show the range in which 95% of the simulation results can be expected for an exact 95% method

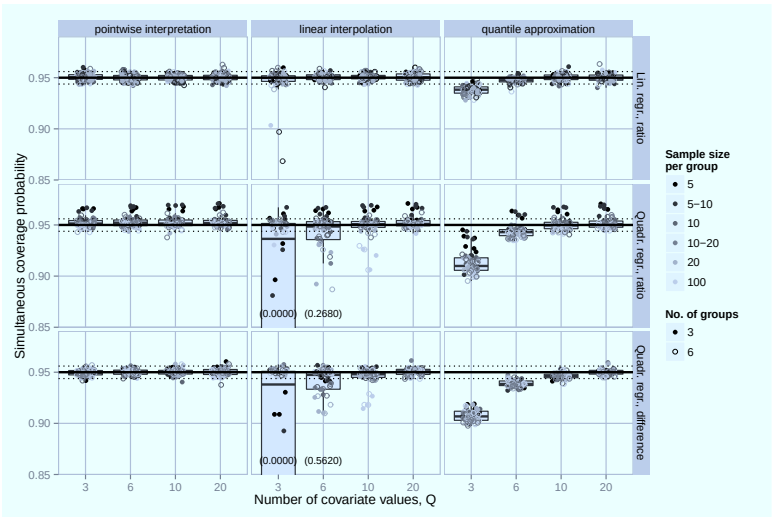


Fig. 3 Simulated simultaneous coverage probabilities (5000 data sets per parameter setting) for the given set of discrete covariate values, confidence bands constructed by linear interpolation, or approximation using a multivariate-t-quantile for $Q = 3, 6, 10, 20$ equidistant covariate values. Dotted lines show the range in which 95% of the simulation results can be expected for an exact 95% method. In four settings, where observed coverage probabilities fell below 0.85, the minimal coverage probability is shown in parentheses

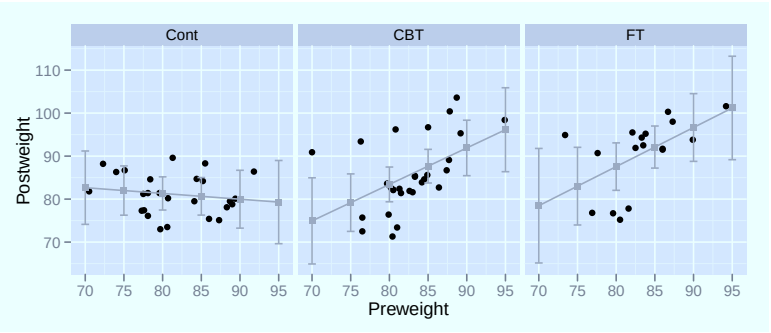


Fig. 4 Observed post-weight and pre-weight, predicted post-weight and confidence intervals for predicted post-weight in three treatments of anorexia and pre-weight six values in the range $[70; 95]$

to severely, or at least slightly too low coverage probabilities. When using the approximate Fieller-type intervals for ratios to compare treatment-specific predictions in the quadratic model with sample sizes as low as 5 or 10 per treatment group, the coverage probabilities appear systematically too high. This is due to the fact that for up to 50% of simulated data sets there was no finite solution for Equation (4).

4 Examples

4.1 All pairwise comparisons with baseline as a covariate

The first data set contains weights (in lbs) of young girls before ('preweight') and after ('postweight') treatment for anorexia (Hand et al., 1994). The first treatment group of 26 girls is the untreated control, the second and third treatment group received a cognitive behavioral treatment (CBT) and family therapy (FT), consisting of 29 and 17 girls, respectively. Analyzing the post-weight in dependency of the treatments, including pre-weight as a possibly interacting covariate, leads to significant main effects for pre-weight and treatment (p=0.0011 and p=0.0004, respectively), as well as to a significant interaction between pre-weight and treatment (p=0.0067) in ANOVA.

Because at least the magnitude of treatment effects depends on the pre-weight value, one may now ask, for which values of pre-weight the treatments differ significantly in post-weight, and if so, by what magnitude. Therefore, the model with is fitted, parametrized as in equation (5), and all pairwise comparisons are specified in the (3×3) matrix $\mathbf{A}$ in equation (9), and $Q = 6$ equidistant pre-weight values are chosen to cover $[70; 95]$. The simultaneous 95% confidence intervals for the resulting $M = 18$ are shown in Figure 5.

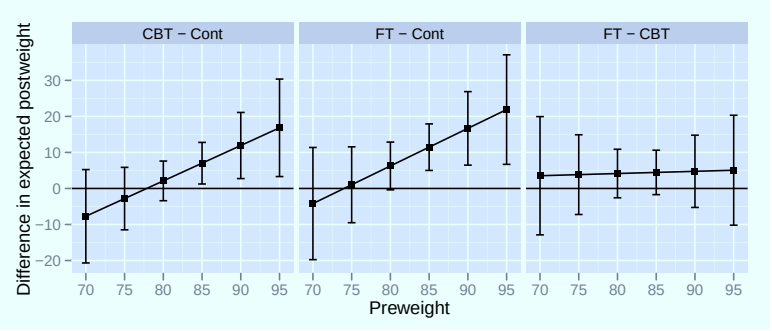


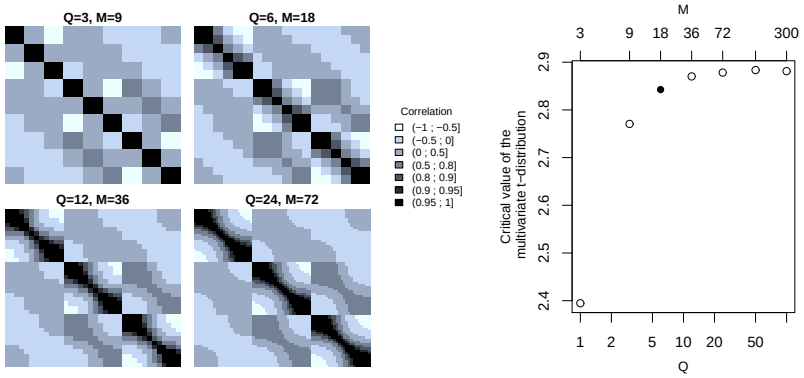
Fig. 5 Simultaneous 95% confidence intervals for all pairwise comparisons between the three anorexia treatment groups at six equidistant values of pre-weight

$$\mathbf{C} = \begin{pmatrix} \text{-1} & 1 & 0 & \text{-70} & 70 & 0 \\ \text{-1} & 0 & 1 & \text{-70} & 0 & 70 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{-1} & 0 & 1 & \text{-95} & 0 & 95 \\ 0 & \text{-1} & 1 & 0 & \text{-95} & 95 \end{pmatrix} = \begin{pmatrix} 1 & 70 \\ 1 & 75 \\ \vdots & \vdots \\ 1 & 95 \end{pmatrix} \otimes \begin{pmatrix} \text{-1} & 1 & 0 \\ \text{-1} & 0 & 1 \\ 0 & \text{-1} & 1 \end{pmatrix}. \tag{9}$$

As can be presumed from Figure 4, the significant interaction between pre-weight and treatment is due to the significant difference in post-weight between the CBT and control as well as FT and control, when pre-weights are 85, 90 and 95 lbs. For none of the six pre-weight values, there is significant difference in expected post-weights between the two treatment groups CBT and FT.

The above intervals are constructed only for interpretations at the chosen $Q = 6$ discrete values of the covariate, $\tilde{\boldsymbol{x}} = (70, 75, 80, 85, 90, 95)$. Figure 6 illustrates the effect of increasing the number of covariate values on the correlation structure (6a), and consequently on the multivariate t-quantiles (6b): The above all pairwise comparison problem is considered for $Q = 3, 6, 12, 24, 50, 100$ equidistant points in $[70; 95]$, resulting in total numbers of parameters of $M = 9, 18, 36, 72, 150, 300$. As a reference point for the critical value, the case $Q = 1$, for all pairwise comparisons at the overall mean of the covariate, $\tilde{x} = \bar{x} = 82.4$ is added.

With $Q = 3$, the correlations between linear combinations with adjacent covariate values for the same treatment contrast, are below 0.5. For $Q = 6$, such linear combinations have already correlations greater than 0.95, when the covariate values are close to the limits of $[70; 95]$, but correlations of 0.5–0.9 for covariate values in the center of the covariate range. Doubling Q from 12 to 24 yields correlations that are always higher than 0.95 for directly adjacent values of x within the same contrast. Figure 6b) shows the quantiles of the multivariate t distribution with $df = 72 - 6$ in dependence of Q, and M. For $Q = 24, 50$ and 100 the quantiles approach 2.88, where the slight changes in



(a) (b)

Fig. 6 Correlation matrices (a) and corresponding multivariate t quantiles (b) for an increasing number Q of equidistant values $\tilde{\boldsymbol{x}}$ in the range $[70; 95]$. The rows and columns of the correlation matrices are ordered primarily by the between-treatment-comparisons, and within each between-treatment-comparison by increasing values of $\tilde{\boldsymbol{x}}$. The entries of the correlation matrices are represented by a gray scale

the values are mainly due to the Monte Carlo error in the computation of the quantiles. For $Q = 6$ the critical value (2.84) is still slightly smaller.

4.2 Treatment interaction with a quadratic regression term

In an experiment discussed by Milliken and Johnson (2002), the yield y_{ij} of a process in dependency of the amount of a substance, x_{ij}, was investigated. The effect of two additives (S1, S2) on that yield is compared to a control group without any additive. Among the $I = 3$ treatment groups, $i = 1$ denotes the control group. (Milliken and Johnson, 2002) assume treatment specific intercepts α_i, an overall linear increase β_1 depending on the substance x_{ij}, as well as treatment specific parameters β_{2i} for the quadratic terms x_{ij}^2 in a general linear model:

$$y_{ij} = \alpha_i + \beta_1 x_{ij} + \beta_{2i} x_{ij}^2 + e_{ij}. \tag{10}$$

The predicted values for the yield y according to the fitted model, as well as the corresponding simultaneous 95% confidence intervals for $Q = 11$ values $\tilde{\boldsymbol{x}} = (0, 1, 2, ..., 10)^T$ as are shown in Figure 7 along with the observations.

The parameter vector is ordered $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_{21}, \beta_{22}, \beta_{23})^T$, similar as in equation (1). Interest is in estimating the gain in expected yield when using one of the two additives compared to running the process without any of the additives ($K = 2$ comparisons to control). Because the yields in the control group are clearly positive (except when substance x is close to 0) one could express the effect of the additives in terms of ratios. That is, expressing the

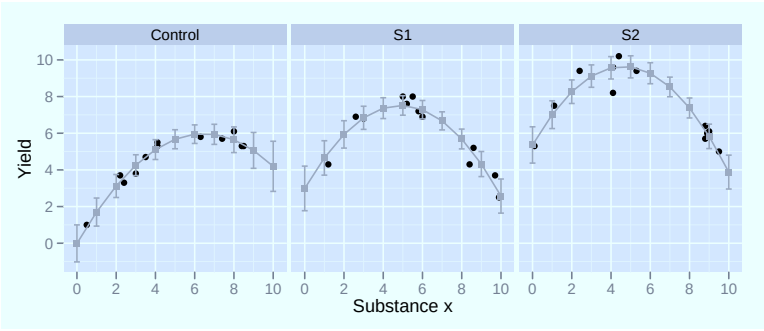


Fig. 7 Observed yield and substance x, predicted yields and confidence intervals for the predicted yield for the process data set (example 4.2)

increase in yield when using additive S1 or S2 as fold–change relative to the yield in the control treatment. Relying on Section 4.3, the matrix $\mathbf{C}$ defines the expected yield of additive S1 and S2 for $\tilde{\boldsymbol{x}}$, and $\mathbf{D}$ defines the yields in the control group for $Q = 9$ values of substance x, $\tilde{\boldsymbol{x}} = (1, 2, ..., 9)$.

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{11}$$

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 9 & 0 & 81 & 0 \\ 0 & 0 & 1 & 9 & 0 & 0 & 81 \end{pmatrix} = \left(\mathbf{1}_{Q\times 1} \otimes \mathbf{A},\, \tilde{\boldsymbol{x}} \otimes \mathbf{1}_{2\times 1},\, \tilde{\boldsymbol{x}}^2 \otimes \mathbf{A} \right), \tag{12}$$

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 9 & 81 & 0 & 0 \\ 1 & 0 & 0 & 9 & 81 & 0 & 0 \end{pmatrix} = \left(\mathbf{1}_{Q\times 1} \otimes \mathbf{B},\, \tilde{\boldsymbol{x}} \otimes \mathbf{1}_{2\times 1},\, \tilde{\boldsymbol{x}}^2 \otimes \mathbf{B} \right). \tag{13}$$

Figure 8 shows that with low concentrations of substance x, the yield is significantly increased with both additives, 1 and 2. For larger concentrations of substance x, the effect of the additives decreases and is not significantly different (at a 5% familywise error rate) for $x = 8, 9$ with additive 1 and $x = 9$ with additive 2. With approximately 95% confidence it can be stated that the mean yield with $x = 1, 2, 3$ using additive S1 is more than 1.96, 1.56, 1.39 times the mean yield in the control. For additive S2 and $x = 1, 2, 3$, the mean yield is at least 2.96, 2.21, 1.89 times that of the control. Increasing the number of points in $\tilde{\boldsymbol{x}}$ from $Q = 10$ ($t_{0.95,M=20,df=N-P=29,\hat{\mathbf{R}}} = 2.8125$) has only small effects on the resulting quantile: for $Q = 20, 40, 80$ equidistant values in the range $[1; 9]$, the corresponding quantiles are 2.8157, 2.8192, 2.8181, respectively.

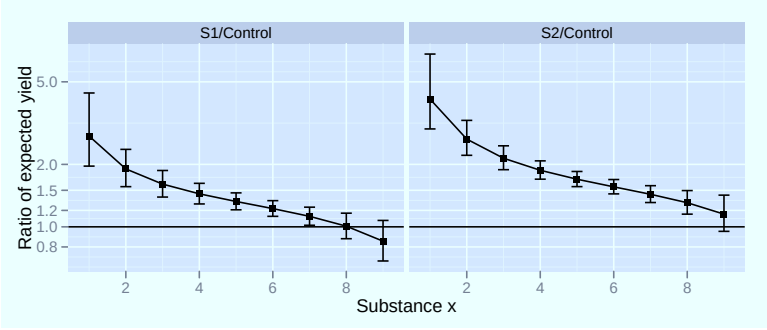


Fig. 8 Simultaneous 95% confidence intervals for the ratios of expected yields between additive S1 and the control as well as additive S2 and the control

4.3 All pairwise comparisons in a binomial generalized linear model

An experiment investigating the mortality of flies exposed to different concentrations of four different compounds containing Selenium is reported in Jeske et al. (2009). In the original publication, the data are analyzed by a generalized linear model assuming the binomial distribution, a probit link with a correction for baseline mortality, and compound specific intercepts and slopes in dependence on the log–concentrations. The data with non-zero concentrations are analyzed here with a simple logit link instead,

$$\begin{aligned} y_{ij} &\sim Bin(n_{ij}, \pi_{ij}), \\ \log\left[\pi_{ij}/\left(1-\pi_{ij}\right)\right] &= \eta_{ij}, \\ \eta_{ij} &= \alpha_i + \beta_i x_{ij}, \end{aligned} \tag{14}$$

where y_{ij} denotes the observed number of dead flies out of n_{ij} flies under observation in the ith compound and dose level j, $j = 1, ..., J_i$. The corresponding unknown mortality is denoted π_{ij}, the linear predictor η_{ij} is modeled with α_i and β_i being the compound specific intercepts and slopes on the logit scale, where x_{ij} are the $\log_{10}$ of the concentrations. Fitting this model and ordering the parameter vector as in Section 2.2, allows to construct asymptotic 95% confidence intervals for the predicted odds at $\log_{10}$-dose levels $\tilde{\boldsymbol{x}} = (0.7, 0.9, 1.1, ..., 2.9)^T$, i.e., $Q = 12$. For this purpose, a (48×8) matrix $\mathbf{C}$ can be constructed by $(\mathbf{1}_Q, \tilde{\boldsymbol{x}}) \otimes \mathbf{A}$, where $\mathbf{A}$ is a (4×4) identity matrix. The confidence intervals for $\mathbf{C}\boldsymbol{\theta}$ are on the scale of the linear predictor and applying the inverse link $\exp(\eta)/[1 + \exp(\eta)]$ on the resulting confidence bounds yields confidence bounds for the predicted mortalities shown in Figure 9. All pairwise comparisons among the four compounds can be performed using the

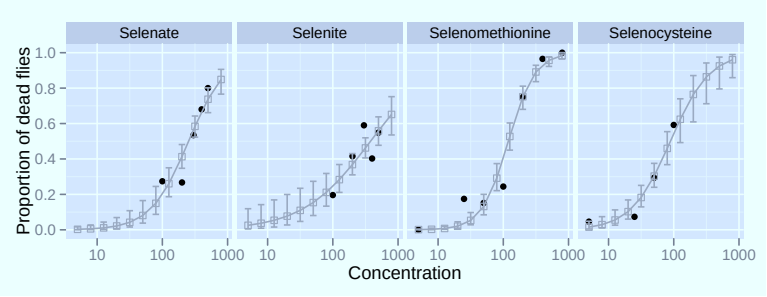


Fig. 9 Observed mortality for the four compounds and asymptotic simultaneous 95% confidence intervals for the predicted mortality based on the fitted model corresponding to equation (14)

matrix $\mathbf{C}$ as defined in equation (15),

$$\mathbf{C} = \begin{pmatrix} 1 & 0.7 \\ 1 & 0.9 \\ 1 & 1.1 \\ \vdots & \vdots \\ 1 & 2.9 \end{pmatrix} \otimes \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{15}$$

Asymptotic confidence limits for ratios among the I compounds with respect to the odds of the probability to die relative to the probability to survive,

$$\frac{\frac{\pi|i,\tilde{x}_q}{(1-\pi|i,\tilde{x}_q)}}{\frac{\pi|i',\tilde{x}_q}{(1-\pi|i',\tilde{x}_q)}}, \text{ for } q = 1, ..., Q, \text{ and all pairs } \{i, i'\} : i \neq i', \tag{16}$$

can be constructed by applying the exp function to the confidence limits for the differences on the scale of the linear predictor defined by $\mathbf{C}\boldsymbol{\theta}$. These intervals are shown in Figure 10.

Figure 9 reveals a number of problems concerning pairwise comparisons among the compounds: the range of concentrations differs among the four compounds, in particular between Selenite and Selenate on the one side and Selenocysteine on the other side, with ranges only overlapping in concentration 100. If one believes in model (14), Figure 10 may lead to the following conclusions: Selenite leads to odds(die/survive) that are roughly 80% that of Selenate for the considered high concentrations ($\tilde{\boldsymbol{x}} > 300$). Selenomethionine leads to increased odds (die/survive) compared to both Selenite and Selenate, for the considered high concentration values ($\tilde{\boldsymbol{x}} > 100$). Most striking is the 5- to 10-fold increase of this odd in Selenomethionine relative to that in Selenite for the considered high concentrations ($\tilde{\boldsymbol{x}} > 300$). Selenocysteine shows an about 2-fold increased odds (die/survive) compared to Selenite for the considered high concentrations ($\tilde{\boldsymbol{x}} > 300$) and also compared to Selenomethionine

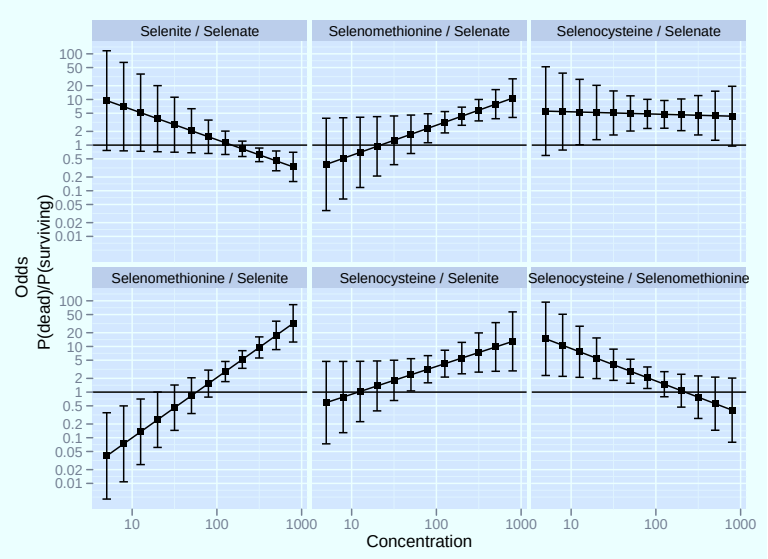


Fig. 10 Asymptotic simultaneous 95% confidence intervals for all pairwise oddsratios among the four Selenium compounds for $Q = 12$ concentration values

but then for the low concentration values ($\tilde{\boldsymbol{x}} < 10$). The two-sided 95% multivariate normal quantiles corresponding to $Q = 6, 12, 24, 48$ and 96 equidistant points in $\{0.7, 2.9\}$ for the given data are $2.9389, 2.9839, 2.9977, 2.9987$ and 2.9975. That is, the intervals on the logit scale would increase in width by about 0.5% if 96 instead of the given 12 values in $\tilde{\boldsymbol{x}}$ would be considered.

5 Discussion

This paper shows how a detailed interpretation of treatment-covariate interactions is possible with standard methods based on simultaneous confidence intervals for user-defined multiple contrast tests in freely available software. Different types of multiple comparisons among several treatments can be interpreted for a pre-specified set of covariate values. The case studies illustrate how to set the methods into practice for a variety of models and experimental questions.

In a strict sense, the simultaneous interpretation is valid only for the pre-specified set of covariate values which have been used for computing the quantile, and not as simultaneous confidence bands, i.e., for any covariate value over the pre-specified range of the covariate. Previously, it has been argued (Bretz et al. 2010; Westfall et al. 2011) that for a sufficiently large set of points that spans a pre-specified range of the covariate, the approach approximates

the corresponding confidence bands. The informal assessment of the correlation structure and the results of simulation studies presented in this paper suggest that already a grid of 20 equidistant points in a given covariate range can be used to construct confidence bands with simultaneous coverage probabilities very close to the nominal level. However, the simulation settings used here are restricted to the general linear model and a well-behaved sampling scheme with the covariate values sampled from the uniform distribution. If model complexity increases, covariates have a skewed distribution or include extreme observations, or the covariate range differs between treatments, the recommendations for the number of covariate values may need further assessment. Also, for the application to generalized linear models an assessment of the small sample performance is needed.

Acknowledgements I thank Prof. L.A. Hothorn and Dr. M. Hasler for their helpful comments on an earlier version of the manuscript. The work was partly supported by the German Science Foundation grant DfG-HO1687.

References

1. Bhargava P, Spurrier J (2004) Exact confidence bounds for comparing two regression lines with a control regression line on a fixed interval. Biom J 46: 720–730.
2. Bretz F, Genz A, Hothorn L (2001) On the numerical availability of multiple comparison procedures. Biom J 43: 645–656.
3. Bretz F, Hothorn T, Westfall P (2010) Multiple Comparisons Using R. Chapman and Hall/CRC, Boca Raton.
4. Dilba G, Bretz F, Guiard V (2006) Simultaneous confidence sets and confidence intervals for multiple ratios. J Statist Plann Inference 136: 2640–2658 .
5. Djira GD (2010) Relative Potency Estimation in Parallel-Line Assays - Method Comparison and Some Extensions. Comm Statist Theory Methods 39: 1180–1189.
6. Djira GD, Hasler M, Gerhard D, Schaarschmidt F (2011) mratios: Inferences for ratios of coefficients in the general linear model, R package version 1.3.16.
7. Genz A, Bretz F (2009) Computation of Multivariate Normal and t Probabilities. Springer, New York.
8. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2011) mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9991.
9. Hand DJ, Daly F, McConway K, Lunn D, Ostrowski E (1994) A Handbook of Small Data Sets. Chapman and Hall/CRC, London.
10. Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50: 346–363.
11. Jamshidian M, Liu W, Bretz F (2010) Simultaneous confidence bands for all contrasts of three or more simple linear regression models over an interval. Comput Statist Data Anal 54: 1475–1483 .
12. Jamshidian M, Liu W, Zhang Y, Jamshidian F (2005) SimReg: A software including some new developments in multiple comparison and simultaneous confidence bands for linear regression models. J Statist Soft 12: 1–22 .
13. Jeske DR, Xu HK, Blessinger T, Jensen P, Trumble J (2009) Testing for the Equality of EC50 Values in the Presence of Unequal Slopes With Application to Toxicity of Selenium Types. J Agric Biol Environ Stat 14: 469–483 .
14. Liu W (2010) Simultaneous Inference in Regression. Chapman & Hall/CRC, Boca Raton.
15. Liu W, Jamshidian M, Zhang Y (2004) Multiple comparison of several linear regression models. J Amer Statist Assoc 99: 395–403 .

16. Lu X, Chen JT (2009) Exact simultaneous confidence segments for all contrast comparisons. J Statist Plann Inference 139: 2816–2822.
17. McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley-Interscience, New York.
18. Milliken G, Johnson D (2002) Analysis of Messy Data, Volume III: Analysis of Covariance. Chapman and Hall/CRC, Boca Raton.
19. R Core Team (2014) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
20. Scheffe H (1959) The Analysis of Variance. Wiley-Interscience, New York.
21. Spurrier J (1999) Exact confidence bounds for all contrasts of three or more regression lines. J Amer Statist Assoc 94: 483–488.
22. Spurrier J (2002) Exact multiple comparisons of three or more regression lines: Pairwise comparisons and comparisons with a control. Biom J 44: 801–812 .
23. Westfall PH, Tobias RD, Wolfinger RD (2011) Multiple Comparisons and Multiple Tests Using SAS, 2nd edition. SAS Institute Inc, Cary.
24. Wickham H. (2009) Elegant Graphics for Data Analysis (Use R). Springer, Dordrecht.
25. Young D, Zerbe G, Hay W (1997) Fieller's theorem, Scheffe simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effects models. Biometrics 53: 838–847.
26. Zerbe G (1978) Fieller Theorem and General Linear-Model. Amer Stat 32: 103–105.

Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions

Frank Schaarschmidt[*,1], **Martin Sill**[2], and **Ludwig A. Hothorn**[1]

[1] Institut für Biostatistik, Leibniz Universität Hannover, Herrenhäuser Str. 2, D-30419 Hannover, Germany

[2] Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received 6 November 2007, revised 19 March 2008, accepted 17 July 2008

Summary

Simultaneous confidence intervals for contrasts of means in a one-way layout with several independent samples are well established for Gaussian distributed data. Procedures addressing different hypotheses are available, such as all pairwise comparisons or comparisons to control, comparison with average, or different tests for order-restricted alternatives. However, if the distribution of the response is not Gaussian, corresponding methods are usually not available or not implemented in software. For the case of comparisons among several binomial proportions, we extended recently proposed confidence interval methods for the difference of two proportions or single contrasts to multiple contrasts by using quantiles of the multivariate normal distribution, taking the correlation into account. The small sample performance of the proposed methods was investigated in simulation studies. The simple adjustment of adding 2 pseudo-observations to each sample estimate leads to reasonable coverage probabilities. The methods are illustrated by the evaluation of real data examples of a clinical trial and a toxicological study. The proposed methods and examples are available in the R package MCPAN.

Key words: Multiple inference; Multivariate normal; Simple adjustment; Small sample.

1 Introduction

Multiple comparison procedures are well established for Gaussian distributed data. In this situation, solutions for a variety of settings are available, e.g., all pairwise comparisons according to Tukey (1953), comparisons to a control group according to Dunnett (1955), and various approaches to test for order-restricted alternatives using multiple contrast tests as described in Bretz (1999, 2006). Moreover, in many practical situations, research questions might be so special that none of these methods appropriately covers the hypotheses of interest. In these cases the estimation of only those particular comparisons which are of interest is more appropriate.

However, for binomial data, only special cases out of this variety of settings are described in the literature. Holford, Walter and Dunnett (1989) describe the construction of large sample simultaneous confidence intervals for comparisons to control for odds-ratios. Piegorsch (1991) describes all pairwise comparisons and comparisons to control for the difference of proportions. In his work, a large sample approximation as well as a method for moderate sample sizes is described and characterized for a number of settings. Bretz and Hothorn (2002) consider trend tests for ordered binomial proportions based on multiple contrast tests with special respect to power calculation, using a large sample approximation.

Often, interest is not in testing the null-hypothesis of zero difference, but in assessing non-inferiority or testing for relevant superiority, or merely in the quantification of an observed treatment effect.

[*] Corresponding author: e-mail: schaarschmidt@biostat.uni-hannover.de, Phone: +49 511 762 5821, Fax: +49 511 762 4966

For this reason, this paper will focus on confidence intervals, serving as a tool for inference in all these situations.

Additionally, sample size is limited in many practical situations; thus, applicability of large sample approximations for binomial proportions is questionable. Exact methods would be the appropriate alternative. However, exact methods are available only for simple cases, as the construction of confidence intervals for a single proportion or the difference of two proportions. These methods are still under controversial discussion because of their conservative performance (Agresti and Coull, 1998; Agresti and Min, 2001), and because of computational problems (Röhmel, 2005).

In contrast, a number of approximate methods for confidence intervals of binomials has been proposed, aiming to have coverage probability close to the nominal level and to be computationally feasible. For the difference of two proportions, Newcombe (1998) provides a comparative simulation study including a newly proposed method. Agresti and Caffo (2000) propose a simple and well performing method by adding one pseudo observation to each cell of the 2×2 table. The use of such simple adjustments is motivated by Agresti and Coull (1998), who simplify, for educational purposes, the confidence interval proposed by Wilson (1927) for a single binomial proportion by adding two pseudo successes and two pseudo failures. The good coverage probabilities of these methods for small and unbalanced sample sizes is shown by Agresti and Coull (1998) and Agresti and Caffo (2000). Brown and Li (2005) compare more recent proposals in a simulation study, recommending Agresti and Caffos approach among two other methods. Finally, Price and Bonett (2004) extend Agresti and Caffos method for estimation of a single linear combination of more than two proportions. They compare two versions of adjustment and provide simulation results for selected settings, but do not provide methods for multiplicity adjustment.

This paper describes the construction of approximate simultaneous confidence intervals for multiple contrasts of binomial proportions. The results of a simulation study for change point contrasts are summarized. Linked to the paper is the add-on package MCPAN for use in the R statistical computing environment (R Development Core Team, 2007).

2 Examples

Eisenberg et al. (2004) investigate the efficacy of Palonosetron application to reduce emesis during chemotherapy. The study comprised five dose groups, 0.3–1.0 μg/kg, 3 μg/kg, 10 μg/kg, 30 μg/kg, 90 μg/kg with moderate sample sizes between 24 and 46. The number of patients with complete response, defined as absence of emesis in the study period, is shown in Table 1.

Eisenberg et al. (2004) treated the lowest dose group as a control, and presented confidence intervals for the difference to this group. However, establishing the presence of a dose-response relationship could be of primary interest. The dose levels are unequally spaced and cover a wide range. The shape of this dose-response relationship for the chosen dose levels is not clear a priori. Hence, it could be suboptimal in terms of power to use methods which are power optimal only in the presence of linear trends, as is the Cochran-Armitage test (Bretz and Hothorn, 2002). Multiple contrasts for order-restricted alternatives according to Bretz and Hothorn (2002) are a powerful option for different shapes of dose-response relationships. When, additionally to a test decision, effect size should be estimated, simultaneous confidence intervals are needed. The approach of Hirotsu and Marumo (2002)

Table 1 Rate and proportion of patients with complete response.

	0.3–1.0 μg/kg	3 μg/kg	10 μg/kg	30 μg/kg	90 μg/kg
Sample size	29	24	25	24	46
Patients with complete response	7	11	10	12	21
Estimated proportion	0.24	0.46	0.40	0.50	0.46

Table 2 Rate and proportion of rats with cancer as presented in Price and Bonett (2004).

	Fiber		No Fiber	
	High Fat	Low Fat	High Fat	Low Fat
Sample size	30	30	30	30
Rats showing cancer	20	14	27	19
Estimated proportion	0.67	0.47	0.90	0.63

to apply change point contrasts for the detection of dose-response relationship is shortly described in Section 3; its application to the data in Table 1 is shown in Section 5.

Cohen et al. (1991) present data of a two-factorial trial, investigating the effect of fiber and fat content in the diet of rats on the development of chemically induced tumors. Table 2 shows the number of rats with tumors in the four treatment groups.

Price and Bonett (2004) estimate confidence intervals for the main effects of fiber and fat, and their interaction separately by three linear combinations of the proportions. However, one could be interested in estimating the three linear combinations simultaneously to preserve the family wise error rate at the 5% level for the complete analysis.

3 Simultaneous Confidence Intervals for Multiple Contrasts

We consider a completely randomized one-way layout with I groups, $i = 1, \cdots, I$, where n_i denotes the number of Bernoulli trials in the i-th group, and Y_i is the number of successes among the n_i trials. The counts Y_i are assumed to be independent binomial random variables $Y_i \sim \text{Bin}(n_i, \pi_i)$, with point estimators $p_i = Y_i/n_i$. The parameters of interest are M linear combinations of the unknown proportions of success π_i:

$$L_m = \sum_{i=1}^{\mathrm{I}} c_{im}\pi_i \,,$$

$m = 1, \ldots, M$, where the coefficients c_{im} are known constants, chosen by the user. We call a vector of coefficients $C_m = (c_{1m}, \cdots, c_{\mathrm{I}m})$ a contrast if the constraint $\sum_{i=1}^{\mathrm{I}} c_{im} = 0$ is fulfilled. Moreover, we will usually define C_m such that $\sum_{i:c_i>0} c_i = \sum_{i:c_i<0} |c_i| = 1$. Then, we can interpret L_m as difference of weighted averages of the π_i. The point estimator for L_m is $\hat{L}_m = \sum_{i=1}^{\mathrm{I}} c_{im}p_i$ and two-sided $(1-\alpha)$-Wald-type intervals for the contrasts can be estimated using Eq. (1):

$$[\hat{L}_m^l; \hat{L}_m^u] = \left[\sum_{i=1}^{\mathrm{I}} c_{im}p_i \pm z\sqrt{\sum_{i=1}^{\mathrm{I}} c_{im}^2 \hat{V}(p_i)}\right] \tag{1}$$

with $\hat{V}(p_i) = p_i(1-p_i)/n_i$, and z is an appropriate critical value. For the univariate problem of estimating a single contrast ($M = 1$), Price and Bonett (2004) use $z = z_{1-\alpha/2}$, based on normal approximation.

Wald intervals for a single linear contrast of binomial proportions are known to keep the $(1-\alpha)$ coverage only for large sample sizes, see, e.g. Price and Bonett (2004). However, limited sample sizes as small as $n_i = 20$ are frequently observed in practice. A simple adjustment can be performed by replacing $\tilde{p}_i = (Y_i + 1)/(n_i + 2)$ for p_i and $\tilde{V}(p_i) = \tilde{p}_i(1-\tilde{p}_i)/(n_i+2)$ for $\hat{V}(p_i)$ in Eq. (1). For a two-sample comparison, this interval is the one proposed by Agresti and Caffo (2000), and will be denoted Add-2 in the following. A less conservative adjustment is achieved by $\tilde{p}_i = (Y_i + 0.5)/(n_i + 1)$ and $\tilde{V}(p_i) = \tilde{p}_i(1-\tilde{p}_i)/(n_i+1)$, in the following denoted as Add-1.

The proposed adjustments to improve small sample performance are not motivated by statistical theory but are determined on a rather heuristic basis. Agresti and Coull (1998) simplified, for educa-

tional purposes, the formula of the score interval (Wilson, 1927) for one binomial proportion. Its simplicity and its good empirical performance motivated a similar adjustment for the difference of two proportions (Agresti and Caffo, 2000). Here, the number of pseudo observations added to each cell of the 2×2 table was found by choosing the one leading to the best coverage probability. Price and Bonett (2004) compared two extensions of these approaches to the estimation of a linear combination of several proportions: Additional to the Add-2 approach they propose replacing $\tilde{p}_i = (Y_i + 2/k)/(n_i + 4/k)$ for p_i and $\tilde{V}(p_i) = \tilde{p}_i(1-\tilde{p}_i)/(n_i + 4/k)$ for $\hat{V}(p_i)$, where k is the number of non-zero coefficients c_{im} in the linear combination of interest. Based on simulation studies for various types of linear combinations, the authors recommend the latter approach, where results are shown for the most common case $I = 3, 4$. However, in simulation studies not shown in this article, we found that the approach favored by Price and Bonett (2004) is more liberal than the Add-2 approach when k becomes large, e.g. $k = 6, 10$. Therefore, their approach is not considered here.

Since our objective is to estimate approximate simultaneous confidence intervals for M contrasts, we have to adjust for multiplicity. I.e., we want to ensure that $P(L_m \in [\hat{L}_m^l; \hat{L}_m^u]; \forall m = 1, \ldots, M) \cong 1 - \alpha$. The adjustment is achieved by using the two-sided equicoordinate critical value $z = z_{M,\boldsymbol{R},1-\alpha}^{\text{two-sided}}$ of an M-variate standard normal distribution such that $P(|\boldsymbol{Z}| \leq z_{M,\boldsymbol{R},1-\alpha}^{\text{two-sided}}) = 1 - \alpha$, where $\boldsymbol{Z}$ is an M-variate standard normal random vector with $(M \times M)$ correlation matrix $\boldsymbol{R}$. One-sided intervals can be obtained by using the critical value $z = z_{M,\boldsymbol{R},1-\alpha}^{\text{one-sided}}$ such that $P(\boldsymbol{Z} \leq z_{M,\boldsymbol{R},1-\alpha}^{\text{one-sided}}) = 1 - \alpha$. Numerically, such quantiles can be obtained from the R-function qmvnorm in the add-on package mvtnorm, introduced by Hothorn, Bretz and Genz (2001). This adjustment takes the number of estimated parameters M as well as the correlation $\boldsymbol{R}$ between them into account. The correlation matrix $\boldsymbol{R}$ depends on the known constants c_{im} and n_i, but additionally on the unknown parameters π_i. Bretz and Hothorn (2002) give the correlation matrix $\boldsymbol{R}$ for the purpose of power calculation explicitly. Elements $\varrho_{mm'}$ of $\boldsymbol{R}$, specifying the correlation between two contrasts m and m', $m \neq m'$ can be computed:

$$\varrho_{mm'} = \frac{\sum_{i=1}^{\mathrm{I}} c_{im} c_{im'} V(p_i)}{\sqrt{\left(\sum_{i=1}^{\mathrm{I}} c_{im}^2 V(p_i)\right)\left(\sum_{i=1}^{\mathrm{I}} c_{im'}^2 V(p_i)\right)}} \tag{2}$$

where $V(p_i) = \pi_i(1-\pi_i)/n_i$. In practical situations, where π_i and therefore $V(p_i)$ are unknown, we use appropriate sample estimates $\hat{V}(p_i)$ to estimate the correlation. Piegorsch (1991) uses the same approach for the special case of comparisons to control.

The contrast coefficients c_{im} are chosen such that they reflect experimental questions. In many cases, standard multiple comparison procedures as all pairwise comparisons or comparisons to a control group achieve this appropriately. For tests on a monotone dose-response relationship one can formulate the global alternative hypothesis: $H_A : \pi_1 \leq \pi_2 \leq \ldots \leq \pi_I$, with at least one inequality strict. Different contrast types have been proposed (Hirotsu and Marumo, 2002; Bretz, 2006). Such tests combine several contrasts, sensitive for different monotone dose-response patterns, in an union-intersection test, taking the correlation between the contrasts into account. The contrast coefficients c_{im} in the different approaches have in common to compare weighted averages of group means between groups of lower order to groups of higher order. One can conclude for the presence of a dose-response relationship, if at least one of the contrasts is significantly larger than zero. For the comparison of I ordered dose groups, change point contrasts according to Hirotsu and Marumo (2002) use $M = I - 1$ contrasts, one for each of the consecutive dose steps. The c_{im} of the mth contrast can be formally defined as $c_{im} = -\frac{n_i}{\sum_{i \leq m} n_i}$ for $i \leq m$ and $c_{im} = \frac{n_i}{\sum_{i > m} n_i}$ for $i > m$. For the simple case of a balanced design with $I = 4$ groups, the contrasts are shown in Table 3.

The application of the change point contrasts to the Palonosetron example is presented in Table 7. However, many other experimental questions may occur in practice that can be expressed as multiple contrasts but are not covered by any standard procedure. A simple example is shown in Table 9.

Table 3 Contrast coefficients c_{im} for change point contrasts in a balanced design with $I = 4$ ordered groups.

Comparison	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$m = 1$	-1	$1/3$	$1/3$	$1/3$
$m = 2$	$-1/2$	$-1/2$	$1/2$	$1/2$
$m = 3$	$-1/3$	$-1/3$	$-1/3$	1

4 Simulation Study

We performed a number of simulation studies to characterize the above proposed confidence intervals for nominal confidence levels of 0.95. We consider the simultaneous coverage probability, defined as $P(L_m \in [\hat{L}_m^l; \hat{L}_m^u t], \forall m = 1, \ldots, M)$, to be the main criterion for recommendation. However, for binomial proportions the coverage probability oscillates in dependence on n_i and π_i (Agresti and Coull, 1998; Agresti and Caffo, 2000; Brown, Cai and DasGupta, 2001; Brown and Li, 2005). Hence, consideration of only a restricted set of parameters can not lead to a general recommendation of a method. Therefore, the simulation studies are divided in two parts.

First, we show results of estimated simultaneous coverage probability, based on 10 000 simulations, for a restricted number of practically reasonable parameter settings (Table 4, 5 and 6). In these tables, we additionally show the simultaneous coverage probability, when no multiplicity adjustment is used but simple quantiles of the univariate standard normal distribution are applied.

In a second step, we performed a more extensive simulation study to explore the dependency of the coverage probability of n_i and π_i with the change point contrast serving as an example. In order to characterize the performance for the whole parameter space, we draw 10 000 samples of $\pi = (\pi_1, \ldots, \pi_I)$ from independent uniform distributions $U(0, 1)$ for the π_is. For each of these settings, the simultaneous coverage probability was estimated based on 10 000 samples drawn independently from binomial distributions Bin (n_i, π_i). To enable computation in a feasible time, we used the true values of π to calculate the correlation, instead of the sample estimates. Hence, these results do not include the variability that is introduced into the methods by using sample estimates for calculation of the quantiles. However, the error that is introduced into the methods by using a multivariate normal approximation for the construction of confidence limits for binomials is still revealed. Sample sizes considered for this part were varied between $n_i = 10$ and $n_i = 100$ in balanced designs with $I = 3, 4, 6$ and 10 groups. Following the philosophy of Agresti and Coull (1998), and Agresti and Caffo (2000), we consider intervals as appropriate which have coverage probability close to the nominal confidence level, but not necessarily equal or above the nominal level for all parameter settings. The results of the second part are summarized in Section 4.2, detailed tables are available from the first author upon request.

Our final recommendation favors the method, which shows highest proportions of settings with coverage probability between 0.94 and 0.96 and a mean coverage close to the nominal level.

4.1 Results for selected parameter settings

Tables 4 and 5 show simultaneous coverage probabilities for lower and upper 95% confidence limits for change point contrasts of $I = 4$ groups and balanced sample size of $n_i = 40$. Table 6 shows results for two-sided 95% confidence intervals for the main effects and interactions in a balanced 2×2 layout with $n_i = 20$. The methods for construction of simultaneous confidence intervals using multivariate normal quantiles, as proposed insection 3, are compared to marginal confidence intervals, i.e., intervals which are not adjusted for multiple comparisons and are constructed using univariate normal quantiles.

Table 4 Estimated simultaneous coverage probability (based on 10 000 simulations) of lower 95% confidence limits with balanced sample size $n_i = 40$ for change point contrasts of four groups.

proportions				simultaneous intervals			marginal intervals		
π_1	π_2	π_3	π_4	Add-2	Add-1	Wald	Add-2	Add-1	Wald
0.01	0.01	0.01	0.01	1.000	1.000	0.789	0.999	0.994	0.721
0.01	0.01	0.05	0.05	0.997	0.993	0.980	0.981	0.977	0.938
0.01	0.01	0.05	0.10	0.994	0.989	0.975	0.980	0.961	0.943
0.01	0.05	0.10	0.20	0.985	0.979	0.962	0.960	0.936	0.911
0.20	0.20	0.20	0.20	0.954	0.944	0.939	0.887	0.883	0.867
0.20	0.20	0.30	0.30	0.957	0.948	0.937	0.893	0.880	0.868
0.20	0.20	0.30	0.40	0.955	0.946	0.933	0.898	0.883	0.868
0.20	0.30	0.40	0.50	0.957	0.947	0.934	0.898	0.875	0.858
0.50	0.50	0.50	0.50	0.951	0.946	0.942	0.887	0.878	0.875
0.50	0.50	0.60	0.60	0.953	0.943	0.938	0.887	0.880	0.868
0.50	0.50	0.60	0.80	0.957	0.945	0.930	0.897	0.875	0.854
0.50	0.90	0.90	0.90	0.963	0.943	0.921	0.916	0.890	0.857

When all the proportions are close to 0 (or close to 1), both the Add-1 and Add-2 result in conservative intervals, while the Wald limits perform severely liberal for some settings. When proportions become closer to 0.5, all methods have coverage probability closer to the nominal level. The Add-2 method is more conservative, while the Wald interval is liberal for nearly all settings considered. When intervals are not adjusted for multiplicity but use simple standard normal quantiles, the probability to exclude at least one of the true parameters increases severely. Then, the simultaneous coverage probability varies between 0.55 and 0.94 for the Wald, 0.88 and 0.99 for the Add-1 and Add-2 intervals for the considered settings. For situations with $M > 3$ and contrasts with low correlation, the liberality of unadjusted methods becomes more severe (results not shown).

Table 5 Estimated simultaneous coverage probability (based on 10 000 simulations) of upper 95% confidence limits with balanced sample size $n_i = 40$ for change point contrasts of four groups.

proportions				simultaneous intervals			marginal intervals		
π_1	π_2	π_3	π_4	Add-2	Add-1	Wald	Add-2	Add-1	Wald
0.05	0.05	0.05	0.05	0.986	0.959	0.889	0.924	0.896	0.820
0.05	0.05	0.01	0.01	0.998	0.994	0.981	0.981	0.977	0.939
0.05	0.05	0.02	0.01	0.998	0.995	0.978	0.984	0.965	0.930
0.05	0.04	0.03	0.02	0.997	0.990	0.957	0.978	0.960	0.885
0.20	0.20	0.20	0.20	0.954	0.946	0.939	0.887	0.883	0.868
0.20	0.20	0.10	0.10	0.963	0.952	0.934	0.914	0.893	0.870
0.20	0.20	0.10	0.05	0.977	0.963	0.940	0.932	0.910	0.867
0.20	0.15	0.10	0.05	0.976	0.964	0.940	0.932	0.906	0.873
0.50	0.50	0.50	0.50	0.953	0.948	0.943	0.887	0.878	0.876
0.50	0.50	0.40	0.40	0.955	0.947	0.940	0.891	0.881	0.868
0.50	0.50	0.40	0.20	0.958	0.946	0.932	0.902	0.879	0.856
0.50	0.40	0.30	0.20	0.957	0.948	0.933	0.903	0.876	0.859
0.50	0.10	0.10	0.10	0.964	0.946	0.924	0.920	0.895	0.857

4.2 Simulations for the whole parameters space

The results of the second part of the simulation study, performed for samples of π from the parameter space, can be summarized as follows: The Add-2 interval can be recommended as a mostly conservative choice for moderate sample sizes, such as $n_i > 40$ and two-sided problems. When sample sizes are as small as $n_i = 10$ or 20, for only 30–40% of the settings an acceptable coverage probability is achieved, with average coverage probabilities between 0.955 and 0.96. For a small number of settings comparing intermediate to extreme proportions we observed liberal performance in two-sided application. In one-sided application with small sample sizes and some proportions close to 0 and others close to 1, the lower and upper bound can be severely asymmetric in the probability to exclude the true parameter. Then, the upper bounds are liberal when differences $\pi_i - \pi_{i'}$ close to 1 are involved and are conservative when differences $\pi_i - \pi_{i'}$ close to -1 are involved. For lower bounds the situation is inverse.

If applied as a one-sided confidence limit, the Add-1 interval is closer to the nominal level than the Add-2 approach for the price of showing liberal performance for a number of settings with sample sizes $n_i = 10, 20$. For samples sizes $n_i = 40, 60, 100$, the Add-1 method attains high proportions of acceptable coverage probabilities faster than the Add-2 method does. Similar to the Add-2 interval, in one-sided application with small sample sizes the lower and upper bound can be asymmetric in the probability to exclude the true parameter. The pattern is the same as for the Add-2 interval. As expected, the Wald interval is severely liberal for small sample sizes and is still liberal for sample sizes as large as $n_i = 100$ per group. For small sample sizes $n_i = 10, 20$ the coverage probability is in the range [0.94; 0.96] for hardly any parameter combination, and mean coverage probability ranges between 0.81 and 0.93. All methods in common perform slightly worse when many groups are compared, i.e., $I = 6, 10$. This effect is most pronounced for the Wald method.

5 Evaluation of Examples

In the Palonosetron example in Table 1, establishing a dose-response relationship could be of primary interest. Since no negative control exists, and the dose-response shape is not known a priori, we apply change point contrasts. The resulting coefficients c_{im} for the $M = 4$ contrasts are shown in Table 7. In

Table 6 Estimated simultaneous coverage probability (based on 10 000 simulations) of two-sided 95% confidence intervals for main effects and their interaction in a 2×2 design with balanced sample size $n_i = 20$.

proportions				simultaneous intervals			marginal intervals		
π_1	π_2	π_3	π_4	Add-2	Add-1	Wald	Add-2	Add-1	Wald
0.01	0.01	0.01	0.01	1.000	1.000	0.550	1.000	1.000	0.547
0.01	0.01	0.10	0.10	0.997	0.977	0.891	0.970	0.952	0.845
0.01	0.10	0.01	0.10	0.998	0.978	0.892	0.973	0.953	0.843
0.01	0.10	0.10	0.10	0.991	0.979	0.909	0.954	0.907	0.825
0.10	0.10	0.10	0.10	0.989	0.976	0.939	0.949	0.910	0.831
0.10	0.10	0.20	0.20	0.974	0.957	0.931	0.912	0.883	0.833
0.10	0.20	0.10	0.20	0.975	0.960	0.935	0.914	0.885	0.834
0.10	0.20	0.20	0.20	0.970	0.956	0.930	0.899	0.874	0.831
0.50	0.50	0.50	0.50	0.944	0.944	0.939	0.838	0.838	0.838
0.50	0.50	0.40	0.40	0.948	0.946	0.933	0.850	0.844	0.840
0.50	0.60	0.50	0.60	0.945	0.943	0.928	0.848	0.840	0.838
0.50	0.40	0.60	0.40	0.950	0.946	0.929	0.861	0.849	0.842

Table 7 Contrast coefficients for change point contrasts in the Palonosetron example (Table 1).

Comparison	m	0.3–1.0 μg/kg	3 μg/kg	10 μg/kg	30 μg/kg	90 μg/kg
(2, 3, 4, 5) − 1	1	−1.00	0.20	0.21	0.20	0.39
(3, 4, 5) − (1,2)	2	−0.55	−0.45	0.26	0.25	0.48
(4, 5) − (1, 2, 3)	3	−0.37	−0.31	−0.32	0.34	0.66
(5) − (1, 2, 3, 4)	4	−0.28	−0.24	−0.25	−0.24	1.00

Table 8 Simultaneous lower 95% Add-2, Add-1 and Wald confidence limits for the contrasts in Table 7 and the data in Table 1.

Comparison	Estimate	Add-2	Add-1	Wald
(2, 3, 4, 5) − 1	0.212	0.006	0.011	0.017
(3, 4, 5) − (1,2)	0.113	−0.064	−0.062	−0.061
(4, 5) − (1, 2, 3)	0.113	−0.062	−0.060	−0.058
(5) − (1, 2, 3, 4)	0.064	−0.123	−0.123	−0.122

Table 8, 95% lower Add-2, Add-1 and Wald confidence limits are shown. The Add-1 and Wald confidence limits are slightly narrower than the Add-2 method. Due to the small sample size, we prefer the conservative Add-2 confidence limits for interpretation.

Based on the Add-2 limits one can conclude with error probability 0.05 that the average proportion of patients with complete response in the dose groups 3 μg, 10 μg, 30 μg, 90 μg is larger than that in the lowest dose group. With 95% confidence, the average proportion in the four higher dose groups is at least 0.006 above that in the lowest dose group. Although we can not infer this with error probability 0.05, the observed pattern of proportions suggests an increase of efficacy when dosage is increased from 0.3–1.0 μg/kg to 3 μg/kg and a plateau for higher doses.

For the second Example in Table 2 we apply the same contrasts as used by Price and Bonett (2004). The 2×2 design is treated as a pseudo one way layout comprising four treatments. The first and second contrasts are differences in average proportions between *Fiber* and *No Fiber*, and *High Fat* and *Low Fat* treatments, respectively. The third contrast is intended to explore interaction, and can be interpreted as the difference of two differences: $(\pi_{\text{HighFat,Fiber}} - \pi_{\text{LowFat,Fiber}}) - (\pi_{\text{HighFat,NoFiber}} - \pi_{\text{LowFat,NoFiber}})$. One can conclude for the presence of an interaction if the difference between fat levels differs significantly depending on fiber level. Table 10 shows two-sided 95% Add-2 confidence intervals for the three effects. For comparison, univariate Wald intervals are shown which use the critical values of the standard normal distribution. These intervals are much more narrow. However, the simulation results in Section 4 indicate that conclusions based on the univariate Wald intervals might be overoptimistic.

Based on the Add-2 confidence interval in Table 10, we can not conclude for an interaction which is significant at the 5% level. However, the presence of a relevant interaction between fat and fiber in

Table 9 Contrast coefficients for example 2.

Comparison	m	High Fat, Fiber	Low Fat, Fiber	High Fat, No Fiber	Low Fat, No Fiber
Fiber − No Fiber	1	0.5	0.5	−0.5	−0.5
Low Fat − High Fat	2	−0.5	0.5	−0.5	0.5
Interaction Fat: Fiber	3	1.0	−1.0	−1.0	1.0

Table 10 Two-sided simultaneous 95% Add-2 and univariate Wald confidence intervals for the contrasts in Table 9 and the data set in Table 2.

Comparison	Estimate	Simultaneous Add-2		Univariate Wald	
		Lower	Upper	Lower	Upper
Fiber − No Fiber	−0.200	−0.378	0.003	−0.359	−0.041
Low Fat − High Fat	−0.233	−0.409	−0.028	−0.393	−0.074
Interaction Fat: Fiber	−0.067	−0.444	0.319	−0.385	0.252

the diet can not be ruled out: the confidence interval ranges from −0.44 to 0.32. Further, with 95% confidence we can state that the average proportion of tumor bearing animals in the two fiber treatments could be up to 0.38 below that of the treatments without fiber, but, as indicated by the upper bound, it can not be ruled out that the proportions are about equal and that there is no effect of fiber in the diet. The average proportion of tumor bearing animals in the two treatments with high fat diet differs from that in the low fat treatments. With 95% confidence one can state that the high fat diets in average lead to an increase of the proportion of tumor bearing animals of at least 0.028. Summarizing, when controlling the family wise error rate is objective of the statistical analysis, a significant effect can only be shown for fat. However, the wide confidence intervals illustrate the notable uncertainty of the estimated effects which is due to the fairly low sample size. Alternatively, if the hypotheses corresponding to the three contrasts can be stated in an a priori order, stepwise approaches might be applied instead of simultaneous testing in a single step.

6 Discussion

In this paper, we show the availability of approximate simultaneous confidence intervals for multiple contrasts of binomial proportions, based on the multivariate normal distribution. The formulation of hypotheses in terms of multiple contrasts permits all pairwise comparisons, comparisons to control (Schaarschmidt, Biesheuvel and Hothorn, accepted) or user defined comparisons, as well as tests for monotone trends among ordered proportions to be performed.

Appropriate adjustments, being extensions of methods proposed for two-sample comparisons, result in acceptable properties for moderate sample sizes. The small sample adjustments are heuristic, motivated rather by their empirical performance than by statistical theory. Hence, there is potential for improvement. When a method is known to perform better for small samples in univariate situations and can be expressed as an adjustment of the Wald formula, a multiplicity adjustment can be introduced by straightforward extension of the formulas above. In practical application of the methods described here, it should be noted that for group wise sample sizes smaller than 40, the proposed methods can still be problematic due to conservative or liberal performance. The described methods are implemented in the R-package MCPAN, providing a number of preformatted contrast types as well as the possibility to estimate user-defined contrasts.

The methods above might be extended to other problems, if experimental questions can be appropriately expressed as contrasts of the group wise parameters. However, for each case it should be investigated for which sample sizes and parameter settings the approximation performs acceptably. For example, in the evaluation of long-term carcinogenicity studies based on poly-*k*-adjusted tumor rates (Bailer and Portier, 1988) interest can be in comparing mortality-adjusted tumor rates of the dose groups to that of the untreated control or, to test for an increasing trend among such tumor rates (Bieler and Williams, 1993; Peddada, Dinse and Haseman, 2005), where the shape of the dose-response relationship is not known a priori. Schaarschmidt, Sill and Hothorn (submitted) propose the use of multiple contrast tests and simultaneous confidence intervals for such rates, and show accepta-

ble performance for moderate sample sizes such as $n_i = 50$. Recently, Hothorn et al. (2008) described methods for asymptotic simultaneous inference in general parametric models. Their approach covers simultaneous confidence intervals for odds ratios of proportions when applied in generalized linear models (McCulloch and Searle, 2001) with the logit link.

Acknowledgements *We are grateful to the reviewers and guest editors whose comments on an earlier version led to tangible improvements of the article. The work of FS was partially funded by Bundesministerium für Bildung und Forschung, grant number 0313269.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* **54**, 280–288.

Agresti, A. and Coull, A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Agresti, A. and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**, 963–971.

Bailer, J. A. and Portier, C. J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics* **44**, 417–431.

Bieler, G. S. and Williams, R. L. (1993). Ratio estimates, the Delta Method, and quantal response tests for increased carcinogenicity. *Biometrics* **49**, 793–801.

Bretz, F. (1999). Powerful modifications of Williams test on trend. Dissertation, Universität Hannover.

Bretz, F. (2006). An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis* **50**, 1735–1748.

Bretz, F. and Hothorn, L. (2002). Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine* **21**, 3325–3335.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–133.

Brown, L. and Li, X. (2005). Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference* **130**, 359–375.

Cohen, L. E., Kendall, M. E., Zang, E., Meschter, C., and Rose, D. P. (1991). Modulation of N-Nitrosomethylurea-induced mammary tumor promotion by dietary er and fat. *Journal of the National Cancer Institute* **83**, 496–501.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.

Eisenberg, P., MacKintosh, F. R., Ritch, P., Cornett, P. A., and Macciocchi, A. (2004). Efficacy, safety and pharmacokinetics of palonosetron in patients receiving highly emetogenic cisplatin-based chemotherapy: a dose-ranging clinical study. *Annals of Oncology* **15**, 330–337.

Hirotsu, C. and Marumo, K. (2002). Changepoint analysis as a method for isotonic inference. *Scandinavian Journal of Statistics* **9**, 125–138.

Holford, T. R., Walter, S. D., and Dunnett, C. W. (1989). Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *Journal of Clinical Epidemiology* **42**, 427–434.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *R News* **1**(2), 27–29.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal* **50**, 346–363.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. John Wiley and Sons, Inc., New York.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**, 873–890.

Peddada, S. D., Dinse, G. E., and Haseman, J. K. (2005). A survival-adjusted quantal response test for comparing tumor incidence rates. *Applied Statistics* **54**, 51–61.

Piegorsch, W. W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics* **47**, 45–52.

Price, R. M. and Bonett, D. G. (2004). An improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis* **45**, 449–456.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Röhmel, J. (2005). Problems with existing procedures to calculate exact unconditional *p*-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal* **47**, 37–47.

Schaarschmidt, F., Biesheuvel, E., and Hothorn, L. A. (accepted). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportionsin randomized clinical trials. *Journal of Biopharmaceutical Statistics*.

Schaarschmidt, F., Sill, M., and Hothorn, L. A. (submitted). Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test.

Tukey, J. (1953). Multiple comparisons. *Journal of the American Statistical Association* **48**, 624–625.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data

Ralph Scherer[*,1,2,**], **Frank Schaarschmidt**[1,**], **Sabine Prescher**[3], and **Kai U. Priesnitz**[4]

[1] Institute of Biostatistics, Leibniz University Hannover, Herrenhäuserstr. 2, 30419 Hannover, Germany
[2] Institute for Biometry, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
[3] Institute for Biosafety of Genetically Modified Plants, Julius-Kühn-Institut, Messeweg 11-12, 38104 Braunschweig, Germany
[4] Institute for Plant Protection, Bavarian State Research Center for Agriculture, Lange Point 10, 85354 Freising, Germany

Received 2 August 2012; revised 8 January 2013; accepted 9 January 2013

Diversity indices might be used to assess the impact of treatments on the relative abundance patterns in species communities. When several treatments are to be compared, simultaneous confidence intervals for the differences of diversity indices between treatments may be used. The simultaneous confidence interval methods described until now are either constructed or validated under the assumption of the multinomial distribution for the abundance counts. Motivated by four example data sets with background in agricultural and marine ecology, we focus on the situation when available replications show that the count data exhibit extra-multinomial variability. Based on simulated overdispersed count data, we compare previously proposed methods assuming multinomial distribution, a method assuming normal distribution for the replicated observations of the diversity indices and three different bootstrap methods to construct simultaneous confidence intervals for multiple differences of Simpson and Shannon diversity indices. The focus of the simulation study is on comparisons to a control group. The severe failure of asymptotic multinomial methods in overdispersed settings is illustrated. Among the bootstrap methods, the widely known Westfall–Young method performs best for the Simpson index, while for the Shannon index, two methods based on stratified bootstrap and summed count data are preferable. The methods application is illustrated for an example.

Keywords: Bootstrap; Extra-multinomial variability; Overdispersion; Shannon index; Simpson index; Simultaneous coverage probability.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

When assessing the environmental risk associated with human intervention in ecosystems, the impact on biodiversity can be of interest. Here, we consider the case where individuals of multiple species are counted. A number of indices have been proposed for measuring biodiversity by summarizing the number of species as well as the proportions of single species in a given community. We focus on two of these indices: Frequently used and controversially discussed is the Shannon index (Magurran, 2004), which is a special case of a family of measures for entropy (Renyi, 1961). The Simpson index (Simpson,

*Corresponding author: e-mail: scherer.ralph@mh-hannover.de, Phone: +49 511 532 4373, Fax: +49 511 532 4295
**These two authors contributed equally to the paper.

1949), derived from the probability that two randomly chosen individuals are from the same species, gives higher weight to species with high relative abundance, and therefore is also called the dominance index (Magurran, 2004).

Studies may involve more than two interventions, e.g., more than two different treatments in an agricultural field trial. These treatments may be genetically modified plants, near-isogenic plants treated with pesticides and untreated near-isogenic plants as shown in our example data set A. As a consequence, the simultaneous estimation of more than one difference between the treatments can be of interest. The example data sets discussed later on in detail could invoke the following questions: Is there at least one treatment that differs in biodiversity compared to the control treatment? And if so, by what magnitude? Is a single novel treatment noninferior to at least one of several standard treatments? Motivated by such questions, we focus on the construction of simultaneous confidence intervals for multiple differences against a control treatment, while the methods discussed can be easily generalized to other types of comparisons between more than two treatments.

A number of methods for the construction of (simultaneous) confidence intervals for diversity indices have been published: In discussion are asymptotic methods that rely on the multinomial assumption of the data for the estimation of variance parameters (Hutcheson, 1970; Pardo et al., 1997; Fritsch and Hsu, 1999; Rogers and Hsu, 2001; From, 2003; Salicru et al., 2005) or bootstrap methods (Fritsch and Hsu, 1999; Rogers and Hsu, 2001; Pla, 2004). The inferential problems discussed until now include: calculating confidence intervals or testing hypotheses for the diversity index of a single sample (Pardo et al., 1997; Fritsch and Hsu, 1999; Rogers and Hsu, 2001; Pla, 2004), for the difference of two samples (Hutcheson, 1970; Fritsch and Hsu, 1999; Rogers and Hsu, 2001) or testing the equality of diversity among several samples by omnibus tests (Pardo et al., 1997; Salicru et al., 2005). Finally, the estimation of simultaneous confidence intervals for pairwise comparisons among several samples has also been considered (Fritsch and Hsu, 1999; Rogers and Hsu, 2001; Salicru et al., 2005). The common assumption in the latter three publications is, tacitly or explicitly, that the counts follow a multinomial distribution, either for the purpose of deriving variance estimators or for the purpose of validating the proposed methods in simulation studies. However, this is a very strict assumption which might be inappropriate for ecological data sets in many practical applications.

For counts of individuals of a single species, it is well accepted that overdispersion (extra-Poisson variability) is a frequent feature of observed count data. That is, the observed variability of species counts is considerably higher than would be expected under the basic assumption of a Poisson distribution. Alternative statistical distributions have been extensively discussed to deal with overdispersed count data on the single species level (e.g., Anscombe, 1949; Bliss and Fisher, 1953; Sileshi, 2006, to name a few).

Whatever the reasons for overdispersed counts may be, the overdispersion can often not be sufficiently explained by the measured covariates at hand. Here, we consider count data of multiple species with the aim to estimate biodiversity indices. That is, in each experimental unit (trap, field plot, mesocosm, etc.), animals of not only one but several species are counted, leading to a vector of counts. It is natural to expect that overdispersion is also present for these vectors of counts, such that the counts do not follow a multinomial distribution, but show higher variability. Additionally, ecological sampling in agricultural field trials is usually not even multinomial sampling: the total number of individuals in each trap is by no means fixed, but is a random variable itself (compare Pla, 2004). Hence, our focus is to compare different options for calculating simultaneous confidence intervals for biodiversity indices in the case of count data with extra-multinomial variability. One possible distribution for such data is the Dirichlet-multinomial distribution (Johnson et al., 1997). The four example data sets are analyzed to motivate that overdispersion may be a relevant problem in practical application.

The remainder of the paper is organized as follows: Section 2 introduces four data sets. Section 3 reviews the properties of the Shannon and Simpson indices and methods for the construction of simultaneous confidence intervals, namely the asymptotic methods based on the multinomial assumption, the Dunnett procedure, and three different approaches to construct simultaneous bootstrap confidence intervals. In Section 4, the data sets are analyzed with respect to plausible distributional assumptions,

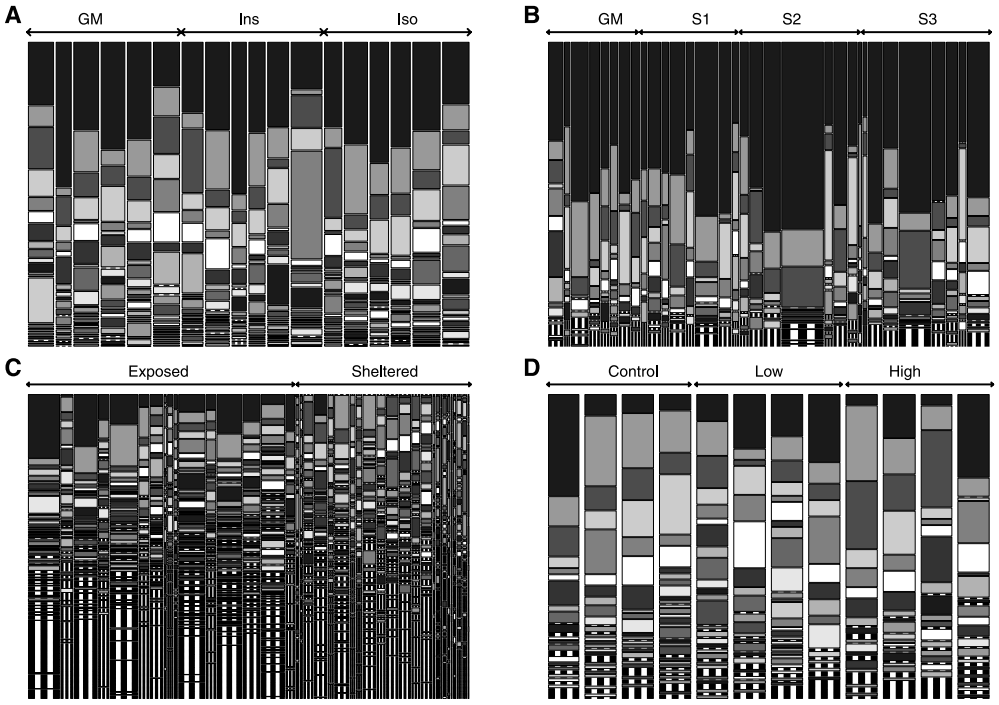


Figure 1 Mosaic plots of the four example data sets. The width of the columns is proportional to the number of individuals caught in each sample and the height of the boxes within each column is proportional to the number of individuals of each species within each sample. Dashed lines indicate species missing in a certain sample. (A) Counts of taxonomic groups of insects from an agricultural field trial; interest is in comparing the genetically modified maize variety, GM, against a conventional variety under two different treatments, Iso and Ins. (B) Counts of insect species from a field trial, with interest in comparing a GM variety to three conventional varieties, S1–S3. The counts of species inhabiting kelp holdfasts are plotted in subfigure (C). An experimental question may be the comparison of biodiversity between exposed and sheltered sites. (D) A mesocosm experiment concerning nematodes in marine sediments with interest in comparing two treatments of enrichment with organic matter (low, high) to an untreated control treatment.

and simultaneous confidence interval methods are applied to one example from marine ecology. The results of the simulation study are shown to illustrate the properties of the simultaneous confidence interval methods for different overdispersion scenarios. Section 5 briefly discusses these results.

2 Data sets

In the following, we give a brief introduction to four publicly available data sets which contain count data of multiple species of some defined habitat or species community. They all include two or more experimental conditions or treatments, with replicated observations available for each of them. Figure 1 gives an impression of the within-treatment variability of the observed counts of a given species, the variability of the total number of individuals, the proportion of rare species and zero counts (dashed

lines in Fig. 1). Moreover, examples A, B, and D motivate the necessity for multiple comparisons to a control group. In Section 4.1, we show results of analyzing some distributional properties of these data sets, including the observed mean-variance dependency of the counts, in order to motivate the simulation of overdispersed count data in Section 3.5.

(A) Saprophagous Diptera in genetically modified maize: A field trial was performed in one location in Germany, 2002, and was arranged as a randomized complete block design with six blocks and three varieties: a genetically modified maize variety (GM), the near-isogenic line (Iso), and the near-isogenic line with a pyrethroid treatment (Ins). Diptera with saprophagous larvae were caught using emergence traps situated in each of the 18 plots. In total 25,308 individuals were captured and classified at the species level. In total 25 species with relatively high abundance are considered, while exceptionally rare species are not contained in the data set. One experimental question in the analysis of these data was whether the biodiversity in the GM variety differs from that in the Iso or Ins treatment.

(B) Predatory community in genetically modified maize: A field trial was performed in 2005 at one location in Germany, arranged as a randomized complete block design with eight blocks and four varieties: a genetically modified line (GM), and three conventional varieties (S1, S2, S3). A trap was situated on each of the 32 plots. In total, 3660 individuals were sampled and classified at the species level, leading to 33 observed species. One experimental question was whether the biodiversity in the GM variety differs from that in the three conventional varieties. Examples A and B are available in the `R` package `simboot` (Scherer, 2012).

(C) Kelp holdfasts in sheltered and wave-exposed environments: Anderson et al. (2005) investigated the biodiversity of the fauna inhabiting kelp holdfasts with respect to different spatial and taxonomic scales. The related published data set in the `R` package `untb` (Hankin, 2007) contains counts of 176 taxonomic units from 40 samples and comprises a total of 8419 individuals. The samples originate from four sheltered and four wave-exposed sites along the New Zealand coast line, i.e., the data contains some clustering of observational units, which is not documented in the published data set. The experimental question for the given data set is whether biodiversity differs between the exposed and sheltered sites.

(D) Marine nematodes under pollution with organic matter: In a mesocosm experiment in marine sediments (Gee et al., 1985) concerning the effects of pollution with organic matter, 12 samples have been subject to three different treatments: a control group, a low and a high enrichment with organic matter, each treatment with four replications. The published data set in the `R` package `mvabund` (Wang et al., 2011) contains counts of 53 species of marine nematodes. The experimental question is whether the enrichment of organic matter (mimicking environmental pollution) does change the diversity indices compared to the control group.

3 Methods

3.1 Notation

Assume that the objective of a study is to compare I treatments or conditions, with index $i = 1, \ldots, I$. The main focus is on multiple comparisons to a control. For simplicity, the last treatment, $i = I$, denotes the control treatment. Each treatment i is randomly assigned to several experimental units (e.g., to several field plots in examples A and B or to several mesocosms in example D). These replications have index $j = 1, \ldots, J_i$ within each i, such that ij identifies the experimental unit. In each experimental unit ij, a vector of counts $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijS})$ is obtained, where y_{ijs} is the observed number of animals of the s-th species (or taxonomic entity), and the species have index $s = 1, \ldots, S$. The total number of animals per experimental unit is denoted $n_{ij} = \sum_{s=1}^{S} y_{ijs}$, the total number of animals observed in all observational units of the i-th treatment is denoted by $n_i = \sum_{j=1}^{J_i} n_{ij}$. The unknown expected proportions of the S species in treatment i is denoted $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iS})$, which we will refer to as the relative abundance.

3.2 Multiple differences of Shannon and Simpson indices

The relative abundances $\boldsymbol{\pi}_i$ might be summarized by indices $\phi_i = f(\boldsymbol{\pi}_i)$. Here, we consider the Shannon index $\phi_i^{(H)}$ and the Simpson index $\phi_i^{(D)}$, given in Eqs. (1) and (2), respectively.

$$\phi_i^{(H)} = -\sum_{s=1}^{S} \pi_{is} \log\left(\pi_{is}\right), \tag{1}$$

$$\phi_i^{(D)} = \sum_{s=1}^{S} \pi_{is}^2. \tag{2}$$

A straightforward estimate for the elements of $\boldsymbol{\pi}_i$ is to sum up species-wise over all observations within treatment i

$$y_{i\cdot s} = \sum_{j=1}^{J_i} y_{ijs} \tag{3}$$

and dividing by the total number of individuals observed in treatment i, n_i: $\hat{\pi}_{is} = y_{i\cdot s}/n_i$. Replacing the relative abundances π_{is} in Eq. (2) by point estimators $\hat{\pi}_{is}$ yields the estimator $\hat{\phi}_i^{(D)}$ of the Simpson index. However, the simple plug-in of point estimates of the relative abundances in Eq. (1) for those species present in the sample leads to a biased estimator of the Shannon index, see, e.g., Hutcheson (1970), Lande (1996), Magurran (2004). In the following, we use the estimator in Eq. (4), correcting for the bias by subtracting the three bias terms of the series given in, e.g., Hutcheson (1970) or Magurran (2004),

$$\tilde{\phi}_i^{(HC)} = -\sum_{s:y_{i\cdot s}>0} \hat{\pi}_{is} \log\left(\hat{\pi}_{is}\right) + \frac{O_i - 1}{2n_i} - \frac{1 - \sum_{s:y_{i\cdot s}>0} \hat{\pi}_{is}^{-1}}{12n_i^2} - \frac{\sum_{s:y_{i\cdot s}>0} \left(\hat{\pi}_{is}^{-1} - \hat{\pi}_{is}^{-2}\right)}{12n_i^3}, \tag{4}$$

where O_i denotes the number of species observed with at least one individual in community i. Though Eq. (4) attempts to correct the bias, preliminary simulations by the authors (available in Supporting Information) as well as previous publications (e.g., Lande, 1996; Fritsch and Hsu, 1999) have shown that such bias corrections do not remove the bias completely and an unbiased estimator for Shannon diversity does not exist (Lande, 1996).

Given that the sums $y_{i\cdot s}$ follow a multinomial distribution with parameters n_i and $\boldsymbol{\pi}_i$, the variance of an estimate $\hat{\phi}_i$ merely depends on $\boldsymbol{\pi}_i$ and n_i. For the Shannon index, Fritsch and Hsu (1999) give Eq. (5) as the variance of the estimator according to the Delta method, where $\boldsymbol{\pi}_i = \left(\pi_{i1}, \ldots, \pi_{iS}\right)'$ and $\boldsymbol{\pi}_i'$ is the transposition of $\boldsymbol{\pi}_i$.

$$\mathrm{Var}(\hat{\phi}_i^{(H)}) = \frac{\log(\boldsymbol{\pi}_i)' \left[\mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i'\right] \log(\boldsymbol{\pi}_i)}{n_i}. \tag{5}$$

Rogers and Hsu (2001) give the variance of the estimator of the Simpson index as

$$\mathrm{Var}(\hat{\phi}_i^{(D)}) = \frac{2}{n_i\left(n_i - 1\right)} \left[\sum_{s=1}^{S} \pi_{is}^2 + 2\left(n_i - 2\right) \sum_{s=1}^{S} \pi_{is}^3 + \left(3 - 2n_i\right) \left(\sum_{s=1}^{S} \pi_{is}^2\right)^2\right]. \tag{6}$$

The sample estimates $\widehat{\mathrm{Var}}(\hat{\phi}_i^{(H)})$ and $\widehat{\mathrm{Var}}(\hat{\phi}_i^{(D)})$ of the variances are obtained by replacing π_{is} by $\hat{\pi}_{is}$ in Eqs. (5) and 6. However, if the variance of the observed counts y_{ijs} is higher than can be expected

from samples of a multinomial distribution, these variance estimators may severely underestimate the variance of the estimated Shannon and Simpson indices.

Our focus is on the $M = I - 1$ differences against the control group,

$$\delta_m = \phi_m - \phi_I, \text{ with } m = 1, \ldots, I-1. \tag{7}$$

We are particularly interested in constructing simultaneous confidence intervals

$$\left[\delta_m^{(L)}, \delta_m^{(U)}\right], m = 1, \ldots, M \tag{8}$$

in which all true differences, $\delta_1, \ldots, \delta_M$, are included with high confidence $(1-\alpha)$, e.g., $\alpha = 0.05$.

3.3 Previously proposed methods for simultaneous confidence intervals

Confidence intervals for diversity indices were constructed by Fritsch and Hsu (1999) and Rogers and Hsu (2001), denoted as FH99 and RH01 in the following, for the special case of multinomial distributed counts. While FH99 constructed unadjusted confidence intervals in an equivalence setting, RH01 presented simultaneous confidence intervals for the comparison of diversity indices between multiple treatments. In contrast to our motivating examples, FH99 and RH01 use a data set without replicates, that is, $J_i = 1$ for all i. Asymptotic simultaneous confidence intervals can then be constructed based on the point estimates given in Section 3.2 and denoted as method AM in the following.

$$\left[\delta_m^{(L)}, \delta_m^{(U)}\right] = \left[\hat{\phi}_m - \hat{\phi}_I \pm q\sqrt{\widehat{\mathrm{Var}}(\hat{\phi})_m + \widehat{\mathrm{Var}}(\hat{\phi})_I}\right]. \tag{9}$$

RH01 is used for these quantiles of the M-variate normal distribution with correlation matrix R depending on the corresponding variance estimates. We apply the same approach for the Shannon index, because the problem is a special case of the methods described in Hothorn et al. (2008). In the two-sided case, this quantile is the value $q = q_{M,R,1-\alpha}^{2-sided}$, such that $P(|z_m| < q, \forall m = 1, \ldots, M) = 1-\alpha$, where $(z_1, \ldots, z_M)$ is an M-variate normal random vector with expectation $\mathbf{0}$ and correlation matrix R. In the one-sided case, the quantile is $q = q_{M,R,1-\alpha}^{1-sided}$, such that $P(z_m < q, \forall m = 1, \ldots, M) = 1-\alpha$. Here, we follow the recommendation of RH01 and compute the intervals defined in Eq. (9) with quantiles of the multivariate normal distribution after summing up over the replications J_i within each treatment i and computing point and variance estimates based on $y_{i.s}$ (Eq. (3)). Quantiles $q_{M,R,1-\alpha}$ are computed using the `R` package `mvtnorm` (Genz et al., 2009).

When replications are available, a technically simple and probably commonly applied approach for analyzing diversity indices is to compute the diversity index of interest for each observation ij, $i = 1, \ldots, I$, $j = 1, \ldots, J_i$ in the data set, resulting in a new variable of interest $\hat{\phi}_{ij} = f(\hat{\pi}_{ijs})$, $\hat{\pi}_{ijs} = y_{ijs} / \sum_{s=1}^{S} y_{ijs}$. One may assume approximately Gaussian error distribution for these data (see, e.g., Magurran, 2004) and consequently use the well-known Dunnett test (Dunnett, 1955) to construct simultaneous confidence intervals for differences between treatments and the control. Let $\bar{\phi}_i$ denote the estimator for the i-th group mean that results from fitting a linear model (assuming independent homoscedastic Gaussian errors). Subtracting $\bar{\phi}_i$ from the corresponding $\hat{\phi}_{ij}$ are the residuals $\hat{\epsilon}_{ij}$ (Eq. (10))

$$\hat{\epsilon}_{ij} = \hat{\phi}_{ij} - \bar{\phi}_i, \tag{10}$$

and to an estimate for the residual variance, $\hat{\sigma}^2 = (\sum_{i=1}^{I} \sum_{j=1}^{J_i} (\hat{\epsilon}_{ij} - \bar{\epsilon}_i)^2) / (\sum_{i=1}^{I} J_i - I)$. Using appropriate two-sided $(1-\alpha)$ quantiles of an M-variate t distribution (details in Dunnett, 1955; Hothorn

et al., 2008) and the residual degree of freedom ν, approximate simultaneous confidence intervals are constructed as given in Eq. (11) and denoted as method AG in the following:

$$\left[\delta_m^L;\, \delta_m^U\right] = \left[\bar{\phi}_m - \bar{\phi}_I \pm t_{1-\alpha, M, R, \nu=\sum_{i=1}^{I} J_i - I}\ \hat{\sigma}\sqrt{\frac{1}{J_m} + \frac{1}{J_I}}\right]. \tag{11}$$

3.4 Bootstrap methods

The method of Westfall and Young (1993) is a straightforward alternative to construct simultaneous confidence intervals for differences of biodiversity indices and makes use of the treatment means $\bar{\phi}_i$, the residuals $\hat{\epsilon}_{ij}$ (Eq. (10)), and the estimator of a common residual variance as defined in the previous section. The procedure of Westfall and Young (1993), denoted as method WY in the following, then is:

(1) Draw a random sample $\hat{\epsilon}^*_{ij}$ with replacement from the residuals $\hat{\epsilon}_{ij}$ (without stratification).
(2) Compute the sample means $\bar{\epsilon}^*_i = (\sum_{j=1}^{J} \hat{\epsilon}^*_{ij})/(J_i)$, $i = 1, \ldots, I$, as well as the common residual variance $\hat{\sigma}^{2*} = (\sum_{i=1}^{I} \sum_{j=1}^{J_i} (\hat{\epsilon}^*_{ij} - \bar{\epsilon}^*_i)^2)/\sum_{i=1}^{I} J_i - I)$ from the bootstrap data.
(3) Compute the statistics

$$t^*_m = \frac{\bar{\epsilon}^*_m - \bar{\epsilon}^*_I}{\hat{\sigma}^*\sqrt{\frac{1}{J_m} + \frac{1}{J_I}}},\ m = 1, \ldots, I-1. \tag{12}$$

(4) Compute $t^*_{\max} = \max(|\, t^*_m\, |)$.
(5) Repeat steps 1–4 B times and store $t^*_{\max}$ for every bootstrap step, $b = 1, \ldots, B$.
(6) $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of the B values $t^*_{\max}$.

Simultaneous confidence intervals are then computed using Eq. (13):

$$\left[\delta_m^L;\, \delta_m^U\right] = \left[\bar{\phi}_m - \bar{\phi}_I \pm q_{1-\alpha}\ \hat{\sigma}\sqrt{\frac{1}{J_m} + \frac{1}{J_I}}\right]. \tag{13}$$

Appropriate one-sided intervals may be obtained by using a $1-\alpha$ quantile of $t^*_{\max} = \max(t^*_m)$ or an α quantile of $t^*_{\min} = \min(t^*_m)$ for upper and lower limits, respectively.

In an alternative bootstrap method, denoted as method TS in the following, we use sums $y_{i.s}$ and n_i and O_i to compute the corresponding point and variance estimates $\hat{\phi}_i$ and $\widehat{\mathrm{Var}}(\hat{\phi}_i)$ given in Section 3.2. As with the WY method, simultaneous confidence intervals may be constructed by bootstrapping the maximum of M test statistics:

(1) Perform a nonparametric bootstrap of the observational units ij, stratified by the i treatments, resulting in bootstrap data y^*_{ijs}. That is, draw I samples of size J_i with replacement from the I sets of indices ij, $(11, 12, \ldots, 1J_1)$, $(21, 22, \ldots, 2J_2)$, $\ldots$, $(I1, I2, \ldots, IJ_I)$, respectively, and build a new data set y^*_{ijs} containing those vectors $(y_{ij1}, \ldots, y_{ijS})$ for which ij has been sampled.
(2) Build the species-wise sums over all J_i observations within each treatment i, $y^*_{i.s}$, and the corresponding terms n^*_i, O^*_i, as well as $\hat{\phi}^*_i$, and $\widehat{\mathrm{Var}}(\hat{\phi})^*_i$ by using Eqs. (4), (5) or (2), (6).
(3) Compute

$$t^*_m = \frac{(\hat{\phi}^*_m - \hat{\phi}^*_I) - (\hat{\phi}_m - \hat{\phi}_I)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\phi}_m)^* + \widehat{\mathrm{Var}}(\hat{\phi}_I)^*}}. \tag{14}$$

(4) Compute $t^*_{\max} = \max(\mid t^*_m \mid)$.
(5) Repeat steps 1–4 B times and store $t^*_{\max}$ for every bootstrap step, $b = 1, \ldots, B$.
(6) $q_{1-\alpha}$ is the $1-\alpha$ quantile of the empirical distribution of the B values $t^*_{\max}$

Simultaneous confidence intervals are then computed as in Eq. (15):

$$\left[\delta^L_m;\, \delta^U_m\right] = \left[\hat{\phi}_m - \hat{\phi}_I \pm q_{1-\alpha}\sqrt{\widehat{\mathrm{Var}}(\hat{\phi})_m + \widehat{\mathrm{Var}}(\hat{\phi})_I}\right]. \tag{15}$$

One-sided intervals may be calculated by using the equivalent one-sided quantiles as described for the WY method above. Note that if the data y_{ijs} follow a multinomial distribution, the statistic t^*_m is a centered and pivotal statistic and follows the rules described by Westfall and Young (1993). If the data are overdispersed, the statistic is not pivotal any more.

Besag et al. (1995) described the construction of simultaneous intervals, directly based on an empirical joint distribution of the parameters of interest. Based on a nonparametric bootstrap of the original observations ij, stratified by the treatments i, simultaneous percentile intervals for multiple differences of diversity indices can be calculated just using the raw estimates.

(1) Resample y_{ijs} stratified by i, as described for the TS method, leading to y^*_{ijs} and compute $\hat{\phi}^*_i$ based on $y^*_{i.s}$, n^*_i and if necessary, O^*_i as above, as well as the differences of interest $\hat{\delta}^*_m = \hat{\phi}^*_m - \hat{\phi}^*_I$, $m = 1, \ldots I-1$.
(2) Repeat step 1 B times and store the $\hat{\delta}^*_m$ in a $(B \times M)$ matrix $\mathbf{\Delta}$ with elements $\hat{\delta}_{bm}$, $b = 1, \ldots, B$.
(3) Order and rank each of the M columns of $\mathbf{\Delta}$ separately. Results are the order statistics $z^{[b]}_m$, and the ranks u_{bm}.
(4) For each $b = 1, \ldots, B$, calculate $u^{(maxmin)}_b = \max\left(B + 1 - \min_m\left(u_{bm}\right), \max_m\left(u_{bm}\right)\right)$.
(5) Order $u^{(maxmin)}_b$ resulting in the order statistics $u^{[b]}$.
(6) Let b^* denote the closest integer to $B(1-\alpha)$ and $u^* = u^{[b*]}$, i.e., the b^*-th value in the ordered vector of $u^{(maxmin)}_b$.

The bounds of the simultaneous confidence region for the M differences of interest, denoted as method PE in the following, are then

$$\left[\delta^L_m;\, \delta^U_m\right] = \left[z^{[B+1-u^*]}_m;\, z^{[u^*]}_m\right]. \tag{16}$$

3.5 Simulation settings

We performed a Monte Carlo simulation study to compare the simultaneous confidence interval methods when applied to count data with extra-multinomial variance. Different patterns of relative abundances π_{is}, satisfying $\sum_{s=1}^{S} \pi_{is} = 1$, were taken from the geometric series (compare Rogers and Hsu, 2001).

$$\pi_s = \frac{k\,(1-k)^{s-1}}{1-(1-k)^s},\ s = 1, \ldots, S,\ 0 < k \leq 1, \tag{17}$$

where k approaching 0 results in relative abundance patterns with high evenness and $k = 1$ indicates absence of all species except one. In the simulations shown here, $S = 30$ is used throughout, while the number of species that are actually present in the data depends on k and n_i. For comparing four treatments, $i = 1, 2, 3, 4$, where $i = 4$ denotes the control group, the eight parameter settings in Table 1 were investigated.

Table 1 Settings of relative abundance patterns in four groups to be compared, defined by the group-wise parameters k_i of the geometric series, the corresponding group-wise Shannon and Simpson indices, and their acronyms used later on.

Acronym	k_1	k_2	k_3	k_4	$\phi_1^{(H)}$	$\phi_2^{(H)}$	$\phi_3^{(H)}$	$\phi_4^{(H)}$	$\phi_1^{(D)}$	$\phi_2^{(D)}$	$\phi_3^{(D)}$	$\phi_4^{(D)}$
k_1	0.1	0.1	0.1	0.1	3.07	3.07	3.07	3.07	0.943	0.943	0.943	0.943
k_2	0.2	0.2	0.2	0.2	2.49	2.49	2.49	2.49	0.889	0.889	0.889	0.889
k_3	0.3	0.3	0.3	0.3	2.04	2.04	2.04	2.04	0.824	0.824	0.824	0.824
k_4	0.4	0.4	0.4	0.4	1.68	1.68	1.68	1.68	0.750	0.750	0.750	0.750
k_{12}	0.2	0.2	0.2	0.1	2.49	2.49	2.49	3.07	0.889	0.889	0.889	0.943
k_{21}	0.1	0.1	0.1	0.2	3.07	3.07	3.07	2.49	0.943	0.943	0.943	0.889
k_{34}	0.3	0.3	0.4	0.3	2.04	2.04	1.68	2.04	0.824	0.824	0.750	0.824
k_{43}	0.4	0.4	0.3	0.4	1.68	1.68	2.04	1.68	0.750	0.750	0.824	0.750

Table 2 Values of c chosen for the simulations to depend on sample size n_{ij} such that 1.01-, 2-, 5-, and 10-fold overdispersion is achieved.

Overdispersion	1.01	2	5	10
$n_{ij} = 100$	9899.00	98.00	23.75	10.00
$n_{ij} = 1000$	99,899.00	998.00	248.75	110.00

We used a Dirichlet-mixture of multinomial distributions (Johnson et al., 1997) to simulate overdispersed count data. The group-wise parameters of the Dirichlet distribution, a_{is}, $s = 1, \ldots, S$, were chosen as $a_{is} = \pi_{is}c$ for the settings of π_{is} in Table 2. Using this distribution to simulate overdispersed multinomial counts, the actual overdispersion decreases with increasing c and additionally depends on the number of animals per observation n_{ij}. For the simulations, c has been chosen as a function of n_{ij}, such that the variance of the data is 1.01, 2, 5 and 10 times that of the corresponding multinomial data with parameters n_{ij} and π_{is} (details in Table 2). In order to assess the effects of different numbers of replications J_i and different numbers of animals per replication n_{ij}, four combinations of J_i and n_{ij} are investigated: $J_i = 5, n_{ij} = 100$; $J_i = 5, n_{ij} = 1000$; $J_i = 20, n_{ij} = 100$; $J_i = 20, n_{ij} = 1000$.

All combinations of the eight settings of group-wise relative abundance, π_{is} listed in Table 1, the four settings of overdispersion, and the four settings of sample sizes (Table 2) have been built. For each of these settings, 1000 data sets have been drawn and nominal 95% simultaneous confidence intervals for comparisons to control have been calculated to assess the simultaneous coverage probability of the intervals. Throughout, the bootstrap methods have been applied with $B = 2000$ bootstraps. Random number generation, calculation of simultaneous confidence intervals, and subsequent graphical representations have been performed in `R-2.9.2` (R Development Core Team, 2009), partially relying on the add-on packages `boot` (Canty and Ripley, 2009), `MCMCpack` (Martin et al., 2009, for Dirichlet random numbers), and `mvtnorm` (Genz et al., 2009). `R`-functions implementing the methods discussed in this paper as well as the example data sets A and B are available in our `R` package `simboot` (Scherer, 2012).

4 Results

To motivate the practical relevance of the simulation study, we first assess the four example data sets in Section 4.1 with respect to distributional properties which are important for the methods AM

and AG. In particular, we address three questions: Do the counts show overdispersion relative to the multinomial distribution? Is the observed variation of biodiversity indices between experimental units in line with those obtained under the multinomial assumption? Do the biodiversity indices calculated for the single experimental units markedly contradict the assumption of a Gaussian error distribution? Section 4.2 illustrates the application of the five methods to example D. Finally, the results of the simulation study are shown.

4.1 Distributional properties of the example data sets

First, we investigate whether the multinomial assumption is a suitable assumption for the count data in the four data sets, after correcting for all recorded covariates. As a first approach, we thus fitted baseline logit models with the multinomial assumption (function multinom in `R` package `nnet`, Venables and Ripley, 2002) to all four data sets. These models included the effects of interest as well as additional variables recorded, such as dummy coded variables for the block effects in examples A and B. Based on these fitted models, we computed the observed variance of the species counts with respect to the predictions of the models. Additionally, the expected variance under the multinomial assumption was calculated based on the overall proportions and mean sample size. Figure 2 shows the observed residual variation of the counts (y-axes), versus the expected variance under the multinomial assumption as defined above (x-axis). Note that the underlying logit models tend to over-fit the data, particularly for the rare species in models with many parameters compared to the number of observational units (e.g., example B). Figure 2 therefore depicts species with expected value per observational unit less than 1 in gray, and species occurring only once in the whole data set in light gray. Due to the possible over-fitting when calculating the residual variance, the current approach should underestimate the magnitude of overdispersion.

Overdispersion compared to the multinomial assumption is most evident in examples A and C. In example B, only the few most abundant species show clear overdispersion. In all examples, the counts of the most abundant species show more than twice the variance expected under multinomial distribution, but is found to be far higher than ten times the expected variance in A and C. Thus, the assumption of multinomial distribution appears inappropriate for all examples. Consequently, the methods of Fritsch and Hsu (1999) and Rogers and Hsu (2001) can be expected to be inappropriate.

A second question is: How does the overdispersion of the counts carry over to the variance estimates for the group-wise biodiversity indices, which mainly influence the width of the simultaneous confidence intervals for the group-wise differences among them? To illustrate this, we calculated the observed Shannon and Simpson index for each vector of observations, $\hat{\phi}_{ij}^{(H)}$ and $\hat{\phi}_{ij}^{(D)}$. These values are fitted by general linear models (homoscedastic Gaussian residuals assumed) accounting for possible treatment and block effects in examples A and B, and for the exposure and treatment effects in examples C and D, respectively. The residuals of those models have been used to compute the empirical variance of the treatment means $\hat{\sigma}_i^2 = (J_i/(J_i - 1)) \sum_{j=1}^{J_i} (\hat{\phi}_{ij} - \bar{\phi}_{ij})^2$. These can be compared to the variance estimates $\widehat{\mathrm{Var}}(\hat{\phi}_i^{(H)})$ and $\widehat{\mathrm{Var}}(\hat{\phi}_i^{(D)})$ provided by the methods in Fritsch and Hsu (1999) and Rogers and Hsu (2001) after treatment-wise summation of the counts and assuming multinomial counts. Table 3 shows the minimal and maximal ratios $\widehat{\mathrm{Var}}(\hat{\phi}_i^{(H)})/\hat{\sigma}_i^2$ for each of the four examples. This illustrates the extent of underestimating the variance when relying on the multinomial assumption for between-group inference concerning diversity indices (as recommended by Fritsch and Hsu, 1999; and Rogers and Hsu, 2001. In this second step, we find that there is also evidence for the inappropriateness of the multinomial assumption on the scale of biodiversity indices. For all groups in the four examples, the variance estimates according to Fritsch and Hsu (1999) and Rogers and Hsu (2001) are smaller than the empirical variance of the treatment means. The ratio is between 0.01 and 0.1 for examples A and C and it ranges between 0.13 and 0.32 in example B (empirically showing the lowest overdispersion).

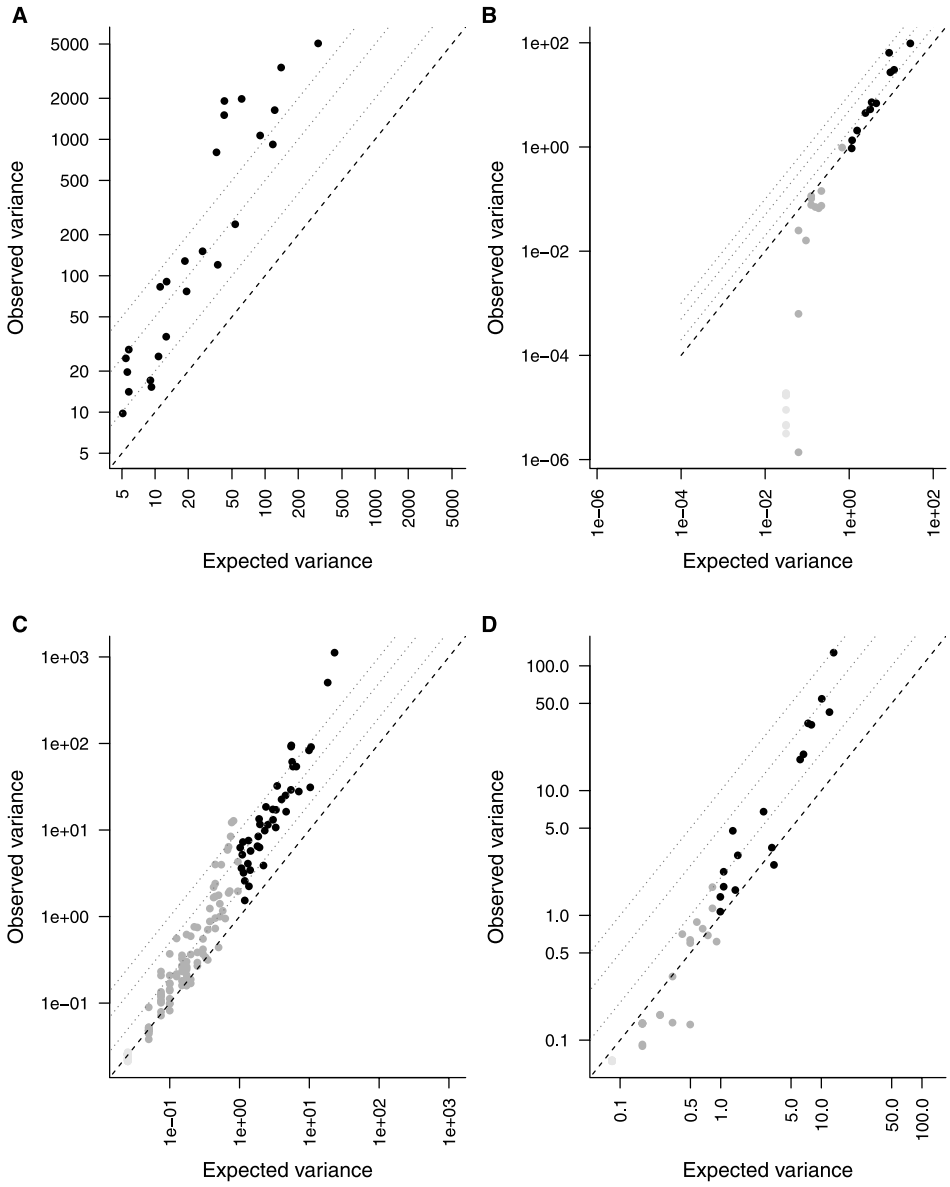


Figure 2 Observed variance of the species counts (after fitting log-linear models with multinomial assumption) plotted versus expected variance under multinomial assumption, with logarithmic scaling for both axes. Rare species (overall expected value per observational unit < 1) are shown in gray, very rare species (occurring only once in the data set) are shown in light gray. The solid line indicates that observed variance equals the expected variance, dotted lines indicate 2-, 5-, and 10-fold overdispersion in the species counts.

Table 3 Minima and maxima of the treatment-wise ratios of variance estimates under multinomial assumption to empirical variance estimate for $\phi^{(H)}$ and $\phi^{(D)}$ in the four example data sets are given in columns 1–4. Columns 5 and 6 show the p-values of the Shapiro–Wilk test on normality applied to the residuals of linear model fits for $\phi_{ij}^{(H)}$ and $\phi_{ij}^{(D)}$, respectively.

	$\widehat{\mathrm{Var}}(\hat{\phi}_i^{(H)})/\hat{\sigma}_i^2(\hat{\phi}_{ij}^{(H)})$		$\widehat{\mathrm{Var}}(\hat{\phi}_i^{(D)})/\hat{\sigma}_i^2(\hat{\phi}_{ij}^{(D)})$		p-value (Shapiro–Wilk)	
Example	Min	Max	Min	Max	$\hat{\phi}_{ij}^{(H)}$	$\hat{\phi}_{ij}^{(D)}$
A	0.033	0.105	0.012	0.111	0.4259	0.4939
B	0.131	0.279	0.189	0.328	0.3985	0.2146
C	0.010	0.027	0.006	0.033	<0.0001	<0.0001
D	0.128	0.815	0.094	0.197	0.7489	0.3285

Finally, the residuals of these linear models can be analyzed with respect to the appropriateness of assuming the normal distribution for the observed biodiversity indices $\hat{\phi}_{ij}^{(H)}$ and $\hat{\phi}_{ij}^{(D)}$. While the residuals in examples A, B, and D provide no evidence against normality in a Shapiro–Wilk test (Table 3), the residuals in example C clearly deviate from normality, where the Simpson index is left-skewed and the Shannon index is right-skewed in the corresponding Q-Q plots (not shown). In the plots, the residuals of the examples A and B also show similar patterns of deviation from normality, which are, however, not significant at the 5% level. Thus, applying classical multiple comparison procedures (assuming homoscedastic Gaussian residuals) could be appropriate in some cases, but may not in others, e.g., in example C.

4.2 Comparing diversity in example D

In example D, we are interested in comparing the low and high dose of enrichment with organic matter to the untreated control group. In Fig. 3, simultaneous confidence intervals for these two differences from a common control are shown for the Shannon and Simpson indices, based on the four methods under consideration. With respect to the practical question at hand, we find no significant difference in Shannon or Simpson diversity between the low enrichment and control and the high enrichment and control (a difference of zero is included in the intervals for all three comparisons). Finding no significant effect here may be due to the true absence of the effect or due to the high residual variation and very small sample size, such that potential small effects can not be detected as significant. The asymptotic method relying on the assumption of multinomial distributed counts (AM) leads to the narrowest confidence intervals, due to underestimating the uncertainty of the diversity estimates in overdispersed count data. All three bootstrap methods yield wider intervals than the AM method. The AG, WY, TS, and AM methods are symmetric around the point estimate by construction, while the intervals yielded by the PE method are slightly asymmetric, taking the skewness of the distribution of the Shannon and particularly the Simpson estimator into account.

4.3 Results of the simulation study

Figures 4 and 5 show the observed simultaneous coverage probabilities of nominal 95% simultaneous confidence intervals for differences to control of Shannon indices and Simpson indices, respectively. The coverage probabilities vary over a wide range between the methods. To show this wide range but still use a common plot range for comparing the five methods, providing a high resolution for coverage probabilities between 0.8 and 1, we use a clog scale in the graphics depicting simulated coverage probabilities. The asymptotic method assuming a multinomial distribution (AM) works well for those

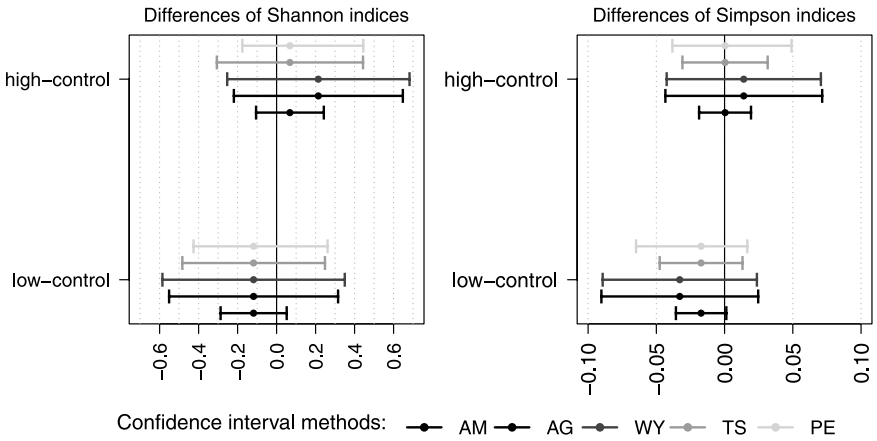


Figure 3 Simultaneous 95% confidence intervals for differences of Shannon and Simpson diversity indices between the low and high dose to the untreated control. Horizontal bars indicate the confidence region, while points indicate the estimates. Different values of gray distinguish intervals computed by the five different methods, AM (asymptotic multinomial), AG (asymptotic Gaussian), WY (Westfall–Young), TS (Bootstrap-t based on sumed counts), and PE (percentile method). The vertical line indicates the null hypothesis of no difference in diversity indices.

settings where the simulated counts are nearly multinomial distributed (1.01-fold overdispersion), while the resulting simultaneous confidence intervals are severely too narrow for all considered cases with clear overdispersion. If one would use this method for a test decision and claim a significant change in diversity if 0 is excluded by one interval, this claim would be erroneous in about 20%, 60%, or 80% of the cases, where the method is applied to data with 2-, 5-, or 10-fold overdispersion, respectively. The WY and the AG methods have very similar coverage probabilities for all simulated parameter settings. When applied with the Shannon index, they show very low coverage probabilities for situations with high but differing diversity in combination with the smaller total number of counts per observational unit n_{ij}. For the Simpson index, WY and AG have very good properties for the situations considered, except for the two situations with relatively high diversity that differs between groups (see settings k_{21}, k_{12} in Table 1 and Fig. 5): The distribution of the group-wise Simpson indices (close to its bound 1) is then skewed differently in the different groups. The observed coverage probability may be higher or lower than the prespecified level. The observed coverage probabilities of the TS and PE methods based on summed counts do not differ clearly between situations with equal (k_1, k_2, k_3, k_4) and unequal (k_{12}, k_{21}, k_{34}, k_{43}) group-wise Shannon indices. Possibly, they are less affected by the bias of the Shannon index. Both methods have (at least slightly) too low coverage probabilities in most situations considered. In particular, the PE method has too low coverage probability if there are only $J_i = 5$ replications per group (first and second row in Figs. 4 and 5).

For a subset of the settings described in Section 3.5, we additionally simulated the performance of the methods for all pairwise comparisons, i.e., Tukey-type multiple comparisons (results not shown). The results support the findings for comparisons to control, whereas the minimal coverage probabilities for situations with a small number of replications and a small total sample size are even somewhat lower. Detailed tables and additional graphics of all simulations are available from the corresponding author upon request. Graphics presenting additional simulation results are available in Supporting Information.

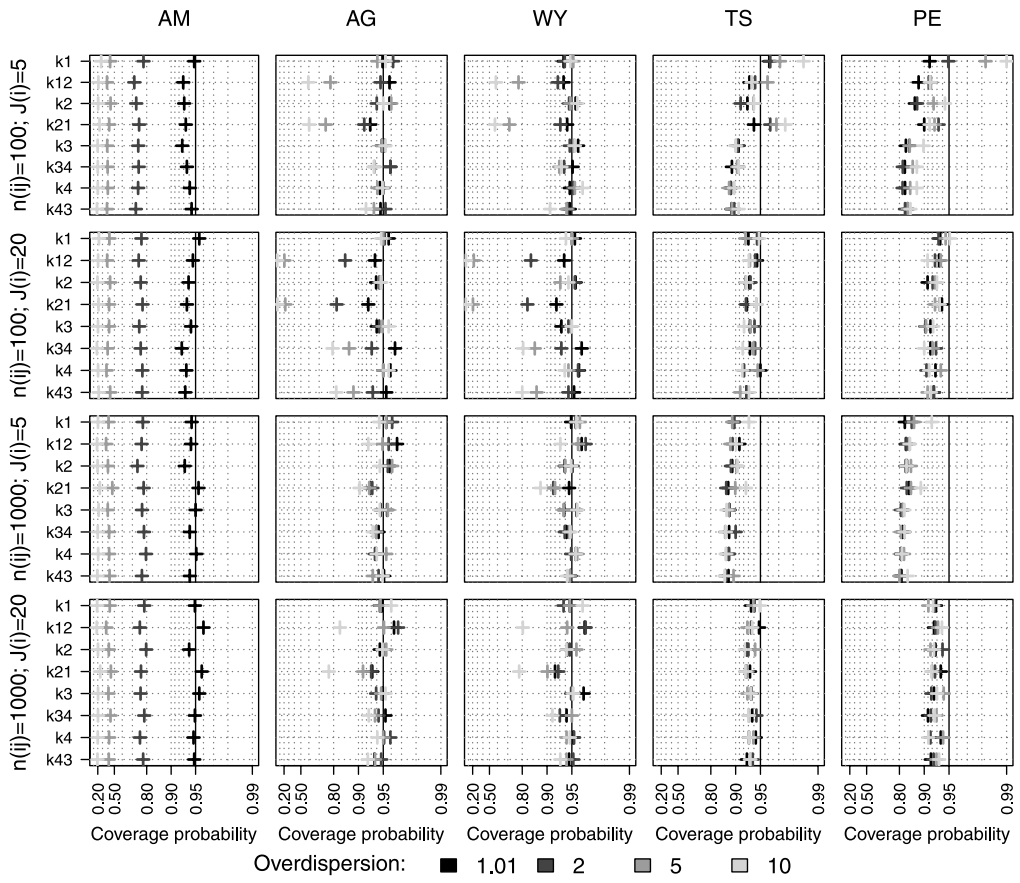


Figure 4 Simultaneous coverage probabilities of nominal 95% simultaneous confidence intervals for Shannon indices. The nominal simultaneous confidence level is represented by the vertical solid line. The coverage probabilities are presented on a clog scale. In the five columns, results for the five methods under consideration (asymptotic multinomial: AM, approximate Gaussian: AG, Westfall–Young: WY, Bootstrap-t based on summed counts: TS, and percentile intervals based on summed counts: PE) are shown. The four rows display results for different numbers of replications (J_i) and different number of total counts per trap (n_{ij}), horizontal lines in subgraphs show results for the different diversity patterns introduced in Table 1. Lighter gray indicates higher overdispersion.

5 Discussion

Motivated by four data sets with background in agricultural or marine ecology, we investigated the performance of simultaneous confidence intervals for differences of diversity indices. The analysis of the example data shows that overdispersion of count data can be of substantial extent, making the assumption of the multinomial distribution for the data implausible. Previously published methods to compute simultaneous confidence intervals based on the multinomial assumption are compared to different bootstrap approaches which do not rely on this assumption and can take overdispersion into account. In a simulation study, drawing overdispersed count data from Dirichlet-multinomial

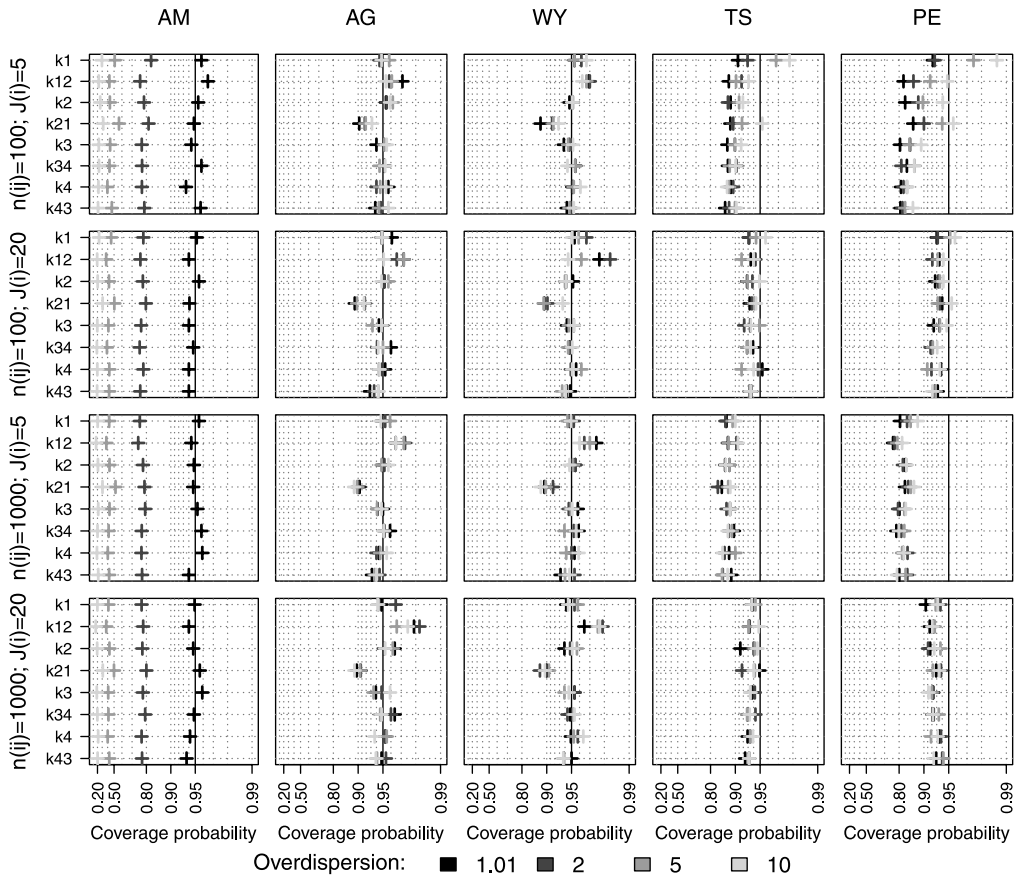


Figure 5 Simultaneous coverage probabilities of nominal 95% simultaneous confidence intervals for Simpson indices. The nominal simultaneous confidence level is represented by the vertical solid line. The coverage probabilities are presented on a clog scale. In the five columns, results for the five methods under consideration (asymptotic multinomial: AM, approximate Gaussian: AG, Westfall–Young: WY, Bootstrap-t based on summed counts: TS, and percentile intervals based on summed counts: PE) are shown. The four rows display results for different numbers of replications (J_i) and different number of total counts per trap (n_{ij}), horizontal lines in subgraphs show results for the different diversity patterns introduced in Table 1. Lighter gray indicates higher overdispersion.

mixtures and relative abundances following geometric series, it is illustrated that methods based on the multinomial assumption have unacceptably low coverage probability for the simulated values of overdispersion. The three bootstrap versions show coverage probabilities closer to the nominal level for all situations with overdispersed data considered in our simulation. Using classical multiple comparison procedures such as the Dunnett procedure to compare samples of observed diversity indices has good coverage probabilities when there is no difference between the samples. However, such intervals may have very low coverage probabilities when samples differ in diversity and the Shannon index is used. Following the previously published recommendation to sum up counts within groups (Rogers and Hsu, 2001) in the context of randomized field trials concerning diversity, and to use methods based

on the multinomial assumption for subsequent statistical inference may have devastating effects on the validity of this inference. Rogers and Hsu (2001) discourage taking replications into account by using linear models or ANOVA when comparing Simpson indices. In contrast to Rogers and Hsu (2001), we recommend using linear models and subsequent multiple comparisons or appropriate bootstrap methods as a first choice if randomized replications are available, and discourage basing conclusions on methods relying on the multinomial assumption without a careful assessment of this strong assumption.

The comparison of three different bootstrap methods is motivated by some specific problems of the diversity indices considered: The estimator of the Simpson index may have a skewed distribution because it is bounded between 0 and 1, the estimator for the Shannon diversity is biased depending on the sample size and the relative abundance pattern, the latter being only observable in parts, and may have a skewed distribution as well. Multiple comparison procedures based on assuming Gaussian errors and the Westfall–Young method of resampling the residuals of a linear model are well known and simple to use. The Westfall–Young method, in principle, has the advantage that it can deal with nonnormality of the data. In the simulation study, however, we found very similar coverage probabilities in the two approaches. Because both are based on diversity indices computed for each observational unit, their coverage probability can be severely decreased when applied for the Shannon index with its biased estimator. The two alternative methods, which are based on group-wise summed counts, are less affected by the bias of the Shannon estimator. However, the percentile method has considerably lower coverage probability than the two other methods, in particular with low group-wise sample sizes as are common in many studies. An advantage of the simple percentile procedures based on approximate normality and the Westfall–Young method is that they can also straightforwardly be applied for other indices of diversity without the need of specifying a variance estimator, e.g., if additional interest would be in the evenness or species richness. Another difference of the percentile method from all other methods considered in this paper is that the intervals are not constructed symmetrically with respect to the point estimates. This is an advantage when the distribution of the estimators is skewed by different magnitudes between samples, and thus the distributions of the estimated differences are skewed as well. However, both other bootstrap methods may be constructed asymmetrically too, while using the $1-\alpha/2$-quantiles of the minimum and maximum test statistics; see pages 58 and 83 in Westfall and Young (1993).

In many practical situations, additional structures such as blocks, more complicated hierarchical structures, or observation of covariates may be contained in the data. The Westfall–Young method of resampling the residuals is the only bootstrap method considered here that could be modified to deal with blocks or covariates. A detailed example and a discussion of bootstrapping residuals in presence of covariates are presented in Westfall and Young (1993, pp. 106–111). Our simulation study does not involve these practically relevant problems. Hence, the methods and conclusions of the simulation studies are restricted to the simplistic case of a completely randomized design. If a data set contains more complex randomization structures or covariates, a more flexible approach is to use linear models (assuming homoscedastic Gaussian residuals) to fit diversity indices computed for each observational unit. Subsequent multiple comparison procedures are described in Hothorn et al. (2008). When following this approach, transformations may be necessary to approach the plausibility of additive effects, homoscedasticity and normality of the residuals. The analysis of the four example data sets here suggests that this has to be assessed on a case-by-case basis.

Although only comparisons to control in a simple one-way treatment structure have been considered in detail, all four methods can be generalized for user-defined multiple contrasts among several treatment groups. This generalization, where particular comparisons of interest are defined in a contrast matrix, is already implemented in the R-programs used for the simulation study above. Generally, the use of diversity indices, summarizing observed abundances in a community by just one number, might be the subject of debate. A more informative approach is that of Di Battista and Gattone (2003) which directly compares relative abundance patterns.

Acknowledgments The work of F.S. was partly funded by the Bundesministerium für Bildung und Forschung (BMBF), grant number 0313269. The work of K.U.P. was funded by the BMBF, grant number 0313279A. The work of S.P. was funded by the BMBF, grant number 0312631G. Further, we thank the associate editor and the referees for their many helpful and constructive comments on earlier versions of the manuscript.

Conflict of interest
The authors have declared no conflict of interest.

References

Anderson, M. J., Connell, S. D., Gillanders, B. M., Diebel, C. E., Blom, W. M., Saunders, J. E. and Landers, T. J. (2005). Relationships between taxonomic resolution and spatial scales of multivariate variation. *Journal of Animal Ecology* **74**, 636–646. DOI:10.1111/j.1365-2656.2005.00959.x

Anscombe, F. J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* **5**, 165–173.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3–66.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* **9**, 176–200.

Canty, A. and Ripley, B. (2009). boot: bootstrap R (S-Plus) functions. R package version 1.2-38.

Di Battista, T. and Gattone, S. A. (2003). Non-parametric tests and confidence regions for intrinsic diversity profiles of ecological populations. *Environmetrics* **14**, 733–741.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.

Fritsch, K. S., and Hsu, J. C. (1999). Multiple comparison of entropies with application to dinosaur biodiversity. *Biometrics* **55**, 1300–1305.

From, S. G. (2003). Confidence intervals for Gini's diversity measure and Shannon's entropy using adjusted proportions. *Communications in Statistics-Theory and Methods* **32**, 935–954.

Gee, J. M., Warwick, R. M., Schaanning, M., Berge, J. A., and Ambrose, W. G. (1985). Effects of organic enrichment on meiofaunal abundance and community structure in sublitoral soft sediments. *Journal of Experimental Marine Biology and Ecology* **91**, 247–262.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2009). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-7.

Hankin, R. K. S. (2007). Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software* **22**, 1–15.

Hothorn, T., Bretz, F., Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

Hutcheson, K. (1970). A test for comparing diversity, based on Shannon formula. *Journal of Theoretical Biology* **29**, 151–154.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. John Wiley and Sons, Inc., New York.

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76**, 5–13.

Magurran, A. E. (2004). *Measuring Biological Diversity*. Blackwell Publishing, Malden, MA.

Martin, A. D., Quinn, K. M. and Park, J. H. (2009). MCMCpack: Markov chain Monte Carlo (MCMC) Package. R package version 1.0-4. Available at http://CRAN.R-project.org/package=MCMCpack

Pardo, L., Salicru, M., Morales, D. and Menendez, M. L. (1997). Large sample behavior of entropy measures when parameters are estimated. *Communications in Statistics-Theory and Methods* **26**, 483–501.

Pla, L. (2004). Bootstrap confidence intervals for the Shannon biodiversity index: a simulation study. *Journal of Agricultural Biological and Environmental Statistics* **9**, 42–56.

R Development Core Team (2009). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org.

Renyi, A. (1961). On measures of information and entropy. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Vol.1, University of California Press, CA, pp. 547–561. Available at http://digitalassets.lib.berkeley.edu/math/ ucb/text/math_s4_v1_article-27.pdf.

Rogers, J. A. and Hsu, J. C. (2001). Multiple comparisons of biodiversity. *Biometrical Journal* **43**, 617–625.

Salicru, M., Vives, S. and Ocana, J. (2005). Testing the homogeneity of diversity measures: a general framework. *Journal of Statistical Planning and Inference* **132**, 117–129.

Scherer, R. (2012). Simboot: simultaneous inference for diversity indices. R package version 0.1-6. Available at http://CRAN.R-project.org/package=simboot

Sileshi, G. (2006). Selecting the right statistical model for analysis of insect count data by using information theoretic measures. *Bulletin of Entomological Research* **96**, 479–488.

Simpson, E. H. (1949). Measurement of diversity. *Nature* **163**, 688.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. (4th edn.). Springer-Verlag New York, Inc., New York.

Wang, Y., Naumann, U., Wright, S. and Warton, D. (2011). mvabund: statistical methods for analysing multivariate abundance data. R package version 2.3. Available at http://CRAN.R-project.org/package=mvabund

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. John Wiley & Sons, New York.

Computational Statistics and Data Analysis 58 (2013) 265–275

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables

Frank Schaarschmidt *

Institute of Biostatistics, Natural Science Faculty, Leibniz Universität Hannover, Herrenhäuser Str.2, 30419 Hannover, Germany

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 29 June 2012
Accepted 16 August 2012
Available online 12 September 2012

Keywords:
Multiple contrasts
One-way layout
Coverage probability
Generalized pivotal quantity

ABSTRACT

In biological and medical research, continuous, strictly positive, right-skewed data, possibly with heterogeneous variances, are common, and can be described by log-normal distributions. In experiments involving multiple treatments in a one-way layout, various sets of multiple comparisons among the treatments and corresponding simultaneous confidence intervals can be of interest. The focus is on multiple contrasts of the expected values of the treatments. Previously published methods based on normal approximations and generalized pivotal quantities are extended to the case of multiple contrasts. These methods are evaluated in a simulation study that involves comparisons to a control group, all pairwise comparisons and, to illustrate more general multiple contrast types, a non-standard type of contrast matrix. A method based on generalized pivotal quantities is recommended because it is superior to all other methods in terms of simultaneous coverage probability and because the type-I-errors are distributed almost equally between lower and upper confidence bounds. Methods based on normal approximations are found to be very liberal and biased with respect to directional type-I-errors. These methods are illustrated with an example from pharmaceutical research.

1. Introduction

In biological or medical research, data are often strictly positive and have right-skewed distributions with variances that increase with increasing means. In particular, these properties can be expected when the observed random variable can be assumed to arise from multiplicative processes. Examples include the mass of cultures or areas of plant leaves in early (exponential) stages of growth, gene expression and metabolite contents in biological systems. One way to take those properties into account is to assume a log-normal distribution for the data. This assumption can be justified theoretically in some situations, e.g., particle size distributions (Johnson et al., 1994). In other cases, such as the area under the curve (AUC) in pharmacokinetics, a log-normal distribution is frequently assumed without theoretical justification (Liu and Weng, 1994).

In controlled studies that include several experimental treatments, multiple comparisons among these treatments are common. This paper considers the case where the familywise error rate (FWER) is controlled for a set of comparisons, and the magnitude and relevance of effects are assessed via simultaneous confidence intervals for those multiple comparisons. Standard procedures under the assumption of normal data are Tukey's method for all pairwise comparisons (Tukey, 1953) and Dunnett's method for comparisons to a control group (Dunnett, 1955). To address more specific experimental questions, simultaneous confidence intervals for user-defined multiple contrasts are available as well (Bretz et al., 2001) under the normality assumption.

A frequently used approach to analyze log-normal data is to log-transform the observations, to assume normality, apply standard methods and interpret the back-transformed confidence intervals (e.g. Steinijans and Hauschke, 1992). In

* Tel.: +49 511 762 5821; fax: +49 511 762 4966.
E-mail address: schaarschmidt@biostat.uni-hannover.de.

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.08.011

cases that involve two or more treatment comparisons of the means of the log-transformed data, this approach leads to confidence intervals for ratios of medians. Another (computationally more difficult) option is to make an explicit comparison of treatments in terms of the differences or ratios of their expected values. To compare two samples, different methods to compute confidence intervals for those parameters have been proposed and evaluated with respect to their coverage probability (Chen and Zhou, 2006; Krishnamoorthy and Mathew, 2003). While a computationally simple asymptotic version has been found to have very poor empirical properties, other methods based on likelihood profiles and generalized pivotal quantities have been recommended (Chen and Zhou, 2006). A number of related problems have been addressed recently: Tian and Wu (2007) considered the estimation of a confidence interval for the common mean of several log-normal samples. Li (2009) addressed the problem of globally testing the equality of means of several samples in a generalized p-value approach. The problem of constructing simultaneous confidence intervals for user-defined sets of ratios or differences of expected values under the log-normal assumption has not been considered until now.

In Section 2, the asymptotic method and the method based on generalized pivotal quantities as described by Krishnamoorthy and Mathew (2003) and Chen and Zhou (2006) are extended to the construction of simultaneous confidence intervals for multiple, user-defined sets of ratios or differences of expected values. The simple standard techniques of Bonferroni adjustments for two-sample comparisons (Chen and Zhou, 2006) are included. The simultaneous coverage probability is compared in a Monte Carlo simulation with various parameter settings including balanced and unbalanced sample sizes and different sets of comparisons in Section 3. In Section 4, these methods are illustrated by applying them to a data set.

2. Methods

2.1. Notation and assumptions

In this work, a completely randomized design is assumed where variables W_{ij} are observable with $i = 1, \ldots, I$ denoting the treatment groups and $j = 1, \ldots, n_i$ denoting the independent replications of the ith treatment. It is also assumed that $W_{ij} = \exp(Y_{ij})$, where $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, such that the variables W_{ij} have two-parameter log-normal distributions where both parameters, μ_i and σ_i^2, can differ among the treatment groups $i = 1, \ldots, I$. Under this assumption, the median of W_{ij} depends only on μ_i via $\exp(\mu_i)$; the coefficient of variation, $CV_i = \sqrt{e^{\sigma_i^2} - 1}$, depends only on σ_i^2; and the expected value $E(W_{ij}) = \exp\left(\mu_i + \frac{\sigma_i^2}{2}\right)$, the variance $V(W_{ij}) = \exp\left(2\mu_i + \sigma_i^2\right)\left(\exp(\sigma_i^2) - 1\right)$, the skewness and the kurtosis depend on both parameters, μ_i and σ_i^2 (Johnson et al., 1994).

The observations are denoted w_{ij}, and the corresponding log-transformed values are $y_{ij} = \log\left(w_{ij}\right)$. The quantities $\hat{\mu}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} y_{ij}$ and $\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(y_{ij} - \hat{\mu}_i\right)^2$ are the maximum likelihood estimates for the parameters μ_i and σ_i^2, respectively, while $\bar{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}\left(y_{ij} - \hat{\mu}_i\right)^2$ denotes the usual unbiased estimate for σ_i^2.

2.2. Parameters of interest

Under the additional assumption of σ_i^2 being equal in all groups $i = 1, \ldots, I$ (i.e., the assumption of a common coefficient of variation in all treatments), it is straightforward to compute simultaneous confidence intervals for multiple differences $\mu_i - \mu_{i'}$ using the methods of Tukey, Dunnett or more general multiple contrast tests. Back-transformation of the resulting intervals leads to intervals for ratios of the medians, $e^{\mu_i - \mu_{i'}} = \frac{e^{\mu_i}}{e^{\mu_{i'}}}$.

This work focuses on differences or ratios of expected values rather than ratios of medians. For simplicity, the notation is $\theta_i = E(W_{ij}) = \exp\left(\mu_i + \frac{\sigma_i^2}{2}\right)$, $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_I)^\top$, and $\psi_i = \log(\theta_i)$, $\boldsymbol{\psi} = (\psi_i, \ldots, \psi_I)^\top$. This work considers multiple comparisons of user-defined sets of M ratios $\rho_{ii'} = \theta_i/\theta_{i'}$ or sets of M differences $\delta_{ii'} = \theta_i - \theta_{i'}$. More generally, the set of the M parameters of interest is defined by an $M \times I$ matrix $\boldsymbol{C}$ with elements c_{mi} in Eqs. (1) and (2),

$$\boldsymbol{\rho} = \exp\left(\boldsymbol{C}\boldsymbol{\psi}\right), \tag{1}$$

$$\boldsymbol{\delta} = \boldsymbol{C}\boldsymbol{\theta}. \tag{2}$$

The conditions $\sum_{i=1}^{I} c_{mi} = 0, \forall m = 1, \ldots, M$ and $\sum_{i:c_{mi}>0} c_{mi} = 1, \forall m = 1, \ldots, M$ simplify the practical interpretations of the parameters as ratios of the expected values (or ratios of weighted geometric means of expected values) in Eq. (1), and differences of the expected values (or differences of weighted arithmetic means of expected values) in Eq. (2).

When observations are assumed to be normally distributed, the expected value and the expected median are the same. In a number of models with non-normal distributions, treatments are compared by using differences or ratios of the expected values or directly related parameters, e.g., in generalized linear models assuming Poisson, binomial or related distributions. Conversely, comparing expected medians in parametric inference is uncommon, except under the log-normal assumption, probably due to its computational convenience. In some applications of the log-normal distribution, such as the comparison of medical costs (e.g. Chen and Zhou, 2006; Tian and Wu, 2007), it is evident that the expected values (the per case costs in

the long run) are more relevant for decision making than median costs. An important practical advantage of using expected medians for inference is that estimators for σ_i^2 are not involved. An approach that uses medians is less vulnerable in the presence of outlying observations than methods that compare expected values.

2.3. Simultaneous confidence intervals

Krishnamoorthy and Mathew (2003) and Chen and Zhou (2006) considered the case of computing intervals for ρ_{12} and δ_{12} when $I = 2$. The computationally most simple method is to calculate the confidence intervals based on the normal approximation. Maximum likelihood estimates for θ_i and ψ_i are $\hat{\theta}_i = \exp\left(\hat{\mu}_i + \frac{\hat{\sigma}_i^2}{2}\right)$ and $\hat{\psi}_i = \hat{\mu}_i + \frac{\hat{\sigma}_i^2}{2}$, respectively (Chen and Zhou, 2006). For a case with I independent samples, the variance of $\hat{\theta}_i$ can be estimated as follows:

$$\hat{V}\left(\hat{\theta}_i\right) = \frac{\hat{\theta}_i^2\hat{\sigma}_i^2}{n_i} + \frac{\hat{\theta}_i^2\hat{\sigma}_i^4}{2n_i}, \tag{3}$$

and the variance of $\hat{\psi}_i$ can be estimated as follows:

$$\hat{V}\left(\hat{\psi}_i\right) = \frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_i^4}{2n_i}, \tag{4}$$

(Chen and Zhou, 2006; Krishnamoorthy and Mathew, 2003). Then, the asymptotic simultaneous two-sided $(1-\alpha)$ confidence intervals for ρ_m are

$$\left[\exp\left(\sum_{i=1}^{I} c_{mi}\hat{\psi}_i \pm z_{1-\alpha/(2M)}\sqrt{\sum_{i=1}^{I} c_{mi}^2\hat{V}\left(\hat{\psi}_i\right)}\right)\right] \tag{5}$$

and for δ_m

$$\left[\sum_{i=1}^{I} c_{mi}\hat{\theta}_i \pm z_{1-\alpha/(2M)}\sqrt{\sum_{i=1}^{I} c_{mi}^2\hat{V}\left(\hat{\theta}_i\right)}\right] \tag{6}$$

where $z_{1-\alpha/(2M)}$ denotes the $(1-\alpha/(2M))$ quantile of a standard normal distribution. This computationally simple standard approach based on asymptotic normality and the Bonferroni adjustment, is referred to as the **ANB** method.

For the comparison of two samples, Chen and Zhou (2006) consider two methods based on likelihood profiles and a method based on generalized pivotal quantities. It is difficult to generalize profile likelihood methods to multiple comparison procedures, because it is computationally demanding to build the profile over an M-dimensional grid and it is unclear which quantile to use in the M-dimensional problem, to take into account the correlations among the multiple contrasts. However, Krishnamoorthy and Mathew (2003) described a method based on generalized pivotal quantities that can be used to construct simultaneous confidence intervals. For the ith sample, the generalized pivotal quantity is given by Eq. (7):

$$T_i = \hat{\mu}_i - \frac{Z_i\bar{\sigma}_i}{U_i\sqrt{n_i}/\sqrt{n_i-1}} + \frac{\bar{\sigma}_i^2}{2U_i^2/(n_i-1)}, \tag{7}$$

where Z_i is a standard normal random variable $Z_i \sim N(0,1)$ and U_i^2 is a χ^2 random variable with $n_i - 1$ degrees of freedom. The quantity T_i is free of unknown parameters in the sense that n_i is determined by the experimental design, $\hat{\mu}_i$ and $\hat{\sigma}_i$ can be calculated from the sample w_{ij}, and Z_i and U_i^2 have well-defined distributions with known parameters. Confidence intervals for ratios and differences of θ_i can be constructed by sampling a large number (K) of values from $Z_i \sim N(0,1)$ and $U_i^2 \sim \chi^2_{n_i-1}$ for each i and denoting them z_{ik} and u_{ik}^2, respectively, with $k = 1,\ldots,K$. Then t_{ik} can be computed in Eq. (8) for all $i = 1,\ldots,I$ and all $k = 1,\ldots,K$

$$t_{ik} = \hat{\mu}_i - \frac{z_{ik}\bar{\sigma}_i}{u_{ik}\sqrt{n_i}/\sqrt{n_i-1}} + \frac{\bar{\sigma}_i^2}{2u_{ik}^2/(n_i-1)}. \tag{8}$$

Extending the approach of Chen and Zhou, 2006, the corresponding quantities for the M ratios $(\rho_1,\ldots,\rho_M)$ and M differences $(\delta_1,\ldots,\delta_M)$ are:

$$r_{mk} = \exp\left(\sum_{i=1}^{I} c_{mi}t_{ik}\right), \quad m = 1,\ldots,M, \text{ and } k = 1,\ldots,K, \tag{9}$$

and

$$d_{mk} = \sum_{i=1}^{I} c_{mi}\exp\left(t_{ik}\right), \quad m = 1,\ldots,M \text{ and } k = 1,\ldots,K. \tag{10}$$

When applying a Bonferroni adjustment to an analysis with M comparisons, the two-sided simultaneous 95% confidence intervals for $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ can be obtained from the empirical $(\alpha/(2M))$ and $(1-\alpha/(2M))$ quantiles in the samples r_{mk}, and d_{mk}, respectively, for $m = 1, \ldots, M$. The simulation study of Chen and Zhou (2006) for the two-sample case ($I = 2$ and $M = 1$) shows that these intervals have much better properties than the asymptotic normal and are easy to compute. However, by applying a Bonferroni adjustment, the correlation structure among the M estimators is ignored, such that they may be more conservative than necessary if M is large and the estimators are positively correlated. In this work, the approach of generalized pivotal quantities with Bonferroni adjustment (**GPQB**) is considered to be a standard.

By using a Bonferroni adjusted standard normal quantile $z_{1-\alpha/(2M)}$, the ANB method above takes into account the number of parameters, M, in the multiple comparisons problem, but ignores possible correlations among the estimators. Treating this situation as a special case of the approach of Hothorn et al. (2008), a method can be defined that uses M-variate normal quantiles $z_{M,\boldsymbol{R},1-\alpha}$ in Eqs. (5) and (6) instead. By the assumption of independence among the groups $i = 1, \ldots, I$, the covariance matrix $\boldsymbol{V}^{\left(\hat{\theta}\right)}$ for $\left(\hat{\theta}_1, \ldots, \hat{\theta}_I\right)^T$ has the diagonal elements $\hat{V}\left(\hat{\theta}_i\right)$ and 0 in all off-diagonal positions. An estimator of the $M \times M$ correlation matrix for the case of multiple differences $\boldsymbol{C\theta}$, denoted $\boldsymbol{R}^{\left(\hat{\theta}\right)}$, follows from standardizing $\boldsymbol{CV}^{\left(\hat{\theta}\right)}\boldsymbol{C}^T$. More explicitly, the element in the mth row and m'th column of $\boldsymbol{R}^{\left(\hat{\theta}\right)}$ is given in Eq. (11):

$$\frac{\sum_{i=1}^{I} c_{mi}c_{m'i}\hat{V}\left(\hat{\theta}_i\right)}{\sqrt{\sum_{i=1}^{I} c_{mi}^2\hat{V}\left(\hat{\theta}_i\right)}\sqrt{\sum_{i=1}^{I} c_{m'i}^2\hat{V}\left(\hat{\theta}_i\right)}}. \tag{11}$$

Given $\boldsymbol{R}^{(\hat{\theta})}$, the R package `mvtnorm` (Genz et al., 2011) can be used to find a two-sided quantile $z_{2,M,\boldsymbol{R}^{(\hat{\theta})},1-\alpha}$ such that $P\left(|Z_m| < z_{2,M,\boldsymbol{R}^{(\hat{\theta})},1-\alpha}, \forall m = 1, \ldots, M\right) = 1-\alpha$, where $\boldsymbol{Z} = (Z_1, \ldots, Z_M)$ is a central, M-variate normal random vector with correlation matrix $\boldsymbol{R}^{(\hat{\theta})}$. Quantiles for the one-sided case, $z_{1,M,\boldsymbol{R}^{(\hat{\theta})},1-\alpha}$, can be chosen to fulfill $P\left(Z_m < z_{2,M,\boldsymbol{R}^{(\hat{\theta})},1-\alpha}, \forall m = 1, \ldots, M\right) = 1-\alpha$. Accordingly, $\boldsymbol{R}^{(\hat{\psi})}$ may be obtained if interest is in $\boldsymbol{C\psi}$, using $\hat{V}\left(\hat{\psi}_i\right)$ from Eq. (4). This method is referred to as **ANM**, as it is based on a normal approximation and a multiplicity adjustment via multivariate normal quantiles.

In the GPQB method in Eqs. (8)–(10), the marginal distributions are well-approximated. However, the joint distribution, which involves potential correlations, is ignored. Both can be accounted for by applying the method of Besag et al. (1995) to construct simultaneous confidence sets for a joint sample of the parameters of interest, here, r_{mk} and d_{mk}. Their method is recalled for r_{mk}:

1. Order r_{mk} separately for each row $m = 1, \ldots, M$, yielding the order statistics $r_m^{[k]}$ and the ranks $r_m^{(k)}$.
2. Build $\max_k = \max\left(\max\left(r_m^{(k)}\right), K + 1 - \min\left(r_m^{(k)}\right)\right)$, for each $k = 1, \ldots, K$.
3. Order $\max_k$, yielding $\max^{[k]}$.
4. Let $q_{1-\alpha}$ denote the nearest integer to $(1-\alpha)K$ and find $k^* = \max^{[q_{1-\alpha}]}$.
5. The lower and upper confidence limits of ρ_m, $m = 1, \ldots, M$ are $\left[r_m^{[K+1-k^*]}; r_m^{[k^*]}\right]$.

Upper confidence limits can be obtained by using $\max_k = \max\left(r_m^{(k)}\right)$ in step 2 and $\left[; r_m^{[k^*]}\right]$ in step 5. Lower confidence limits can be obtained by using $\min_k = \min\left(r_m^{(k)}\right)$ in step 2, ordering $\min_k$, yielding $\min^{[k]}$ in step 3, and $k^* = \min^{[q_\alpha]}$ and $\left[r_m^{[k^*]}; \right]$ in steps 4 and 5 (compare Mandel and Betensky, 2008). Due to the use of ranks, it is not important whether the algorithm is applied for a given sample of r_{mk} on the scale of $\boldsymbol{\rho}$ or $\boldsymbol{\psi}$. Additionally, the correlations of the M dimensions of the joint distribution are taken into account by choosing the order statistics in the dimensions $m = 1, \ldots, M$ according to the quantile of the maximum ranks. If the sampled values r_{mk} show a clearly positive correlation among the parameters of interest $m = 1, \ldots, M$, then this method will yield narrower intervals than the GPQB method. In this paper, this last method is called **GPQ**; on a theoretical basis, it is expected to perform the best among the methods considered.

With the GPQ and GPQB methods it is straightforward to construct simultaneous confidence intervals for more complicated parameters, e.g., ratios of weighted arithmetic means (Hothorn and Djira, 2011). Two $(M \times I)$ contrast matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$ (with elements a_{mi} and b_{mi}, respectively), may define linear combinations of the expected values θ_i, to analyze the M ratios of those linear combinations, $\boldsymbol{A\theta}/\boldsymbol{B\theta}$. Instead of using Eqs. (9) or (10), one may compute $\left(\sum_{i=1}^{I} a_{mi}\exp t_{ik}\right) / \left(\sum_{i=1}^{I} b_{mi}\exp t_{ik}\right)$, for $m = 1, \ldots, M$, and $k = 1, \ldots, K$, and apply the percentile confidence interval methods (Besag et al., 1995) to yield simultaneous confidence bounds. However, this approach is not considered any further in this paper.

Various types of bootstrap can also be used to construct simultaneous confidence intervals. Chen and Zhou (2006) considered a parametric bootstrap approach in their simulation study for comparing two samples. This method can be

adapted to multiple contrasts by obtaining an appropriate quantile by means of bootstrapping the maximum test statistic (Westfall and Young, 1993) over the M contrasts. However, this approach relies on the same parametric assumptions, but approximates the distribution of interest with less precision than the GPQ method (Chen and Zhou, 2006). The simulation results of Chen and Zhou (2006) do not motivate to consider this approach any further. Another option is to use non-parametric bootstrapping, for example, resampling from the observations w_{ij} with replacement and stratification by the treatment groups. For each of the resulting data sets, the parameter of interest can be estimated, and the percentile methods by Besag et al. (1995) or Mandel and Betensky (2008) can be used to construct simultaneous confidence intervals. If the log-normal assumption holds, such an approach should be suboptimal due to resampling a non-pivotal statistic. If the log-normal assumption is clearly violated, it should outperform the remaining approaches. However, the investigation of these problems is beyond the scope of this paper.

2.4. Monte Carlo simulation

A simulation study was performed to assess empirically the properties of two-sided nominal 95% confidence intervals. As a primary criterion, the simultaneous coverage probability is considered, which is defined as the probability that all true parameters ρ_m are included in their respective lower and upper confidence limits ρ_m^l, ρ_m^u, i.e., $P(\rho_m^l \leq \rho_m \leq \rho_m^u,\ \text{for all } m = 1, \ldots, M)$. Because confidence intervals are used for decisions upon directional hypotheses, it can be of interest to assess whether the probability to exclude the true parameter is equal for both the upper and lower limits. For this purpose, the difference between the probability of lower bounds to exclude any true parameters and that of the upper bounds in relation to the overall probability to exclude any true parameter,

$$\frac{P\left(\exists m = 1, \ldots, M : \rho_m^l > \rho_m\right) - P\left(\exists m = 1, \ldots, M : \rho_m > \rho_m^u\right)}{1 - P\left(\rho_m^l \leq \rho_m \leq \rho_m^l \forall m = 1, \ldots, M\right)}, \tag{12}$$

is estimated in the simulation study. In accordance with Chen and Zhou (2006), this quantity is referred to as the relative bias. The above criteria were assessed for $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ and the 23 parameter settings of μ_i and σ_i^2 with $I = 4$, as shown in Table 1. The parameter settings comprise settings with no difference in the expected values among the four groups, due to the equality of μ_i and σ_i^2 (settings 1–5) and despite the inequality of μ_i and σ_i^2 (settings 6–11). Patterns of decreasing and increasing values of θ_i are invoked by either differences in μ_i, while the σ_i^2s are held to be equal (settings 12–15), or by differences in σ_i^2 while the μ_is are held to be equal (settings 16–19), or by differences in both μ_i and σ_i^2 (settings 20–23). All parameter settings are simulated for balanced sample sizes with $n_i = 5, 10, 20, 40, 100$ and four sets of unbalanced sample sizes $(n_1, \ldots, n_4) = (5, 10, 10, 10), (20, 10, 10, 10), (10, 20, 20, 20), (40, 20, 20, 20)$. Three types of contrast matrices $\boldsymbol{C}$ were applied (Eq. (13)), which implement many-to-one comparisons to the first group, all pairwise comparisons and a particular set of pairwise comparisons to the first and second group. Many-to-one comparisons, $\boldsymbol{C}^{(1)}$, are a very common practical problem, leading to only positive correlations among the $M = I - 1$ estimators. All pairwise comparisons, $\boldsymbol{C}^{(2)}$, induce a more complicated, singular correlation matrix, as the contrast matrix contains implicit redundancy. The last contrast matrix, $\boldsymbol{C}^{(3)}$, is included merely to illustrate that the given methods can be used for more general multiple comparison problems customized by the user for particular experimental questions.

$$\boldsymbol{C}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{C}^{(2)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \quad \boldsymbol{C}^{(3)} = \begin{pmatrix} -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}. \tag{13}$$

For each parameter setting, 10 000 random samples were drawn from log-normal distributions, and the above criteria were recorded. The GPQ and GPQB method were applied with $K = 10\,000$. The simulation study was performed using R-2.12.0 (R Development Core Team, 2010). To make the methods defined here available for users, the GPQ and ANM method have been included in the R package `MCPAN 1.1-13` (Schaarschmidt et al., 2011), where the functions `lnrci` and `lndci` implement the methodology for the ratios and the differences, respectively.

3. Results

Fig. 1 shows that, summarizing over all the settings of μ_i, σ_i^2, n_i, contrast types and for both the ratios and differences, the GPQ method provides simultaneous coverage probabilities that are closest to the nominal level. The Bonferroni corrected GPQB has relatively good properties, but it is always more conservative than the GPQ method. The two asymptotic methods are severely liberal in the majority of the cases, but can be conservative in some settings. Violations of the nominal confidence level (or the type-I-error of hypothesis tests) of similar magnitude have been reported for the asymptotic method in the two-sample case (Chen and Zhou, 2006; Krishnamoorthy and Mathew, 2003). The most severe violations are due to small sample settings (see Fig. 2) and situations with large σ_i^2, or σ_i^2s differing among groups (not shown). In general, the asymptotic methods perform better for ratios than for differences.

Table 1
Settings of the parameters μ_i and σ_i^2 in the simulation study.

Setting	μ_1	μ_2	μ_3	μ_4	σ_1^2	σ_2^2	σ_3^2	σ_4^2
1	1.90	1.90	1.90	1.90	0.2	0.2	0.2	0.2
2	3.75	3.75	3.75	3.75	0.5	0.5	0.5	0.5
3	1.00	1.00	1.00	1.00	2.0	2.0	2.0	2.0
4	3.00	3.00	3.00	3.00	2.0	2.0	2.0	2.0
5	−2.00	−2.00	−2.00	−2.00	0.5	0.5	0.5	0.5
6	1.90	1.90	1.90	1.75	0.2	0.2	0.2	0.5
7	1.90	1.90	1.90	1.95	0.2	0.2	0.2	0.1
8	3.75	3.75	3.50	3.50	0.5	0.5	1.0	1.0
9	3.75	3.75	3.90	3.90	0.5	0.5	0.2	0.2
10	1.00	1.00	0.50	0.50	2.0	2.0	3.0	3.0
11	1.00	1.00	1.50	1.50	2.0	2.0	1.0	1.0
12	1.90	1.90	3.90	3.90	0.2	0.2	0.2	0.2
13	1.90	1.90	−1.85	−1.85	0.2	0.2	0.2	0.2
14	1.00	3.00	3.00	3.00	2.0	2.0	2.0	2.0
15	3.00	1.00	1.00	1.00	2.0	2.0	2.0	2.0
16	1.00	1.00	1.00	1.00	2.0	2.0	1.0	0.5
17	1.90	1.90	1.90	1.90	0.2	0.2	0.5	1.0
18	−2.00	−2.00	−2.00	−2.00	0.5	0.5	0.5	0.1
19	−2.00	−2.00	−2.00	−2.00	0.5	0.5	0.5	2.0
20	−1.85	1.75	3.75	3.75	0.2	0.5	0.5	0.5
21	1.95	1.90	3.75	3.50	0.1	0.2	0.5	1.0
22	3.00	1.75	1.90	1.90	2.0	0.5	0.2	0.1
23	2.50	3.00	3.50	1.50	3.0	2.0	1.0	1.0

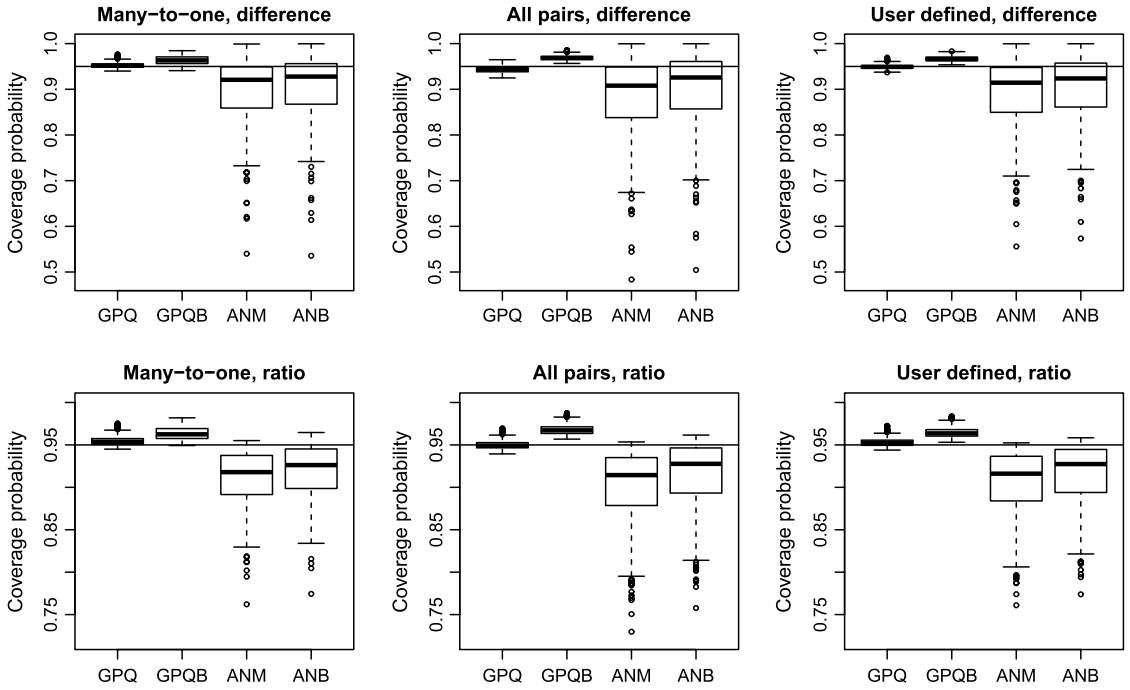


Fig. 1. Simultaneous coverage probability of the four methods, for nominal 95% confidence intervals for ratios and differences as defined by the contrast matrices in Eq. (13). The boxplots summarize results for all settings of μ_i, σ_i^2 and n_i introduced in Section 2.4.

Fig. 2 provides a more detailed comparison of the estimated simultaneous coverage probabilities of the GPQ method relative to the other methods, with restriction to many-to-one comparisons. The dotted lines mark the least extreme values of observed coverage probabilities, for which hypotheses tests against a null of coverage probability =0.95 would be rejected at the 5% level. Thus, for an exact method, only 5% of the settings should be smaller or larger than the dotted lines. Fig. 2(a)

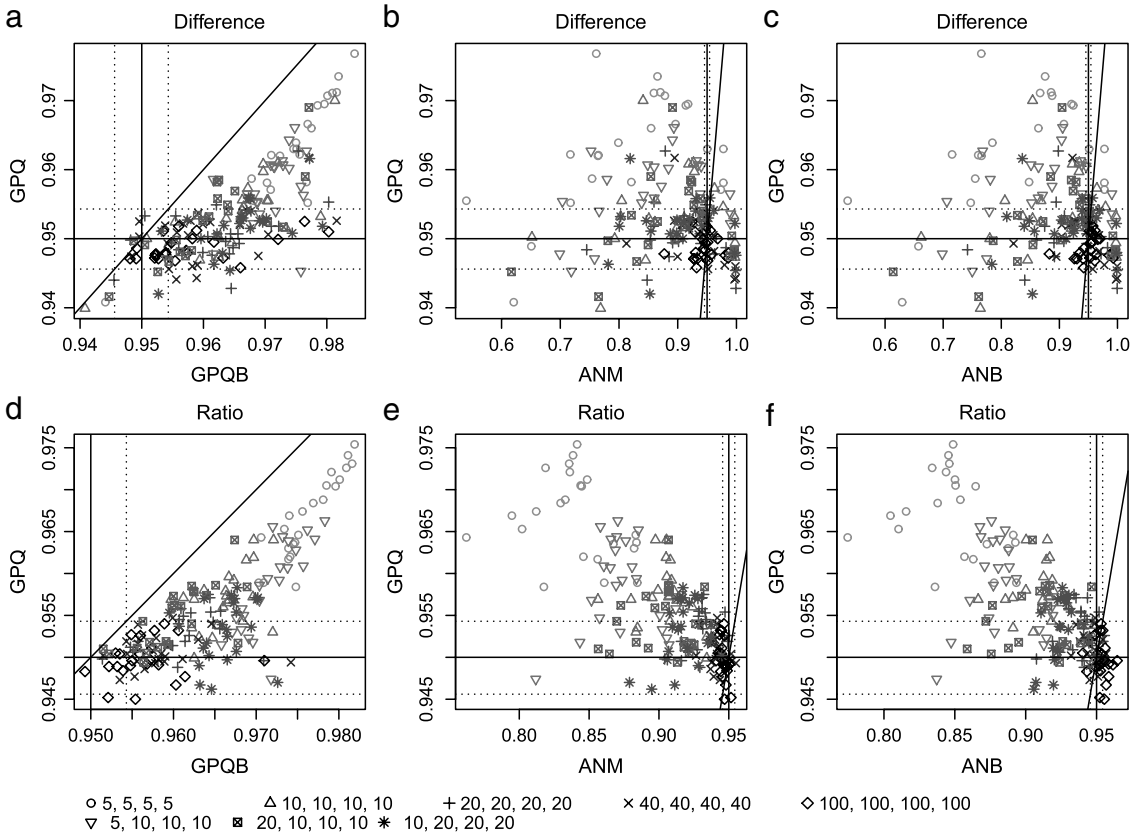


Fig. 2. The simultaneous coverage probability for the GPQ method (y-axis) is plotted against the simultaneous coverage probability of the GPQB, ANM and ANB methods for many-to-one comparisons. Different symbols distinguish the various sample size settings, where a darker gray indicates higher average sample size.

and (d) illustrate that using the GPQ instead of the GPQB method leads to a coverage probability closer to the nominal level in nearly all cases. Nevertheless, the GPQ method is more conservative than would be allowed for an exact method, if sample sizes $n_i = 5, 10$ are involved. Fig. 2(b) and (e) show that those difficult settings with small or unbalanced sample size are mainly responsible for the severely liberal performance of the ANM method, whereas with $n_i = 100$ also this method shows coverage probabilities close to the nominal level. Finally, Fig. 2(c) and (f) show that the unacceptable performance of the ANM method is not due to using estimates of the variances to estimate the correlation matrix; this performance is also evident when the simple Bonferroni correction is used with the normal approximation. The Bonferroni correction, however, may lead to an unnecessarily high coverage probability when sample sizes are larger ($n_i = 40, 100$).

Fig. 3 shows the results for the relative bias (Eq. (12)) of the simultaneous confidence intervals with respect to the occurrence of type-I-errors in the lower and upper tails. Values close to 0 indicate that type-I-errors are equally likely in lower and upper tails, while negative values indicate that true parameters are more likely to be excluded erroneously by upper bounds than by lower bounds. In general, the magnitude of the bias is smaller for ratio intervals than for intervals for multiple differences, and the ANM method shows extreme values of relative bias. Most extreme bias values are found in situations of unbalanced or small sample sizes, and in situations where the σ_i^2 parameters (the groupwise coefficients of variation) are high in value or differ among the groups. In the ANM method, the magnitude and the direction of the relative bias depends strongly on the particular parameter setting; in the GPQ approach, this dependency is less pronounced. This illustrates that the symmetric ANM intervals are inadequate with respect to the skewed distributions.

Additional simulations (results not shown) have been run using unbiased variance estimates $\tilde{\sigma}_i^2$ and the denominator $(n_i - 1)$ in the computation of the ANM and ANB methods (Krishnamoorthy and Mathew, 2003). With this modification, the general patterns of the coverage probability and the relative bias are the same. However, the methods are slightly less liberal, and the amount of relative bias is reduced. The minimal coverage probabilities are approximately 0.85 and 0.68, whereas with ML estimates, values of only 0.75 and below 0.60 are observed. Thus, such an adjustment improves the ANM and ANB methods, but the improvements are not large enough to recommend ANM or ANB for small or moderate sample sizes.

For a limited number of parameter settings, the properties of the four methods have been compared under slight violations of the log-normal assumption. With a probability of 0.1, individual values y_{ij} were replaced by values sampled

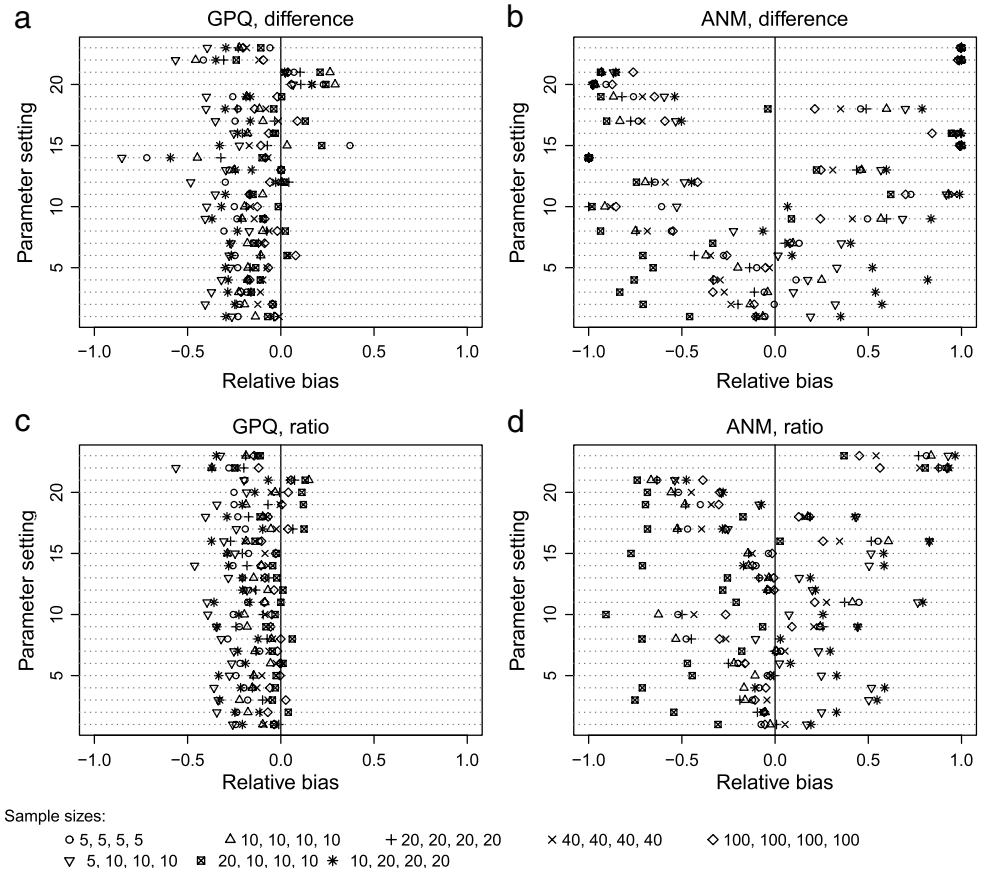


Fig. 3. Relative bias (Eq. (12)) of the GPQ and ANM method for the 23 parameter settings given in Table 1 (y-axis), where different symbols distinguish the different sample size settings.

from a normal distribution with identical μ_i but doubled values of σ_i^2, thus adding a small amount of outlying observations to both tails of the log-normal distributions specified in Table 1. For the GPQ method, this results in increased absolute values of the relative bias and more liberal performances for large sample sizes ($n_i = 100$) in single parameter settings. The minimal coverage probability observed for the GPQ method was 0.92 for the nominal 0.95 confidence intervals. The ANM method and ANB method show unacceptably low coverage probabilities and extreme relative bias values under these conditions. The details of these additional simulations are available from the author upon request.

4. Example

The example data set (Hand et al., 1994) consists of 57 observations of nitrogen bound bovine serum albumin in mice, where animals are assigned to $I = 3$ treatment groups: normal mice ($n_1 = 20$), alloxan-induced diabetic mice ($n_2 = 18$) and alloxan-induced diabetic mice treated with insulin ($n_3 = 19$). Fig. 4 (a) shows boxplots of the three treatment groups; Fig. 4 (b) and (c) present Q–Q plots of the residuals of a linear model (accounting for differences among treatment groups) based on the original serum albumin values and the log-transformed serum albumin values, respectively.

The original values are obviously right-skewed, while the log-transformed data do not contradict the assumption of a normal distribution. Further, applying the Shapiro–Wilk test to those residuals leads to the rejection of the normality assumption for the original data ($p = 7 * 10^{-6}$); after a log-transformation, the hypothesis of normality cannot be rejected ($p = 0.315$). Hence, the log-normal assumption appears to be reasonable for the given data set.

Table 2 lists sample estimates for μ_i, σ_i and expected values θ_i. A possible experimental question is whether the mean serum albumin levels in the two treated groups are changed with respect to the mean serum albumin level in the control group of normal mice. This leads to applying the contrast matrix $\mathbf{C}_{(\mathbf{1})}$ from (13) with the last column and the last row omitted. Because the ANB and GPQB methods are unnecessarily conservative in some cases, while the ANB does not avoid the severely liberal performance of the normal approximation in other cases, this example is restricted to showing results for the ANM and GPQ methods. To illustrate the empirical distribution of the two ratios and two differences obtained using the GPQ method, Fig. 5 shows scatterplots of $K = 10000$ values of r_{mk} and d_{mk} as well as histograms for the marginal distributions

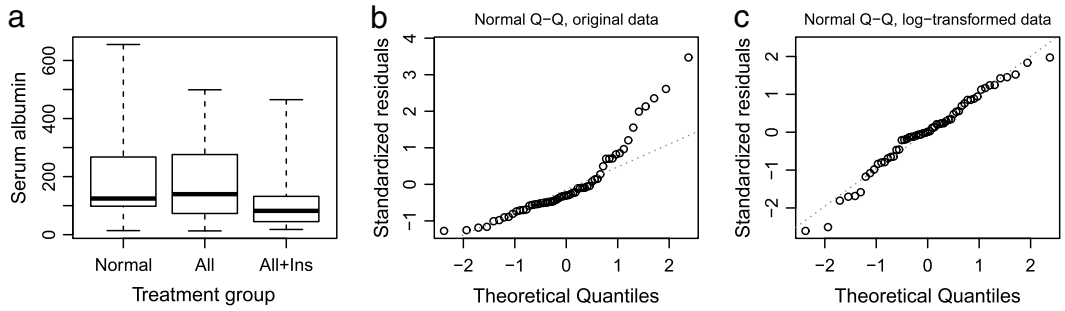


Fig. 4. Boxplot (a), Normal Q–Q plot of the residuals of a 1-way-ANOVA model for the original (b), and log-transformed serum albumin values (c).

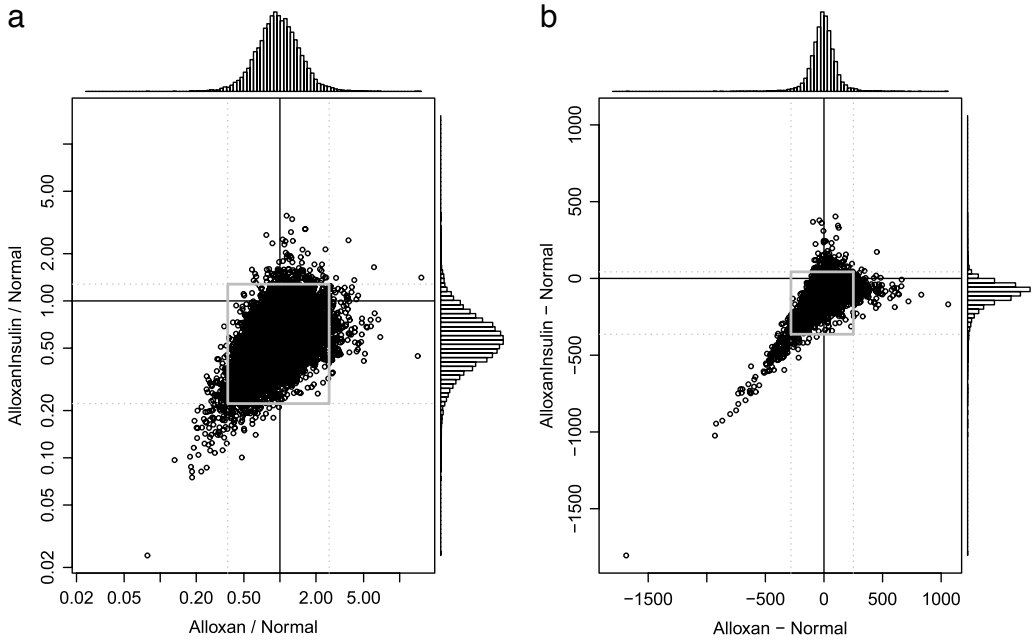


Fig. 5. Scatterplots for the joint distribution and histograms of the marginal distributions of for (a) the two ratios, r_{mk}, and (b) the two differences, d_{mk}. Solid gray boxes and dotted gray lines show the confidence set obtained by the GPQ method and the projection to the axes, respectively.

Table 2
Groupwise parameter estimates for the example.

Group i	n_i	$\hat{\mu}_i$	$\bar{\sigma}_i$	$\hat{\theta}_i$
Normals	20	4.859	0.963	205.1
Alloxan	18	4.867	0.922	198.7
AlloxanInsulin	19	4.397	0.834	115.0

for each $m = 1, 2$. It is clear that the joint distribution underlying the GPQ confidence intervals is not elliptic (as is assumed when using the normal approximation), and the marginal distributions are skewed, particularly for the differences to the control. The gray boxes are the confidence regions obtained by the GPQ method; they contain the central 9500 sampled values r_{mk} and d_{mk}. Table 3 presents the two-sided nominal 95% simultaneous confidence intervals for the resulting ratios and differences to the control group, according to the ANM and GPQ methods. At the 5% significance level, no change between the two treated groups and the control group can be inferred. In general, the ANM intervals are narrower than the GPQ intervals, which is consistent with the results of the simulation study.

5. Discussion

This paper describes methods to compute simultaneous confidence intervals in multiple comparisons for ratios or differences of expected values of several log-normal samples in a one-way layout. Simple asymptotic methods, which rely

Table 3
Two-sided nominal 95% simultaneous confidence intervals for ratios and differences to the control group.

		ANM		GPQ	
Comparison	Estimate	Lower	Upper	Lower	Upper
Alloxan/Normal	0.9686	0.4272	2.1958	0.3665	2.5759
AlloxanInsulin/Normal	0.5608	0.2617	1.2016	0.2216	1.2787
Alloxan - Normal	−6.4481	−170.4314	157.5352	−281.7700	250.5247
AlloxanInsulin - Normal	−90.0815	−220.9288	40.7658	−363.5005	42.8610

only on sample estimates for the first and second moments, fail to cover the true parameter vector with the prespecified confidence probability, except in cases where sample sizes are large, e.g., 40 or 100 observations per group. Methods based on generalized pivotal quantities (Weerahandi, 1993) clearly perform better, even for sample sizes as small as 5 observations per group. In addition, taking into account the correlations among the parameter estimates avoids the unnecessarily conservative performance that is a consequence of applying the simple Bonferroni-correction. The use of contrast matrices allows for standard multiple comparisons such as comparisons to a control group, all pairwise comparisons, but also for user-defined comparisons related to non-standard experimental questions. It is straightforward to generalize this approach to other contrast types such as groupwise comparisons to the overall mean. The approaches discussed here are restricted severely in that they are described only for the one-way layout, i.e., they cannot be applied in the presence of covariates, block effects or other secondary sources of variation.

The method based on generalized pivotal quantities is computationally simple, it does not depend on the availability of algorithms to compute multivariate normal quantiles and it approximates a joint distribution of parameters, which is not available analytically. In this respect, it is an interesting option for other multiple comparison problems. For example, multiple contrasts of means of unbalanced, heteroscedastic normal samples follow a multivariate t-distribution. However, the degrees of freedom of the χ^2 denominators defining the multivariate t distribution differ among the multiple contrasts, i.e., among the dimensions of the multivariate t distribution. Probabilities and quantiles of such a multivariate t-distribution are not available analytically. Hasler and Hothorn (2008) used multiple multivariate t quantiles with different degrees of freedom to address this problem. Related approaches have been proposed recently (Li et al., 2011; Xiong and Mu, 2009); a combination of their approaches with the simple algorithm here could yield competitive performance compared to the method of Hasler and Hothorn (2008) with respect to computation time, theoretical background and computational availability.

Acknowledgments

I thank the anonymous referees as well as L.A. Hothorn and M. Hasler for their helpful and constructive comments on earlier versions of this manuscript. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° HEALTH-F5-2008-201619, and the German Science Foundation DFG-H01678/9.

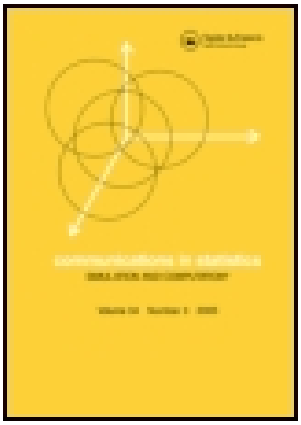
Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2012.08.011.

References

Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and Stochastic-systems. Statistical Science 10, 3–41.
Bretz, F., Genz, A., Hothorn, L., 2001. On the numerical availability of multiple comparison procedures. Biometrical Journal 43, 645–656.
Chen, Y.-H., Zhou, X.-H., 2006. Interval estimates for the ratio and difference of two lognormal means. Statistics in Medicine 25, 4099–4113.
Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096–1121.
Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2011. mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9991.
Hand, D.J., Daly, F., McConway, K., Lunn, D., Ostrowski, E., 1994. A Handbook of Small Data Sets, first ed. Chapman and Hall/CRC, London.
Hasler, M., Hothorn, L.A., 2008. Multiple contrast tests in the presence of heteroscedasticity. Biometrical Journal 50, 793–800.
Hothorn, L.A., Djira, G.D., 2011. A ratio-to-control Williams-type test for trend. Pharmaceutical Statistics 10, 289–292.
Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. Biometrical Journal 50, 346–363.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, second ed. In: Wiley Series in Probability and Statistics, vol. 1. John Wiley & Sons, New York.
Krishnamoorthy, K., Mathew, T., 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of Statistical Planning and Inference 115, 103–121.
Li, X., 2009. A generalized p-value approach for comparing the means of several log-normal populations. Statistics & Probability Letters 79, 1404–1408.
Li, X., Wang, J., Liang, H., 2011. Comparison of several means: A fiducial based approach. Computational Statistics & Data Analysis 55, 1993–2002.
Liu, J., Weng, C., 1994. Evaluation of log-transformation in assessing bioequivalence. Communications in Statistics-Theory and Methods 23, 421–434.
Mandel, M., Betensky, R.A., 2008. Simultaneous confidence intervals based on the percentile bootstrap approach. Computational Statistics & Data Analysis 52, 2158–2165.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.R-project.org/.
Schaarschmidt, F., Gerhard, D., Sill, M., 2011. MCPAN: Multiple comparisons using normal approximation. R package version 1.1–13.
Steinijans, V., Hauschke, D., 1992. Update on the statistical analysis of bioeqivalence studies. International Journal of Clinical Pharmacology and Therapeutics 30, 45–50.
Tian, L., Wu, J., 2007. Inferences on the common mean of several log-normal populations: The generalized variable approach. Biometrical Journal 49, 944–951.
Tukey, J., 1953. The problem of multiple comparisons, unpublished manuscript. In: Braun, H. (Ed.), The Collected Works of J.W. Tukey (1994), Vol. 8. Chapman and Hall, New York, pp. 1–300.
Weerahandi, S., 1993. Generalized confidence-intervals. Journal of the American Statistical Association 88, 899–905.
Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment, first ed. In: Wiley Series in Probability and Statistics, Wiley-Interscience.
Xiong, S., Mu, W., 2009. Simultaneous confidence intervals for one-way layout based on generalized pivotal quantities. Journal of Statistical Computation and Simulation 79, 1235–1244.



Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Simultaneous Confidence Intervals For Ratios of Fixed Effect Parameters in Linear Mixed Models

Frank Schaarschmidt[a] & Gemechis D. Djira[b]

[a] Institute of Biostatistics, Leibniz Universit at Hannover, Herrenh auser Str.2, D-30419, Hannover Germany,

[b] Department of Mathematics and Statistics, South Dakota State University, Brookings, South Dakota, USA

Accepted author version posted online: 23 Jun 2014.

To cite this article: Frank Schaarschmidt & Gemechis D. Djira (2014): Simultaneous Confidence Intervals For Ratios of Fixed Effect Parameters in Linear Mixed Models, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2013.849741

To link to this article: http://dx.doi.org/10.1080/03610918.2013.849741

SIMULTANEOUS CONFIDENCE INTERVALS FOR RATIOS OF FIXED EFFECT PARAMETERS IN LINEAR MIXED MODELS

Short title: Simultaneous inference for ratios in mixed models

Frank Schaarschmidt[a] and Gemechis D. Djira[b]

[a] Institute of Biostatistics, Leibniz Universität Hannover, Herrenhäuser Str.2, D-30419 Hannover, Germany, schaarschmidt@biostat.uni-hannover.de

[b] Department of Mathematics and Statistics, South Dakota State University, Brookings, South Dakota, USA

Key Words: multiple comparisons; Fieller; Gibbs Sampler; coverage probability.

ABSTRACT

In multiple comparisons of fixed effect parameters in linear mixed models, treatment effects can be reported as relative changes or ratios. Simultaneous confidence intervals for such ratios had been previously proposed based on Bonferroni adjustments or multivariate normal quantiles accounting for the correlation among the multiple contrasts. We propose Fieller-type intervals using multivariate t quantiles and the application of Markov Chain Monte Carlo techniques to sample from the joint posterior distribution and construct percentile-based simultaneous intervals. The methods are compared in a simulation study including bioassay problems with random intercepts and slopes, repeated measurements designs and multicenter clinical trials.

1. INTRODUCTION

Inferences for ratios of mean parameters in linear models **are needed** when the parameters of interest are ratios of the original model parameters or ratios of linear combinations of the **original model parameters. For example, this problem arises in relative potency estimations (Zerbe, 1978; Djira, 2010)**. In other cases, expressing the change between treatment means in terms of fold-change (Dilba et al., 2006) instead of absolute differences may be more convenient for interpretations of relevancy as is needed for assessing non-inferiority or superiority.

Trials involving multiple treatments may lead to multiple comparisons and the estimation of simultaneous confidence intervals for multiple, correlated ratio parameters. Commonly used methods to account for multiple comparisons (e.g. Bonferroni, Sidak or Scheffe adjustment) are based on conservative assumptions concerning the correlation structures. Such methods can be improved by taking the correlations into account (Djira, 2010; Dilba et al., 2006; Piepho et al., 2005) through quantiles of the related multivariate *t* or multivariate normal distribution. However, the correlation structure in these problems depends on the unknown ratio parameters. For special cases, exact confidence sets can be constructed (Dilba et al., 2006; Hare and Spurrier, 2007). In more general approaches, methods based on the plug-in of ratio estimates have been proposed (Djira, 2010; Dilba et al., 2006; Piepho et al. 2005).

If trials involve clustered observations their analysis by linear mixed models leads to further complications for estimating simultaneous confidence intervals of fixed effects parameters. Depending on how treatments are assigned to or within clusters, the random effects accounting for the clusters may lead to correlations among the estimates for the fixed effects parameters. The construction of simultaneous confidence intervals for ratios of fixed effect parameters has been considered by Young, Zerbe and Hay (1997) and Djira (2010). Young et al. (1997) adjust for multiple comparisons based on the conservative Scheffe approach, ignoring the correlation structure and imposing some restrictions on the denominators of the ratios. Djira (2010) proposed a method based on multivariate normal quantiles, using an estimate of the correlation structure that depends

both on the estimated covariance matrix of the fixed effects parameters and the estimated ratios.

Different approximations are involved in the construction of simultaneous confidence intervals for ratios in mixed linear models, but their practical impact on the coverage probability has not yet been assessed. In this paper, the simultaneous coverage probability of such intervals is estimated for a number of different models as well as parameter and sample size settings. In particular, the properties of the methods described by Djira (2010) are investigated using multivariate normal and multivariate *t* quantiles as well as Bonferroni adjusted *t*-quantiles. Alternatively, one can directly deal with the uncertainty concerning the covariance structure and the ratio parameters by using objective priors and base inference on samples from the joint posterior distribution of the ratio parameters. Based on this sample, simultaneous confidence intervals can be constructed (Besag et al., 1995).

The rest of the paper is organized as follows. Sections 2.1 and 2.2 lay down the notations and methods for simultaneous confidence intervals for ratios in linear mixed effects models. Section 2.3 describes the simulation setup. And a variety of models of practical interest are described in Section 2.4. Section 3 compares the methods in terms of the empirical coverage probabilities and interval widths. In Section 4, the methods are applied to a real data example. Finally, the findings are discussed in Section 5.

2. METHODS

Consider linear mixed effects models of the general form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{Y}$ is the response variable, $\boldsymbol{X}$ is the design matrix of the fixed effects, and the corresponding vector of fixed effects is $\boldsymbol{\theta} = (\theta_1, ..., \theta_I)'$. $\boldsymbol{Z}$ is the random effects design matrix with corresponding random effects $\boldsymbol{u}$. The residual error $\boldsymbol{e}$ is assumed to be independent of $\boldsymbol{u}$ and both $\boldsymbol{u}$ and $\boldsymbol{e}$ are normally distributed.

Although the model in (1) assumes additivity of effects, in some contexts, it is of interest to further express differential effects in terms of ratio parameters (Young et al., 1997; Djira, 2010). In general, we consider M ratios of linear combinations of the fixed effects parameters. That is,

$$\gamma_m = \frac{\boldsymbol{c}_m\boldsymbol{\theta}}{\boldsymbol{d}_m\boldsymbol{\theta}}, \quad m = 1, ..., M \tag{2}$$

The row vectors $\boldsymbol{c}_m = (c_{m1}, ..., c_{mI})$ and $\boldsymbol{d}_m = (d_{m1}, ..., d_{mI})$ contain known constants determining which (linear combinations of) elements of $\boldsymbol{\theta}$ are used to define the mth ratio. The M ratios are defined by two matrices (numerator and denominator) $\boldsymbol{C}$ and $\boldsymbol{D}$ of dimension $(M \times I)$, and with elements c_{mi} and d_{mi}, respectively.

2.1. SIMULTANEOUS FIELLER CONFIDENCE INTERVALS

From a linear mixed effects model fit, estimates for the fixed effects parameters, $\hat{\boldsymbol{\theta}}$, and its covariance matrix $\widehat{V(\hat{\boldsymbol{\theta}})}$ can be obtained. Approximate simultaneous confidence intervals for $\gamma_1, ..., \gamma_M$ can be obtained as follows (Djira, 2010). Let

$$W_m = (\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\hat{\boldsymbol{\theta}}. \tag{3}$$

Following the asymptotic normality of estimates of the fixed effects parameters (e.g., Verbeke and Molenberghs, 2000), the joint distribution of W_m, $m = 1, ..., M$ is approximately M-variate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, with elements $\sigma_{mm'}$

$$\sigma_{mm'} = Cov\,(W_m, W_{m'}) = (\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,V(\hat{\boldsymbol{\theta}})\,(\boldsymbol{c}_{m'} - \gamma_{m'}\boldsymbol{d}_{m'})'\,. \tag{4}$$

Simultaneous $(1-\alpha)100\%$ Fieller confidence intervals can be constructed by solving the equations

$$\frac{\left[(\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)\,\hat{\boldsymbol{\theta}}\right]^2}{(\mathbf{c}_m - \gamma_m\boldsymbol{d}_m)\,\widehat{V(\hat{\boldsymbol{\theta}})}\,(\boldsymbol{c}_m - \gamma_m\boldsymbol{d}_m)'} = z^2_{M,1-\alpha,\hat{R}}, \tag{5}$$

for γ_m, $m = 1, ..., M$.

The critical point $z_{M,1-\alpha,\hat{R}}$ in the above equations is a two-sided $(1-\alpha)100\%$ equicoordinate

quantile of a central M-variate normal distribution with mean $\mathbf{0}$ and correlation matrix $\hat{\boldsymbol{R}}$. $\hat{\boldsymbol{R}}$ is obtained by evaluating Equation (4) at the estimates $\hat{\gamma}_m = \boldsymbol{c}_m\hat{\boldsymbol{\theta}}/\boldsymbol{d}_m\hat{\boldsymbol{\theta}}$ and $\widehat{V(\hat{\boldsymbol{\theta}})}$, resulting in an estimate $\hat{\boldsymbol{\Sigma}}$, which is then standardized by its diagonal elements. The elements $\hat{\rho}_{mm'}$ of the correlation matrix $\hat{\boldsymbol{R}}$ are then: $\hat{\rho}_{mm'} = \hat{\sigma}_{mm'}/\left(\sqrt{\hat{\sigma}_{mm}}\sqrt{\hat{\sigma}_{m'm'}}\right)$. The correlations occur due to the interrelated parameter estimates from the model or several comparisons involving the same parameter estimates.

Note that the above computation involves some approximations. First, the critical value for inverting the test statistic in Equation (5) depends on the unknown ratio parameters and the unknown covariance matrix. Both quantities are replaced by their estimates. For small samples, a multivariate t-distribution should provide a better approximation than the multivariate normal distribution, that is replacing the multivariate normal quantile $z_{M,1-\alpha,\hat{R}}$ by a multivariate t quantile, $t_{M,1-\alpha,\hat{R},\nu}$, with a common number of degrees of freedom ν. However, different options exist for computing the degrees of freedom in mixed linear models. The situation for ratios is complicated by the fact that estimating a degree of freedom should involve not only sample sizes and variance components but also the unknown ratio parameter. Easily, situations arise where M different degrees of freedom would be most appropriate for the M dimensions of the multivariate t distribution. The computation of quantiles for this type of multivariate t distributions is not yet available. In the following, an adjustment of the degrees of freedom for small sample size will be used (Pinheiro and Bates, 2000). For all cases considered here, this adjustment corresponds to the containment method in SAS (Littell et al., 1996), i.e., only the structure of nesting of the fixed effect parameters in the random effects is taken into account. Note that, depending on the nesting of treatments within fixed effects, the degrees of freedom can be equal to the residual degrees of freedom or much smaller. In the following, the method based on multivariate normal quantiles will be called Fieller multivariate z, while the modification using multivariate t quantiles will be called Fieller multivariate t.

An approach that avoids the plug-in of estimates to obtain quantiles is to use Bonferroni adjusted quantiles of univariate t-distribution with ν degrees of freedom, i.e., $t_{1-\alpha/(2M),\nu}$. Bonferroni adjustment takes the number of comparisons M into consideration but does not account for the

correlations directly hence yielding conservative results. This method will be referred to as Fieller Bonferroni *t*.

2.2. SIMULTANEOUS CONFIDENCE INTERVALS BASED ON MCMC

The linear mixed effects model in Equation (1) is inherently hierarchical in nature. An alternative and flexible approach that accounts for the hierarchy and also allows as to directly deal with the distribution of ratio estimates is to use Bayesian statistics. Markov Chain Monte Carlo (MCMC) methods allow to sample from the joint posterior distribution of the ratio parameters (e.g., Gelman and Hill, 2007). By this approach, the uncertainty concerning the covariance matrix, the ratios, and their combined effect on the correlation is implicitly included in the joint posterior distribution of the parameters of interest. If the priors on $\boldsymbol{\theta}$ and hyperpriors on $\boldsymbol{u}$ are chosen to be non-informative, the posterior distribution of $\boldsymbol{\theta}$ is primarily influenced by the data and the model assumptions. In this case, the Bayesian credible intervals will be generally comparable to that of frequentist confidence intervals.

Suppose that we have K draws from the joint posterior of $\boldsymbol{\theta}$. Samples from the joint posterior distribution of the ratios can be generated as $g_{mk} = \boldsymbol{c}_m\boldsymbol{\theta}^k / \boldsymbol{d}_m\boldsymbol{\theta}^k$, $m = 1, ...M$, $k = 1, ..., K$, where $\boldsymbol{\theta}^k$ is the kth draw from the posterior distribution of $\boldsymbol{\theta}$. A method to establish simultaneous $(1-\alpha)100\%$ credible intervals for $\gamma_m, ..., \gamma_M$ from such a sample has been described by Besag et al. (1995).

2.3 SIMULATION STUDY

A frequentist simulation study was performed to assess the performance of the rectangular confidence sets described in Sections 2.1 and 2.2. The simultaneous coverage probability $P\left(\gamma_m^{(l)} \leq \gamma_m \leq \gamma_m^{(u)}, \forall m = 1, ..., M\right)$ was estimated for two-sided 95% confidence intervals. The number of simulation runs for each parameter was set at 10000 for the simultaneous Fieller intervals in Section 2.1. Due to infeasible computing time, 1000 simulation runs were used for the MCMC. With 1000 replications, the standard error of the estimated coverage probabilities would be 0.0069 for an exact 95% confidence interval. To gauge the uncertainties in the simulated cover-

age probabilities, the thresholds for which an observed coverage probability deviates significantly from the nominal 95% level, are shown as dotted lines in Figures 1-4. Comparing the Fieller-type and MCMC based methods can be tricky since depending on the parameter settings, there may be a non-negligible probability that the Fieller-type intervals have no solution, or that the set consists of two separate semi-finite intervals (Young et al., 1997). In such cases, the confidence limits are set to $[-\infty, \infty]$.

As a secondary criterion, the mean confidence interval width was recorded. The widths of the confidence intervals severely differ between the different models and parameter settings introduced below. The distribution of interval widths is highly skewed, in particular for ratios larger than 1. Occasionally, very wide intervals may occur that are still bounded and have high influence on the observed mean interval width. Thus, the different methods are compared by showing the ratios of the 1% trimmed means of interval widths, for each parameter setting.

2.4 THE MODELS

Four linear mixed models of practical interests that involve ratio estimation problems are considered here. They are (i) relative potency estimation, (ii) slope ratio estimation, (iii) multiple comparisons in repeated measurements designs, and (iv) multiple treatment comparisons in terms of ratios in multicenter clinical trials. For all models, the fixed effects parameters have been varied such that the case of equality of all treatments as well as cases of decreasing or increasing treatment effects are considered. Since inferences in linear mixed effects models are based on Wald tests, the effect of sample size has also been investigated.

(i) Relative potency estimation in parallel-line assays. Consider parallel-line assays with random batch effects. If there are various batches to which treatments have been applied, random effects on slope and intercepts are introduced to model the between batch variability as

$$y_{jsq} = \alpha_j + \beta x_{jsq} + a_s + b_s x_{jsq} + e_{jsq} \tag{6}$$

where y_{jsq} is the response of the sth subject in the jth treatment at time q, α_j is the population intercept for the jth treatment, β is the common population slope for all treatments. Here, $\boldsymbol{\theta} = (\alpha_1, \alpha_2, ..., \alpha_J, \beta)'$, a_s and b_s are random intercept and slope for the sth subject and they jointly follow a bivariate normal distributed with zero means, variances σ_a^2 and σ_b^2, and covariance $\rho\sigma_a\sigma_b$. The interest is the relative potencies among the J treatments, e.g. multiple relative potencies compared to a standard treatment, say $j = 1$. With $J = 4$ treatments, the three ($M = 3$) relative potency parameters of interest are $\gamma_i = (\alpha_i - \alpha_1)/\beta$, $i = 2, 3, 4$. The numerator ($\boldsymbol{C}$) and denominator ($\boldsymbol{D}$) contrast matrices described in Section 2 will be

$$\boldsymbol{C} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Using block matrices, the contrast matrices are given by $\boldsymbol{C} = (-\mathbf{1}_{3\times1}\ \boldsymbol{I}_3\ \mathbf{0}_{3\times1})$ and $\boldsymbol{D} = (\mathbf{0}_{3\times4}\ \mathbf{1}_{3\times1})$. Simulations were run for balanced samples sizes of $10, 20$ and 40 subjects per treatment, five variance parameters, four configurations of intercepts (all intercepts equal, intercepts decreasing compared to control and two sets with decreasing intercepts compared to the control). The results are succinctly displayed in Figure 1.

(ii) Slope ratio assays. A different situation arises if a common intercept α is assumed in combination with J treatment-specific slopes β_j:

$$y_{jsq} = \alpha + \beta_j x_{jsq} + a_s + b_s x_{jsq} + e_{jsq} \tag{7}$$

Here, $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2, ..., \beta_J)'$ and treatments will be compared using multiple slope ratios. For $J = 4$ treatments and $M = 3$ comparisons to a control treatment ($j = 1$), the associated numerator and denominator contrast matrices will be $\boldsymbol{C} = (\mathbf{0}_{3\times2}\ \boldsymbol{I}_3)$ and $\boldsymbol{D} = (\mathbf{0}_{3\times1}\ \mathbf{1}_{3\times1}\ \mathbf{0}_{3\times3})$.

For this model, balanced sample sizes of 10, 20, and 40 subjects per group, four variances as well as three configurations of slope parameters have been considered (all slopes equal, slopes

increasing or decreasing compared to the control). The results are summarized in Figure 2.

(iii) Repeated measures designs. This model involves multiple treatments that are randomly assigned to subjects (e.g., patients). For each experimental unit, one observation is obtained at time points $q = 1, ..., Q$. Of interest is treatment, time, or treatment by time interaction effects. Of particular interest is the time-by-treatment interaction. In this model, the correlation of measurements within subjects is modeled by including subject-specific random intercepts, namely $a_{js} \sim N(0, \sigma_a^2)$. This will induce a covariance structure known as exchangeable or compound-symmetry (CS).

$$y_{jsq} = \mu_{jq} + a_{js} + e_{jsq} \tag{8}$$

In this model, the fixed effects are ordered primarily according to time and then by treatment as $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, ..., \mu_{1Q}, \mu_{21}, \mu_{22}, ..., \mu_{2Q}, ..., \mu_{J1}, ..., \mu_{JQ})'$. With this definition, ratios of two treatments to a control treatment (say, $j = 1$), separately for each of three time points, would be defined by the contrast matrices $\boldsymbol{C} = (\boldsymbol{0}_{6\times3}\ \boldsymbol{I}_6)$ and $\boldsymbol{D} = \boldsymbol{I}_{2\times1} \otimes (\boldsymbol{I}_3\ \boldsymbol{0}_{3\times6})$.

For this model, three types of treatment numbers and contrasts have been considered: the above type of contrast is considered for $J = 4$ treatments with $Q = 3$ time points, resulting in $M = 9$ ratios; the comparisons of time points to the first time point within each treatment ($J = 4$, $Q = 3$, $M = 8$); a three arm trial aiming at the estimation of the ratio of (treatment - placebo) to (active - placebo) (Pigeot et al., 2003) for each of three time points ($J = 3$, $Q = 3$, $M = 3$) are additional types of comparisons considered in the simulation study. The number of subjects per group was hold to be equal at 5, 10, 20 or 40 subjects in each treatment group; additionally the following unbalanced settings were considered: $(32, 56, 56, 56)$, $(8, 24, 24, 24)$, $(4, 12, 12, 12)$ subjects in $J = 4$ treatment groups and, $(30, 60, 60)$, $(10, 25, 25)$, $(6, 12, 12)$ for $J = 3$. The two variance components were set $\sigma_a^2 = 2$ and $\sigma_e^2 = 2$ such that the within subject correlation is 0.5 for all settings. Three configurations of the fixed effect parameter were considered: The treatment means over time were hold to be equal for all treatments and time points; the treatment effects with increasing or decreasing means compared to control were simulated, were the magnitude of effects

changed over time.

(iv) Multiple comparisons in multicenter trails. This scenario differs with respect to the structure of nesting. The response of the sth patient in center h assigned to receive treatment j is denoted by y_{jhs}. The population effects of the treatments $j = 1, ..., J$ are denoted μ_j. The random effects contain the variability among centers, modeled by $a_h \sim N(0, \sigma_a^2)$ and a putative variability in efficacy of the treatments among the centers, modeled by the treatment-center interaction term, $b_{jh} \sim N(0, \sigma_b^2)$, $h = 1, ..., H$. Again, e_{jhs} models the residual error of the sth patient in center h randomized in treatment arm j.

$$y_{jhs} = \mu_j + a_h + b_{jh} + e_{jhs} \tag{9}$$

For the fixed effect parameter $\boldsymbol{\theta} = (\mu_1, ..., \mu_J)'$, $M = 3$ comparisons to control ($j = 1$) among $J = 4$ treatments in terms of ratios is defined using the contrast matrices $\boldsymbol{C} = (\mathbf{0}_{3\times1}\ \boldsymbol{I}_3)$ and $\boldsymbol{D} = (\mathbf{1}_{3\times1}\ \mathbf{0}_{3\times3})$.

In the simulation, three different sets of variance components (see Figure 4) as well as four configurations of fixed effect parameters: all treatments hold to be equal, decreasing means compared to control and two settings of increasing means compared to the control treatment. Sample sizes within center were balanced with 10, 20, 50, or 100 in a setting with five centers, or 10, 20, and 50 in settings with 10 centers. Five settings with unbalanced samples sizes between centers were considered: $(150, 150, 100, 50, 50)$, $(60, 60, 40, 20, 20)$, $(30, 30, 20, 10, 10)$, $(120, 120, 40, 40, 40, 40, 40, 20, 20, 20)$ and $(48, 48, 16, 16, 16, 16, 16, 8, 8, 8)$.

2.5 SOFTWARE AND TECHNICAL DETAILS

For data generation, the R software (R Development Core Team, 2010) was used. Linear mixed models were fitted by the add-on package nlme (Pinheiro et al., 2010), and the Fieller-type intervals (Section 2.1) were computed using the packages mratios (Djira et al., 2012) and mvtnorm (Genz et al., 2010). The Gibbs Sampler as implemented in OpenBUGS 3.0.3 (Spiegelhalter et al., 2007) and the R-package R2WinBUGS (Sturtz et al., 2005) was used to obtain samples of the posterior distri-

bution of the fixed effects parameters. When defining the above models in OpenBUGS, the priors for fixed effects parameters were independent normal with expectation 0 and precision parameter $\tau = 1/\sigma^2 = 0.0001$. Priors for the squared variance parameters were taken as uniform distributed $U(0, 100)$ (Gelman and Hill, 2007). The BUGS models used in the simulations can be found in the supplementary materials **available at http://www.biostat.uni-hannover.de/software.html**. For each simulated data set, 70000 updates were drawn, with initial 20000 discarded, and 1 out of 10 values retained in the remaining chain. Thus, the simultaneous intervals described in Section 2.2 are based on a sample of $K = 5000$ values from the joint posterior distribution.

3. RESULTS

The scatter plots in Figures 1 to 4 compare the coverage probabilities of the four methods, with reference to the Fieller Bonferroni *t* method. Summarizing the results below, the Fieller multivariate *z* method can be clearly liberal, while the two Fieller-type methods using *t* approximations and the MCMC-based intervals both are closer to the nominal level. The Fieller Bonferroni *t* method is slightly conservative. It rarely exhibits too low coverage probabilities except for the slope ratio setting with sample size 10 per group (Figure 2), but shows too high coverage probabilities in a number of parameter settings for all models. The Fieller multivariate *z* method is particularly liberal when either the number of observations is small or the number of clusters to estimate the variance components is small (Figure 4). The Fieller multivariate *t* method is less liberal, except for a number of small sample settings in the estimation of slope ratios (Figure 2) and relative potencies (Figure 1). In the remaining problematic cases the coverage probability of the Fieller multivariate *t* method is very close to the nominal level (Figures 2 and 3) or tends to be conservative for small sample sizes and the low number of centers (Figure 4). The MCMC-based intervals do not show systematically liberal performance for most of the parameter constellations considered. Occasionally, too low coverage probabilities are observed for estimating slope ratios and relative potencies, which are also problematic with the Fieller-type intervals. However, the MCMC-based intervals

are slightly more conservative than the computationally much simpler Fieller multivariate *t* method in settings with small sample sizes in the relative potency and repeated measures situations (Figures 1 and 3) and in the multicenter setting with small sample sizes and the lower number of centers (Figure 4). For the multicenter setting with only five centers and small total sample sizes, the MCMC based method is even slightly more conservative than the Fieller Bonferroni *t* method.

Additional simulations (supplementary material) suggest, that using 'non-informative' inverse gamma priors instead of the uniform priors above, leads to MCMC-based intervals that are less conservative but show occasionally slightly too low coverage probabilities. However, with the few parameter settings considered, this can only be a hint for further investigations.

Problems with convergence and unbounded intervals occurred only in the relative potency and the slope ratio model. About 60-70% of the parameter settings for these models did not show any convergence problems or unbounded solutions; for the remaining parameter settings, usually less than 0.5% of the simulated data sets caused at least one of the two problems. The observed maximum rate of convergence problems occurred in the slope ratio model, Eq.(7), with $n_i = 10$ and $\sigma_a^2 = 1$, $\sigma_b^2 = 0.5$, $\sigma_e^2 = 1$ with about 1.5% of the data sets. Simulations were run for a number of additional parameter settings of the relative potency and slope ratio model, where larger proportions of simulations lead to convergence problems or yielded unbounded solutions for the Fieller-type methods. Such problems occurred in settings with small sample sizes, as 5 subjects per treatment. For these cases, a meaningful comparison of frequentist coverage probabilities is difficult. When treating unbounded intervals in the simulation as covering the true parameter, high proportions of unbounded cases will increase the coverage probability. Using this definition of coverage probability leads to similar interpretations as above (supplementary material): the Fieller multivariate *z* method is even more liberal for such small sample settings, while the Fieller multivariate *t* method can be liberal or conservative to similar or slightly extremer extend as shown above.

Figure 5 shows the ratios of trimmed mean interval widths over all models and parameter

settings between the Fieller multivariate *t* and Bonferroni *t*, as well as between the MCMC and the Fieller Bonferroni *t* method. The simplistic plug-in of the covariance matrix and the ratio estimates to obtain multivariate *t* quantiles results for nearly all considered cases in shorter intervals as compared to ignoring correlations with the Bonferroni t approach. The achieved reduction in interval width rarely exceeds 10%. The MCMC intervals have similar widths as the Fieller Bonferroni *t* method, however, occasionally have been observed to be 10% narrower or wider for some parameter settings. High relative deviations in interval width occur when intervals are wide, usually due to small sample sizes. In the relative potency model and the repeated measurement settings, the MCMC intervals are usually narrower than the Bonferroni *t* intervals. In the slope ratio and the multicenter setting, where the estimation of the variance components is difficult, the MCMC method often yields wider intervals when sample sizes or the number of centers is small.

4. EXAMPLE

Pinheiro and Bates (2000) present an example of 16 rats assigned to three different diets, with treatment groups comprising 8, 4, and 4 rats (package nlme). The body weight of individual rats was observed over a period of 64 days. Although the bodyweights do differ already at the initial time points, it might be of interest to compare the average daily increase in bodyweight in the different diets.

Equation (10) shows the model fitted for this data set,

$$y_{jsq} = \alpha_j + \beta_j x_{jsq} + a_{js} + b_{js} x_{jsq} + e_{jsq}, \tag{10}$$

with y_{jsq} observed bodyweight of subject s within diet j at time q, x_{jsq} the time point, α_j and β_j the population intercepts and slopes for the three diets $j = 1, 2, 3$. Finally, a_{js} and b_{js} denote the rat specific random deviation of slopes and intercepts, for which a bivariate normal distribution is assumed as in Equations (6,7), and the residuals e_{jsq} are assumed to be independently normally distributed. Note that this type of model is not included in the simulation study, and the related model

and multiple comparison problem in Eq.(7) has not been simulated for so small and unbalanced sample sizes and parameter configurations as suggested by this example.

Table 1 gives fixed effects estimates and the estimated covariance matrix obtained by fitting this model using the R package `nlme` (Pinheiro et al., 2010). Whether the average daily increase in weight differs between any of the diets and, if so, to what extent, may be addressed by all pairwise comparisons between diet specific population slopes. With respect to the fixed effects parameter vector $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)'$ these ratios are defined by the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ in Equation (11):

$$\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{11}$$

Applying Equation (4) and the correlation formula in Section 2.1 yields a correlation matrix $\hat{\boldsymbol{R}}$ with the off-diagonal elements $\hat{\rho}_{12} = 0.700$, $\hat{\rho}_{13} = -0.262$, $\hat{\rho}_{23} = 0.505$. The resulting two-sided quantile of the 3-variate *t*-distribution with $\nu = 158$ is $t_{M=3,0.95,\hat{R},\nu=158} = 2.07$. To illustrate the gain in precision of the confidence intervals by accounting for correlations, the quantile resulting from using the simple Bonferroni method is $t_{1-0.05/(3*2),\nu=158} = 2.42$. Whereas ignoring multiple comparisons and using single two-sided *t*-tests for all three comparisons would result in the quantile $t_{1-0.05/2,\nu=158} = 1.98$. The confidence intervals of the Fieller multivariate *t* and Bonferroni *t* method accounting for correlations are shown in Table 2.

Alternatively, one may sample from the joint posterior distribution of the parameters of interest using code close to that used for the models in Eq. (6) and Eq. (7), with a diet specific intercept and slope on the population level (Gelman and Hill, 2007). In the given example, independent normal priors $\alpha_j \sim N(0, 10000)$, $\beta_j \sim N(0, 10000)$, for the diets $j = 1, .., 3$ and uniform priors for the variance components and correlation parameter $\sigma_a^2 \sim U(0, 100)$, $\sigma_b^2 \sim U(0, 100)$, $\sigma_e^2 \sim U(0, 100)$ and $\rho \sim U(-1, 1)$ were used. A single MCMC chain was updated 70000 times, initial 20000 draws discarded and 1 out of 5 values of the remaining chain retained. The resulting sample was free of

autocorrelations assessed by graphical methods (not shown).

The three possible scatter plots of the joint empirical posterior for the ratios β_2/β_1, β_3/β_1, and β_3/β_2 are shown in Figure 6. The posterior means and medians, as well as the limits of two-sided 95% credible sets based on this sample are shown in Table 2.

Comparing the limits resulting from the two methods (Table 2) shows that the MCMC based intervals are slightly wider than the Fieller multivariate *t* and Fieller Bonferroni *t* intervals for all three parameters. The average daily increase in weight in diet 2 is about 1.5 - 6.9 times increased over that in diet 1, with 95% confidence (Fieller multivariate *t*). Based on this data, there is no clear difference detectable with 95% confidence between diet 2 and diet 3, or diet 2 and diet 1.

5. DISCUSSION

In this paper, previously described methods to construct simultaneous confidence intervals for ratios of fixed effect parameters in mixed models have been evaluated by simulating their simultaneous coverage probabilities. The simulation study includes four different linear mixed models, various constellations of sample size, fixed effects and variance components, and also a number of different types of comparisons, representing different multiple comparisons problems. In line with straightforward expectations it is found that using Fieller-type confidence intervals with multivariate normal quantiles may lead to too narrow confidence intervals when sample sizes are small. Using multivariate *t* quantiles with a simple adjustment of the degree of freedom yields confidence sets with coverage close to the nominal level for most of the simulated cases. Though, this method showed coverage probabilities as low as 91% for nominal 95% intervals for some small sample settings. Running MCMC with non-informative priors and constructing simultaneous intervals based on samples of the posterior distribution yields intervals that have appropriate or slightly too high frequentist simultaneous coverage probabilities, and wider intervals than the Fieller multivariate *t* approach.

The Fieller-type intervals have the clear advantage to be computationally simple. A docu-

mented function (gsci.ratio) to compute such intervals for a user defined set of ratios, given estimates of the fixed effects, the corresponding estimate of the covariance matrix and the degree of freedom is available in the R package mratios (Djira et al., 2010). Thus, using this function, a much wider range of models and multiple comparison problems can be addressed than illustrated here. The MCMC based method with non-informative priors has been shown to be comparable to frequentist results. However, applying this method in practice requires careful choice of technical parameters, check of convergence and autocorrelations for any new model and data set, problems which have not been illustrated in the example above. Hence it is technically much more demanding. Moreover, there is generally no really non-informative prior. The particular choices of the distributional assumption and the parameters considered as non-informative may differ between authors, (e.g., Browne and Draper, 2005; Gelman, 2006; Zhao et al., 2006), and of course the choice of the parameters depends on the scaling of the variables in the data set. Hence, using the models in the supplementary material and the example with the given prior distribution for new data still requires thinking and careful check of imposed assumptions, data and results.

ACKNOWLEDGEMENTS

The authors thank D. Gerhard and M. Hasler for helpful comments on an earlier version.

BIBLIOGRAPHY

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10(1)**, 3–66.

Browne, W. J. and Draper, D. (2005). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, **1**, 437–524.

Dilba, G., Bretz, F., and Guiard, V. (2006). Simultaneous confidence sets and confidence intervals for multiple ratios. *Journal of Statistical Planning and Inference*, **136(8)**, 2640–2658.

Djira, G. D. (2010). Relative potency estimation in parallel-line assays - method comparison and

some extensions. *Communications in Statistics - Theory and Methods*, **39(7)**, 1180–1189.

Djira, G. D., Hasler, M., Gerhard, D., and Schaarschmidt, F. (2012). mratios: Inferences for ratios of coefficients in the general linear model. R package version 1.3.17.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, **1**, 515–534.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel Hierarchical Models.* Cambridge: Cambridge University Press.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2010). mvtnorm: Multivariate normal and t distributions. R package version 0.9-95. URL http://CRAN.R-project.org/package=mvtnorm

Hare, D. R. and Spurrier, J. D. (2007). Simultaneous inference for ratios of linear combinations of general linear model parameters. *Biometrical Journal* **49(6)**, 854–862.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

Piepho, H.-P. and Emrich, K. (2005). Simultaneous confidence intervals for two estimable functions and their ratio under a linear model. *The American Statistician*, **59(4)**, 292–300.

Pigeot, I., Schäfer, J., Röhmel, J., and Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, **22(6)**, 883–899.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus.* New York: Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Development Core Team (2010). nlme: Linear and nonlinear mixed effects models. R package version 3.1-97.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). OpenBUGS, Version 3.0.3, http://www.openbugs.info/w/

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12(3)**, 1–16.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models For Longitudinal Data.* New York: Springer.

Young, D. A., Zerbe, G. O., and Hay, W. W. (1997). Fiellers theorem, Scheffs simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effect models. *Biometrics*, **53(3)**, 835–847.

Zerbe, G. O. (1978). On Fieller's theorem and the general linear model. *The American Statistician*, **32**, 103–105.

Zhao, Y., Staudemayer, J., Coull, B. A., and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science* **21**, 35–51.

Table 1: Parameter estimates for the example data.

Parameter	Estimate	Estimated covariance matrix					
		α_1	α_2	α_3	β_1	β_2	β_3
α_1	251	171	0.00	0.00	-0.19	0.00	0.00
α_2	452	0.00	342	0.00	0.00	-0.38	0.00
α_3	503	0.00	0.00	342	0.00	0.00	-0.38
β_1	0.36	-0.19	0.00	0.00	0.01	0.00	0.00
β_2	0.97	0.00	-0.38	0.00	0.00	0.02	0.00
β_3	0.66	0.00	0.00	-0.38	0.00	0.00	0.02

Table 2: Two-sided 95% Fieller multivariate *t* confidence intervals and two-sided 95% credible intervals based on an MCMC sample of 10000 values, for all pairwise ratios of the 3 slope parameters.

		Fieller				MCMC			
		Bonferroni *t*		multivariate *t*		posterior		credible intervals	
Ratio	estimate	lower	upper	lower	upper	mean	median	lower	upper
β_2/β_1	2.685	1.445	7.160	1.467	6.883	2.636	2.660	1.263	8.485
β_3/β_1	1.830	0.825	5.039	0.845	4.845	1.829	1.811	0.673	6.246
β_3/β_2	0.682	0.340	1.182	0.348	1.165	0.700	0.679	0.291	1.307

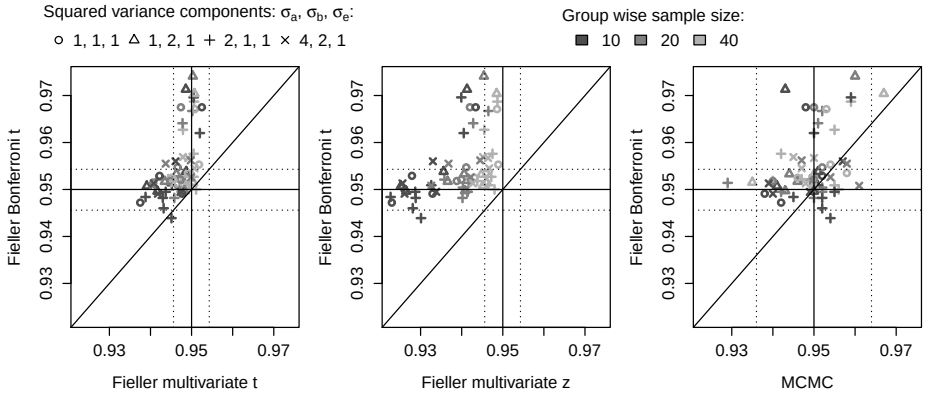


Figure 1: Simultaneous coverage probability for comparison to control of relative potencies, Eq. (6), estimates are based on 10000 and 1000 simulations for Fieller-type and MCMC methods, respectively. Dotted lines indicate thresholds for which the coverage probability differs significantly from the nominal level.

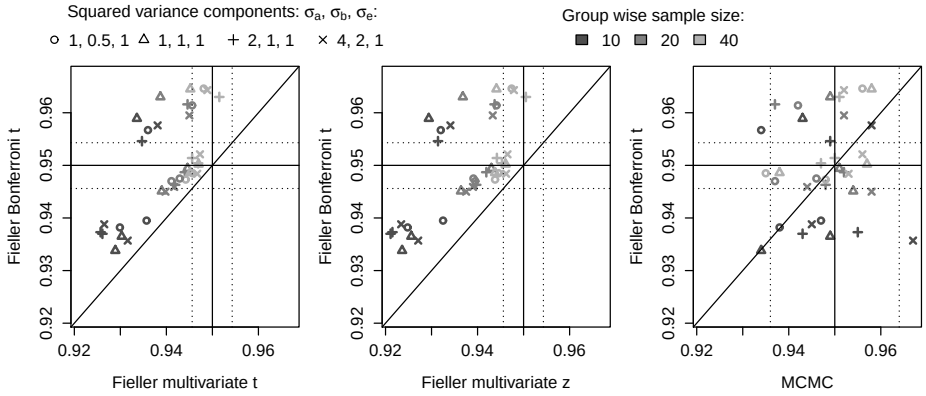


Figure 2: Simultaneous coverage probability for comparison to control in the slope ratio assay, Eq. (7), estimates are based on 10000 and 1000 simulations for Fieller-type and MCMC methods, respectively. Dotted lines indicate thresholds for which the coverage probability differs significantly from the nominal level.

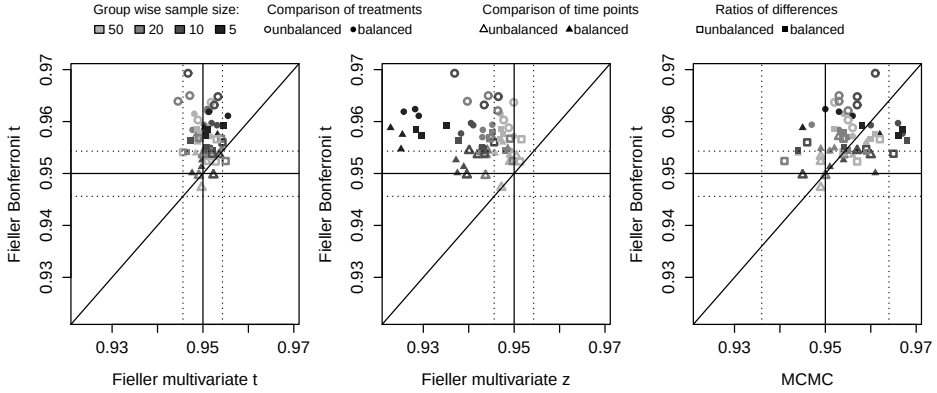


Figure 3: Simultaneous coverage probability for different types of comparisons in a repeated measurement setting, Eq. (8), estimates are based on 10000 and 1000 simulations for Fieller-type and MCMC methods, respectively. Dotted lines indicate thresholds for which the coverage probability differs significantly from the nominal level.

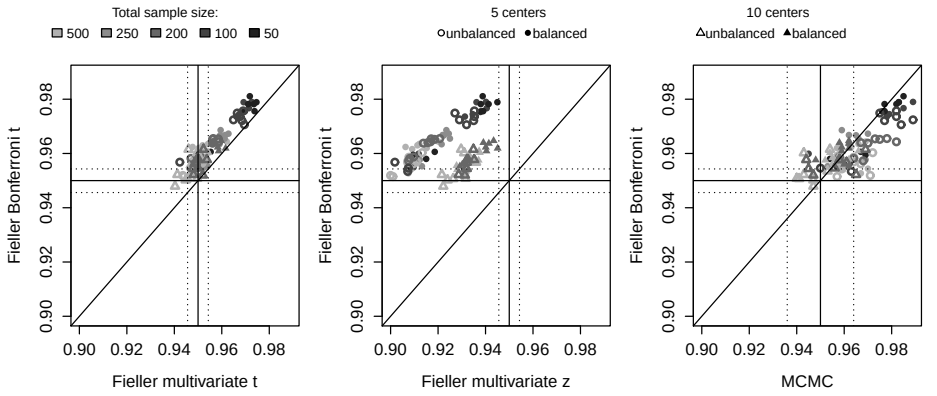


Figure 4: Simultaneous coverage probability for comparison to control in terms of among treatments in multicenter clinical trial, Eq. (9), estimates are based on 10000 and 1000 simulations for Fieller-type and MCMC methods, respectively. Dotted lines indicate thresholds for which the coverage probability differs significantly from the nominal level.

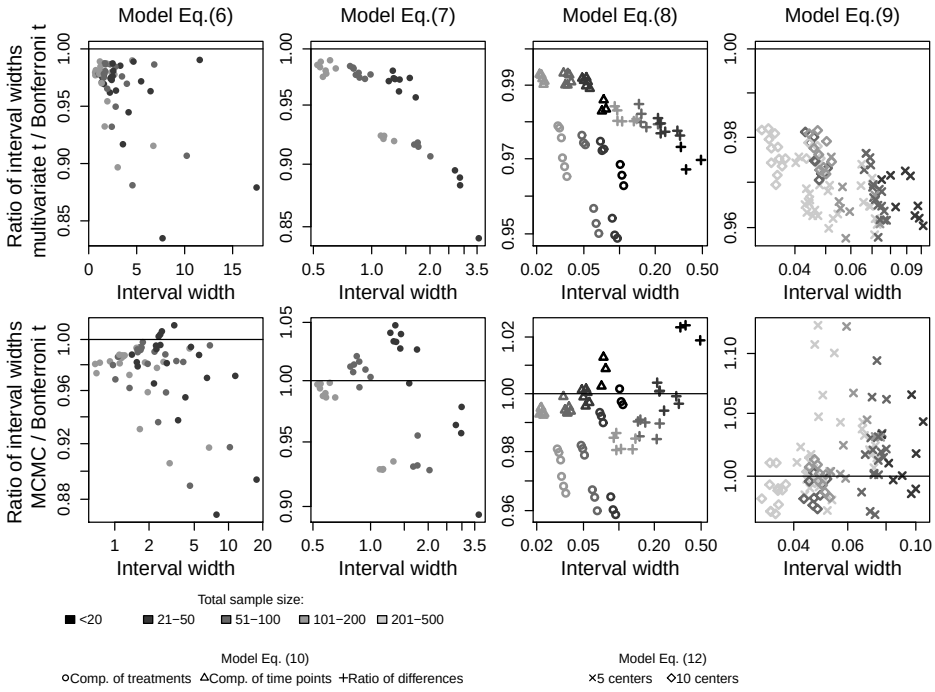


Figure 5: Ratios of (trimmed) mean interval widths of two selected methods (y-axes) vs. the mean interval width averaged across the two methods (x-axes).

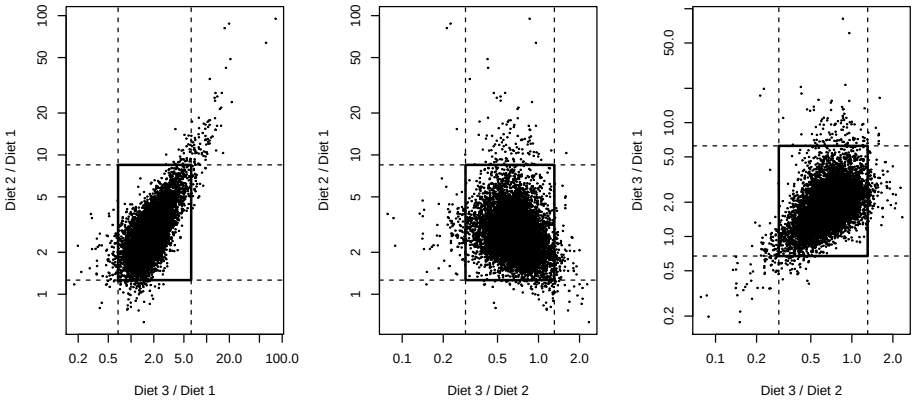


Figure 6: Scatter plots of all pairwise ratios of joint posterior of the slope parameters $\beta_1, \beta_2, \beta_3$. Solid rectangles show the 95% credible sets based on this sample.

Simultaneous confidence intervals for comparisons of several multinomial samples

Frank Schaarschmidt[1], Daniel Gerhard[2], Charlotte Vogel[1]

December 12, 2015

[1] Leibniz Universitaet Hannover, Institute of Biostatistics, Herrenhaeuserstr.2, 30419 Hannover, Germany

[2] University of Canterbury, School of Mathematics & Statistics, Private Bag 4800, Christchurch 8041, New Zealand

Abstract Multinomial data occur if the major outcome of an experiment is the classification of experimental units into more than two mutually exclusive categories. In experiments with several treatment groups, one may then be interested in multiple comparisons between the treatments w.r.t several definitions of odds between the multinomial proportions. We describe asymptotic methods for constructing simultaneous confidence intervals for this inferential problem. Further, alternative methods based on sampling from Dirichlet posterior distributions with vague Dirichlet priors are described. Monte Carlo simulations are performed to compare these methods w.r.t. their frequentist simultaneous coverage probabilities for a wide range of sample sizes and multinomial proportions: The methods have comparable properties for large samples and no rare events involved. In small sample situations or when rare events are involved in the sense that the expected values in some cells of the contingency table are as low as 5 or 10, the method based on sampling from the Dirichlet posterior yields simultaneous coverage probabilities closest to the nominal confidence level. The methods are provided in an R-package and their application is illustrated for examples from developmental toxicology and differential blood counts.

Keywords: multiple comparisons; polytomous data; Dirichlet; baseline logit; coverage probability

1 Introduction

In a number of toxicological assays, the major outcome is the classification of each experimental unit into one of several categories. For example cells may be classified by visual assessment into several categories, where categories distinguish undamaged cells from different types of unusual characteristics or malformation. In clinical trials, the primary outcome may be the classification of individual patients into one of several categories reflecting disease severity, or clinical subtypes of a certain disease. Often, such categories are ordinal. In some applications, however, the order of categories can be ambiguous, that is, there is no clear order of severity among categories, or there may be no order at all, such that the categories are best described as a nominal variable.

In such trials, multiple treatments can be of interest, for example, multiple dose groups compared to a control group in toxicological assays or different therapeutic interventions in a clinical trial. Counting the number of individuals in each category and each treatment group gives rise to a 2-dimensional contingency table with several rows and columns. In the following, we will assume that the individual experimental units are assigned to treatment groups in a completely randomized design and that the sample size per treatment group is fixed by the experimental design (i.e., it is not the result of a random process as, for example, in an epidemiological exposure study). Under these conditions, we may assume that the counts of the different categories follow a multinomial distribution, independently in each treatment group.

Such contingency tables may be analyzed by applying the χ^2 tests for independence. Significance of such a test will only produce the rather general interpretation: The probability to fall into some of the categories does significantly differ between some of the treatment groups. In practice, this will rarely be an exhaustive interpretation of the data. On the contrary, interest will be in more detailed interpretation: Which categories increase or decrease in probability between which of the treatment groups, and if so, by what extent? If multiple comparisons between treatments with respect to several categories contribute to an overall hypothesis in the sense of a union intersection test (e.g. Casella

and Berger, 2002), simultaneous confidence intervals are necessary for such interpretations. But, depending on the application, not all possible comparisons between categories are of interest and not all comparisons between treatments may play a role for the overall hypothesis. Rather, particular categories and treatments in a given assay or trial will give rise to a special set of comparisons which are of interest.

Methods for simultaneous confidence intervals (SCI) in multiple comparisons in contingency tables have been proposed by Gold (1963) and Goodman (1964). Gold (1963) describes an asymptotic Scheffe-type-approach for SCI suitable for all possible linear combinations of the proportions of several multinomial vectors by using a χ^2-quantile with degrees of freedom as in the corresponding global test. Such approaches are inherently two-sided, and the resulting intervals will be unnecessarily large if only a small subset of comparisons (out of all possible comparisons) is of interest. Goodman (1964) considers asymptotic methods for all possible comparisons as well as a selected subset of comparisons of multinomial proportions on the log-scale, assuming a single multinomial distribution for a contingency table with multiple rows and columns (as suitable, e.g. for epidemiological studies). He shows that Bonferroni-adjusted standard normal quantiles may yield narrower intervals than the Scheffe-type approach, when only few comparisons are of interest. Still this approach can be improved because the Bonferroni-adjustment ignores the correlation between the estimators (or the related test statistics).

Since then, numerous authors have considered simultaneous confidence intervals for proportions or pairwise comparisons of proportions in a single multinomial sample (e.g. Glaz and Sison, 1999; Piegorsch and Richwine, 2001; Hou et al., 2003; Wang, 2000; Chafa and Concordet, 2009). To our knowledge, simultaneous confidence intervals for the comparison of multiple odds between multiple multinomial samples have not been considered any further, although there is room for improvement compared to the seminal methods of Gold (1963) and Goodman (1964): The test statistics related to comparisons of multiple logits of multinomial proportions asymptotically follow a multivariate normal distribution (e.g., Agresti, 2013) and multiple multinomial samples can be considered as a special case for the application of multivariate generalized linear models (e.g. McCullagh and Nelder, 1989; Agresti, 2013). One can thus use quantiles of the multivariate normal distribution (Bretz et al., 2001) based on a sample estimate of the correlation structure to construct asymptotic simultaneous confidence intervals according to Hothorn et al. (2008). Such intervals will be narrower than Bonferroni-adjusted intervals in cases where only a limited subset of parameters with correlated estimators is of interest, because their quantiles account for the correlation structure that is ignored by Bonferroni or Scheffe-type approaches. Although all necessary computational methods are available, these methods have so far not been investigated w.r.t. their properties when applied with small sample sizes. Also, they suffer from infinite interval bounds, when single cells of the contingency table happen to contain zeros. Further improvements compared to these asymptotic methods might be achievable by sampling from the joint distribution of interest, e.g. from the posterior of a Bayesian model with a vague prior. Simultaneous confidence intervals can then be computed from such samples by percentile methods as described in Besag et al. (1995), or Mandel and Betensky (2008).

In the remaining part of the paper, we will first describe asymptotic simultaneous confidence intervals for user-defined sets of logits compared between several multinomial samples. Additionally, we will consider simultaneous percentile intervals applied on samples of Dirichlet posteriors with vague Dirichlet priors. The small sample performance of these methods will be compared in frequentist simulation studies. Finally, the methods are applied to two data sets.

2 Material and Methods

2.1 Data structure and notation

We consider $g = 1, ..., G$ treatment groups in a randomized design, where n_g is the sample size in group g that has been fixed by the experimental design. As the experimental outcome, each individual or experimental unit in group g is categorized into exactly one of C possible categories, with index $c = 1, ..., C$. Furthermore we assume that due to the randomized assignment of treatments to individuals or experimental units, there is no further subgrouping of individuals or heterogeneity among individuals and also, that there are no secondary factors or covariates that affect the outcome. Thus we assume that the counted number of individuals of categories $c = 1, ..., C$ in group g, $\mathbf{x}_g = (x_{g1}, x_{g2}, ..., x_{gC})$, follows a multinomial distribution

$$(x_{g1}, x_{g2}, ..., x_{gC}) \sim multinomial\,(n_g, (\pi_{g1}, \pi_{g2}, ..., \pi_{gC}))$$

where π_{gc} is the unknown probability of an individual in treatment group g to fall into category c. Usually, such observed counts are summarized in a contingency table, $\mathbf{X}_{(G\times C)}$.

2.2 Parameters of interest

A simple choice for the analysis of such data is to compare the baseline logits between the groups. That is, the ratios of the latter proportions, $\pi_{g2}, ..., \pi_{gC}$, to that of the first category π_{g1} (the baseline category) are of interest. Treatment effects are then expressed as the relative change of these ratios between the treatment groups. For only two treatment groups, $g = 1, 2$, the odds ratios of interest are then

$$\left(\frac{\pi_{22}/\pi_{21}}{\pi_{12}/\pi_{11}}, \frac{\pi_{23}/\pi_{21}}{\pi_{13}/\pi_{11}}, ..., \frac{\pi_{2C}/\pi_{21}}{\pi_{1C}/\pi_{11}}\right).$$

Depending on the practical meaning of the different categories, more or less parameters than these comparisons to the baseline categories might be of interest. Either, the comparisons of certain categories to baseline may be not of primary interest, or, additional odds, referring to ratios between the proportions of categories $c = 2, ..., C$, may be important. On the log scale, all possible pairwise logits can be written as

$$\begin{pmatrix} \delta_{g1} \\ \delta_{g2} \\ \vdots \\ \delta_{gI} \end{pmatrix} = \mathbf{A}_{(I\times C)} \log\left(\boldsymbol{\pi}_g^T\right) = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -1 & 0 & 0 & \dots & 0 & 1 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix} \begin{pmatrix} \log(\pi_{g1}) \\ \log(\pi_{g2}) \\ \vdots \\ \log(\pi_{gC}) \end{pmatrix}$$

Note that on the scale of baseline logits, $\psi_{gc} = \log(\pi_{gc}) - \log(\pi_{g1})$, $c = 2, ..., C$, all pairwise logits can be written as

$$\begin{pmatrix} \delta_{g1} \\ \delta_{g2} \\ \vdots \\ \delta_{gI} \end{pmatrix} = \mathbf{A}^*_{(I\times C-1)} \boldsymbol{\psi}_g^T = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix} \begin{pmatrix} \psi_{g2} \\ \psi_{g3} \\ \vdots \\ \psi_{gC} \end{pmatrix}.$$

2.3 Between-group comparisons of interest

Similarly, comparisons to a control treatment ('Dunnett-type'), all pairwise comparisons ('Tukey-type') between treatments or a particular subset of these may be of interest, depending on the practical meaning of the G treatment groups for a given experimental question. We can thus write the between-group-comparisons in a contrast matrix $\mathbf{B}_{(J\times G)}$ for the ith logit defined above

$$\boldsymbol{\theta}_i = \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \vdots \\ \theta_{Ji} \end{pmatrix} = \mathbf{B}_{(J\times G)} \boldsymbol{\delta}_i^T = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -1 & 0 & 0 & \dots & 0 & 1 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix} \begin{pmatrix} \delta_{1i} \\ \delta_{2i} \\ \vdots \\ \delta_{Gi} \end{pmatrix}.$$

If these between-group-comparisons $j = 1, ..., J$ are the same for all logits $i = 1, .., I$, the parameter vector can be briefly written as

$$\boldsymbol{\theta} = (\mathbf{B} \otimes \mathbf{A}) \begin{pmatrix} \log \boldsymbol{\pi}_1 \\ \log \boldsymbol{\pi}_2 \\ \vdots \\ \log \boldsymbol{\pi}_G \end{pmatrix},$$

where the elements θ_{ij} of $\boldsymbol{\theta}$ are primarily ordered by between group comparison $j = 1, ..., J$ and then, for each j,, by odds ratio $i = 1, ..., I$ (inner order).

2.4 Simultaneous Wald-type confidence intervals

The statistical model outlined above is a special case of a multivariate generalized linear model (Agresti, 2013; McCullagh and Nelder, 1989), for which the baseline logits ψ_{gc}, $c = 2, ..., C$ are the natural parameter (Agresti, 2013). One can thus use the methods of Hothorn et al. (2008) to construct simultaneous confidence intervals for $\boldsymbol{\theta}$ based on the estimated baseline logits $\hat{\boldsymbol{\psi}}$ and the corresponding estimated variance covariance matrix $\hat{\boldsymbol{\Sigma}}$. In more general settings, such estimates could be obtained by fitting baseline logit models. Then, also secondary factors or covariates might be included in such a model. In the simple case considered here, these estimates can be obtained from the contingency table $\mathbf{X}_{(G\times C)}$, using the asymptotic variance of baseline logits under multinomial sampling (derived using the Delta Method in Agresti, 2013, p.590-591): Denote the vector of sample estimators of the log-proportions in group g by $\log(\hat{\boldsymbol{\pi}}_g) = (\log(x_{g1}/n_g), ..., \log(x_{gC}/n_g))$. The corresponding estimators of the I linear combinations of interest are $\hat{\boldsymbol{\delta}}_g = \mathbf{A}\log(\hat{\boldsymbol{\pi}}_g)$, which have the asymptotic covariance matrix (Agresti, 2013, p.591)

$$\boldsymbol{\Sigma}_g = n_g^{-1}\left(\mathbf{A}Diag(\boldsymbol{\pi}_g)^{-1}\mathbf{A}^T - \mathbf{A}\mathbf{1}\mathbf{1}^T\mathbf{A}^T\right),$$

where $Diag(\boldsymbol{\pi}_g)^{-1}$ is the inverse of a diagonal matrix containing the true proportions $\boldsymbol{\pi}_g$, and $\mathbf{1}$ is an $(C \times 1)$ vector with all elements = 1.

Since we assumed independence between the treatment groups $g = 1, ..., G$, we can assemble the logits of interest for all treatment groups $g = 1, ..., G$ by stacking the column vectors $\boldsymbol{\delta}_g$, such that the corresponding covariance matrix can be written as a block-diagonal matrix:

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_G \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_G \end{pmatrix}.$$

The between-group comparisons (outer order) for all logits of interest (inner order) can then be written as

$$\boldsymbol{\theta} = (\mathbf{B} \otimes \mathbf{I}_{I\times I})\,\boldsymbol{\delta},$$

where $\boldsymbol{I}_{I\times I}$ is the identity matrix with I rows and columns. The corresponding covariance matrix is

$$\mathbf{V} = (\mathbf{B} \otimes \mathbf{I}_I)\,\boldsymbol{\Sigma}\,(\mathbf{B} \otimes \mathbf{I}_I)^T.$$

Estimators for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_g$, $\boldsymbol{\Sigma}$ and $\mathbf{V}$, may be obtained by plugging-in of the sample proportions $\hat{\boldsymbol{\pi}}_g$ instead of $\boldsymbol{\pi}_g$, and are denoted as $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Sigma}}_g$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{V}}$. Approximate simultaneous confidence intervals for the $M = IJ$ corresponding odds ratios are then

$$\exp\left[\hat{\theta}_m \pm z_{two-sided,1-\alpha,M,\hat{\mathbf{R}}}\sqrt{\hat{v}_m}\right], m = 1, ..., M,$$

where $\hat{\theta}_m$ is the mth element of $\hat{\boldsymbol{\theta}}$, $\hat{v}_m$ is the mth element of diagonal of $\hat{\boldsymbol{\Sigma}}$, $z_{two-sided,1-\alpha,M,\hat{\mathbf{R}}}$ is the two-sided $1-\alpha$-quantile of the M-variate normal distribution (Genz and Bretz, 2009) with correlation matrix $\hat{\mathbf{R}}$, and $\hat{\mathbf{R}}$ is obtained by standardizing $\hat{\boldsymbol{\Sigma}}$ by its diagonal elements (Hothorn et al., 2008).

Clearly, this approach has a number of problems: The plug-in of $\hat{\boldsymbol{\pi}}_g$ to obtain $\hat{\boldsymbol{\Sigma}}$, and the plug-in of $\hat{\boldsymbol{R}}$ to obtain the multivariate normal quantile $z_{two-sided,1-\alpha,M,\hat{\mathbf{R}}}$ are only justified for large samples (Hothorn et al., 2008). Moreover, $\boldsymbol{\Sigma}$ is only the asymptotic variance. The confidence intervals are symmetric w.r.t. $\hat{\theta}_m$, but the sampling distribution of $\hat{\theta}_m$ may be skewed if some expected cell counts, $n_g\pi_{gc}$, are small and π_{gc} differ, that is, if some sample sizes are moderate and/or the proportions are close to the border of the parameter space. Finally, the plug-in of π_{gc} with extreme observations as $x_{gc} = 0$ yields unreasonable estimated variances (∞) for the parameters on the log-scale; this leads to the failure of computing $z_{two-sided,1-\alpha,M,\hat{\mathbf{R}}}$, based on $\hat{\mathbf{R}}$, and even when using some ad-hoc adjustment for computing $\hat{\mathbf{R}}$, the intervals involving the corresponding π_{gc} will be uninformative due to spanning the complete parameter space. In parameter settings, where such events occur frequently, we can expect that the Wald-type simultaneous confidence intervals are unnecessarily conservative, that is, cover the true parameters too often.

In order to deal with the last problem, we apply the following ad-hoc adjustments: To compute the correlation matrix and contrasts of interest when the contingency table contains zeros, these are replaced by 0.5 (e.g. Plackett, 1962; Goodman, 1964). This approach is referred to as $\mathbf{W}$. Alternatively, one may use $\tilde{x}_{gc} = x_{gc} + 0.5$, $\tilde{n}_g = \sum_{c=1}^{C}\tilde{x}_{gc}$ and $\tilde{\boldsymbol{\pi}}_g = (\tilde{x}_{g1}/\tilde{n}_g, \tilde{x}_{g2}/\tilde{n}_g, ..., \tilde{x}_{gC}/\tilde{n}_g)$ instead of $\hat{\boldsymbol{\pi}}_g$

in all computations above. That is, 0.5 is added to each cell of the $G \times C$ contingency table, and all subsequent computations are performed based on this altered contingency table. This adjusted method is referred to as **W0.5**.

2.5 Sampling from the posterior distribution with a weakly informative prior

Under the assumption of G independent multinomial samples, one can make use of the fact that the Dirichlet distribution is a conjugate prior for the assumption of multinomial data. In the Bayesian model

$$\begin{aligned}(\pi_{g1}, \pi_{g2}, ..., \pi_{gC}) &\sim Dirichlet\left((\alpha_{g1}, \alpha_{g2}, ..., \alpha_{gC})\right), \\ (x_{g1}, x_{g2}, ..., x_{gC}) &\sim multinomial\left(n_g, (\pi_{g1}, \pi_{g2}, ..., \pi_{gC})\right),\end{aligned}$$

we can easily draw samples from the joint posterior distribution,

$$P\left((\pi_{g1}, ..., \pi_{gC}) \,|\, (x_{g1}, ..., x_{gC})\right) \sim Dirichlet\left((x_{g1} + \alpha_{g1}, ..., x_{gC} + \alpha_{gC})\right).$$

To construct simultaneous intervals for $\boldsymbol{\theta}$, many (say K) samples are drawn from this posterior, independently for each group g: Denote with $\boldsymbol{p}_k$ the stacked vectors of all groups $g = 1, .., G$ in the kth sample, that is, $\boldsymbol{p}_k = (p_{11}, ..., p_{1C}, p_{21}, ..., p_{2C}, p_{G1}, ..., p_{GC})^T$. For each sample $k = 1, ..., K$, the corresponding sample for the $M = IJ$ parameters of interest can be computed by:

$$\boldsymbol{t}_k = (\mathbf{B} \otimes \mathbf{A}) \log \boldsymbol{p}_k.$$

Rectangular sets containing the central 95% of the K sampled vectors $\boldsymbol{t}_k, k = 1, ..., K$ are described by Besag et al. (1995). For a $(K \times M)$ matrix $\boldsymbol{T}$, containing the K samples of the parameter vector of interest, $\boldsymbol{t}_k$, the main steps of this procedure are recalled here in close relation to the descriptions in Schaarschmidt and Djira (in press) or Schaarschmidt (2003):

1. Rank each column, $m = 1, ..., M$ of $\boldsymbol{T}$ separately and record the resulting ranks r_{km} and order statistics $t_m^{(k)}$.

2. For each row, $k = 1, ..., K$, of the resulting matrix $(K \times M)$ matrix of ranks with elements r_{km}, compute $max_k = \max(\max_{m=1,..,M}(r_{km}), K + 1 - \min_{m=1,...,M}(r_{km}))$.

3. Order max_k, resulting in the order statistics $max^{[k]}$ and find $k^* = max^{[q_{0.95}]}$, where $q_{0.95}$ is the nearest integer to $K * 0.95$.

The lower and upper interval bounds for each parameter of interest, $m = 1, ..., M$, are then obtained from the order statistics and back-transformation to the scale of odds-ratios: $\exp\left[t_m^{(K+1-k^*)}, t_m^{(k^*)}\right]$. Corresponding one-sided 95% simultaneous percentile intervals (Mandel and Betensky, 2008) can be calculated to obtain upper and/or lower limits for each element of $\boldsymbol{\theta}$. When the prior is chosen such that it has nearly no impact on the posterior, one can expect that the resulting intervals have good frequentist properties, e.g. simultaneous coverage probability close to 95%. Choosing the prior parameters $\alpha_{gc} = 1$ for all g, c results in an uniform prior distribution, while $\alpha_{gc} = 0.5$ for all g, c is known as Jeffreys prior. In the following, such intervals will be called **DP0.5** and **DP1**, if used with the Jeffreys prior or the uniform prior, respectively.

2.6 Simulation study

In order to compare the frequentist coverage probabilities between the different methods, a Monte Carlo Simulation has been performed for the following parameter settings: for $C = 3$ or $c = 5$ categories and $G = 4$ treatment groups, balanced sample sizes per treatment group of $n_g = 10, 20, 50, 100, 1000$ are considered. Three different sets of odds ratios have been considered: Baseline logits are compared between treatments ($g = 2, 3, 4$) and the control group ($g = 1$), as well as all pairwise comparisons between treatment groups for baseline logits, and all pairwise logits compared between treatments and control group. The true proportions of the categories are varied from the case that all categories appear equally often (1/3, 1/3, 1/3) to settings where the earlier categories (serving as baseline) are dominating (up to $\pi_{g1} = 0.9$) and the remaining categories are rare (down to $\pi_{g3} = 0.01$), and conversely, the earlier categories being rare ($\pi_{g1} = 0.01$) and the remaining categories abundant ($\pi_{g3} = 0.9$). For $C = 3$, 59 different configurations of $\boldsymbol{\pi}_g$ have been simulated. In 21 of these, all logits are equal between all treatment groups, in the remaining 38 settings some logits differ between

some of the treatment groups. For $C = 5$, 35 different parameter settings for $\boldsymbol{\pi}_g$ were considered (13 implying equal logits between treatment groups and 22 implying differences); with five categories, only samples sizes $n_g = 50, 100, 1000$ per group have been considered. A complete list of parameter settings is available as supplementary material. For each resulting parameter setting, 1000 data sets have been drawn from the multinomial distribution. In the methods based on sampling from the Dirichlet distribution, $K = 10,000$ values were drawn from the posterior to compute the simultaneous intervals for each data set.

2.7 Software

An implementation of the Wald-type intervals is available in the R-package `MCPAN 1.1-20` (Schaarschmidt et al., 2015) relying on multivariate normal quantiles obtained from the R-package `mvtnorm` (Genz et al., 2015). The methods based on sampling from the Dirichlet-posterior make use of the R-package `MCMCpack` (Martin et al., 2011) for Dirichlet random numbers, and the percentile intervals (Besag et al., 1995; Mandel and Betensky, 2008) implemented in package `MCPAN` .

3 Results

3.1 Simultaneous coverage probabilities

Figure 1 shows the simulated simultaneous coverage probabilities (SCP) of nominal 95% simultaneous confidence intervals. For all methods, there is a clear dependency of SCP on the minimal expected cell count ($\min(n_{gc}\pi_{gc})$): Intervals cover the parameters too often if the minimal expected cell count is small, e.g. below 5 or 2, while SCPs are close to the nominal level when the minimal expected cell count is equal or larger than 50. The intervals based on sampling from the Dirichlet posterior with uniform priors (DP1) show SCPs close to or above the nominal levels, whereas using Jeffreys prior may result in SCPs below the nominal level. The DP1 interval shows improved SCP compared to the Wald-type interval for intermediate values of the minimal expected cell count: While the Wald-type intervals start to be too conservative for minimal expected cell counts in the range of 10 or 20, the DP1 method shows SCPs close to the nominal level for minimal expected cell counts of 5 or 10. With the exception of a few parameter settings, the ad-hoc approach of adding 0.5 to each cell and using the Wald-type intervals afterwards does not show tangible differences of the SCP. Figure 2 illustrates the improvement of SCP when using the DP1 method instaed of the Wald-type interval: With samples sizes such as 100, 50, or 20 per group, the DP1 is less conservative than the Wald-type interval for the majority of parameter settings but rarely shows observed SCP larger than that of the Wald-type interval.

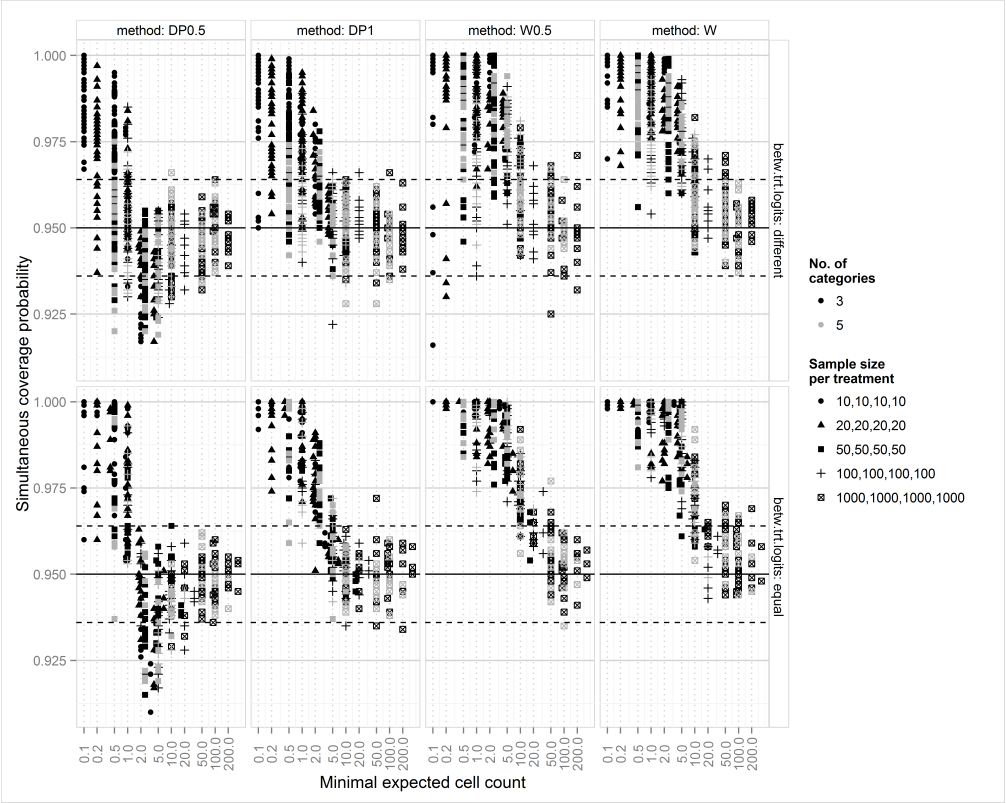


Figure 1: Simultaneous coverage probabilities of nominal 95% simultaneous confidence intervals, given the minimal expected cell count $\min(n_{gc}\pi_{gc})$. Symbols distinguish sample size per treatment group g, grayscale distinguishes parameter settings with $C = 3$ or $C = 5$ categories. Column panels show results for different confidence intervals methods, while row panels distinguish parameter settings where at least one logit differs between treatment groups (upper row) and all logits of interest are equal in all treatment groups (lower row). Dashed horizontal lines show the range in which 95% of all simulation results (based on 1000 simulations per setting) would fall if a methods had exactly 95% true simultaneous coverage probability.

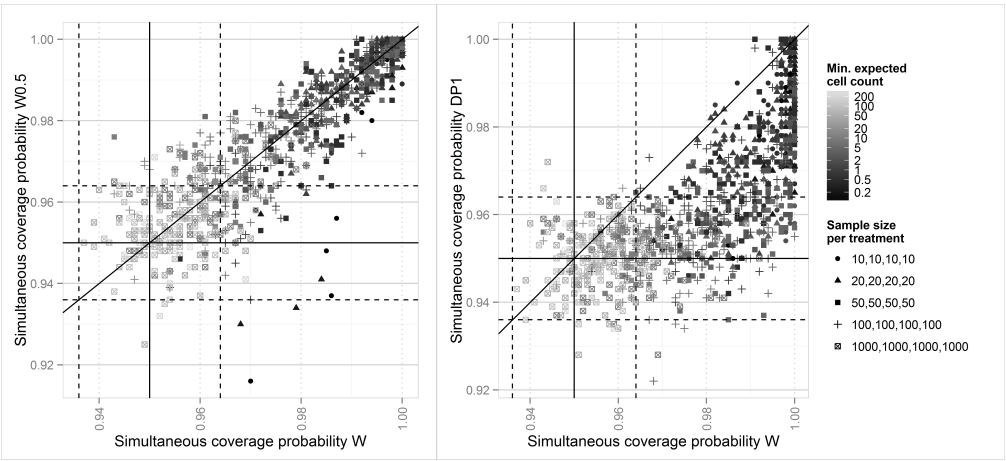


Figure 2: Simulated (1000 simulation runs per parameter setting) simultaneous coverage probabilities of the Wald-type interval (x-axis) plotted against that of the Wald-add-0.5 interval and the interval based on Dirichlet sampling with uniform prior. Gray scale is used to show each settings minimal expected cell count $\min(n_{gc}\pi_{gc})$; symbols distinguish sample size per treatment group g. Dashed horizontal and vertical lines show the range in which 95% of all simulation results (1000 simulations) would fall if a methods had exactly 95% true coverage probability.

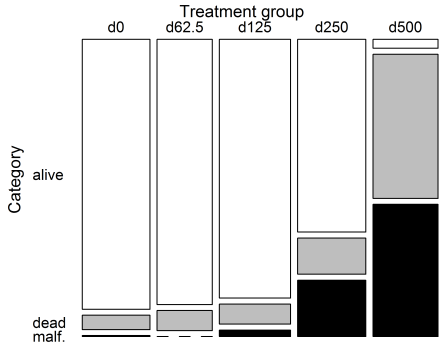


Figure 3: Mosaicplot of the (5×3) table of developmental toxicity data (Agresti, 1990)

4 Examples

4.1 Developmental toxicity

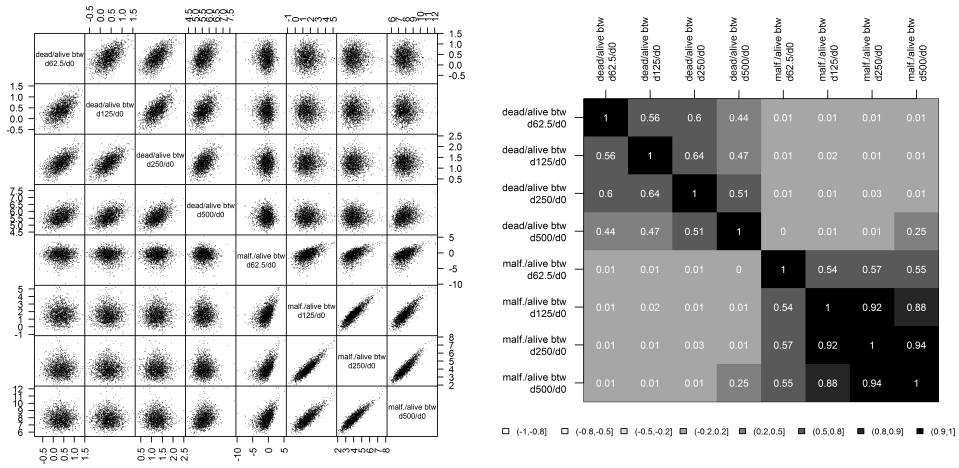
Agresti (1990, p.320, Tab.9.7 therein) shows results of a study on developmental toxicity in mice. After exposure to $G = 5$ treatments (control d0, and 4 different dosages, d62.5, d125, d150, d500) of a compound during pregnancy, the offspring of mice $(n_1 = 297, n_2 = 242, n_3 = 312, n_4 = 299, n_5 = 285)$ is classified into $C = 3$ categories: alive, dead, malformation. Figure 3 shows a mosaic plot derived from the (5×3) contingency table data. To investigate for which dose groups there is an increase of π_{dead}/π_{alive} or $\pi_{malformation}/\pi_{alive}$ over that of the control, one can consider simultaneous confidence intervals for baseline logits (baseline = alive) compared between the 4 dose groups and the control.

Figure 4a shows a scatter plot matrix of 2000 sampled values for the $(M = 8)$ corresponding logits based on a sample of the joint posterior with prior $(\pi_{alive}, \pi_{dead}, \pi_{malformation}) \sim Dirichlet((1, 1, 1))$ on each sample $g = 1, ..., 5$. It is obvious that those parameters referring to comparisons to the control group for the same odds are positively correlated (parameters $1, ...4$ and $5, ..., 8$, respectively) and that the magnitude of correlation further depends on the estimated proportions (higher positive correlations in malformed/alive than in dead/alive). Figure 4b shows the estimated correlation matrix $(\hat{\boldsymbol{R}})$ underlying the quantiles Wald-type-intervals (W) for this example. Table 1 shows the lower and upper limits of the corresponding 95% simutaneous intervals for the oddsratios: The odds dead/alive are significantly increased compared to control in d250 and d500 according to the DP1 method this ratio is increased by factor 1.5 - 8.3 in d250 and by factor 97 - 930 in d500. The odds malformation/alive also show a significant increase in dose groups d250 and d500, at least by factor 10 and 390 (DP1), respectively. The R code to reproduce these calculations (up to uncertainties due to sampling) is provided as supplementary material.

The corresponding (two-sided) 95% quantile of an 8-variate normal distribution is $z_{0.95,M=8,\hat{\boldsymbol{R}}} = 2.638$. Compared to the Scheffe-type quantile adjusting for all possible contrasts (Gold, 1963), $\sqrt{\chi^2_{df=8}} = 3.938$, the Wald-type intervals have considerably reduced width, whereas the reduction of width is relatively little compared to the Bonferroni adjustment of Goodman (1964): $z_{1-0.05/(8*2)} = 2.734$.

Table 1: Simultaneous 95% confidence intervals for comparisons to control for the baseline odds dead/alive and malformed/alive, rounded to the second significant digit.

Odds	Betw-group-ratio	Estimate	Lower(W)	Upper(W)	Lower(DP1)	Upper(DP1)
dead/alive	d62.5/d0	1.4	0.54	3.7	0.55	3.7
dead/alive	d125/d0	1.5	0.59	3.6	0.59	3.7
dead/alive	d250/d0	3.5	1.5	8.2	1.5	8.3
dead/alive	d500/d0	300	95	940	97	930
malf./alive	d62.5/d0	0.62	0.01	60	0.00	18
malf./alive	d125/d0	6.9	0.41	120	0.66	88
malf./alive	d250/d0	82	5.7	1200	10	890
malf./alive	d500/d0	4100	250	67000	390	45000



(a) Sample of 2000 values from the Dirichlet posterior with uniform prior (DP1).

(b) Estimated correlation matrix $\hat{\boldsymbol{R}}$ corresponding underlying the Wald-type interval (W).

Figure 4: Observed correlation between estimators of the eight parameters of interest in the developmental toxicity example.

4.2 Differential blood count (WBC) in rats

Table 2 (Hothorn et al., technical report) shows counts of white blood cells of 4 categories, LY, MO, NE, EO (lymphocytes, monocytes, neutrophil and eosinephil granulocytes); other cell types occurred only with one cell and are omitted. Counts have been obtained from rats (females and males) under four different treatments: an untreated control (C) and three dose groups (L, M, H, for low, mid and high dose). Note that the counts in Table 2 are obtained by pooling ten individuals per sex and treatment group (exception: eight animals for males in high dose).

Table 2: Differential count of white blood cells in rats of both sexes and four treatment groups.

sex	Group	LY	MO	NE	EO
female	C	1668	41	272	19
female	L	1633	47	305	15
female	M	1699	39	244	18
female	H	1643	37	299	21
male	C	1594	32	340	34
male	L	1593	25	356	26
male	M	1510	34	431	25
male	H	1196	33	351	19

One may now be interested, whether any of the relative proportions of the single categories change between control and dose groups in males or females. We express this as all ($I = 6$) pairwise odds between the $C = 4$ categories. These odds are then compared between the L, M and H dose and the control, separately for males and females, resulting in $J = 6$ between-group-comparisons. The corresponding matrices A and B are then:

$$\mathbf{A}_{(I\times C)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \mathbf{B}_{(J\times G)} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

The counts in Table 2 are relatively large, thus all considered methods can be expected to perform well and to yield very similar results. The quantiles between the Goodman approach ($z_{1-0.05/(36*2)} = 3.197$) and the Wald-type intervals with plug-in of estimated correlations ($z_{0.95,M=36,\hat{\boldsymbol{R}}} = 3.085$) again differ only slightly.

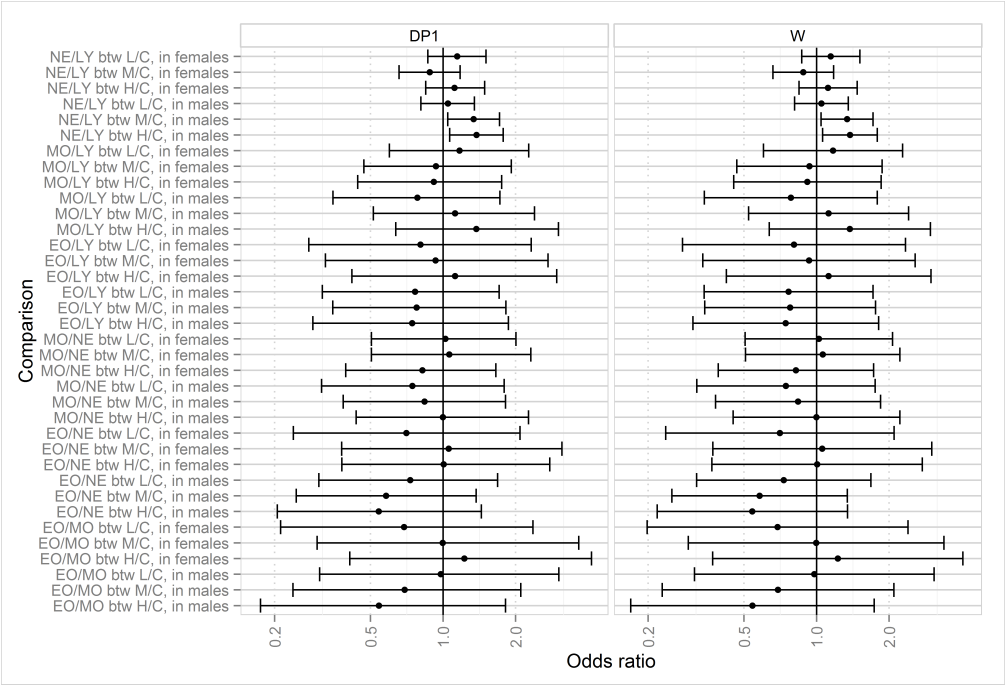


Figure 5: Simultaneous 95% confidence intervals for 36 odds ratios defined in the differential blood count example.

Figure 5 shows the 95% confidence intervals for the 36 odds ratios defined above for the methods DP1 and W. The intervals hardly differ between the methods. The only significant differences w.r.t. these 36 odds ratios are found for the mid and high dose groups (H, M) in males where the proportion of neutrophile granulocytes relative to lymphocytes (π_{NE}/π_{LY}) is significantly increased in treatment groups M and H compared to the control group, C. Table 3 shows estimates and confidence limits of the W and DP1 method for those odds (π_{NE}/π_{LY}): in males, the ratio (π_{NE}/π_{LY}) is increaaed by factor [1.041, 1.720] in group M, and by factor [1.065, 1.792] in group H, relative to that of the control group. The R code to reproduce these calculations is provided as supplementary material.

Table 3: Simultaneous 95% confidence intervals for comparisons to control for the baseline odds dead/alive and malformed/alive (subset out of a total of 36 comparisons), rounded to the 3rd digit.

Oddsratio	Estimate	Lower(W)	Upper(W)	Lower(DP1)	Upper(DP1)
NE/LY btw L/C, in females	1.145	0.868	1.511	0.868	1.511
NE/LY btw M/C, in females	0.881	0.659	1.178	0.658	1.183
NE/LY btw H/C, in females	1.116	0.845	1.474	0.844	1.472
NE/LY btw L/C, in males	1.048	0.810	1.355	0.810	1.357
NE/LY btw M/C, in males	1.338	1.044	1.716	1.041	1.720
NE/LY btw H/C, in males	1.376	1.059	1.787	1.065	1.792

5 Discussion

We described methods for the computation of simultaneous confidence intervals for user defined sets of pairwise between-treatment comparisons w.r.t. user defined sets of odds ratios based on the assumption of several independent multinomial samples. The asymptotic method accounts for the correlation between estimators by the plug-in of an estimated covariance matrix. A small sample approach, based on sampling from a Dirichlet posterior with vague priors, is considered as an alternative.

In a simulation study, the coverage probability of these methods is assessed for different sets of multinomial proportions, different sample sizes per treatment group, three (or five) multinomial categories, as well as different types of comparisons between groups and categories. The method based

on sampling from the Dirichlet posterior with a vague prior assigning parameter $\alpha = 1$ to all categories performs best in the considered settings: When the minimal expected cell count of the contingency table is moderate (e.g. at least five) the simultaneous coverage probability is close to the nominal level. If rare proportions or small sample sizes lead to smaller expected cell counts, the method is conservative. The asymptotic method is more conservative as it reaches coverage probabilities close to the nominal level for expected cell counts of 20 or above, and covers the true parameter too often otherwise. Note that these recommendations may not hold when comparing multinomial samples with much more categories than considered here, e.g. 10 or 20.

All methods considered are conservative for small sample size and/or rare events. That is, with either method it will be hard to detect relatively small changes in the proportions of rare categories, or when sample sizes are small. The method based on sampling from the Dirichlet posterior can easily be extended to include informative priors. For example, when historical control data are available for bioassays, the Dirichlet prior for untreated control groups may be chosen to reflect the expected values and the plausible range for the proportions of the categories under control conditions. Moreover, it would be computationally simple, to extend the methods based on Dirichlet posteriors to simultaneous confidence intervals for differences or ratios between multinomial proportions.

Clearly the statistical model used here (i.e. all methods considered) has a number of problems. In the statistical model, many parameters are fitted to the data. Such approaches may over-fit the data in situations where simpler models would be appropriate. For example, when in dose-response analysis linear or log-linear relations to baseline logits are plausible, regression models for baseline logits are a sparse alternative to estimating extra parameters for each dose group (see, e.g. Agresti, 2013). When there are several ordinal categories, cumulative logit models or related approaches can be more appropriate (see, e.g. Ryu, 2009; Agresti, 2013). Furthermore, the methods described here as well as the simulation settings apply only to highly controlled lab experiments or randomized trials with no further substructures. However, the Wald-type intervals can likewise be applied when baseline logits and corresponding covariance matrix are estimated from generalized linear model fits. Then similar inferential procedures can be performed while including covariates, secondary factors of interest or stratification. Moreover, experiments or studies will often involve replicated biological units per treatment group, for example, several animals, litters, or cultures per treatment group in toxicology or clustered observations in clinical trials or exposure studies. If variation between these units is larger than expected under multinomial distribution (over-dispersion): all methods considered here will have (severely) too narrow confidence intervals, that is too low coverage probability, because they do not account for such over-dispersion.

References

Agresti, A. (1990). Categorical Data Analysis, New York: John Wiley & Sons, New York.

Agresti, A. (2013). Categorical Data Analysis (3rd ed.). John Wiley & Sons, Inc., Hoboken, New Jersey.

Besag, J., Green, P., Higdon, D., Mengersen, K. (1995). Bayesian computation and Stochastic Systems. Statistical Science 10, 3-41.

Bretz, F., Genz, A., Hothorn, L. (2001). On the numerical availability of multiple comparison procedures. Biometrical Journal 43, 645-656.

Casella, G., and Berger, R.L. (2002). Statistical Inference (2nd Ed.). Duxbury, Pacific Grove, CA, USA.

Chafa, D. and Concordet, D. (2009). Confidence Regions for the Multinomial Parameter With Small Sample Size. Journal of the American Statistical Association 104, 1071-1079, DOI: 10.1198/jasa.2009.tm08152

Genz, A. and Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195., Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. (2015). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-3. URL http://CRAN.R-project.org/package=mvtnorm

Glaz, J. and Sison, C.P. (1999). Simultaneous confidence intervals for multinomial proportions. Journal of Statistical Planning and Inference 82, 251-262.

Gold, R.Z. (1963). Test Auxiliary to χ^2 in a markov chain. The Annals of Mathematical Statistics 34, 5674.

Goodman, L.A. (1964). Simultaneous Confidence Limits for Cross-Product Ratios in Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological), 26, 86-102.

Hothorn, T., Bretz, F., Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal 50, 346-363.

Hothorn, L.A., Gerhard,D., Pras-Raves, M. (2009): Statistical evaluation of the differential blood count in toxicological studies. Technical report, Institute of Biostatistics, Hannover.

Hou, C.-D., Chiang, J., Tai, J.J. (2003). A family of simultaneous confidence intervals for multinomial proportions. Computational Statistics & Data Analysis 43, 29-45.

Mandel, M., Betensky, R.A. (2008). Simultaneous confidence intervals based on the percentile bootstrap approach. Computational Statistics & Data Analysis 52, 2158-2165.

Martin, A.D., Quinn, K.M., Park, J.H.(2011). MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 42(9): 1-21. URL http://www.jstatsoft.org/v42/i09/.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Second Edition. Chapman & Hall/CRC.

Piegorsch, W.W. and Richwine, K.A. (2001) Large-Sample Pairwise Comparisons among Multinomial Proportions with an Application to Analysis of Mutant Spectra. Journal of Agricultural, Biological, and Environmental Statistics 6 (3), 305-325.

Plackett, R.L. (1962). A note on interactions in contingency tables, Journal of the Royal Statistical Society B 24, 162-166.

Ryu, E. (2009). Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. Statistics in Medicine 28, 3179-3188.

Schaarschmidt, F., Gerhard, D., Sill, M. (2015). MCPAN: Multiple Comparisons Using Normal Approximation. R package version 1.1-20. http://CRAN.R-project.org/package=MCPAN

Schaarschmidt, F. (2013). Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. Computational Statistics and Data Analysis 58, 265-275.

Schaarschmidt, F., Djira, G.D. Simultaneous confidence intervals for ratios of fixed effect parameters in linear mixed models. Accepted for publication in Communications in Statistics - Simulation and Computation. DOI: 10.1080/03610918.2013.849741

Wang, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. Journal of Multivariate Analysis 99, 896-911.

Danksagung

Ich danke, in randomisierter Reihenfolge: Mario Hasler, Ruth Schaarschmidt, Ralph Scherer, Cornelia Frömke, Daniel Gerhard, meinen Eltern, Andreas Kitsche, Gemechis Dilba Djira, Clemens Buczilowski, Martin Sill, Ludwig A. Hothorn, Philip Pallmann, Charlotte Vogel, Hanne Visser, Kornelius Rohmeyer, Lea Vaas.

Appendix

PERSÖNLICHER UND BERUFLICHER WERDEGANG

FRANK SCHAARSCHMIDT

PRIVATANSCHRIFT:
DIANASTR. 1,
31275 LEHRTE

DIENSTANSCHRIFT:
INSTITUT FÜR BIOSTATISTIK,
NATURWISSENSCHAFTLICHE FAKULTÄT,
HERRENHÄUSERSTR. 2,
30419 HANNOVER
TEL: 0511-762-5821
EMAIL: schaarschmidt@biostat.uni-hannover.de

Persönliche Daten

	geboren am 27. Februar 1979 in Jena verheiratet, drei Kinder

Schulbildung/ Wehrdienst

09/1985 - 06/1993	Grundschule und Gymnasium in Jena
09/1993 - 06/1998	Gymnasium in Diepholz, Abitur in Diepholz
07/1998 - 08/1999	Zivildienst Kreiskrankenhaus Diepholz

Studium

10/1999 - 04/2005	Gartenbau an der Universität Hannover, Abschluss Diplom-Agraringenieur mit Gesamtnote 1,1. Vertiefung der Fächer Biometrie, Angewandte Genetik und Pflanzenzüchtung, Gemüsebau, Phytomedizin, Marktlehre und Agrarpolitik, Systemtheorie im Pflanzenbau, Wissenschaftliche Hilfskraft in den Bereichen Angewandte Pflanzenzüchtung und Biometrie, Diplomarbeit im Fach Biometrie, Thema: Binomial group testing – design and analysis, in Zusammenarbeit mit der KWS Saat AG, Einbeck (Abteilung Zuckerrübenzüchtung). Ausgezeichnet mit dem Bernd Streitberg Preis der Deutschen Region der Internationalen Biometrischen Gesellschaft

Studienunterbrechung: 03/2002 - 03/2003 Pflichtpraktikum im Rahmen des Studiums bei Firma Juliwa-Enza GmbH & Co KG, Heidelberg 03/2002-09/2002, sowie Firma Enza Zaden bv, Enkhuizen, Niederlande (10/2002-03/2003): Tätigkeiten in den Bereichen Feldversuchswesen, Pflanzenzüchtung sowie angewandte Genetik im Feldgemüsebau

Promotion

05/2005 – 01/2009

Institut für Biostatistik an der Gottfried-Wilhelm-Leibniz-Universität Hannover, Gesamtnote ausgezeichnet.

Thema: Marginal and simultaneous confidence intervals for abundance data with application to safety assessment of non-target species.
Referent: Prof. Dr. L. A. Hothorn, Universität Hannover
Korreferent: Prof. Dr. H.-P.Piepho, Universität Hohenheim

Forschungsschwerpunkte während der Promotion: Simultane Konfidenzintervalle für multiple Kontraste binomialer Proportionen und Zähldaten, Frequentistische Eigenschaften simultaner Bayesianischer Intervalle auf Basis von MCMC sampling, Äquivalenznachweis für Zähldaten aus ökologischen Feldversuchen, approximative Güteberechnung für multiple Kontrasttests.

Weitere Tätigkeiten am Institut für Biostatistik:

Erstellen von R-Paketen, statistische Beratung von Bachelor-, Master- und Promotionsstudenten an der Fakultät Naturwissenschaften (ehem. Fachbereich Gartenbau), Betreuung von Bachelor- und Masterarbeiten am Institut für Biostatistik, Mitarbeit an Drittmittelprojekten (EU, BmBF), Lehrtätigkeit für Studenten der Fachrichtungen Gartenbauwissenschaften und Pflanzenbiotechnologie.

Auslandsaufenthalt:09/2005, Biometrics Dept., Firma N.V. Organon, Oss, Niederlande, Arbeit an Vergleichen zur Kontrolle für binomiale Proportionen.

Post-Doc

ab 02/2009

Institut für Biostatistik an der Gottfried-Wilhelm-Leibniz-Universität Hannover
Forschungsschwerpunkte: Frequentistische Eigenschaften simultaner Bayesianischer Intervalle auf Basis von MCMC sampling in generalisierten und gemischten linearen Modellen, simultane Konfidenzintervalle zum Vergleich von Diversitätsindizes und von Erwartungswerten lognormalverteilter Daten, simultane Konfidenzbänder in generalisierten linearen Modellen, Analyse von Zähldaten in Metagenomanalysen,

Konfidenzintervalle für Quotienten von Proportionen für binomiale Daten mit Überdispersion

Weitere Tätigkeiten: Lehrtätigkeit für Studenten der Fachrichtungen Gartenbauwissenschaften und Pflanzenbiotechnologie, statistische Beratung an der naturwissenschaftlichen Fakultät, Betreuung von B.Sc und M.Sc Arbeiten am Institut für Biostatistik, Erstellen von R-Paketen für das Institut für Biostatistik, Mitarbeit an Drittmittelprojekten (EU, DFG), Leitung der lokalen Organisation für das 55. Biometrische Kolloquium 2009, Hannover

Elternzeit: 02/2011 – 03/2011 und 11/2011 – 04/2012

Hannover, den 11. Dezember 2015

Publikationsliste Frank Schaarschmidt, Stand 12.12.2015

1. Zeitschriften (peer-reviewed)

Bredemeier, B., von Haaren, C., Rüter, S., **Schaarschmidt, F.,** Reich, M.(2015): Spatial congruence between organic farming and biodiversity related landscape features in Germany. *International Journal of Biodiversity Science, Ecosystem Services & Management* 11(4) 330-340. http://dx.doi.org/10.1080/21513732.2015.1094515

Baeßler, B., **Schaarschmidt, F.,** Schnackenburg, B., Stehning, C., Giolda, A., Maintz, D, Bunck, A.C. Cardiac T2-mapping using a fast gradient echo spin echo sequence - first in vitro and in vivo experiences. Accepted for publication in *Journal of Cardiovascular Magnetic Resonance*. 17:67, DOI 10.1186/s12968-015-0177-2

Konietschke, F., Placzek, M., **Schaarschmidt, F.,** Hothorn, L. A. (2015). nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software* 46, 9. http://www.jstatsoft.org/

Schaarschmidt, F., Hofmann, M., Jaki, T., Grün, B., Hothorn, L.A. (2015). Statistical approaches for the determination of cut points in anti-drug antibody bioassays. *Journal of Immunological Methods* 418, 84–100. DOI: 10.1016/j.jim.2015.02.004

Radespiel, U., Schaber, K., Kessler, S.E., **Schaarschmidt, F.,** Strube, C. (2015): Variations in the excretion patterns of helminth eggs in two sympatric mouse lemur species (Microcebus murinus and M. ravelobensis) in northwestern Madagascar. *Parasitology Research* 114 (3): 941-954. DOI: 10.1007/s00436-014-4259-0

Kitsche, A., **Schaarschmidt, F**. (2015): Analysis of statistical interactions in factorial experiments. *Journal of Agronomy and Crop Science* 201 (1): 69-79, DOI: 10.1111/jac.12076

Matthies, S., Rüter, S., Prasse, R., **Schaarschmidt, F.** Factors driving the vascular plant species richness in urban green spaces: Using a multivariable approach. *Landscape and Urban Planning* (2015) 177-187. DOI information: 10.1016/j.landurbplan.2014.10.014

Schaarschmidt, F., and Hothorn, L.A. (2014). Statistical Methods and Software for Validation Studies on New In Vitro Toxicity Assays. *ATLA* 42, 319–326.

Priesnitz, K.-U., Benker, U., **Schaarschmidt, F.** (2013). Assessment of the potential impact of a Bt maize hybrid expressing Cry3Bb1 on ground beetles (Carabidae). *Journal of Plant Diseases and Protection* 120 (3), 131–140.

Scherer, R., **Schaarschmidt, F.**, Prescher, S., Priesnitz, K.U. (2013) Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data. *Biometrical Journal* 55 (2), 246–263. DOI: 10.1002/bimj.201200157

Schaarschmidt, F. (2013). Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. *Computational Statistics and Data Analysis* 58, 265–275. doi:10.1016/j.csda.2012.08.011

Schlingmann, B., Schadzek, P., Hemmerling, F., **Schaarschmidt, F.,** Heisterkamp, A., Ngezahayo, A. (2013). The role of the C-terminus in functional expression and internalization of rat connexin46 (rCx46). *Journal of Bioenergetics and Biomembranes*, 45, 59–70. DOI 10.1007/s10863-012-9480-x

Pallmann, P., **Schaarschmidt, F.,** Hothorn, L., Fischer, C., Nacke, H., Priesnitz, K., Schork, N. (2012). Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices. *Molecular Ecology Resources* 12(6): 1068-1078. doi: 10.1111/1755-0998.12004

Kitsche, A., Hothorn, L.A., **Schaarschmidt, F.** (2012). The use of historical controls in estimation of simultaneous confidence intervals for comparisons against a concurrent control. *Computational Statistics and Data Analysis* 56, 3865–3875.

Bunck A.C., Kröger J.-R., Juettner A., Brentrup A., Fiedler B., **Schaarschmidt F.,** Crelier G., Schwindt W., Heindel W., Niederstadt T., Maintz D. (2011). Magnetic resonance 4D flow characteristics of cerebrospinal fluid at the craniocervical junction and the cervical spinal canal. *European Radiology* 21, 1788–1796.

Rauschen, S., Schultheis, E., Hunfeld, H., **Schaarschmidt, F.,** Schuphan, I. and Eber, S. (2010). Diabrotica-resistant Bt-maize DKc5143 event MON88017 has no impact on the field densities of the leafhopper *Zyginidia scutellaris*. *Environmental Biosafety Research* 9, 87–99.

Bilder, C.R., Zhang, B., **Schaarschmidt, F.,** Tebbs, J.M. (2010). binGroup: A Package for Group Testing. *The R Journal* 2/2, 56-60.

Rauschen, S., **Schaarschmidt, F.**, Gathmann, A. (2010). Occurrence and field densities of Coleoptera in the maize herb layer: implications for Environmental Risk Assessment of genetically modified Bt-maize. *Transgenic Research* 19:727-744. DOI: 10.1007/s11248-009-9351-3

Djira, G.D., **Schaarschmidt, F.,** Fayissa, B. (2010). Inferences for Selected Location Quotients with Applications to Health Outcomes. *Geographical Analysis* **42**:288-300.

Perry, J.N., ter Braak, C.J.F., Dixon, P.M., Duan, J.J., Hails, R.S., Huesken, A., Lavielle, M., Marvier, M., Scardi, M., Schmidt, K., Tothmeresz, B., **Schaarschmidt, F.,** van der Voet, H. (2009): Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. *Environmental Biosafety Research* **8**:65-78.

Schaarschmidt, F. and Vaas, L. (2009). Analysis of trials with complex treatment structure using multiple contrast tests. *HortScience* **44**(1):188-195.

Schaarschmidt, F., Biesheuvel, E., Hothorn, L.A. (2009). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials, *Journal of Biopharmaceutical Statistics* **19**(2):292-310.

Schaarschmidt, F., Sill, M., and Hothorn, L.A. (2008). Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions. *Biometrical Journal* **50**(5):782-792.

Schaarschmidt, F., Sill, M., and Hothorn, L.A. (2008). Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test. *Journal of Biopharmaceutical Statistics* **18**(5):934-948.

Rauschen, S., Eckert, J., **Schaarschmidt, F.,** Schuphan, I., Gathmann, A. (2008). An evaluation of methods for assessing the impacts of *Bt*-maize MON810 cultivation and pyrethroid insecticide use on Auchenorrhyncha (Planthoppers and Leafhoppers). *Agricultural and Forest Entomology* **10**:331-339.

Dilba, G., **Schaarschmidt, F.** and Hothorn, L.A. (2007). Inferences for Ratios of Normal Means. *R News* **7**(1):20-23 http://CRAN.R-project.org/doc/Rnews/Rnews_2007-1.pdf.

Schaarschmidt, F. (2007). Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Communications in Biometry and Crop Science* **2**(1):32-40.

2. Zeitschriften (peer-reviewed, accepted for publication)

Schaarschmidt, F.: Confidence Intervals for the Risk Ratio when Analyzing Bioassays in the Presence of Overdispersion. *Biometrics and Biostatistics International Journal.* http://medcraveonline.com/BBIJ

Baeßler, B., **Schaarschmidt,** F., Stehning, C., Schnackenburg,B., Maintz, D. Bunck, A.C. A Systematic Evaluation of Three Different Cardiac T2-mapping Sequences at 1.5 and 3T in Healthy Volunteers. Accepted for publication in *European Journal of Radiology.* In press.

Schaarschmidt, F., Djira, G.D. Simultaneous confidence intervals for ratios of fixed effect parameters in linear mixed models. Accepted for publication in *Communications in Statistics – Simulation and Computation*.

Pallmann, P. and **Schaarschmidt, F.** Common pitfalls when testing additivity of treatment mixtures with χ^2. Accepted for publication in *Journal of Applied Entomology*.

Wimalasekera, R., **Schaarschmidt, F.,** Angelini, R., Cona, A., Tavladoraki, P., Scherer, G.F.E.: POLYAMINE OXIDASE2 of Arabidopsis Contributes to ABA Mediated Plant Developmental Processes. Accepted for publication in *Plant Physiology and Biochemistry.*

Baeßler, B., **Schaarschmidt, F.,** Dick, A., Stehning, C., Schnackenburg, B., Michels, G., Maintz, D., Bunck, A.C.: Mapping tissue inhomogeneity in acute myocarditis: a novel analytical approach to quantitative myocardial edema imaging by T2-mapping. Accepted for publication in *Journal of Cardiovascular Magnetic Resonance.*

3. Beitrag Tagungsband

Gerhard, D. and **Schaarschmidt, F.** (2007). Proof of safety for non-target species: a confidence interval based approach. In: Piepho, H.-P. and Bleiholder, H. (Eds.). *Agricultural Field Trials – Today and Tomorrow.* Proceedings of the International Symposium 08-10 October 2007, Stuttgart-Hohenheim. Verlag Grauer, Beuren.