Multiple Vergleiche mit der SAS-Prozedur MIXED

Erich Schumacher
Institut für Angewandte Mathematik und Statistik
Universität Hohenheim
70593 Stuttgart
schumach@uni-hohenheim.de

Marc Weimer
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280
69120 Heidelberg
m.weimer@dkfz-heidelberg.de

Zusammenfassung

Mit Hilfe der SAS-Prozeduren GLM und MIXED können allgemeine lineare Modelle analysiert werden, wobei alle zufälligen Effekte des Modells als normalverteilt angenommen werden.
Mit Hilfe der Option Adjust = SIMULATE und geeigneten Unteroptionen des SAS-Statements LSMEANS können die Testverfahren Simulated, Tukey-Kramer, Dunnett, DunnettU, DunnettL, Bonferroni, Sidak, GT-2, Scheffe zur Durchführung multipler paarweiser Vergleiche hinsichtlich des nominellen multiplen Niveaus α sowohl in PROC GLM als auch in PROC MIXED miteinander verglichen werden. In der Regel erweisen sich – falls alle paarweisen Vergleiche durchgeführt werden sollen, die Tests Simulated und Tukey-Kramer als die trennschärfsten Verfahren, bei Vergleichen mit einer Kontrolle sind in der Regel die Dunnett-Tests die trennschärfsten. Die Konservativität des Tukey-Kramer Tests ist jedoch an die Einhaltung einer sog. „Hayter"- Bedingung geknüpft, siehe Hochberg, Tamhane [15] und Hayter [13,14], die im allgemeinen linearen Modell häufig nicht erfüllt ist und weder in GLM noch in MIXED verifiziert wird. Ist die Hayter- Bedingung erfüllt, spricht Hsu [17] von Modellen mit einer „Ein-Weg-Struktur" (one-way structure). In diesem Falle sind dann ein- und zweiseitige Dunnett-Tests zum Vergleich von Behandlungsgruppen mit einer Kontrolle exakte Tests, die das multiple Niveau α einhalten.
Eine Vorgehensweise ist, eine Liste von Designs anzugeben, bei denen die Hayter-Bedingung auf jeden Fall eingehalten wird. Dies stößt insbesondere bei der Analyse unbalancierter Daten, bei denen Beobachtungen „missing at random" fehlen, sehr rasch an Grenzen.
Die Autoren haben ein SAS-Macro „*macro_hayter.sas*" auf der Basis von PROC MIXED erarbeitet, mit deren Hilfe die Einhaltung oder Nichteinhaltung der Hayter-Bedingung direkt für die auszuwertenden Daten überprüft werden kann. Damit hat man ein Hilfsmittel zur Verfügung, bei der jeweilig anstehenden Analyse in vielen Fällen unter den von SAS angebotenen Tests eine trennscharfe, aber auch das multiple Niveau einhaltende Testprozedur auszuwählen.
Neben den multiplen paarweisen Vergleichen wird auch auf die Analyse allgemeiner linearer Kontraste eingegangen, hierbei wird vor allem auf SAS – Macros zurückgegriffen, die Westfall et al. [31, 32] entwickelt und Frömke und Bretz [9] weiterentwickelt haben.

Schlüsselworte: Multiple Vergleiche, Lineares Modell, PROC MIXED, Tukey-Kramer Test, Dunnett Tests, SAS-Macros für multiple Vergleiche.

1 Einleitung

1.1 Historische Entwicklung

Es existiert eine Fülle von Publikationen zum Thema multipler Vergleiche (MCP: Multiple Pairwise Comparisons). Da einige Tests in SAS den Namen von bekannten Statistikern tragen, erwähnen wir Tukey [28] , Dunnett [6, 7] , Scheffé [23, 24] und Kramer [19]. Bis zur Mitte der 80-er Jahre war ein weitverbreitetes Nachschlagewerk das Buch von R. G. Miller Jr. [21]. Im Jahre 1984 bewies Hayter [13] analytisch die schon lange vermutete Konservativität des Tukey-Kramer Tests im Falle der einfachen Varianzanalyse bei unbalancierten Daten. Hochberg und Tamhane [15] haben 1987 ihre Monografie „Multiple Comparison Procedures“ veröffentlicht, das die wesentlichen Entwicklungen multipler Vergleiche bis zu diesem Zeitpunkt zusammenfasste und das bis heute ein wichtiges Nachschlagewerk ist.
Durch die rasante Entwicklung verfügbarer Rechenkapazität wurde auch die Weiterentwicklung multipler Verfahren mittels Computer – Intensiver Methoden beschleunigt.
Diese Entwicklung fand in dem Buch von Westfall und Young [30] aus dem Jahre 1993 seinen Niederschlag, dort werden vor allem „Resampling“-Verfahren zur Durchführung multipler Vergleiche benutzt. In SAS wurde die in diesem Buch beschriebene Theorie gemeinsam durch Westfall und den SAS-Entwickler Wolfinger in der PROC MULTTEST umgesetzt. Es können sowohl Daten unter Normal- als auch unter Nichtnormal-Verteilungsannahmen analysiert werden, wobei die Analyse jedoch auf im wesentlichen auf einfaktorielle Designs beschränkt ist. Einen weiteren Fortschritt brachte 1996 die Monografie von Hsu [17], wobei zum einen eine konsequente Trennung in All-Paarvergleiche (MCA), multiple Vergleiche mit einer Kontrolle (MCC), multiple Vergleiche mit der „besten Gruppe“ (MCB) und multiple Vergleiche mit dem „Mittel“ (MCM) vorgenommen wurde. Zum anderen wird als Analysemethode die Berechnung von simultanen Vertrauensintervallen bevorzugt, da diese gegenüber multiplen Tests einen höheren Informationsgehalt aufweisen. Als multiples Testverfahren wird ein Simulationsverfahren empfohlen, das in der Regel noch trennschärfer als der Tukey-Kramer Test bei All-Paarvergleichen ist. In SAS fand dieser multiple Test seinen Niederschlag in der Option ADJUST = SIMULATE des LSMEANS-Statements, die Hsu aufgrund einer Arbeit von Edwards und Berry [8] aus dem Jahre 1987 zusammen mit dem SAS-Entwickler Tobias entwickelt hat.

Da die Analyse allgemeiner linearer Kontraste unter Verwendung der Statements ESTIMATE und CONTRAST in GLM und auch in MIXED etwas zu wünschen übrig lässt, haben Westfall et al. [31] im Jahre 1999 ein SAS-Kompendium geschrieben und dazu eine Sammlung von SAS-Macros [32] entwickelt. Diese Macros kann man „downloaden“ unter

http://ftp.sas.com/samples/A56648

Eine andere Weiterentwicklung zur Auswertung allgemeiner linearer Modelle wurde vor allem von Bretz und Genz in den Jahren von 1999 bis 2004 vorangetrieben. Sie

verwenden zur Berechnung geeigneter Quantile der multivariaten t-Verteilung mit allgemeiner Korrelationsmatrix R anstelle von Simulationsverfahren sog. Quasi-Monte-Carlo Methoden, die mit wesentlich geringerer Rechenzeit eine zufriedenstellende Genauigkeit bei der Berechnung der Quantile erzielen, siehe Somerville und Bretz [26, 27], Bretz [2], Bretz et al. [3], Bretz et al. [4], Genz und Bretz [10, 11].
Einige Verbesserungen der Macros von Westfall et al. [32], inbesondere die Weiterentwicklung des Macros „ %SimInterval“ zu „%SimulationsInterval“ werden von Fröbe und Bretz [9] im Internet zur Verfügung gestellt unter

http://bioinf.uni-hannover.de/~froemke/art/ag04/index.html

Zu erwähnen bleibt noch ein „Konkurrenzprodukt“, das „multcomp“-Package von Frank Bretz, Torsten Hothorn, Peter Westfall: MCP in R [1].

R: free software
http://www.R-project.org
multicomp-Package: R archive network
http://cran.R-project.org

1.2 Was wird in diesem Vortrag behandelt ?

Es werden ein- und mehrfaktorielle Designs mit balancierten und unbalanciertem Datenmaterial analysiert, wobei die im Modell vorkommenden zufälligen Effekte als normalverteilt angenommen werden. Wir behandeln All-Paarvergleiche, Vergleiche mit einer Kontrolle und lineare Kontraste, die in der Form von simultanen Vertrauensintervallen dargestellt werden, da diese unserer Ansicht nach höheren Informationsgehalt als multiple Tests mit adjustierten „pvalues“ bieten , siehe auch Hsu [17] und Schaarschmidt [22]. Mit Hilfe des Macros *%macro_hayter.sas* können wir bei allgemeinen linearen Modellen verifizieren, ob die hinreichende Bedingung von Hayter [13,14] zur Einhaltung des multiplen Niveaus beim Tukey-Kramer Test erfüllt ist oder nicht. Hsu [17] untersucht auf den Seiten 185-188 seiner Monografie, welche Konsequenzen die Einhaltung der Hayter-Bedingung hat und nennt ein allgemeines lineares Modell mit verifizierter Hayter-Bedingung ein Modell, das eine „Einweg-Struktur“ (one-way structure) besitzt. Im Kapitel 2 gehen wir auf multiple Verfahren bei der einfachen Varianzanalyse sowie deren Verallgemeinerungen auf allgemeine lineare Modelle ein. Außerdem stellen wir das Macro *%macro_hayter.sas* vor und betrachten sog. „Modelle mit Ein-Weg-Struktur“. Kapitel 3 befasst sich mit Beispielen zur einfaktoriellen Varianz- und Kovarianzanalyse bei balancierten und unbalancierten Daten. Darüber hinaus wird der Einsatz des Macros *SimultanIntervals* von Frömke und Bretz [9] demonstriert.

2 Einfaktorielle Varianzanalyse – fixe Effekte

Wir geben im folgenden eine kurze Einführung zur einfaktoriellen Varianzanalyse, eine ausführlichere Darstellung findet man in Schumacher [25].

Daten. Es liegen k Stufen eines Einflußfaktors vor. Auf der i-ten Stufe (Gruppe) werden n_i Beobachtungen y_{ij} einer zu untersuchenden Zielvariablen ermittelt. Sind die Stichprobenumfänge der k (>2) Gruppen alle gleich groß ($n_i = n$), dann spricht man von balancierten Daten. Unbalancierte Daten liegen vor, wenn die Stichprobenumfänge n_i unterschiedlich sind.

Einfaktorielles Modell mit fixen Effekten

$\mathbf{y}_{ij} = \mu_i + \mathbf{e}_{ij}, \quad i = 1,2,...,k, \quad j = 1,2,...,n_i, \quad N = \sum n_i$

$\mu_1, \mu_2, ..., \mu_k$: Unbekannte Erwartungswerte der k Gruppen,

$\mathbf{e}_{ij}$: unabhängig $N(\mu_i; \sigma_i^2)$-verteilte Zufallsvariablen mit

$\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = \sigma^2, \quad \sigma^2$ unbekannt (Homoskedastizität)

Die Hypothese $H_0: \mu_1 = \mu_2 = ... = \mu_k$ kann zum vorgegebenen Niveau α mittels eines F-Tests geprüft werden. In der Notation der SAS-Prozeduren GLM und MIXED wird die Überschreitungswahrscheinlichkeit ***Pr > F*** angegeben, die zu folgender Entscheidungsvorschrift verwendet werden kann:

$$\text{Ist } \Pr > F \text{ kleiner als } \alpha \text{, dann verwerfe } H_0.$$

2.1 Klassische multiple Tests und simultane Vertrauensintervalle

Multiples Niveau. Würde man für eine Gesamtzahl m (≥ 2) von paarweisen Vergleichen jeweils einen t-Test zum selben Niveau α^* durchführen, dann würde die multiple Irrtumswahrscheinlichkeit α (multiples Niveau, Type I experimentwise error rate, Familywise Error Rate FWE) einen weit höheren Wert als α^* annehmen. Dieses multiple Niveau α ist definiert als die Wahrscheinlichkeit, mit der mindestens eine der m Nullhypothesen H_0^r, r = 1,2,...,m irrtümlicherweise abgelehnt wird.
Eine Abschätzung für das multiple Niveau α ist gegeben durch :

$$\alpha^* \leq \alpha \leq 1-(1-\alpha)^m \leq m \cdot \alpha \quad , \quad m \geq 2.$$

Für die Berechnung von simultanen Vertrauensintervalle für die Differenzen $\mu_r - \mu_t$ zur univariaten Vertrauenswahrscheinlichkeit $1-\alpha$ würde dies bedeuten, dass

$$P\left(\left(\mu_r - \mu_t\right) \in \left(\bar{y}_{r\cdot} - \bar{y}_{t\cdot} \mp t_{1-\frac{\alpha}{2}, N-k} \cdot s\sqrt{\frac{1}{n_r} + \frac{1}{n_t}} \right), \quad m \text{ Intervalle} \right) \geq 1 - m \cdot \alpha \ ,$$

d.h. die multiple Überdeckungswahrscheinlichkeit kann bis auf $1 - m \cdot \alpha$ absinken.

Wir fordern jedoch, dass die multiple Überdeckungswahrscheinlichkeit (*ÜW*) falls möglich genau gleich $1-\alpha$ sein soll.
Es wird ein Wert c_α gesucht mit der Eigenschaft

$$\ddot{U}W := P\left((\mu_r - \mu_t) \in \left(\overline{y}_{r\cdot} - \overline{y}_{t\cdot} \mp c_\alpha \cdot s\sqrt{\frac{1}{n_r} + \frac{1}{n_t}} \right), \quad \text{für m Intervalle} \right) = 1-\alpha$$

Simultane Vertrauensintervalle (bzw. Tests) nach Bonferroni und Sidak

Für $c_\alpha = t_{1-\frac{1}{2}\alpha_{bon}}$ *mit* $\alpha_{bon} = \dfrac{\alpha}{m}$ gilt im allgemeinen: $\ddot{U}W > 1-\alpha$

Für $c_\alpha = t_{1-\frac{1}{2}\alpha_{sid}}$ *mit* $\alpha_{sid} = 1-(1-\alpha)^{\frac{1}{m}}$ gilt im allgemeinen: $\ddot{U}W \geq 1-\alpha$

In SAS werden alle m = k(k–1)/2 simultanen Vertrauensintervalle $\mu_r - \mu_t, 1 \leq r < t \leq k$ berechnet. Sowohl der Sidak-Test und in stärkerem Maße auch der Bonferroni-Test sind konservative Verfahren, da sie das vorgegebene multiple Niveau α in der Regel nicht voll ausschöpfen. Verbesserungen dieser Verfahren können durch Verwendung von Wahrscheinlichkeitsungleichungen nach Hunter [18] und Worsley [33] erreicht werden, diese sind jedoch nicht in SAS implementiert, siehe auch Genz et al. [12]. Das Bonferroni-Verfahren lässt sich ohne weiteres auf allgemeinere lineare Modelle übertragen.

Simultane Vertrauensintervalle (bzw. Tests) nach Scheffe

Für $c_\alpha = \sqrt{F_{1-\alpha,k-1,N-k}}$ gilt im allgemeinen: $\ddot{U}W > 1-\alpha$

Zur Berechnung von simultanen Vertrauensintervallen für alle Paarvergleiche ist der Scheffe-Test nicht zu empfehlen, er ist viel zu konservativ. Seine Stärke liegt in multiplen Vergleichen von vielen (sogar unendlich vielen) sog. linearen Kontrasten $C = \sum_{i=1}^{k} c_i \mu_i$ mit $\sum_{i=1}^{k} c_i = 0$. Auch die Berechnung simultaner Scheffe-Intervalle lässt sich ohne weiteres auf allgemeinere lineare Modelle übertragen.

Simultane Vertrauensintervalle (bzw. Tests) nach Tukey bzw. Tukey-Kramer

Bei balanciertem Datenmaterial, d.h. bei k Gruppen mit gleich großen Stichproben umfängen $n_i \equiv n$, gilt nach Tukey [28]:

$$P\left((\mu_r - \mu_t) \in \left(\overline{y}_{r\cdot} - \overline{y}_{t\cdot} \mp \frac{1}{\sqrt{2}} q_{1-\alpha,k,N-k} \cdot s\sqrt{\frac{2}{n}} \right), \quad 1 \leq r < t \leq k \right) = 1-\alpha,$$

d.h. $c_\alpha = \dfrac{1}{\sqrt{2}} q_{1-\alpha,k,N-k}$, wobei $q_{1-\alpha,k,N-k}$ das Quantil der studentisierten Spannweiten-verteilung ist, siehe Miller [21]. Hierbei ist die Korrelationsmatrix R der verwendeten k-dimensionalen t-Verteilung die Einheitsmatrix I.

Hayter [13] hat 1984 den Beweis geliefert, daß die untenstehende Tukey-Kramer-Variante auch bei unbalancierten Daten das multiple Niveau α einhält.

$$P\left((\mu_r - \mu_t) \in \left(\bar{y}_{r\cdot} - \bar{y}_{t\cdot} \mp \frac{1}{\sqrt{2}} q_{1-\alpha,k,N-k} \cdot s\sqrt{\frac{1}{n_r} + \frac{1}{n_r}} \right), \quad 1 \le r < t \le k \right) \le 1-\alpha$$

Bei balancierten Daten wird das multiple Niveau α voll ausgeschöpft. Wird bei unbalancierten Daten der Tukey-Kramer-Test verwendet, wird das multiple Niveau nicht voll ausgeschöpft. Nach Hayter [14] lässt sich die Berechnung simultaner Tukey-Kramer-Intervalle nur für k = 3 Gruppen auf allgemeinere lineare Modelle übertragen, siehe auch Brown [5]. Im Falle k > 3 ist eine hinreichende Bedingung für die Konservativität der Tukey-Kramer-Variante, dass die Modelle eine „Ein-Weg-Struktur“ besitzen, näheres siehe Abschnitt 2.3.

Simultane Vertrauensintervalle nach Dunnett (Vergleiche mit einer Kontrolle)
Wir stellen nur die Berechnung zweiseitiger Vergleiche mit einer Kontrolle vor, es lassen sich auch ohne weiters auch einseitige Vergleiche durchführen.
Bei unbalanciertem Datenmaterial, d.h. bei k Gruppen mit ungleich großen Stichproben umfängen, gilt nach Dunnett [7]:

$$P\left((\mu_i - \mu_1) \in \left(\bar{y}_{i\cdot} - \bar{y}_{1\cdot} \mp |\, d \,|_{1-\alpha,k-1,N-k,\boldsymbol{\lambda}} \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_1}} \right), \quad 2 \le i \le k \right) = 1-\alpha$$

d.h. $c_\alpha = |\, d \,|_{1-\alpha,k,N-k,\boldsymbol{\lambda}}$, wobei $d\,|_{1-\alpha,k,N-k,\boldsymbol{\lambda}}$ das Quantil der einer „many-to-one“ t-Verteilung ist, siehe auch Westfall et al. [31].
Eine Verallgemeinerung der Dunnett-Tests bei allgemeinen linearen Modellen ist mit Hilfe sog. „Faktoranalytischer Methoden“ möglich, siehe Hsu [17]. Die Einhaltung des multiplen Niveaus durch den Dunnett-Tests und dessen exakte Berechnung ist jedoch nur bei Modellen mit „Ein-Weg-Struktur“ garantiert, näheres siehe Abschnitt 2.3.

Simultane Vertrauensintervalle nach Hochberg und Gabriel (GT2-Verfahren)
Dieses Methode wird vor allem bei balancierten Daten und unabhängigen oder orthogonalen Vergleichen verwendet, wir gehen hier nicht näher darauf ein, siehe Westfall et al. [31]. Sie ist nicht so trennscharf wie die Tukey-Kramer Prozedur, hält jedoch im allgemeinen linearen Modell das multiple Niveau ein, siehe Genz et al. [12] .

2.2 Computer-intensive multiple Prozeduren

Sucht man nach möglichst trennscharfen multiplen Verfahren nicht nur im einfaktoriellen Varianzanalysemodell mit unbalancierten Daten, sondern auch in verallgemeinerten linearen Modellen, die das multiple Niveau einhalten und wenn möglich voll ausschöpfen, dann benötigt man als Basis die k-dimensionale multivariate t-Verteilung und insbesondere die multivariaten t-Wahrscheinlichkeiten

$$T_k(\mathbf{a},\mathbf{b}) = T_k(\mathbf{a},\mathbf{b},\mathbf{R},\nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\det(\mathbf{R})(\nu\pi)^q}} \int_{a_1}^{b_1}\int_{a_2}^{b_2} \dots \int_{a_1}^{b_1} \left(1+\frac{\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}{\nu}\right)^{-\frac{\nu+k}{2}} d\mathbf{x}$$

$$= \int_0^\infty \chi_\nu(s)\cdot\Phi(\frac{s\mathbf{a}}{\sqrt{\nu}},\frac{s\mathbf{b}}{\sqrt{\nu}},\mathbf{R})\,ds \quad \text{mit} \quad \mathbf{x}=(x_1,x_2,\ldots,x_k)^T, a=(a_1,a_2,\ldots,a_k)^T$$

$$\mathbf{b}=(b_1,b_2,\ldots,b_k)^T, \quad \chi_\nu(s) = 2^{1-\frac{\nu}{2}}\cdot s^{\nu-1}\cdot e^{-\frac{s^2}{2}}$$

Diese Integrale numerisch in den Griff zu bekommen, war lange Zeit bei beliebiger Korrelationsmatrix **R** (symmetrisch, positiv definit) ohne geeignete Rechenhilfsmittel nicht möglich, deshalb war es auch so schwierig, den Tukey-Test auf unbalancierte Datenstrukturen oder allgemeinere lineare Modelle zu übertragen. Für einfach strukturierte Korrelationsmatrizen wie **R** = Einheitsmatrix **I** (balancierter Fall) konnte man schon früher die Integrationen durchführen, da sich die k-dimensionalen Integrale dann auf zumeist 3- bzw. 2-dimensionale Integrale zurückführen ließen.
Edwards und Berry [8] sowie Hsu [17] konnten die obigen Integrale mittels Simulationsverfahren numerisch in den Griff bekommen. Dieses Verfahren ist mittels der Option ADJUST=SIMULATE in den SAS-Prozeduren GLM und MIXED umgesetzt worden. Außerdem haben Westfall et al. [32] eine Reihe von SAS-Macros entwickelt, welche in allgemeinen linearen Modellen nicht nur simultane Vertrauensintervalle für All-Paarvergleiche oder Vergleiche mit einer Kontrolle ermitteln, sondern dies auch für allgemeine (lineare oder auch orthogonale) Kontraste ermöglichen. In die Simulation für die Quantilberechnung geht auch die Struktur der Korelationsmatrix R ein.
Das Verdienst von Bretz et al. [2, 3, 4] und Genz et al. [10, 11] ist es, obige Integrale mittels Quasi-Monte Carlo-Methoden noch schneller und trotzdem genau zu berechnen. Froemke und Bretz [9] haben die entsprechenden SAS-Macros von Westfall et al. [32] weiterentwickelt, wir werden diese an Beispielen im Kapitel 3 verwenden.
Sowohl die Simulations- als auch die Quasi-Monte-Carlo Methoden zur Berechnung von (im numerischen Sinne) „beliebig“ genauen Quantilen haben natürlich auch ihren Preis, sie benötigen, falls die Zahl der Gruppen 20 übersteigt, doch erhebliche Rechenzeiten, siehe Genz et al. [12]. Für sehr große Gruppenzahlen können dann auch wieder Rechenzeit sparende konservative Verfahren wie Tukey-Kramer Tests als auch verbesserte Bonferroni-Tests eingesetzt werden, siehe Genz et al. [12].

2.3 Das Macro „macro_hayter.sas”

Hayter [14] hat 1989 den Beweis geliefert, daß die Tukey-Kramer-Variante auch bei allgemeinen linearen Modellen mit einer allgemeineren Kovarianzstruktur, die auch nichtdiagonale Kovarianzmatrizen zulässt, das multiple Niveau α einhält, falls die Hayter-Bedingung erfüllt ist.

Im allgemeinen linearen Modell gelte:

$\hat{\mu}_r$ sei eine (erwartungstreue) Schätzung von μ_r mit der allgemeinen Kovarianzmatrix $COV\left(\left(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k\right)^T\right) = \sigma^2 \cdot V = \sigma^2 \cdot \left(v_{ij}\right)_{i,j=1,2,\ldots,k}$, dann ist $Var(\hat{\mu}_i - \hat{\mu}_j) = \sigma^2 \cdot \left(v_{ii} + v_{jj} - 2v_{ij}\right) =: d_{ij}$

Gilt folgende hinreichende Bedingung nach Hayter [14]:

(H) $d_{ij} = a_i + a_j$ für alle $1 \le i < j \le k$ und alle a_i, a_j positiv, dann halten die simultanen Vertrauensintervalle nach Tukey-Kramer die multiple Überdeckungswahrscheinlichkeit 1-α ein.

Soll diese Bedingung verifiziert werden, dann muss man ein lineares Gleichungssystem mit k(k-1)/2 Gleichungen für k Unbekannte berechnen.
Im Falle k = 3 sind das 3 lineare Gleichungen für 3 Unbekannte, dieses System hat eine eindeutige Lösung mit i. a. positiven Lösungen a_1, a_2, a_3. Das bedeutet, dass für k = 3 im allgemeinen linearen Modell die Tukey-Kramer Prozedur das multiple Niveau einhält, siehe auch Brown [5].
Im Falle k > 3 liegt ein überbestimmtes Gleichungssystem vor, das im allgemeinen keine Lösung besitzt. Mit Hilfe der SAS-Prozedur MIXED (nicht mit GLM möglich) kann man sich mit Hilfe des ODS (Output Delivery System) die SAS-Datei „diffs" beschaffen, in der die Standardfehler $\sqrt{d_{ij}}$ aufgelistet sind. Mit Hilfe folgenden mathematischen Tricks lässt sich die Bedingung von Hayter verifizieren:

Wir lösen das analoge lineare Regressionsproblem

$$\begin{pmatrix} d_{12} \\ d_{13} \\ \ldots \\ d_{23} \\ \ldots \\ d_{k-1,k} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & .. & .. & 0 \\ 1 & 0 & 1 & 0 & 0 & .. & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 1 & 1 & 0 & 0 & .. & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ .. \\ a_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{13} \\ \ldots \\ \varepsilon_{23} \\ \ldots \\ \varepsilon_{k-1,k} \end{pmatrix}, \text{ d.h. } \quad \vec{d} = X \cdot \vec{a} + \vec{e}$$

Die Lösung ergibt sich gemäß $\vec{a} = \left(X^T X\right)^{-1} X^T \vec{d}$ und $\vec{e} = \vec{d} - X\vec{a}$.

Falls $\vec{e} = \vec{0}$ <u>und</u> alle a_i, i=1,2,...,k(≥ 3) positiv, dann ist die hinreichende Bedingung von Hayter erfüllt.
Die Autoren des Vortrags haben unter Berücksichtigung dieser theoretischen Überlegungen ein SAS-Macro entwickelt, das diese Rechenschritte nachvollzieht.

Nach Hsu [17] werden Modelle, welche die Hayter-Bedingung (H) erfüllen, Modelle mit einer „Ein-Weg-Struktur“ bezeichnet („one-way-structure“). Bei solchen Modellen haben die Schätzungen $\hat{\mu}_r$, r =1,2,…,k eine Korrelationsmatrix R mit einer sog. „Produkt-Korrelations-Struktur“, es lässt sich die k-dimensionale t-Verteilung mit Korrelationsmatrix R auf das Produkt von k Integralen über die standardisierte Normalverteilung zurückführen, siehe Hsu [17].
Ein- und zweiseitige Vergleiche mit einer Kontrolle werden mittels geeigneter Dunnett-Tests durchgeführt. Bei „Ein-Weg-Struktur“-Modellen halten die Dunnett-Tests exakt das multiple Niveau ein. Außerdem ist die Hayter-Bedingung eine notwendige Bedingung für sog. „varianzbalancierte“ Modelle, z.B. „balancierte unvollständige Blockanlagen, „Lateinische Quadrate“, „Youden Squares“ sowie für Modelle mit „orthogonalen“ Effekten, siehe Hsu [17].
Der Vorteil des Macros *macro_hayter.sas* ist, für jede Analyse, auch mit unbalancierten Daten und/oder Kovarianzanalysen, falls alle LSMEANS schätzbar sind, die „Ein-Weg-Struktur“ des Modells (numerisch) zu verifizieren oder zu falsifizieren.

3 SAS-Programme und Macros

3.1 Einfaktorielle unbalancierte Varianzanalyse

Es sollen vier Gruppen miteinander verglichen werden, wobei die erste Gruppe nur eine Beobachtung, die zweite Gruppe zwei, die dritte Gruppe vier und die vierte Gruppe sieben Beobachtungen enthält. Es wird das klassische einfaktorielle Varianzanalysemodell unterstellt, siehe Kapitel 2. Im folgenden wird unter anderem die Einhaltung der Hayter-Bedingung gezeigt und orthogonale Kontraste untersucht.

SAS-Programm

```
DATA unb_1fak;
DO gruppe=1 TO 4; DO j=1 TO 7;
INPUT y@@;OUTPUT;END;END;
CARDS;
3 . . . .  .  .
4 5 . . .  .  .
4 5 6 7 .  .  .
5 6 7 8 9 10 11
PROC PRINT; RUN;
TITLE 'ALL-PAARVERGLEICHE';
PROC MIXED DATA=unb_1fak;
CLASS gruppe ;
MODEL y=gruppe ;
LSMEANS gruppe /PDIFF=ALL CL COV
ADJUST=SIMULATE(SEED=121211 ACC=0.0005 EPS=0.01 REPORT CVADJUST);
ODS OUTPUT diffs=lsm;
RUN;
```

```
     /* k(k-1)/2 Varianzen aller Paardifferenzen werden erzeugt */
DATA var_diffs;SET lsm;
dij=StdErr**2; KEEP dij;
RUN;

%INCLUDE 'C:\macro_hayter.sas';        /* Pfad evtl. anpassen */
%hayter(var_diffs);

TITLE 'Zwei-seitige VERGLEICHE mit Kontrolle';
PROC MIXED DATA=unb_1fak;
CLASS gruppe ;
MODEL y=gruppe  ;
LSMEANS gruppe  /PDIFF=CONTROL
 ADJUST=SIMULATE(SEED=121211 ACC=0.01 EPS=0.01 REPORT CVADJUST);
RUN; QUIT;
```

Output (stark gekürzt)

```
The Mixed Procedure
Covariance Parameter
      Estimates
Cov Parm     Estimate
Residual       3.3500
```

Hieraus entnehmen wir die Schätzung von $\sigma^2 : \hat{\sigma}^2 = 3.35$

```
Type 3 Tests of Fixed Effects
              Num     Den
Effect         DF      DF    F Value    Pr > F
gruppe          3      10       3.97    0.0421
```

Globaler F-Test liefert auf dem 0.05-Niveau Signifikanz

```
    Details for Quantile Simulation
Random number seed                121211
Comparison type                      All
Sample size                      1260622
Target alpha                        0.05
Accuracy radius (target)          0.0005
Accuracy radius (actual)          0.0003
Accuracy confidence                  99%
```

```
                     Simulation Results
                                Estimated     99% Confidence
Method          95% Quantile        Alpha         Limits
Simulated           3.024570       0.0500    0.0497    0.0503
Tukey-Kramer        3.059352       0.0473    0.0470    0.0476
Bonferroni          3.276841       0.0336    0.0333    0.0338
Sidak               3.264268       0.0342    0.0339    0.0345
GT-2                3.199288       0.0380    0.0377    0.0382
Scheffe             3.335385       0.0305    0.0303    0.0308
T                   2.228139       0.1738    0.1733    0.1743
```

Aus obigem Output ersehen wir, dass die Simulationsmethode nach Edwards, Berry und Hsu die trennschärfsten Tests bzw. die kürzesten simultane Vertrauensintervalle liefert. Es wird diese Methode empfohlen, sie ist auch im allgemeinen die trennschärfste bei einfachen Varianzanalysen mit unbalancierten Daten.

```
Löse das Gleichungssystems d_ij=a_i + a_j
Lösung existiert , falls alle e_i gleich 0
Sind die Lösungen a_i positiv ?
       E_I          A_I
        0           3.35
        0           1.675
        0           0.8375
        0           0.4785714
        0
        0
Alle a_i größer 0 und Hayterzahl gleich 0:
Es liegt ein 1-Weg_Struktur_Modell vor

HAYTERZAHL
     0
```

Die Anwendung von *macro_hayter.sas* zeigt, dass ein „1-Weg_Struktur"-Modell vorliegt. Somit ist der folgende zweiseitige DUNNETT-Test exakt !

```
Zwei-seitige VERGLEICHE mit Kontrolle
             Simulation Results
                                       Exact
Method                 95% Quantile     Alpha
Simulated                  2.628661    0.0500
Dunnett, two-sided         2.628661    0.0500
Bonferroni                 2.870073    0.0336
Sidak                      2.860154    0.0342
GT-2                       2.828869    0.0360
Scheffe                    3.335385    0.0156
T                          2.228139    0.0958
```

Die Simulationsmethode und der exakte Dunnett-Test liefern identische Resultate.

Macros von FROEMKE UND BRETZ, 2004

Nach der Demonstration der Anwendung des hayter.sas - macros gehen wir auch noch auf von Westfall et al. [32] entwickelte, von Frömke und Bretz [9] weiterentwickelte Macros ein. Wir speichern die „download"-baren Macros *MakeGLMStats.sas*, *Contrasts.sas*, *Estimate.sas*, *Simultan_Intervals.sas* z.B. auf dem Verzeichnis C:\ ab. Beachtet man die Syntax- Anweisungen von Westfall et al. [32] , S. 92-96 sowie für das Beispiel einer Split-Plot Anlage bei Frömke und Bretz [9], S. 1326-1327, dann können mit dem Macro *Simultan_Intervals.sas* simultene Vertrauensintervalle für allgemeine Kontraste (unter Einbezug des Allgemeinmittels) als auch lineare oder auch orthogonale Kontraste berechnet werden. Natürlich müssen diese Kontraste schätzbare Funktionen sein, siehe Little et al. [20].

```
TITLE 'FROEMKE und BRETZ: Orthogonale Kontraste';
%INC 'C:\MakeGLMStats.sas';
%MakeGLMStats(dataset=unb_1fak,
               classvar=gruppe,
               yvar=y,
               model=gruppe); /*greift auf SAS-Datei unb_1fak zu*/
%macro Contrasts;
 C = {0  -1    1    0    0 ,
      0   0    0   -1    1 ,
      0  -0.5 -0.5 0.5  0.5 };   /* 1.Null für Allgemeinmittel */
   C = C` ;                                /* transponieren */
   Clab = {"Gruppe 2 - 1", "Gruppe 4 - 3",
   "Mittel aus Gruppe 3+4 - Mittel aus Gruppe 1+2" };
/* Contrast labels */
%mend;
%SimultanIntervals (seed=121211,side=U);
```

Output (gekürzt)

Das Statement `%SimultanIntervals (seed=121211,side=U);`
bewirkt die Berechnung einseitiger simultaner Vertrauensintervalle:

```
Estimated 95% Quantile = 2.426206
                            Standard          ---- Pr > t ---
Contrast          Estimate   Error    t Value   Raw     Adjusted
Gruppe 2 - 1      1.5000     2.2417    0.67    0.2593   0.5730
Gruppe 4 - 3      2.5000     1.1472    2.18    0.0272   0.0746
Mi3+4 - Mi1+2     3.0000     1.2591    2.38    0.0192   0.0535
  95% Confidence
     Interval
 -3.9387    Infty
 -0.2833    Infty
 -0.0548    Infty
```

Man beachte: Die größten multiplen Irrtumswahrscheinlickeiten treten gerade dann auf, wenn die betrachteten multiplen Vergleiche (nahezu) unabhängig sind. Deshalb ist insbesondere beim multiplen Vergleich orthogonaler Kontraste eine Adjustierung der einzelnen Irrumswahrscheinlichkeiten notwendig, siehe Westfall et al. [31], S. 62-64.

Simultane Intervalle für die Mittelwerte

```
TITLE 'FROEMKE und BRETZ: Simultane Intervalle für Mittelwerte';
%MakeGLMStats(dataset=unb_1fak,classvar=gruppe,
              yvar=y,model=gruppe);
%macro Contrasts;
   C = {1   1  0  0  0 ,
        1   0  1  0  0 ,
        1   0  0  1  0 ,
        1   0  0  0  1 };
   C = C` ;
   Clab = {"Gruppe1", "Gruppe2", "Gruppe3", "Gruppe4"};
%mend;
%SimultanIntervals(seed=121211,conf=0.95,side=B);
```

Das Statement `%SimultanIntervals (seed=121211,side=B);`
bewirkt die Berechnung zweiseitiger simultaner Vertrauensintervalle.

Output

```
                    Estimated 95% Quantile = 2.984014
                      Standard        --- Pr > |t| --     95%
                                                        Confidence
Contrast  Estimate Error t Value     Raw Adjusted        Interval
Gruppe1   3.0000  1.8303   1.64   0.1322  0.4006  -2.4616   8.4616
Gruppe2   4.5000  1.2942   3.48   0.0060  0.0216   0.6380   8.3620
Gruppe3   5.5000  0.9152   6.01   0.0001  0.0005   2.7692   8.2308
Gruppe4   8.0000  0.6918   11.56  <.0001  <.0001   5.9357  10.0643
```

3.2 Einfaktorielle Kovarianzanalyse

Es sollen vier Gruppen miteinander verglichen werden, wobei zu einer Zielvariablen y auch noch eine Kovariable x simuliert wird. Es wird ein einfaktorielles Kovarianz-analysemodell unterstellt mit normalverteilten Fehlern.
#

```
DATA kova;
DO a=1 TO 4;
DO j=1 TO 5;
y=2*a+RANNOR(12347); x=a**2+j;OUTPUT;END;END;
CARDS;
PROC PRINT; RUN;
PROC MIXED DATA=kova;
CLASS a  ;
MODEL y=a x / ;
LSMEANS a  /PDIFF=ALL ADJUST=SIMULATE
           (ACC=0.001 EPS=0.01 REPORT CVADJUST) COV;
ODS OUTPUT diffs=lsm;
RUN; QUIT;
DATA var_diffs_a;SET lsm;
dij=StdErr**2;
KEEP dij;
%INCLUDE 'C:\macro_hayter.sas';
%hayter(var_diffs_a);
```

Output (stark gekürzt)

```
#####################The Mixed Procedure
                     Simulation Results
                                Estimated      99% Confidence
Method            95% Quantile      Alpha           Limits
Simulated             2.680922     0.0500     0.0491     0.0509
Tukey-Kramer          2.882149     0.0343     0.0335     0.0352
Bonferroni            3.036283     0.0258     0.0250     0.0265
Sidak                 3.025851     0.0263     0.0256     0.0271
.   .    .    .    .    .    .    .    .
```

Die Simulationsmethode liefert auch hier das trennschärfste Verfahren. Die Anwendung von *macro_hayter.sas* ergibt, dass kein Modell mit „Ein-Weg-Struktur“ vorliegt. Deshalb halten Vergleiche nach Dunnett im allgemeinen das multiple Niveau nur approximativ ein.

##

```
Löse das Gleichungssystems d_ij=a_i + a_j
Lösung existiert , falls alle e_i gleich 0
Sind die Lösungen a_i positiv ?
   E_I              A_I
-1.720762         2.1434055
0.5433986         0.3320769
1.1773636         -0.271699                 /* negativer Wert  */
1.1773636         3.9547341
0.5433986
-1.720762

Alle a_i größer 0 und Hayterzahl gleich 0:
Es liegt ein 1-Weg_Struktur_Modell vor      /* Hier: NEIN       */

HAYTERZAHL
3.0471263                      /* Auch Hayterzahl nicht gleich 0 */ ############
```

Bemerkungen

1. Bei der Auswertung mehrfaktorieller Modelle mit fixen Effekten und „schwach“ unbalancierten Daten, siehe Schumacher [25], sollte die Analyse der „Ein-Weg-Struktur“ des Modells für jeden Faktor getrennt vorgenommen werden. Am besten verwendet man für jeden Effekt ein eigenes LSMEANS-Statement, nur dann kann die ODS-Datei „diffs“ ohne weiteres direkt zur Analyse mit *macro_hayter.sas* verwendet werden. Da alle LSMEANS schätzbar sind, macht aber die Erweiterung auf mehrere Effekte keine Schwierigkeiten. Eine ausführliche Darstellung der Analyse mehrfaktorieller Modelle mit fixen Effekten unter Verwendung von geeigneten Macros findet man bei Westfall et al. [31], Kapitel 9 und [32].

2. Auf die Analyse „stark“ unbalancierter Daten (d.h. fehlenden Zellen und damit nicht schätzbaren LSMEANS) gehen wir hier nicht ein.

3. Einige Bemerkungen zu multiplen Vergleichen bei mehrfaktoriellen Modellen mit gemischten Effekten, insbesondere auch unter Heteroskedastizität, findet man im Kapitel 10 von Westfall et al. [31]. Weitere Beiträge zu multiplen Vergleichen bei gemichten Modellen siehe Hochberg [16], Hsu [17] sowie Voss und Hsu [29].

Das Macro *macro_hayter.sas* sowie etliche Beispiele findet man zum Downloaden in

http://www.uni-hohenheim.de/inst110/mitarbeiter/schumach

Literatur

[1] Bretz F., Hothorn T. and Westfall P.H. (2002): On multiple comparisons in R. R Newsletter, 2(3), 14-17.

[2] Bretz F. (2001): On the numerical availability of multiple comparison procedures. Medizinische Informatik, Biometrie und Epidemiologie, 32(2-3): 123-124.

[3] Bretz F., Hayter A. and Genz A. (2002): Critical point and power calculations for the studentized range test for generally correlated means. Journal of Statistical Computation and Simulation, 71:85-99.

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

[5] Brown L. D. (1979): A proof that the Tukey-Kramer multiple comparison procedure for differences between means is level α for 3, 4, or 5 treatments. Technical report, Statistics Center, Cornell University, Ithaca, NY.

[6] Dunnett C. W. (1955): A multiple comparison procedure for comparing several treatments with a control. Journal of the Am. Statistical Association, 50:1096-1121.

[7] Dunnett C. W. (1980): Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. Journal of the American Stat. Association, 75:789-795.

[8] Edwards D. G. and Berry J. J. (1987): The efficiency of simulation-based multiple comparisons. Biometrics, 43:913-928.

[9] Frömke C. and Bretz F. (2004): Simultaneous tests and confidence intervals for the evaluation of agricultural field trials, Agronomy Journal, 96:1323-1330.
URL: http://bioinf.uni-hannover.de/~froemke/art/ag04/index.html

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

[11] Genz A. and Bretz F. (2002): Methods for the computation of multivariate t-probabilities. Journal of Computational and Graphical Statistics, 11:1-22.

[12] Genz A. , Bretz F. , Hochberg Y. (2004): Approximations to multivariate t-integrals with application to multiple comparison procedures. Recent Developments in MCP, Institute of Mathematical Statistics LNMS 47: 24-32

[13] Hayter A. J. (1984): A proof of the conjecture, that the Tukey-Kramer multiple comparisons procedure is conservative. Annals of Statistics, 12:61-75.

[14] Hayter A. J. (1989): Pairwise comparisons of generally correlated means. Journal of the American Statistical Association, 84:208-213.

[15] Hochberg Y. and Tamhane A.C. (1987): Multiple Comparison Procedures, John Wiley, Mew York.

[16] Hochberg Y. and Tamhane A. C. (1983): Multiple comparisons in a mixed model. American Statistician, 37:305-307.

[17] Hsu J. (1996): Multiple Comparisons. Theory and Methods. Chapman and Hall.

[18] Hunter D. (1976): An upper Bound for the Probability of a Union. Journal of Applied Probability, 13:597-603.

[19] Kramer C. Y. (1956): Extension of multiple range tests to group means with unequal numbers of replications. Biometrics, 12: 309-310.

[20] Little R. J. and Rubin D. B. (1987): Statistical Analysis with Missing Data, New York: Wiley.

[21] Miller R. G. (1981): Simultaneous Statistical Inference. Springer-Verlag, Heidelberg and Berlin, second Edition.

[22] Schaarschmidt F. (2004): Konfidenzintervalle als interpretatorisches Mittel zum Vergleich von 2 oder k Stichproben in Ein- oder Mehrweganlagen gartenbaulicher Versuche.

URL http://www.bioinf.uni-hannover.de/teaching/fallstudien/schaarschmidt1.pdf

[23] Scheffé H. (1953): A method for judging all contrasts in the analysis of variance. Biometrika, 40:87-104.

[24] Scheffé H. (1959): Analysis of Variance, New York: Wiley.

[25] Schumacher E. (2004): Einführung in die Biometrie 3. Vergleich von mehr als zwei Parametern (Varianzanalyse). 2. Auflage, Saphir-Verlag. ISBN 3-930037-17-3.

[26] Somerville P. N. and Bretz F. (2001) : Obtaining critical values for simultaneous confidence intervals and multiple testing. Biometrical Journal, 43:657-663.

[27] Somerville P. N. and Bretz F. (2001) FORTRAN 90 and SAS-IML programs for computation of critical values for multiple testing and simultaneous confidence intervals. Journal of Statistical Software, 6:1-14.

[28] Tukey J. W. (1953): The problem of Multiple Comparisons. Dittoed manuscript of 396 pages, Department of Statistics, Princeton University.

[29] Voss D. T. and Hsu J. C. (1995): Multiple comparisons for an unbalanced a x b design under mixed models with interaction. Technical Report 560, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247.

[30] Westfall P. H. and Young S. S. (1993): Resampling-Based Multiple Testing; Examples and Methods for P-Value Adjustment. John Wiley & Sons Inc., New York.

[31] Westfall P. H., Randall T. D., Rom D., Wolfinger R. D. and Hochberg Y. (1999a): Multiple comparisons and multiple tests using the SAS system. SAS Institute Inc. Cary, North Carolina.

[32] Westfall P. H., Tobias R.D., Rom D., Wolfinger R. D. and Hochberg Y. (1999b): Multiple comparisons and multiple tests. SAS publication 56648. SAS Institute Inc. Cary, North Caroline. URL http://ftp.sas.com/samples/A56648.

[33] Worsley K. J. (1982): An improved Bonferroni inequality and applications. Biometrika, 69:297-302.