# Multiple Mittelwertvergleiche
# - parametrisch und nichtparametrisch -
# sowie $\alpha$-Adjustierungen
# mit
# praktischen Anwendungen mit R und SPSS

Version 0.90
(30.10.2014)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@uni-koeln.de

Universität zu Köln

# Vorwort

## Entstehung

In den letzten Jahren hatte ich mehrfach Kurse zum Thema „nichtparametrische Methoden mit SPSS“ bzw. Kurse zur Programmiersprache S und dem System R am RRZK gehalten. Dort hatte sich gezeigt, dass ein großes Interesse an nichtparametrischen statistischen Verfahren besteht, insbesondere im Bereich Varianzanalyse. Dazu hatte ich ein umfangreiches Skript erstellt, in dem auch das Thema „nichtparametrische multiple Mittelwertvergleiche und $\alpha$-Adjustierungen“ behandelt werden sollte. Allerdings merkte ich schnell bei der Sichtung der Verfahren und der aktuellen Literatur, dass dies ein eigenes „Thema“ sein muss. Denn bei der Behandlung der nichtparametrischen Methoden kommt man auf der einen Seite nicht umhin, die generelle Problematik multipler Vergleiche zu besprechen. Auf der anderen Seite gibt es nicht so viele in Frage kommenden Verfahren, so dass man sich - wie bei der nichtparametrischen Varianzanalyse - häufig der parametrischen Methoden bedienen muss. Somit musste das Skript letztlich das ganze Themensprektrum behandeln. Und ich war doch überrascht, wie viele Verfahren, insbesondere im Gebiet der $\alpha$-Adjustierungen, in den letzten Jahren entwickelt worden sind und noch laufend neue hinzukommen.

Dass dieses Thema sehr kontrovers ist, wenn nicht sogar das am meisten diskuitierte der angewandten Statistik, ist eigentlich schon der Veröfentlichung der klassischen Tests „Tukey“, „Newman-Keuls“ und „Duncan“ in den 50er Jahren bekannt. Leztztlich geht es dabei um die Kontroverse zwischen den konservativen Tests mit strikter Einhaltung des Fehlers 1. Art und den liberalen Tests mit größerer Stärte (Power). Ich selbst favorisiere die zweite Kategorie, versuche mich aber bei den Empfehlungen objektiv zu verhalten.

## Umfang

Das Skript setzt voraus, dass der Leser zum einen mit Varianzanalysen (mehr oder weniger) vertraut ist und zum anderen mit R bzw. SPSS umgehen kann. So werden z.B. bei SPSS weitgehend die Angaben zu den Menüs, über die die einzelnen Funktionen erreichbar sind, zugunsten der SPSS-Syntax ausgespart.

Eine generelle Einführung in die Varianzanalyse, simple effects-Analysen, Ziehen der richtigen Schlüsse etc behandelt werden, ist geplant.

## Spätere Versionen

Das vorliegende Skript ist erst als ein Anfang zu verstehen. Es sollen noch weitere Verfahren berücksichtigt werden. Auch sollen noch Beispiele folgen, wie mit den vorhandenen Methoden Verfahren zu realisieren sind, die nicht softwareseitig unterstützt werden. Das gilt natürlich in besonderem Maße für SPSS, das auf diesem Gebiet nicht sonderlich üppig und aktuell ausgestattat ist.

## Lesehinweise

Die Ergebnistabellen aus R und SPSS sind zum Teil verkürzt wiedergegeben, d.h. Teile, die nicht zum Verständnis erforderlich sind, fehlen hier.

# Inhaltsverzeichnis

# 1. Einleitung

Wilcox [102] schreibt zu Beginn seines Kapitels über multiple Mittelwertvergleiche. „The choice of a multiple comparison procedure is a complex issue...“.

In keinem anderen Gebiet der Angewandten Statistik gibt es so viele Kontroversen über den richtigen Weg wie bei den hier besprochenen Methoden der multiplen Mittelwertvergleiche. Letztlich geht es um die Frage: Muss die Gesamtheit der Vergleiche immer das vorgegebene $\alpha$ einhalten, auch auf die Gefahr hin, dass die Wahrscheinlichkeit für den Fehler 2. Art $\beta$ immens hoch wird und viele Unterschiede nicht nachgewiesen werden können?

Ein Indiz dafür ist die stetig steigende Anzahl der Methoden und noch mehr der Publikationen für multiple Vergleiche. Beim Lesen fast jeder Publikation stößt man immer wieder auf neue, andere Verfahren. Entsprechend viele Verfahren gibt es inzwischen, nicht nur parametrische sondern auch nichtparametrische. Hinzu kommt, dass einige Verfahren unter verschiedenen Autorennamen zu finden sind, oder umgekehrt, dass sich in der Sekundärliteratur verschiedene Tests als derselbe entpuppen. Bei den hier vorgestellten Methoden wurde versucht, sich auf die zu beschränken,

- die inzwischen eine gewisse „Bedeutung“ haben,
- die als empfehlenswert anzusehen sind und
- die in R oder SPSS verfügbar sind.

An dieser Stelle sei vermerkt, dass man in dieser Hinsicht manchmal neidisch auf das Softwaresystem SAS schauen muss, das auf diesem Gebiet - im Gegensatz zu SPSS - wirklich uptodate ist.

## 1. 1 Die Ausgangssituation

Es wird im Folgenden angenommen, dass die Werte einer abhängigen Variablen (Kriteriumsvariablen) x für $k$ Gruppen mit Stichprobenumfängen $n_i$ ($i=1,...,k$) vorliegen, also das Modell einer Varianzanalyse mit unabhängigen Stichproben zugrunde liegt. Ziel ist es, die $k$ Gruppenmittelwerte $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ auf Gleichheit zu überprüfen, genauer: im Fall dass die Hypothese gleicher Mittelwerte verworfen wird, welche Mittelwerte sich von welchen anderen unterscheiden. Hierbei werden sowohl die Fälle gleicher wie auch ungleicher Varianzen $s_i^2$ berücksichtigt. Der Fall abhängiger Stichproben, also von Messwiederholungen, wird an manchen Stellen separat behandelt. Ebenso werden nichtparametrische Methoden vorgestellt, insbesondere für den Fall nicht normalverteilter Residuen oder ordinal skalierter Kriteriumsvariablen.

Diese Mittelwertvergleiche werden üblicherweise im Anschluss an eine Varianzanalyse durchgeführt. Denn signifikante Effekte besagen nur, dass zwischen irgendwelchen Gruppen Mittelwertunterschiede bestehen, geben aber keinen weiteren Aufschluss darüber, welche Gruppen oder Stufen dies nun sind. Deswegen werden diese Vergleiche auch häufig mit *post-hoc-Tests* bezeichnet. Die meisten der vorgestellten Verfahren können aber auch unabhängig von einer Varianzanalyse angewandt werden oder ersetzen sogar eine solche. An dieser Stelle sei ausdrücklich auf das Kapitel 14.1 hingewiesen, worin die Bedeutung eines vorangehenden globalen Tests, also z.B. eines F-Tests, mittels einer Varianzanalyse dargelegt wird.

Für diese Fragestellung unterscheidet man grundsätzlich:

- *geplante* Vergleiche, *apriori-Vergleiche* oder *Kontraste*, die als Hypothesen bereits *vor* der Untersuchung, d.h. vor Erhebung des Datenmaterials, vorliegen, und
- *multiple Mittelwertvergleiche* oder *posthoc-Tests*, für die keine speziellen Hypothesen vorliegen und die üblicherweise durchgeführt werden, wenn die Varianzanalyse einen signifikanten Effekt aufzeigt, der dann näher analysiert werden soll. Das allgemeinste, aber auch häufig das schwächste Verfahren in dieser Kategorie sind die *paarweisen Vergleiche mit α-Adjustierungen.*

## 1. 2 multiple Vergleiche

Neben den hier im Vordergrund stehenden paarweisen Vergleichen sei an dieser Stelle noch auf eine andere Gruppe von multiplen Vergleichen aufmerksam gemacht: Den Vergleich mehrerer Merkmale für zwei oder mehrere Gruppen. So werden z.B. in der Gendatenanalyse für Tausende von Variablen eine Experimental- mit einer Kontrollgruppe verglichen. Hierbei geht es hauptsächlich um die Methoden der α-Adjustierung, die in Kapitel 2 ausführlich behandelt werden und größtenteils für beide Arten von multiplen Vergleichen anwendbar sind.

## 1. 3 Klassifizierung der Verfahren

Bei den verschiedenen Methoden unterscheidet man häufig die *Ein-Schritt-Verfahren* (*single step procedures*) und die s*chrittweisen Verfahren* (*stepwise procedures*). Bei den ersten werden alle Paarvergleiche unabhängig voneinander durchgeführt. Hierunter fallen die meisten bekannten Methoden wie z.B. Tukey, Scheffé und Bonferroni. Bei schrittweisen Vnerfahren wird meist mit dem Test einer globalen Hypothese begonnen (*step-down procedures*). Im Fall einer Signifikanz wird die Hypothese verfeinert und getestet usw. Hierzu zählen z.B. die Methoden von Student-Newman-Keuls, Holm und REGWQ. Darüber hinaus gibt es auch Verfahren, bei denen mit Paarvergleichen angefangen wird und in Abhängigkeit vom Ergebnis die Hypothesen verallgemeinert werden (*step-up procedures*), so z.B. die Methoden von Hochberg und Tamhane.

Unabhängig von dieser Klassifizierung gibt es auch die Einteilung in solche Verfahren, die klassischerweise Mittelwertdiffernzen mittels eines kritischen Wertes oder über die Berechnung eines p-Wertes auf Signifikanz überprüfen, und in solche, die für jede Mittelwertdifferenz ein Konfidenzintervall aufstellen. Hier sind dann zwei Mittelwerte verschieden, wenn das Konfidenzintervall der Differenz nicht die Null enthält. Beide Methoden sind natürlich inein ander überführbar.

## 1. 4 Besondere Verteilungen

Bei den multiplen Vergleichen werden für die Signifikanztests neben den bekannten t- und F-Verteilungen noch weitere sonst weniger bekannte benutzt. Basis sind wie immer normalverteilte Variablen $x_i$.

Die Tests der klassischen Verfahren basieren auf der *studentized range-Verteilung*. Dies ist die Verteilung der (standardisierten) Spannweite (engl. *range*) der *k* Mittelwerte, also von $(max(\bar{x}_i) - min(\bar{x}_i))/(s/\sqrt{n})$ bzw. der größten aller *m* Mittelwertdifferenzen. Hierbei ist *s* die (einheitliche) Standardabweichung der untersuchten Variablen. Es werden sowohl gleiche $n_i$ wie auch gleiche $s_i$ vorausgesetzt. Die Quantilswerte $q_a(k,df)$ hängen neben dem vorgegebenen α ab von der Spannweite *k* und den Freiheitsgraden *df* der Fehlervarianz $s^2$.

Für $n \to \infty$ geht diese Verteilung über in die *normal range-Verteilung*.

Für den Fall ungleicher $n_i$ wird verschiedentlich die *studentized maximum modulus-Verteilung* benutzt. Dies ist die Verteilung von $(max|x_i - \bar{x}|)/(s/\sqrt{n})$. Hier hängen die Quantilswerte $r_a(m,df)$ neben dem vorgegebenen $\alpha$ von der Anzahl der Vergleiche *m* und den Freiheitsgraden *df* der Fehlervarianz $s^2$ ab.

Bei der Verwendung dieser Verteilungen, wie auch bei der F-Verteilung, gibt es keine Unterscheidung zwischen ein- und zweiseitigen Tests. Dagegen wird man bei den anderen Verfahren, deren kritische Werte sich über die t-Verteilung errechnen, in der Regel zweiseitige Tests durchführen.

## 1. 5 Statistische Fehler

Da sich hier vieles um die Raten für den Fehler 1. Art und den Fehler 2. Art dreht, hier zunächst noch mal deren Definition:

- Fehler 1. Art (*type I error*):
  Zwei Mittelwerte werden durch einen Test als unterschiedlich erkannt, obwohl sie in der Grundgesamtheit tatsächlich gleich sind. Die Fehlerrate dafür ist $\alpha$ und wird i.a. vor der Durchführung des Tests vom Untersucher, z.B. in Abhängigkeit vom Stichprobenumfang *n*, festgelegt.
- Fehler 2. Art (*type II error*):
  Zwei Mittelwerte werden durch einen Test als nicht unterschiedlich erkannt, obwohl sie in der Grundgesamtheit tatsächlich verschieden sind. Die Fehlerrate dafür ist $\beta$ und ist i.a. unbekannt, genauer: Sie kann für die hier besprochenen Tests kaum errechnet werden. Man weiß lediglich, dass $\beta$ mit größer werdendem $\alpha$ abnimmt und dass mit steigendem Stichprobenumfang *n* die Fehlerrate abnimmt.
  An dieser Stelle muss auch noch an einen anderen Begriff erinnert werden: die *Macht* (engl. *power*) eines Tests. Dies ist die Wahrscheinlichkeit, mit der ein Fehler 2. Art vermieden wird. Diese beträgt also $1-\beta$ .

Bei der Auswahl eines Tests aus den zahlreichen nachfolgend aufgeführten Verfahren geht es primär um die Abwägung zwischen den Fehlern 1. und 2. Art, d.h. welcher der beiden Fehler akzeptabel ist und damit größer sein darf.

Bei multiplen Vergleichen werden allerdings mehrere Tests gleichzeitig durchgeführt, d.h. es geht es um mehrere Testhypothesen, die zusammen eine Familie (*family*) oder ein Experiment bilden. Dies führt zu einer Reihe weiterer Definitionen von Fehlerraten. Angenommen, bei einem Signifikanztest treten die beiden Fehler 1. und 2. Art mit folgenden Häufigkeiten auf (in Klammern die „üblichen“ Wahrscheinlichkeiten):

| | | Test-Entscheidung | | |
|---|---|---|---|---|
| | | $H_0$ wahr | $H_0$ falsch ($H_1$) | Summe |
| Wirklichkeit | $H_0$ wahr | A | B ($\alpha$) | $m_0$ |
| | $H_0$ falsch ($H_1$) | C ($\beta$) | D | $m_1$ |
| | Summe | A+C | B+D | m |

- PCER (*per-comparison error rate)* oder CER *(comparisonwise error rate)*
  Dies ist die erwartete Anzahl falscher $H_0$-Ablehnungen B (Anzahl der Fehler 1. Art) bezogen auf die Anzahl der Tests *m*:
  $PCER = E(B)/m$
  Dies entspricht der Rate $\alpha$ für den Fehler 1. Art.
- PFER (*per-family error rate)*
  Dies ist die erwartete Anzahl falscher $H_0$-Ablehnungen B (Anzahl der Fehler 1. Art) :
  $PCER = E(B)$
- FWER (*familywise error rate*) oder EER (*experimentwise error rate*):
  Dies ist die Wahrscheinlichkeit, mindestens einmal $H_0$ fälschlich abzulehnen, also mindestens einen Fehler 1. Art zu machen:
  $FWER = P(B > 0)$
  Die FWER oder EER ist die Rate, mit der beim Testen der Hypothesen der Familie oder des Experiments der Fehler 1. Art auftritt, also die Fehlerrate, mit der Nullhypothesen für irgendeinen der Tests der Familie fälschlicherweise abgelehnt werden.
- FDR (*false discovery rate*, gelegentlich auch *false detection rate*)
  Dies ist die zu erwartende Rate falscher Annahmen von $H_1$:
  $FDR = E(B / (B+D))$
  Damit diese Rate aber auch für den Fall, dass $H_1$ kein einziges Mal angenommen wurde, also $B+D=0$, definiert ist, gibt es die folgende Variante für FDR:
  $FDR = E(B / (B+D) \mid (B+D) > 0) \cdot P((B+D) > 0)$

Die FWER ist im Kontext der multiplen Vergleiche die Wichtigste. Deswegen noch eine kurze anschauliche Darstellung: Bei einem Test, z.B.dem t-Test, bedeutet ein $\alpha$=0.05 als Rate für den Fehler 1. Art, dass bei 100 Vergleichen etwa 5 zufällig und damit falsch signifikant sind. Bei einem multiplen Vergleich von 3 Gruppen, also 6 Paarvergleichen, z.B. mit dem Tukey-Test, bedeutet eine FWER=0.05, dass bei 100 solchen multiplen Vergleichen, also insgesamt 600 Paarvergleichen, auch nur etwa 5 zufällig und damit falsch signifikant sind.

Während bei fast allen Verfahren die Devise gilt: In erster Linie muss $\alpha$ unter Kontrolle gehalten werden, verfolgt die FDR ein anderes Ziel verfolgen, nämlich die Rate falscher Signifikanzen (*false discoveries*) bezogen auf die Rate aller Signifikanzen zu minimieren (vgl. Wikipedia [121]). Eine ausführliche Beschreibung mit mathematischem Hintergrund bietet Genovese [122]. Die Berücksichtigung der FDR ist weniger stringent als die von $\alpha$, d.h. Verfahren, die dieses berücksichtigen, verfügen über mehr Power, halten auf der anderen Seite nicht unbedingt ein vorgegebenes $\alpha$ ein.

## 1. 6 Die Problematik der multiplen Vergleiche

Die übliche Ausgangssituation ist: Der globale Test der Varianzanalyse zum Vergleich der Gruppenmittelwerte ergibt ein signifikantes Resultat, d.h. irgendwelche der *k* Gruppenmittelwerte unterscheiden sich. Eine nahe liegende Lösung auf die Frage, welche der Gruppen sich unterscheiden, könnte sein, alle Gruppen paarweise miteinander zu vergleichen, z.B. einfach mit dem vorher gewählten Verfahren der Varianzanalyse. In diesem Fall würden *k(k-1)/2* Vergleiche durchgeführt. Bei z.B. 4 Gruppen sind dies schon 6 Vergleiche und bei 6 Gruppen schon 15 Vergleiche. Würden nun alle Vergleiche mit derselben vorgegebenen Irrtumswahrscheinlichkeit $\alpha$, z.B. $\alpha$=0,05 , bewertet, so erhielte man weitaus mehr zufällige und damit falsche Signifikanzen, als das $\alpha$ suggeriert. Denn nach der Wahrscheinlichkeitsrechnung beträgt die Wahrscheinlichkeit, dass bei *m* durchgeführten Tests $H_0$ fälschlich abgelehnt wird (sofern

die Tests unabhängig voneinander sind):

$$1 - ( 1 - \alpha )^m$$

Das bedeutet, dass z.B. bei 4 Gruppen, also 6 Tests, die reale Irrtumswahrscheinlichkeit schon 0,265 und bei 6 Gruppen und damit 15 Vergleichen 0,537 beträgt. D.h. man erhält schon bei „normaler“ Gruppenanzahl *k* so viele zufällig falsche Signifikanzen, dass das Prozedere in Frage gestellt werden muss. Das Ziel ist es dagegen, dass für eine solche *Familie von Vergleichen* insgesamt das Fehlerrisiko (*familywise error rate*, *FWER*) dem vorgegebenen $\alpha$ entspricht. Dieses Problem ist auch ausführlich in Wikipedia beschrieben [104].

Im Folgenden wird mit *m* die Anzahl der Vergleiche einer Familie bezeichnet. Damit ist *m*, wie oben erwähnt, z.B. *k(k-1)/2* , wenn die Mittelwerte alle paarweise oder *k-1*, wenn alle nur gegen eine Kontrollgruppe verglichen werden.

## 1. 7 Darstellung der Ergebnisse

Das Ziel ist ja, die *k* Gruppen paarweise zu vergleichen. Daraus resultieren *k(k-1)/2* Vergleiche, die zweckmäßigerweise in Form einer Matrix dargestellt werden, genau genommen nur in einer „halben“ Matrix, einer Dreiecksmatrix, da ja der Vergleich der Gruppen *i* mit *j* mit dem der Gruppen *j* mit *i* identisch ist. Eine solche Matrix ist z.B. unten beim SNK-Verfahren (Kapitel 4.3) skizziert.

Eine Variante dieser Darstellung wird u.a. bei SPSS praktiziert: Eine Tabelle, die als Zeilen die *k* Gruppen enthält, und in jeder Zeile dann die Ergebnisse dieser Gruppe mit allen anderen Gruppen wiedergegeben werden. Ein Beispiel ist in S-1 (Kapitel 16) zu finden.

Eine alternative Darstellung ist die der *homogenen Untergruppen*. Hier werden nicht die signifikanten Vergleiche ausgewiesen, sondern die nichtsignifikanten. Dazu werden die *k* Mittelwerte der Größe nach sortiert: $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \ldots \leq \bar{x}_{(k)}$ . (Hier wird mit $..._{(i)}$ der Index innerhalb der Reihenfolge angegeben.) Es werden dann die Folgen von Mittelwerten ermittelt, die sich nicht voneinander unterschieden, also gleich oder *homogen* sind. Nachfolgend ein Beispiel auf Basis des SNK-Tests (Student-Newman-Keuls), das mit SPSS erzeugt wurde:

| | | N | Untergruppe | |
|---|---|---|---|---|
| | | | 1 | 2 |
| Student-Newman-Keuls | 2 | 7 | 4,00 | |
| | 1 | 9 | 5,11 | |
| | 3 | 8 | 5,63 | 5,63 |
| | 4 | 9 | | 7,11 |
| | Sig. | | ,104 | ,063 |

Die ersten 4 Zeilen dieser Tabelle entsprechen den 4 zu vergleichenden Gruppen, die von oben nach unten bzgl. der Größe der Mittelwerte angeordnet sind. Die erste Spalte enthält die Gruppennummer, die zweite den Stichprobenumfang $n_i$. In den Spalten drei und vier werden nun zwei homogene Untergruppen ausgewiesen. Die erste umfasst die Gruppen 2, 1 und 3. Die zweite die Gruppen 3 und 4. Das heißt: Zum einen sind die Mittelwerte der Gruppen 1, 2 und 3 als gleich anzusehen. Zum anderen sind auch die Mittelwerte der Gruppen 3 und 4 gleich.

Möchte man nun schließen, welche Mittelwerte sich von welchen anderen unterscheiden, so sind dies alle Paarvergleiche bis auf die, die oben als gleich etmittelt worden sind. Da bleiben noch übrig: die Vergleiche 2 mit 4 sowie 1 mit 4.

Das Niveau, auf dem diese Vergleiche signifikant sind, entspricht dem vorgebenen α. Individuelle p-Werte für jeden Vergleich werden nicht ermittelt. Die letzte Zeile in o.a. Tabelle gibt allerdings einen p-Wert aus: das Sigifikanzniveau für den Test auf Gleichheit der jeweiligen homogenen Untergruppe. Das beträgt z.B. für die zweite Untergruppe p=0,063 und liegt nur knapp über dem α=0,05. D.h. es ist zu vermuten, dass bei etwas größeren Stichprobenumfängen die Mittelwerte der Gruppen 3 und 4 nicht mehr als gleich anzusehen gewesen wären.

## 1. 8 Hinweise auf Literatur

In Wikipedia sind die klassischen Verfahren dargestellt [104]. Ein umfassenden und insbesondere auch leicht verständlichen Überblick über die Methoden, deren Unterschiede und Eigenschaften geben Day & Quinn [101]. Eine umfassende Beschreibung der Verfahren (allerdings auf dem Stand von 1990) mit Angabe sämtlicher Formeln (allerdings in einer schlecht lesbaren Typographie) und umfassende Kriterien zur Auswahl der Verfahren bietet Wilcox [102]. Eine sehr gute Einführung in die Problematik, Beschreibung der bekanntesten Tests sowie deren Unterschiede bieten Rafter, Abell & Braselton [106], Gonzalez [103] und Tamhane [111] (für den Fall inhomogener Varianzen). Eine gute Erläuterung der Unterschiede ist auch bei Dallal [107] zu finden. Eine Übersicht der verschiedenen Methodengruppen gibt Shaffer [108]. Eine kompakte Übersicht der multiple Mittelwertvergleiche einschließlich Formeln bieten die SAS-Dokumentation [182] sowie die SPSS-Dokumentation [110], natürlich soweit die Verfahren in SAS bzw. SPSS verfügbar sind. Eine umfassende Literaturliste zu den Verfahren ist bei Rao & Swarupchand [109] zu finden.

Einen guten Überblick der α-Adjustierungen bietet das Benutzerhandbuch von SAS [123]. Einige neuere Verfahren sind auch bei Garcia et al [124] kurz beschrieben sowie bei Blakesley et al [125], die auch einen Vergleich der Methoden erörtern.

## 1. 9 Beispieldatensätze

Weitestgehend werden die Datensätze aus [99] verwendet, zumal die dort gefundenen Anova-Ergebnisse hier mittels multipler Mittelwertvergleiche weiter untersucht werden. Allerdings werden noch zusätzlich folgende Datensätze benutzt:

**Beispieldaten 9 (mydata9):**

10 Personen mussten unter 4 verschiedenen Bedingungen einen Reaktionstest absolvieren. Die Werte der abhängigen Variablen können zwischen 1 und 20 liegen.

| Versuchsbedingung | | | |
|---|---|---|---|
| V1 | V2 | V3 | V4 |
| 3 | 4 | 5 | 5 |
| 4 | 3 | 5 | 3 |
| 4 | 5 | 6 | 4 |
| 3 | 6 | 3 | 7 |
| 5 | 9 | 6 | 9 |
| 5 | 7 | 8 | 12 |
| 4 | 9 | 8 | 10 |
| 6 | 8 | 9 | 13 |
| 7 | 6 | 10 | 16 |
| 7 | 9 | 11 | 15 |

**Beispieldaten 10 (mydata10):**

Insgesamt 35 Personen in 6 Gruppen mussten unter verschiedenen Bedingungen einen Reaktionstest absolvieren. Die Werte der abhängigen Variablen können zwischen 1 und 20 liegen.

| Versuchsbedingung | | | | | |
|---|---|---|---|---|---|
| Gruppe 1 | Gruppe 2 | Gruppe 3 | Gruppe 4 | Gruppe 5 | Gruppe 6 |
| 10 | 12 | 8 | 14 | 16 | 13 |
| 11 | 12 | 10 | 14 | 16 | 15 |
| 11 | 13 | 10 | 15 | 18 | 15 |
| 13 | 15 | 11 | 16 | 19 | 16 |
| 14 | 15 | 12 | 18 | 20 | 18 |
| 14 | 16 | | 19 | 20 | |
| | | | | 20 | |

# 2. paarweise Vergleiche mit α-Adjustierungen

Ein einfacher Weg, das Problem der Verletzung des α-Risikos zu lösen, besteht darin, das α für die multiplen Vergleiche von vorneherein so klein zu wählen, dass nach der o.a. Formel das gewünschte α als Ergbenis herauskommt. Und wie in dem Zahlenbeispiel in Kapitel 1.6 demonstriert, je größer die Gruppenzahl, desto stärker muss das vorgegebene α verkleinert werden und damit wird es immer schwieriger, einen signifikanten Unterschied nachzuweisen. Ein kleiner Vorteil dieses Verfahrens gegenüber anderen liegt darin, dass zum einen keine zusätzlichen Voraussetzungen zu beachten sind und zum anderen das Verfahren im Zusammenhang mit beliebigen Tests angewandt werden kann, also z.B. t-Test, U-Test oder $\chi^2$-Test.

Generell muss allerdings gesagt werden, dass fast alle Methoden der α-Adjustierung, insbesondere die bekannteren, klassischen Verfahren, sich sehr konservativ verhalten. Dennoch wurden in den letzten Jahren Methoden entwickelt, die mit multiplen Mittelwertvergleichen nicht nur konkurrieren können, sondern manchmal auch überlegen sind, insbesondere den nichtparametrischen Vergleichen.

α-Adjustierungen werden auch zahlreich bei den multiplen Mittelwertvergleichen in den Kapiteln 4 bis 9 eingesetzt. Sie haben eine zentrale Bedeutung. Allgemein hat die Bedeutung der Adjustierungen in den letzten Jahren zugenommen, insbesondere durch die Gendatenanalyse (micro array data), bei denen mehrere Tausend Gene für zwei oder mehr Gruppen verglichen werden. Dem entsprechen ebenso viele Hypothesen, von denen es gilt, möglichst wenig falsche Alternativhypothesen anzunehmen. Allerdings hat man bei Vergleichen von Tausenden Variablen verstärkt das Problem, dass wegen deren Korrelationen die Hypothesen nicht mehr unabhängig voneinander sind. Dieses Problem ist bei den hier im Vordergrund stehenden paarweisen Vergleichen nicht gegeben. α-Adjustierungen, die Abhängigkeiten der Hypothesen berücksichtigen, z.B. die von *Benjamini & Yekutieli* oder einige von *Blanchard & Roquain*, sind für die hier im Vordergrund stehenden paarweisen Vergleiche weniger geeignet, da sie bei unabhängigen Hypothesen sehr konservativ reagieren.

In der englisch-sprachigen Literatur wird meistens von *p adjustment* anstatt von α *adjustment* gesprochen. Dabei wird dem Rechnung getragen, dass heutzutage vielfach die p-Werte interpretiert werden, ohne ein konkretes α vorgegeben zu haben. Denn als Ergebnis für einen Test ist ein p-Wert informativer als die Aussage, dass der Test auf einem vorgebenen α-Niveau signifikant ist oder nicht. Dementsprechend wird ein p-Wert in ein adjustiertes p´ umgerechnet.

## 2. 1 Verfahren von Bonferroni und Sidak

Das einfachste Verfahren ist das von *Bonferroni* [126]: Basierend auf der o.a. Wahrscheinlichkeit, dass bei *m* durchgeführten Tests $H_0$ fälschlich abgelehnt wird, ergibt sich die Gleichung

$$\alpha = 1 - ( 1 - \alpha' )^m$$

aus der das α' so bestimmt wird, dass das gewünschte α-Risiko eingehalten wird:

$$\alpha' = 1 - ( 1 - \alpha )^{1/m} \qquad (2\text{-}1)$$

Hierbei ist *m* die Anzahl der durchgeführten Vergleiche. α' lässt sich allerdings näherungsweise sehr gut und einfach errechnen als

$$\alpha' = \alpha /m \qquad (2\text{-}2)$$

Die Vorgehensweise ist dann:

- Man legt fest, welche Vergleiche vorgenommen werden sollen. Je weniger desto besser. Diese Anzahl sei *m*.
- Man bestimmt das $\alpha' = \alpha / m$ .
- Alle p-Werte werden mit $\alpha'$ anstatt $\alpha$ verglichen.

Verwendet man anstatt dieses näherungsweise errechneten $\alpha'$ das aus o.a. Gleichung 2-1 exakt errechnete $\alpha'$ , so resultiert daraus die *Sidak-Adjustierung*.

Oder in Gestalt der p-Wert-Adjustierung:

- *Bonferroni*:
  $p' = m \cdot p$ (streng genommen $p' = \min(1, m \cdot p)$ )
- *Sidak*:
  $p' = 1 - ( 1 - p )^m$

Diese Verfahren bleiben in der Praxis nur für den „Notfall“ vorbehalten, zumal andere bessere Verfahren, wie das von Li, ebenso einfach mit der Hand durchzuführen sind.

## 2. 2 schrittweise Verfahren

Eine Verbesserung, d.h. weniger konservatives Verhalten, bringen die schrittweisen Verfahren. Während bei den o.a. klassischen Adjustierungen von Bonferroni und Sidak alle Vergleiche mit demselben korrigierten $\alpha'$ durchgeführt werden, werden hier die Vergleiche in mehreren Schritten durchgeführt und in jedem Schritt die α-Korrektur angepasst, indem einige Hypothesen ausgeschlossen werden können und sich dadurch das *m* verkleinert, also das $\alpha'$ vergrößert. Allerdings sind einige wenig transparent und sind nicht mehr mit der Hand zu rechnen. Dazu zählen insbesondere die Adjustierungen von *Hommel*, *Rom* und *Shaffer*, die allerdings diejenigen mit der meisten Power sind.

Bei den *step-down*-Verfahren wird üblicherweise mit dem Vergleich des größten mit dem kleinsten Mittelwert begonnen, bzw. dem Vergleich, der den niedrigsten p-Wert erzeugt. In den nachfolgenden Schritten reduziert sich die Anzahl der Vergleiche und damit die Anzahl der möglicherweise falsch abgelehnten Nullhypothesen, da sich einige aus den bisherigen Ergebnissen ableiten lassen. Damit kann auch für die ausstehenden Vergleiche z.B. ein $\alpha' = \alpha / m'$ mit einem deutlich kleineren $m' < m$ gewählt werden. Die Prozedur endet, wenn ein Vergleich nicht mehr signifikant ist.

Bei den *step-up*-Verfahren wird mit dem Vergleich begonnen, der den größten p-Wert erbrachte, der in der Regel einem nichtsignifikanten Vergleich entspricht. Bei den nachfolgenden Vergleichen zu den nächst größeren p-Werten wird z.B. ein $\alpha' = \alpha / m'$ mit einem deutlich größeren $m'$ gewählt. Die Prozedur endet, wenn ein Vergleich signifikant ist.

Die Verfahren sind u.a. bei Garcia et al [124] kurz beschrieben.

### 2. 2. 1 step-down-Verfahren: Holm und Hochberg

Eine marginale Verbesserung gegenüber der Bonferroni-Adjustierung bieten die Verfahren von *Holm* [127] und *Hochberg* [128]. Hierbei werden die Vergleiche hinsichtlich des p-Wertes aufsteigend sortiert, so dass $p_1 \le \ldots \le p_i \le \ldots \le p_m$ . Die entsprechenden Null-Hypothesen werden mit $H_1$,..., $H_m$ bezeichnet. Beide vergleichen jeweils $p_i$ mit $\alpha/(m-i+1)$.

Holm geht der Reihe nach vor: $p_1$ wird nun mit $\alpha/m$ verglichen. Ist dieser Vergleich ($H_1$) signifikant, kann $p_2$ verglichen werden und zwar mit $\alpha/(m-1)$ . Ist auch dieser Vergleich ($H_2$) signifikant, darf $p_3$ verglichen werden und zwar mit $\alpha/(m-2)$ usw. Oder in Gestalt adjustierter p-Werte:

$$p_i' = min(1,(m-i+1)p_i) .$$

In diesem Fall muss allerdings darauf geachtet werden, dass auch für die $p_i'$ gilt: $p_1' \le \ldots \le p_i' \le \ldots \le p_m'$ . Sind für $j<i$ irgendwelche $p_j>p_i$, so muss $p_i=p_j$ gesetzt werden.

Da hier mit $\alpha' = \alpha/m'$ im Wesentlichen auch eine Art Bonferroni-Adjustierung benutzt wird, heißt dieses Varfahren auch häufig *Bonferroni-step-down*.

Bei Hochberg wird das größte $i$ ($1 \le i \le m$) gesucht, so dass noch $p_i \le \alpha/(m-i+1)$ . Alle Hypothesen $H_1,..,H_i$ werden dann abgelehnt und alle $H_{i+1},..,H_m$ angenommen. Das adjustierte p' ist dasselbe wie bei Holm.

Es gibt zwei Varianten dieser Methoden: *Holland* bietet eine Variante des Verfahrens von Holm: Anstatt des Vergleichs von $p_i$ mit $\alpha/(m-i+1)$ wird $p_i$ mit $1-(1-\alpha)^{(m-i+1)}$ verglichen. Dieser Wert ergibt sich, wenn man für $\alpha'$ die Formel (2-2) durch (2-1) ersetzt. *Finner* bietet eine weitere Variante: Anstatt des Vergleichs von $p_i$ mit $\alpha/(m-i+1)$ wird $p_i$ mit $1-(1-\alpha)^{m/i}$ verglichen. Diese Verfahren werden auch mit *Sidak step-down* bezeichnet.

### 2. 2. 2 Benjamini & Hochberg

Eine deutliche Verbesserung stellt dagegen das weniger bekannte Verfahren von *Benjamini & Hochberg* [128] dar (vgl. auch Wikipedia [121]). Dieses basiert auf einem anderen Ansatz: Die Kontrolle der falschen Signifikanzen (*false discovery rate, FDR*) anstatt der Kontrolle des α (FWER) (vgl. Kapitel 1.5). Das Verfahren verläuft ähnlich dem o.a. von Hochberg. Lediglich die Vergleichswerte für die $p_i$ sind hier andere: $p_i \le (i \cdot \alpha)/m$ , die, wie man leicht sieht, durch den Faktor $i$ im Zähler deutlich größer sind. Simulationen haben gezeigt, dass mit dieser Methode im Schnitt 25% mehr Signifikanzen nachgewiesen werden als bei den anderen o.a. α–Adjustierungen. Die adjustierten p-Werte errechnen sich als

$$p_i' = (m/i)p_i$$

Eine Variante dieses Verfahrens für den Fall, dass die Tests nicht unabhängig voneinander sind, etwa für den Vergleich von zwei Gruppen für eine große Anzahl von Variablen, haben Benjamini, Krieger und Yekutieli erarbeitet, das bei Benjamini [128] und in Wikipedia [121] beschrieben ist.

Das Verfahren von Simes (vgl. Kapitel 7.6) ist weitgehend mit dem hier beschriebenen von Benjamini & Hochberg identisch.

### 2. 2. 3 step-up-Verfahren: Hochberg, Hommel und Rom

*Hochbergs* step-up-Verfahren ähnelt dem o.a. step-down-Verfahren von Holm. Hierbei werden die Vergleiche hinsichtlich des p-Wertes wieder aufsteigend sortiert, so dass $p_1 \le \ldots \le p_i \le \ldots \le p_m$ . Die entsprechenden Null-Hypothesen werden mit $H_1,..., H_m$ bezeichnet. Zunächst wird der größte p-Wert $p_m$ mit α verglichen, dann $p_{m-1}$ mit $\alpha/2$, $p_{m-2}$ mit $\alpha/3$ usw. Das Verfahren endet, wenn ein Vergleich signifikant ist. Die entsprechende Hypothese wird dann abgelehnt, ebenso alle Hypothesen zu allen kleineren $p_i$ . Oder in Gestalt adjustierter p-Werte:

$$p_i' = min(1,(m-i+1)p_i) .$$

In diesem Fall muss allerdings wieder darauf geachtet werden, dass auch für die $p_i'$ gilt:

$p_1' \le \ldots \le p_i' \le \ldots \le p_m'$ .

Eine Verbesserung dieses Verfahrens von Hochberg bietet *Rom* [129]. Die $\alpha$ , mit denen die p-Werte zu vergleichen sind, werden allerdings relativ aufwändig rekursiv berechnet. Zunächst wird wie oben $p_m$ mit $\alpha_m=\alpha$ und $p_{m-1}$ mit $\alpha_{m-1}=\alpha/2$ verglichen. Die nächsten $\alpha_i$ (z.B. $\alpha_{m-2}$ für $i=2$) errechen sich wie folgt:

$$\alpha_{m-i} = \left(\sum_{j=1}^{i-1} \alpha^j - \sum_{j=1}^{i-2} \binom{i}{m} \alpha_{m-j}^{i-j}\right) / i$$

Das Verfahren endet, wenn ein Vergleich signifikant ist. Die entsprechende Hypothese wird dann abgelehnt, ebenso alle Hypothesen zu allen kleineren $p_i$ .

Das Verfahren von *Hommel* ähnelt dem im vorigen Kapitel vorgestellten Verfahren von Hochberg. Es wird das größte $j$ ($1 \le j \le m$) gesucht, so dass noch für alle $i \le j$ gilt $p_{m-j+i} > i\alpha/j$ . Alle Hypothesen zu $p_i$ mit $p_i < \alpha/j$ werden dann verworfen. Existiert kein solches $j$, werden alle Hypothesen verworfen. Dieses Verfahren gilt als eines der stärksten.

### 2. 2. 4 Jianjun Li

Vergleichsweise einfach ist dagegen das Verfahren von *Jianjun Li* [130]. Im ersten Schritt wird $H_m$ geprüft: Ist $p_m \le \alpha$ , werden alle Hypothesen abgelehnt. Andernfalls werden die Hypothesen $H_i$ abgelehnt, für die gilt: $p_i \le \alpha(1-p_m)/(1-\alpha)$ . Li hat selbst theoretisch wie auch durch Simulationen gezeigt, dass sein Verfahren denen von *Hochberg*, *Hommel* und *Rom* leicht überlegen ist. Für die adjustierten p-Werte gilt:

$$p_i' = \frac{p_i}{p_i + 1 - p_{(m)}}$$

Zum einen ist die Berechnung der $\alpha'$ so einfach, dass Lis Methode mit der Bonferroni-Adjustierung konkurrieren kann. Zum anderen hat das Verfahren eine sehr hohe Power, insbesondere wenn $p_m$ nahe bei $\alpha$ liegt. In dem Fall liegt $\alpha'$ nahe bei $\alpha$ .

### 2. 2. 5 Shaffers S1 und S2

Eine weitere Verbesserung bringen die $\alpha$-Adjustierungen von Juliet Shaffer. Hier kann nur eine grobe Idee der Verfahren vermittelt werden. Die Handhabung der Verfahren selbst ist so kompliziert, dass sie ohnehin nur mit Computerprogrammen angewandt werden können.

Das S1-Verfahren ist ein schrittweises Verfahren, bei dem in jedem Schritt Teilmengen von Hypothesen $H_{i1}+H_{i2}+...$ geprüft und weiter verfeinert werden. In jedem Schritt wird die maximale Anzahl $m'$der Nullhypothesen ermittelt, die wahr sein könnten (und irrtümlich abgelehnt werden könnten), wenn im vorigen Schritt die Nullhypothese abgelehnt worden war. (Wenn im vorigen Schritt die Nullhypothese angenommen worden war, erübrigen sich weitere Schritte der Verfeinerung.) Dieses $m'$ wird dann für die $\alpha$-Adjustierung dieses Schrittes verwandt.

Bei dem S2-Verfahren wird genauer berücksichtigt, welche Hypothese im vorangegangenen Schritt verworfen worden war. Dazu werden die Mittelwerte in Partitionen aufgeteilt, die sich jeweils signifikant unterscheiden, und untersucht, wieviele nicht redundante Hypothesen in diesem Schritt überhaupt aufgestellt werden können. Hierdurch wird zwar die Anzahl der Nullhypothesen, die wahr sein könnten, reduziert, das Verfahren selbst wird noch wesentlich komplizierter und für größere Gruppenzahl $k$ kaum mehr praktizierbar.

Bei Shaffer werden i.a. transitive Schlüsse gezogen. Z.B. wenn $\mu_1 \le \mu_2 \le \mu_3$ und es wurde bereits $\mu_1 < \mu_2$ bewiesen, dann kann normalerweise daraus auch $\mu_1 < \mu_3$ geschlossen werden.

Allerdings nur im Fall gleicher Varianzen. Doch hat Donoghe [180] einen Algorithmus für das S2-Verfahren von Shaffer entwickelt, der auch für den Fall inhomogener Varianzen anwendbar ist. (Der Algorithmus ist zwar in [180] aufgelistet, aber dennoch schwer verständluch und ohne Computer nicht durchführbar.)

## 2. 3 Auf Resampling basierende Verfahren

Klassischerweise werden Tests auf der Basis eines Modells mittels Wahrscheinlichkeitsrechnung entwickelt. Beim *Resampling* (vgl. Wikipedia [131]) wird dagegen die vorliegende Stichprobe wie eine Grundgesamtheit angesehen, aus der sehr viele, z.B. 1000 oder 100.000 Unterstichproben gezogen werden, um für diese die Fragestellung zu untersuchen oder den Test durchzuführen. Hieraus wird dann eine empirische Verteilung, z.B. der Testgröße, ermittelt, so dass daraus eine Wahrscheinlichkeitsaussage möglich ist. Letztlich kann man das Resampling als eine nichtparametrische Simulation auffassen. Während bei der Simulation Stichproben aus parametrischen Verteilungen, etwa der Normalverteilung, erzeugt werden, um daraus z.B. Quartilswerte zu ermitteln, geschieht dies beim Resampling über Unterstichproben aus den vorliegenden Daten.

Ein Vorteil dieser Vorgehensweise ist u.a. der, dass keine Voraussetzungen wie z.B. Varianzhomogenität oder Unabhängigkeit eingehalten werden müssen, weil bei der Ermittlung des Testergebnisses ja alle Eigenschaften aus der vorliegenden Stichprobe berücksichtigt werden. Ein Nachteil ist der erhebliche Rechenaufwand, weil die Berechnungen ja nicht nur einmal, sondern tausende Male durchgeführt werden müssen. Darüber hinaus sollten die Stichproben nicht zu klein sein - $n>100$ ist schon wünschenswert - damit genügend verschiedene Unterstichproben ausgewählt werden können.

Wegen des o.a. Vorteils werden diese Verfahren hauptsächlich beim Gruppenvergleich von mehreren Merkmalen angewandt. So insbesondere bei der Gendatenanalyse, bei der häufig tausende Variablen gleichzeitig verglichen werden müssen und daher eine gute α-Adjustierung erforderlich, um die „Spreu vom Weizen zu trennen“.

### 2. 3. 1 Westfalls free step-down-Verfahren

Das *free step-down-Verfahren* von *Westfall* ähnelt vom Ansatz her dem o.a. Verfahren von Holm. Im ersten Schritt wird der kleinste p-Wert $p_1$ der $m$ Vergleiche bzgl. $m$ Tests adjustiert, aber nicht mit $\alpha/m$ verglichen, sondern mit einem Quantil der Verteilung p-Werte, die durch Resampling ermittelt wird. Im nächten Schritt wird $p_2$ bzgl. $m-1$ Tests adjustiert usw.

Das free step-down-Verfahren ist ausführlich von Reilly [132] beschrieben und mit R-Beispielen ergänzt worden.

### 2. 3. 2 min.P- und max.T-Verfahren

Das min.P- und max.T-Verfahren von Westfall & Young [134] sind zunächst einmal ein allgemeiner Rahmen zur Adjustierung. So kann z.B. das o.a. Verfahren von Holm auch als min.p-Variante dargestellt werden. Doch in der Regel werden diese Verfahren mit Resampling assoziiert. Aber auch davon gibt es mehrere Varianten.

Kurz die Idee des min.P-Verfahrens. Anstatt z.B. den kleinsten p-Wert $p_1$ der $m$ Vergleiche mit $\alpha/m$ zu vergleichen, wird dieser mit dem entsprechenden Quantil der Verteilung der min(p)-Werte auf Basis des Resampling verglichen. D.h. für alle Resampling-Unterstichproben werden die $m$ Vergleiche durchgeführt und der jeweils kleinste p-Wert ermittelt. Hieraus resultiert dann

eine empirische Verteilung der min(p)-Werte, woraus das α-Quantil zum Vergleich für $p_1$ ermittelt werden kann. In weiteren Schritten werden dann analog $p_2$, $p_3$,... verglichen.

Das max.T-Verfahren verläuft fast identisch. Nur, dass anstatt der p-Werte die Teststatistiken, hier mit T bezeichnet, als Kriterium verwendet werden. Beide Verfahren sind ohnehin identisch, wenn die sog. *subset pivotaliy* erfüllt ist (vgl. Kapitel 2.6.2).

Die Algorithmen zu diesen Verfahren und einigen Varianten sind beschrieben von Ge et al [135].

## 2. 4 Adaptive Verfahren

Bei *adaptiven Verfahren* geht es um die Einhaltung der FDR (vgl. Kapitel 1.5). Sei $\pi_0$ der Anteil der wahren Nullhypothesen an allen Hypothesen. Dieser Anteil ist natürlich unbekannt. Die adaptiven Verfahren benötigen eine Schätzung für diese Größe. Diese kann entweder vorgegeben werden, oder es können Verfahren zur Schätzung von $\pi_0$ verwendet werden. Letztendlich sind die adaptiven Verfahren eine Weiterentwicklung des o.a. Verfahrens von *Benjamini & Hochberg*.

Es gibt mehrere Varianten von adaptiven Verfahren:

- *plugin-Verfahren*: Bei diesen wird eine Anfangsschätzung von $\pi_0$ in bekannte Verfahren (z.B. von Benjamini & Hochberg) eingearbeitet. Das bekannteste ist das von *Storey.*
- *two-step-Verfahren*: Bei diesen werden zunächst eine Teilmenge der Hypothesen betrachtet, umd damit $\pi_0$ zu schätzen. Im zweiten Schritt werden dann alle Hypothesen überprüft.
- *one-stage-Verfahren*: Bei diesen wird für ein vorgebenes $\pi_0$ ein step-down- oder step-up-Verfahren eingesetzt.

Eine gute Übersicht bieten *Blanchard & Roquain* [136], die selbst auch eine Reihe von adaptiven Verfahren beigesteuert haben. So u.a. ein relativ einfaches *one-stage-Verfahren*, das eine kleine Verbesserung gegenüber dem klassischen von Benjamini & Hochberg bietet. Der Ablauf ist identisch, jedoch sind die Vergleichswerte

$$p_i \leq \alpha \cdot min\left((1-\alpha)\frac{i}{m-i+1}, 1\right)$$

## 2. 5 Andere Verfahren

### 2. 5. 1 Sidak-Variante von Tukey-Ciminera-Heyse

Für den Fall, dass die *m* Tests nicht unabhängig voneinander sind, z.B. bei korrelierenden Variablen, etwa beim Vergleich der Mittelwerte von abhängigen Stichproben, haben *Tukey, Ciminera* und *Heyse* eine Variante der Basiskorrektur von Sidak vorgeschlagen:

$$p_i' = 1-(1-p_i)^{\sqrt{m}}$$

Zu berücksichtigen ist allerdings, dass bei unabhängigen Variablen, z.B. beim multiplen Mittelwertvergleich bei unabhängigen Stichproben, das α-Risiko nicht mehr strikt eingehalten wird.

### 2. 5. 2 Weitere Verfahren

Eine gute Zusammenstellung weiterer Adjustierungsverfahren ist in der Dokumentation des

Statistiksystems SAS zu finden [123].

## 2. 6 Grundlegende Methoden

Zum Abschluss zwei Verfahren, die eigentlich keine α-Adjustierung beinhalten, aber Grundlagen für andere Adjustierungsverfahren sind.

### 2. 6. 1 Simes-Verfahren

Das Verfahren von *Simes* [137] prüft die Fragestellung: Wie kann aus den einzelnen $p_i$-Werten der *m* Vergleiche geschlossen werden, ob die globale Hypothese gleicher Mittelwerte $H_0$ falsch oder richtig ist. Andere Adjustierungen, z.B. die o.a. von *Hommel,* bedienen sich allerdings des Simes-Verfahrens, so dass es hier angeführt wird.

Sind die p-Werte der *m* Paarvergleiche wieder der Größe nach sortiert: $p_1 \le \ldots \le p_i \le \ldots \le p_m$. $H_0$ wird abgelehnt, wenn $p_i \le (i/m)\alpha$ für irgendein $1 \le i \le m$. Das Interessante daran ist, dass diese Bedingung nicht ausschließlich für *i=1* erfüllt sein muss, sondern es genügt auch ein anderes *i*.

### 2. 6. 2 Closure-Prinzip

Das *Closure*-Prinzip oder das *closed testing* ist ebenfalls ein Hilfsmittel, dessen sich zahlreiche neuere Adjustierungen bedienen. Es lässt sich am einfachsten anhand eines Beispiels erläutern.

Angenommen, es sollen 3 Hypothesen $H_1$, $H_2$ und $H_3$ geprüft werden, z.B. für eine Variable der Vergleich von 3 Experimentalgruppen mit einer Kontrollgruppe oder für 3 Variablen der Vergleich einer Experimentalgruppe mit einer Kontrollgruppe. Das Prinzip umfasst 4 Schritte:

1. Jede der 3 Hypothesen wird mit einem passenden Test geprüft.
2. Es wird die *Closure* (die abgeschlossene Hülle) der 3 Hypothesen ermittelt. Diese umfasst deren Schnittmengen, die mit $H_{12}$ , $H_{13}$ , $H_{23}$ und $H_{123}$ bezeichnet werden. So bezeichnet z.B. $H_{12}$ die Hypothese "$H_1$ und $H_2$ sind wahr".
3. Jede dieser Schnittmengen wird nun mit einem passenden Test geprüft, z.B. mit dem F-Test der Anova (im univariaten Fall) oder einer Manova (im multivariaten Fall)
4. Eine Hypothese $H_i$ wird unter Einhaltung der FWER (familywise error rate) verworfen, wenn die folgenden beiden Bedingungen erfüllt sind:
   - $H_i$ wurde im ersten Schritt abgelehnt.
   - Alle Schnittmengen, die $H_i$ enthalten, wurden im dritten Schritt abgelehnt.

Der wesentliche Vorteil des closure-Prinzips liegt darin, dass die Anzahl der Vergleiche reduziert werden kann. Auf der anderen Seite gibt es bei *m* Hypothesen $2^m$-1 solcher Schnittmengen. Unter einigen Annahmen lassen sich diese allerdings deutlich reduzieren. Eine davon ist die sog. *subset pivotality*. Diese besagt, dass die Verteilung einer Teststatistik T dieselbe ist, egal ob alle Nullhypothesen (einer Familie) wahr sind oder nur eine Teilmenge.

Mit Hilfe des closure-Prinzips lassen sich relativ einfache schrittweise Verfahren konstruieren, ähnlich den Verfahren von Hochberg oder schrittweisen von Dunnett. Liegen z.B. *m* Nullhypothesen vor und sind $t_{(1)} \ge t_{(2)} \ge \ldots \ge t_{(m)}$ die der Größe nach geordneten Testgrößen zu deren Test (also Realisationen entsprechender Zufallsvariablen $T_{(1)}$, $T_{(2)}$,...,$T_{(m)}$). Bezeichnen

$H_{(1)}$, $H_{(2)}$,..., $H_{(m)}$ die dazugehörigen Hypothesen. Dann lässt sich folgendes schrittweises Verfahren ableiten:

- Schritt 1:
  $H_{(1)}$ wird abgelehnt, wenn $P(max(T_1, \ldots, T_m) \geq t_{(1)}) \leq \alpha$ ,
  andernfalls werden alle Nullhypothesen angenommen.
- Schritt i (i=2,...,m-1):
  $H_{(i)}$ wird abgelehnt, wenn $P(max(T_i, \ldots, T_m) \geq t_{(i)}) \leq \alpha$ ,
  andernfalls werden die Nullhypothesen $H_{(i)}$,..., $H_{(m)}$ angenommen.
- Schritt m:
  $H_{(m)}$ wird abgelehnt, wenn $P(T_m \geq t_{(m)}) \leq \alpha$ ,
  andernfalls wird $H_{(m)}$ angenommen.

Die Wahrscheinlichkeiten *P* werden z.B. über die maximum-Verteilung (vgl. Kapitel 1.4) berechnet oder über Resampling (vgl. Kapitel 2.3) ermittelt.

## 2. 7 Adjustierungen in R und SPSS

SPSS bietet α-Adjustierungen lediglich im Zusammeng mit Mittelwertvergleichen an:

- in `oneway` und `glm` Fishers LSD mit den Adjustierungen von Bonferroni und Sidak,
- in `nptests` die Adjustierung von Bonferroni

R bietet α-Adjustierungen sowohl unabhängig von irgendwelchen Vergleichsverfahren an:

- `p.adjust` die Standard-Adjustierungen (Bonferroni, Holm, Hochberg und Bejamini),
- `mutoss` neben den o.a. Standard-Adjustierungen auch speziellere Adjustierungen (u.a. Rom, adaptive Verfahren von Storey sowie von Blanchard & Rouquain sowie eine Reihe weiterer Verfahren von Benjamini)
- `multcomp` die Verfahren von Westfall und Shaffer,

sowie α-Adjustierungen im Zusammenhang mit multiplen Mittelwertvergleichen, die vielfach das o.a. `p.adjust` benutzen und somit die darin verfügbaren Verfahren, so u.a.

- `pairwise.t.test` für paarweise t-Tests ,
- `pairwise.wilcox.test` für paarweise Rangsummen-Tests,
- `agricolae` für paarweise nichtparametrische Vergleiche.

# 3. Kontraste und Kodierungen

## 3. 1 Grundlagen

Vielfach existieren bei der Varianzanalyse eines Merkmals zusätzlich zur globalen Hypothese gleicher Mittelwerte noch spezielle Hypothesen. Liegen z.B. 3 Gruppen vor, etwa eine Kontrollgruppe K sowie 2 Experimentalgruppen A und B, so könnten diese lauten: Vergleich der Mittelwerte von K gegen A sowie K gegen B. Solche Hypothesen müssen allerdings bereits *vor* der Untersuchung festliegen. Solche speziellen Vergleiche heißen *apriori-Vergleiche* oder *Kontraste*. Hierbei können nicht nur jeweils die Mittelwerte von zwei Gruppen verglichen werden, sondern allgemein eine Linearkombination der Mittelwerte auf den Wert 0. Bei o.a. Beispiel etwa den Mittelwert von K gegen den Durchschnitt der Mittelwerte von A und B, d.h. die beiden Experimentalgruppen unterscheiden sich „im Schnitt“ von der Kontrollgruppe hinsichtlich der Mittelwerte. Die Linearkombination ist dann $1*\mu_K - 0.5*(\mu_A + \mu_B)$. Theoretisch können sogar bei der Zusammenfassung von Gruppen gewichtete Mittel gebildet werden, etwa $(0.333*\mu_A + 0.667*\mu_B)$, wenn etwa die B-Gruppe doppelt so stark berücksichtigt werden soll wie die A-Gruppe.

Hat ein Faktor *k* Gruppen (Schichten), so ist ein Kontrast C über *k* Koeffizienten $c_j$ definiert:

$$C = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k \qquad (3-1)$$

wobei die Nebenbedingung $c_1+c_2+..+c_k = 0$ eingehalten werden muss. Diese Summe C wird dann auf den Wert 0 getestet. Im parametrischen Fall errechnet sich die Testgröße dann als

$$SS_C = \frac{(c_1\bar{x}_1 + c_2\bar{x}_2 + \ldots + c_k\bar{x}_k)^2}{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \ldots + \frac{c_k^2}{n_k}} \qquad (3-2)$$

und entspricht dem Anteil der Streuung $SS_{Effekt}$, der durch diesen Kontrast erklärt wird. Somit lässt sich diese Streuung $SS_C$ analog mit dem F-Test auf Signifikanz überprüfen :

$$F = \frac{SS_C}{MS_{Fehler}} \qquad (3-3)$$

wobei dieser F-Wert 1 Zähler-Fg hat und Nenner-Fg dem Test von $SS_{Effekt}$ zu entnehmen sind.

Es gibt aber noch eine andere, in R bevorzugte, Darstellung dieses Tests, und zwar mittels eines t-Tests, wobei in Erinnerung gerufen wird, dass allgemein $t_n = \sqrt{F_{1,n}}$ gilt:

$$t = \frac{C}{s_e} = \sqrt{F}$$

wobei *C* der o.a. Kontrastschätzer und $s_e$ der Standardfehler (des Kontrastschätzers) ist.

Es sei noch erwähnt, dass die Skalierung der $c_j$ ohne Bedeutung ist, d.h. Kontraste $c'_j = a \cdot c_j$ ergeben dasselbe Resultat wir die Kontraste $c_j$ .

In der Regel hat der Untersucher mehrere Hypothesen, aus denen dann mehrere Kontraste resultieren. Hierfür gelten dann folgende Regeln bzw. Eigenschaften:

- Es dürfen nur (*k-1*) Kontraste getestet werden.
- Zwei Kontraste $C_1$ mit Koeffizienten $c_{11}+c_{12}+..+c_{1k}$ und $C_2$ mit Koeffizienten $c_{21}+c_{22}+..+c_{2k}$ heißen *orthogonal*, d.h. sind unabhängig voneinander, wenn die folgende

Bedingung erfüllt ist:

$$\frac{c_{11}c_{21}}{n_1} + \frac{c_{12}c_{22}}{n_2} + \ldots + \frac{c_{1k}c_{2k}}{n_k} = 0$$

- Eine Menge von Kontrasten heißt *orthogonal*, wenn alle Paare orthogonal sind.
- Werden (*k-1*) orthogonale Kontraste $C_1$, $C_2$,..,$C_k$ mit Streuungen $SS_{C1}$ , $SS_{C2}$ ,.., $SS_{C(k-1)}$ getestet, dann gilt $SS_{C1} + SS_{C2} +..+ SS_{C(k-1)} = SS_{Effekt}$ , d.h. die gesamte durch den Faktor erklärte Streuung lässt sich in (*k-1*) einzeln erklärbare Streuungen unterteilen.

Sind die zu untersuchenden Kontraste nicht orthogonal oder sollen mehr als (*k-1*) Kontraste geprüft werden, so sind die einzelnen Testergebnisse nicht mehr unabhängig voneinander. In solchen Fällen ist eine α-Korrektur (s. Kapitel 2) vorzunehmen. Speziell hierfür ist u.a. das Verfahren von *Dunn & Bonferroni* (s. Kapitel 7.1) konzipiert.

**Beispiel:**

Für die o.a. Situation eines Faktors mit den Gruppen K, A und B werden 2 Kontraste definiert: K-A sowie K-B. Daraus resultieren folgende Koeffizienten $c_j$ :

| | Kontraste | |
|---|---|---|
| Gruppe | $C_1$ | $C_2$ |
| K | 1 | 1 |
| A | - 1 | 0 |
| B | 0 | - 1 |

Diese beiden Kontraste sind nicht orthogonal, denn $1 \cdot 1 + (-1) \cdot 0 + 0 \cdot (-1) = 1$.

Wird dagegen zum einen die Kontrollgruppe K gegen das Mittel von A und B verglichen und zum anderen die beiden Experimentalgruppen A und B gegeneinander, dann resultieren daraus die Koeffizienten $c_j$ :

| | Kontraste | |
|---|---|---|
| Gruppe | $C_1$ | $C_2$ |
| K | 2 | 0 |
| A | - 1 | 1 |
| B | - 1 | - 1 |

Diese beiden Kontraste sind orthogonal, denn $2 \cdot 0 + (-1) \cdot 1 + (-1) \cdot (-1) = 0$.

Die Kontraste oder Kodierungen haben auch eine andere Funktion: Bei der Regression müssen Prädiktoren mit nominalem Skalenniveau dichotomisiert werden. Die „naive“ Art, ein nominales Merkmal *f* mit k Ausprägungen in mehrere dichotome $d_1$,.., $d_k$ zu transformieren, ist normalerweise so, dass $d_j$ genau dann den Wert 1 hat, wenn *f* den Wert j hat, und sonst 0. Da von diesen *k* Variablen zwangsläufig eine redundant ist - jede beliebige von diesen lässt sich aus den übrigen errechnen, z.B. $d_k = 1-d_1-d_2-...-d_{k-1}$, muss eine weggelassen werden. Diese Kodierung, das *dummy coding*, ist nicht die einzige Möglichkeit, ein nominales Merkmal zu transformieren. Nachfolgend werden die Standardmethoden für die Kodierung und Kontrastbildung vorgestellt.

## 3. 2 Standard-Kontraste

Prinzipiell kann der Benutzer natürlich individuelle Kontraste festlegen, was sowohl in R als auch in SPSS mit ein wenig Aufwand verbunden ist. Es gibt aber eine Reihe von „Standard"-Kontrasten, die für einen Faktor vereinbart werden können. Allerdings ist die Namensgebung nicht einheitlich. Hierbei sind Kontraste und Kodierungen (nominaler Variablen) zu unterscheiden. Bei Kontrasten muss die Nebenbedingung $c_1+c_2+..+c_k = 0$ eingehalten werden.

**Dummy Coding / Indikator / Einfach bzw. Simple (SPSS)/ contr.treatment (R)**

Statistisch werden alle Gruppen gegen eine vorgegebene, üblicherweise die erste oder letzte, verglichen, nämlich die, die bei den oben erwähnten $d_j$ nicht repräsentiert ist. Die „Referenzgruppe" kann sowohl bei R als auch bei SPSS festgelegt werden. Dies wird angewandt, wenn eine Gruppe die Vergleichsgruppe ist, meist die sog. *Kontrollgruppe*. Anzumerken ist, dass bei SPSS die Koeffizienten dieselben sind, wie beim Effekt-Kodierung bei R, aber die Ergebnisse denen eines Vergleichs mit einer vorgegebenen Gruppe entsprechen:

| | Kontraste R | | | | Kontraste SPSS | | | |
|---|---|---|---|---|---|---|---|---|
| Gruppe | 1 | 2 | ... | (k-1) | 1 | 2 | ... | (k-1) |
| 1 | 1 | 0 | | 0 | 1 | 0 | | 0 |
| 2 | 0 | 1 | | 0 | 0 | 1 | | 0 |
| ... | 0 | 0 | | | | | | |
| k-1 | 0 | 0 | | 1 | 0 | 0 | | 1 |
| k | 0 | 0 | | 0 | - 1 | - 1 | | - 1 |

**Effekt-Kodierung / Abweichung bzw. Deviation (SPSS) / contr.sum (R)**

Dies sind orthogonale Kontraste, die letztlich der Varianzanalyse zugrunde liegen. Durch diese werden nämlich die Abweichungen vom Gesamtmittelwert getestet. Da nur (*k-1*) Vergleiche erlaubt sind, muss der Test für eine Gruppe entfallen. Dies ist üblicherweise (in R und SPSS) die letzte Gruppe. Die Koeffizienten:

| | Kontraste R | | | | Kontraste SPSS | | | |
|---|---|---|---|---|---|---|---|---|
| Gruppe | 1 | 2 | ... | (k-1) | 1 | 2 | ... | (k-1) |
| 1 | 1 | 0 | | 0 | (k-1)/k | - 1/k | | - 1/k |
| 2 | 0 | 1 | | 0 | - 1/k | (k-1)/k | | - 1/k |
| ... | 0 | 0 | | | | | | |
| k-1 | 0 | 0 | | 1 | - 1/k | - 1/k | | (k-1)/k |
| k | -1 | -1 | | -1 | - 1/k | - 1/k | | - 1/k |

**Helmert-Kodierung / Differenz bzw. Difference (SPSS) / contr.helmert (R)**

Bei dieser Bildung von orthogonalen Kontrasten werden sukzessive aufeinander folgende Gruppen miteinander vergleichen: 1-2, (1,2)-3, (1,2,3)-4 usw. wobei mit (.,.) der Mittelwert der entsprechenden Gruppen bezeichnet wird.

| | Kontraste R und SPSS | | | |
|---|---|---|---|---|
| Gruppe | 1 | 2 | ... | (k-1) |
| 1 | - 1 | - 1/2 | | - 1/(k-1) |
| 2 | 1 | - 1/2 | | - 1/(k-1) |
| ... | 0 | 1 | | |
| k-1 | 0 | 0 | | - 1/(k-1) |
| k | 0 | 0 | | 1 |

**umgekehrte Helmert-Kodierung / Helmert (SPSS)**

Bei dieser Bildung von orthogonalen Kontrasten werden sukzessive die erste gegen alle folgenden Gruppen miteinander vergleichen, die zweite gegen alle folgenden usw. (Diese Kontraste sind in R nicht verfügbar.)

| | Kontraste SPSS | | | |
|---|---|---|---|---|
| Gruppe | 1 | 2 | ... | (k-1) |
| 1 | 1 | 0 | | 0 |
| 2 | - 1/(k-1) | 1 | | 0 |
| ... | - 1/(k-1) | - 1/(k-2) | | |
| k-1 | - 1/(k-1) | - 1/(k-2) | | 1 |
| k | - 1/(k-1) | - 1/(k-2) | | - 1 |

**Wiederholt bzw. Repeated (SPSS) / sequen (R)**

Bei dieser Kodierung werden sukzessive zwei aufeinander folgende Gruppen miteinander vergleichen: 1-2, 2-3, 3-4 usw. Diese werden sinnvollerweise bei Messwiederholungsfaktoren eingesetzt. (Diese Kontraste sind in R nicht universell verfügbar.)

| | Kontraste SPSS | | | |
|---|---|---|---|---|
| Gruppe | 1 | 2 | ... | (k-1) |
| 1 | 1 | 0 | | 0 |
| 2 | - 1 | 1 | | 0 |
| ... | 0 | - 1 | | |
| k-1 | 0 | 0 | | 1 |
| k | 0 | 0 | | - 1 |

**Polynomial**

Diese Kontraste dienen der Trendanalyse und setzen ordinales Skalenniveau des Faktors voraus. Die Kontrastkoeffizienten errechnen sich aus den sog. ortogonalen Polynomen. In dieser Version des Skripts wird nicht näher darauf eingegangen.

Ausführliche Erläuterungen der Standard-Kontraste sind unter [151] für R bzw. unter [152] für SPSS zu finden.

## 3. 3 Auswahl der Kontraste

R bietet die o.a. Standard-Kontraste, die über die folgenden Funktionen erreichbar sind:

`contr.treatment(k,base=j)` (*j*=Nummer der Vergleichsgruppe)
`contr.sum(k)`
`contr.helmert(k)`
`contr.poly(k)`

wobei *k* die Anzahl der Gruppen ist. Die Auswahl erfolgt über das Kommando

`contrasts(Faktorname) <- contr.name`

Es gibt auch eine Voreinstellung für Objekte vom Typ „factor“:

`contr.treatment(k,base=k)` für „normale“ Faktoren
`contr.poly(k)` für „ordered factors“

die dann z.B. bei der Verwendung von „factor“-Variablen bei der Regression verwendet werden. Die Voreinstellung kann über

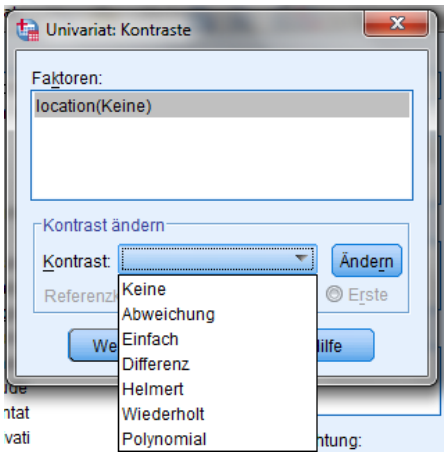
`options(contrasts=c("contr.name1","contr.name2"))`

geändert werden und über `getOption("contrasts")` abgefragt werden.

Bei SPSS gibt es in den Routinen zur Varianzanalyse sowie zur binär logistischen Regression zum einen das Unterkommando

`/Contrast(Faktorname)=name`

wobei *name* einer der oben für SPSS angeführten *englischen* Kontrastnamen ist, zum anderen in den Eingabemasken den Button „Kontraste“, der zu der folgenden Ausawhl führt:




Dabei darf allerdings nicht der „Ändern“-Button vergessen werden.

## 3. 4 nichtparametrische Kontraste für die RT-, ART- und KWF-Verfahren

Einige der bei Lüpsen [99] vorgestellten nichtparametrischen Varianzanalysen lassen sich ja auf die parametrischen Standardverfahren zurückführen, so insbesondere die RT-, die ART- sowie die verallgemeinerte Kruskal-Wallis- und Friedman-Analysen (KWF). Die Analyse von Kontrasten ist darin problemlos möglich.

Als erstes sollen Kontrast-Vergleiche in Verbindung mit dem RT-Verfahren, und zwar am Beispiel des Datensatzes 2 (`mydata2`) mit dem Faktor „drug“ demonstriert werden. Zunächst einmal wird angenommen, dass die erste Gruppe eine Vergleichsgruppe ist, gegen die die anderen drei Gruppen getestet werden sollen.

**Beispiel R-0.1:**

Die Tabelle 4.6 in Kapitel 4.3.4 [99] zeigt für den Faktor „drugs“ einen signifikanten Effekt an, der nun weiter untersucht werden soll. Dabei besteht die Hypothese, dass der Mittelwert der ersten Gruppe sich von allen anderen unterscheidet. Diese kann mit den „einfach“-Kontrasten (`contr.treatment`) geprüft werden. Dazu ist `gls` aus dem Paket `nlme` als Varianzanalysefunktion zu verwenden, die zwar keine Anova-Tabelle ausgibt, aber die Kontraste:

```
contrasts(mydata2$drug) <- contr.treatment(4,base=1)
aovc   <- gls(rx~group*drugs,mydata2)
summary(aovc)
```

Neben ein paar weiter nicht interessierenden Ergebnissen wird eine Tabelle aller Kontraste mit Tests ausgegeben. Hierbei ist anzumerken, dass bedingt durch die 2-faktorielle Analyse auch Kontraste für den anderen Faktor (group) sowie für die Interaktion ausgegeben werden. Die Zeilen `drugs2`,..,`drugs4` enthalten die Vergleiche mit `drug1`:

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 8.2500 | 2.514377 | 3.2811303 | 3.043817e-03 |
| group1 | 5.2500 | 2.514377 | 2.0879920 | 4.714492e-02 |
| drugs2 | 5.9750 | 3.346511 | 1.7854415 | 8.632831e-02 |
| drugs3 | 9.3750 | 3.426519 | 2.7360130 | 1.127545e-02 |
| drugs4 | 16.7125 | 3.346511 | 4.9940068 | 3.785352e-05 |
| group1:drugs2 | 1.7250 | 3.346511 | 0.5154622 | 6.107586e-01 |
| group1:drugs3 | -1.3750 | 3.426519 | -0.4012819 | 6.916220e-01 |
| group1:drugs4 | -7.9125 | 3.346511 | -2.3644026 | 2.613481e-02 |

*Tabelle 3-1*

**Beispiel S-0.1:**

Die Tabelle 4.8 in Kapitel 4.3.4 zeigt für den Faktor „drug“ einen signifikanten Effekt an, der nun weiter untersucht werden soll. Dabei besteht die Hypothese, dass der Mittelwert der ersten Gruppe sich von allen anderen unterscheidet. Diese kann mit den „simple“-Kontrasten geprüft werden. Dazu ist bei den Anweisungen für die oben erwähnet Analyse die Zeile

```
/Contrasst(drugs)=Simple(1)
```

einzufügen, wobei das „(1)“ die Nummer der Vergleichsgruppe angibt, also hier die erste. Die Ausgabe dazu sollte selbsterklärend sein:

| Kontrastergebnisse (K-Matrix) | | | |
|---|---|---|---|
| Einfacher Kontrast[a] | | | Abhängige Variable |
| | | | Rx |
| Niveau 2 vs. Niveau 1 | Kontrastschätzer | | 5,975 |
| | Hypothesenwert | | 0 |
| | Differenz (Schätzung - Hypothesen) | | 5,975 |
| | Standardfehler | | 3,347 |
| | Sig. | | ,086 |
| | 95% Konfidenzintervall für die Differenz | Untergrenze | -,917 |
| | | Obergrenze | 12,867 |
| Niveau 3 vs. Niveau 1 | Kontrastschätzer | | 9,375 |
| | Hypothesenwert | | 0 |
| | Differenz (Schätzung - Hypothesen) | | 9,375 |
| | Standardfehler | | 3,427 |
| | Sig. | | ,011 |
| | 95% Konfidenzintervall für die Differenz | Untergrenze | 2,318 |
| | | Obergrenze | 16,432 |
| Niveau 4 vs. Niveau 1 | Kontrastschätzer | | 16,713 |
| | Hypothesenwert | | 0 |
| | Differenz (Schätzung - Hypothesen) | | 16,713 |
| | Standardfehler | | 3,347 |
| | Sig. | | ,000 |
| | 95% Konfidenzintervall für die Differenz | Untergrenze | 9,820 |
| | | Obergrenze | 23,605 |
| a. Referenzkategorie = 1 | | | |

*Tabelle 3-2*

Das Vorgehen ist im Zusammenhang mit dem ART-Verfahren (vgl. Kapitel 4.3.6 in [99]) völlig identisch.

Ein wenig anders ist es bei Verwendung des KWF-Verfahrens (vgl. Kapitel 4.3.5 in [99]). Hier müssen die $\chi^2$-Werte für jeden Vergleich „mit der Hand“ ausgerechnet werden, was ein wenig mühselig ist, zumal SPSS nicht die Testgröße ausgibt:

$$\chi^2 = t^2 \cdot \frac{MS_{Fehler}}{MS_{total}} \qquad t = \frac{C}{s_e}$$

wobei

- $t$ die t-verteilte Teststatistik ist, die bei SPSS erst errechnet werden muss aus
- $C$ der Kontrastwert (in SPSS: Kontrastschätzer) und
- $s_e$ der Standardfehler (des Kontrastschätzers),
- $MS_{Fehler}$ die Fehlervarianz (aus der Anova-Tabelle zu entnehmen)
- $MS_{total}$ die Gesamtvarianz, die bereits für die Anova-Tests ermittelt worden war (vgl. Kapitel 4.3.5).

Die $\chi^2$-Werte haben jeweils 1 Fg und müssen anhand der Tabellen der $\chi^2$-Verteilung auf Signifikanz überprüft werden. Aus Tabelle 4-8 (Kapitel 4.3.5 in [99]) lässt sich $MS_{Fehler}$ = 43,35 sowie $MS_{total}$ =2904,5/32 = 90,77 errechnen.

**Beispiel R-0.2:**

In der Anova-Tabelle für diese Daten (Tabelle 4-6) fehlt ein Wert für $MS_{Fehler}$ . Dieser muss gegebenenfalls mit `aov` neu errechnet werden und ergibt `msfehler` mit dem Wert 43,35. Zur Berechnung der $\chi^2$-Werte müssen die t-Werte aus der Tabelle 9-1 quadriert, mit $MS_{Fehler}$ sowie durch $MS_{total}$ dividiert werden. Das kann in R programmiert werden. (Die Berechnung „per Hand“ kann dem Abschnitt „SPSS“ entnommen werden.) Wenn `aovc` das oben ermittelte Ergebnisobjekt von `gls` ist, dann lässt sich mit folgenden Anweisungen daraus zunächst die Kontrasttabelle `ctabelle`, die t-Werte `twerte` und schließlich die $\chi^2$-Werte `chisq`:

```
ctabelle<- as.data.frame(summary(aovc)$tTable)
twerte  <- ctabelle$"t-value"
names(twerte)<- row.names(ctabelle)
aov2r   <- anova(aov(rx~group*drugs,mydata2))
mstotal <- sum(aov2r[,2])/sum(aov2r[,1])
msfehler<- aov2r[4,3]
chisq   <- twerte^2*msfehler/mstotal
pvalues <- 1-pchisq(chisq,1)
data.frame(chisq,pvalues)
```

mit der nachfolgenden Ausgabe, worin die Zeilen `drugs2`,..,`drugs4` die gewünachten Testergebnisse enthalten:

```
                   chisq      pvalues
(Intercept)    5.14197182 0.0233541081
group1         2.08228611 0.1490168492
drugs2         1.52255843 0.2172327363
drugs3         3.57535389 0.0586429521
drugs4        11.91189867 0.0005577652
group1:drugs2  0.12690430 0.7216636075
group1:drugs3  0.07690983 0.7815296246
group1:drugs4  2.67008813 0.1022503615
```

*Tabelle 3-3*

**Beispiel S-0.2:**

Die Berechnung soll nur für den ersten Vergleich (drug1 - drug2) gezeigt werden:

$$\chi^2 = \left(\frac{5,975}{3,347}\right)^2 \cdot \frac{43,35}{90,77} = 1,52$$

Der kritische $\chi^2$-Wert bei 1 Fg beträgt 3,84, so dass kein Unterschied zwischen drug1 und drug2 nachgewiesen werden kann.

Das vorige Beispiel wird dahingehend modifiziert, dass drug1 und drug2 als etablierte Präparate angenommen werden, während drug3 und drug4 als neu angesehen werden. Daher sollen zum einen die beiden alten Präparate (1-2) sowie die beiden neuen Prärparate (3-4) verglichen werden, zum anderen die alten zusammen gegen die neuen zusammen ((1,2)-(3,4)). Daraus resultiert folgende Kontrastmatrix:

| | Kontraste | | |
|---|---|---|---|
| Gruppe | 1 | 2 | 3 |
| drug1 | 1 | 0 | 1 |
| drug2 | - 1 | 0 | 1 |
| drug3 | 0 | 1 | - 1 |
| drug4 | 0 | - 1 | - 1 |

*Tabelle 3-4*

Nachfolgend werden nur die Anweisungen für die Benutzer-spezifischen Kontraste aufgeführt. Die Ausgabe ist praktisch identisch mit der der Standard-Kontraste im vorigen Beispiel.

**Beispiel R-0.3:**

Auch hier dient natürlich wieder die Funktion `gls` aus dem Paket `nlme` zur Analyse der Kontraste. Lediglich die Spezifikation der Koeffizienten differiert. Die Werte müssen spaltenweise eingegeben, als Matrix mit 3 Spalten definiert und dann übergeben werden:

```
cont <- matrix( c(1,-1,0,0,  0,0,1,-1,  1,1,-1,-1), ncol=3)
contrasts(mydata2$drugs) <- cont
aovc <- gls(rx~group*drugs,mydata2)
summary(aovc)
```

**Beispiel S-0.3:**

Auch hier ist nur eine kleine Modifikation der Anweisungen des letzten Beispiels erforderlich. Die Kontrast-Anweisung lautet:

```
/Contrast(drugs) = Special(1 -1 0 0   0 0 1 -1   1 1 -1 -1)
```

Die Ausführungen dieses Abschnitts gelten gleichermaßen für Analysen mit Messwiederholungen.

## 3. 5 universelles Verfahren für Kontraste

Wenn die nichtparametrische Varianzanalyse nicht auf die parametrische zurückgeführt werden kann, steht damit auch nicht mehr die Kontrastfunktionalität der Standardroutinen von R und SPSS zur Verfügung. D.h. man verfügt nur über die Funktion zur Durchführung einer Varianzanalyse. Damit lassen sich aber immerhin durch passendes Umkodieren der Gruppen/Faktorvariablen sowohl zwei Gruppen vergleichen als auch Gruppen von Gruppen vergleichen. Das soll wieder am oben verwendeten Datensatz 2 (`mydata2`) erläutert werden.

Es sollen die Kontraste aus Tabelle 3-4 getestet werden. Vor jedem der drei Vergleiche muss die Gruppenvariable `drugs` so umkodiert werden, dass jeweils nicht verwendete Werte auf Missing gesetzt werden. Dies erfolgt mit einer Hilfsvariablen `d`.

**Beispiel R-0.4:**

Die Kontraste sollen im Anschluss an eine Kruskal-Wallis-Varianzanalyse durchgeführt werden. Es wird darauf aufmerksam gemacht, dass die levels-Angaben aus der factor-Definition der Gruppierungsvariablen (hier `drugs`) auf `d` übertragen werden, aber anschließend nicht mehr stimmen, da die Anzahl der Stufen von `d` auf zwei reduziert wurde. Das kann bei verschiedenen Funktionen zu Problemen führen. Gegebenenfalls muss dies in einer factor-Anweisung korrigiert werden.

```
kruskal.test(mydata2$x,drugs) # gloabler Vergleich

d <- mydata2$drugs            # Vergleich 1-2
d[d==3|d==4] <- NA
d<-factor(d,levels=c(1,2))
kruskal.test(mydata2$x,d)

d <- mydata2$drugs            # Vergleich 3-4
d[d==1|d==2] <- NA
d<-factor(d,levels=c(3,4))
kruskal.test(mydata2$x,d)

d <- mydata2$drugs            # Vergleich (1,2)-(3,4)
d[d==1|d==2] <- 1
d[d==3|d==4] <- 4
d<-factor(d,levels=c(1,4))
kruskal.test(mydata2$x,d)
```

Der globale $\chi^2$-Wert beträgt 11,2 . Die $\chi^2$-Werte der drei Kontraste: 1,97 (1-2), 2,61 (3-4) und 7,32 ((1,2)-(3,4)) mit der Summe von 11,9, die ungefähr dem globalen Wert entspricht, da die Kontraste orthogonal sind.

**Beispiel S-0.4:**

Die Kontraste sollen im Anschluss an eine Kruskal-Wallis-Varianzanalyse durchgeführt werden.

```
NPtests /independent test (x) group (drugs) Kruskal_Wallis.

* Vergleich 1-2 .
Recode drugs (1=1) (2=2) (3,4=sysmis) into d.
NPtests /independent test (x) group (d) Kruskal_Wallis.

* Vergleich 3-4 .
Recode drugs (3=3) (4=4) (1,2=sysmis) into d.
NPtests /independent test (x) group (d) Kruskal_Wallis.

* Vergleich (1,2)-(3,4) .
Recode drugs (1,2=1) (3,4=4) into d.
NPtests /independent test (x) group (d) Kruskal_Wallis.
```

Der globale $\chi^2$-Wert beträgt 11,2 . Die $\chi^2$-Werte der drei Kontraste: 1,97 (1-2), 2,61 (3-4) und 7,32 ((1,2)-(3,4)) mit der Summe von 11,9, die ungefähr dem globalen Wert entspricht, da die Kontraste orthogonal sind.

Aus diesem Beispiel geht das generelle Prozedere hervor. So lassen sich auch die im vorigen Abschnitt vorgenommenen Vergleiche der `drug2,..,drug4` gegen `drug1` durchführen.

## 3. 6 Kontraste bei logistischen Regressionen

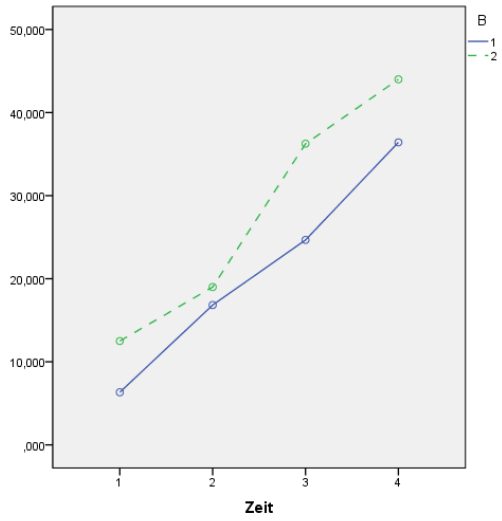
Bei der logistischen Regression gibt es für nominale Prädiktoren Standard-Kontraste. Wenn in R ein Prädiktor als „factor“ deklariert ist, wird für diesen automatisch die Kodierung gewählt, die in der `options(contrasts...)`-Anweisung festgelegt wurde (vgl. Kapitel 3.3). In SPSS kann bei der binär-logistischen Regression wie oben in 3.3 dargestellt die Kodierung gewählt werden. Speziellere Kontraste müssen wie oben in Kapitel 3.5 skizziert über Umkodierungen analysiert werden. Beispiele sind in Kapitel ?? zu finden.

## 3. 7 Kontraste für Messwiederholungen und Interaktionen

Aus dem eingangs (Kapitel 3.1) angeführten Signifikanztest für einen Kontrast kann abgelesen werden, dass dafür lediglich die Varianz $MS_{Error}$ erforderlich ist, die praktisch den Nenner des entsprechenden F-Tests für den untersuchten Effekt darstellt. Somit sind zumindest im Fall der RT-, ART- und KWF-Analysen Kontrastanalysen gleichermaßen für Versuchspläne mit Messwiederholungen durchführbar.

Sind für zwei Faktoren A und B Kontraste festgelegt worden, $k_A$*-1* Kontraste für A sowie $k_B$*-1* Kontraste für B, so resultieren aus den Produkten der jeweiligen Kontraste *(*$k_A$*-1)(*$k_B$*-1)* Kontraste für die Interaktion A*B. Damit lassen sich auch Interaktionen im Detail untersuchen. Sind in R bzw. SPSS für zwei Faktoren A und B Kontraste definiert worden, so werden automatisch auch diese Kontraste für die Interaktion A*B ausgegeben.

Dies soll am Datensatz 6 (`winer568`) demonstriert werden. Dieser umfasst die Gruppierungsfaktoren A und B sowie den Messwiederholungsfaktor Zeit. Tabelle 6-6 (Kapitel 6.5.3 in [99]) enthielt die Anova-Tabelle für das RT-Verfahren. Die Signifikanzen waren dort mittels des ART-Verfahrens verifiziert worden, so dass problemlos die einfach rangtransformierten Daten verwendet werden können. Hier soll jetzt die Interaktion B*Zeit näher betrachtet werden. Hierbei besteht die Vermutung, dass zwischen je zwei aufeinanderfolgenden Zeitpunkten der Anstieg der Werte für die Gruppen von B unterschiedlich stark verläuft.




*Interaktionsplot B*Zeit*

Hierzu werden für den Faktor Zeit die Standard-Kontraste „wierderholt“ festgelegt, bei denen die Zeitpunkte 1-2. 2-3 und 3-4 verglichen werden, sowie für Faktor B die Effekt-Kodierung

**Beispiel S-0.5:**

Hierzu werden zunächst analog den Berechnungen in Kapitel 6.3 [99] die Daten umstrukturiert, so dass aus den Variablen `V1,..,V4` eine Variable `V` entsteht. Anschließend wird diese Kriteriumsvariable `V` über alle Faktoren A, B und Zeit hinweg in Ränge transformiert (Variable `RV`) und schließlich die Daten wieder in die ursprüngliche Form zurücktransformiert, woraus u.a. die Messwiederholungsvariablen `RV.1,..,RV.4` gebildet werden. Mit diesen Daten kann nun die Varianzanalyse durchgeführt werden. Im Unterkommando `wsfactor` werden mit `Repeated` die gewünschten Kontraste für Zeit festgelegt, im Unterkommando `contrast` für die Gruppierungsfaktoren A und B.

```
GLM RV.1 RV.2 RV.3 RV.4 by A B
  /wsfactor=Zeit 4 Repeated
  /contrast(A)=Deviation
  /contrast(B)=Deviation
  /plot=profile(Zeit*B)
  /wsdesign=Zeit
  /design=A B A*B.
```

Die Ergebnisse der Varianzanalyse sind in Tabelle 6-6 [99] zusammengefasst (dort allerdings in der Ausgabe von R). Hier nun die Ausgabe der Kontraste für den Faktor Zeit:

| Tests der Innersubjektkontraste | | | | | | |
|---|---|---|---|---|---|---|
| Quelle | Zeit | Quadratsumme vom Typ III | df | Mittel der Quadrate | F | Sig. |
| Zeit | Niveau 1 vs. Niveau 2 | 867,000 | 1 | 867,000 | 71,383 | ,000 |
| | Niveau 2 vs. Niveau 3 | 1887,521 | 1 | 1887,521 | 122,932 | ,000 |
| | Niveau 3 vs. Niveau 4 | 1140,750 | 1 | 1140,750 | 86,777 | ,000 |
| Zeit * A | Niveau 1 vs. Niveau 2 | 800,333 | 1 | 800,333 | 65,894 | ,000 |
| | Niveau 2 vs. Niveau 3 | 379,688 | 1 | 379,688 | 24,729 | ,001 |
| | Niveau 3 vs. Niveau 4 | 280,333 | 1 | 280,333 | 21,325 | ,002 |
| Zeit * B | Niveau 1 vs. Niveau 2 | 48,000 | 1 | 48,000 | 3,952 | ,082 |
| | Niveau 2 vs. Niveau 3 | 266,021 | 1 | 266,021 | 17,326 | ,003 |
| | Niveau 3 vs. Niveau 4 | 48,000 | 1 | 48,000 | 3,651 | ,092 |

Hier interessieren die Ergebnisse des letzten Blocks Zeit*B. Daraus geht hervor, dass (vermutlich wegen der geringen Fallzahl) nur zwischen den Zeitpunkten 2 und 3 („Niveau 2 vs. Niveau 3“) ein unterschiedlich starker Anstieg der Werte nachgewiesen werden kann.

**Beispiel R-0.5:**

Ausgangsbasis ist der in Kapitel 6.5.3 [99] erstellte Datensatz `winer568t`.

- Zunächst müssen für die Faktoren die Kontraste festgelegt werden. Da die Standard-Kontraste „wiederholt“ in R standardmäßig nicht verfügbar sind, müssen diese als Koeffizienten-Matrix vorgegeben werden.
- Für A und B bietet `contr.sum` die Effekt-Kodierung.
- Die Kontraste werden wie oben über die Funktion `gls` des Pakets `nlme` getestet. Allerdings muss in diesem Fall der Faktor Zeit als Messwiederholungsfaktor deklariert werden. Dies erfolgt in `gls` über die Spezifikation der Fallkennung (`Vpn`) sowie der Struktur für die Kovarianzen der Messwiederholungsvariablen, die hier mit „*compound symmetry*“ festgelegt wird, was der sonst üblichen Sphärizität entspricht (vgl. Kapitel 5.2 in [99]):
  `corr = corCompSymm(, form= ~ 1 | Vpn)`

Die Kommandos lauten dann:

```
cont4 <- matrix( c(1,-1,0,0,  0,1,-1,0,  0,0,1,-1), ncol=3)
contrasts(winer568t$Zeit) <- cont4
contrasts(winer568t$A)    <- contr.sum
contrasts(winer568t$B)    <- contr.sum
aovgls <- gls(Rx~A*B*Zeit, data=winer568t,
          corr = corCompSymm(, form= ~ 1 | Vpn))
summary(aovgls)
```

Zunächst vorab die oben erzeugte Kontrastmatrix `cont4`:

```
> cont4

     [,1] [,2] [,3]
[1,]    1    0    0
[2,]   -1    1    0
[3,]    0   -1    1
[4,]    0    0   -1
```

Hier der Teil der Ausgabe, der die Kontrast-Tests enthält:

```
Coefficients:
                 Value Std.Error    t-value p-value
(Intercept)  24.500000 1.2012621  20.395216  0.0000
A1            2.187500 1.2012621   1.821001  0.0780
B1           -3.437500 1.2012621  -2.861574  0.0074
Zeit1       -15.083333 0.7663867 -19.681101  0.0000
Zeit2       -21.666667 0.8849471 -24.483573  0.0000
Zeit3       -15.708333 0.7663867 -20.496616  0.0000
A1:B1         0.500000 1.2012621   0.416229  0.6800
A1:Zeit1     -2.104167 0.7663867  -2.745568  0.0098
A1:Zeit2      3.958333 0.8849471   4.472960  0.0001
A1:Zeit3      4.395833 0.7663867   5.735790  0.0000
B1:Zeit1      0.354167 0.7663867   0.462125  0.6471
B1:Zeit2      2.708333 0.8849471   3.060447  0.0044
B1:Zeit3      0.354167 0.7663867   0.462125  0.6471
A1:B1:Zeit1   0.750000 0.7663867   0.978618  0.3351
A1:B1:Zeit2   1.500000 0.8849471   1.695017  0.0998
A1:B1:Zeit3   0.875000 0.7663867   1.141721  0.2620
```

Hier interessieren die Ergebnisse der Zeilen `B1:Zeit`. Daraus geht hervor, dass (vermutlich wegen der geringen Fallzahl) nur zwischen den Zeitpunkten 2 und 3 (`B1:Zeit2`) ein unterschiedlich starker Anstieg der Werte nachgewiesen werden kann.

Anzumerken ist noch, dass über `anova(aovgls)` auch eine Anova-Tabelle erzeugt werden kann.

# 4. Die klassischen Verfahren von Fisher, Tukey, Newman-Keuls und Duncan

Der Ansatz der Testverfahren weicht in der Regel vom klassischen Modell ab: Hier wird zu einem vorgegebenen $\alpha$ ein kritischer Wert $c_d$ errechnet, derart dass alle Mittelwertdifferenzen, die diesen Wert $c_d$ überschreiten, als signifikant auf dem $\alpha$-Niveau einzustufen sind. Während bei den paarweisen Vergleichen, etwa 2-Stichproben-t-Tests, der Schätzfehler für jeden Vergleich individuell berechnet wird, wird hier für alle Vergleiche nur *ein* Schätzfehler genommen, nämlich den über alle Gruppen gepoolten Schätzfehler. Dieser ist im parametrischen Fall üblicherweise $\sqrt{MS_{Fehler}/n}$ und aus der Anova-Tabelle abzulesen ebenso wie die dazugehörenden Freiheitsgrade $df_{Fehler}$. Hierbei ist zunächst einmal wie auch im Folgenden $n = n_1 = ... = n_k$ der einheitliche Stichprobenumfang für alle $k$ Gruppen. Aus dem Poolen des Schätzfehlers ergibt sich allerdings auch die Voraussetzung der Varianzhomogenität. Die kritischen Werte $c_d$ haben die Gestalt

$$c_d = q_\alpha(df_{Fehler})\sqrt{2}\sqrt{MS_{Fehler}/n} \qquad (4-1)$$

mit einem von $\alpha$ abhängigen kritischen oder Quantilswert $q_a$, der von Test zu Test verschieden ist. (Dass bei einigen der Tests der Faktor $\sqrt{2}$ in der Berechnung für $c_d$ fehlt, liegt daran, dass für die dort verwendete Verteilung gilt: $q = t\sqrt{2}$.)

Zunächst einmal wird der Fall unabhängiger Stichproben besprochen. Hier kommt es naturgemäß häufiger zu unterschiedlichen Stichprobenumfängen $n_i$. Die meisten der klassischen Verfahren sind zwar für gleiche $n_i$ konzipiert, sind aber auch noch valide, wenn man das einheitliche $n$ durch das harmonische Mittel $\tilde{n}_h$ der $n_i$ ersetzt. Schaut man sich die Formeln der Tests genau an, so wird in diesen häufig bei unterschiedlichen $n_i$ die Größe $\frac{1}{n}$ durch $\frac{1}{2}\cdot\left(\frac{1}{n_i}+\frac{1}{n_j}\right)$ ersetzt.

## 4. 1 Fishers LSD (least significant difference)

Dies ist wohl das älteste dieser Verfahren. Hier wird lediglich beim t-Test die Fehlervarianz $s_{\bar{x}_2-\bar{x}_1}$ durch die o.a. gepoolte Varianz ersetzt. Der kritische Wert errechnet sich dann als

$$c_d = t_{\alpha'}(k(n-1))\sqrt{2}\sqrt{MS_{Fehler}/n} \qquad (4-2)$$

mit dem kritischen Wert $t_a$ der t-Verteilung mit *k(n-1)* Fg. Ursprünglich hatte Fisher den Test als „echten" post-hoc-Test im Anschluss an einen signifikanten F-Wert der Varianzanalyse konzipiert und somit für die Ermittlung der kritischen Werte der t-Verteilung keine $\alpha$-Adjustierung vorgesehen. Simulationen haben gezeigt, dass in diesem Fall das $\alpha$-Risiko auch annähernd eingehalten wird, solange die Varianzen homogen sind (vgl. Wilcox [102]).

Da jedoch wie in Kapitel 1.6 dargelegt bei mehrfacher Anwendung dieses Tests die Rate für den Fehler 1. Art verletzt wird, modifizierte Fisher später diesen Test zum PLSD (*protected least significant difference*), indem das $\alpha$ vorher wie im vorigen Kapitel beschrieben hinsichtlich der Anzahl der Vergleiche $m$ adjustiert wird. Dieses Verfahren ist bedingt durch die $\alpha$-Adjustierung extrem konservativ, so dass es kaum mehr angewandt wird.

R: `agricolae, ExpDes`
SPSS: `oneway. glm`

## 4. 2 Tukey HSD (honestly significant differnce)

Das wohl bekannteste Verfahren ist das von Tukey (für gleiche $n_i$) bzw. von Tukey & Kramer (für ungleiche $n_i$). Geprüft wird wie bei der Varianzanalyse: alle *k* Mittelwerte sind gleich. Theoretisch sollte deswegen hier auch dassselbe Ergebnis herauskommen. Das Verfahren geht aber einen anderen Weg als die Varianzanalyse: Es wird die Verteilung der Spannweite (engl. *range*) der *k* Mittelwerte untersucht, bzw. der größten aller *m* Mittelwertdifferenzen. Sind alle Mittelwerte gleich, so müsste die Spannweite 0 sein. Die daraus resultierende Verteilung heißt *range distribution* (vgl. Kapitel 1.4). Der kritische Wert, mit dem die Mittelwertdifferenzen zu vergleichen sind, errechnet sich als:

$$c_d = q_\alpha(k, df_{Fehler})\sqrt{MS_{Fehler}/n} \qquad (4-3)$$

wobei $q_a(k,df_{Fehler})$ der kritische Wert der studentized range-Verteilung zur vorgegebenen Irrtumswahrscheinlichkeit $\alpha$ ist. Dabei sind $df_{Fehler}$ die Freiheitsgrade von $MS_{Fehler}$ und $k$ ist in der Terminologie dieser Verteilung die Spannweite der Mittelwerte, aber zugleich die Gruppenzahl. Mit diesem kritischen Wert $c_d$ werden nun alle Differenzen verglichen, auch die kleineren. Dadurch ist dieser Test als sehr konservativ einzustufen, was Vergleiche ausgenommen $max(\bar{x}) - min(\bar{x})$ anbetrifft.

Games und Howell haben diesen Test für heterogene Varianzen und ungleiche Zellenbesetzungszahlen erweitert (vgl. Kapitel 6.1).

Der Tukey-Test testet zwar dieselbe Hypothese wie der F-Test der parametrischen Varianzanalyse, ist aber deutlich konservativer und hat daher eine geringere Power. Im Vergleich zu den anderen nachfolgenden Tests, die auch auf der studentized range-Verteilung basieren, ist er der Einzige, beim dem das tatsächliche $\alpha$-Risiko exakt $\alpha$ ist.

R: `agricolae, multcomp, TukeyHSD, EspDes`
SPSS: `oneway. glm`

## 4. 3 Student-Newman-Keuls (SNK)

Die Autoren *Newman* und *Keuls* hatten festgestellt, dass, wenn die größte Mittelwertdifferenz signfikant ist, die zweitgrößten Differenzen eine andere Verteilung haben, nämlich die der Spannweite *k-1*. Daher werden die Mittelwertdifferenzen der Ordnung (*k-1*) mit dem entsprichen kritischen Wert der studentized range-Verteilung verglichen. Entsprechend, wenn diese signfikant sind, die Mittelwertdifferenzen der Ordnung (*k-2*) mit dem entsprechenden kritischen Wert der studentized range-Verteilung. Der kritische Wert, mit dem die Mittelwertdifferenzen zu vergleichen sind, errechnet sich hier als:

$$c_d = q_\alpha(r, df_{Fehler})\sqrt{MS_{Fehler}/n} \qquad (4-4)$$

wobei $q_a(r,df_{Fehler})$ derselbe kritische Wert der studentized range-Verteilung zur vorgegebenen Irrtumswahrscheinlichkeit $\alpha$ wie beim o.a. Tukey-Test ist. Allerdings wird hier nicht für alle Mittelwertdifferenzen dasselbe $q_a$ genommen, sondern der Parameter *r*, der die Spannweite angibt, der Mittelwertdifferenz angepasst.

Dazu werden die *k* Mittelwerte der Größe nach sortiert: $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \ldots \le \bar{x}_{(k)}$ und wie unten dargestellt in einer Matrix angeordnet. (Hier wird mit $..._{(i)}$ der Index innerhalb Reihenfolge angegeben.) Die größte Differenz $\bar{x}_{(k)} - \bar{x}_{(1)}$ hat die Spannweite *k*, also ist *r=k*, so dass zum Vergleich dieser Differenz mit $c_d$ der kritische Wert $q_a(k,df_{Fehler})$ verwendet wird. Die beiden

nächstgrößeren Differenzen $\bar{x}_{(k-1)} - \bar{x}_{(1)}$ und $\bar{x}_{(k)} - \bar{x}_{(2)}$ haben die Spannweite *k-1*, also ist *r=k-1*, so dass zum Vergleich mit $c_d$ der kritische Wert $q_a(k\text{-}1, df_{Fehler})$ verwendet wird. Das Prozedere wird fortgesetzt mit *r=k-2* usw, bis ein Vergleich, beginnend bei *r=k* rechts oben, nicht mehr signifikant ist. Alle übrigen Mittelwertdifferenzen sind als nicht signifikant zu werten. Dadurch dass kleinere Mittelwertdifferenzen auch mit kleineren kritischen Werten $c_d$ verglichen werden, sind bei dem SNK-Verfahren in der Regel mehr Signifikanzen zu erwarten, als bei dem Tukey-Test.:

| | $\bar{x}_{(1)}$ | $\bar{x}_{(2)}$ | ... | $\bar{x}_{(k-1)}$ | $\bar{x}_{(k)}$ |
|---|---|---|---|---|---|
| $\bar{x}_{(1)}$ | - | $\bar{x}_{(2)} - \bar{x}_{(1)}$ | | $\bar{x}_{(k-1)} - \bar{x}_{(1)}$ | $\bar{x}_{(k)} - \bar{x}_{(1)}$ |
| $\bar{x}_{(2)}$ | | - | | | $\bar{x}_{(k)} - \bar{x}_{(2)}$ |
| ... | | | - | | |
| $\bar{x}_{(k-1)}$ | | | | - | $\bar{x}_{(k)} - \bar{x}_{(k-1)}$ |
| $\bar{x}_{(k)}$ | | | | | - |

r = k
r = k-1
r = 2

Dem Newman-Keuls-Test wird vorgeworfen, dass er das α-Risiko nicht konsequent einhält. Peritz hat eine Modifikation dieses Tests mittels der Verfahren von Ryan (vgl. Kapitel 5) entwickelt, derart dass der experimentweise Fehler unterhalb α bleibt (vgl. Kapitel 5.3).

R: `agricolae, ExpDes, mutoss`
SPSS: `oneway. glm`

## 4. 4 Duncan

Auch bei dem SNK-Verfahren können nicht alle Unterschiede nachgewiesen werden. Duncan hat das Verfahren mit dem *New Multiple Range Test* liberalisiert. Eine Skizzierung der Methode ist bei [107] nachzulesen. Die dahinter steckende Idee ist letztlich diese: Bei *k* Gruppen sind nur (*k-1*) Mittelwertvergleiche unabhängig. Wenn aber, wie bei Tukey und Newman-Keuls, *k(k-1)* Vergleiche durchgeführt werden, sind diese nicht mehr unabhängig und die (implizit vorgenommene) α-Korrektur muss entsprechend reduziert werden. Dazu wird wie bei den anderen Tests die studentized range-Verteilung verwendet, aber das α gemäß

$$\alpha' = 1 - (1 - \alpha)^{k-1}$$

korrigiert. D.h. bei k=4 wird bei α=0,05 ein α‘=0,143 genommen. Dies resultiert in eine andere Verteilung, die der o.a. studentized range-Verteilung sehr ähnlich ist, aber kleinere kritische Werte $q_a(k, df_{Fehler})$ erzeugt. Insbesondere ist der kritische Wert für die Spannweite *k*, also den Vergleich $max(\bar{x}) - min(\bar{x})$, deutlich kleiner. Dadurch ist der Duncan-Test deutlich liberaler ist als die von Tukey und Newman-Keuls. Verzichtet man allerdings auf den Test für die Spannweite *k* und führt statt dessen den F-Test der Varianzanalyse durch (die beide die identische Hypothese „alle Mittelwerte sind gleich“ testen), ist der Duncan-Test ein legitimes und brauchbares Verfahren (vgl. Kapitel 14.1).

Die Vorgehensweise ist, abgesehen von der Verwendung dieser kritischen Werte, identisch mit der beim o.a. SNK-Verfahren. Wenn auch das Verhältnis von Fehler 1. und 2. Art bei dem

Duncan-Test ausgewogener ist als bei den zuvor aufgeführten Tukey- und SNK-Verfahren, gilt der Duncan-Test als verpönt, da der $\alpha$-Fehler nicht korrekt eingehalten wird.

R: `agricolae, ExpDes`
SPSS: `oneway. glm`

## 4. 5 Tukey-b

Dieser Test, auch Tukey *Wholly Significant Difference* (*WSD*) genannt, ist im wahrsten Sinne des Wortes ein Kompromiss-Test: Für jeden Vergleich einer Mittelwertdifferenz werden die kritischen Werte des Tukey HSD- und des SNK-Tests gemittelt. Tukey hatte selbst eingesehen, dass sein o.a. HSD-Test sehr konservativ ist, und sich zu diesem Kompromiss durchgerungen. Allerdings hat er sich in der Praxis kaum durchgesetzt.

R: `-`
SPSS: `oneway. glm`

# 5. Ryans schrittweise Verfahren

Ryan hat ein schrittweises Verfahren entwickelt, das sich recht allgemein einsetzen lässt, da es zunächst nicht an einen bestimmten statistischen Test gebunden ist. Das Prozedere ähnelt dem des Newman-Keuls-Tests. Dazu werden zunächst ähnlich wie beim SNK-Test die Mittelwerte der Größe nach sortiert: $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \ldots \leq \bar{x}_{(k)}$. (Hier wird mit $..._{(i)}$ der Index innerhalb Reihenfolge angegeben.) Es beginnt mit dem globalen Test für die Hypothese $\mu_1 = ... = \mu_k$. Wird diese abgelehnt, sind also nicht alle Mittelwerte gleich, so sind weitere Tests erforderlich. Als nächstes werden zweimal *r=k-1* Gruppen verglichen: $\mu_{(1)} = ... = \mu_{(k-1)}$ und $\mu_{(2)} = ... = \mu_{(k)}$. Ergibt einer der Tests ungleiche Mittelwerte, so sind im nächsten Schritt jeweils *r=k-2* Gruppen zu vergleichen: $\mu_{(1)} = ... = \mu_{(k-2)}$, $\mu_{(2)} = ... = \mu_{(k-1)}$ bzw. $\mu_{(3)} = ... = \mu_{(k)}$. Ab diesem Schritt ist allerdings eine $\alpha$-Korrektur erforderlich:

$$\alpha' = 1 - ( 1 - \alpha )^{r/k}$$

wobei *r* für die Anzahl der in dem jeweiligen Schritt zu vergleichenden Gruppen steht, also zuletzt *r=k-2*. Dieses schrittweise Verfahren endet, wenn in einem Schritt keine Signifikanzen mehr auftreten.

Aus diesen Schritten resultieren nun sog. *homogene* Untergruppen der *k* Gruppen, deren Mittelwerte sich *nicht* voneinander unterscheiden. Möchte man umgekehrt wissen, welche Gruppen sich von welchen anderen bzgl. der Mittelwerte unterscheiden, so sind von allen Paarvergleichen alle die zu eliminieren, bei denen die beiden beteiligten Gruppen zur selben homogenen Untergruppen gehören. Das Signifikanzniveau entspricht dann dem vorgegebenen $\alpha$. p-Werte für die signifikanten Vergleiche werden hierbei nicht ermittelt.

Zum Test der Hypothesen können beliebige Tests verwendet werden, die mehrere Mittelwerte simultan auf Gleichheit prüfen. Im parametrischen Fall sind dies in erster Linie der F-Test und der Q-Test (studentized range-Verteilung). Im nichtparametrischen Fall können dies z.B. der Kruskal-Wallis-Test oder die Friedman-Varianzanalyse sein.

Die nachfolgend auf dem Verfahren von Ryan basierenden Vergleichsmethoden haben eine vergleichsweise hohe Power: Sie sind dem Tukey-Test überlegen. Dabei schneidet das Verfahren von Peritz noch besser ab, kann aber nicht an den SNK-Test reichen.

## 5. 1 REGWF (Ryan-Einot-Gabriel-Welsch mit F-Tests)

Bei dem REGWF-Verfahren wird die oben beschriebene Methode von Ryan mit dem F-Test angewandt, es werden also schrittweise mehrere Mittelwerte mittels des klassischen F-Tests der Varianzanalyse auf Gleichheit getestet.

R: -
SPSS: `oneway. glm`

## 5. 2 REGWQ (Ryan-Einot-Gabriel-Welsch mit Q-Tests)

Bei dem REGWF-Verfahren wird die oben beschriebene Methode von Ryan mit dem Q-Test (der studentized range-Verteilung) angewandt. Im Gegensatz zu dem o.a. REGWF-Verfahren mit dem F-Test ist hier bei dem Q-Test die Spannweite *r* zu berücksichtigen. Daher nachfolgend ein paar Ausführungen dazu.

Im ersten Schritt wird der globale Test für die Hypothese $\mu_1 = ... = \mu_k$ mittels des Q-Tests, also letztlich des Tukey-Tests (vgl. Kapitel 4.2) durchgeführt. Wird diese abgelehnt, werden als nächstes zweimal *k-1* Gruppen verglichen: $\mu_{(1)} = ... = \mu_{(k-1)}$ sowie $\mu_{(2)} = ... = \mu_{(k)}$, wiederum mit der studentized range-Verteilung, diesmal allerdings mit der Spannweite *r=k-1*. Ergibt einer der Tests ungleiche Mittelwerte, so sind im nächsten Schritt jeweils *r=k-2* Gruppen zu vergleichen: $\mu_{(1)} = ... = \mu_{(k-2)}$, $\mu_{(2)} = ... = \mu_{(k-1)}$ bzw. $\mu_{(3)} = ... = \mu_{(k)}$, allerdings mit der eingangs aufgeführten $\alpha$-Korrektur. Generell entspricht bei den nachfolgenden Q-Tests die Spannweite *r* der Anzahl der zu vergleichen Gruppen.

Wegen der größeren Power des F-Tests gegenüber der studentized range-Verteilung (Q-Test) ist der REGWF dem REGWQ vorzuziehen.

R: `mutoss`
SPSS: `oneway. glm`

## 5. 3 Peritz

*Peritz* hat ein Verfahren entwickelt, das eine Mischung aus dem Newman-Keuls-Test (vgl. Kapitel 4.3) und dem o.a. Verfahren von Ryan darstellt. Auch hier werden homogene (sich nicht signifikant unterscheidende) Untergruppen der *k* Gruppenmittelwerte ermittelt. Ziel ist es, das $\alpha$-Risiko besser unter Kontrolle zu halten, als es der SNK-Test tut. Das ursprünglich von Peritz beschriebene Verfahren erwies sich als extrem unpraktikabel und aufwändig. Verschiedene Autoren entwickelten dann Algorithmen zur Durchführung der Peritz-Vergleiche. Daher kursieren von diesem Verfahren in der Literatur mehrere Varianten. Nachfolgend die Version, wie sie von *Begun & Gabriel* sowie von *Ramsey* beschrieben wurde (vgl. Braun & Tukey [161]).

- Im ersten Schritt werden alle Paardifferenzen als nicht signifkant eingestuft, die auf Grund des SNK-Tests als nicht signifikant erkannt werden.
- Im zweiten Schritt werden alle Paardifferenzen als signifkant eingestuft, die auf Grund des REGWFQ-Tests als signifikant erkannt werden.
- Alle übrigen Paardifferenzen werden als „verdächtig“ (engl. *contentious*) eingestuft und müssen weiter untersucht werden. Wenn die beiden betrachteten Mittelwerte bereits in einer homogenen Untergruppe enthalten sind, so ist die Paardifferenz ebenfalls als nicht signifikant anzusehen. Nur wenn alle Teilmengen der übrigen *k-2* Mittelwerte mittels des REGWQ-Verfahrens als nicht homogen erkannt werden, ist die betrachtete Paardifferenz als signifikant einzusteufen.

Das Verfahren ist weder in R noch in SPSS verfügbar, allerdings gibt es im Internet Fortran-Code dafür.

# 6. Vergleiche bei inhomogenen Varianzen

## 6. 1 Games & Howell

Games und Howell haben den o.a. Test von Tukey und Kramer so erweitert, dass er nicht nur für ungleiche $n_i$ , sondern auch für ungleiche Gruppenvarianzen $s_i^2$ anwendbar ist. Dabei wird die Korrektur von *Welch und Satterthwaite*, die zur Berechnung der Fehlervarianz sowie der Freiheitsgrade für den t-Test bei inhomogenen Varianzen verwandt wird, auf die Berechnung von $MS_{Fehler}$ und $df_{Fehler}$ bei der Berechnung von $c_d$ angewandt:

$$MS_{Fehler} = \left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)/2 \qquad (6-1a)$$

$$df_{Fehler} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{s_j^2}{n_j}\right)^2}{n_j - 1}} \qquad (6-1b)$$

Dies erfordert allerdings eine separate Berechnung von $MS_{Fehler}$ und $df_{Fehler}$ bei jedem Vergleich.

Dieser Test gilt als leicht liberal. Simulationen haben gezeigt, dass bei kleinen bis mittleren $n_i$ (etwa $n_i \le 40$ ) ausgerechnet bei annähernd gleichen $s_i^2/n_i$ das α-Risiko leicht verletzt wird. Auf der anderen Seite hat er eine deutlich größere Power als die u.a. T2- und T3-Tests.

Dieses Verfahren kann gleichermaßen auf andere Tests wie z.B. die von Newman-Keuls oder Duncan angewandt werden.

R: -
SPSS: `oneway. glm`

## 6. 2 Dunnetts C

Eine andere Verallgemeinerung des Tukey-Tests hinsichtlich ungleichen $n_i$ und ungleichen Gruppenvarianzen $s_i^2$ bietet Dunnett mit dem C-Test. Auch hier erfordert es eine separate Berechnung von $MS_{Fehler}$ und $df_{Fehler}$ bei jedem Vergleich.

Für den Vergleich der Gruppen *i* und *j* errechnet sich der kritische Wert $c_d$ als:

$$c_d = \frac{q_\alpha(k, n_i - 1)\frac{s_i^2}{n_i} + q_\alpha(k, n_j - 1)\frac{s_j^2}{n_j}}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

wobei $q_a(k,df)$ wie beim Tukey-Test die kritischen Werte der studentized range-Verteilung für die Spannweite *k* ist.

R: `DTK.test`
SPSS: `oneway. glm`

## 6. 3 Dunnetts T3 und Hochbergs GT2

Dunnett hat eine weitere Verallgemeinerung des Tukey-Tests entwickelt hat: den T3-Test. Dieser basiert auf einer anderen Variante der studentized range-Verteilung: dem *studentized maximum modulus*. Für den Vergleich der Gruppen *i* und *j* errechnet sich der kritische Wert $c_d$ als:

$$c_d = r_\alpha(m, df)\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}$$

wobei $r_a(m,df)$ der kritische Wert der s*tudentized maximum modulus*-Verteilung ist, *m* die Anzahl der Vergleiche ist und sich die Freiheitsgrade *df* wie im Test von Games & Howell errechnen.

Für kleine bis mittlere $n_i$ (etwa $n_i \leq 40$ ) ist dieser T3-Test gegenüber dem o.a. C-Test überlegen. Für größere $n_i$ ist dagegen der C-Test vorzuziehen.

Für den Fall gleicher Varianzen ist dieser Test identisch mit dem GT2-Test von Hochberg.

Sind auch die $n_i$ gleich, so kann der o.a. Quantilswert $r_a(m,df)$ durch $q_\alpha(m, df)/\sqrt{2}$ , den Quartilswert der studentized range-Verteilung, ersetzt werden und man erhält exakt den HSD-Test von Tukey.

R: `drsmooth`
SPSS: `oneway. glm`

## 6. 4 Tamhane T2

Dahinter verbirgt sich lediglich der t-Test für inhomogene Varianzen kombiniert mit einer α-Korrektur nach Sidak (vgl. 2.1) bzgl. aller *m=k(k-1)/2* möglichen Vergleiche. Dadurch ist dieser Test extrem konservativ

R: `-`
SPSS: `oneway. glm`

## 6. 5 Hubers Sandwich-Schätzer

Lineare Kontraste oder im einfachsten Fall ein Mittelwertvergleich stellen einen Spezialfall der linearen Modelle dar. Damit sind auch die dafür entwickelten robusten Schätzmethoden anwendbar. Eine davon ist Hubers *Sandwich*-Schätzung (vgl. Wikipedia [162] sowie etwas ausführlicher Freedman [163]). Die Testergebnisse müssen dann anschließend adjustiert werden. Allerdings: Im Gegensatz zu den o.a. robusten Varianten des t-Tests von *Welch* und *Satterthwaite,* die im Fall gleicher Varianzen das Ergebnis des t-Tests liefern, sind die Ergebnisse bei der Sandwich-Schätzung in jedem Fall schwächer als die der „normalen“ (kleinste-Quadrat-) Schätzung.

R: `multcomp`
SPSS: `-`

# 7. Verallgemeinerte Kontraste

In Kapitel 3.1 waren Kontrastvergleiche $C_j$ ($j$=1,..,$m$) von Mittelwerten vorgestellt worden, die jedoch zwei Restriktionen unterlagen: die Anzahl *k-1* und die Orthogonalität. Hierbei sei *m* die Anzahl der Vergleiche. Es gibt jedoch auch Verfahren, um beliebige Vergleiche durchzuführen, allerdings auf Kosten der Effizienz. Durch die Verwendung von $MS_{Fehler}$ werden auch hier homogene Varianzen vorausgesetzt. Allerdings gelten hier die gleichen Aussagen zur Robustheit wie bei der Varianzanalyse (vgl. Kapitel 4.1 in [99]). Darüber hinaus gibt es auch eine Reihe von Verfahren für den Fall ungleicher Varianzen.

Es sei hier ausdrücklich noch einmal darauf aufmerksam gemacht, dass mittels der Kontraste auch beliebige Paarvergleiche möglich sind.

## 7. 1 Dunn-Bonferroni

Hierunter verbirgt sich lediglich das in Kapitel 3.1 besprochene Verfahren unter Verwendung der Bonferroni-α-Adjustierung, womit die eingehaltenen Restriktionen kompensiert werden. Die Korrektur erfolgt mit *m = Anzahl der zu prüfenden Kontraste*.

Es sei hier darauf aufmerksam gemacht, dass es, wie in Kapitel 2 aufgeführt, deutlich bessere α-Adjustierungen gibt als die von Bonferoni. Solche können dann benutzt werden, um eine größere Anzahl von durchgeführten Vergleichen zu kompensieren.

Somit können zur Durchführung alle Funktionen benutzt werden, die Benutzer-definierte Kontraste erlauben.

## 7. 2 Scheffé

Dieses Verfahren sieht auf den ersten Blick aus wie das der orthogonalen Kontraste in Kapitel 3. Allerdings wird bei Scheffé die Testgröße leicht modifiziert. Ausgangsbasis sind die in Kapitel 3.1 definierte Testgröße $SS_C$ (3-2) und die Fehlervarianz der Varianzanalyse $MS_{Fehler}$. Die Testgröße der Scheffé-Vergleiche

$$F = \frac{SS_C}{(k-1)MS_{Fehler}}$$

ist dann F-verteilt mit *k-1* Zähler-FG und $df_{Fehler}$ Nenner-FG, wobei $df_{Fehler}$ die FG von $MS_{Fehler}$ ist.

Der Scheffé-Test gilt als eines der konservativsten Verfahren und ist daher nicht unbedingt zu empfehlen.

R: `agricolae`, `DescTools`
SPSS: `oneway. glm`

## 7. 3 Tukey

Tukeys HSD-Test lässt sich auf Kontraste verallgemeinern. Der kritische Wert, mit dem ein Kontrast C zu vergleichen sind, errechnet sich wie beim Tukey-Test (vgl. Kapitel 4.2) als:

$$c_d = q_\alpha(k, df_{Fehler})\sqrt{MS_{Fehler}/n} \cdot \sum_i^k |c_i|/2$$

wobei $q_a(k,df_{Fehler})$ der kritische Wert der studentized range-Verteilung zur vorgegebenen Irr-

tumswahrscheinlichkeit α ist. Dabei sind $df_{Fehler}$ die Freiheitsgrade von $MS_{Fehler}$ und $k$ ist in der Terminologie dieser Verteilung die Spannweite der Mittelwerte, aber zugleich die Gruppenzahl. Analog zum Tukey-Kramer-Test für ungleiche $n_i$ gibt es auch für diesen Kontrast-Test in einer Variante für ungleiche Zellenbesetzungszahlen.

## 7. 4 Welch, Ury & Wiggins

Die beim o.a. Verfahren von *Games & Howell* verwendete Korrektur der Freiheitsgrade von *Welch* (6-1a und 6-1b) lässt sich auch auf den Test von Scheffé anwenden. Dazu sind in der o.a. Formel die $s_i^2$ durch $c_i s_i^2$ zu ersetzen und die Summation über alle $k$ Gruppen auszudehnen. Dieses Verfahren ist auch unter den Namen *Ury* und *Wiggins* bekannt.

Die in Kapitel 3.1 vorgestellte Berechnung von $SS_C$ (3-2) für einen Kontrast $C_j$ wird bzgl. der Varianzen gewichtet :

$$SS_{C_j} = \frac{(c_1\bar{x}_1 + c_2\bar{x}_2 + \ldots + c_k\bar{x}_k)^2}{\frac{c_1^2 s_1^2}{n_1} + \frac{c_2^2 s_2^2}{n_2} + \ldots + \frac{c_k^2 s_k^2}{n_k}} \qquad (7-1a)$$

Die Freiheitsgrade errechnen wie folgt:

$$df_{Fehler} = \frac{\left(\frac{c_1^2 s_1^2}{n_1} + \frac{c_2^2 s_2^2}{n_2} + \ldots + \frac{c_k^2 s_k^2}{n_k}\right)^2}{\frac{\left(\frac{c_1^2 s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{c_2^2 s_2^2}{n_2}\right)^2}{n_2 - 1} + \ldots + \frac{\left(\frac{c_k^2 s_k^2}{n_k}\right)^2}{n_k - 1}} \qquad (7-1b)$$

Schließlich ist zum Test der Kontraste $C_j$ entweder der F-Test wie im vorigen Abschnitt oder der t-Test wie in Kapitel 3.1 beschrieben mit einer α-Korrektur durchzuführen. Die Nenner-FG sind die o.a. $df_{Fehler}$.

## 7. 5 Brown & Forsythe und Kaiser & Bowden

*Brown & Forsythe* ([181], vgl. dazu auch Kapitel 4.3.3 in [99]) haben eine Variante des *Scheffé*-Tests für inhomogene Varianzen entwickelt. Im Prinzip wird ein ähnlicher Ansatz verfolgt wie oben von Welch. Dieser zeigte bei Simulationen jedoch verschiedentlich eine Verletzung der α-Fehlerrate. *Kaiser* und *Bowden* (vgl. Wilcox [102]) haben das Verfahren dann weiter modifiziert, so dass der Fehler 1. Art nicht verletzt wird. Die Tests von *Brown & Forsythe* bzw. *Kaiser & Bowden* sind gegebenenfalls denen von *Dunn-Bonferroni* oder *Welch* vorzuziehen, auch wenn sie vergleichsweise konservativ sind (vgl. Rafter, Abell & Braselton [106]).

## 7. 6 Simes

Das Verfahren von *Simes* ähnelt den α-Korrekturen von *Holm, Hochberg* und *Benjamini* (vgl. Kapitel 2.2) hinsichtlich des Prozedere, beinhaltet letztlich die Methode von Benjamini & Hochberg. Nachfolgend das Verfahren in der allgemeineren Version für inhomogene Varianzen mit der Korrektur der Freiheitsgrade von *Welch*, die bereits beim o.a. Verfahren von *Games & Howell* (6-1) benutzt wurde.

Zunächst wird mit jedem Kontrast $C_j$ ein F- oder t-Test durchgeführt.

$$t_j = \sqrt{SS_{C_j} / MS_{Fehler}}$$

und zwar im Falle homogener Varianzen mit $SS_{Cj}$ wie in Kapitel 3.1 bzw. im Falle inhomogener Varianzen wie oben beim Welch-Test beschrieben. Die Freiheitsgrade $df_{Fehler}$ sind im ersten Fall $(\sum n_i) - k$ bzw. im zweiten Fall wie oben beim Welch-Test (7-1b) beschrieben. Hieraus ergeben sich p-Werte $p_j$. Nun werden die p-Werte der Größe nach sortiert: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ . Anschließend werden diese schrittweise verglichen.

Es beginnt mit dem kleinsten: $p_{(1)}$ . Ist $p_{(1)} \leq \alpha / m$ , so wird der dazugehörige Vergleich als signifikant angesehen und es wir als nächstes $p_{(2)}$ geprüft. Ist $p_{(2)} \leq (2\alpha) / m$ , so wird der dazugehörige Vergleich als signifikant angesehen. Allgemein wird $p_{(j)} \leq (j\alpha) / m$ verglichen. Das Verfahren wird fortgesetzt, bis ein Vergleich nicht signifikant ist.

Dieses Verfahren von *Simes* ist weniger konservativ als die Verfahren von *Dunn* und *Bonferroni*, halten aber dennoch das $\alpha$-Risiko korrekt ein.

# 8. Weitere multiple Vergleiche

## 8. 1 Waller-Duncan

Das Verfahren von *Waller & Duncan* geht einen anderen Weg: Im Gegensatz zu dem sonst üblichen Ansatz, die Fehlerrate 1. Art korrekt einzuhalten, wird hier die gesamte Anzahl der falschen Signifikanzen, also die Rate für den Fehler 1. und 2. Art möglichst klein gehalten.

Entsprechend dem Bayeschen Ansatz wird eine Verlustfunktion aufgestellt, die minimiert wird. Und zwar wird dazu für jeden Vergleich eine Funktion angenommen, die die Mittelwertdifferenz mit einer Konstante $k_0$ multipliziert, wenn $H_0$ irrtümlich angenommen wurde bzw. mit einer Konstante $k_1$ multipliziert, wenn $H_1$ irrtümlich angenommen wurde. Die Summe dieser Funktionen wird dann minimiert. Über das Verhältnis $K = k_0/k_1$ kann eine Abwägung von Fehler 1. Art zum Fehler 2. Art vorgenommen werden, von Duncan *seriousness* genannt. Daraus resultiert wie bei Fishers LSD ein kritischer Wert $c_d$ zur Beurteilung der Mittelwertdifferenzen, der neben der Fehlervarianz und *K* noch vom F-Wert der Varianzanalyse abhängt. Und zwar in der Weise, dass bei kleinen F-Werten Fehler 1. Art vermieden werden und bei großen F-Werten Fehler 2. Art. Hierfür gibt es spezielle Tabellen, weswegen der Test auch kaum mit der Hand durchzuführen ist. Wegen der Ähnlichkeit von $c_d$ zu dem des Tests von Fisher wird er auch verschiedentlich *Baysian least significant difference* (*BLSD*) genannt.

Anstatt eines $\alpha$ ist hier *K* vorzugeben. Übliche Werte für *K* sind 50, 100 und 500, die in etwa einem $\alpha$ von 0,10 , 0,05 bzw. 0,01 entsprechen. Praktische Hinweise für die Wahl von *K* sind in der Literatur rar. Einige Ausführungen sind bei *Carmer & Walker* [164] zu finden.

R: `agricolae`
SPSS: `oneway. glm`

## 8. 2 Scott & Knott

Wieder einen anderen Weg haben Scott & Knott [165] eingeschlagen: Es werden homogene Untergruppen mittels einer Clusteranalyse ermittelt, und zwar mittels einem divisivem hierarchischen Verfahren. Im ersten Schritt wird versucht, die *k* Mittelwerte in zwei möglichst homogene Untergruppen aufzuteilen. Dazu wird für alle möglichen Partitionen in zwei Gruppen das Verhältnis von Zwischengruppen- zu Innergruppenstreuung mittels eines Likelihood-Ratio-Tests überprüft. (Dieser ist letztlich dem F-Test der Varianzanalyse sehr ähnlich.) Verläuft dieser Test negativ, endet das Verfahren. Andernfalls wird in den folgenden Schritten versucht, auf dieselbe Weise die im vorigen Schritt gefundenen Untergruppen wieder zu teilen. Das Verfahren endet, wenn kein Likelihood-Ratio-Test mehr signifikant ist.

Simulationen haben gezeigt, dass auf der einen Seite das $\alpha$-Risiko häufig nicht eingehalten wird, aber auf der anderen Seite gegenüber anderen Verfahren auch kleinere Mittelwertunterschiede, insbesondere bei kleinen Stichproben, erkannt werden (vgl. Carmer & Walker [164]).

R: `ExpDes, Laercio`
SPSS: -

## 8. 3 Gabriel

*Gabriel* [166] hat ein Verfahren entwickelt, bei dem für jeden Mittelwert ein Konfidenzintervall erstellt wird. Zwei Mittelwerte unterscheiden sich dann, wenn deren Konfidenzintervalle sich

nicht überschneiden. Die untere Intervallgrenze $l_i$ bzw. die obere $u_i$ für das Konfidenzintervall eines Mittelwerts $\bar{x}_i$ errechnen sich wie folgt:

$$l_i = \bar{x}_i - r_\alpha(m, df)\sqrt{\frac{MS_{Fehler}}{2n_i}} \qquad u_i = \bar{x}_i + r_\alpha(m, df)\sqrt{\frac{MS_{Fehler}}{2n_i}}$$

Hierbei ist $r_a(m,df)$ der kritische Wert der s*tudentized maximum modulus*-Verteilung ist, $m$ die Anzahl der Vergleiche und $df$ die Freiheitsgrade der Fehlervarianz $MS_{Fehler}$.

R: -
SPSS: `oneway. glm`

# 9. Vergleich mit einer Kontrollgruppe

Die Ausgangssituation ist hier eine etwas andere: Es liegen zwar wieder *k* Gruppen (unabhängige Stichproben) vor. Allerdings zeichnet sich eine dadurch aus, dass sie eine Kontrollgruppe ist. In diesem Fall interessieren nur die Vergleiche der anderen Gruppen mit dieser Kontrollgruppe. Dadurch werden anstatt *k(k-1)/2* nur noch *k-1* Vergleiche durchgeführt. An dieser Stelle könnte man auf die Idee kommen, die Fragestellung über *k-1* orthogonale Kontraste (vgl. Kapitel 3) zu lösen. Aber man kann schnell überprüfen, dass die Vergleiche nicht orthogonal sind. Somit kommt man um eine Korrektur bzgl. der Anzahl von Vergleichen nicht herum. Allerdings wird durch die kleinere Anzahl der negative Effekt der erforderlichen $\alpha$-Korrektur deutlich reduziert. Ein weiterer Aspekt: Im Gegensatz zu den Paarvergleichen liegen hier meist einseitige Hypothesen vor, was bei der Wahl des $\alpha$ bzw. der Interpretation des p-Wertes zu berücksichtigen ist.

Vielfach wird an Kontrollgruppenvergleiche eine Bedingung gestellt, mehr aus Vernunftsgründen, denn aus mathematischen: Die Kontrollgruppe sollte ein größeres *n* haben als die anderen Gruppen. Als Faustregel wird genannt: $n_i \sim n\sqrt{k-1}$.

Es wird darauf aufmerksam gemacht, dass es neben den hier aufgeführten Tests gegen eine Kontrollgruppe auch solche gibt, die gegen die „beste“ oder „schlechteste“ Gruppe vergleichen.

Noch abschließend eine Bemerkung zu den Alternativhypothesen. Während bei den o.a. paarweisen Vergleichen üblicherweise 2-seitig getestet wird, also $\mu_i = \mu_j$ gegen $\mu_i \neq \mu_j$, bieten die Programme häufig bei den Kontrollgruppenvergleichen einen 1-seitigen Test, also z.B. $\mu_i \leq \mu_j$ gegen $\mu_i > \mu_j$.

## 9. 1 Dunnett

*Dunnetts* Test zum Vergleich mit einer Kontrollgruppe ist wohl der bekannteste. Das Prozedere ist dasselbe wie bei den klassischen Tests (vgl. Kapitel 4): Es wird ein kritischer Wert errechnet, mit dem alle Mittelwertdifferenzen zu vergleichen sind.

$$c_d = t_\alpha(k, df_{Fehler})\sqrt{2MS_{Fehler}/n}$$

Hierbei ist $t_a$ ein kritischer Wert einer speziell von Dunnett modifizierten t-Verteilung zu *k* zu vergleichenden Mittelwerten und zu einem vorgegebenen $\alpha$. $MS_{Fehler}$ ist wie üblich die Fehlervarianz, die der Anova-Tabelle zu entnehmen ist, und $df_{Fehler}$ die dazugehörigen Freiheitsgrade. Auch hier geht der Test eigentlich von gleichen Zellenbesetzungszahlen aus. Das *n* kann aber ohne Probleme durch das harmonische Mittel $\tilde{n}_h$ der $n_i$ ersetzt werden.

Auch hier wird wie bei den anderen klassischen Tests Homogenität der Varianzen vorausgesetzt. Allerdings hat *Rudolph* [167] durch Simulationen festgestellt, dass Dunnetts Test robust gegen leichte Verletzungen der Normaltäts- und der Varianzhomogenitäts-Voraussetzungen ist. Für kleinere Stichproben ist er sogar nichtparametrischen Tests, wie dem u.a. Test von *Steel*, überlegen.

R: `nparcomp`
SPSS: `oneway. glm`

## 9. 2 schrittweise Dunnett-Verfahren

Marcus et al [168] haben mittels des closure-Prinzips (vgl. Kapitel 2.6.2) schrittweise Verfahren des o.a. Tests von *Dunnett* entwickelt. Diese verhalten sich zum o.a. einfachen (single-step) Verfahren in etwa der *Newman-Keuls*-Test zum *Tukey* HSD-Verfahren. Sie lassen sich auch mit den schrittweisen α-Korrekturen von *Holm* (absteigend) oder *Hochberg* (aufsteigend) vergleichen. Während beim o.a. klassischen Dunnett-Test alle Mittelwertdifferenzen mit demselben kritischen Wert $c_d$ verglichen werden, variieren die hier je nach Rangstufe *r* der Testgröße. Die schrittweisen Verfahren sind dem klassischen Test vorzuziehen, da sie deutlich stärker sind.

Beim step-down-Verfahren werden zunächst die Mittelwertdifferenzen $d_i$ zur Vergleichsgruppe der Größe nach sortiert: $d_{(k-1)} \leq \ldots \leq d_{(2)} \leq d_{(1)}$ . Im ersten Schritt wird die größte Differenz $d_1$ mit dem kritischen Wert des single-step-Verfahrens für *k* Gruppen („normaler" Dunnett-Test) verglichen. Ist dieser Vergleich signifikant, wird danach $d_2$ mit dem kritischen Wert für *k-1* Gruppen verglichen. Allgemein wird $d_i$ mit $c_d = t_\alpha(k - i + 1, df_{Fehler}) \sqrt{2MS_{Fehler}/n}$ verglichen. Sobald ein Vergleich nicht signifikant ist, werden alle übrigen Nullhypothesen angenommen.

Analog verläuft das step-up-Verfahren: Die Vergleiche sind dieselben wie beim o.a. step-down-Verfahren, beginnen jedoch mit der kleinsten Differenz $d_{k-1}$ . Sobald ein Vergleich signifikant ist, werden alle Nullhypothesen zu den größeren Differenzen abgelehnt.

R: `DunnettTests`
SPSS: -

## 9. 3 Steel

*Steel* hat einen nichtparametrischen Test zum Vergleich mit einer Kontrollgruppe konstruiert (vgl. Winer [13]). Die Berechnung erfolgt natürlich über eine Rangtransformation der beobachteten Werte $x_i$. Allerdings erfolgt die Berechnung und somit auch die Rangtransformation für jeden Vergleich separat. Wird also die Kontrollgruppe K mit einer Experimentalgruppe $E_i$ verglichen, so werden die Werte beider Gruppen gemeinsam in Ränge transformiert und anschließend für beide Gruppen die Rangsummen $T_i(K)$ und $T(E_i)$ ermittelt. Steel hat für das Miminum beider Werte $\min(T_i(K), T(E_i))$ kritische Werte tabelliert.

## 9. 4 Gao und Nemenyi

Für die nichtparametrischen Tests von Gao und Nemenyi (vgl. Kapitel 12) gibt es Varianten, bei denen alle Gruppen mit einer Kontrollgruppe verglichen werden.

R: `mutoss, Nemenyi`
SPSS: -

# 10. Versuchspläne mit Messwiederholungen

Sucht man nach multiplen Mittelwertvergleichen speziell für abhängige Stichproben, so wird man kaum welche finden. Ausnahme: einige nichtparametrische Tests, die in Kapitel 12 aufgeführt sind. Auch die meisten Bücher, die sich dem Thema der Multiplen Mittelwertvergleiche widmen, sparen den Fall abhängiger Stichproben aus. So gibt es letztlich nur folgende Möglichkeiten:

- die Übertragung der in den vorangegangenen Kapiteln vorgestellten Methoden auf den Fall der Messwiederholungen,
- das schrittweise Verfahren von Ryan (vgl. Kapitel 5),
- die Durchführung von Paarvergleichen, z.B. mittels gepaartem t-Test, zusammen mit einer $\alpha$-Adjustierung.

Verschiedentlich wird empfohlen, die Messwiederholungen zu ignorieren und die Gruppen wie unabhängige Stichproben zu behandeln. Hierbei verliert man zum einen den „Vorteil“ der Messwiederholungen mit dem gegenüber $MS_{Fehler}$ kleineren $MS_{Residuen}$, wodurch, wenn überhaupt, weniger Unterschiede nachgewiesen werden können. Zum anderen können dann die Beobachtungen nicht als unabhängig angesehen werden, was weitere Unsicherheiten bei den Ergebnissen verursacht.

Theoretisch können die o.a. Verfahren, sofern sie als Standardfehler die Fehlervarianz $MS_{Fehler}$ aus der Anova-Tabelle benutzen, auf den Fall abhängiger Stichproben übertragen werden. Dazu ist lediglich in den Formeln $MS_{Fehler}$ durch die kleinere $MS_{Residuen}$ (allgemein: die Varianz im Nenner des F-Wertes des zu untersuchenden Effekts) zu ersetzen. Dies wird u.a. von Winer [13] empfohlen. Maxwell [169] hat sich auch intensiv mit diesem Thema auseinandergesetzt. Allerdings wird ausdrücklich darauf hingewiesen, dass die zentrale Voraussetzung der Varianzanalyse mit Messwiederholung (vgl. Kapitel 5.1 in [99]), nämlich die Homogenität der Varianz-Kovarianzstruktur (*Spherizität*) strikt eingehalten werden muss. Die Korrekturen der Freiheitsgrade von *Huynh-Feldt* oder *Greenhouse-Geisser* im Falle nicht vorhandener Spherizität lassen sich leider nicht auf die Tests multipler Mittelwertvergleiche übertragen.

Doch praktisch gibt es bei der Umsetzung Probleme: Weder R noch SPSS bieten die in Frage kommenden Tests bei Messwiederholungen an. Das betrifft die Verfahren des Kapitels 4, Dunnetts Kontrollgruppenvergleich (Kapitel 9.1) sowie die verallgemeinerten Kontraste von Scheffé und Tukey in Kapitel 7. Allerdings lässt sich das sowohl in R als auch in SPSS mit relativ wenig Aufwand auch per Hand erledigen (vgl. Kapitel 15.2.1 und 16.2)

Eine andere Lösung bietet das schrittweisen Verfahren von Ryan (vgl. Kapitel 5), das zunächst einmal mit einem beliebigen globalen Test der $k$ Mittelwerte durchgeführt werden kann. Also auch mit dem F-Test für abhängige Stichproben. Dies hat sogar den Vorteil, dass dieser etwas robuster ist als die in Kapitel 7 aufgeführten Tests, die auf der studentized range-Verteilung basieren. Zum anderen kann hier bei fehlender Spherizität ohne Weiteres eine der Korrekturen der Freiheitsgrade von *Huynh-Feldt* oder *Greenhouse-Geisser* angewandt werden (vgl. Kapitel 5.2 in [99]). Doch auch hier: Weder R noch SPSS bieten das Verfahren von Ryan für Messwiederholungen an.

So bleibt als einzige leicht umsetzbare Lösung der paarweise Vergleich mittels gepaartem t-Test unter Verwendung einer $\alpha$-Adjustierung. Dies hat immerhin den Vorteil, dass dabei keinerlei Varianzhomogenität gefordert wird. Bestenfalls müssten die Paardifferenzen einem Test auf Normalverteilung unterzogen werden.

# 11. Mehrfaktorielle Versuchspläne

Warum muss hierüber überhaupt geschrieben werden? In Kapitel 4.3.1.3 [99] war erläutert worden, dass der F-Test eines Faktors durch die Hinzunahme weiterer Faktoren in das zu analysierende Design beeinflusst wird. Und zwar kann, wie dort an einem Beispiel demonstriert wurde, unter Umständen ein Effekt erst dadurch signifikant werden, wenn durch die Berücksichtigung weiterer Faktoren der statistische Fehler reduziert wird. Liegt nun ein solcher Fall vor, d.h. ein Effekt ist 1-faktoriell nicht signifikant, jedoch bei einer mehrfaktoriellen Analyse, dann wird vielfach einer der oben vorgestellten Mittelwertvergleiche den Effekt nicht näher analysieren können, weil bei den Berechnungen meistens nur die Daten des zu betrachteten Faktors einfließen.

Es gibt jedoch Ausnahmen: Insbesondere die klassischen Verfahren (vgl. Kapitel 4), aber auch einige andere, so sämtliche Kontrast-Verfahren sowie die von Simes und von Gabriel, verwenden in der Berechnung des kritischen Wertes $c_d$, mit dem die Mittelwertdifferenzen zu vergleichen sind, die Größe $MS_{Fehler}$ (*pooled variance estimate*) aus der Varianzanalyse zur Schätzung des Standardfehlers. Diese Fehlerstreuung ist aber genau die, die durch die Hinzunahme weiterer Faktoren reduziert wird und dadurch einen Effekt signifikant werden lässt. Wird diese dann für einen Mittelwertvergleich aus der Anova-Tabelle der mehrfaktoriellen Analyse entnommen, so kann der Effekt korrekt analysiert und die Gruppenunterschiede erkannt werden. Allerdings setzt die Verwendung von $MS_{Fehler}$ homogene Varianzen voraus.

Leider wird dies nicht von allen Programmen korrekt umgesetzt.

# 12. nichtparametrische Methoden

Üblicherweise werden bei nichtparametrischen Methoden die Werte der abhängigen Variablen (Kriteriumsvariablen) in Ränge transformiert. Grundsätzlich sind dabei im Falle multipler Mittelwertvergleiche zwei Methoden zu unterscheiden:

- *joint ranking*
  Die Werte aller k Gruppen zusammen werden in Ränge transformiert und anschließend werden damit die paarweisen Vergleiche durchgeführt, d.h. die Ränge sind für alle Vergleiche immer dieselben.
- *pairwise ranking*
  Für jeden paarweisen Vergleich werden die Werte der beiden betrachteten Gruppen in Ränge transformiert und hiermit der Vergleich durchgeführt, d.h. für jeden Vergleich können die Ränge andere sein.

Ein entscheidender Unterschied, der meistens als nachteilig angesehen wird, ist, dass bei dem *joint ranking* die Ränge, die für einen Vergleich zweier Gruppen benutzt werden, von den anderen *k-2* nicht betrachteten Gruppen abhängen. Das gibt es bei den parametrischen Verfahren nicht.

Ein Begriff, der bei nichtparametrischen Varianzanalysen eine wichtige Rolle spielt, ist der *relative Effekt*. Er dient zur Unterscheidung zwischen zwei Verteilungen, etwa der Zufallsvariablen $X_1$ und $X_2$. Der relative Effekt von $X_2$ zu $X_1$ist definiert als $p^+ = P(X_1 \leq X_2)$, d.h. durch dieWahrscheinlichkeit, dass $X_1$ kleinere Werte annimmt als $X_2$. Dabei hat $X_1$ eine stochastische Tendenz zu größeren Werten als $X_2$, falls $p^+ < 1/2$ und eine stochastische Tendenz zu kleineren Werten, falls $p^+ > 1/2$ ist. Ausführliche Ausführungen hierzu sind bei E. Brunner & U. Munzel [3] zu finden.

## 12. 1 Anwendung parametrischer Verfahren

Wie häufig in der nichtparametrischen Statistik erhält man ein praktikables Verfahren, wenn man ein parametrisches mit vorhergehender Rangtransformation benutzt und gegebenenfalls den Signifikanztest anpasst (vgl. Kapitel 4.3.4 und 4.3.6 in [99]). Dies empfehlen auch Conover & Iman [6] sowie Day & Quinn [101] für den Fall multipler Mittelwertvergleiche. Damit können u.a. die klassischen Methoden aus Kapitel 4 benutzt werden. Allerdings resultiert für die Rangbildung daraus das problematische *joint ranking* .

Cochran und Lunnay hatten gezeigt, dass parametrische statistische Verfahren sich meistens auch auf dichotome Kriteriumsvariablen anwenden lassen. (Vgl. dazu Kapitel 7.1 in Lüpsen [99] sowie W.G. Cochran [ 7 ] und G.H. Lunney [ 8 ].) Speziell für die klassischen Mittelwertvergleiche haben Chuang-Stein und Tong [170] dies bestätigt, sofern die Fallzahl hinreichend groß ist (mindestens 40 FG) und die relativen Häufigkeiten zwischen 0,25 und 0,75 liegen, da andernfalls die Varianzen zu unterschiedlich werden können.

## 12. 2 Aligned Rank Transform-Verfahren

Eine Variante der Anwendung parametrischer Verfahren auf Variablen, die die parametrischen Voraussetzungen nicht erfüllen, ist das Aligned Rank Transform-Verfahren (ART). Bei diesem wird zum Test eines Faktors die Kriterumsvariable um die Effekte anderer Faktoren bereinigt und dann in Ränge transformiert (vgl. dazu Kapitel 2.3 in Lüpsen [99]). Dies ist ein sehr effizientes Verfahren für nichtparametrische Varianzanalysen. Da üblicherweise den multiplen Mit-

telwertvergleichen eine Varianzanalyse vorausgeht, ist es somit ein Leichtes, im Anschluss an eine ART-Analyse multiple Mittelwertvergleiche für die transformierte Variable, z.B. mittels der klassischen Verfahren (vgl. Kapitel 4), durchzuführen. Ein ausführliche Beschreibung des ART-Verfahrens mit Beispielen dazu ist bei Abundis [179] zu finden.

## 12. 3 Nemenyi

Der Test von Nemenyi kann als der Klassiker unter den nichtparametrischen multiplen Mittelwertvergleichen angesehen werden. Er wird häufig als Analogon zum Scheffé-Test für die rangtransformierte Kriteriumsvariable angesehen. Allerdings ist er wesentlich konservativer als die Anwendung des Scheffé- oder Tukey-Tests auf die rangtransformierte Variable. Der Nemenyi-Test verwendet das *joint ranking*.

Das Prozedere ist dasselbe wie beim Tukey-Test: Die Mittelwertdifferenzen der Ränge werden alle mit demselben kritischen Wert $c_d$ verglichen, der sich im Fall unabhängiger Stichproben errechnet als

$$c_d = \frac{q_\alpha(k)}{\sqrt{2}} \sqrt{\left(\frac{n(n+1)}{12}\right) \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Hierbei ist $q_a(k)$ der kritische Wert der normal range-Verteilung zu einem vorgegenen $\alpha$ und zur Spannweite *k*. Alternativ kann auch die studentized range-Verteilung verwendet werden, wenn bei dieser die größt mögliche Anzahl von FG für die Fehlervarianz gewählt wird. Für den Fall von Bindungen bei der abhängigen Variablen wird die Prüfung mittels der $\chi^2$-Verteilung vorgezogen:

$$c_d = \sqrt{C\chi^2_\alpha(k-1)} \sqrt{\left(\frac{n(n+1)}{12}\right) \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

wobei $\chi^2{}_\alpha(k\text{-}1)$ der Quantilswert zum vorgegebenen $\alpha$ für *k-1* FG ist. $C$ ist hierbei die Bindungskorrektur, wie sie beim Kruskal-Wallis-Test benutzt wird.

Es gibt auch einen entsprechenden Test für abhängige Stichproben. Auch hier werden die Mittelwertdifferenzen der rangtransformierten Werte der abhängigen Variablen mit einem einheitlichen kritischen Wert $c_d$ verglichen

$$c_d = \frac{q_\alpha(k)}{\sqrt{2}} \sqrt{\frac{k(k+1)}{6n}}$$

mit $q_a(k)$ wie oben. Eine spezielle Variante für den Fall von Bindungen wie bei unabhängigen Stichproben ist nicht bekannt. Daher ist der Test bei abhängigen Stichproben und bei größerer Anzahl von Bindungen nicht exakt.

R: `Nemenyi, PMCMR, NSM3`
SPSS: -

## 12. 4 Dwass-Steel-Critchlow-Fligner

Beim Verfahren von Dwass, Steel und Critchlow-Fligner werden werden im Gegensatz zu den meisten anderen für jeden Vergleich nur die Daten der beiden betrachteten Gruppen verwendet (*pairwise ranking)*. Für den Test werden die Werte beider Gruppen *i* und *j* zusammen in Ränge transformiert, und für beide Gruppen die Rangsummen $R_i$ bzw. $R_j$ errechnet. Die Testgröße ergibt sich, indem für eine der beiden Gruppen, z.B. *i*, die Rangsumme $R_i$ z-transformiert und mit $\sqrt{2}$ multipliziert wird:

$$c_d = \sqrt{2} \cdot z(R_i)$$

Diese wird mit dem kritischen Wert der normal range-Verteilung zu einem vorgegenen $\alpha$ verglichen. Alternativ kann auch die studentized range-Verteilung verwendet werden, wenn bei dieser die größt mögliche Anzahl von FG für die Fehlervarianz gewählt wird.
Ein Vergleich mit dem o.a. Nemenyi-Test hat ergeben, dass beide Verfahren als gleichwertig anzusehen sind, insbesondere für kleinere Gruppenzahlen *k*.

Dieser Test sollte nicht mit dem *Fligner-Killeen-Test* auf Gleichheit von Varianzen verwechselt werden.

R: `NSM3`
SPSS: -

## 12. 5 Conover & Iman

Auch das Verfahren von *Conover & Iman* ist als Transskription eines „Klassikers“ anzusehen: als Adaption des LSD-Tests von Fisher (vgl. Kapitel 4.1) auf die gemäß *joint ranking* rangtransformierte Daten.

## 12. 6 Gao

Gao, Alvo, Chen und Li [171] haben ein Verfahren für multiple Vergleiche entwickelt, das auf dem Modell für beliebige Verteilungen ordinaler oder metrischer Variablen von Akritas, Arnold und Brunner (vgl. [22]) basiert und das auch Basis für die nichtparametrische ATS-Varianzanalyse (vgl. Kapitel 2.5 in [99]) ist. Wie bei letzterem wird überprüft, ob alle *relativen Effekte* kleiner oder gleich 0,5 sind (vgl. dazu die einleitenden Bemerkungen dieses Kapitels). Dieses Verteilungsmodell wird auf die Verfahren von Scheffé und Tukey übertragen. Es ist daher leicht konservativ, kann allerdings im Gegensatz zu vielen anderen Verfahren bedenkenlos auf ordinal skalierte Kriteriumsvariablen angewandt werden.

Das Verfahren von Gao gibt es auch zum Vergleich mit einer Kontrollgruppe. Darüber hinaus haben Munzel und Tamhane [172] ein äquivalentes Verfahren für abhängige Stichproben entwickelt, das aber bislang weder in R noch in SPSS verfügbar ist.

R: `nparcomp`
SPSS: -

## 12. 7 Campbell & Skillings

Das Verfahren von Campbell und Skillings ist das von Ryan (vgl. Kapitel 5), bei dem als Test die Kruskal-Wallis-Varianzanalyse benutzt wird. Das Prozedere wird hier noch einmal kurz skizziert.

Es beginnt mit dem globalen Test für die Hypothese $\mu_1 = ... = \mu_k$ mittels des Kruskal-Wallis-Tests. Wird diese abgelehnt, sind also nicht alle Mittelwerte gleich, so sind weitere Tests erforderlich. Dazu werden zunächst ähnlich wie beim SNK-Test die Mittelwerte der Größe nach sortiert: $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \ldots \leq \bar{x}_{(k)}$. (Hier wird mit $..._{(i)}$ der Index innerhalb Reihenfolge angegeben.) Im nächsten Schritt werden zwei Kruskal-Wallis-Varianzanalysen durchgeführt zum Test von $\mu_{(1)} = ... = \mu_{(k-1)}$ sowie $\mu_{(2)} = ... = \mu_{(k)}$. Ergibt einer der Tests ungleiche Mittelwerte, so sind im anschließenden Schritt jeweils *r=k-2* Gruppen zu vergleichen:
$\mu_{(1)} = ... = \mu_{(k-2)}$, $\mu_{(2)} = ... = \mu_{(k-1)}$ bzw. $\mu_{(3)} = ... = \mu_{(k)}$. Ab diesem Schritt ist allerdings eine $\alpha$-Korrektur erforderlich:

$$\alpha' = 1 - ( 1 - \alpha )^{r/k}$$

wobei *r* für die Anzahl der in dem jeweiligen Schritt zu vergleichenden Gruppen steht, also zuletzt *r=k-2*. Dieses schrittweise Verfahren endet, wenn in einem Schritt keine Signifikanzen mehr auftreten.

Aus diesen Schritten resultieren nun sog. *homogene* Untergruppen der *k* Gruppen, deren Mittelwerte sich *nicht* voneinander unterscheiden. Die signifikanten Unterschiede können wie bei den o.a. Verfahren von Ryan (vgl. Kapitel 5) ermittelt werden.

Dieses schrittweise Verfahren von Ryan, das hier mit dem Kruskal-Wallis H-Test zum globalen Test der *k* Mittelwerte angewandt wurde, kann auch für abhängige Stichproben benutzt werden. Dann ist anstatt des H-Tests die Friedman-Varianzanalyse einzusetzen, zweckmäßigerweise mit der Korrektur von Iman & Davenport (vgl. Kapitel 5.3.2 in [99]). Die Tests werden verschiedentlich auch mit anderen Verfahren durchgeführt.

R: `nparcomp`
SPSS: `nptests`

## 12. 8 nichtparametrische Kontraste

Konietschke, Hothorn und Brunner [173] haben einen Test von nichtparametrischen Kontrasten entwickelt, mit denen sich natürlich auch paarweise Vergleiche durchführen lassen, sowohl für abhängige wie auch unabhängige Stichproben. Das Modell ist dasselbe, das auch beim o.a. Test von Gao verwendet wird: Der Vergleich der relativen Effekte, ein Modell, das sowohl für metrische als auch für ordinale Variablen anwendbar ist (vgl. Akritas, Arnold und Brunner [22]).

R: `nparcomp`
SPSS: -

## 12. 9 Quade

Der Test von *Quade* (vgl. Garcia et al [124]) ist eigentlich kein multipler Mittelwertvergleich, sondern ein globaler Test auf Gleichheit der Mittelwerte bei Messwiederholungen, ähnlich dem Friedman-Test. Er kann natürlich auch für paarweiser Vergleiche eingesetzt werden, wenn anschließend die p-Werte adjustiert werden.

Die Idee ist folgende: Bei der Rangbildung $R_{ji}$ für die Friedman-Analyse, bei der pro Fall/ Merkmalsträger *j* (*j=1,...,n* ) die Werte *i=1,..., k* vergeben werden, ist nur eine geringe Differenzierung zwischen den *k* Gruppen möglich. Daher wird eine Fallgewichtung $Q_j$ eingeführt, die Fälle mit einem größeren Wertesprektrum bevorzugt. $Q_j$ errechnet sich aus der Spannweite $D_j$ der Werte eines Falls (Differenz von Maximum und Minimum der $x_{ji}$ ), die dann in Ränge umgerechnet wird. Aus beiden Rängen $R_{ji}$ und $Q_j$ zusammen wird dann das Produkt $W_{ji} = Q_j * R_{ji}$ errechnet. Zum Vergleich zweier Gruppen werden schließlich die Rangsummen von $W_{ji}$ verwendet:

$$T_i = \left( \sum_{j=1}^{n} W_{ji} \right) / (n(n+1)/2)$$

die dann in einen t- oder z-Test umgerechnet werden.

R: `quade.test`
SPSS: -

## 12. 10 van der Waerden

Der wenig bekannte Test von *van der Waerden* ist eine Alternative zum Kruskal-Wallis-H-Test, also ein globaler Test und kein multipler Mittelwertvergleich. Zunächst werden wie beim H-Test die Werte in Ränge gewandelt. Diese werden danach aber in Quantile der Normalverteilung transformiert, *normal scores* genannt. Auch hier wird dann ein $\chi^2$-Test durchgeührt, der die Gleichheit der Mittelwerte testet. Im Falle einer Signifikanz werden dann wiederum paarweise van der Waerden-Tests durchgeführt, üblicherweise allerings ohne eine $\alpha$-Adjustierung. Daher ist dieser Test als multipler Mittelwertvergleich liberal. Der Test ist ausführlich in Wikipedia [174] beschrieben.

R: `agricolae`
SPSS: -

## 12. 11 Fligner-Policello

Der Test von *Fligner und Policello* ist eine robuste Alternative zum bekannten *Mann-Whitney-U-Test* (auch *Wilcoxon-Rangsummen*-Test genannt). Er gehört daher eigentlich nicht in diese Sammlung von Tests. Dennoch soll auf ihn aufmerksam gemacht werden, da er im Gegensatz zum U-Test nur Symmetrie der Verteilungsform voraussetzt, nicht aber dass die Form in beiden Gruppen die gleiche ist. So dürfen insbesondere die Varianzen ungleich sein.

Der Test basiert auf *placement scores*. Dabei ist ein solcher score $P(x_{1i})$ für eine Beobachtung $x_{1i}$ der ersten Gruppe definiert als Anzahl der Beobachtungen $x_{2j}$ der zweiten Gruppe, die kleiner als $x_{1i}$ sind. Analog ist ein score $P(x_{2i})$ für eine Beobachtung $x_{2i}$ der zweiten Gruppe definiert als Anzahl der Beobachtungen $x_{1j}$ der ersten Gruppe, die kleiner als $x_{2i}$ sind. Mit diesen scores errechnet sich die Teststatistik als

$$z = \left(\sum_i^{n_2} P(x_{2i}) - \sum_i^{n_1} P(x_{1i})\right) / (2\sqrt{V_1 + V_2 + \bar{P}_1\bar{P}_2})$$

wobei $\bar{P}_1$ und $\bar{P}_2$ die Mittelwerte der placement scores für die Gruppen 1 und 2, sowie $V_1$ und $V_2$ die Abweichungsquadratsummen sind. Diese Testgröße wird über die Normalverteilung auf Verschiedenheit von 0 überprüft. Der Test ist anschaulich beschrieben in der SAS-Dokumentation [183].

R: `NSM3`
SPSS: -

## 12. 12 Dunn

Dunns Test ähnelt dem Wilcoxon-Rangsummen-Test (bzw. dem Mann-Whitney-U-Test). Im Unterschied zu diesen verwendet er allerdings die Rangsummen des *joint ranking* (vgl. einführende Bemerkungen) aller k Gruppen. Üblicherweise wird er im Zusammenhang mit der Kruskal-Wallis-Varianzanalyse durchgeführt, bei der die gruppenweisen Rangsummen ohenhin errechnet werden. Da hier jeweils zwei Gruppen verglichen werden, ist eine anschließende $\alpha$-Adjustierung ratsam.

R: `dunn.test`
SPSS: -

# 13. Voraussetzungen

## 13. 1 Ungleiche $n_i$ und ungleiche Varianzen

Bei den meisten multiplen Mittelwertvergleichen werden gleiche Zellenbesetzungszahlen $n_i$ sowie gleiche Varianzen $s_i^2$ gefordert. Einer der wesentlichen Gründe dafür ist die Absicht, transitive Schlüsse hinsichtlich der Mittelwertrelationen ziehen zu können. D.h. wenn z.B. für 4 Mittelwerte folgende Relationen gelten: $\bar{x}_1 \leq \bar{x}_2 \leq \bar{x}_3 \leq \bar{x}_4$ , dann möchte man schließen: Wenn $\bar{x}_2 < \bar{x}_3$ dann gilt auch $\bar{x}_1 < \bar{x}_3$ , da ja $\bar{x}_1 \leq \bar{x}_2$ gilt. Dass sowohl ungleiche $n_i$ als auch heterogene Varianzen unter Umständen einen solchen Schluss nicht erlauben, sollen die folgenden beiden, zugegebenermaßen extremen, Beispiele demonstrieren. Im ersten sind die Varianzen gleich, aber die $n_i$ stark verschieden, im zweiten ist es umgekehrt:.

| | | Gruppe 1 | Gruppe 2 | Gruppe 3 | Gruppe 4 |
|---|---|---|---|---|---|
| | $\bar{x}_i$ | 39,5 | 40 | 42 | 43 |
| ungleiche $n_i$ | $n_i$ | 2 | 25 | 25 | 2 |
| | $s_i^2$ | 4 | 4 | 4 | 4 |
| ungleiche $s_i^2$ | $n_i$ | 10 | 10 | 10 | 10 |
| | $s_i^2$ | 25 | 4 | 4 | 25 |

In beiden Fällen unterscheiden sich $\bar{x}_2$ und $\bar{x}_3$ signifikant voneinander, aber nicht $\bar{x}_1$ und $\bar{x}_4$ , obwohl deren Mittelwertdifferenz weitaus größer ist. Der Grund: Im ersten Beispiel liegen für den Vergleich der beiden mittleren Gruppen zusammen 50 Werte vor, so dass eine relativ kleine Differenz bereits zu signifikanten Ergebnissen führt, während für den anderen Vergleich nur 4 Werte vorliegen und somit eine größere Unsicherheit, die zur Annahme der Hypothese gleicher Mittelwerte führt. Im zweiten Beispiel haben die Gruppen 2 und 3 relativ kleine Streuungen, so dass eine relativ kleine Differenz bereits zu signifikanten Ergebnissen führt, bzw. die Gruppen 1 und 4 relativ große Streuungen und somit eine größere Unsicherheit, die zur Annahme der Hypothese gleicher Mittelwerte führt.

Dies legt nahe, im Falle ungleicher $n_i$ nur dann das harmonische Mittel für $n$ zu verwenden (wie oben und in der Literatur häufig empfohlen), wenn die $n_i$ nicht allzu unterschiedlich sind. Das gleiche gilt auch für den Fall heterogener Varianzen. Moderate Unterschiede führen nicht zu Problemen.

## 13. 2 Parametrische Tests

Die Voraussetzungen der Tests hängen weitestgehend von der im Test verwendeten Prüfverteilung ab. Dies sind im Wesentlichen die studentized range-Verteilung sowie die F- bzw. t-Verteilung. Generell wird gefordert:

- Normalverteilung der Residuenvariable $e_{ij}$ im linearen Modell,
- Homogenität der Gruppenvarianzen ( $\sigma_1^2 = ... = \sigma_k^2$ ),
- Unabhängigkeit der beobachteten Werte.

Bei den multiplen Mittelwertvergleichen kommt noch hinzu, dass bei kleinen Stichproben

($n_i$<9) , insbesondere mit ungleichen $n_i$ ,Verletzungen der Voraussetzungen sich stärker auswirken als bei großen (vgl. Day und Quinn [101] ).

Verschiedene Autoren melden Vorbehalte, multiple Mittelwertvergleiche anzuwenden, wenn die Mittelwerte in Folge einer Kovarianzanalyse adjustiert worden sind. Diese erfüllen dann nicht mehr die Voraussetzung der Unabhängigkeit. Akzeptabel sind hier bestenfalls die Standardverfahren, wenn kein Zweifel an der Varianzhomogenität besteht (vgl. Day und Quinn [101] ).

### 13. 2. 1 studentized range-Verteilung

Die Tests von Tukey, Student-Newman-Keuls und von Duncan basieren alle auf der studentized range-Verteilung. Diese hat dieselben o.a. Voraussetzungen wie der F-Test der Varianzanalyse. Allerdings verfügt sie nicht über dieselbe Robustheit gegenüber Verletzungen der Voraussetzungen. Hier geht es im Wesentlichen um die Einhaltung des Fehlers 1. Art, wenn eine oder beide Voraussetzungen nicht erfüllt sind.

Die Einhaltung des $\alpha$-Risikos wurde von verschiedenen Autoren untersucht, u.a. als einer der Ersten von Ramseyer und Tcheng [175]. Solange nur eine der beiden Voraussetztungen erfüllt ist, wird die Fehlerrate weitgehend eingehalten. So zeigte sich hinsichtlich der Verteilungsform, dass lediglich bei gleichverteilten Daten eine marginale Vergrößerung des Fehlers 1. Art auftritt. Auch bei heterogenen Varianzen sind die Tests noch gültig. Leider verdoppelt sich allerdings die Fehlerrate, wenn beide Voraussetzungen gleichzeitig nicht erfüllt sind. Dies ist zu bedenken, wenn z.B. Variablen rangtransformiert werden, weil dadurch gleichverteilte Daten erzeugt werden. Schließlich ist noch zu erwähnen, dass bei größeren Stichproben ($n_i$>10) die Verletzung der Fehlerrate geringer ausfällt als bei kleinen.

Andere Autoren, z.B. Day und Quinn [101] zeigten, dass u.a. Tukeys Test allergisch auf heterogene Varianzen in der Art reagiert, dass in solchen Fällen die Fehlerrate 1. Art bis auf 10% ansteigen kann. Dies gilt in gleicher Weise auch für die anderen beiden Tests. Dagegen wirkt sich die Schiefe einer Verteilung, z.B. bei der log-Normalverteilung, weniger unangenehm aus. In diesem Fall wird empfohlen, die Daten vorher entsprechend logarithmisch zu transformieren, da dadurch insbesondere die Power erhöht wird.

Insbesondere für Tukeys Test gibt es eine Reihe von Alternativen für den Fall heterogener Varianzen, u.a. die Tests von Games & Howell sowie die C- und T3-Tests von Dunnett (vgl. Kapitel 6). Die Anpassung des Tukey-Tests für inhomogene Varianzen von Games & Howell (vgl. Kapitel 6.1) lässt sich wie oben erwähnt auch auf die anderen Tests, die die Fehlervarianz $MS_{Fehler}$ aus der Varianzanalyse verwenden, anwenden, so z.B. auf die Tests von Newman-Keuls und Duncan. Allerdings sind diese Varianten weder in SPSS noch in R verfügbar.

### 13. 2. 2 F- bzw. t-Verteilung

Fast alle übrigen Tests basieren auf der F- bzw. t-Verteilung. Für diese gelten dieselben Voraussetzungen und Robustheit-Eigenschaften wie bei der Varianzanalyse. Diese wurden ausführlich in Lüpsen (Kapitel 4.1. in [99]) besprochen. Hier noch einmal in Kürze die wichtigsten Eigenschaften:

- Je größer die Stichproben, desto weniger sind die Voraussetzungen noch relevant. Insbesondere ist nach dem zentralen Grenzwertsatz die Normalverteilungsvoraussetzung nur für kleinere Stichproben ($n_i$ < 50) bedeutsam.
- Bei annähernd gleichgroßen Stichprobenumfängen $n_i$ wirken sich weder nichtnormalverteil-

te Residuen noch inhomogene Varianzen störend aus.

- Stark heterogene Varianzen können das $\alpha$-Risiko vergrößern.

### 13. 2. 3 Weitere Hinweise

Noch ein paar Anmerkungen zu speziellen Tests:

- Mehrfaktorielle Versuchspläne:
  Für multiple Vergleiche eines Faktors werden häufig dessen Stufen über die der anderen Faktoren zusammengefasst (*Poolen*). Z.B. bei zwei Faktoren A und B werden zum Vergleich der Gruppen von A alle Werte, unabhängig von Faktor B, herangezogen. Dadurch können aber mehrgipflige Verteilungsformen entstehen, die für multiple Vergleiche ungeeignet sind. Dies ist noch gravierender in dem Fall, dass die Varianzen heterogen sind.
- Dunnetts Test mit einer Kontrollgruppe:
  Rudolph [167] hat durch Simulationen festgestellt, dass dieser robust gegen leichte Verletzungen der Normaltäts- und der Varianzhomogenitäts-Voraussetzungen ist. Für kleinere Stichproben ist er sogar nichtparametrischen Tests, wie dem von Steel, überlegen.

## 13. 3 Nichtparametrische Tests

Eine generelle Voraussetzung für nichtparametrische Mittelwertvergleiche, so u.a. auch für die Anova, betrifft die Verteilungsformen. Die nichtparametrischen Tests wie z.B. der U-Test oder die Kruskal-Wallis-Varianzanalyse können gelegentlich wie Omnibus-Tests reagieren, d.h. sie sprechen nicht nur auf Mittelwertunterschiede an, sondern auch auf Unterschiede anderer Verteilungsparameter wie Streuung oder Schiefe. Daher dürfen sich - streng genommen - die Verteilungen zwischen den einzelnen Gruppen nur hinsichtlich der Lage unterscheiden, während die Verteilungsform, z.B. Schiefe, Streuung etc in allen in etwa gleich ist.

Day & Quinn [101] haben gezeigt, dass heterogene Varianzen bei den nichtparametrischen Vergleichsverfahren kaum störenden Einfluss haben. Lediglich bei stark ungleichen Zellenbesetzungszahlen und zugleich stark ungleichen Varianzen (mehr als Faktor 4) kann die Fehlerrate verletzt werden, insbesondere bei Nemenyis Test.

# 14. Zur Auswahl eines Tests oder die Qual der Wahl

Die Auswahl eines der oben beschriebenen Verfahren hängt zwar davon ab, in wieweit die Voraussetzungen erfüllt sind, insbesondere gleiche $n_i$ und gleiche Varianzen, doch hier spielt ein anderer Begriff eine wichtigere Rolle: die Power, d.h. der Wunsch ungleiche Mittelwerte erkennen zu können (Ziel 1). Dabei spielt, wie unten dargelegt, eine entscheidende Rolle: Liegt bereits ein signifikantes Ergebnis einer Varianzanalyse vor oder nicht. Das Ergebnis legt nahe, in jedem Fall zunächst einen globalen Test der Mittelwerte durchzuführen. Dann steht nämlich eine Auswahl von liberalen Tests zur Verfügung, die eine deutlich größere Power besitzen.

Neben diesem Wunsch, Unterschiede erkennen zu können, die zwangsläufig auch zufällig zustande kommen können, kann natürlich auch das Ziel sein, einen eventuellen Mittelwertunterschied quasi hundertprozentig zu belegen (Ziel 2). In diesem Fall sind die konservativen Tests vorzuziehen, die das $\alpha$-Risiko unter allen Umständen korrekt einhalten.

## 14. 1 Mittelwertvergleiche mit oder ohne eine Varianzanalyse

In der Regel, wie auch eingangs angeführt, dienen multiple Mittelwertvergleiche zur Analyse eines signifikanten Haupteffekts im Anschluss an eine Varianzanalyse. Dies ist dann eine völlig andere Ausgangssituation, als wenn man die Mittelwertvergleiche *ohne* vorherige Varianzanalyse durchführt, da ja bereits mindestens ein signifikanter Unterschied besteht. Man kann sich dies leicht vor Augen führen: Testet man zunächst global die Mittelwerte über eine Varianzanalyse und im Falle einer Signifikanz anschließend über einen multiplen Mittelwertvergleich, so wird die Ausgangshypothese gleicher Mittelwerte, insbesondere der Vergleich des größten gegen den kleinsten Mittelwert, zweimal gestestet. Aber mit verschiedenen Tests, die nicht dasselbe Ergebnis bringen müssen. Das bedeutet, dass bei einem $\alpha$=0,05 der reale Fehler 1. Art bei ca. 0,025 liegt.

Bernhardson [176] hat diesen Unterschied untersucht. Dessen Ergebnisse und weitere Ausführungen zu diesem Thema sind u.a. bei Wilcox [102] nachzulesen. Diverse andere Autoren weisen auch dezent auf diese notwendige Unterscheidung hin. Zur Veranschaulichung sind in nachfolgender Tabelle die tatsächlichen Raten (auf Basis von Simulationen) bei verschiedenen Gruppenanzahlen *k* für den Fehler 1. Art für Fishers („ungeschützten“) LSD-, Tukeys HSD-, den Newman-Keuls- und den Duncan-Test aufgeführt, einmal mit $\alpha_1$, der Fehlerrate ohne eine vorherige signifikante Varianzanalyse und einmal mit $\alpha_2$, der Fehlerrate mit vorheriger signifikanter Varianzanalyse.

| | k=2 | | k=4 | | k=8 | | k=10 | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha_1$ | $\alpha_2$ | $\alpha_1$ | $\alpha_2$ | $\alpha_1$ | $\alpha_2$ | $\alpha_1$ | $\alpha_2$ |
| Fishers LSD | 0,055 | 0,055 | 0,211 | 0,051 | 0,50 | 0,050 | 0,591 | 0,050 |
| Tukeys HSD | 0,055 | 0,055 | 0,053 | 0,046 | 0,049 | 0,043 | 0,049 | 0,034 |
| Newman-Keuls | 0,055 | 0,055 | 0,053 | 0,046 | 0,049 | 0,034 | 0,049 | 0,034 |
| Duncan | 0,055 | 0,055 | 0,139 | 0,051 | 0,308 | 0,049 | 0,363 | 0,050 |

Wie man sieht, hält selbst der liberalste Test, Fishers LSD, die vorgegebene Fehlerrate von 5% im Falle eines signifikanten F-Tests für den untersuchten Haupteffekt weitgehend ein. (Fisher selbst hat seinen Test an ein signifikantes Anova-Ergebnis verknüpft.) Und bei Tukeys HSD liegen die Fehlerraten mit zunehmnder Gruppenzahl deutlich unter 5%, so dass der Test zu-

nehmend konservativ reagiert. Allerdings wurden die o.a. Ergebnisse unter der Annahme erzielt, dass alle Mittelwerte gleich sind.

Allerdings haben Day & Quinn [101] Simulationen für die Situation durchgeführt, dass nicht alle Mittelwerte gleich sind. Diese zeigten nur einen geringen Schutz der FWER, wenn die hier betrachteten liberaleren Tests nur im Anschluss an eine signifikante Anova durchgeführt werden. Auf der anderen Seite gibt es auch Autoren, wie z.B. Carmer und Walker [164], die ganz von α-Adjustierungen und FWER bzw. EER abraten und empfehlen, z.B. im Fall eines signifikanten F-Tests für die Einzelvergleiche den (ungeschützten) t-Test zu verwenden, insbesondere wenn man nur die Vergleiche durchführt, für die Hypothesen vorliegen.

Umgekehrt: Wenn man der Argumentation von Bernhardson und Wilcox nicht folgen und sich strikt an die Einhaltung der FWER bzw. EER halten will, so sollte man die multiplen Mittelwertvergleiche oder α-Adjustierungen auch dann durchführen, wenn die Anova keinen signifikanten F-Wert ergeben hat. Denn andernfalls verschenkt man etwas (der ohnehin geringen) Power. Es gibt einige Autoren, die ohnehin den zweiten Weg favorisieren, z.B. Zolman [177].

Dieser Unterschied wird allgemein ignoriert. Vielmehr werden in fast allen Veröffentlichungen zum Thema multipler Mittelwertvergleiche insbesondere die Tests von Duncan und Newman-Keuls verteufelt, weil sie das α-Risiko nicht einhielten. Allerdings ohne darauf hinzuweisen, dass das nur für den Fall einer fehlenden Anova gilt. Die Basis aller oben beschriebenen Tests ist die, dass *nicht* vorher eine Varianzanalyse durchgeführt wurde. Eine Herleitung der Tests unter der Prämisse eines signifikanten F-Tests ist nicht bekannt.

## 14. 2 Zur Stärke einiger Tests

Die Stärke (Power) eines Tests hängt von verschiedenen Größen ab, z.B. $n$ , $\alpha$ und den Mittelwerten selbst. Daher sind die nachfolgenden Werte nur relativ zu interpretieren. Die meisten Untersuchungen der Power beschränken sich auf die „klassischen“ Tests. So werden z.B. bei Einot & Gabriel [178] die Tests von Tukey, Newman-Keuls, Duncan, Ryan und Peritz verglichen sowie bei Liu [184] die Tests von Tukey, Newman-Keuls, Peritz und eigenen.

Abgesehen vom Duncan-Test hat der SNK-Test von allen Tests die größte Macht (ca. 80%), kann also am besten Unterschiede erkennen. Dafür muss man unter Umständen geringe Vergrößerungen des α-Risikos in Kauf nehmen. Von den übrigen Tests schneidet der von Ryan mit 60% am besten ab. Dagegen haben z.B. die Tests von Tukey eine Power von 53% und der von Scheffé nur 35% (vgl. Day & Quinn [101]). Das Verfahren von Peritz wird nur selten untersucht, es schneidet dann aber noch etwas besser als das von Ryan ab.

Ein Vergleich der o.a. kritischen Werte $c_d$ ermöglicht allerdings auch einen Vergleich der Verfahren:

Fishers LSD > Duncan > SNK > Tukey b > Peritz > Ryan > Tukey HSD > Scheffé

So steht man vor einem Dilemma: Auf der einen Seite möchte man den Fehler 1. Art vollkommen unter Kontrolle halten, aber viele der Tests, die das erfüllen, sind sehr konservativ, haben also eine geringe Power. Hierzu zählt z.B. Tukeys HSD oder für den Fall beliebiger Kontraste der Scheffé-Test oder der nichtparametrische Nemenyi-Test. Auf der anderen Seite möchte man auch Unterschiede erkennen, was bei der geringen Power relativ schwierig ist. Day & Quinn [101] schlagen als Ausweg vor, insbesondere bei explorativen Studien, von vorneherin das α größer zu wählen, etwa 0,10. Dies erachten die Autoren als besser als ein α von 0,05, das möglicherweise nicht konsequent eingehalten wird.

## 14. 3 Parametrisch im Fall von Varianzhomogenität

Im Folgenden wird angenommen, dass die in Kapitel 13.2 beschriebenen Voraussetzungen unter Berücksichtigung der dort aufgeführten Robustheitseigenschaften soweit erfüllt sind, dass die parametrischen Tests für homogene Varianzen durchgeführt werden können.

Zunächst wird vorausgesetzt, dass der F-Test ein signifikantes Resultat erbracht hat und Ziel 1 verfolgt wird. Dann sind der Newman-Keuls-Test sowie der Duncan-Test die erste Wahl. Eine Alternative bietet auch das Verfahren von Waller & Duncan, das einen Ausgleich zwischen den Fehler 1. und 2. Art bietet. Soll weniger liberal getestet werden, so sind die Verfahren von Ryan empfehlenswert. (Das Verfahren von Peritz steht allerdings noch nicht zur Verfügung.)

Wurde kein globaler Mittelwertvergleich durchgeführt oder wird Ziel 2 verfolgt, so ist Tukeys HSD-Test die erste Wahl. Aber auch hier sind die Verfahren von Ryan empfehlenswert.

## 14. 4 Parametrisch im Fall von Varianzheterogenität

Zunächst wird wieder vorausgesetzt, dass der F-Test ein signifikantes Resultat erbracht hat und Ziel 1 verfolgt wird. Dann sind zwar der Newman-Keuls-Test sowie der Duncan-Test in der Version für ungleiche Varianzen (vgl. Kapitel 6.1) eine gute Wahl. Doch diese stehen derzeit in den Softwarepaketen nicht zur Verfügung.

In allen Fällen sind der Test von Games & Howell sowie die C- und T3-Tests von Dunnett angebracht. Hiervon besitzt der von Games & Howell die größte Power und sollte die erste Wahl sein. Von den Tests von Dunnett ist für kleine bis mittlere $n_i$ (etwa $n_i \leq 40$) der T3-Test gegenüber dem o.a. C-Test überlegen. Für größere $n_i$ ist dagegen der C-Test vorzuziehen. Sollen nicht nur paarweise Vergleiche, sondern Kontraste getestet werden, so ist das Verfahren von Brown & Forsythe empfehlenswert. Einen Vergleich der Verfahren für inhomogene Varianzen, insbesondere hinsichtlich der Power, bietet Tamhane [111].

Da aber einige von diesen Verfahren nicht in R und/oder SPSS zur Verfügung stehen, bieten robuste paarweise t-Tests für inhomogene Varianzen, wie sie in R und SPSS in den Standardroutinen für t-Tests angeboten werden, zusammen mit einer guten p-Wert-Adjustierung (vgl. Abschnitt 14.6) eine gute Alternative.

## 14. 5 Nichtparametrisch

Sowohl der Nemenyi- als auch der Dwass, Steel, Critchlow-Fligner-Test gelten als extrem konservativ (vgl. Day & Quinn [101]). Mehr Power bieten z.B. das Ryan-Verfahren mit dem Kruskal-Wallis-Test bzw. der Friedman-Analyse, das als Campbell & Skillings-Verfahren bekannt ist und sogar von SPSS angeboten wird. Besser sind auch die Paarvergleiche von Gao, die problemlos auf ordinale Merkmale angewandt werden können. Gao et al [171] weisen darauf hin, dass viele Verfahren für Daten mit Bindungen, wie sie typischerweise bei ordinalen Variablen auftreten, ungeeignet sind. Dies gilt allerdings weniger für die Verfahren von Ryan und Campbell & Skillings.

Wie auch in den Beispielen in Kapitel 15 gezeigt wird, schneiden auch einige Paarvergleiche, wie der Wilcoxon-Rangsummen-Test oder der Fligner-Policello-Test, mit $\alpha$-Adjustierungen deutlich besser ab. So kann z.B. die Adjustierung von Li mit der Hand gerechnet werden.

## 14. 6 α-Adjustierungen

Wenn alle Stricke reißen, gibt es noch die α-Adjustierungen. Tatsächlich lassen sich diese ja auf einen beliebigen Test anwenden, mittels dem Vergleiche durchgeführt werden, auf einen t-Test, einen U-Test oder was auch immer. Doch was ist hier erste Wahl, d.h. welche Adjustierungen sind nicht allzu konservativ? Hier hängt die Wahl allerdings von dem Angebot des jeweiligen Statistikprogramms ab.

Erste Wahl sind mit Sicherheit die Verfahren von Shaffer, Hommel, Rom, Li sowie Benjamini & Hochberg, wovon lediglich die letzten beiden auch (mit ein wenig Rechnerei) per Hand durchgeführt werden können. Danach kommen erst die von Holm und Hochberg, die allerdings bequem auch per Hand zu erledigen sind. Von Bonferroni und Sikdak ist in jedem Fall abzuraten, da sie extrem konservativ sind.

Doch noch ein Wort für α-Adjustierungen: Sollen nicht alle Paare sondern nur eine Auswahl davon auf Gleichheit getestet werden, so können die α-Korrekturen deutlich besser abschneiden als z.B. der Tukey HSD-Test.

# 15. Anwendungen mit R

Bei R sind „naturgemäß“ fast alle in den vorigen Kapiteln besprochenen Verfahren verfügbar. Natürlich in zu installierenden Paketen, von denen nur die berücksichtigt werden, die über `cran.r-project.org` zur Verfügung gestellt werden. Und das sind viele. In den nachfolgenden Kapiteln werden die Verfahren „paketeweise“ vorgestellt. Um die Suche nach Beispielen für bestimmte Verfahren zu erleichtern, nachfolgend eine Übersicht, aus der die Beispielnummern, die mit den Kapitelnummern korrespondieren, zu entnehmen sind:

| Datensatz | mydata2 | mydata3 | mydata10 | winer518 | mydata9 |
|---|---|---|---|---|---|
| Stichproben | unabh. | unabh. | unabh. | abh. | abh. |
| Varianzen | homogen | heterogen | homogen | homogen | heterogen |
| Tukey HSD | R-1.3 | | | | |
| Newman-Keuls | R-1.1 | | | R-2.1 | |
| Dunnett C / Dunnetts stepwise | R-1.5 | R-1.4 | | | |
| Ryans REGWQ | | | R-1.7 | | |
| Scott & Knox | R-1.6 | | | | |
| paarweise Vergleiche - α-Korrektur t-Test / U-Test / Wilcoxon-Test / Fligner-Policello | | R-1.2b<br>R-6.1a<br>R-3.4b | R-1.2a<br>R-5.2 | | R-6.1b<br>R-6.2 |
| Kontraste | | R-3.2 | | R-2.2 | |
| Kruskal-Wallis - paarweise Vergleiche | R-3.1a | | | | |
| Friedman - paarweise Vergleiche / Quade | | | | R-4.1 | R-4.2<br>R-3.9 |
| Nemenyi | | | R-3.3 | | R-4.3 |
| Dwass-Steel-Critchlow-Fligner / Gao | R-3.2 | | R-3.4 | | |
| Rangtransformation - LSD / HSD | | R-3.1b | R-3.3 | | |

## 15. 1 parametrische Vergleiche - unabhängige Stichproben

### 15. 1. 1 agricolae

In dem Paket `agricolae` sind die klassischen Verfahren (u.a. die des Kapitel 4) verfügbar, die allerdings alle Varianzhomogenität voraussetzen (siehe unten angeführte Tabelle)

Der Aufruf der einzelnen Funktionen ist weitgehend identisch. In der aktuellen Version (von 9/2014) sind zwei Eingabevarianten möglich:

- *Funktion* (*Anova-Objekt*, „*Faktor*“, alpha=*Wert*, group=T/F, console=T/F)
- *Funktion* (*abh.Variable, Faktor, $df_{Fehler}$, $MS_{Fehler}$*, alpha=*Wert*, group=T/F, console=T/F, p.adj=“*Kürzel*“)

Hierbei sind:

- *funktion*: eine der u.a. Funktionsnamen
- *Anova-Objekt*: das Ergebnis einer Varianzanalyse (Objekttyp aov oder lm) , darf mehrfaktoriell sein
- *Faktor*: der Name des zu untersuchenden Faktors (in "...")
- *abh.Variable*: der Name der abhängigen Variablen (Kriteriumsvariablen)
- group: bei T erfolgt die Ausgabe von homogenen Untergruppen, bei F erfolgt die Ausgabe paarweiser Vergleiche (vgl. Kapitel 1.7)
- console: bei T werden die Ergebnisse angezeigt.
- $df_{Fehler}$: Anzahl der FG der Fehlervarianz
- $MS_{Fehler}$: Fehlervarianz (vgl. Kapitel 4)
- p.adj: (nur bei LSD.test und kruskal): α-Adjustierung nach einer der folgenden Methoden, unter Angabe des Kürzels: "none" (keine Korrektur), "holm", "hochberg", "bonferroni", "BH" (Benjamini-Hochberg), "BY" (Benjamini-Yekutieli), "fdr" (Benjamini-Hochberg), "hommel" .

Neben der üblichen Dokumentation gibt es auch ein Tutorial für agricolae [191]..

| Funktionsname | Test |
|---|---|
| duncan.test | Duncan-Test |
| HSD.test | Tukeys HSD-Test |
| LSD.test | Fishers LSD-Test |
| scheffe.test | Scheffé-Test |
| SNK.test | Student-Newman-Keuls-Test |
| waerden.test | v.d.Waerden-Test |
| waller.test | Waller-Duncan-Test |
| kruskal | Kruskal-Wallis-Test mit paarweisen Vergleichen |
| friedman | Friedman-Varianzanalyse mit paarweisen Vergleichen |

**Beispiel R-1.1:**

Zunächst soll ein Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen gerechnet werden. Dazu werden die Beispieldaten 2 (mydata2) verwendet (vgl. Kapitel 4 in [99]). Das Ergebnis der Varianzanalyse enthält Tabelle 4-2 (in [99]). Hier soll jetzt mittels des Newman-Keuls-Test der Effekt des Faktors drugs im Detail untersucht werden.

```
options (contrasts=c("contr.sum","contr.poly"))
mydata2 <-within(mydata2,{drugs<-factor(drugs);
                 group<-factor(group)})
aov2    <- aov(x~group*drugs,mydata2)
SNK.test(aov2,"drugs", group=T, console=T)
```

Nachfolgend die Ausgabe, hier mit homogenen Untergruppen:

```
Study: aov2 ~ "drugs"

Student Newman Keuls Test for x

Mean Square Error:  1.636667

drugs,  means

          x      std r Min Max
1 4.000000 1.414214 7   2   6
2 5.111111 1.536591 9   3   7
3 5.625000 1.685018 8   3   8
4 7.111111 1.536591 9   5   9

alpha: 0.05 ; Df Error: 25

Critical Range
       2        3        4
1.304272 1.577402 1.741938

Harmonic Mean of Cell Sizes  8.161943

Different value for each comparison
Means with the same letter are not significantly different.

Groups, Treatments and means
a        4       7.111
b        3       5.625
bc       2       5.111
c        1       4
```

Die letzten 4 Zeilen enthalten das wesentliche Ergebnis: Unter `Groups` sieht man die Buchstaben a, b und c. Jeder Buchstabe entspricht einer homogenen Untergruppe. Daraus ergibt sich: Die Behandlungen 3 und 2 unterscheiden sich nicht (Untergruppe b). Und die Behandlungen 2 und 1 unterscheiden sich nicht (Untergruppe c).

Stellt man nun eine Dreiecksmatrix auf, in der sowohl die Zeilen als auch die Spalten Mittelwerte entsprechen, werden die o.a. nicht signifkanten Vergleiche mit einem - gekennzeichnet. Die verbliebenen Felder entsprechen signifikanten Vergleichen und werden mit x markiert:

| | 4,00 | 5,11 | 5,63 | 7,11 |
|---|---|---|---|---|
| 4,00 | | - | **x** | **x** |
| 5,11 | | | - | **x** |
| 5,63 | | | | **x** |

Alternativ hier der Aufruf mit der Übergabe der Vektoren der abhängigen und der Gruppenvariablen sowie mit der Ausgabe der paarweisen Vergleiche, die hier nur partiell wiedergegeben wird:

```
options (contrasts=c("contr.sum","contr.poly"))
mydata2  <-within(mydata2,{drugs<-factor(drugs);
               group<-factor(group)})
aov2     <- aov(x~group*drugs,mydata2)
dffehler <- df.residual(aov2)
msfehler <- deviance(aov2)/dffehler
SNK.test(mydata2$x, mydata2$drugs, dffehler, msfehler,
   group=F, console=T)
```

```
    Difference   pvalue sig.       LCL         UCL
1-2 -1.1111111 0.091592    . -2.415383  0.19316114
1-3 -1.6250000 0.042571    * -3.202402 -0.04759781
1-4 -3.1111111 0.000258  *** -4.853049 -1.36917340
2-3 -0.5138889 0.424756      -1.818161  0.79038336
2-4 -2.0000000 0.011079    * -3.577402 -0.42259781
3-4 -1.4861111 0.027169    * -2.790383 -0.18183886
```

Hier werden die p-Werte für jeden Vergleich angegeben sowie das 5%-Konfidenzintervall.

Ein Beispiel für die Anwendung bei Messwiederholungen ist in Kapitel 15.2.1 zu finden.

### 15. 1. 2 multcomp

Zum multiplen Mittelwertvergleich bietet das Paket `multcomp` zum einen die Möglichkeit, Kontraste zu bilden und zu testen, zum anderen stehen eine Reihe von α-Adjustierungen für multiple Tests dieser Kontraste zur Verfügung. Im Rahmen dieser Möglichkeiten lassen sich allerdings auch die klassischen Verfahren wie Tukey HSD und Dunnetts Kontrollgruppenvergleich durchführen und sogar für den Fall inhomogener Varianzen.

Folgende α-Adjustierungen werden angeboten:

| adjusted-Option | α-Adjustierung |
|---|---|
| `bonferroni` | Bonferroni |
| `holm` | Holm (step-down) |
| `hochberg` | Hochberg (step-down) |
| `hommel` | Hommel |
| `BH` | Benjamini-Hochberg |
| `BY` | Benjamini & Yekutieli |
| `Westfall` | |
| `free` | Westfalls free step-down |
| `Shaffer` | Shaffer S1 |
| `single-step` | Adjustierung über multivariate t-Verteilung |
| `none` | keine Adjustierung |

Das `single-step`-Verfahren (der default-Wert) ist im eigentlichen Sinn keine α-Adjustierung, sondern vielmehr eine Transformation eines t-Wertes in einen multivariaten t-Wert, der der studentized range-Verteilung entspricht. Dadurch erhält man mit diesem Adjustierungstyp im Falle von Tukey-Kontrasten Tukeys HSD-Test und im Falle von Dunnett-Kontrasten den ent-

sprechenden Vergleich mit einer Kontrollgruppe.

Auf folgende Merkmale sei hingewiesen:

- Für den Fall inhomogener Gruppenvarianzen kann für die Schätzung der Kontraste und damit für die Tests eine *Sandwich*-Schätzung (vgl. Kapitel 6.5) vorgenommen werden.
- Mehrfaktorielle Versuchspläne werden zwar berücksichtigt, beeinträchtigt allerdings die p-Wert-Adjustierung und sollten vermieden werden. Allerdings sind Analysen der „simple effects“ (Vergleiche eines Faktors für die Stufen des anderen Faktors) möglich.
- Allerdings werden keine Versuchspläne mit Messwiederholungen unterstützt.
- Bei der Standard-Adjustierung (`single-step`) werden Zufallszahlen erzeugt. Das führt dazu, dass bei wiederholten Aufrufen der Funktion leicht unterschiedliche Ergebnisse erzeugt werden.

Zunächst müssen für den zu untersuchenden Faktor Kontraste definiert werden. Dazu stehen eine Reihe von Standardkontraste zur Verfügung, u.a.

| Kontrast-Option | Anzahl Kontraste | Kontrast |
|---|---|---|
| `Tukey` | k(k-1) | paarweise Vergleiche |
| `Dunnett` | (k-1) | Vergleich mit einer Kontrollgruppe (1. Gruppe) |
| `Sequen` | (k-1) | 1-2, 2-3, 3-4,... |
| `GrandMean` | k | Differenz zum geschätzten Gesamtmittelwert |
| `AVE` | k | Differenz zum Gesamtmittelwert |

Es können aber auch individuelle Kontraste definiert werden, wie dies bereits am Ende von Kapitel 3.4 vorgestellt wurde.

Der Standardaufruf für paarweise Vergleiche mit einer α-Adjustierung erfolgt über die Funktion `glht`, die als Eingabe das Ergebnis einer Varianzanalyse erhält. Darin wird über `mcp` (*multiple comparison procedure*) der Faktor (aus dem Anova-Modell) und der Kontrastname spezifiziert. Im anschließenden `summary` kann die α-Adjustierung über den Parameter `test = adjusted("...")` festgelegt werden.

```
aov1 <- aov(Anova-Modell)
aov1_glht <- glht(aov1, linfct = mcp(drugs="Tukey")
summary(aov1_glht, test=adjusted(„Adjustierung“)
```

**Beispiel R-1.2a:**

Zunächst soll ein Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen gerechnet werden. Dazu werden die Beispieldaten 10 (`mydata10`) verwendet. Hier soll jetzt mittels paarweiser Vergleiche, genauer mittels Tukey-Kontrasten (vgl. Kapitel 7.3), und Hommels α-Adjustierung der Effekt des Faktors `drugs` im Detail untersucht werden. Nachfolgend Eingabe und Ausgabe:

```
options (contrasts=c("contr.sum","contr.poly"))
mydata10 <-within(mydata2, {Gruppe<-factor(Gruppe))
aov10    <- aov(x~Gruppe, mydata10)
glht10 <- glht(aov10, linfct = mcp(drugs = "Tukey"))
summary(glht10, test = adjusted("hommel"))
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = x ~ Gruppe, data = mydata10)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0   1.6667     1.0360   1.609 0.240100
3 - 1 == 0  -1.9667     1.0866  -1.810 0.240100
4 - 1 == 0   3.8333     1.0360   3.700 0.008974 **
5 - 1 == 0   6.2619     0.9983   6.272 1.06e-05 ***
6 - 1 == 0   3.2333     1.0866   2.976 0.040900 *
3 - 2 == 0  -3.6333     1.0866  -3.344 0.020628 *
4 - 2 == 0   2.1667     1.0360   2.091 0.197521
5 - 2 == 0   4.5952     0.9983   4.603 0.000840 ***
6 - 2 == 0   1.5667     1.0866   1.442 0.320133
4 - 3 == 0   5.8000     1.0866   5.338 0.000129 ***
5 - 3 == 0   8.2286     1.0507   7.831 1.84e-07 ***
6 - 3 == 0   5.2000     1.1349   4.582 0.000891 ***
5 - 4 == 0   2.4286     0.9983   2.433 0.128351
6 - 4 == 0  -0.6000     1.0866  -0.552 0.585052
6 - 5 == 0  -3.0286     1.0507  -2.882 0.051509 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- hommel method)
```

In der Ergebnismatrix werden pro Zeile links die beiden zu vergleichenden Gruppennummern ausgewiesen, bzw. Labels, sofern für den Faktor vereinbart, gefolgt von dem Hypothesenwert (hier 0), der Schätzung des Kontrasts C, dem Standardfehler, dem daraus resultierenden t-Wert sowie dem adjustierten p-Wert.

**Beispiel R-1.2b:**

Hier wird ein Beispiel für den Fall unabhängiger Stichproben mit inhomogenen Varianzen gerechnet. Dazu werden die Beispieldaten 3 (`mydata3`) verwendet (vgl. Kapitel 4 in [99]). Bei der Varianzanalyse hatte sich gezeigt, dass die Varianzen heterogen sind. (Für den Levene-Test wird ein p=0,012 ausgewiesen (vgl. Kapitel 4.3.3 in [99]). Bei dieser Gelegenheit sollen auch benutzerdefinierte Kontraste demonstriert werden.

Es wird der Faktor `dosis` untersucht, der bei der Anova als signifikant ausgewiesen wurde (je nach Verfahren etwa p=0,046 ). Eine Interaktion mit dem Faktor `Gruppe` bestand nicht. Die angestrebten Vergleiche sind: Dosis 1-2, 1-3, 1-4 sowie 2-(3,4). Die Kontraste werden mittels `rbind` erstellt und bilden das Objekt `contr`. Da keine signifkante Interaktion vorliegt, werden die Vergleiche auf Basis einer 1-faktoriellen Varianazanalyse durchgeführt. Wegen der Varianzinhomogenität wird die sandwich-Schätzung gewählt: `vcov=sandwich` und zur Adjustierung das Verfahren von Westfall. Nachfolgend Eingabe und Ausgabe:

```
contr <- rbind( "Dosis 1-2" = c(1,-1,0,0),
                "Dosis 1-3" = c(1,0,-1,0),
                "Dosis 1-4" = c(1,0,0,-1),
                "Dosis 2-(3,4)" = c(0,2,-1,-1))

aov2 <- aov(x~dosis, mydata3)
glht_aov2 <- glht(aov2, linfct=contr, vcov=sandwich)
summary(glht_aov2, test=adjusted(„Westfall"))
```

```
Linear Hypotheses:
                    Estimate Std. Error t value Pr(>|t|)
Dosis 1-2 == 0       -1.3333     0.5048  -2.641   0.0258 *
Dosis 1-3 == 0       -2.5833     1.0281  -2.513   0.0320 *
Dosis 1-4 == 0       -3.2778     1.0399  -3.152   0.0133 *
Dosis 2-(3,4) == 0   -3.1944     1.4209  -2.248   0.0323 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- Westfall method)
```

Hier sind alle 4 angeforderten Kontraste signifikant. Eine grafische Darstellung hierzu ist in Kapitel 15.9 zu finden.

### 15. 1. 3 stats: TukeyHSD

Tukeys HSD-Test (für unabhä#ngige Stichproben) steht über das Standardpaket `stats` immer über die Funktion `TukeyHSD` zur Verfügung. Bei mehrfaktoriellen Analysen wird der entsprechende Fehlerterm berücksichtigt. Der Standardaufruf lautet:

```
TukeyHSD(aov(Anova-Modell), „Faktor")
```

Die Angabe des Faktors kann natürlich bei einer 1-faktoriellen Analyse entfallen. Fehlt die Angabe bei einer mehrfaktoriellen Analyse, so wird zum einen der Tukey-Test für jeden Faktor durchgeführt und darüber hinaus für alle Zellen aller Interaktionen.

**Beispiel R-1.3:**

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Das Ergebnis der Varianzanalyse enthält Tabelle 4-2 (in [99]). Nachfolgend Aufruf und Ausgabe, die mit der von `glht` aus o.a. `multcomp`-Paket weitgehend identisch ist:

```
options (contrasts=c("contr.sum","contr.poly"))
aov2    <- aov(x~group*drugs,mydata2)
TukeyHSD(aov2,"drugs")
```

```
         diff        lwr      upr     p adj
2-1 1.1289242 -0.6444653 2.902314 0.3197979
3-1 1.7051587 -0.1160764 3.526394 0.0723640
4-1 3.1289242  1.3555347 4.902314 0.0003003
3-2 0.5762346 -1.1336730 2.286142 0.7907677
4-2 2.0000000  0.3411460 3.658854 0.0138862
4-3 1.4237654 -0.2861421 3.133673 0.1273813
```

Die Ausgabe enthält pro Zeile die Kodierungen der beiden zu vergleichenden Gruppen, die Mittelwertdifferenz, das 95%-Konfidenzintervall sowie den p-Wert. Eine grafische Darstellung hierzu ist in Kapitel 15.9 zu finden.

### 15. 1. 4 DTK: DTK.test

`DTK.test` führt Dunnetts C-Test durch, eine Verallgemeinerung von Tukeys HSD für heterogene Varianzen. Die Eingabe verlangt die Rohwerte anstatt eines aov-Objekts:

```
DTK.test(y-Variable, Faktor, alpha)
```

**Beispiel R-1.4:**

Hier wird das o.a. Beispiel 4 für den Fall unabhängiger Stichproben mit inhomogenen Varianzen wieder aufgegriffen (Beispieldaten 3 (`mydata3`), (vgl. Kapitel 4 in [99]). Nachfolgend Ein- und Ausgabe:

```
with(mydata3, DTK.test(x,dosis,0.05))
```

```
          Diff   Lower CI Upper CI
2-1 1.3333333 -0.6128042 3.279471
3-1 2.5833333 -1.3393918 6.506058
4-1 3.2777778 -0.6642344 7.219790
3-2 1.2500000 -2.2364898 4.736490
4-2 1.9444444 -1.5616883 5.450577
4-3 0.6944444 -3.9551577 5.344047
```

Auf den ersten Blick fehlen in der Ausgabe die p-Werte und damit die Signifikanzen. Diese sind aber aus den Konfidenzintervallen abzulesen: Enthält das Intervall die „0“ - und das ist hier bei allen Intervallen der Fall, so wird die Nullhypothese angenommen und es liegt kein signifikanter Unterschied vor. Ein Vergleich mit den Ergebnissen aus Beispiel 4 zeigt, dass Dunnetts C-Test schwächer abschneidet als die Sandwich-Schätzung mit der α-Adjustierung von Westfall.

### 15. 1. 5 DunnettsTests

Das Paket `DunnettsTests` bietet die beiden schrittweisen Varianten von Dunnetts Vergleich mit einer Kontrollgruppe. Die Benutzung erfordert jedoch ein wenig Vorbereitung, da die Testgrößen $t_i$ (vgl. Kapitel 9.2) vorher selbst errechnet werden müssen:

$$t_i = \frac{\overline{x_i} - \overline{x_1}}{\sqrt{2MS_{Fehler}/n}}$$

Dies beinhaltet:

- Berechnung der Mittelwertdifferenzen `mdiff` zur Kontrollgruppe,
- Berechnung des harmonischen Mittels `nhm` der $n_i$,
- Ermittlung der Fehlervarianz $MS_{Fehler}$ aus der Anova-Tabelle.

Die adjustierten p-Werte liefern dann

- `qvSDDT (Testgrößen, Freiheitsgrade)` für den step-down-Test
- `qvSUDT (Testgrößen, Freiheitsgrade)` für den step-up-Test

**Beispiel R-1.5:**

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Als Kontrollgruppe wird hier die erste Gruppe (drug 1) angenommen.

Die Kommandos sowie das Ergebnis:

```
aov2     <- aov(x~group*drugs, mydata2)
mdiff    <- with(mydata2, tapply(x,drugs,mean)-mean(x[drugs==1]))
ni       <- with(mydata2, table(drugs))
nhm      <- 1/mean(1/ni)
msfehler <- deviance(aov2)/df.residual(aov2)
twerte   <- mdiff/sqrt(2*msfehler/nhm)
qvSDDT(twerte[-1],df=df.residual(aov2))
```

```
$`ordered test statistics`
      H3       H2       H1
4.912665 2.565990 1.754523

$`Adjusted P-values of ordered test statistics`
[1] 0.000 0.015 0.046
```

Hierbei beziehen sich `H1` auf die erste Differenz, also Gruppe 2-1, `H2` auf die zweite usw. Das Ergebnis weist alle drei Vergleiche als signifikant aus. Nachfolgend zum Vergleich die Ergebnisse des single-step-Dunnett-Tests (erstellt mit `glht` aus `multcomp`):

```
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0   1.1111     0.7810   1.423  0.35336
3 - 1 == 0   1.6250     0.8021   2.026  0.12410
4 - 1 == 0   3.1111     0.7810   3.983  0.00121 **
```

### 15. 1. 6 ExpDes

In dem Paket `ExpDes` sind einige der klassischen Verfahren (u.a. die des Kapitel 4) verfügbar, die allerdings alle Varianzhomogenität voraussetzen:

| Funktionsname | Test |
|---|---|
| `duncan` | Duncan-Test |
| `snk` | Student-Newman-Keuls-Test |
| `tukey` | Tukeys HSD-Test |
| `lsd` | Fishers LSD-Test |
| `lsdb` | Fishers LSD-Test mit Bonferroni-Adjustierung |
| `scottknott` | Scott-Knott-Verfahren |

Der Aufruf der einzelnen Funktionen ist weitgehend identisch:

*Funktion* (*abh.Variable*, *Faktor*, $df_{Fehler}$, $SS_{Fehler}$, alpha=*Wert*)

Hierbei sind:

- *Funktion*: eine der o.a. Funktionsnamen
- *abh.Variable*: der Name der abhängigen Variablen (Kriteriumsvariablen)
- *Faktor*: der Name des zu untersuchenden Faktors
- $SS_{Fehler}$: Streuungsquadratsumme der Fehlervarianz

- $df_{Fehler}$: Anzahl der FG der Fehlervarianz
- `alpha`: Irrumswahrscheinlichkeit $\alpha$

Da die Fehlervarianz vorgegeben werden muss, können die Tests auch bei mehrfaktoriellen Versuchsplänen sowie bei Messwiederholungen durchgeführt werden. Allerdings bietet `ExpDes` gegenüber dem umfangreicheren und komfortableren Paket `agricolae` lediglich das Scott-Knott-Verfahren zusätzlich.

**Beispiel R-1.6:**

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Die Ein- und Ausgabe des Verfahrens von Scott & Knott:

```
aov2     <- aov(x~group*drugs,mydata2)
dffehler <- df.residual(aov2)
ssfehler <- deviance(aov2)
scottknott(mydata2$x, mydata2$drugs, dffehler, ssfehler, 0.05)
```

```
Scott-Knott test
---------------------------------------
  Groups Treatments    Means
1      a          4 7.111111
2      b          3 5.625000
3      b          2 5.111111
4      b          1 4.000000
---------------------------------------
```

Die Ausgabe ähnelt der des Pakets `agricolae`. Allerdings ist bei dem Verfahren von Scott & Knott zu beachten, dass die Partition der Gruppen immer disjunkt ist. Hier unterscheidet sich also Gruppe 1 (markiert mit „a“) von den Gruppen 2, 3 und 4 (markiert mit „b“).

### 15. 1. 7 mutoss

Das Paket `mutoss` (derzeit noch einem „experimentellem“ Stadium) bietet primär Methoden zur p-Wert-Adjustierung (vgl. Kapitel 2), darüber hinaus aber auch die Mittelwertvergleiche von Newman-Keuls und das REGWQ-Verfahren von Ryan (vgl. Kapitel 5.2). Beide Funktionen verlangen als Eingabe ein Anova-Modell der Form *`y~Faktor`* , allerdings mit nur einem Faktor, erlauben allerdings optional die Vorgabe von $MS_{Fehler}$ und den dazugehörigen FG, so dass diese auch bei mehrfaktoriellen Analysen oder solchen mit Messwiederholungen anwendbar sind.

```
regwq (Anova-Modell, Dataframe, alpha=Wert, MSE, df)
snk   (Anova-Modell, Dataframe, alpha=Wert, MSE, df)
```

Vor der Installation von `mutoss` ist die Installation des Pakets `multtest` aus dem Bioconductor-Projekt mittels der beiden folgenden Anweisungen erforderlich:

```
source("http://bioconductor.org/biocLite.R")
biocLite("multtest")
```

**Beispiel R-1.7:**

Für die Beispieldaten 10 (`mydata10`) wird der Test von Ryan, Einot, Gabriel und Welsch in der Variante mit der studentized range-Verteilung durchgeführt. Varianzhomogenität ist gegeben. Die hier implementierte Testvariante ist auch anwendbar für ungleiche $n_i$. Ein- und Ausgabe:

```
regwq (x~Gruppe, mydata10, alpha=0.05)
```

```
#----REGWQ - Ryan / Einot and Gabriel / Welsch  test procedure

Number of hyp.:  15
Number of rej.:  9
  rejected pValues adjPValues
1        1       0          0
2        3       0          0
3        5  0.0047     0.0047
4        7  0.0063     0.0063
5        8  0.0156     0.0156
6       10  0.0195     0.0195
7        2   1e-04      1e-04
8        4   4e-04      4e-04
9        6   4e-04      4e-04
```

Es wird angezeigt, dass von den 15 Paarvergleichen 9 als unterschiedlich anzusehen sind, wobei die 9 Hypothesennummern angegeben sind. Aber welche Hypothesen sind das? Dies ist nur mit einem Teil der übrigen Ausgabe herauszufinden

```
$rejected
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
FALSE FALSE
[14] FALSE FALSE

$confIntervals
      [,1] [,2] [,3]
5-3 8.2286   NA   NA
4-3 5.8000   NA   NA
5-1 6.2619   NA   NA
6-3 5.2000   NA   NA
4-1 3.8333   NA   NA
5-2 4.5952   NA   NA
2-3 3.6333   NA   NA
6-1 3.2333   NA   NA
4-2 2.1667   NA   NA
5-6 3.0286   NA   NA
1-3 1.9667   NA   NA
2-1 1.6667   NA   NA
6-2 1.5667   NA   NA
4-6 0.6000   NA   NA
5-4 2.4286   NA   NA
```

Unter `$rejected` werden für 15 Hypothesen angezeigt, ob sie abgelehnt werden (`TRUE`) oder angenommen werden (`FALSE`). (Das `T/F` bezieht sich auf die Vorgabe „rejected“). Dieses deckt sich mit der Angabe in dem ersten Teil der Ausgabe, in der die 9 Nummern der abgelehnten Hypothesen vermerkt wurden. Worauf sich die 15 Hypothesen beziehen, ist aus dem Teil `$confIntervals` zu entnehmen. Dort sind alle 15 untereinander aufgelistet und sich als durchnummeriert vorzustellen. In der ersten Spalte ist vermerkt, welchen Vergleich die jeweilige Hypothese beinhaltet. Also $H_1$ vergleicht 5-3, $H_2$ vergleicht 4-3, usw. Also werden abgelehnt (in der Reihenfolge des ersten Teils der Ausgabe): $H_1$ (5-3), $H_3$ (5-1), $H_5$ (4-1) usw.

## 15. 2 parametrische Vergleiche - abhängige Stichproben

In R stehen für Messwiederholungsfaktoren keine speziellen multiplen Mittelwertvergleiche zur Verfügung. Es gibt folgende Möglichkeiten:

- Durchführung der klassischen Tests (u.a. der aus Kapitel 4). Dies ist möglich bei Funktionen, die die Vorgabe von $MS_{Fehler}$ bzw. $df_{Fehler}$ erlauben, die dann der Varianzanalyse mit Messwiederholungen entnommen werden. Hierfür stehen die Pakete `agricolae` , `ExpDes` und `mutoss` zur Verfügung.
- orthogonale Kontraste, z.B. mittels Standardkontrast „wiederholt“ (repeated). Dies ist z.B. mit dem Paket `multcpmp` möglich.
- paarweise Vergleiche mit α-Adjustierungen (vgl. Kapitel 15.5 und 15.6).

### 15. 2. 1 agricolae

Das Paket `agricolae` wurde bereits in 15.1.1 vorgestellt. Hier nun ein Beispiel für die Anwendung im Fall von Messwiederholungen.

**Beispiel R-2.1:**

Hierzu wird der Beispieldatensatz 5 (`winer518`) verwendet (vgl. Kapitel 5 in [99]). Für den Faktor `Zeit` wurde dort ein signifikanter Haupteffekt (p<0,001) nachgewiesen (vgl. Tabelle 6-1 in [99]). Dieser soll nun näher untersucht werden. Dies soll mit dem Newman-Keuls-Test erfolgen, wobei, wie in Kapitel 10 beschrieben, die Fehlervarianz $MS_{Fehler}$ durch $MS_{Residuen}$ ersetzt wird. Voraussetzung ist, dass die Varianzhomogenität, hier also die Spherizität, erfüllt ist, was aus Tabelle 6-2 [99] zu entnehmen ist. $MS_{Residuen}$ sowie die dazugehörigen FG $df_{Residuen}$ müssen nun ermittelt werden, um sie als Parameter der Funktion `SNK.test` zu übergeben.

Ausgangsbasis ist wieder der in Kapitel 5.3.1 [99] erstellte Dataframe `winer518t`. Zunächst wird die Anova mit der „Standardfunktion“ `aov` durchgeführt:

```
data(winer518t)
aov1 <- aov(score~Geschlecht*Zeit+Error(Vpn/Zeit),winer518t)
```

Ein Weg - manchmal der einfachste - ist, die beiden Werte aus der Anova-Tabelle `summary(aov1)` einfach abzulesen, die hier noch einmal wiedergegeben wird:

```
Error: Vpn
           Df Sum Sq Mean Sq F value Pr(>F)
Geschlecht  1   3.33   3.333   0.472  0.512
Residuals   8  56.53   7.067

Error: Vpn:Zeit
                Df Sum Sq Mean Sq F value   Pr(>F)
Zeit             2  58.07  29.033   22.05 2.52e-05 ***
Geschlecht:Zeit  2  44.87  22.433   17.04 0.000109 ***
Residuals       16  21.07   1.317
```

Demnach sind $MS_{Residuen}$ = `1.317` bzw. $df_{Residuen}$ = 16.

Der andere Weg: die beiden Werte aus `aov1` zu extrahieren, was etwas umständlich ist:

```
aov1r    <- anova(aov1[["Vpn:Zeit"]])
dffehler <- aov1r[2,1]
msfehler <- aov1r[2,3]
```

Hierbei kann auch `aov1[[3]]` anstatt `aov1[["Vpn:Zeit"]]` geschrieben werden. Schließlich kann der Mittelwertvergleich durchgeführt werden:

```
SNK.test(winer518t$score, winer518t$Zeit, dffehler, msfehler,
         group=T, console=T)
```

```
Study: winer518t$score ~ winer518t$Zeit

Student Newman Keuls Test for winer518t$score

Mean Square Error:  1.316667

winer518t$Zeit,  means

  winer518t.score      std  r Min Max
1             5.9 1.911951 10   3   9
2             4.4 2.633122 10   1   9
3             2.5 1.840894 10   1   6

alpha: 0.05 ; Df Error: 16

Critical Range
       2        3
1.087851 1.324123

Means with the same letter are not significantly different.

Groups, Treatments and means
a        1       5.9
b        2       4.4
c        3       2.5
```

Aus dieser Ausgabe der homogenen Untergruppen ist zu entnehmen, dass jeder Mittelwert eine (1-elementige) Untergruppe - bezeichnet mit a, b und c - bildet, daher sich alle 3 Mittelwerte voneinander unterscheiden.

### 15. 2. 2 nlme

Das Paket `nlme` bietet die Analyse von Kontrasten, insbesondere auch für den Fall von Messwiederholungen. Dafür bieten sich die „wiederholt“-Kontraste an, die jeweils zwei aufeinander folgende Stufen miteinander vergleichen.

**Beispiel R-2.2**

Hierzu werden wieder wie im letzten Beispiel die Beispieldaten 4 (`winer518`) verwendet. Es werden die Stufen des Faktors `Zeit` verglichen. Eine Varianzhomogenität, d.h. Spherizität, wie bei dem vorherigen Beispiel ist hier nicht erforderlich. Zunächst wird eine Kontrastmatrix definiert und Zeit zugewiesen. Die Messwiederholungen werden über `corr=corCompSymm` deklariert. Basis ist wie in Beipiel R-2.1 der Dataframe `winer518t`. Die Ein- und Ausgabe:

```
cont3<-matrix(c(1,-1,0, 0,1,-1), ncol=2)
contrasts(winer518t$Zeit) <- cont3
aovgls <- gls(score~Geschlecht*Zeit, data=winer518t,
corr = corCompSymm(, form= ~ 1 | Vpn))
summary(aovgls)
```

```
Coefficients:
                     Value Std.Error   t-value p-value
(Intercept)        4.600000 0.6863753  6.701872  0.0000
Geschlecht2       -0.666667 0.9706813 -0.686803  0.4988
Zeit1              0.400000 0.4189935  0.954669  0.3493
Zeit2              2.200000 0.4189935  5.250678  0.0000
Geschlecht2:Zeit1  2.466667 0.5925463  4.162825  0.0003
Geschlecht2:Zeit2 -0.866667 0.5925463 -1.462614  0.1565
```

Die Zeilen `Zeit1` und `Zeit2` beinhalten die Vergleiche `Zeit1-Zeit2` bzw. `Zeit2-Zeit3`. Demnach unterscheiden sich die Mittelwerte der ersten beiden Zeitpunkte nicht, jedoch die der letzten beiden. Die Zeilen `Geschlecht2:Zeit1` bzw. `Geschlecht2:Zeit2` zeigen an, in wieweit sich der Vergleich `Zeit1-Zeit2` bzw. `Zeit2-Zeit3` für die beiden Geschlechtsgruppen unterscheidet. Hier also: der Vergleich `Zeit1-Zeit2` fällt für Männer und Frauen unterschiedlich aus.

Das Beispiel R-0.1 in Kapitel 3.7. enthält ein weiteres Beispiel.

## 15. 3 nichtparametrische Vergleiche - unabhängige Stichproben

### 15. 3. 1 agricolae

Das Paket `agricolae` war bereits in Kapitel 15.1 vorgestellt worden. Neben den parametrischen Verfahren werden dort auch nichtparametrische 1-faktorielle Varianzanalysen angeboten: die Tests von *Kruskal-Wallis* und von *van der Waerden* für unabhängige Stichproben sowie die *Friedman*-Varianazanalyse für abhängige Stichproben. Alle schließen automatisch Paarvergleiche an, wobei diese mit der jeweiligen Varianzanalyse, allerdings für zwei Gruppen, durchgeführt werden. Standardmäßig wird keine $\alpha$-Korrektur vorgenommen. Die Funktion für den Kruskal-Wallis-Tests bietet allerdings auch diverse Adjustierungen (vgl. Kapitel 15.1). Die Eingabe muss zwangsläufig die Vektoren der abhängigen Variablen sowie gegebenenfalls des Faktors enthalten. Nachfolgend Beispiele für alle drei Tests.

**Beispiel R-3.1a:**

Zunächst ein Beispiel für unabhängige Stichproben. Wie in Beispiel R-1.1 werden die Beispieldaten 2 (`mydata2`) verwendet. Für Mittelwertvergleiche im Anschluss an die Kruskal-Wallis-Varianzanalyse wird die $\alpha$-Korrektur von Benjamini & Hochberg `BH` (vgl. Kapitel 2.2.1) gewählt.

```
kruskal(mydata2$x, mydata2$drugs, p.adj="BH", group=F, console=T)
```

```
Kruskal-Wallis test's

Value: 11.19504
degrees of freedom: 3
Pvalue chisq  : 0.01071663

mydata2$drugs,  means of the ranks

  mydata2.x r
1   9.00000 7
2  15.00000 9
3  17.62500 8
4  24.66667 9
```

```
P value adjustment method: BH
Comparison between treatments mean of the ranks

      Difference   pvalue sig.       LCL         UCL
1 - 2  -6.000000 0.181049      -14.31720  2.31719564
1 - 3  -8.625000 0.095888    . -17.16659 -0.08340812
1 - 4 -15.666667 0.003576   ** -23.98386 -7.34947103
2 - 3  -2.625000 0.508496      -10.64447  5.39446531
2 - 4  -9.666667 0.049968    * -17.44669 -1.88664254
3 - 4  -7.041667 0.124416      -15.06113  0.97779864
```

Die Ausgabe enthält zunächst das Ergebnis für die Kruskal-Wallis-Varianzanalyse ($p = 0.01071$), die mittleren Ränge für die Gruppen sowie die paarweisen Mittelwertvergleiche. Vergleicht man die Ergebnisse für die zur Verfügung stehenden Adjustierungen, so kann man sehen, dass die hier verwendete Adjustierung von Benjamini & Hochberg die kleinsten p-Werte liefert.

Nachfolgend Eingabe und Ausgabe für den van der Waerden-Test, der zumindestens in diesem Fall stärker ist ($p = 0.00852$) als der Kruskal-Wallis-Test:

```
waerden.test(mydata2$x, mydata2$drugs, group=F, console=T)
```

```
Study: mydata2$x ~ mydata2$drugs
Van der Waerden (Normal Scores) test's

Value : 11.69112
Pvalue: 0.008519736
Degrees of freedom:  3

mydata2$drugs,  means of the normal score

    mydata2.x       std r
1 -0.76661807 0.7118198 7
2 -0.19006498 0.7472355 9
3  0.05463753 0.8210328 8
4  0.73269877 0.7000067 9

Comparison between treatments means
mean of the normal score

      Difference   pvalue sig.        LCL         UCL
1 - 2 -0.5765531 0.136046      -1.3456161  0.19250993
1 - 3 -0.8212556 0.042082    * -1.6110678 -0.03144341
1 - 4 -1.4993168 0.000414  *** -2.2683799 -0.73025382
2 - 3 -0.2447025 0.505074      -0.9862354  0.49683040
2 - 4 -0.9227637 0.013736    * -1.6421563 -0.20337116
3 - 4 -0.6780612 0.071586    . -1.4195942  0.06347167
```

Ein Beispiel für nichtparametrische Vergleiche bei Messwiederholungen mittels multipler Friedman-Tests folgt in Kapitel 15.4.

**Beispiel R-3.1b:**

Eine Alternative bietet die Durchführung der „klassischen“ Verfahren angewandt auf die rangtransformierte abhängige Variable, wie es von Conover & Iman [6] vorgeschlagen wird.

Hier wird z.B. Fishers LSD-Test durchgeführt, auch als Test von Conover & Iman bekannt (vgl. Kapitel 12.5), zunächst in der „ungeschützten“ Version. Anschließend kann man eine α-Adjustierung vornehmen, etwa die Benjamini & Hochberg, indem vom Ergebnis `lsd2` die Vergleiche extrahiert werden und davon die 2. Spalte der p-Werte einer Adjustierung mit `p.adjust` unterzogen wird:

```
mydata2 <- within(mydata2, Rx<-rank(x))
aov1r<-aov(Rx~drugs,mydata2)
lsd2 <- LSD.test(aov1r,"drugs",console=T,group=F)
lsd2$comparison
pwerte <- lsd2$comparison[,2]
p.adjust(pwerte,"BH")
```

Nachfolgend die „ungeschützten“ Vergleiche des LSD-Tests sowie die adjustierten p-Werte:

```
      Difference       pvalue sig.       LCL         UCL
1 - 2  -6.000000 0.1508746685      -14.31720  2.31719564
1 - 3  -8.625000 0.0479448796    * -17.16659 -0.08340812
1 - 4 -15.666667 0.0005962466  *** -23.98386 -7.34947103
2 - 3  -2.625000 0.5084961909      -10.64447  5.39446531
2 - 4  -9.666667 0.0166555675    * -17.44669 -1.88664254
3 - 4  -7.041667 0.0829438696    . -15.06113  0.97779864

[1] 0.18104960 0.09588976 0.00357748 0.50849619 0.04996670 0.12441580
```

Wählt man z.B. als Adjustierung das Verfahren von Li, so erhält man als
$\alpha' = \alpha(1-0.5085)/(1-\alpha) = 0.026$. Sowohl bei Benjamini & Hochberg als auch bei Li werden die Vergleiche 1-4 und 2-4 als signifikant ausgewiesen.

### 15. 3. 2 nparcomp

Hinweis: Dieser Abschnitt muss noch ergänzt und überarbeitet werden.

Das Paket `nparcomp` ist speziell für nichtparametrische multiple Mittelwertvergleiche konzipiert. Es bietet u.a. die folgenden Verfahren (im Wesentlichen für unabhängige Stichprobe):

| Funktionsname | Verfahren |
|---|---|
| `gao` | Gaos Vergleich mit einer Kontrollgruppe |
| `gao_cs` | Gaos paarweise Vergleiche |
| `mctp` | Schätzung der relativen Effekte und deren paarweise Vergleiche |
| `mctp.rm` | Schätzung der relativen Effekte und deren paarweise Vergleiche für abhängige Stichproben |
| `nparcomp` | Schätzung der relativen Kontrasteffekte und deren paarweise Vergleiche |

Folgende Hinweise:

- Es werden nur 1-faktorielle Analysen durchgeführt.
- Für `mctp` und `nparcomp` stehen dieselben Kontraste zur Verfügung wie im Paket `multcomp`. Standardmäßig werden auch hier die sog. Tukey-Kontraste gewählt, also paarweise Vergleiche. Alternativ können auch individuelle Kontraste vorgegeben werden.

- Bei dem Verfahren von Gao werden zwei Varianten ausgegeben:
  `Single.Analysis`: *pairwise ranking* - paarweise Vergleiche, individuelle Ränge
  `CS.Analysis`: *joint ranking* - gemeinsame Ränge (eine Variante des Verfahrens von *Campbell & Skillings*)

Der Aufruf aller Funktionen ist weitgehend identisch:

```
Funktion(Anova-Modell, Dataframe, type="Kontraste")
```

Darüber hinaus gibt es noch eine Reihe weiterer Parameter, die hier vorläufig nicht berücksichtigt werden.

**Beispiel R-3.2:**

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Zunächst wird ein paarweiser Vergleich der Mittelwerte mit dem Verfahren von Gao durchgeführt, anschließend mittels paarweiser Kontraste:

```
gao_cs (x~drugs, mydata2)
mctp   (x~drugs, mydata2, type="Tukey")
```

```
$Info
  Order Sample Size    Effect   Variance
1     1      1    7 0.2575758 0.04101622
2     2      2    9 0.4393939 0.07042011
3     3      3    8 0.5189394 0.08361209
4     4      4    9 0.7323232 0.04241276

$Single.Analysis
  Comp Effect Statistic      DF  P.RAW p.BONF p.HOLM
1  4-1 0.4747    4.6173 13.1518 0.0005 0.0028 0.0028
2  3-1 0.2614    2.0465 12.4746 0.0624 0.3743 0.2495
3  4-2 0.2929    2.6162 15.0714 0.0194 0.1164 0.0970
4  2-1 0.1818    1.5543 14.0000 0.1424 0.8546 0.3231
5  3-2 0.0795    0.5884 14.3613 0.5654 1.0000 0.5654
6  4-3 0.2134    1.7328 12.5102 0.1077 0.6461 0.3231

$CS.Analysis
  Comp Effect Statistic      DF Quantiles  Adj.P  Alpha Rejected Layer
1  4-1 0.4747    6.5298 13.1518    4.1443 0.0023 0.0500     TRUE     1
2  3-1 0.2786    2.8109 12.9794    3.7349 0.1546 0.0500    FALSE     2
3  4-2 0.3077    3.6721 15.8816    3.6518 0.0486 0.0500     TRUE     2
4  2-1 0.2063    2.0838 13.5960    3.5528 0.1634 0.0253    FALSE     3
5  3-2 0.0903    0.8782 14.3148    3.5312 0.5444 0.0253    FALSE     3
6  4-3 0.2292    2.4048 14.3365    3.5306 0.1106 0.0253    FALSE     3
```

Zur Ausgabe von `gao`: Zunächst werden unter `Info` die relativen Effekte der vier Gruppen protokolliert. (Diese liegen immer zwischen 0 und 1. Deren Verhältnis zueinander entspricht in etwa dem der Mittelwerte zueinander.) Anschließend werden unter `Single.Analysis` die Ergebnisse der Vergleiche bei paarweiser Rangbildung ausgegeben: hinter den Codenummern der beiden Gruppen zunächst drei Statistiken, die den Gao-Test betreffen, dann den unkorrigierten, den Bonferroni-adjustierten und den Holm-korrigierten p-Wert. Schließlich unter `CS.Analysis` die Ergebnisse der Vergleiche bei gemeinsamer Rangbildung mit dem Verfahren von Campbell & SKillings ausgegeben: hinter den Codenummern der beiden Gruppen zunächst vier Statis-

tiken, die den Gao-Test betreffen, dann den korrigierten p-Wert, der mit dem dahinter ausgegebenen `Alpha` zu vergleichen ist. Zu beachten ist, dass die Reihenfolge der Vergleiche sich aus den Ergebnissen ableitet und nicht aus der Reihenfolge der Gruppen.

```
 #----Data Info-------------------------------------#
  Sample Size    Effect
1      1    7 0.2703373
2      2    9 0.4534006
3      3    8 0.5322421
4      4    9 0.7440201

 #----Contrast--------------------------------------#
       1  2  3 4
2 - 1 -1  1  0 0
3 - 1 -1  0  1 0
4 - 1 -1  0  0 1
3 - 2  0 -1  1 0
4 - 2  0 -1  0 1
4 - 3  0  0 -1 1

 #----Analysis--------------------------------------#
      Estimator  Lower Upper Statistic     p.Value
2 - 1     0.183 -0.154 0.479     1.562 0.426118151
3 - 1     0.262 -0.096 0.551     2.057 0.215786884
4 - 1     0.474  0.235 0.609     5.037 0.001353779
3 - 2     0.079 -0.308 0.443     0.583 0.933250369
4 - 2     0.291 -0.012 0.532     2.669 0.080089308
4 - 3     0.212 -0.135 0.508     1.743 0.337882248
```

Zur Ausgabe von `mctp`: Zunächst werden unter `Data Info` die relativen Effekte der vier Gruppen sowie die Kontrastmatrix protokolliert. Unter `Analysis` folgen dann die Schätzung des Kontrastwertes, ein Konfidenzintervall dafür sowie ein Test auf Verschiedenheit von 0.

### 15. 3. 3 PMCMR

In `PMCMR` wird nur ein Verfahren angeboten: Nemenyis Test, allerdings sowohl für unabhängige als auch für abhängige Stichproben. Der Aufruf bei unabhängigen Stichproben:

```
posthoc.kruskal.nemenyi.test(abh.Variable, Gruppe, method=...)
```

Bei `method=` kann gewählt werden zwischen "`Tukey`" und "`Chisq`" (vgl. dazu Kapitel 12.3).

**Beispiel R-3.3:**

Zunächst soll ein Beispiel für den Fall unabhängiger Stichproben gerechnet werden. Dazu werden die Beispieldaten 10 (`mydata10`) verwendet. Es soll der Einfluss des Faktor `Gruppe` untersucht werden. Zunächst wird der Kruskal-Wallis-Test durchgeführt (was allerdings statistisch nicht erforderlich ist, um die Paarvergleiche durchzuführen), anschließend der paarweise Vergleich der Mittelwerte mittels Nemenyis Test, wobei wegen der Bindungen die Prüfung über die $\chi^2$-Verteilung gewählt wird.

```
with(mydata10, kruskal.test (x, Gruppe))
with(mydata10, posthoc.kruskal.nemenyi.test(x,Gruppe,method="Chisq"))
```

Die Ausgabe beider Tests:

```
        Kruskal-Wallis rank sum test

data:  x and Gruppe
Kruskal-Wallis chi-squared = 25.2664, df = 5, p-value = 0.0001238

       Pairwise comparisons using Nemenyi-test with Chi-squared
                        approximation for independent samples

data:  x and Gruppe

  1       2       3       4       5
2 0.90362 -       -       -       -
3 0.95824 0.43815 -       -       -
4 0.23228 0.85163 0.03665 -       -
5 0.00386 0.11058 0.00022 0.76587 -
6 0.42789 0.95479 0.09906 0.99987 0.64728
```

Danach zeigt der Kruskal-Wallis zunächst an, dass zwischen den Mittelwerten Unterschiede bestehen ($p<0.01$). Für das Verfahren von Nemenyi werden die p-Werte für die Paarvergleiche ausgegeben. Demnach unterschieden sich nur die Gruppen 1 und 5, 3 und 4 sowie 3 und 5.

Führt man dagegen den Tukey-Test auf die rangtransformierten Daten durch, was nach Conover & Iman [6] durchaus legitim ist, solange die $n_i$ nicht stark variieren, erhält man deutlich mehr sigifikante Unterschiede:

```
mydata10 <- within(mydata10, Gruppe<-factor(Gruppe); Rx<-rank(x))
aov10r<-aov(Rx~Gruppe,mydata10)
TukeyHSD(aov10r)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Rx ~ Gruppe, data = mydata10)

$Gruppe
          diff        lwr        upr     p adj
2-1   6.166667  -3.678492 16.0118255 0.4168680
3-1  -5.266667 -15.592356  5.0590231 0.6331456
4-1  13.083333   3.238174 22.9284922 0.0042651
5-1  20.690476  11.203444 30.1775086 0.0000038
6-1  11.533333   1.207644 21.8590231 0.0217328
3-2 -11.433333 -21.759023 -1.1076436 0.0233343
4-2   6.916667  -2.928492 16.7618255 0.2950741
5-2  14.523810   5.036777 24.0108420 0.0008294
6-2   5.366667  -4.959023 15.6923564 0.6147920
4-3  18.350000   8.024310 28.6756897 0.0001076
5-3  25.957143  15.972329 35.9419563 0.0000001
6-3  16.800000   6.015169 27.5848312 0.0006649
5-4   7.607143  -1.879890 17.0941753 0.1745553
6-4  -1.550000 -11.875690  8.7756897 0.9972207
6-5  -9.157143 -19.141956  0.8276706 0.0867844
```

Ein weiteres Beispiel zum Nemenyi-Test folgt im nächsten Kapitel.

### 15. 3. 4 NSM3

Das Paket `NSM3` bietet eine Fülle von nichtparametrischen Tests, die im Buch von *Hollander, Wolfe and Chicken* [97] beschrieben sind, u.a. auch die folgenden für unabhängige Stichproben:

| Funktionsname | Verfahren |
|---|---|
| `pNDWol` | Nemenyi, Damico-Wolfe Y statistic zum Vergleich mit einer Kontrollgruppe |
| `pSDCFlig` | Dwass, Steel, Critchlow, Fligner paarweise Vergleiche |
| `pFligPoli` | Fligner-Policello Vergleich von zwei Stichproben |
| `pJCK` | Jonckheere-Terpstra J statistic: Varianzanalyse für ordinal skalierten Faktor |

`NSM3` bietet für fast alle Funktionen drei Arten der Signifikanzberechnung:

- "Exact" auf Basis der Berechnung von Permutationen,
- "Monte Carlo" auf Basis von Simulationen oder
- "Asymptotic" näherungsweise für größere *n*

Es wird darauf aufmerksam gemacht, dass bei den ersten beiden Optionen auch im Fall kleinerer Stichproben gelegentlich nicht nur Minuten, sondern auch Stunden Rechenzeit benötigt werden. Die Ausgabe der Funktionen ist generell etwas unübersichtlich. Häufig genügt auch die Ausgabe zweier Ergebnisvektoren:

- ...`$p.val` die p-Werte
- ...`$obs.stat` die Testgrößen

Der Aufruf ist einheitlich:

```
Funktion (abh.Variable, Faktor, method=Signifikanzberechnung)
```

**Beispiel R-3.4a:**

Es wird das vorige Beispiel mit den Beispieldaten 10 (`mydata10`) aufgegriffen. Und es sollen wieder die Mittelwerte der 6 Gruppen mittels des Tests von Dwass, Steel, Critchlow, Fligner verglichen werden.

```
with(mydata10,pSDCFlig(x,Gruppe,method="Asymptotic"))
```

Zunächst die Ausgabe auszugsweise (nur die beiden ersten Vergleiche)

```
Ties are present, so p-values are based on conditional null distribution.
Group sizes:  6 6 5 6 7 5
Using the Asymptotic method:

For treatments 1 - 2,the Dwass,Steel,Critchlow-Fligner W Statistic is 2.1704
The smallest experimentwise error rate leading to rejection is 0.6421 .

For treatments 1 - 3,the Dwass,Steel,Critchlow-Fligner W Statistic is -2.3728
The smallest experimentwise error rate leading to rejection is 0.5467 ..
```

Nachfolgend der Ergebnisvektor `$p.val` , woraus im Vergleich zu den Ergebnissen des o.a. Beispiels R-3.3 zu entnehmen ist, dass dieser Test zwar besser als der Nemenyi-Test abschneidet, aber nicht an die des Tukey-Tests angewandt auf die rangtransformierte Kriteriumsvariable herankommt:

```
 [1] 0.64213204 0.54667833 0.09419970 0.02957835 0.19280440
     0.09933739 0.63842351
 [8] 0.04374035 0.77574913 0.06563977 0.04695106 0.09089353
     0.32486812 0.99908941
[15] 0.18825121
```

**Beispiel R-3.4b:**

Es werden die Beispieldaten 3 (`mydata3`) aufgegriffen, die einen 2-faktoriellen Versuchsplan (Faktoren `Gruppe` und `Dosis`) beinhalten, allerdings mit heterogenen Varianzen. Hier soll ein nichtparametrischer Vergleich der 4 Dosierungen durchgeführt werden, aber wegen der heterogenen Vraianzen mit einem robusten Test: dem Fligner-Policello-Test. Da dieser nur zwei Gruppen vergleicht, können die paarweisen Vergleiche nicht in einem Funktionsaufruf angefordert werden. Hier werden die beiden in Kapitel 15.8 vorgestellten Lösungen für eigene Funktionen verwendet: `p.collect` und `my.pairwise.test` mit `pairwise.table`. Zunächst `p.collect`, bei der der Ergebnisvektor der p-Werte noch adjustiert werden muss, hier nach der Hommel-Methode:

```
pwerte <- p.collect(mydata3$x,mydata3$dosis,4)
p.adjust (pwerte, "hommel")
```

```
    1 - 2     1 - 3     1 - 4     2 - 3     2 - 4     3 - 4
0.1798202 0.2877123 0.1222777 0.6794000 0.6249000 0.6794000
```

Nun die Lösung mittels `my.pairwise.test` und `pairwise.table`, bei der zur Adjustierung die Benjamini-Hochberg-Methode verwendet wird:

```
my.pairwise.test(mydata3$x,mydata3$dosis,"BH")
```

```
          1         2         3
2 0.1135459        NA        NA
3 0.1560318 0.4657602        NA
4 0.1135459 0.2948929 0.6983705
```

## 15. 4 nichtparametrische Vergleiche - abhängige Stichproben

### 15. 4. 1 agricolae

Das Paket `agricolae` wurde bereits im Kapitel 15.1.1 sowie im vorigen Kapitel vorgestellt. Hier nun noch ein Beispiel für nichtparametrische Vergleiche bei Messwiederholungen.

**Beispiel R-4.1:**

Für ein Beispiel mit Messwiederholungen werden wie in o.a. Beispiel R-2.1 die Daten 5 (`winer518`) verwendet. Ausgangsbasis ist wieder der in Kapitel 5.3.1 [99] erstellte Dataframe `winer518t`. Als Ausgabemodus werden die homogenen Untergruppen gewählt. Die Eingabe unterscheidet sich in Folge der Messwiederholungen ein wenig von der für die anderen `agricolae`-Funktionen:

- 1. Parameter: die Fallkennung, hier `Vpn`
- 2. Parameter: der Messwiederholungsfaktor, hier `Zeit`
- 3. Parameter: dei abhängige Variable: hier `score`.

```
with(winer518t, friedman(Vpn, Zeit, score, group=T, console=T))
```

```
Study: score ~ Vpn + Zeit

Zeit,  Sum of the ranks

  score  r
1    26 10
2    21 10
3    13 10

Friedman's Test
===============
Adjusted for ties
Value: 9.555556
Pvalue chisq : 0.008414677
F value : 8.234043
Pvalue F: 0.002888671

Alpha     : 0.05
t-Student : 2.100922
LSD       : 6.789732

Means with the same letter are not significantly different.
GroupTreatment and Sum of the ranks
a        1       26
a        2       21
b        3       13
```

Auch hier wird zunächst das Ergebnis der Friedman-Varianzanalyse protokolliert ($p = 0.00289$), gefolgt von dem Ergebnis der Mittelwertvergleiche. Daraus ist abzulesen, dass die Zeitpunkte 1 und 2 sich nicht unterscheiden, da sie derselben Untergruppe a angehören. Folglich unterscheidet sich Zeitpunkt 3 (Untergruppe b) von 1 und 2.

### 15. 4. 2 PMCMR

`PMCMR` für die Durchführung von Nemenyis Test wurde bereits im vorigen Kapitel vorgestellt. Für den Aufruf bei abhängigen Stichproben gibt es zwei Varianten:

```
posthoc.friedman.nemenyi.test(abh.Variable, Gruppe, blocks=Vpn)
posthoc.friedman.nemenyi.test(Matrix)
```

In der ersten Variante wird eine umstrukturierte Datenmatrix benötigt (vgl. Kapitel 5 in [99]). Dabei ist *Vpn* die Fallkennzeichnung. In der zweiten Variante werden die Daten in der eigentlichen Struktur für Messwiederholungen benötigt, allerdings nicht als Dataframe, sondern als Matrix. (Dies entspricht der Eingabe der Funktion `friedman.test` .) Diese erhält man z.B. durch `as.matrix(`*Dataframe*`)`.

**Beispiel R-4.2:**

Die Beispieldaten 9 (`mydata9`, vgl. Kapitel 1.9) beinhalten einen Versuchsplan mit 4 Messwiederholungen `V1,...,V4` (Faktor `Bedingung`). Es sollen die Mittelwerte der 4 Bedingungen verglichen werden. Zunächst wird der Friedman-Test durchgeführt (was allerdings statistisch nicht erforderlich ist, um die Paarvergleiche durchzuführen), anschließend der paarweise Vergleich der Mittelwerte mittels Nemenyis Test.

```
friedman.test (as.matrix(mydata9))
posthoc.friedman.nemenyi.test (as.matrix(mydata9))
```

Die Ausgabe beider Tests:

```
        Friedman rank sum test

data:  as.matrix(mydata9)
Friedman chi-squared = 12.3474, df = 3, p-value = 0.006283

   Pairwise comparisons using Nemenyi post-hoc test with
   q approximation for unreplicated blocked data

data:  as.matrix(mydata9)

   V1     V2     V3
V2 0.4544 -      -
V3 0.0365 0.6189 -
V4 0.0099 0.3531 0.9728
```

Danach zeigt die Friedman-Varianzanalyse zunächst an, dass zwischen den Mittelwerten Unterschiede bestehen ($p<0.01$). Für das Verfahren von Nemenyi werden die p-Werte für die Paarvergleiche ausgegeben. Demnach unterschieden sich die Bedingungen 1 und 3 sowie 1 und 4 ($p<0.05$).

### 15. 4. 3 NSM3

Das Paket `NSM3` war bereits in Kapitel 15.3.4 vorgestellt worden. Für abhängige Stichproben werden u.a.die folgenden Verfahren angeboten:

| Methodenname | Verfahren |
|---|---|
| `pWNMT` | Nemenyi, Wilcoxon, McDonald-Thompson zum paarweisen Vergleich |
| `pNWWM` | Nemenyi, Wilcoxon-Wilcox, Miller zum Vergleich mit einer Kontrollgruppe |
| `pSkilMack` | Skillings-Mack SM statistic: Varianzanalyse bei fehlenden Werten |

Der Aufruf ist einheitlich:

*`Funktion`* (*`Matrix`*, method=*`Signifikanzberechnung`*)

wobei *`Matrix`* die (Teil)Matrix der zu untersuchenden abhängigen Variablen ist, die z.B. über `as.matrix(`*`dataframe`*`[,c(`*`Indizes`*`)])` erzeugt werden kann.

**Beispiel R-4.3:**

Es wird das vorige Beispiel mit den Beispieldaten 9 (`mydata9`) aufgegriffen. Es sollen wieder die 4 Versuchsbedingungen verglichen werden, hier mit dem Nemenyi-Test:

```
pWNMT(as.matrix(mydata9),method="Asymptotic")
```

Nachfolgend wird lediglich der Vektor der p-Werte wiedergegeben, die mit dem Ergebnis des Beipiels R-4.2 übereinstimmen.:

```
[1] 0.45461396 0.03652119 0.00986959 0.61897668 0.35325958 0.97291497
```

### 15. 4. 4 stats: quade.test

Der Quade-Test (vgl. Kapitel 12.9) entspricht eigentlich einer Varianzanalyse, führt also einen globalen Test auf Gleichheit der Mittelwerte durch. Möchte man paarweise Vergleiche durchführen, so müssen diese anschließend einer α-Adjustierung unterzogen werden. Der Aufruf entspricht dem der anderen Tests für abhängige Stichproben:

```
quade.test (as.matrix(dataframe[ ,ausgewählte Variablen]))
```

Der Ergebnisverktor `$p.value` enthält die p-Werte, die gegebenenfalls anschließend bearbeitet werden müssen.

**Beispiel R-4.4:**

Es sollen wie in den vorangegangenen Beispielen dieses Kapitels die 4 Bedingungen der Beispieldaten 9 (`mydata9`) verglichen werden. Zunächst wird der globale Test durchgeführt, anschließend über zwei `for`-Schleifen jeweils zwei Bedingungen verglichen, die p-Werte in `pwerte` gesammelt, diesen die beiden Bedingungs-Nummern als Namen zugewiesen, um schließlich diese mittels der Adjustierung von Hommel in `p.adjust` zu korrigieren.

```
quade.test(as.matrix(mydata9))

k<-4; i<-0; pwerte<-0
for (i1 in 1:(k-1)) { for (i2 in (i1+1):k)
 {i<-i+1 ;
  erg <- quade.test(as.matrix( ,[c(i1,i2]))
  pwerte[i] <- erg$p.value
  names(pwerte)[i] <- paste(i1,i2,sep=" - ")
 }}

pwerte
p.adjust(pwerte,"hommel")
```

Die Ausgabe enthält nacheinander das Ergebnis des globalen Quade-Tests, die nichtadjustierten p-Werte der Paarvergleiche und schließlich die Hommel-adjustierten p-Werte:

```
        Quade test

data:  as.matrix(mydata9)
Quade F = 9.1559, num df = 3, denom df = 27, p-value = 0.0002401

------------------------------------------------------------------------
     1 - 2      1 - 3      1 - 4      2 - 3      2 - 4      3 - 4
0.01559136 0.00041458 0.00821927 0.15643384 0.03032801 0.02931011
------------------------------------------------------------------------
     1 - 2      1 - 3      1 - 4      2 - 3      2 - 4      3 - 4
0.04549202 0.00248749 0.03791001 0.15643384 0.06065602 0.05862023
```

## 15. 5 allgemeine α-Adjustierungen

### 15. 5. 1 stats: p.adjust

`p.adjust` ist eine R-Standardfunktion, die verschiedene bekannte p-Wert-Adjustierungen anbietet:

| Methodenname | Verfahren |
|---|---|
| `BH` | Benjamini-Hochberg step-up procedure |
| `BY` | Benjamini and Yekutieli |
| `hochberg` | Hochberg step-up procedure |
| `holm` | Holm's step-down-procedure |
| `hommel` | Hommel's step-up-procedure |
| `bonferroni` | Bonferroni Korrektur |

Der Aufruf dieser Funktion:

```
p.adjust (p-Werte, „Methode")
```

d.h. die zu adjustierenden p-Werte sind vorher in einem Vektor zu sammeln. Ein Beispiel R-5.2 folgt im nächsten Kapitel.

### 15. 5. 2 mutoss

Das Paket `mutoss` wurde bereits in Kapitel 15.1.7 kurz vorgestellt. Dessen Hauptthema sind allerdings die p-Wert-Adjustierungen. Neben einer Reihe von Funktionen zur Konstruktion von step-up- und step-down Verfahren sowie einigen weniger bekannten Adjustierungen werden u.a. die folgenden Methoden angeboten:

| Funktionsname | Adjustierungsverfahren |
|---|---|
| `BH` | Benjamini-Hochberg step-up procedure |
| `BL` | Benjamini-Liu's step-down procedure |
| `BY` | Benjamini and Yekutieli<br>(für abhängige Vergleiche) |
| `adaptiveBH` | adaptive Benjamini-Hochberg step-up procedure |
| `hochberg` | Hochberg step-up procedure |
| `holm` | Holm's step-down-procedure |
| `hommel` | Hommel's step-up-procedure |
| `bonferroni` | Bonferroni Korrektur |
| `sidak` | Sidak Korrektur |
| `SidakSD` | Sidak step-down procedure (Variante von Holm) |
| `rom` | Rom's step-up procedure |
| `multiple.down` | Benjamini-Krieger-Yekutieli multi-stage step-down procedure |

| Funktionsname | Adjustierungsverfahren |
|---|---|
| `two.stage` | schrittweises Varfahren zur FDR-Kontrolle |
| `twostageBR` | Blanchard-Roquain 2-stage adaptive step-up procedure |
| `BlaRoq` | Blanchard-Roquain step-up procedure<br>(für abhängige Vergleiche) |
| `indepBR` | Blanchard-Roquain 1-stage adaptive step-up procedure<br>(für unabhängige Vergleiche) |

Der Aufruf dieser Funktionen ist weitgehend identisch:

```
Funktion (p-Werte, alpha, silent=T/F)
```

d.h. die zu adjustierenden p-Werte sind vorher in einem Vektor zu sammeln. Die Angabe eines `alpha` dient zur Kennzeichnung der Signifikanzen, ist aber erforderlich. Der Parameter `silent` dient zur Unterdrückung der Ausgabe, z.B. bei Verwendung in eigenen Funktionen. Die Darstellung der Ergebnisse ist allerdings nicht einheitlich. Insbesondere werden manchmal die adjustierten p-Werte ausgegeben, manchmal auch nur die Nummern der abgelehnten Hypothesen. Dabei ist es ein wenig irritierend, dass die abgelehnten (Null-) Hypothesen mit `TRUE` und die angenommenen mit `FALSE` gekennzeichnet werden. Dies bezieht sich allerdings auf die Ausgabe `rejected`. Und `TRUE` bezogen auf „rejected“ bedeutet eben „abgelehnt“.

**Beispiel R-5.2:**

Es sollen die 6 Gruppen des Beispieldatensatzes 10 (`mydata10`, vgl. Kapitel 1.9) hinsichtlich der Mittelwerte verglichen werden. Allerdings soll der Mann Whitneys U-Test zusammen mit der α-Korrektur nach dem Blanchard-Roquains 1-stage-Verfahren verwendet werden.

Die paarweisen Vergleiche werden zunächst in einer doppelten Schleife durchgeführt. Dabei werden die p-Werte gesammelt und diesen die beiden Gruppennummern als Namen zugewiesen. Es ist ratsam, sich die p-Werte ausgeben zu lassen, da damit häufig eine Identifikation der (abgelehnten) Hypothesen leichter möglich ist. Anschließend wird noch eine Adjustierung nach dem Verfahren von Hommel mittels `p.adjust` durchgeführt. Ein- und Ausgabe:

```
k<-6
i<-0
for (i1 in 1:(k-1))
  { for (i2 in (i1+1):k)
    {i<-i+1
     erg <- with(mydata10, wilcox.test(x[Gruppe==i1],x[Gruppe==i2]))
     pwerte[i] <- erg$p.value
     names(pwerte)[i] <- paste(i1,i2,sep=" - ")
  }}
pwerte
indepBR (pwerte,0.05)
p.adjust(pwerte,"hommel")
```

```
      1 - 2       1 - 3       1 - 4       1 - 5       1 - 6
0.145968903 0.113049979 0.011369265 0.003105035 0.027030077
       2 - 3       2 - 4       2 - 5       2 - 6       3 - 4
 0.012452138 0.144531980 0.004731213 0.221167571 0.007827028
       3 - 5       3 - 6       4 - 5       4 - 6       5 - 6
0.005267286 0.011667312 0.050184328 0.780762910 0.025630938
```

```
Blanchard-Roquain 1-stage step-up under independence (2009)

Number of hyp.:  15
Number of rej.:  9
  rejected     pValues
1        4 0.003105035
2        8 0.004731213
3       11 0.005267286
4       10 0.007827028
5        3 0.011369265
6       12 0.011667312
7        6 0.012452138
8       15 0.025630938
9        5 0.027030077
```

9 der 15 Hypothesen (gleicher Mittelwerte) werden abgelehnt. Die Identifikation der abgelehnten Hypothesen erfolgt wie in Beispiel R-1.7. Hier noch die Ausgabe von `p.adjust` mit der Hommel-Adjustierung, bei der lediglich zwei Vergleiche als unterschiedlich ausgewiesen werden, nämlich 1-5 und 2-5:

```
     1 - 2      1 - 3      1 - 4      1 - 5      1 - 6
0.33175136 0.33175136 0.09095412 0.03726042 0.17564515
     2 - 3      2 - 4      2 - 5      2 - 6      3 - 4
0.09961711 0.33175136 0.04731213 0.44233514 0.07044325
     3 - 5      3 - 6      4 - 5      4 - 6      5 - 6
0.05267286 0.09333850 0.24328151 0.78076291 0.17564515
```

## 15. 6 Mittelwertvergleiche mit α-Adjustierungen

### 15. 6. 1 stats: pairwise.t.test

Im Standardumfang von R gibt es die sehr hilfreiche Funktion `pairwise.t.test` zur Durchführung von paarweisen Mittelwertvergleichen sowohl für unabhängige Stichproben, in Varianten für homogene (`pool.sd=T`) und für inhomogene Varianzen (Welch-Methode, `pool.sd=F`), als auch für abhängige Stichproben (`paired=T`). p-Wert-Adjustierungen werden automatisch vorgenommen, wobei die Methoden der Funktion `p.adjust` (vgl. 15.5.1) zur Verfügung stehen. Der Aufruf für unabhängige bzw. abhängige Stichproben:

```
pairwise.t.test (abh.Variable, Gruppe, "Adj.methode", pool.sd=T/F)
pairwise.t.test (abh.Variable, Gruppe, "Adj.methode", paired=T)
```

Für den Fall abhängiger Stichproben muss auch die transformierte Datenmatrix benutzt werden, wie sie für die Varianzanalyse mit Messwiederholungen in R benötigt wird.

**Beispiel R-6.1a:**

Für die Beispieldaten 3 (`mydata3`), einem Versuchsplan mit zwei unabhängigen Faktoren (Gruppe und Dosis) aber heterogenen Varianzen soll ein Vergleich der 4 Dosierungen mittels des Tests von Simes (vgl. Kapitel 7.6) vorgenommen werden. Dieser beinhaltet paarweise t-Tests für ungleiche Varianzen mit anschließender Korrektur nach Benjamini & Hochberg. Ein- und Ausgabe:

```
with(mydata3, pairwise.t.test(x,dosis,"BH",pool.sd=F))
```

```
        Pairwise comparisons using t tests with non-pooled SD

data:  x and dosis

  1     2     3
2 0.083 -     -
3 0.083 0.312 -
4 0.078 0.139 0.630

P value adjustment method: BH
```

Hiernach ist keiner der Vergleiche signifikant. Eine andere Testmöglichkeit bieten nichtparametrische (hinsichtlich heterogenen Varianzen) robuste Mittelwertvergleiche mittels des Test von Fligner-Policello, der im Paket `NSM3` enthalten ist. Ein Beipiel dafür ist R-3.4b, das zwar qualitativ dieselben Ergebnisse erbringt, allerdings quantitativ etwas größere p-Werte.

**Beispiel R-6.1b:**

Die Beispieldaten 9 (`mydata9`) beinhalten einen Versuchsplan mit 4 Messwiederholungen `V1,...,V4` (Faktor `Bedingung`). Es sollen die Mittelwerte der 4 Bedingungen verglichen werden. Dazu muss zunächst die Datenmatrix `mydata9` in die Matrix `mydata9t` umstrukturiert werden (vgl. Kapitel 5 in [99]). Anschließend wird mittels der Funktion `ezANOVA` aus dem Paket `ez` die Varianzhomogenität, hier also die Spherizität überprüft, um eventuell einen der klassischen Tests anwenden zu dürfen. Da diese nicht gegeben ist (Mauchly-Test: p=0.041), werden paarweise t-Tests mit der Korrektur von Hommel durchgeführt:

```
library(ez)
mydata9t <- reshape(mydata9, direction="long", timevar="Bedingung",
    v.names="score", varying=c("V1","V2","V3","V4"), idvar="Vpn")
mydata9t <- within(mydata9t, Bedingung<-factor(Bedingung);
    Vpn<-factor(Vpn))
ezANOVA(mydata9t, score, Vpn, within=.(Bedingung))
with(mydata9t, pairwise.t.test(score,Bedingung,"hommel",paired=T))
```

Ausgabe von `ezANOVA`:

```
$ANOVA
     Effect DFn DFd        F            p p<.05       ges
2 Bedingung   3  27 9.304303 0.0002158305     * 0.2373559

$`Mauchly's Test for Sphericity`
     Effect        W          p p<.05
2 Bedingung 0.220048 0.04079565     *
```

Ausgabe von `pairwise.t.test`, wonach nur die Vergleiche 1-3 und 1-4 signifikant sind:

```
data:  score and Bedingung
  1      2      3
2 0.0774 -      -
3 0.0025 0.2100 -
4 0.0199 0.0810 0.1080

P value adjustment method: hommel
```

Ein Vergleich diverser α-Adjustierungen für dieses Beispiel bietet das Beispiel S-8 in Kapitel 16.4.

### 15. 6. 2 stats: pairwise.wilcox.test

Zur o.a. Funktion `pairwise.t.test` gibt es das nichtparametrische Analogon `pairwise.wilcox.test`, womit zum einen für unabhängige Stichproben der Mann-Whitney-U-Test sowie für abhängige Stichproben der Wilcoxon-Rangsummen-Test durchgeführt werden können. p-Wert-Adjustierungen werden automatisch vorgenommen, wobei die Methoden der Funktion `p.adjust` (vgl. 15.5.1) zur Verfügung stehen. Der Aufruf für unabhängige bzw. abhängige Stichproben:

```
pairwise.wilcox.test (abh.Variable, Gruppe, "Adj.methode")
pairwise.wilcox.test (abh.Variable, Gruppe, "Adj.methode", paired=T)
```

Für den Fall abhängiger Stichproben muss die transformierte Datenmatrix benutzt werden, wie sie für die Varianzanalyse mit Messwiederholungen in R benötigt wird.

**Beispiel R-6.2a:**

Es werden wie im vorangegangenen Beispiel die Beispieldaten 9 (`mydata9`) mit 4 Messwiederholungen `V1,...,V4` (Faktor `Bedingung`) benutzt. Es sollen die Mittelwerte der 4 Bedingungen mittels Wilcoxon-Test verglichen werden. Dazu muss zunächst die Datenmatrix `mydata9` in die Matrix `mydata9t` mittels der Funktion `reshape` umstrukturiert werden (vgl. dazu Kapitel 5.1 in [99]). Ein- und Ausgabe, die keine signifikanten Unterschiede anzeigt:

```
mydata9t <- reshape(mydata9,direction="long",timevar="Bedingung",
            v.names="score", varying=c("V1","V2","V3","V4"),idvar="Vpn")

with (mydata9t, pairwise.wilcox.test
                (score, Bedingung, paired=T, p.adjust.method="BH"))
```

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  score and Bedingung

  1     2     3
2 0.066 -     -
3 0.052 0.184 -
4 0.062 0.084 0.088
```

Möchte man jedoch „progressivere“ α-Adjustierungen einsetzen, z.B. die von *Blanchard & Roquain* aus dem Paket `mutoss`, die die FDR unter Kontrolle hält, oder die von Li, die die FWER unter Kontrolle hält, dann ist man gezwungen die paarweisen Wilcoxon-Tests einzeln zusammenzustellen. Dazu das folgende Beispiel:

**Beispiel R-6.2b:**

Die gleiche Aufgabe mit denselben Daten wie im vorangegangenen Beispiel. Es werden in einer doppelten `for`-Schleife die paarweisen Wilcoxon-Tests durchgeführt, im Vektor `pwerte` die p-Werte gesammelt und diesen die Kodierungen der verglichenen Gruppen als Namen gegeben. Anschließend können darauf die p-Wert-Adjustierungen, z.B. `indepBR`, angewandt werden. Zuvor werden noch einmal die rohen p-Werte ausgegeben:

```
k<-4; i<-0; pwerte<-0
for (i1 in 1:(k-1)) { for (i2 in (i1+1):k)
 {i<-i+1 ;
  erg <- with(mydata9t,
        wilcox.test(score[Bedingung==i1],score[Bedingung==i2],paired=T))
  pwerte[i] <- erg$p.value
  names(pwerte)[i] <- paste(i1,i2,sep=" - ")
 }}
pwerte
indepBR (pwerte, alpha=0.05)
```

```
     1 - 2      1 - 3      1 - 4      2 - 3      2 - 4      3 - 4
0.02965429 0.00872904 0.02055030 0.18422515 0.04198132 0.04401464

     Blanchard-Roquain 1-stage step-up under independence (2009)

Number of hyp.:  6
Number of rej.:  5
  rejected     pValues
1        2 0.008729039
2        3 0.020550302
3        1 0.029654287
4        5 0.041981322
5        6 0.044014643
$rejected
[1]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE

$criticalValues
[1] 0.00791667 0.01900000 0.03562500 0.05000000 0.05000000 0.05000000
```

Während bei dem Benjamini & Hochberg-Verfahren im vorigen Beispiel die adjustierten p-Werte einiger Vergleiche knapp über $\alpha$=0.05 lagen, werden nun hier bei dem adaptiven Verfahren von *Blanchard & Roquain* 5 von den 6 Vergleichen als signifikant ausgewiesen. Lediglich der 4. Vergleich (2-3) fällt gemäß `rejected` negativ aus. In der Spalte `pValues` werden die adjustierten p-Werte errechnet.

Eine gute Alternative ist auch das Verfahren von Li, das leicht mit der Hand gerechnet werden kann. Die $p_i$ sind mit folgenden $\alpha'$ zu vergleichen: $\alpha(1-p_{(6)})/(1-\alpha)$. Mit $\alpha$=0.05 $p_{(6)}$=0.1842 (maximales p bei Vergleich 2-3) ergibt sich $\alpha'$= 0.043. Damit werden alle Hypothesen bis auf 4 und 6 abgelehnt, die den Vergleichen 2-3 und 3-4 entsprechen.

### 15. 6. 3 multcomp

Das Paket `multcomp` wurde bereits in 15.1.2 sowie 15.2.2 vorgestellt. Es bietet eine Reihe von $\alpha$-Adjustierungen, die in anderen Paketen nicht zu finden sind, u.a. die Verfahren von *Westfall* und *Shaffer*, in Zusammenhang mit Mittelwert-Kontrasten und -Paarvergleichen.

## 15. 7 Weitere Pakete

Es gibt für R noch eine Reihe weiterer Pakete, die multiple Mittelwertvergleiche und $\alpha$-Adjustierungen anbieten, von denen einige hier noch kurz erwähnt werden:

- `laercio`
  Dieses bietet die Tests von Tukey, Duncan und Scott & Knott an. Als Eingabe dient ein `aov`-Objekt.

- `coin`
  Dieses bietet u.a. die Varianzanalysen von Kruskal-Wallis und Friedman an sowie den Wilcoxon-Rangsummen-Test.
- `pgirmess`
  Dieses bietet sowohl die Kruskal-Wallis- als auch die Friedman-Varianzanalyse jeweils mit multiplen Mittelwertvergleichen nach einem Verfahren von *Siegel & Catellan.*
- `multtest`
  Dieses auf `www.bioconductor.org` angebotene Paket bietet die p-Wert-Adjustierungen min.P und max.P für den Vergleich größerer Variablenzahlen für zwei oder mehr Gruppen.
- `dunn.test`
  Dieses bietet lediglich den nichtparametrischen *Dunns* Test an.
- `DescTools`
  Hierin befindet sich zum einen eine Funktion `PostHocTest` zur Durchführung paarweiser Vergleiche: Fishers LSD, Dunn-Bonferroni, Newman-Keuls, Scheffé und Tukey. Die Funktionalität ist etwa dieselbe wie im Paket `agricalae`. Zum anderen die Fuktion `ScheffeTest` zur Durchführung von Scheffé-Kontrasten.

## 15. 8 Hilfsfunktionen

Die o.a. Beispiele zeigten, dass man in vielen Fällen auf 2-Stichprobenvergleiche zurückgreifen muss, die dann für alle Paare von Gruppen durchgeführt werden. Eine Möglichkeit besteht darin (wie oben gezeigt), über `for`-Schleifen jeweils die Paare auszuwählen und die p-Werte dabei in einem Vektor zu sammeln. Insbesondere wenn man solche paarweisen Vergleiche für mehrere Analysen wiederholt durchführen möchte, sucht man nach einer Vereinfachung. Nachfolgend werden zwei Verfahren vorgestellt.

### 15. 8. 1 p.collect

Genau diese oben erwähnten Schleifenanweisungen lassen sich in eine Funktion auslagern. Allerdings ist eine universelle Funktion, die Paarvergleiche mit verschiedenen statistischen Tests durchführt, wenig sinnvoll. Zwar lässt sich der Funktionsname des gewünschten Tests, z.B. `wilcox.test`, als Parameter übergeben, aber die Parameter für den gewünschten Test, z.B. `paired` für gepaarte oder nichtgepaarte Tests, variieren zu sehr von Funktion zu Funktion.

Nachfolgend ein Beispiel für eine solche Funktion, hier `p.collect` genannt, zur Durchführung paarweiser Tests mittels `pFligPoli` aus dem Paket `NSM3` (vgl. Kapitel 15.3.3) nach dem Verfahren von Fligner-Policello (vgl. Kapitel 12.11). Die Funktion hat die Parameter `x` (abh. Variable), `g` (Faktor) sowie `k` (Gruppenanzahl). Diese Anweisungen wurden bereits in den Beispielen R-4.4, R-5.2 sowie R-6.2b erläutert.

```
p.collect <- function (x,g,k)
{ i<-0
  pwerte<-0
  for (i1 in 1:(k-1)) { for (i2 in (i1+1):k)
   {i<-i+1
    erg <- pFligPoli(x[as.numeric(g)==i1],x[as.numeric(g)==i2])
    pwerte[i] <- erg$two.sided
    names(pwerte)[i] <- paste(i1,i2,sep=" - ")
  }}
 pwerte
}
```

### 15. 8. 2 pairwise.table

R bietet eine Funktion `pairwise.table` zur Durchführung paarweiser Vergleiche einschließlich p-Wert-Adjustierung mittels der in `p.adjust` (vgl. Kapitel 15.5.1) zur Verfügung stehenden Verfahren. Allerdings muss diese innerhalb einer eigenen Funktion aufgerufen werden. Auch hier kann letztlich diese Funktion nur mit einem einzigen statistischen Verfahren verknüpft werden. Auf diese Weise arbeiten auch die Standardfunktionen `pairwise.t.test` und `pairwise.wilcox.test`. (Durch den Aufruf dieser Funktionen ohne `(..)` kann man sich den Code anzeigen lassen.)

Nachfolgend ein Beispiel für dieselbe Aufgabe wie im vorigen Abschnitt. Der wesentliche Teil steckt in der Funktion `compare.levels`, die als Ergebnis den p-Wert des gewünschten statistischen Verfahrens ausgibt, hier die Anweisung `pFligPoli(x1,x2)$two.sided`. Anschließend wird diese Funktion der Funktion `pairwise.table` als Parameter übergeben.

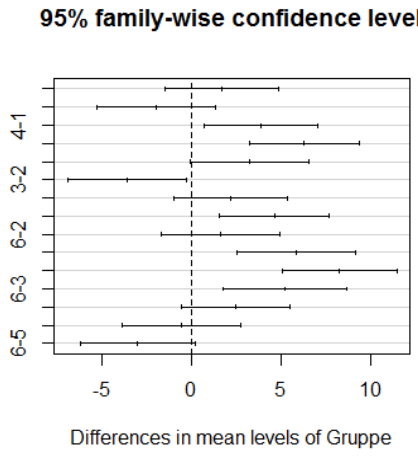
```
my.pairwise.test <- function(x,g,p.adjust.method)
{ g <- factor(g)
  compare.levels <- function(i, j) {
    x1<-x[as.integer(g) == i]
    x2<-x[as.integer(g) == j]
    pFligPoli(x1,x2)$two.sided }
  PVAL <- pairwise.table(compare.levels, levels(g), p.adjust.method)
  PVAL
}
```

## 15. 9 Grafische Darstellungen

Eine Form der grafischen Darstellung bei multiplen Mittelwertvergleichen ist die der Konfidenzintervalle, sofern diese für das jeweilige Verfahren berechnet werden können. Und zwar werden Konfidenzintervalle für die Mittelwertdifferenz aufgestellt. Für diese gilt die Regel: Enthält ein Intervall die „0“, so unterscheiden sich die Mittelwerte nicht. Enthält es nicht die „0“, so gilt die Differenz als signifikant auf dem entsprechenden Niveau des Konfidenzintervalls.

Die Darstellung wird mittels der Standardfunktion `plot` erzeugt, die als Argument das Ergebnis eines multiplen Mittelwertvergleichs erwartet, z.B.

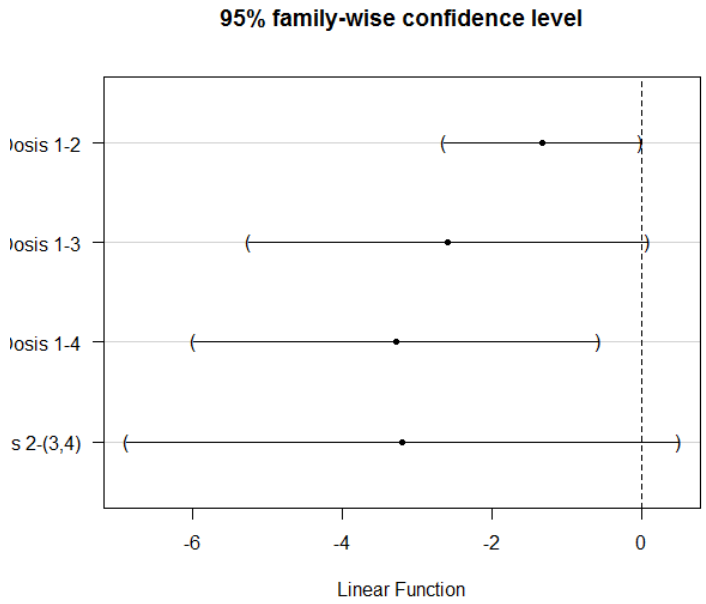
```
result <- TukeyHSD(aov(x~Gruppe,mydata10))
plot(result)
```



Eine andere Funktion, die Konfidenzintervalle erzeugt, ist z.B. `glht` aus dem Paket `multcomp`. Die grafische Darstellung des Ergebnisses aus Beispiel R-1.2b kann erzeugt werden mit:

```
plot(glht_aov2)
```

die sich allerdings auf die Ergebnisse mittels `single-step`-Adjustierung bezieht:



# 16. Anwendungen mit SPSS

## 16. 1 parametrische Vergleiche - unabhängige Stichproben

Bei SPSS gibt es keine Probleme mit der Auswahl der SPSS-Prozedur. Alle zur Verfügung stehenden Methoden können sowohl in `GLM` (Menü: Allgemeines lineares Modell) als auch in `Oneway` (Menü: Mittelwerte vergleichen -> einfaktorielle ANOVA) über den Button `Post Hoc` angefordert werden. Bei beiden Prozeduren sind Eingabemaske und Ausgabe völlig identisch. Bei mehrfaktoriellen Versuchsplänen wird (korrekterweise) die Fehlervarianz $MS_{Fehler}$ aus der Anova-Tabelle übernommen, sofern der Test diese Streuung verwendet.

In Kapitel 1.7 wurde bereits darauf aufmerksam gemacht, dass SPSS zwei Ausgabemodi hat: *paarweise Vergleiche* und *homogene Untergruppen*. Die Wahl hängt von dem jeweilig angeforderten Test ab. Manche, wie z.B. Tukeys HSD und Scheffé werden in beiden Varianten ausgegeben. Allerdings sollte sich die Wahl nicht nach der Art der Ausgabe richten, wenn auch die paarweisen Vergleiche den Vorteil haben, zum einen leichter verständlich zu sein, da jeder Vergleich „direkt“ abgelesen werden kann, und zum anderen p-Werte auszugeben. Dagegen erhält man bei der Ausgabe homogener Untergruppen nur die Information, ob sich zwei Mittelwerte auf dem vorgegebenen Signifikanzniveau unterscheiden. Letzteres wird übrigens über den Button `Optionen`, und dort bei „Signifikanzniveau“ eingestellt (Standard: 0,05). Dunnetts Kontrollgruppenvergleich erlaubt die Wahl zwischen ein- und zweiseitigen Tests.

Folgende Verfahren werden in SPSS angeboten:

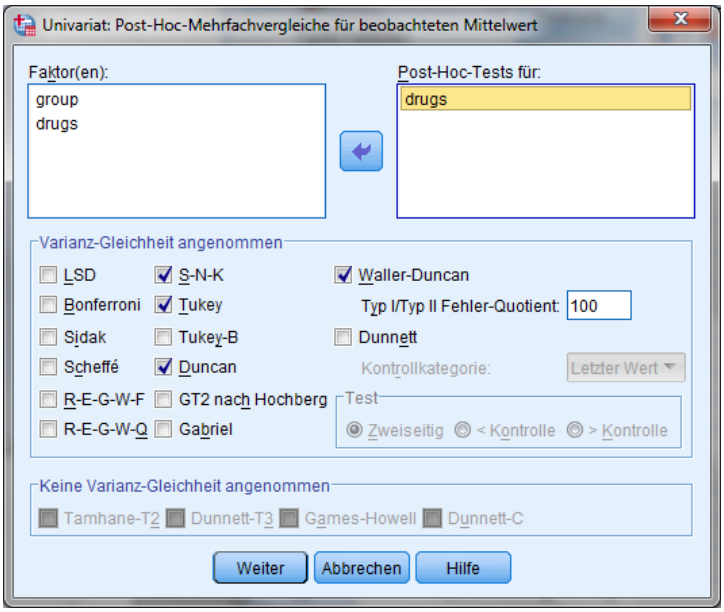
| für homogene Varianzen | | |
|---|---|---|
| | paarw. Vergl. | homog. Untergr. |
| Fisher LSD (ungeschützt) | x | |
| Fisher LSD mit Bonferroni-Adj. | x | x |
| Fisher LSD mit Sidak-Adj. | | x |
| Tukey HSD | | x |
| Tukey B | | x |
| Newman-Keuls S-N-K | | x |
| Duncan | | x |
| R-E-G-W-F | x | x |
| R-E-G-W-Q | x | x |
| Hochberg GT2 | | x |
| Gabriel | x | x |
| Waller-Duncan | x | |
| Scheffé | x | |
| Dunnett | x | |

| für inhomogene Varianzen | |
|---|---|
| | paarw. Vergl. |
| Tamhane T2 | x |
| Dunnett T3 | x |
| Dunnett C | x |
| Games-Howell | x |

Hinweis: Die nachfolgenden Referenzen auf Ergebnisse beziehen sich auf das Skript [99].

**Beispiel S-1:**

Zunächst soll ein Beispiel für den Fall homogener Varianzen gerechnet werden. Dazu werden die Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Das Ergebnis der Varianzanalyse enthält dort die Tabelle 4-4. Danach hat der Faktor `drug` einen signifikanten Einfluss ($p<0.001$). Dieser wird nun mittels eines multiplen Mittelwertvergleichs näher untersucht, wenn das auch angesichts der signifikanten Interaktion nicht unbedingt hilfreich ist. Da alle Voraussetzungen erfüllt sind, kann man sich z.B. für den Newman-Keuls-Test (SNK) entscheiden. Hier werden zu Demonstrationszwecken eine Reihe von Tests angefordert, was natürlich in der Praxis nicht zulässig ist, damit man sich nicht das „schönste“ Ergebnis aussuchen kann und damit ein weiteres α-Risiko verletzt. Nachfolgend die Eingabemaske:




Nachfolgend zunächst die Ausgabe der paarweisen Vergleiche, in der lediglich das Ergebnis des Tukey HSD-Tests erscheint:

Multiple Comparisons

Abhängige Variable: x

| | (I) | (J) | Mittlere Differenz (I-J) | Standard fehler | Sig. | 95%-Konfidenzintervall | |
|---|---|---|---|---|---|---|---|
| | | | | | | Untergrenze | Obergrenze |
| Tukey-HSD | 1 | 2 | -1,11 | ,645 | ,333 | -2,88 | ,66 |
| | | 3 | -1,63 | ,662 | ,093 | -3,45 | ,20 |
| | | 4 | -3,11* | ,645 | ,000 | -4,88 | -1,34 |
| | 2 | 1 | 1,11 | ,645 | ,333 | -,66 | 2,88 |
| | | 3 | -,51 | ,622 | ,841 | -2,22 | 1,20 |
| | | 4 | -2,00* | ,603 | ,014 | -3,66 | -,34 |
| | 3 | 1 | 1,63 | ,662 | ,093 | -,20 | 3,45 |
| | | 2 | ,51 | ,622 | ,841 | -1,20 | 2,22 |
| | | 4 | -1,49 | ,622 | ,105 | -3,20 | ,22 |
| | 4 | 1 | 3,11* | ,645 | ,000 | 1,34 | 4,88 |
| | | 2 | 2,00* | ,603 | ,014 | ,34 | 3,66 |
| | | 3 | 1,49 | ,622 | ,105 | -,22 | 3,20 |

Der Fehlerterm ist Mittel der Quadrate(Fehler) = 1,637

*. Die mittlere Differenz ist auf dem 0,05-Niveau signifikant.

Die Ausgabe ist weitgehend selbsterklärend. Anzumerken ist, dass signifikante Unterschiede bei den Mittelwertdifferenzen durch * markiert sind.

Anschließend die Ausgabe homogener Untergruppen für alle vier angeforderten Tests:

| | | N | Untergruppe | | |
|---|---|---|---|---|---|
| | | | 1 | 2 | 3 |
| Student-Newman-Keuls[a,b,c] | 1 | 7 | 4,00 | | |
| | 2 | 9 | 5,11 | 5,11 | |
| | 3 | 8 | | 5,63 | |
| | 4 | 9 | | | 7,11 |
| | Sig. | | ,092 | ,425 | 1,000 |
| Tukey-HSD[a,b,c] | 1 | 7 | 4,00 | | |
| | 2 | 9 | 5,11 | | |
| | 3 | 8 | 5,63 | 5,63 | |
| | 4 | 9 | | 7,11 | |
| | Sig. | | ,074 | ,114 | |
| Duncan[a,b,c] | 1 | 7 | 4,00 | | |
| | 2 | 9 | 5,11 | 5,11 | |
| | 3 | 8 | | 5,63 | |
| | 4 | 9 | | | 7,11 |
| | Sig. | | ,092 | ,425 | 1,000 |
| Waller-Duncan[a,b,d] | 1 | 7 | 4,00 | | |
| | 2 | 9 | 5,11 | 5,11 | |
| | 3 | 8 | | 5,63 | 5,63 |
| | 4 | 9 | | | 7,11 |

Mittelwerte für Gruppen in homogenen Untergruppen werden angezeigt.
Grundlage: beobachtete Mittelwerte.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 1,637.

a. Verwendet Stichprobengrößen des harmonischen Mittels = 8,162

b. Die Größen der Gruppen ist ungleich. Es wird das harmonische Mittel der Größe der Gruppen verwendet. Fehlerniveaus für Typ I werden nicht garantiert.

c. Alpha = 0,05

d. Quotient der Schwere des Fehlers für Typ 1/Typ 2 = 100.

Ein paar Erläuterungen zu dieser Ausgabe. Die Ergebnisse für den SNK- und den Duncan-Test sind identisch. Diese besagen:

- Die Mittelwerte 4,00 und 5,11 unterschieden sich nicht (da gleiche Spalte, also gleiche Untergruppe).
- Die Mittelwerte 5,11 und 5,63 unterschieden sich nicht.
- Der Mittelwert 7,11 bildet eine 1-elementige Untergruppe.

Stellt man nun eine Dreiecksmatrix auf, in der sowohl die Zeilen als auch die Spalten den Mittelwerte entsprechen, werden die o.a. nicht signifkanten Vergleiche mit einem - gekennzeichnet. Die verbliebenen Felder entsprechen signifikanten Vergleichen und werden mit x markiert:

| | 4,00 | 5,11 | 5,63 | 7,11 |
|---|---|---|---|---|
| 4,00 | | - | **x** | **x** |
| 5,11 | | | - | **x** |
| 5,63 | | | | **x** |

Beim Tukey-Test entfallen die signifikanten Unterschiede 4,00 - 5,63 sowie 5,63 - 7,11. Beim Test von Waller-Duncan entfällt nur der letzte Vergleich.

Die letzte Zeile eines Tests wird mit „sig." gekennzeichnet. Sie enthält das Signifikanzniveau für den Unterschied der Mittelwerte der jeweiligen Spalte, d.h. der homogenen Untergruppe. Dieser p-Wert liegt zwangsläufig über dem gewählten $\alpha$, da sich ja die Mittelwerte einer Untergruppe nicht mehr signifikant unterschieden. Ein Wert nahe $\alpha$ deutet darauf hin, dass die entsprechende Untergruppe nicht mehr ganz homogen ist, so z.B. der Wert 0,074 für die erste Untergruppe bei dem Test von Tukey. Der darin enthaltene Mittelwert 5,63 ist auch bei allen anderen Tests in dieser Untergruppe nicht mehr enthalten.

**Beispiel S-2:**

Dazu werden die Beispieldaten 3 (`mydata3`) verwendet (vgl. Kapitel 4 in [99]). Da die Daten keine Varianzhomogenität aufwiesen, wurde dort eine 1-faktorielle Varianzanalyse für inhomogene Varianzen (in der Prozedur `Oneway`) mit Tests von Brown & Forsythe sowie von Welch durchgeführt, mit folgendem Ergebnis für den Faktor `dosis` der Beispieldaten 3:

| Robuste Testverfahren zur Prüfung auf Gleichheit der Mittelwerte | | | | |
|---|---|---|---|---|
| x | | | | |
| | Statistik[a] | df1 | df2 | Sig. |
| Welch-Test | 3,879 | 3 | 13,308 | ,034 |
| Brown-Forsythe | 3,218 | 3 | 18,618 | ,047 |

Bei den multiplen Mittelwertvergleichen gibt es auch Verfahren, die bei heterogenen Varianzen anwendbar sind (vgl. Kapitel 6). In SPSS ist das allerdings nur bei einer 1-faktoriellen Analyse möglich. Nur dann sind - wie im vorigen Beispiel erläutert - in der Eingabemaske die Tests unter „Keine Varianzgleichheit angenommen" auswählbar. Nachfolgend die Ausgabe für den Test von Games & Howell sowi Dunnetts C-Test:

| | (I) dosis | (J) dosis | Mittlere Differenz (I-J) | Standard fehler | Sig. | 95%-Konfidenzintervall | |
|---|---|---|---|---|---|---|---|
| | | | | | | Untergrenze | Obergrenze |
| Games-Howell | 1 | 2 | -1,33 | ,548 | ,145 | -3,08 | ,41 |
| | | 3 | -2,58 | 1,104 | ,154 | -5,97 | ,80 |
| | | 4 | -3,28 | 1,109 | ,054 | -6,61 | ,05 |
| | 2 | 1 | 1,33 | ,548 | ,145 | -,41 | 3,08 |
| | | 3 | -1,25 | 1,031 | ,637 | -4,55 | 2,05 |
| | | 4 | -1,94 | 1,037 | ,302 | -5,17 | 1,28 |
| | 3 | 1 | 2,58 | 1,104 | ,154 | -,80 | 5,97 |
| | | 2 | 1,25 | 1,031 | ,637 | -2,05 | 4,55 |
| | | 4 | -,69 | 1,412 | ,960 | -4,77 | 3,38 |
| | 4 | 1 | 3,28 | 1,109 | ,054 | -,05 | 6,61 |
| | | 2 | 1,94 | 1,037 | ,302 | -1,28 | 5,17 |
| | | 3 | ,69 | 1,412 | ,960 | -3,38 | 4,77 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dunnett-C | 1 | 2 | -1,33 | ,548 | | -3,28 | ,61 |
| | | 3 | -2,58 | 1,104 | | -6,32 | 1,15 |
| | | 4 | -3,28 | 1,109 | | -6,93 | ,38 |
| | 2 | 1 | 1,33 | ,548 | | -,61 | 3,28 |
| | | 3 | -1,25 | 1,031 | | -4,65 | 2,15 |
| | | 4 | -1,94 | 1,037 | | -5,26 | 1,37 |
| | 3 | 1 | 2,58 | 1,104 | | -1,15 | 6,32 |
| | | 2 | 1,25 | 1,031 | | -2,15 | 4,65 |
| | | 4 | -,69 | 1,412 | | -5,29 | 3,90 |
| | 4 | 1 | 3,28 | 1,109 | | -,38 | 6,93 |
| | | 2 | 1,94 | 1,037 | | -1,37 | 5,26 |
| | | 3 | ,69 | 1,412 | | -3,90 | 5,29 |
| Der Fehlerterm ist Mittel der Quadrate(Fehler) = 4,864 | | | | | | | |

Obwohl von den vier in SPSS für heteroge Varianzen verfügbaren Tests der von Games & Howell der stärkste ist, wird kein Einzelvergleich als signifikant ausgewiesen. (Der kleinste p-Wert ist 0,054.) Bei Dunnetts C-Test fehlen die p-Werte. Hier kann man Unterschiede auf dem 5%-Niveau daran erkennen, dass das 95%-Konfidenzintervall für die Mittelwertdifferenz (die beiden letzten Spalten) nicht die „0“ enthalten. Letzteres ist aber bei allen Vergleichen der Fall, so dass keiner signifikant ist.

In Kapitel 4.3.3 in [99] wurde die abhängige Variable logarithmiert, um die Varianzen zu „stabilisieren“ und damit eine „normale“ Varianzanalyse durchzuführen. Der p-Wert für den Effekt von `dosis` beträgt 0,039. Die Logarithmierung der Kriteriumsvariablen wird auch von Day & Quinn [101] favorisiert. Nachfolgend der Newman-Keuls-Test für `ln(x)`:

| lnx | | | | |
|---|---|---|---|---|
| | dosis | N | Untergruppe | |
| | | | 1 | 2 |
| Student-Newman-Keuls[a,b,c] | 1 | 6 | 1,6215 | |
| | 2 | 10 | 1,8640 | 1,8640 |
| | 3 | 8 | 1,9866 | 1,9866 |
| | 4 | 9 | | 2,0767 |
| | Sig. | | ,054 | ,340 |
| Der Fehlerterm ist Mittel der Quadrate(Fehler) = ,088. | | | | |
| a. Verwendet Stichprobengrößen des harmonischen Mittels = 7,956 | | | | |

der genau einen Unterschied nachweist: Gruppe 1 (1,6215) mit Gruppe 4 (2,0767).

## 16. 2 parametrische Vergleiche - abhängige Stichproben

In SPSS stehen für Messwiederholungsfaktoren keine multiplen Mittelwertvergleiche zur Verfügung. Es gibt folgende Alternativen:

- orthogonale Kontraste, z.B. mittels Standardkontrast „wiederholt“ (repeated)
- Durchführung der klassischen Tests (u.a. der aus Kapitel 4) „per Hand“
- paarweise Vergleiche mit α-Adjustierungen (siehe Kapitel 16.4)

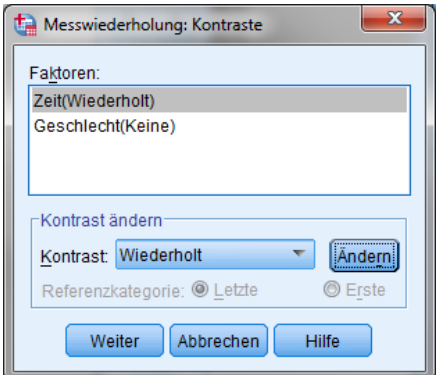
**Beispiel S-3:**

Dazu werden die Beispieldaten 5 (`winer518`) verwendet (vgl. Kapitel 5 in [99]). Für den Faktor `Zeit` wurde dort ein signifikanter Haupteffekt (p<0,001) nachgewiesen (vgl. dort Tabelle 6-3). Dieser soll nun näher untersucht werden.

Über den orthogonalen Standardkontrast „wiederholt“ (*repeated*) werden die Stufen 1 mit 2 sowie 2 mit 3 verglichen. Wenn die Stufen eine natürliche Reihenfolge haben, wie hier die `Zeit`, dann sind über diese Kontraste bereits nahezu erschöpfende Informationen erhältlich. Dazu ist entweder in der Syntax `polynomial` durch `repeated` zu ersetzen:

```
GLM t1 t2 t3
  /wsfactor=Zeit 3 repeated
```

oder über den Button `Kontraste` für den Faktor `Zeit` der Kontrast von `polynomial` auf `repeated` zu ändern (`Ändern`-Button nicht vergessen!).




Die Ergebnisse für diese Kontraste:

| Tests der Innersubjektkontraste | | | | | | |
|---|---|---|---|---|---|---|
| Quelle | Zeit | Quadratsumme vom Typ III | df | Mittel der Quadrate | F | Sig. |
| Zeit | Niveau 1 vs. Niveau 2 | 22,500 | 1 | 22,500 | 10,976 | ,011 |
| | Niveau 2 vs. Niveau 3 | 36,100 | 1 | 36,100 | 13,885 | ,006 |
| Zeit * Geschlecht | Niveau 1 vs. Niveau 2 | 84,100 | 1 | 84,100 | 41,024 | ,000 |
| | Niveau 2 vs. Niveau 3 | 44,100 | 1 | 44,100 | 16,962 | ,003 |
| Fehler(Zeit) | Niveau 1 vs. Niveau 2 | 16,400 | 8 | 2,050 | | |
| | Niveau 2 vs. Niveau 3 | 20,800 | 8 | 2,600 | | |

Im ersten Block „`Zeit`“ werden die Stufen 1 mit 2 sowie 2 mit 3 verglichen. Beide Vergleiche sind signifikant. Da bei der Anova auch die Interaktion signifikant war (p<0,001), gibt der Block „`Zeit*Geschlecht`“ weitere Informationen. Auch hier werden die Stufen 1 mit 2 sowie 2 mit 3 verglichen, allerdings unter dem Aspekt, wieweit diese Unterschiede vom Faktor `Geschlecht` beeinflusst werden. Dies ist sozusagen die Detailanalyse der Interaktion. Und dieser Einfluss ist ebenfalls hoch signifikant.

**Beispiel S-4:**

Eine Alternative bieten die klassischen Tests, die die Varianz $MS_{Fehler}$ verwenden, u.a. die aus Kapitel 4. Diese kann bei Analysen mit Messwiederholungen durch $MS_{Residuen}$ ersetzt werden. Dies ist die Fehlervarianz für den zu testenden Effekt und wird in SPSS direkt unter den Ergebnissen für den Effekt ausgegeben: `Fehler(Zeit)`. Voraussetzung ist allerdings die strikte

Varianzhomogenität, hier die Spherizität, die über den Mauchly-Test geprüft wird. Dieser Test verlief negativ (p=0,778 in Tabelle 6-3 [99]), so dass dieses Vorgehen möglich ist.

Es wird der Newman-Keuls-Test durchgeführt. Dazu werden nun die kritischen Werte $c_d$ für die Mittelwertdifferenzen errechnet. Nach Kapitel 4.3 Formel 4-3 ist dies

$$c_d = q_\alpha(r, df_{Fehler})\sqrt{MS_{Fehler}/n}$$

Zunächst werden die Quantilswerte für die studentized range-Verteilung $q_\alpha(r,df_{Fehler})$ bestimmt, und zwar für die Abstände $r$=2 und $r$=3. Diese sind in SPSS über die Funktion `IDF.SRANGE` erhältlich. Zur Berechnung von $c_d$:

- Man erzeugt sich im Dateneditor einen kleinen Datensatz mit einer Variablen `r`,
- gibt für diese 2 Werte ein: 2 und 3,
- errechnet mittels der o.a. Funktion `IDF.SRANGE(1-α, r, dfFehler)` die q-Werte in der Variablen `qr` für α=0,05 und $df_{Fehler}$ = 16 ,
- dividiert diese durch $\sqrt{MS_{Fehler}/10}$ mit der Ergebnisvariablen `cd`, wobei für Faktor `Zeit` $MS_{Fehler} = MS_{Residuen}$ =1,317.

Die hier verwendeten Werte sind der Tabelle 6-3 in [99] zu entnehmen. Die SPSS-Syntax hierfür:

```
Compute qr=IDF.SRANGE(0.95 , r ,16).
Compute cd=qr*sqrt(1.317/10)
Execute.
```

Die Ergebnisse sind dem Dateneditor zu entnehmen:

| | r | qr | cd |
|---|---|---|---|
| 1 | 2,00 | 3,00 | 1,09 |
| 2 | 3,00 | 3,65 | 1,32 |

Die Mittelwerte für die 3 Zeitpunkte sind nun der Größe nach zu ordnen:
2,50 (t3) - 4,40 (t2) - 5,90 (t1).

Der erste `cd`-Wert 1,09 (zu r=2) dient zum Vergleich benachbarter Mittelwerte, der zweite 1,32 (zu r=3) zum Vergleich von Mittelwerten, die zwei Stufen auseinander liegen. Somit erhält man folgende Ergebnisse für die Vergleiche:

| Vergleich | Mittelwertdifferenz | Abstand | cd | Ergebnis |
|---|---|---|---|---|
| t1 - t2 | 1,50 | 1 | 1,09 | signifikant |
| t1 - t3 | 3,40 | 2 | 1,32 | signifikant |
| t2 - t3 | 1,90 | 1 | 1,09 | signifikant |

Für die Durchführung des Tukey HSD-Tests wird nur der `cd`-Wert für r=3 (allgemein: r=$k$) benötigt, da alle Differenzen nur mit diesem verglichen werden. Das Ergebnis ist dasselbe. Für den Duncan-Test müsste man für die q-Werte auf Tabellen in der Lietratur zurückgreifen, da diese Verteilung nicht in SPSS verfügbar ist.

## 16. 3 nichtparametrische Vergleiche

SPSS bietet in seinen beiden 1-faktoriellen nichtparametrischen Varianzanalysen, dem Kruskal-Wallis-H-Test bei unabhängigen Stichproben bzw. der Friedman-Varianzanalyse bei abhängigen Stichproben, multiple Mittelwertvergleiche an:

- das Verfahren von Campbell und Skillings (vgl. Kapitel 12.7)
- paarweise Vergleiche mit der α-Adjustierung von Bonferroni (vgl. Kapitel 2.1).

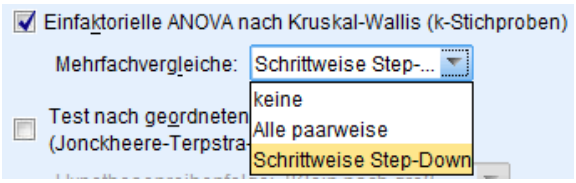
Beide Verfahren werden kurz vorgestellt, wenn auch das erste unbedingt vorzuziehen ist.

**Beispiel S-5:**

Es werden wieder die Beispieldaten 2 (`mydata2`) verwendet (vgl. Kapitel 4 in [99]). Und zwar wird der Faktor `drugs` untersucht, zunächst über einen globalen Test mittels Kruskal-Wallis-H-Test, zusätzlich mit Mittelwertvergleichen über das Verfahren von Campbell und Skillings. Die SPSS-Syntax dafür:

```
NPTests
  /Independent Test (x) Group (drugs) Kruskal_Wasllis(compare=stepwise).
```

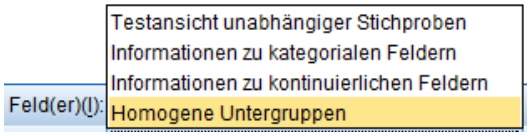
Erfolgt die Testauswahl über die Eingabemasken, so ist wie folgt auszuwählen:




Zunächst die Ausgabe des H-Tests für den globalen Mittelwertunterschied:

| | Nullhypothese | Test | Sig. | Entscheidung |
|---|---|---|---|---|
| 1 | Die Verteilung von ist über Kategorien von Drugs gleich. | Kruskal-Wallis-Test unabhängiger Stichproben | ,011 | Nullhypothese ablehnen. |

Nach Doppelklick auf dieses Ergebnis im Ausgabefenster öffnet sich ein weiteres. Um die Mittelwertvergleiche anzuzeigen, muss im rechten Fensterteil rechts unten




„Homogene Untergruppen“ ausgewählt werden. Die Ausgabe dazu folgt weiter unten.

Es werden zwei homogene Untergruppen (vgl. Kapitel 1.7) ermittelt: zum einen die Gruppen 1, 2 und 3 (türkis) und zum anderen die Gruppen 2, 3 und 4 (rot). Daraus resultiert genau ein signifikanter Unterschied: die Gruppen 1 und 4. In der jeweiligen Gruppenzeile wird der mittlere Rang dieser Gruppe angezeigt. Die in den letzten Zeilen angezeigten Ergebnisse beziehen sich auf einen Kruskal-Wallis-H-Test, der die Homogenität der jeweiligen Untergruppe überprüft. So zeigt der p-Wert 0,067 in der roten Untergruppe an, der nur knapp über dem Signifikanzniveau von 0,05 liegt, dass diese vergleichsweise inhomogen ist.

**Homogene Untergruppen auf der Basis von**

| | | Untergruppe | |
|---|---|---|---|
| | | 1 | 2 |
| Stichprobe[1] | 1,000 | 9,000 | |
| | 2,000 | 15,000 | 15,000 |
| | 3,000 | 17,625 | 17,625 |
| | 4,000 | | 24,667 |
| Teststatistik | | 3,614 | 5,415 |
| Sig. (2-seitiger Test) | | ,164 | ,067 |
| Angepasste Sig. (2-seitiger Test) | | ,164 | ,067 |

Homogene Untergruppen beruhen auf asymptotischen Signifikanzen. Das Signifikanzniveau ist ,05.

[1]Jede Zelle zeigt den durchschnittlichen Stichprobenrang von .

Ersetzt man in der Syntax `stepwise` durch `pairwise` bzw. in der Eingabemaske „Schrittweise Step-down“ durch „Alle paarweise“, so werden paarweise Vergleiche mit der Bonferroni-Adjustierung durchgeführt. In diesem Fall unterscheiden sich die Ergebnisse qualitativ nicht voneinader.

| Stichprobe1-Stichprobe2 | Test-statistik | Std.-Fehler | Std. Test-statistik | Sig. | Angep. Sig. |
|---|---|---|---|---|---|
| 1,000-2,000 | -6,000 | 4,801 | -1,250 | ,211 | 1,000 |
| 1,000-3,000 | -8,625 | 4,931 | -1,749 | ,080 | ,482 |
| 1,000-4,000 | -15,667 | 4,801 | -3,263 | ,001 | ,007 |
| 2,000-3,000 | -2,625 | 4,629 | -,567 | ,571 | 1,000 |
| 2,000-4,000 | -9,667 | 4,491 | -2,152 | ,031 | ,188 |
| 3,000-4,000 | -7,042 | 4,629 | -1,521 | ,128 | ,769 |

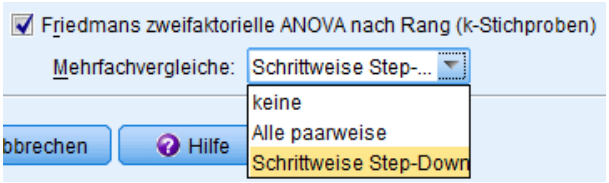
Jede Zeile testet die Nullhypothese, dass die Verteilungen von Stichprobe 1 und Stichprobe 2 gleich sind.
Asymptotische Signifikanzen (2-seitige Tests) werden angezeigt. Das Signifikanzniveau ist ,05.

**Beispiel S-6:**

Es werden wieder die Beispieldaten 4 (`winer518`) verwendet (vgl. Kapitel 5 in [99]), die Messwiederholungen auf dem Faktor Zeit aufweisen. Und zwar wird der Faktor `Zeit` untersucht, zunächst über einen globalen Test mittels Frieman-Test, zusätzlich mit Mittelwertvergleichen über das Verfahren von Campbell und Skillings. Die SPSS-Syntax dafür:

```
NPTests
  /Related Test (t1 t2 t3) Friedman (compare=stepwise) .
```

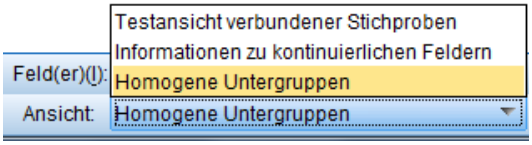
Erfolgt die Testauswahl über die Eingabemasken, so ist wie folgt auszuwählen:



Zunächst die Ausgabe der Friedman-Varianzanalyse für den globalen Mittelwertunterschied:

| | Nullhypothese | Test | Sig. | Entscheidung |
|---|---|---|---|---|
| 1 | Die Verteilungen von Zeit 1, Zeit 2 and Zeit 3 sind gleich. | Friedmans Zweifach-Rangvarianzanalyse verbundener Stichproben | ,008 | Nullhypothese ablehnen. |

Nach Doppelklick auf dieses Ergebnis im Ausgabefenster öffnet sich ein weiteres. Um die Mittelwertvergleiche anzuzeigen, muss im rechten Fensterteil rechts unten




„Homogene Untergruppen“ ausgewählt werden. Nachfolgend die Ausgabe dazu:

**Homogene Untergruppen**

| | | Untergruppe 1 | Untergruppe 2 |
|---|---|---|---|
| Stichprobe[1] | Zeit 3 | 1,300 | |
| | Zeit 2 | 2,100 | 2,100 |
| | Zeit 1 | | 2,600 |
| Teststatistik | | 1,600 | ,400 |
| Sig. (2-seitiger Test) | | ,206 | ,527 |
| Angepasste Sig. (2-seitiger Test) | | ,206 | ,527 |

Homogene Untergruppen beruhen auf asymptotischen Signifikanzen. Das Signifikanzniveau ist ,05.

[1]Jede Zelle zeigt den durchschnittlichen Stichprobenrang.

Es werden zwei homogene Untergruppen ermittelt: zum einen die Zeitpunkte 2 und 3 (türkis) und zum anderen die Zeitpunkte 1 und 2 (rot). Daraus resultiert genau ein signifikanter Unterschied: die Zeitpunkte 1 und 3. In der jeweiligen Gruppenzeile wird der mittlere Rang dieser Gruppe angezeigt. Die in den letzten Zeilen angezeigten Ergebnisse beziehen sich auf eine Friedman-Varianzanalyse, die die Homogenität der jeweiligen Untergruppe überprüft.

Vorsicht: Für die ausgewählten Variablen müssen „Variable Labels“ definiert sein, damit die Ergebnisse „sichtbar“ sind.

Auch hier können wie im vorigen Beispiel alternativ paarweise Vergleiche mit einer Bonferroni-Korrektur angefordert werden.

## 16. 4 α-Adjustierungen

**Beispiel S-7:**

Noch einmal zurück zu dem Beispieldatensatz 3 (`winer518`). In den Beipielen 3-1 und 3-2 wurden zwei Methoden vorgestellt, die Mittelwerte paarweise zu vergleichen. Hier nun die dritte Variante: paarweise t-Tests mit α-Adjustierungen. Die Syntax fierfür:

```
T-Test Pairs=t2 t3 t3 with t1 t1 t2 (paired).
```

mit folgendem Ergebnis:

| | | Mittelwert differenz | T | df | Sig. (2-seitig) |
|---|---|---|---|---|---|
| Paaren 1 | t2 - t1 | -1,500 | -1,419 | 9 | ,189 |
| Paaren 2 | t3 - t1 | -3,400 | -5,667 | 9 | ,000 |
| Paaren 3 | t3 - t2 | -1,900 | -2,237 | 9 | ,052 |

Einzig der Vergleich t3-t1 erbringt einen p-Wert (< 0,001) kleiner als $\alpha$=0,05. Da dieser sogar kleiner als $\alpha' = 0{,}05/3 = 0{,}0167$ ist, dem nach Bonferroni korrigierten $\alpha$, kann dieser Vergleich als signifikant angesehen werden. Andere, bessere Adjustierungen brauchen nicht probiert zu werden, da das von Bonferroni das konservativste und damit schlechteste Verfahren ist.

**Beispiel S-8:**

Es sollen die 4 Versuchsbedingungen des Beispieldatensatzes 9 (`mydata9`, vgl. Kapitel 1.9) verglichen werden. Eine Varianzanalyse ergibt zunächst, dass die Varianzhomogenität (Sphärizität) nicht gegeben ist. Der Test von Mauchly ist signifikant (p=0,041). Der F-Test der Varianzanalyse unter Verwendung der Huynh-Feldt-Korrektur zeigt einen signifikanten Einfluss der Versuchsbedingungen an (p=0,001). Damit scheiden die klassischen Verfahren, wie in Kapitel 16.2 demonstriert, aus. Orthogonale Kontraste schränken zu sehr ein, da letztlich alle Vergleiche von Interesse sind. So bleibt nur noch der Ausweg paarweiser Vergleiche mittels t-Test und $\alpha$-Korrektur. Einige der in Kapitel 2 vorgestellten Methoden sollen hier angewandt werden, sofern sie per Hand leicht durchführbar sind.

Zunächst werden die t-Tests durchgeführt .:

| | Mittelwert differenz | T | df | Sig. (2-seitig) |
|---|---|---|---|---|
| Vergleich V1 - V2 | -1,500 | -2.666 | 9 | ,02581 |
| Vergleich V1 - V3 | -2.300 | -5.438 | 9 | ,00041 |
| Vergleich V1 - V4 | -4.400 | -3.836 | 9 | ,00399 |
| Vergleich V2 - V3 | -0.800 | -1.350 | 9 | .20995 |
| Vergleich V2 - V4 | -2.900 | -2.567 | 9 | .03036 |
| Vergleich V3 - V4 | -2.100 | -2.215 | 9 | .05401 |

und die p-Werte der Größe nach aufsteigend sortiert:

| i | p-Wert | Vergleich | Bonferroni | Holm | Hochberg | Benjamini Hochberg | Li | Blanchard Rochain |
|---|---|---|---|---|---|---|---|---|
| | | | $\alpha$/6 | $\alpha$/(6-i+1) | $\alpha$/(6-i+1) | $\alpha$ i/6 | $\alpha(1-p_6)/(1-\alpha)$ | $\alpha(1-\alpha)$ i/(6-i+1) |
| 1 | ,00041 | V1 - V3 | 0.00833 | 0.00833 | 0.00833 | 0.00833 | 0.04158 | 0.00792 |
| 2 | ,00399 | V1 - V4 | 0.00833 | 0.01000 | 0.01000 | 0.01667 | 0.04158 | 0.01900 |
| 3 | .02581 | V1 - V2 | 0.00833 | 0.01250 | 0.01250 | 0.02500 | 0.04158 | 0.03563 |
| 4 | .03036 | V2 - V4 | 0.00833 | 0.01667 | 0.01667 | 0.03333 | 0.04158 | 0.06333 |
| 5 | .05401 | V3 - V4 | 0.00833 | 0.02500 | 0.02500 | 0.04167 | 0.04158 | 0.11875 |
| 6 | .20995 | V2 - V3 | 0.00833 | 0.05000 | 0.05000 | 0.05000 | 0.05000 | 0.28500 |

Bei *Holm* werden nacheinander die $p_i$ mit den errechneten $\alpha'$ verglichen. Beim ersten negativen Vergleich wird abgebrochen. Danach sind nur die beiden ersten Vergleiche (V1-V3 und V1-V4) signifikant.

Bei *Hochberg* werden zwar dieselben $\alpha'$ benutzt, aber es darf geprüft werden, ob für größere *i* doch noch ein Vergleich signifikant ist. Das ist aber nicht der Fall, so dass dasselbe Resultat herauskommt.

Bei *Benjamini* & *Hochberg* sind zunächst die ersten beiden Vergleiche signifikant. Da aber auch der vierte Vergleich signifikant ist, gelten die Vergleiche 1 bis 4 als signifikant.

Bei *Li* ist schnell zu sehen, dass ebenfalls die Vergleiche 1 bis 4 signifikant sind.

Bei *Blanchard & Roquain* sind ebenfalls die Vergleiche 1 bis 4 signifikant.

Zum Vergleich ist die (nicht empfehlenswerte) Adjustierung von *Bonferroni* angeführt, die zahlenmäßig mit der von *Sidak* ähnlich ist.

# 17. Fazit

Dieses Kapitel befindet sich noch in Arbeit.

## Literaturhinweise

### Allgemeines zur nichtparametrischen Statistik und Varianzanalyse

(Die Nummern sind identisch mit denen des u.a. Skripts von Lüpsen [99].)

[ 3 ] E. Brunner & U. Munzel (*2002): Nichtparametrische Datenanalyse - unverbundene Stichproben*, Springer, ISBN 3-540-43375-9

[ 6 ] Conover, W. J. & Iman, R. L. (1981): *Rank transformations as a bridge between parametric and nonparametric statistics*. American Statistician 35 (3): 124–129.

[ 7 ] Cochran, W.G. (1950): *The comparison of percentages in matched samples*. Biometrika 3

[ 8 ] Lunney, G.H. (1970): *Using Analysis of Variance with a dichtomous dependent variable: an empirical study*. Journal of Educational Measurement Volume 7, Issue 4

[13] B.J. Winer et.al. (1991): *Statistical Principles in Experimental Design*, Wiley, New York

[22] Michael G. Akritas, Steven F. Arnold & Edgar Brunner (1997): *Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs*, Journal of the American Statistical Association, Volume 92, Issue 437 , pages 258-265

[97] Myles Hollander, Douglas A. Wolfe, Eric Chicken (2014): *Nonparametric Statistical Methods*, Wiley

[99] Haiko Lüpsen (2014): Varianzanalysen - Prüfung der Voraussetzungen und Übersicht der nichtparametrischen Methoden sowie praktische Anwendungen mit R und SPSS, UNiversität zu Köln, http://www.uni-koeln.de/~luepsen/statistik/buch/nonpar-anova.pdf

### Übersichten Mittelwertvergleiche

[101] R.W. Day & G.P. Quinn (1989): Comparisons of Treatments after an Analysis of Variance in Ecology, Ecological Monographs, Vol. 59, No. 4, pp 433-463

[102] Rand R. Wilcox (2013): *New Statistical Procedures for the Social Sciences: Modern Solutions To Basic Problems*, Psychology Press, Lawrence Erlbaum Assoc

[103] Richard Gonzalez: *Contrasts and Post Hoc tests (Lecture Notes)* (2009), University of Michigan, Ann Arbor, http://www-personal.umich.edu/~gonzo/coursenotes/file3.pdf

[104] http://en.wikipedia.org/wiki/Multiple_comparisons_problem

[105] http://en.wikipedia.org/wiki/Post-hoc_analysis

[106] John A. Rafter, Martha L. Abell, James P. Braselton (2002): *Multiple Comparison Methods for Means*, SIAM Review, Vol. 44, No. 2, pp 259-278

[107] Gerard E. Dallal (2001/2012): *Multiple Comparison Procedures* aus *The Little Handbook of Statistical Practice*
http://www.jerrydallal.com/LHSP/mc.htm

[108] Juliet P. Shaffer (1995): *Multiple Hypothesis Testing*, Annual Reviews Psychology, 46, pp. 561-584

[109] C.V. Rao & U. Swarupchand (2009): *Multiple Comparison Procedures - a Note and a Bibliography* , Journal of Statistics, Volume 16, 2009, pp. 66-109

[110] SPSS: *Appendix 10: Post Hoc Tests* ftp://ftp.boulder.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/app10_post_hoc_tests.pdf

[111] A.C. Tamhane (1979): *A Comparison of Procedures for Multiple Comparisons of Means With unequal Variances*, Journal of the American Statistical Association, Vol 74, Number 365

[112] Jason Hsu (1999): *Multiple Comparisons: Theory and Methods*, Chapman & Hall

**α-Adjustierungen**

[121] http://en.wikipedia.org/wiki/False_discovery_rate

[122] Christopher R. Genovese (2003): *A Tutorial on False Discovery Control*, http://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf

[123] SAS/STAT(R) 9.22 User's Guide: *The MULTTEST Procedure: p-Value Adjustments*, http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_multtest_sect014.htm

[124] S. Garcia, A. Fernandez, J.Luengo und F. Herrera (2010): *Advanced nonparametric Tests for Multiple Comparisons in the Design of Experiments in Computational Intelligence and Data Mining: Experimental Analysis of Power*, Information Sciences 180, pp 2044-2064

[125 ]R.E. Blakesley, S. Mazumdar, M.A. Dew, P. R. Houck, G. Tang, C. F. Reynolds, III, and M.A. Butters (2009): *Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research*, Neuropsychology, Mar 2009; 23(2), pp 255–264

[126] http://en.wikipedia.org/wiki/Bonferroni_correction

[127] http://en.wikipedia.org/wiki/Holm-Bonferroni_method

[128] Benjamini & Hochberg (1995): *Controlling the false discovery rate: A practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society, Serie B,57, 289–300.

[129] Heyse, J. and Rom, D. (1988), *Adjusting for Multiplicity of Statistical Tests in the Analysis of Carcinogenicity Studies*, Biometrical Journal, 30, pp 883–896.

[130] Jianjun Li (2007): *A two-step rejection procedure for testing multiple hypotheses*, Journal of Statistical Planning and Inference 138, pp 1521 – 1527

[131] http://en.wikipedia.org/wiki/Resampling_(statistics)

[132] Cavan Reilly (2013): *Multiple comparison procedures*, University of Minesota, http://www.biostat.umn.edu/~cavanr/asfChap4-2.pdf

[134] P. H. Westfall, S. S. Young (1993): *Resampling based Multiple Testing: Examples and Methods for p-value Adjustment*, New York, Wiley

[135] Yongchao Ge, Sandrine Dudoit and Terence P. Speed (2003): *Resampling-based multiple testing for microarray data analysis*, Berkeley University, Technical Report 633 http://statistics.berkeley.edu/sites/default/files/tech-reports/633.pdf

[136] Gilles Blanchard and Etienne Roquain (2009): *Adaptive False Discovery Rate Control under Independence and Dependence*, Journal of Machine Learning Research 10, pp 2837-2871

[137] R.J. Simes (1986): An improved Bonferroni procedure for multiple tests of significance, Biometrika, Vol 73, pp 751-754

**Kodierungen und Kontraste**

[151] R Library: Contrast Coding Systems for categorical variables: http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

[152] *Regression with SPSS: Chapter 5: Additional coding systems for categorical variables in regression analysis* : http://www.ats.ucla.edu/stat/spss/webbooks/reg/chapter5/spssreg5.htm

**Multiple Mittelwertvergleiche**

[161] Henry I. Braun & John W. Tukey (1983). *Multiple comparisons through orderly partitions: The maximum subrange procedure*, in H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement, A festschrift for Frederic M. Lord, pp. 55-65.

[162] http://en.wikipedia.org/wiki/Heteroscedasticity-consistent_standard_errors

[163] David A. Freedman (2006): *On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors"*, The American Statistician, Vol. 60, No. 4, pp. 299-302 http://www.stat.berkeley.edu/~census/mlesan.pdf

[164] S.G. Carmer und W.M. Walker (1985): *Pairwise Multiple Comparisons of Treatment Means in Agronomic Research*, Journal of Agronimic Education, Vol. 14

[165] A. J. Scott and M. Knott (1974): *A Cluster Analysis Method for Grouping Means in the Analysis of Variance*, Biometrics, Vol. 30, No. 3, pp. 507-512

[166] K.R. Gabriel (1978): *A simple Method of Multiple Comparisons of Means*, Journal of the American Statistical Association, Vol 73, Number 364

[167] P.E.Rudolph (1988): *Robustness of Multiple Comparison Procedures: Treatment versus Control*, Biometrical Journal 50, 1, 41-45

[168] R. Marcus, E. Peritz, K. Gabriel (1976). *On closed testing procedures with special reference to ordered analysis of variance*. Biometrika 63, pp 655–660

[169] Scott E. Maxwell (1989): *Pairwise Multiple Comparisons in Repeated Measures Designs*, Journal of Educational Statistics, Vol. 5, No. 3, pp. 269-287

[170] Christy Chuang-Stein and Donald M. Tong (1995): *Multiple comparisons procedures for comparing several treatments with a control based on binary data*, Statistics in Medicine, Volume 14, Issue 23, pp 2509–2522

[171] Gao, X., Alvo, M., Chen, J. and Li, G (2008): *Nonparametric Multiple Comparison Procedures For Unbalanced One-Way Factorial Designs*. Journal of Statistical Planning and Inference. Vol 138, pp 2574-2591

[172] Ullrich Munzel and Ajit C. Tamhane (2002): *Nonparametric Multiple Comparisons in Repeated Measures Designs for Data with Ties*, Biometrical Journal 44 Nr. 6, pp 762–779

[173] Frank Konietschke, Ludwig A. Hothorn, Edgar Brunner (2012): *Rank-based multiple test procedures and simultaneous confidence intervals*, Electronic Journal of Statistics Vol. 6, 738–759

[174] http://en.wikipedia.org/wiki/Van_der_Waerden_test

[175] Gary Ramseyer & Tse-Kia Tcheng (1973): T*he Robustness of the Studentized Range Statistic to Violations of the Normality and the Homogeneity of Variance Assumptions*, American Educational Research Journal, Vo. 10, No 3, pp 235-240

[176] Clemens S. Bernhardson (1975): T*ype I Error Rates When Multiple Comparison Procedures Follow a Significant F Test of ANOVA*, Biometrics, Vol. 31, No. 1, pp. 229-232

[177] James F. Zolman (1993): *Experimental Design and Statistical Inference*, Oxford University Press

[178] I. Einot and K.R. Gabriel (1975): *A Study of Powers of Several Methods of Multiple Comparisons*, Journal of the American Statistical Association, Vol 70, Number 361

[179] Marisela Abundis (2001): Multiple Comparison Procedures in Factorial Designs using the Alogned Rank Transformation, Thesis, Texas Tech University

[180] John R. Donoghue (1998): Implementing Shaffer‘s Multiple Comparison Procedure for Large Number of Groups, Educational Testing Services, Princeton, RR-98-47

[181] R.A. Brown & A.B. Forsythe (1974): *The small sample behaviour of some statistics which test the equality of several means*, Technometrics 16, 129-302

[182] SAS/STAT(R) 13.2 User's Guide: The GLM Procedure - Multiple Comparisons http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glm_details29.htm

[183] SAS/STAT(R) 13.2 User's Guide: The GLM Procedure - Multiple Comparisons http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_npar1way_details20.htm

[184] W. Liu (1997): *On Step-up Tests for Comparing Several Treatments*: Statistica Sinica 7, pp 957-972

**Sonstiges**

[191] http://cran.r-project.org/web/packages/agricolae/vignettes/tutorial.pdf