Multiple Mittelwertvergleiche
- parametrisch und nichtparametrisch -
sowie α-Adjustierungen
mit
praktischen Anwendungen mit R und SPSS

Version 2.0
(10.10.2020)

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 sowohl am RRZK als auch an anderen Einrichtungen 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 α-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 α-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 diskutierte der angewandten Statistik, ist eigentlich schon seit der Veröfentlichung der klassischen Tests von „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 Stärke (Power).

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.

Lesehinweise

Die Ergebnistabellen aus R und SPSS sind zum Teil verkürzt wiedergegeben, d.h. Teile, die nicht zum Verständnis erforderlich sind, fehlen hier. Die Beispieldatensätze sind alle im Anhang aufgeführt zusammen mit Anova-Tabelle und Ergebnis der Varianz-Homogenitätstests.

Versionen

Version 2.0 (28.4.2019): Ergänzung zahlreicher, nicht notwendigerweise neuerer Methoden für Mittelwertvergleiche (z.B. Hayter) und α-Adjustierungen (z.B. Holland-Copenhaver oder Gatekeeping), Berücksichtigung diverser neuer R-Pakete.

Version 0.92 (19.2.2019): diverse kleine Fehlerkorrekturen

Version 0.91 (31.10.2016): individuelle Kontraste in R korrigiert (Kapitel 3.4 und 3.7)

Inhaltsverzeichnis

1. Einleitung 1

1. 1 Die Ausgangssituation 1
1. 2 Multiple Vergleiche 2
1. 3 Klassifizierung der Verfahren 2
1. 4 Besondere Verteilungen 2
1. 5 Statistische Fehler 3
1. 6 Die Problematik der multiplen Vergleiche 4
1. 7 Darstellung der Ergebnisse 5
1. 8 Hinweise auf Literatur 6
1. 9 Beispiele und Datensätze 6

2. Paarweise Vergleiche mit α-Adjustierungen 7

2. 1 Verfahren von Bonferroni und Sidak 7
2. 2 Schrittweise Verfahren 8
2. 2. 1 step-down-Verfahren: Holm und Hochberg 9
2. 2. 2 Benjamini & Hochberg 9
2. 2. 3 step-up-Verfahren: Hochberg, Hommel und Rom 10
2. 2. 4 Jianjun Li 10
2. 2. 5 Shaffers S1 und S2 10
2. 2. 6 Holland & Copenhaver 11
2. 3 Auf Resampling basierende Verfahren 11
2. 3. 1 Westfalls free step-down-Verfahren 12
2. 3. 2 min.P- und max.T-Verfahren 12
2. 4 Adaptive Verfahren 12
2. 5 Andere Verfahren 13
2. 5. 1 Sidak-Variante von Tukey-Ciminera-Heyse 13
2. 5. 2 Fixed sequence 13
2. 5. 3 Fallback 13
2. 5. 4 Weitere Verfahren 14
2. 6 Grundlegende Methoden 14
2. 6. 1 Simes-Verfahren 14
2. 6. 2 Closure-Prinzip 14
2. 7 Multiple endpoints - mehrere abhängige Variablen 15
2. 7. 1 Gatekeeping 16
2. 7. 2 Parametrisches multiples Testen 17
2. 8 Adjustierungen in R und SPSS 17

3. Kontraste und Kodierungen 19

3. 1 Grundlagen 19
3. 2 Standard-Kontraste 21
3. 3 Auswahl der Kontraste 23
3. 4 Nichtparametrische Kontraste für die RT-, ART- und Puri & Sen-Verfahren 24
3. 5 Universelles Verfahren für Kontraste 27
3. 6 Kontraste bei logistischen Regressionen 29
3. 7 Kontraste für Messwiederholungen und Interaktionen 29

4. Die klassischen Verfahren von Fisher, Tukey, Newman-Keuls und Duncan 32

4. 1 Fishers LSD (least significant difference) 32
4. 2 Tukey HSD (honestly significant differnce) 33
4. 3 Student-Newman-Keuls (SNK) 33
4. 4 Duncan 34
4. 5 Tukey-b 35
4. 6 Hayter 35

5. Schrittweise Verfahren von Ryan und Welsch 36

5. 1 REGWF (Ryan-Einot-Gabriel-Welsch mit F-Tests) 36
5. 2 REGWQ (Ryan-Einot-Gabriel-Welsch mit Q-Tests) 36
5. 3 Peritz 37
5. 4 Welsch 37

6. Vergleiche bei inhomogenen Varianzen 39

6. 1 Games & Howell 39
6. 2 Dunnetts C 39
6. 3 Dunnetts T3 und Hochbergs GT2 40
6. 4 Tamhane T2 40
6. 5 Hubers Sandwich-Schätzer 40

7. Verallgemeinerte Kontraste 41

7. 1 Dunn-Bonferroni 41
7. 2 Scheffé 41
7. 3 Tukey 41
7. 4 Welch, Ury & Wiggins 42
7. 5 Brown & Forsythe und Kaiser & Bowden 42
7. 6 Simes 42

8. Weitere multiple Vergleiche 44

8. 1 Waller-Duncan 44
8. 2 Scott & Knott 44
8. 3 Gabriel 45
8. 4 Daytons Modell-Tests 45

9. Vergleich mit einer Kontrollgruppe 46

9. 1 Dunnett 46
9. 2 Schrittweise Dunnett-Verfahren 47
9. 3 Steel 47
9. 4 Gao und Nemenyi 47

10. Versuchspläne mit Messwiederholungen 48

11. Mehrfaktorielle Versuchspläne 50

12. Nichtparametrische Methoden 51

12. 1 Anwendung parametrischer Verfahren 51
12. 2 Aligned Rank Transform-Verfahren 52
12. 3 Nemenyi 52
12. 4 McSweeney 53
12. 5 Dwass-Steel-Critchlow-Fligner 53
12. 6 Conover & Iman 53
12. 7 Gao 54
12. 8 Campbell & Skillings 54
12. 9 Quade 55
12. 10 van der Waerden 55
12. 11 Fligner-Policello 56
12. 12 Dunn 56
12. 13 Hayter-Nasimoto-Wright 57
12. 14 Hayter-Stone 57
12. 15 Baumgartner-Weiß-Schindler 57
12. 16 Durbin 58
12. 17 Jonckheere-Terpstrata 58
12. 18 Nichtparametrische Kontraste 58

13. Voraussetzungen 60

13. 1 Ungleiche n_i und ungleiche Varianzen 60
13. 2 Messwiederholungen und Split Plot Designs 61
13. 3 Parametrische Tests 61
13. 3. 1 studentized range-Verteilung 61
13. 3. 2 F- bzw. t-Verteilung 62
13. 3. 3 Weitere Hinweise 62
13. 4 Nichtparametrische Tests 63

14. Zur Auswahl eines Tests oder die Qual der Wahl 64

14. 1 Mittelwertvergleiche mit oder ohne eine Varianzanalyse 64
14. 2 Zur Stärke einiger Tests 65
14. 3 Parametrisch im Fall von Varianzhomogenität 66
14. 4 Parametrisch im Fall von Varianzheterogenität 66
14. 5 Nichtparametrisch: allgemein 67
14. 6 Nichtparametrisch: Auswahl passend zur Varianzanalyse 67
14. 7 α-Adjustierungen 68

15. Anwendungen mit R 69

15. 1 Parametrische Vergleiche - unabhängige Stichproben 70
15. 1. 1 agricolae 70
15. 1. 2 PMCMRplus 73
15. 1. 3 multcomp 75
15. 1. 4 stats: TukeyHSD 78
15. 1. 5 DTK: DTK.test 79
15. 1. 6 DunnettsTests 79
15. 1. 7 ExpDes 80
15. 1. 8 mutoss 81

15. 1. 9 npmv 83
15. 1. 10 rgabriel 84
15. 2 Parametrische Vergleiche - abhängige Stichproben 85
15. 2. 1 agricolae 85
15. 2. 2 nlme 87
15. 3 Nichtparametrische Vergleiche - unabhängige Stichproben 88
15. 3. 1 agricolae 88
15. 3. 2 nparcomp 90
15. 3. 3 PMCMRplus 92
15. 3. 4 NSM3 93
15. 4 Nichtparametrische Vergleiche - abhängige Stichproben 95
15. 4. 1 agricolae 95
15. 4. 2 PMCMRplus 96
15. 4. 3 NSM3 99
15. 5 Allgemeine α-Adjustierungen 99
15. 5. 1 stats: p.adjust 99
15. 5. 2 mutoss 100
15. 5. 3 Mediana 102
15. 5. 4 multcomp 104
15. 6 Paarweise Mittelwertvergleiche mit α-Adjustierungen 104
15. 6. 1 stats: pairwise.t.test 104
15. 6. 2 stats: pairwise.wilcox.test 106
15. 6. 3 ggstatsplot: pairwise_p 108
15. 6. 4 asht: tukeyWelsch 109
15. 6. 5 multcomp 109
15. 7 Andere Hilfsmittel - Verfahren von Ryan und Welsch 109
15. 8 Weitere Pakete und Funktionen 111
15. 9 Hilfsfunktionen 111
15. 9. 1 p.collect 112
15. 9. 2 pairwise.table 112
15. 10 Grafische Darstellungen 112

16. Anwendungen mit SPSS 114

16. 1 Parametrische Vergleiche - unabhängige Stichproben 114
16. 2 Parametrische Vergleiche - abhängige Stichproben 118
16. 3 Nichtparametrische Vergleiche 121
16. 4 α-Adjustierungen 125

Anhang **128**

1 Beispieldatensätze

1. 1 Beispieldaten 2 (mydata2) 128
1. 2 Beispieldaten 3 (mydata3) 129
1. 3 Beispieldaten 4 (winer518) 130
1. 4 Beispieldaten 6 (winer568) 131
1. 5 Beispieldaten 9 (mydata9) 132
1. 6 Beispieldaten 10 (mydata10) 133

Literaturhinweise **134**

1. Einleitung

Wilcox (2013) 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 α einhalten, auch auf die Gefahr hin, dass die Wahrscheinlichkeit für den Fehler 2. Art β 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 Verfahren 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 Ryan. Darüber hinaus gibt es auch Verfahren, bei denen mit Paarvergleichen angefangen wird und in Abhängigkeit vom Ergebnis die Hypothesen verallgemeinert werden (*step-up procedures*), z.B. die Methoden von Hochberg und Tamhane.

Unabhängig von dieser Klassifizierung gibt es auch die Einteilung in solche Verfahren, die klassischerweise Mittelwertdifferenzen 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 ineinander 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*, häufig mit Q abgekürzt. 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 (für alle Gruppen 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 α von der Anzahl der Vergleiche *m* und den Freiheitsgraden *df* der Fehlervarianz s^2 ab.

Darüberhinaus hat Hayter eine Variante der o.a. studentized range-Verteilung eingeführt: die *one-sided studentized range Verteilung.* Die Idee geht auf Bohrer zurück, der engere Konfidenzintervalle konstruieren konnte, wenn alle Konstrastkoeffizienten nichtnegativ sind.

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 α 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 β und ist i.a. unbekannt, genauer: Sie kann für die hier besprochenen Tests kaum errechnet werden. Man weiß lediglich, dass β mit größer werdendem α 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 (α)	m_0
	H_0 falsch (H_1)	C (β)	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 α 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 α=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 α 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). Eine ausführliche Beschreibung mit mathematischem Hintergrund bietet Genovese (2003). Die Berücksichtigung der FDR ist weniger stringent als die von α, d.h. Verfahren, die dieses berücksichtigen, verfügen über mehr Power, halten auf der anderen Seite nicht unbedingt ein vorgegebenes α 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 α, z.B. α=0,05 , bewertet, so erhielte man weitaus mehr zufällige und damit falsche Signifikanzen, als das α 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. Auf der anderen Seite ist diese Rechnung nur dann richtig, wenn über das Ergebnis der globalen Varianzanalyse nichts bekannt ist. Wieviele falsche Signifikanzen man erhält, wenn das Ergebnis des globalen Tests signifikant ist, ist ein noch ungelöstes Problem. Das Ziel ist es dagegen, dass für eine solche *Familie von Vergleichen* insgesamt das Fehlerrisiko (*familywise error rate, FWER*) dem vorgegebenen α entspricht. Dieses Problem ist auch ausführlich in Wikipedia beschrieben.

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. Zunächst sind einige Bücher zu nennen: Eine umfassende Beschreibung der Verfahren mit Angabe sämtlicher Formeln und umfassende Kriterien zur Auswahl der Verfahren bietet Wilcox (2013), desweiteren Kirk (1995) und Hsu (1996). An Artikeln sind zu nennen: Day & Quinn (1989), die einen umfassenden und insbesondere auch leicht verständlichen Überblick über die Methoden, deren Unterschiede und Eigenschaften geben, Cribbie (2000), der die wichtigsten Methoden einschließlich der Formeln und Methodenvergleiche beschreibt, sowie Dmitrienko et al. (2012) im Rahmen eines Buches über klinische Studien. Eine sehr gute Einführung in die Problematik, Beschreibung der bekanntesten Tests sowie deren Unterschiede bieten Rafter, Abell & Braselton (2002), Gonzalez (2009) und Tamhane (1979) für den Fall inhomogener Varianzen. Eine gute Erläuterung der Unterschiede ist auch bei Dallal (2001/2012) zu finden. Eine Übersicht der verschiedenen Methodengruppen gibt Shaffer (1995). Eine kompakte Übersicht der multiple Mittelwertvergleiche einschließlich Formeln bieten die SAS-Dokumentation sowie die SPSS-Dokumentation, natürlich soweit die Verfahren in SAS bzw. SPSS verfügbar sind. Eine umfassende Literaturliste zu den Verfahren ist bei Rao & Swarupchand (2009) zu finden.

Einen guten Überblick der α-Adjustierungen bietet das Benutzerhandbuch von SAS für die Prozedur Multtest. Einige neuere Verfahren sind auch bei Garcia et al. (2010) kurz beschrieben sowie bei Blakesley et al. (2009), die auch einen Vergleich der Methoden erörtern.

Hinsichtlich der generellen Voraussetzungen für die Varianzanalyse, die ja auch für die Mittelwertvergleiche gelten, sei auf Skript über Varianzanalysen (Lüpsen, 2020) verwiesen.

1. 9 Beispiele und Datensätze

Weitestgehend werden die Datensätze aus dem Skript über Varianzanalysen (Lüpsen, 2019) verwendet, zumal die dort gefundenen Anova-Ergebnisse hier mittels multipler Mittelwertvergleiche weiter untersucht werden. Die Datensätze sind im Anhang aufgeführt, dazu die wichtigsten Anova-Tabellen, auf die in diesem Skript Bezug genommen wird. Die Tabellennummern wurden aus dem Skript über Varianzanalysen übernommen, jedoch um die Eindeutigkeit zu wahren, mit einem V versehen. Also mit Tabelle “V 4-6“ ist eine Anova-Tabelle aus dem Skript über Varianzanalysen gemeint, die sowohl dort als auch im Anhang dieses Skripts zu finden ist. Die Beispiele sind durchnummeriert. Anwendungen mit R haben das Präfix “R“, solche mit SPSS das Präfix “S“.

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 χ^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*, *Blanchard & Roquain*, oder das *parametric multiple testing*, 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* (siehe Wikipedia): Sei α‘ die vorzugebende Irrtumswahrscheinlichkeit, dann ist (1- α‘) die Wahrscheinlichkeit, H_0 *nicht* fälschlich auf dem α‘-Niveau abzulehnen, und $(1 - \alpha')^m$ die Wahrscheinlichkeit, H_0 bei allen *m* Vergleichen fälschlich auf dem α‘-Niveau abzulehnen. Damit ist die Irrtumswahrscheinlichkeit, H_0 bei *m* Vergleichen fälschlich abzulehnen $1 - (1 - \alpha')^m$. Daraus ergibt sich folgende 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 α' anstatt α verglichen.

Verwendet man anstatt dieses näherungsweise errechneten α' das aus o.a. Gleichung 2-1 exakt errechnete α' , 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, weil andere bessere Verfahren, wie das von Li, ebenso einfach mit der Hand durchzuführen sind. Allerdings gewinnt die Bonferroni-Adjustierung im Zusammenhang mit neueren Adjustierungsmethoden wie z.B. Gatekeeping (vgl. Kapitel 2.6) wieder an Bedeutung, bei denen meistens Hypothesen für mehrere Variablen (*muliple endpoints* genannt) getestet werden and anschließend die p-Werte adjustiert werden müssen. Dort wird die Bonferroni-Adjustierung in einer gewichteten Form angewandt. Dazu werden den *k* Hypothesen Gewichte $w_1,...,w_k$ mit $0<w_i<1$ sowie $w_1+...+w_k=1$ entsprechend ihrer Bedeutung gegeben, und jede wird mit $\alpha' = \alpha * w_i$ verglichen. Sind alle Gewichte gleich, also $w_i=1/k$, so resultiert daraus die „nor male“ Korrektur.

2. 2 Schrittweise Verfahren

Eine Verbesserung, d.h. ein weniger konservatives Verhalten, bringen die schrittweisen Verfahren. Während bei den o.a. klassischen Adjustierungen von Bonferroni und Sidak alle Vergleiche mit demselben korrigierten α' 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 α' vergrößert. Allerdings sind einige wenig transparent und nicht mehr mit der Hand zu rechnen. Dazu zählen insbesondere die Adjustierungen von *Hommel, Rom*, *Shaffer* sowie *Holland-Copenhaver*, 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. (2010) kurz beschrieben.

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

Eine marginale Verbesserung gegenüber der Bonferroni-Adjustierung bieten die Verfahren von *Holm* (siehe Wikipedia) und *Hochberg* (Benjamini & Hochberg, 1995). 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 $\alpha/(m\text{-}i\text{+}1)$.

Holm geht der Reihe nach vor: p_1 wird nun mit α/m verglichen. Ist dieser Vergleich (H_1) signifikant, kann p_2 verglichen werden und zwar mit $\alpha/(m\text{-}1)$. Ist auch dieser Vergleich (H_2) signifikant, darf p_3 verglichen werden und zwar mit $\alpha/(m\text{-}2)$ 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*.

Holm hat dieses Verfahren verallgemeinert auf den Fall gewichteter Hypothesen mit Gewichten $w_i>0$. Dazu werden berechnet $q_i=p_i/w_i$. Die adjustierten p_i' errechnen sich dann:

$$p'_1 = min(1, q_1) \qquad p'_i = min(1, max(p'_{i-1}, (w_i + \ldots + w_m)q_i)) \qquad (i > 1)$$

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 $\alpha/(m\text{-}i\text{+}1)$ wird p_i mit $1 - (1-\alpha)^{(m-i+1)}$ verglichen. Dieser Wert ergibt sich, wenn man für α' die Formel (2-2) durch (2-1) ersetzt. *Finner* bietet eine weitere Variante: Anstatt des Vergleichs von p_i mit $\alpha/(m\text{-}i\text{+}1)$ 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* (Benjamini & Hochberg, 1995) dar (vgl. auch Wikipedia). 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 (Benjamini & Hochberg, 1995) und in Wikipedia 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* (Heyse & Rom, 1988). Die α‘, 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 α_i (z.B. α_{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* (Li, 2007), bei dem auch die Vergleiche hinsichtlich des p-Wertes wieder aufsteigend sortiert werden, so dass $p_1 \le \ldots \le p_i \le \ldots \le p_m$ mit entsprechenden Null-Hypothesen H_1,..., H_m . 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:

$$p_i' = \frac{p_i}{p_i + 1 - p_m}$$

wobei sich eine Ordnung der p-Werte erübrigt, weil $p_m = \max(p_i)$ ist. Zum einen ist die Berechnung der α' 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 α liegt. In dem Fall liegt α' nahe bei α .

2. 2. 5 Shaffers S1 und S2

Auch die α-Adjustierungen von *Juliet Shaffer* bringen eine Verbesserung. 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 *i* 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 α-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 (1998) einen Algorithmus für das S2-Verfahren von Shaffer entwickelt, der auch für den Fall inhomogener Varianzen anwendbar ist. Der Algorithmus ist zwar in Donoghe (1998) aufgelistet, aber dennoch schwer verständlich und ohne Computer nicht durchführbar. Eine einigermaßen verständliche Beschreibung bietet Cribbie (2000).

2. 2. 6 Holland & Copenhaver

Eine Verbesserung des Verfahrens von Holm bringt die α-Adjustierung von *Holland & Copenhaver*. Es basiert auf der o.a. Methode von Shaffer. Auch hier kann nur eine grobe Idee des Verfahrens vermittelt werden. Die Handhabung selbst ist so kompliziert, dass sie ohnehin nur mit Computerprogrammen angewandt werden können.

Auch hier 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. Beginnend mit H_1 wird eine Hypothese H_i abgelehnt, wenn

$$p_i \le \alpha_i = 1-(1-\alpha)^{1/C_k}$$

wobei C_k wie beim Verfahren von Shaffer die maximale Anzahl der Nullhypothesen ist, die wahr sein könnten (und irrtümlich abgelehnt werden könnten), wenn im vorigen Schritt die Nullhypothese abgelehnt worden war. Das Verfahren endet, wenn im *i*-ten Schritt $p_i > \alpha_i$. Eine einigermaßen verständliche Beschreibung bietet Cribbie (2000).

2. 3 Auf Resampling basierende Verfahren

Klassischerweise werden Tests auf der Basis eines Modells mittels Wahrscheinlichkeitsrechnung entwickelt. Beim *Resampling* (vgl. Wikipedia) 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 für die 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. Eine gut verständliche Beschteibung der Resampling-Verfahren geben Dmitrienko et al. (2012).

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 α/m verglichen, sondern mit einem Quantil der Verteilung der 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 bei Reilly (2013) 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 (1993) 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 α/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 sowie Varianten sind beschrieben von Ge et al. (2003).

2. 4 Adaptive Verfahren

Bei *adaptiven Verfahren* geht es um die Einhaltung der FDR (vgl. Kapitel 1.5). Sei π_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 π_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 π_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 π_0 zu schätzen. Im zweiten Schritt werden dann alle Hypothesen überprüft.
- *one-stage-Verfahren*: Bei diesen wird für ein vorgebenes π_0 ein step-down- oder step-up-Verfahren eingesetzt.

Eine gute Übersicht bieten *Blanchard & Roquain* (2009), 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

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

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 beim multiplen Mittelwertvergleich bei unabhängigen Stichproben, das α-Risiko nicht mehr strikt eingehalten wird.

2. 5. 2 Fixed sequence

Beim fixed sequence-Verfahren werden die Hypothesen apriori, z.B. hinsichtlich ihrer Bedeutung, geordnet: $H_1,...,H_m$, wobei H_1 die bedeutendste und H_m die unbedeutendste ist. Diese werden dann auch genau in dieser Reihenfolge getestet. Sind $p_1,...,p_m$ die entsprechenden p-Werte, dann wird H_i abgelehnt, wenn $p_j \le \alpha$ für $j \le i$, d.h. wenn bereits alle vorangegangenen Hypothesen abgelehnt worden sind. Hierbei wird ie FWER korrekt eingehalten.

Ein Beispiel: Wird ein Präparat in 4 verschieden starken Dosierungen ($D_1,...,D_4$) verabreicht, dann kann davon ausgegangen werden, dass bei einem Vergleich mit einer Placebo-Gruppe der Unterschied umso größer ausfällt, je stärker die Dosierung gewählt wurde. In diesem Fall wäre H_1 der Vergleich Placebo-D_4, H_2 der Vergleich Placebo-D_3, usw.

Es gibt auch ein reversed fixed sequence-Verfahren mit derselebn Ausgangsbasis wie oben. H_1 wird anhand α geprüft. Wird H_1 abgelehnt, so werden auch alle übrigen H_i abgelehnt. In den folgenden Schritten werden $H_1,...,H_{i-1}$ angenommen und $H_i,...,H_m$ verworfen, wenn $p_j>\alpha$ für alle $j<i$ und $p_j \le \alpha$ für $j \ge i$. Da $p_i \ge p_j$ für $i>j$ gilt allgemein, dass die p_i zugleich die adjustierten p-Werte sind.

2. 5. 3 Fallback

Die Ausgangsbasis beim fallback-Verfahren ist ähnlich wie beim o.a. fixed sequence-Verfahren: die Hypothesen $H_1,...,H_m$ sind apriori geordnet, allerdings ähneln die α–Adjustierungen

der Holm-Methode. Den k Hypothesen werden wie bei der gewichteten Bonferroni-Adjustierung Gewichte $w_1,...,w_k$ mit $0<w_i<1$ sowie $w_1+...+w_k=1$ entsprechend ihrer Bedeutung gegeben. Sind $p_1,...,p_m$ die entsprechenden p-Werte, dann werden für die H_i folgende Entscheidungen getroffen:

- Schritt 1: H_1 ist mit $\alpha_1=w_1\alpha$ zu testen. H_1 wird abgelehnt, wenn $p_1 \le \alpha_1$.
- Schritte i=2,...,m: H_i ist mit $\alpha_i=\alpha_{i-1} + w_i\alpha$ zu testen, wenn H_{i-1} abgelehnt worden war, bzw. mit $\alpha_i=w_i\alpha$, wenn H_{i-1} angenommen worden war. H_i wird abgelehnt, wenn $p_i \le \alpha_i$.

Im Gegensatz zur Holm-Strategie kann hier weitergetestet werden, auch wenn in einem Schritt die Null-Hypothesen angenommen worden war. Daher auch der Name *fallback*. Da das closure-Prinzip hier angewandt wird, wird auch die FWER eingehalten. Das fallback-Verfahren ist stärker als das gewichtete Bonferroni-Verfahren. Und bei w_1=1 resultiert daraus die fixed sequence-Methode.

2. 5. 4 Weitere Verfahren

Eine gute Zusammenstellung weiterer Adjustierungsverfahren ist in der Dokumentation des Statistiksystems SAS zu finden.

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* (1986) 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: $p_1 \le \ldots \le p_i \le \ldots \le p_m$, dann wird H_0 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 $t_{(1)} \geq t_{(2)} \geq \ldots \geq t_{(m)}$ die der Größe nach geordneten Testgrößen zu deren Test (also Realisationen entsprechender Zufallsvariablen $T_{(1)}, T_{(2)},\ldots,T_{(m)}$). Bezeichnen $H_{(1)}, H_{(2)},\ldots, H_{(m)}$ 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 $H_{(i)},\ldots, H_{(m)}$ 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 Multiple endpoints - mehrere abhängige Variablen

Bei den bisher aufgeführten Verfahren standen meistens multiple Hypothesen im Vordergrund, die sich alle auf dieselbe Variable bezogen, bei der Mittelwertunterschiede unter verschiedenen Bedingungen untersucht wurden. Hier resultieren dagegen die multiplen Hypothesen aus einer Hypothese, z.B. eines Mittelwertvergleichs, die für mehrere Variablen untersucht wird. Und insbesondere bei einer größeren Anzahl von Variablen, in diesem Zusammenhang meistens *multiple endpoints* genannt, im Deutschen am besten mit *K.O-Kriterien* zu ums chreiben, wird eine α-Adjustierung der Ergebnisse erforderlich. Im Gegensatz zu den bislang aufgeführten Verfahren wird hier eine Gewichtung der Hypothesen vorgenommen, z.B. wenn das Ergebnis einer Variable wichtiger für den Ausgang einer Untersuchung ist als das anderer Variablen. Die wichtigsten und für den Ausgang einer Studie entscheidenden werden *primary endpoints* genannt. Gibt es nur zwei oder drei, aber von gleicher Bedeutung, so heißen diese *co-primary endpoints.* Üblicherweise wird für deren Tests keine α-Adjustierung vorgenommen. Die darauf basierenden Verfahren zur α-Adjustierung kommen aus den medizinischen Anwendungen, wo es z.B. um den Test der Wirksamkeit von neuen Medikamenten oder Methoden geht. Haupteinsatzgebiet sind die α-Adjustierungen beim Test einer Hypothese, z.B. dem Vergleich einer Versuchs- mit einer Kontrollgruppe, bei mehreren Variablen. Andere Anwendungen sind z.B. der Vergleich einer oder mehrerer Versuchsgruppen mit einer Kontroll- und einer Placebo-

Gruppe für eine Variable. Einen ausführlichen Text zu diesem Thema bietet das US Department of Health (2017).

2. 7. 1 Gatekeeping

Maßgeblichen Anteil an der Entwicklung des Gatekeeping-Verfahrens hat Alex Dmitrienko (2007). Ausgangspunkt sind eine Reihe von Hypothesen, die entsprechend ihrer Bedeutung hierarchisch in Familien gruppiert werden: F_1 für die Gruppe der wichtigsten Hypothesen $H_{11},...,H_{1n1}$, danach F_2 mit den nächstwichtigeren $H_{21},...,H_{2n2}$, usw. Insgesamt *m* Familien mit $n=n_1+...+n_m$ Hypothesen. Innerhalb jeder Familie muss das α-Risiko, genauer der FWER, eingehalten werden, wodurch eine α-Adjustierung erforderlich wird. Allerdings spielen beim Test der Familie *i* die Tests der Familien $j>i$ keine Rolle, während umgekehrt die Tests der Familien $j<i$ berücksichtigt werden. Damit wird sichergestellt, dass der Test der wichtigen Hypothesen nicht von den Ergebnissen der unwichtigen Hypothesen beeinflusst wird. Als Basis dienen die (nicht adjustierten) p-Werte aller *n* Hypothesen: $p_{11}, p_{12},...,p_{m1}, p_{m2},...$ Man beginnt mit der ersten Familie F_1 . Hier werden nun zwei Fälle unterschieden:

- Serielles Gatekeeping: Nur wenn alle Hypothesen von F_1 abgelehnt worden sind, d.h. wenn $(max(\tilde{p}_{11}, \tilde{p}_{12}, \ldots) \leq \hat{\alpha}$, werden die Hypothesen von F_2 überprüft.
- Paralleles Gatekeeping: Wenn mindestens eine Hypothese von F_1 abgelehnt worden ist, d.h. wenn $min(\tilde{p}_{11}, \tilde{p}_{12}, \ldots) \leq \hat{\alpha}$, werden die Hypothesen von F_2 überprüft.

Hier sind die $\tilde{p}_{ij}$ die im jeweiligen Schritt adjustierten p_{ij}-Werte. Dieses Verfahren wird nun für $i=2,..,m$ fortgesetzt. Hieraus ergibt sich auch der Name des Verfahrens: Jede Hypothesen-Familie F_i ist der Pförtner (engl. *gatekeeper*) für die darauf folgende Hypothesen-Familie F_{i+1}. Die adjustierten α bzw. p-Werte werden dann nach dem o.a.closure-Prinzips ermittelt. Das Gatekeeping bietet u.a. noch die folgenden Optionen:

- Die einzelnen Hypothesen einer Familie dürfen gewichtet werden.
- In jedem Schritt kann eine andere Methode zur α-Adjustierung eingesetzt werden.
- Für die ausgewählte α-Adjustierung kann eine Mittelung mit der Bonferroni-Adjustierung festgelegt werden. Dies geschieht über den *truncation*-Parameter γ $(0 \leq \gamma \leq 1)$ derart, dass für $\gamma=0$ die Bonferroni-Adjustierung und für $\gamma=1$ die ausgewählte α-Adjustierung entsteht.

Ein kleines Beispiel dazu: Wird ein neues Präparat an einer Versuchsgruppe getestet und die Wirkung über 4 Parameter (Variablen) A, B, C und D gemessen. Zum Vergleich werden diese 4 Parameter auch an Personen einer Kontrollgruppe ermittelt. Daraus werden 4 Hypothesen H_A, H_B, H_C und H_D aufgestellt: $\mu_V^{(A)}=\mu_K^{(A)},...$ also für jeden Parameter der Vergleich der Versuchsgruppe V mit der Kontrollgruppe K. Angenommen, die Variablen B und D sind von besonderer Bedeutung, d.h. wenn bei diesen kein Unterschied festgestellt werden kann, dann wird das Präparat komplett abgelehnt. Dann bilden H_B und H_D die Familie F_1 und H_A und H_C die Familie F_2. Für die 4 Hypothesen werden nun über einen Mittelwertvergleich, z.B. t-Test, p-Werte ermittelt. p_B und p_D werden nun adjustiert, allerdings nur mit einem $k=2$, da die anderen p-Werte an dieser Stelle keine Rolle spielen. Je nachdem welches Gatekeeping-Verfahren benutzt wird, muss nun ein oder auch beide adjustierten p-Werte kleiner gleich α sein, damit die übrigen Hypothesen, H_A und H_C berücksichtigt werden können. In diesem Fall werden allerdings bei der Adjustierung von p_A und p_C die p-Werte der Familie F_1 mitberücksichtigt.

Für das parallele Gatekeeping hat Dmitrienko ein besonders effizientes Konzept entwickelt: *rejection gain factor*. Im ersten Schritt, für Familie F_1 wird wie oben beschrieben mit α verglichen. In den folgenden Schritten *i* werden die $\tilde{p}_{ij}$ jedoch mit $b_i\alpha$ verglichen mit den *rejection*

gain factors $0 \le b_i \le 1$. Diese b_i, werden wie folgt berechnet, wobei hier alle Hypothesen als gleichgewichtet angenommen werden:

$$b_1 = 1 \qquad b_i = \prod_j^{i-1} \left(\frac{r_j}{n_j}\right)$$

wobei r_i die Anzahl der im Schritt *i* abgelehnten Hypothesen aus F_i sind.

2. 7. 2 Parametrisches multiples Testen

Werden eine Hypothese für mehrere Variablen (multiple endpoints) verglichen, so ist es naheliegend, dass die Variablen korrelieren und damit die Tests nicht unabhängig sind. Diesem kann man z.B. durch ein parametrisches Modell begegnen, indem man für die untersuchten Variablen eine multivariate Normalverteilung annimmt, deren Korrelations- bzw. Kovarianzmatrix mit Hilfe der erhobenen Daten geschätzt werden kann. Diese sollte allerdings positiv-definit sein. Hierzu gibt es inzwischen eine Reihe von Methoden, u.a. die folgende von Xie (2011).

Hier werden auch wieder den Hypothesen Gewichte $w_1,...,w_k$ mit $0<w_i<1$ sowie $w_1+...+w_k=1$ entsprechend ihrer Bedeutung gegeben. Die adjustierten *p*-Werte p' errechnen sich wie folgt:

$$p'_i = 1 - P\left(\bigcap_{j=1}^{m} a_i \le X_j \le b_i\right)$$

wobei die X_j ($1 \le j \le m$) multivariat verteilt sind, *P* die Wahrscheinlichkeit der multivariaten Normalverteulung ist, sowie

$$a_i = \phi^{-1}\left(\frac{p_i w_i}{2w_i}\right) \qquad b_i = \phi^{-1}\left(1-\frac{p_i w_i}{2w_i}\right)$$

Das Symbol $\bigcap$ für den Durchschnitt von Mengen ist so zu verstehen, dass die dahinter aufgeführten Bedingungen für alle *j* erfüllt sein müssen. Dieses Verfahren hält die FWER ein. Je größer die Korrelationen zwischen den Variablen ist, desto stärker ist die Überlegenheit gegenüber den „normalen" Verfahren ohne Berücksichtigung der Korrelationen, etwa Holm und Hochberg. Durch die Berechnung von Wahrscheinlichkeiten für die multivariate Normalverteulung ist diese Methode sehr rechenaufwändig.

2. 8 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, Hommel, Hochberg und Bejamini, Benjamini & Yekutieli),
- `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,
- `Mediana` mehrere Gatekeeping-Verfahren, die Parametric multiple testing procedure sowie mehrere Verfahren von Westfall, Wiens und Dmitrienko,

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,
- `ggstatsplot` für diverse paarweise Tests (u.a. t-Test, Games-Howell und Dwass-Steel),
- `asht` für parametrische und nichtparametrische paarweise Vergleiche mittels der Verfahren von Ryan & Welsch,
- `agricolae` für paarweise nichtparametrische Vergleiche.
- `PMCMRplus` für paarweise parametrische und nichtparametrische Vergleiche.

Mehr dazu in den Kapiteln 15 (R) bzw. 16 (SPSS).

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 in Dokumenten der UCLA sowohl für R als auch 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

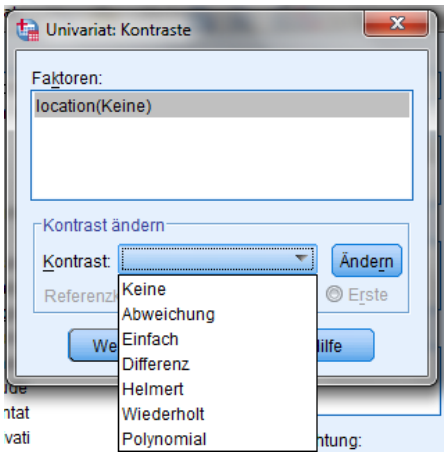
`options(contrasts=c("contr.name1","contr.name2"))`

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 Puri & Sen-Verfahren

Einige der bei Lüpsen (2019) vorgestellten nichtparametrischen Varianzanalysen lassen sich ja auf die parametrischen Standardverfahren zurückführen, so insbesondere die RT-, die ART- sowie die Puri & Sen- oder 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 Anova-Tabelle V 4.6 (Kapitel 4.3.4 in Lüpsen, 2019) 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 Anova-Tabelle V 4.8 (Kapitel 4.3.4 in Lüpsen, 2019) 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 Lüpsen, 2019) völlig identisch.

Ein wenig anders ist es bei Verwendung des Puri & Sen-Verfahrens (vgl. Kapitel 4.3.5 in Lüpsen, 2019). Hier müssen die χ^2-Werte für jeden Vergleich aus den parametrischen t-Werten „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.

Die χ^2-Werte haben jeweils 1 Fg und müssen anhand der Tabellen der χ^2-Verteilung auf Signifikanz überprüft werden. Aus der Anova-Tabelle für den Datensatz 2 `mydata2` (Kapitel 4.3.5 in Lüpsen, 2019) 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 V 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 χ^2-Werte müssen die t-Werte aus der o.a. Tabelle 3-1 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 χ^2-Werte `chisq` ermitteln:

```
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 (vgl. o.a. Tabelle 3-2):

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

Der kritische χ^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 die Funktion `lm` dient zur Analyse der Kontraste. Die Spezifikation der Koeffizienten differiert erheblich, falls diese vorgegeben werden. Die Werte müssen spaltenweise eingegeben, und z.B. mittels `cbind` zu einer Matrix mit 3 Spalten zusammengefasst werden. Doch Vorsicht: eigene Kontraste können in R nicht einfach über die Koeffizienten c_{ij} spezifiziert werden. Diese müssen zusätzlich als erste Spalte die Werte (1/*k*,...,1/*k*) enthalten. Anschließend wird die Inverse der transponierten Matrix gebildet. Schließlich werden daraus die Spalten 2,..,k als Kontrastmatrix genommen. (Dies ist auch ausführlich in dem Skript des Institute for Digital Research and Education der UCLA beschrieben.)

```
cmatrix <- cbind(rep(1/4,4), "A1-A2"=c(1,-1,0,0), "A3-A4"=c(0,0,1,-1),
   "A12-A34"=c(1,1,-1,-1))
icmatrix <- solve(t(cmatrix))
cont     <- icmatrix[ ,2:4]
contrasts(mydata2$drugs) <- cont
aovc <- lm(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 o.a. 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 χ^2-Wert beträgt 11,2 . Die χ^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 χ^2-Wert beträgt 11,2 . Die χ^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

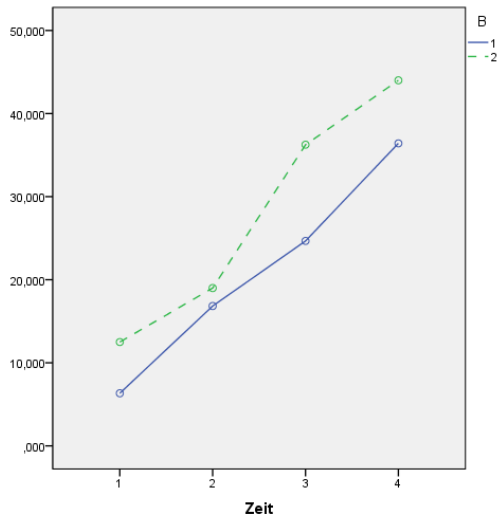
Bei der logistischen Regression gibt es für nominale Prädiktoren Standard-Kontraste. Wenn in R ein Prädiktor als „factor“ deklariert ist, wird für diesen automatisch die Kodierung gewählt, die in der `options(contrasts...)`-Anweisung festgelegt wurde (vgl. Kapitel 3.3). In SPSS kann bei der binär-logistischen 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 Puri & Sen-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`, s. Anhang) demonstriert werden. Dieser umfasst die Gruppierungsfaktoren A und B sowie den Messwiederholungsfaktor Zeit. Tabelle 6-7 (s. Anhang) 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 „wiederholt“ 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 in Lüpsen (2019) 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 6-4 (Lüpsen, 2019) zusammengefasst. 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 Datensatz `winer568t` (s.Anhang).

- 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 für den Faktor Zeit werden über eine Matrix cont4 spezifiziert, die als erste Spalte die Werte 1/4 enthalten muss. Anschließend wird die Inverse der transponierten Matrix als Matrix icont4 errechnet. Die Spalten 2 bis 4 beinhalten die für R erforderlichen Kontrastkoeffizienten.
- 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:
`corr = corCompSymm(, form= ~ 1 | Vpn)`

Die Kommandos lauten dann:

```
cont4 <- cbind( rep(1/4,4), "Z 1-2"=c(1,-1,0,0),
         "Z 2-3"=c( 0,1,-1,0), "Z 3-4"=c( 0,0,1,-1))
icont4<- solve(t(cont4))
contrasts(winer568t$Zeit) <- icont4[,2:4]
contrasts(winer568t$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 erzeugten Kontrastmatrizn `cont4` und `icont4`:

```
> cont4                          icont4

          Z 1-2 Z 2-3 Z 3-4           Z 1-2 Z 2-3 Z 3-4
[1,] 0.25     1     0     0    [1,] 1  0.75   0.5  0.25
[2,] 0.25    -1     1     0    [2,] 1 -0.25   0.5  0.25
[3,] 0.25     0    -1     1    [3,] 1 -0.25  -0.5  0.25
[4,] 0.25     0     0    -1    [4,] 1 -0.25  -0.5 -0.75
```

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

```
Coefficients:
                       Value Std.Error    t-value p-value
(Intercept)         24.500000  1.201262  20.395216  0.0000
A1                   2.187500  1.201262   1.821001  0.0780
B1                  -3.437500  1.201262  -2.861574  0.0074
ZeitZeit 1-2        -8.500000  1.251504  -6.791827  0.0000
ZeitZeit 2-3       -12.541667  1.251504 -10.021274  0.0000
ZeitZeit 3-4        -9.750000  1.251504  -7.790625  0.0000
A1:B1                0.500000  1.201262   0.416229  0.6800
A1:ZeitZeit 1-2     -8.166667  1.251504  -6.525481  0.0000
A1:ZeitZeit 2-3      5.625000  1.251504   4.494591  0.0001
A1:ZeitZeit 3-4      4.833333  1.251504   3.862019  0.0005
B1:ZeitZeit 1-2     -2.000000  1.251504  -1.598077  0.1199
B1:ZeitZeit 2-3      4.708333  1.251504   3.762139  0.0007
B1:ZeitZeit 3-4     -2.000000  1.251504  -1.598077  0.1199
A1:B1:ZeitZeit 1-2   0.000000  1.251504   0.000000  1.0000
A1:B1:ZeitZeit 2-3   1.375000  1.251504   1.098678  0.2801
A1:B1:ZeitZeit 3-4   0.250000  1.251504   0.199760  0.8429
```

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 α ein kritischer Wert c_d errechnet, derart dass alle Mittelwertdifferenzen, die diesen Wert c_d überschreiten, als signifikant auf dem α-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 α 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, vorausgesetzt die n_i variieren nicht allzu stark. 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 α-Adjustierung vorgesehen. Simulationen haben gezeigt, dass in diesem Fall das α-Risiko auch annähernd eingehalten wird, solange die Varianzen homogen sind (vgl. Wilcox, 2013).

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 α vorher wie im vorigen Kapitel beschrieben hinsichtlich der Anzahl der Vergleiche m adjustiert wird. Dieses Verfahren ist bedingt durch die α-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 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. 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 $max(x_i) - min(x_i)$ 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 α-Risiko exakt α 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 α 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 Dallal (2001) 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. Daraus resultiert 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 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 α-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`

4. 6 Hayter

Das vergleichsweise weniger bekannte Verfahren von Hayter, *one-sided studentized range test*, basiert zum einen auf dem Ansatz von Tukey-Kramer (siehe 4.2), zum anderen auf der Anordnung der Mittelwerte ($\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \ldots \leq \bar{x}_{(k)}$), hält aber die (eingangs erwähnte) FWER korrekt ein, wodurch andere kritische Werte h_a zu berechnen sind:

$$c_d = h_\alpha(k, df_{Fehler}) \sqrt{MS_{Fehler} / n} \qquad (4-5)$$

Liu et al. (2002) haben dieses Verfahren verbessert, insofern es eine höhere Power hat: Für zwei Gruppen *i* und *j* wird nicht die Mittelwertdifferenz mit einem kritischen Wert verglichen, sondern das Maximum aller Mittelwertdifferenzen von Gruppen *l* und *m* mit $i \leq l \leq m \leq j$. Zwei Mittelwerte sind dann als unterschiedlich anzusehen, wenn

$$max_{i \leq l < m \leq j}(\overline{x_m} - \overline{x_l}) \geq h_\alpha(k, df_{Fehler}) \sqrt{MS_{Fehler} / n} \qquad (4-6)$$

Nashimoto & Wright haben dieses Verfahren auch in einer nichtparametrischen Variante entwickelt (siehe Kapitel 12). Alle Varianten sind in Nashimoto & Wright (2005) beschrieben.

R: `MTest (PMCMRplus)`
SPSS: -

5. Schrittweise Verfahren von Ryan und Welsch

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 α-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 α. 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: `agricolae`
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 α-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 α-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, 1983).

- 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.

R: `ssnonpartest (npmv), AdjustPvalues (Mediana)`
SPSS: -

5. 4 Welsch

Das Verfahren von Welsch (1977) stellt gewissermaßen das Gegenstück zum o.a. Verfahren von Ryan dar. Allerdings werden hier mehr Relationen durch Implikationen erschlossen. Auch hier werden zunächst die Mittelwerte der Größe nach sortiert: $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \ldots \le \bar{x}_{(k)}$. Im ersten Schritt werden jeweils zwei aufeinander folgende Mittelwerte $\bar{x}_{(j)} \le \bar{x}_{(j+1)}$ verglichen.

- Sind diese als verschieden anzusehen, so sind auch alle größeren Sequenzen aufeinander folgender Mittelwerte als verschieden anzusehen, die diese beiden enthalten.
- Sind diese als gleich anzusehen, werden die beiden Sequenzen $\bar{x}_{(j-1)} \le \bar{x}_{(j)} \le \bar{x}_{(j+1)}$ und $\bar{x}_{(j)} \le \bar{x}_{(j+1)} \le \bar{x}_{(j+2)}$ jeweils auf Gleichheit geprüft. Auch hier gilt dann: Wenn eine Se-

quenz signifikant verschieden ist, werden auch alle größeren Sequenzen als unterschiedlich eingestuft, die diese drei Mittelwerte enthalten. Wenn auch die Mittelwerte dieser 3er-Sequenzen als gleich eingestuft werden, so werden im nächsten Schritt alle die 4er Sequenzen überprüft, die die o.a. 3er-Sequenzen enthalten, usw.

Auf diese Weise wird nun erschlossen, ob zwei Mittelwerte gleich oder verschieden sind. Dieses Verfahren hat nur einen praktischen Haken: Während beim o.a. Verfahren von Ryan praktisch beliebige Varianzanalyse-Tests zum Vergleich mehrerer Mittelwerte eingesetzt werden dürfen, können hier keine kritischen Werte der Standardverteilungen zur Beurteilung der Vergleiche verwendet werden, wenn das α-Risiko, genauer gesagt der FWER (*familywise error rate*), unter Kontrolle gehalten werden soll. Welsch hat für den Fall normalverteilter Variablen Tabellen für die kritischen Werte erstellt.

R: `tukeyWelsch (asht)`

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, was allerdings eine separate Berechnung von MS_{Fehler} und df_{Fehler} bei jedem Vergleich erfordert:

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

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

Dieser Test gilt als leicht liberal. Simulationen haben gezeigt, dass bei kleinen bis mittleren n_i (etwa $n_i \leq 40$) 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: `gamesHowellTest (PMCMRplus), games_howell (ggstatsplot)`
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: 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: `dunnettT3Test (PMCMRplus)`
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: `tamhaneT2Test (PMCMRplus)`
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 sowie etwas ausführlicher Freedman, 2006). 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 Lüpsen, 2019). 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 Bonferroni. 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: `scheffeTest (PMCMRplus), scheffe.test (agricolae), DescTools`
SPSS: `oneway, glm`

7. 3 Tukey

Tukeys HSD-Test lässt sich auf Kontraste verallgemeinern. Der kritische Wert, mit dem ein Kontrast C (3-1) 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 Irrtumswahrscheinlichkeit 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 Gruppen-

zahl. 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}.

R: `uryWigginsHochbergTest (PMCMRplus)`
SPSS: `oneway, glm`

7. 5 Brown & Forsythe und Kaiser & Bowden

Brown & Forsythe (1974, vgl. dazu auch Kapitel 4.3.3 in Lüpsen, 2019) 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, 2013) 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 et al., 2002).

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*, hält aber dennoch das α-Risiko korrekt ein.

`R: pairwise.t.test`

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 α ist hier *K* vorzugeben. Übliche Werte für *K* sind 50, 100 und 500, die in etwa einem α 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* (1985) zu finden.

R: `waller.test (agricolae)`
SPSS: `oneway, glm`

8. 2 Scott & Knott

Wieder einen anderen Weg haben Scott & Knott (1974) 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 α-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, 1985).

R: `scottknott (ExpDes), LScottKnott (laercio)`
SPSS: -

8. 3 Gabriel

Gabriel (1978) 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: `rgabriel`
SPSS: `oneway, glm`

8. 4 Daytons Modell-Tests

Der Ansatz von Dayton (1998) ähnelt etwas dem o.a. Verfahren von Scott & Knott: Es werden die Partitionen der k Gruppen mittels eines Likelihood-Ratio-Tests (LR) überprüft. (Dieser ist letztlich dem F-Test der Varianzanalyse sehr ähnlich.) Angenommen es liegen 3 Gruppen vor, dann lassen sich aus den entsprechenden 3 Mittelwerten $\mu_1 \leq \mu_2 \leq \mu_3$ 4 Partitionen bilden: $\{(\mu_1\ \mu_2\ \mu_3)\}$, $\{\mu_1\ (\mu_2\ \mu_3)\}$, $\{(\mu_1\ \mu_2), \mu_3\}$, $\{(\mu_1), (\mu_2), (\mu_3)\}$, wobei in (...) jeweils gleiche Mittelwerte eingeschlossen sind. Allgemein gibt es bei k Gruppen $2^{(k-1)}$ Partitionen. Für jede Partition wird der LR bzw. der dazu äquivalente AIC (Akaike Information Criterion) berechnet. Der kleinste AIC gibt die beste Partition an. Es ist allerdings erforderlich, vorher einen globalen Test, z.B. den F-Test, durchzuführen, weil das Ergebnis des Modell-Tests ja nicht sicherstellt, dass überhaupt Mittelwertunterschiede vorhanden sind. Vorteile dieses Verfahrens sind zum einen die Vermeidung von Intransitäten, zum anderen die fehlende Voraussetzung von homogenen Varianzen und ungleichen Zellenbesetzungszahlen sowie die Anwendung auch auf nichtmetrische Variablen.

R: `multComp (AICcmodavg)`

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 α-Korrektur deutlich reduziert. Ein weiterer Aspekt: Im Gegensatz zu den Paarvergleichen liegen hier meist einseitige Hypothesen vor, was bei der Wahl des α 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 α. 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* (1988) 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.

Eine gute Beschreibung der Verfahren von Dunnett (vgl. auch nachfolgende Abschnitte) ist bei Dmitrienko et al. (2012) zu finden.

R: `mtcp (nparcomp), AdjustPvalues (Mediana)`
SPSS: `oneway, glm`

9. 2 Schrittweise Dunnett-Verfahren

Marcus et al. (1976) 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 wie 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.

Die Verfahren sind gut verständlich beschrieben in Dmitrienko et al.(2010).

R: `DunnettTests, paradjp (multxpert)`
SPSS: -

9. 3 Steel

Steel hat einen nichtparametrischen Test zum Vergleich mit einer Kontrollgruppe konstruiert (vgl. Winer, 1991). 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.

R: `Steel.test (kSamples)`
SPSS: -

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: `gao (nparcomp), (gao) mutoss`
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 α-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 $MS_{Residuen}$ (allgemein: die Varianz im Nenner des F-Wertes des zu untersuchenden Effekts) zu ersetzen. Dies wird u.a. von Winer (1991) empfohlen. Maxwell (1989) 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 Lüpsen, 2019), 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, die verallgemeinerten Kontraste von Scheffé und Tukey in Kapitel 7 sowie die meisten der Kontrollgruppenvergleiche in Kapitel 9. Allerdings bietet R Funktionen für diese Tests, bei denen die Fehlervarianz und die dazugehörenden Freiheitsgrade vorgegeben werden können. Auch in SPSS lässt sich das mit relativ wenig Aufwand 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 Lüpsen, 2019). Doch auch hier: SPSS bietet das Verfahren von Ryan für Messwiederholungen nicht an, während vom Autor (Lüpsen, 2019) eine entsprechende Funktion für R angeboten wird

So bleibt als einzige leicht umsetzbare Lösung der paarweise Vergleich mittels gepaartem t-Test

unter Verwendung einer α-Adjustierung. Dies hat immerhin den Vorteil, dass dabei keinerlei Varianzhomogenität, hier also Spherizität, 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 (Lüpsen, 2019) war erläutert worden, dass durch die Hinzunahme weiterer Faktoren in das zu analysierende Design der F-Test eines Faktors 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.

R bietet Funktionen für diese Verfahren, bei denen die Fehlervarianz und die dazugehörenden Freiheitsgrade vorgegeben werden können. Auch in SPSS lässt sich das mit relativ wenig Aufwand per Hand erledigen (vgl. Kapitel 15.2.1 und 16.2).

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_1ist 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 (2002) zu finden.

Im Folgenden sei $N = \sum n_i$. D.h. bei abhängigen Stichproben ist $N = n$.

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 Lüpsen, 2019). Dies empfehlen auch Conover & Iman (1981) sowie Day & Quinn (1989) für den Fall multipler Mittelwertvergleiche. Da Ramseyer und Tcheng (1973) erkannt haben, dass bei gleichverteilten Daten die klassischen Verfahren in Kapitel 4 das α-Risko nicht mehr unter Kontrolle halten, ist es vorteilhafter, anstatt der reinen Rangtransformation die *inverse normal transformation* (INT), auf die abhängige Variable anzuwenden. Hierbei werden die Ränge zusätzlich in *normal scores* umgerechnet (vgl. Lüpsen, 2019), so dass man annähernd eine Normalverteilung erhält. 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, 2019, sowie W.G. Cochran, 1950, und G.H. Lunney, 1970.) Speziell für die klassischen Mittelwertvergleiche haben Chuang-Stein und Tong (1995) 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, 2019). Dies wird vielfach als ein sehr effizientes Verfahren für nichtparametrische Varianzanalysen angesehen, wenn es auch deutliche Bedenken hinsichtlich der Anwendbarkeit gibt (vgl. Lüpsen, 2017). Da üblicherweise den multiplen Mittelwertvergleichen 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 (2001) 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 Verfahren ist anschaulich beschrieben bei Marascuilo & McSweeney (1977, Kapitel 12.2. und 14.5, ohne allerdings den Test zu nennen).

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 α 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 χ^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 α 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: `posthoc.friedman.nemenyi.test, posthoc.kruskal.nemenyi.test (PMCMR), kwAllPairsNemenyiTest, frdAllPairsNemenyiTest (PMCMRplus), Nemenyi, NSM3`

12. 4 McSweeney

McSweeney (vgl. Marascuilo & McSweeney, 1977) hat analog zum o.a. Nemenyi-Test ein Verfahren zum paarweisen Vergleich bei Verwendung der normal scores von van der Waerden entwickelt. Auch hier werden die Mittelwertdifferenzen (der normal scores) mit einer kritischen Größe c_d verglichen:

$$c_d = q_\alpha(k) \sqrt{\frac{S}{N-1} \cdot \frac{1}{n}}$$

wobei $q_a(k)$ der kritische Wert der normal range-Verteilung (vgl. Kapitel 1.4) zur Spannweite k und einem voregebenen α ist (alternativ der kritische Wert der studentized range-Verteilung mit ∞ FG) und $S = \sum z_i^2$ die Quadratsumme aller normal scores z_i.

Hierzu gibt es auch eine Variante für abhängige Stichproben. Hier errechnet sich c_d als:

$$c_d = q_\alpha(k) \sqrt{\frac{k}{k-1} \cdot \frac{S}{k \cdot n} \cdot \frac{1}{n}}$$

wobei $q_a(k)$ wie oben und S wieder die Quadratsumme aller kn normal scores z_i ist.

12. 5 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 α 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: `pSDCFlig(NSM3), dscfAllPairsTest (PMCMRplus), pairwise_p (ggstatsplot), NSM3`
SPSS: -

12. 6 Conover & Iman

Auch das Verfahren von *Conover & Iman* (1979) ist als Transskription eines „Klassikers“ anzusehen: als Adaption des LSD-Tests von Fisher (vgl. Kapitel 4.1) auf die gemäß *joint ranking* rangtransformierte Daten: Die Differenz der mittleren Ränge zweier Gruppen *i* und *j* ist als unterschiedlich anzusehen, wenn

$$\left|\overline{R}_i - \overline{R}_j\right| > t_{1-\alpha/2}(n-k)\sqrt{s^2\frac{n-1-H}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

wobei H die Testgröße des Kruskal-Wallis-Test ist und

$$s^2 = \frac{1}{n-1}\left(\sum R_i^2 - n\frac{(n+1)^2}{4}\right)$$

Diesen Test gibt es für abhängige Stichproben. Dann ist die Differenz der mittleren Ränge zweier Gruppen *i* und *j* ist als unterschiedlich anzusehen, wenn

$$\left|\overline{R}_i - \overline{R}_j\right| > t_{1-\alpha/2}((n-1)(k-1))\sqrt{\frac{2k\left(1-\frac{F}{n(k-1)}\right)\left(\sum_i^n\sum_j^k R_{ij}^2 - \frac{nk(k+1)^2}{4}\right)}{(k-1)(n-1)}}$$

wobei F die Testgröße des Friedman-Tests ist. Conover & Iman haben keine α-Adjustierung vorgesehen. Vgl. dazu auch die Ausführungen in Kapitel 14.1.

R: `kwAllPairsConoverTest, frdAllPairsConoverTest(PMCMRplus)`

12. 7 Gao

Gao, Alvo, Chen und Li (2008) haben ein Verfahren für multiple Vergleiche entwickelt, das auf dem Modell für beliebige Verteilungen ordinaler oder metrischer Variablen von Akritas, Arnold und Brunner (1997) basiert und das auch Basis für die nichtparametrische ATS-Varianzanalyse ist (vgl. Kapitel 2.5 in Lüpsen, 2019). Das Modell ist gut beschrieben in Brunner & Munzel (2002). 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 und dichotom skalierte Kriteriumsvariablen angewandt werden.

Das Verfahren von Gao gibt es auch zum Vergleich mit einer Kontrollgruppe. Darüber hinaus haben Munzel und Tamhane (2002) ein äquivalentes Verfahren für abhängige Stichproben entwickelt, das aber bislang weder in R noch in SPSS verfügbar ist.

R: `nparcomp, mutoss`
SPSS: -

12. 8 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: $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \ldots \le \bar{x}_{(k)}$. (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 α-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 Lüpsen, 2019). Die Tests werden verschiedentlich auch mit anderen Verfahren durchgeführt.

R: `nparcomp, asht`
SPSS: `nptests`

12. 9 Quade

Der Test von *Quade* (vgl. Garcia et al., 2010) 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 , quadeAllPairsTest (PMCMRplus)`
SPSS: -

12. 10 van der Waerden

Der wenig bekannte Test von *van der Waerden* (vgl. Lüpsen, 2017) 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 χ^2-Test durchgeührt, der die Gleichheit der Mittelwerte testet. Im Falle einer Signifikanz werden dann

wiederum paarweise van der Waerden-Tests durchgeführt. Der Vergleich ist identisch mit dem von Conover & Iman (s.o.), wobei lediglich die Rangmittelwerte durch die Mittelwerte der normal scores und die Testgröße des H durch die des van der Waerden-Tests zu ersetzen sind. Üblicherweise wird keine α-Adjustierung vorgenommen. Daher ist dieser Test als multipler Mittelwertvergleich liberal. .

R: `agricolae, vanWaerdenAllPairsTest (PMCMRplus)`
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.

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 ohnehin errechnet werden. Die Differenz der mittleren Ränge zweier Gruppen *i* und *j* ist als unterschiedlich anzushene, wenn

$$\left|\bar{R}_i - \bar{R}_i\right| > z_{1-\alpha/2}\sqrt{\left(\frac{n(n-1)}{12} - B\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

wobei B eine Bindungskorrektur beinhaltet. Da hier jeweils zwei Gruppen verglichen werden, ist eine anschließende α-Adjustierung ratsam. Dieser Test kann natürlich auch zum Vergleich mit einer Kontrollgruppe verwendet wwerden.

R: `dunn.test, kwAllPairsDunnTest (PMCMRplus)`
SPSS: -

12. 13 Hayter-Nasimoto-Wright

Das Verfahren von Hayter, *one-sided studentized range test*, wurde in Kapitel 4.6 kurz beschrieben. Es basiert zum einen auf dem Ansatz von Tukey-Kramer (siehe 4.2), zum anderen auf der Anordnung der Mittelwerte ($\bar{x}_{(1)} \le \bar{x}_{(2)} \le \ldots \le \bar{x}_{(k)}$), hält aber die (eingangs erwähnte) FWER korrekt ein. Liu et al. (2002) haben dieses Verfahren verbessert und Nashimoto & Wright (2007) für nichtparametrische Vergleiche angepasst. Hierbei werden zunächst, wie beim Kruskal-Wallis-Test alle Werte in Ränge umgerechnet. Für zwei Gruppen *i* und *j* wird nicht die Rangmitteldifferenz mit einem kritischen Wert verglichen, sondern das Maximum aller Rangmitteldifferenzen von Gruppen *l* und *m* mit $i \le l \le m \le j$. Zwei Mittelwerte sind dann als unterschiedlich anzusehen, wenn

$$max_{i \le l < m \le j}(\overline{R_m} - \overline{R_l}) \ge h_\alpha(k, \infty)\sqrt{MS/n}$$

wobei

$$MS = \sqrt{N(N+1)/12} \quad \text{mit } N = kn$$

Die speziellen kritischen Werte h_a sind dieselben wie in 4.6.

R: `osrtTest (PCMCMRplus)`
SPSS: -

12. 14 Hayter-Stone

Ein weiteres Verfahren von Hayter basiert auf Mann-Whitneys U-Test. Für zwei Gruppen *i* und *j* wird die Testgröße U_{ij} als Anzahl der x-Werte bestimmt, für die $x_{(i)} < x_{(j)}$ ist, wobei $x_{(i)}$ bzw. $x_{(j)}$ ein Wert der Gruppe *i bzw. j* ist. Aus diesen wird errechndet

$$S_{ij} = \frac{2\sqrt{6}(U_{ij} - n_i n_j/2)}{\sqrt{n_i n_j (n_i + n_j + 1)}}$$

Die globale Hypothese $\mu_1=\mu_2=...=\mu_k$ wird abgelehnt, wenn

$$max_{1 \le i < j \le k}(S_{ij}) \ge h_\alpha(k, \infty)$$

und die Hypothese $\mu_i=\mu_j$ für zwei Gruppen *i* und *j* wird abgelehnt, wenn

$$S_{ij} \ge h_\alpha(k, \infty)$$

Das Verfahren ist beschrieben in Nashimoto & Wright (2007) .

R: `pHaySton (NSM3)`
SPSS: -

12. 15 Baumgartner-Weiß-Schindler

Der Test von Baumgartner-Weiß-Schindler (z.B. Neuhäuser, 2007) ist eigentlich ein 2-Stichproben-Vergleich. Er gilt allerdings als besser als der klassische Mann-Whitney U-Test oder der Steel-Dwass-Test. Für die Durchführung multipler Vergleiche ist dann eine α-Adjustierung erforderlich.

Seien $x_i^{(1)}$ bzw. $x_i^{(2)}$ die Werte der beiden zu vergleichenden Gruppen mit n bzw. m Beobachtungen und $R_i^{(1)}$ bzw. $R_i^{(2)}$ die entsprechenden Ränge bei gemeinsamer Rangbildung. Dann werden daraus die beiden Teststatistiken errechnet:

$$B_1 = \frac{1}{n}\sum_i^n \frac{\left(R_i^{(1)} - \frac{m+n}{n}i\right)^2}{\frac{i}{n+1}\left(1-\frac{i}{n+1}\right)\frac{m(m+n)}{n}}$$

$$B_2 = \frac{1}{m}\sum_j^m \frac{\left(R_j^{(2)} - \frac{m+n}{m}j\right)^2}{\frac{j}{m+1}\left(1-\frac{j}{m+1}\right)\frac{n(m+n)}{m}}$$

Die Prüfgröße $B = (B_1 + B_2)/2$ hat eine spezielle Verteilung. Der asymptotische Test ist für kleinere Stichproben nicht immer zuverlässig, was durch einen exakten Test aufgefangen werden kann. Von diesem gibt es einige Varianten, u.a. eine von Murakami (2006) und eine unter Verwendung eines Verfahrens von Lepage.

R: `bwsAllPairsTest (PMCMRplus)`
SPSS: -

12. 16 Durbin

Während beim Friedman-Test jede der n Versuchspersonen jede der k Versuchsbedingungen (Messwiederholungen) „bewertet“, ist dies beim Test von *Durbin* (1951) nicht erforderlich. Allerdings muss die Anzahl der Bewertungen bei allen Versuchspersonen gleich sein. Dieser Test stellt somit eine Verallgemeinerung des Friedman-Tests dar. Die Berechnungen können bei Pohlert (2016) nachgelesen werden. Der Durbin-Test testet natürlich zunächst die globale Hypothese gleicher Mittelwerte. Es gibt ihn allerdings auch als multiplen paarweisen Vergleich, der mit einer α-Adjustierung angewandt werden sollte.

R: `durbinAllPairsTest (PMCMRplus)`
SPSS: -

12. 17 Jonckheere-Terpstrata

Der Test von Jonckheere & Terpstrata unterscheidet sich insofern von allen übrigen, als dass er keine paarweisen Vergleiche durchführt, sondern lediglich eine globale Hypothese:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{gegen } H_1: \quad \mu_1 \le \mu_2 \le \ldots \le \mu_k$$

wobei in H_1 mindestens eine “<“-Relation, d.h. mindestens eine Ungleichheit, bestehen muss.

R: `jonckheere.test (PMCMR)`
SPSS: -

12. 18 Nichtparametrische Kontraste

Konietschke, Hothorn und Brunner (2012) 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, 1997).

R: `nparcomp`
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: $\bar{x}_1 \le \bar{x}_2 \le \bar{x}_3 \le \bar{x}_4$, dann möchte man schließen: Wenn $\bar{x}_2 < \bar{x}_3$ dann gilt auch $\bar{x}_1 < \bar{x}_3$, da ja $\bar{x}_1 \le \bar{x}_2$ 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. Bislang gibt es keine Verfahren für ungleiche Stichprobenumfänge, nicht einmal für k=3 oder 4.

Cribbie (2000) zitiert mehrere Studien, in denen der Einfluss ungleicher Varianzen und Stichprobenumfänge untersucht wird. Danach verhalten sich die meisten Verfahren robust gegenüber ungleichen Varianzen, u.a. Ryans und Peritz Methode mit F- oder Q-Test, die Tests von SNK, Hayter, Scheffe und Dunnett. Ledglich, wenn zugleich die n_i ungleich sind, kommt es zu Verletzungen des α-Risikos. Allerdings wird vielfach empfohlen, statt dessen paarweise t-Tests mit Welch-Korrektur für inhomogene Varianzen und anschließender α-Adjustierung, z.B. nach Holland-Copenhaver oder auch Shaffer, durchzuführen, alleine schon wegen der Einfachheit der Durchführung. Auch wird darauf hingewiesen, dass die nichtparametrischen Methoden nicht notwendigerweise robust gegen Heterogenität sind. Korhonen (1982) hat speziell die Verfahren für heterogene Varianzen verglichen. Die Ergebnisse sind allerdings so vielfältig und

insbesondere von den Stichprobenumfängen abhängig, dass keine generelle Empfehlung gegeben werden kann.

13. 2 Messwiederholungen und Split Plot Designs

Bei Messwiederholungs-Designs gibt es bekanntermaßen für die parametrische Varianzanalyse noch die zusätzliche Voraussetzung der Spherizität. Diese beinhaltet nicht nur gleiche Varianzen aller Messwiederholungsvariablen, sondern auch gleiche Korrelationen zwischen diesen. Wie oben bereits erwähnt, gibt es kaum multiple Mittelwertvergleiche speziell für Messwiederholungen, so dass man, wie in Kapitel 10 erwähnt, entweder die für unabhängige Stichproben konzipierten Tests mit einer Modifikation der Fehlervarianz anwenden muss oder Kontraste bzw. paarweise t-Tests. Boik (1981) und Maxwell (1989) haben gezeigt, dass fehlende Spherizität das α-Risiko stark beeinträchtigt, wenn, wie üblich, eine über alle *k* Bedingungen gepoolte Fehlervarianz verwendet wird. Er schlägt daher vor, für jeden Vergleich für die Fehlervarianz nur die betroffenen Schichten zu benutzen. Dies ist dann z.B. erfüllt, wenn paarweise t-Tests verwendet werden.

13. 3 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, 1989).

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, 1989).

13. 3. 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 α-Risikos wurde von verschiedenen Autoren untersucht, u.a. als einer der Ersten von Ramseyer und Tcheng (1973). Solange nur eine der beiden Voraussetzungen 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 (1989) 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. 3. 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 Lüpsen, 2019) 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 nichtnormalverteilte Residuen noch inhomogene Varianzen störend aus.
- Stark heterogene Varianzen können das α-Risiko vergrößern.
- Von besonderer Brisanz ist der Fall des *pairings*, d.h. wenn die s_i^2 mit den n_i korrelieren. Beim negativen pairing, d.h. wenn große Varianzen bei kleineren Zellenbesetzungszahlen auftreten, wird der Fcehler 1. Art in der Regel krass verletzt. Bei positivem pairing reagieren die Tests konservativ.

13. 3. 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 (1988) hat durch Simulationen festgestellt, dass dieser robust gegen leichte Verletzungen der Normaltäts- und der Varianzhomogenitäts-Voraussetzungen ist.

13. 4 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 (1989) 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 α-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 α=0,05 der reale Fehler 1. Art bei ca. 0,025 liegt.

Bernhardson (1975) hat diesen Unterschied untersucht. Dessen Ergebnisse und weitere Ausführungen zu diesem Thema sind u.a. bei Wilcox (2013) 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 α_1, der Fehlerrate ohne eine vorherige signifikante Varianzanalyse und einmal mit α_2, der Fehlerrate mit vorheriger signifikanter Varianzanalyse.

	k=2		k=4		k=8		k=10	
	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_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 (1989) 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 (1985), Conover & Iman (1979) oder Carmer & Swanson (1973), 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. Dennoch hat Dijkstra (1987) gezeigt, dass, wenn z.B. 5 Gruppen denselben Mittelwert m_1 und andere 5 Gruppen den Mittelwert m_2 haben, das α-Risiko verletzt wird.

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 (1993).

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. Allerdings haben Smith & Han (1981) errechnet, dass die Fehlerrate stark sinkt, wenn zuvor eine Varianzanalye durchgeführt wurde. Und der Unterschied zwischen dem Tests von Tukey und dem von Hayter zugunsten von Hayter rührt daher, dass dieser einen signifikanten globalen Test vorausgesetzt hat.

14. 2 Zur Stärke einiger Tests

Die Stärke (Power) eines Tests hängt von verschiedenen Größen ab, z.B. n , α 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 (1975) die Tests von Tukey, Newman-Keuls, Duncan, Ryan und Peritz verglichen sowie bei Liu (1997) 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, 1989). Das Verfahren von Peritz wird nur selten untersucht, es schneidet dann aber noch etwas besser als das von Ryan ab wie auch das von Hayter (vgl. Seaman et al., 1991). Kromrey & La Rocca (1995) kommen zu demselben Resultat, auch unter der Bedingung ungleicher Varianzen.

Ein Vergleich der o.a. kritischen Werte c_d ermöglicht allerdings auch einen Vergleich einiger 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 (1989) 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.

Wurde kein globaler Mittelwertvergleich durchgeführt oder wird Ziel 2 verfolgt, so ist Tukeys HSD-Test keine schlechte Wahl, wenn auch die Verfahren von Ryan, Welsch und Peritz besser abschneiden, allerdings nicht immer in der Software zur Verfügung stehen.

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, bieten u.a. Cribbie (2000), Tamhane (1979) und Korhonen (1982). Zu erwähnen ist noch, dass in der schwierigen Situation des negativen pairings die Tests T2 und T3 das α–Risiko einhalten und dennoch eine gute Teststärke besitzen.

Für den Fall von Messwiederholungen gibt es, wie oben erwähnt kaum spezielle Verfahren. Allerdings haben Keselman & Lix (1995), die zahlreiche Untersuchungen zu diesem Thema durchgeführt haben, gezeigt, dass der Duncan-Test relativ robust gegen fehlende Spherizität ist. Wegen der vielfältigen Schwierigkeiten bei Messwiederholungen (vgl. Kapitel 13.2) wird aber allgemein zu paarweisen t-Tests mit einer α–Adjustierung geraten.

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: allgemein

Zunächst einmal muss festgestellt werden, dass es zwar viele vergleichende Studien zu den parametrischen Verfahren gibt, aber praktisch keine zu nichtparametrischen Mittelwertvergleichen.

Sowohl der Nemenyi- als auch der Dwass, Steel, Critchlow-Fligner-Test gelten als extrem konservativ (vgl. Day & Quinn, 1989). 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. (2008) 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. Liberaler sind dagegen die Verfahren von Conover & Iman sowie das auf normal scores basierende von McSweeney. Zu beachten ist auch die Methode von Fligner & Policello, weil diese auch gegen heterogene Varianzen robust ist.

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 α-Adjustierungen deutlich besser ab. So kann z.B. die Adjustierung von Li mit der Hand gerechnet werden.

Am Beispiel des Datensatzes 10 (mydata10, s. Anhang), der im Kapitel 15 in vielen Beipielen benutzt wurde, lässt sich gut ablesen, wie unterschiedlich die Ergebnisse ausfallen, wenn verschiedene Paarvergleiche durchgeführt werden. Sehr konservativ sind sind die Methoden von Nemenyi (R-5.3), Dwass-Steel (R-3.4) und Blanchard & Rouquain (R-5.2), die nur 3 Unterschiede, von möglichen 15, erkennen. Besser schneiden ab Welsch (R-6.2) und Peritz (R-1.9) mit 6 bzw. 7 signifikanten Vergleichen und schließlich normal scores mit Tukey HSD (R-2.3) und Tukey-Kontrasten mit Hommel-Adjustierung (R-1.3) mit jeweils 8 Mittelwertunterschieden.

14. 6 Nichtparametrisch: Auswahl passend zur Varianzanalyse

In der Regel wird eine nichtparametrische Varianzanalyse vorab durchgeführt, vielfach mittels einer Variablentransformation wie *rank transform* (RT), *inverse normal transformation* (INT) oder *aligned rank transform* (ART). Dann ist es sowohl naheliegend als auch empfehlenswert, mit der transformierten Variablen eines der klassischen Verfahren anzuwenden, ausgenommen das RT-Verfahren, weil die erzeugten Ränge gleichverteilt und damit für multiple Vergleiche ungeeignet sind. Und: wenn die Varianzanalyse mit normal scores durchgeführt wurde (INT oder v.d.Waerden), dann sollte auf keinen Fall ein Verfahren gewählt werden, das auf Rängen basiert, weil dann der Vorteil der Umrechnung der Ränge in normal scores verloren geht.

Nachfolgend eine Tabelle, in der abzulesen ist, welche Mittelwertvergleiche zu welcher Varianzanalyse passen.

Varianzanalyse	liberal / konservativ homogen / heterogen	Mittelwertvergleich-Methode
F-Test	liberal	Duncan, SNK
	konservativ	Tukey HSD, Scheffé, Welch-Ury-Wiggins
	mittel	Ryan-Welsch, Peritz
	liberal-heterogen	Games & Howell, Simes
	konservativ-heterogen	Tamhane T2
Puri & Sen (Kruskal-W)	konservativ	Nemenyi, Dwass-Steel
	konservativ-heterogen	Fligner-Policello
	konservativ- beliebig	paarweise Wilcoxon mit α-Adjust.
Rank Transform	liberal	Conover & Iman, Ryan-Welsch
Inverse Normal Transf.	liberal	McSweeney
		INT-scores mit Tests wie oben bei F-Test
van der Waerden	mittel	van der Waerden
		INT-scores mit Tests wie oben bei F-Test
Akritas-Brunner ATS	konservativ	Gao
Friedman	konservativ	Nemenyi
	liberal	Conover & Iman

14. 7 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, Holland & Copenhaver, 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 die hinreichend attraktiv sind. Und das sind viele. In den nachfolgenden Kapiteln werden die Verfahren „paketweise“ 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:

Stichproben	unabh.	unabh.	abh.	abh.
Varianzen	homogen	heterogen	homogen	heterogen
Tukey HSD	R-1.4			
Newman-Keuls	R-1.1, R-1.2		R-2.1	
Dunnett C / Dunnetts stepwise	R-1.6	R-1.5		
Dunnett Kontrollgruppe	R-1.2			
Ryans REGWQ	R-1.8	R-6.7		
Scott & Knox	R-1.7			
Peritz	R-1.9			
Gabriel	R-1.10			
Welsch	R-6.4			
paarweise Vergleiche - α-Korrektur t-Test / U-Test / Wilcoxon-Test / Fligner-Policello / Simes	R-1.2a R-5.2	R-1.2b R-6.1a R-3.4b R-5.2		R-6.1b R-6.2
Jianjun Li α-Korrektur		R-6.1a	R-4.2b	
Gatekeeping α-Korrektur			R-5.3	
Kontraste	R-1.3	R-3.2	R-2.2	
Kruskal-Wallis - paarweise Vergleiche	R-3.1a			
Friedman - paarweise Vergleiche / Quade			R-4.1	R-4.2a R-3.9
Nemenyi	R-3.3		R-4.2b	R-4.3
Dwass-Steel-Critchlow-Fligner / Gao	R-3.2, R-3.4			
Rangtransformation - LSD / HSD	R-3.3	R-3.1b		

Zunächst kurze Informationen zu den größten und wichtigsten Paketen:

- `agricolae` bietet im Wesentlichen die klassischen parametrischen Tests, darüber hinaus einige wenige nichtparametrische. Der Vorteil der darin enthaltenen Funktionen liegt darin, dass zum einen Mittelwertvergleiche auch bei mehrfaktoriellen Analysen durchgeführt werden können und zum anderen MS_{Fehler} sowie df_{Fehler} vorgegeben werden können, wodurch

auch Vergleiche bei Messwiederholungen möglich sind.

- Dieselben Möglichkeiten bietet das Paket `ExpDes` , allerdings für nur wenige Verfahren.
- `PMCMRplus` bietet fast alle in den vorigen Kapiteln aufgeführten parametrischen und insbesondere nichtparametrischen Verfahren an, allerdings mit der Einschränkung, dass sie nicht bei mehrfaktoriellen Analysen eingesetzt werden können. Und für Messwiederholungen stehen nur die speziell dafür konzipierten nichtparametrsichen Vergleiche zur Verfügung. Da viele der angebotenen Verfahren 2-Stichproben-Vergleiche sind, sind α-Adjustierungen vorgesehen, und zwar zum einen für alle Paarvergleiche und zum anderen für Vergleiche mit einer Kontrollgruppe. Leider muss negativ vermerkt werden, dass einige Funktionen, z.B. für auf der studentized range Verteilung basierenden Methoden, nicht die korrekten Ergebnisse errechnen.
- Auch `NSM3` bietet überwiegend nichtparametrische, insbesondere relativ unbekannte Verfahren an.

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 3/2019) 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`) , das mehrfaktoriell sein darf
- *Faktor*: der Name des zu untersuchenden Faktors (gegebenenfalls 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:` nur 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` innerhalb der R-Dokumentation.

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
Median.test	Median test mit paarweisen Vergleichen
REGW.test	Ryan, Einot and Gabriel and Welsch Test 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. Anhang). Das Ergebnis der Varianzanalyse enthält Tabelle 4-2. 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 PMCMRplus

Das Paket `PMCMRplus` ersetzt das frühere `PMCMR`, das nicht mehr unterstützt wird. Allerdings haben sich die Namen derjenigen Funktionen geändert, die schon in `PMCMR` enthalten waren. U.a. werden die folgenden Verfahren für Analysen ohne Messwiederholungen angeboten:

Funktionsname	*param*	*α–Adj*	*Test*
bwsAllPairsTest		x	Baumgartner-Weiß-Schindler (paarweise)
cochranTest			Cochran‘s Test auf Varianhomogenität
kwAllPairsConoverTest			Conover & Iman (paarweise)
kwManyOneConoverTest			Conover & Iman (mit Kontrollgruppe)
dscfAllPairsTest			Dwass, Steel, Critchlow and Fligner
duncanTest	x		Duncan
kwAllPairsDunnTest		x	Dunn (paarweise)
kwManyOneDunnTest		x	Dunn (mit Kontrollgruppe)
dunnettT3Test	x		Dunnett‘s T3 Test für inhomogene Varianzen
dunnettTest	x		Dunnett‘s Test (mit Kontrollgruppe)
lsdTest	x		Fishers least significant difference
gamesHowellTest	x		Games & Howell Test für heterog. Varianzen
osrtTest	x		Hayter
MTest			Hayter-Nashimoto-Wright
hartleyTest			Hartley‘s F-max-Test auf Varianhomogenität
jonckheereTest			Jonckheere-Terpstrata Anova
kruskalTest			Kruskal Wallis Anova
manyOneUTest		x	Mann-Whitney‘s U-Test (mit Kontrollgrp)
kwAllPairsNemenyiTest			Nemenyi (paarweise)
kwManyOneNdwTest			Nemenyi (mit Kontrollgruppe)
normalScoresTest			normal scores Anova nach Lu & Smith
normalScoresAllPairsTest		x	normal scores (paarweise)
normalScoresManyOneTest		x	normal scores (mit Kontrollgruppe)
scheffeTest	x		Scheffé
siegelTukeyTest			Siegel-Tukey Test auf Varianhomogenität
snkTest	x		Student-Newman-Keuls
steelTest			Steel (mit Kontrollgruppe)
tamhaneDunnettTest			Tamhane-Dunnett (mit Kontrollgruppe)
tamhaneT2Test	x		Tamhane T2 Test für heterogene Varianzen
tukeyTest	x		Tukey HSD
uryWigginsHochbergTest	x		Welch, Ury & Wiggins
vanWaerdenTest			van der Waerden Anova
vanWaerdenAllPairsTest		x	van der Waerden (paarweise)
vanWaerdenManyOneTest		x	van der Waerden (mit Kontrollgruppe)
welchManyOneTTest		x	t-Test (mit Kontrollgruppe)

Der Aufruf der einzelnen Funktionen ist weitgehend identisch, und zwar für 1-faktorielle Analysen ohne Messwiederholungen:

```
Funktion (abh.Variable, Faktor, data=Datensatz, p.adjust.method="Name",
          dist="Methode")
```

wobei

- *Funktion* eine der u.a. Funktionsnamen ist. Die Variablen könnn hierbei auch in der Form *abh.Variable~Faktor* angegeben werden.
- Bei dem Parameter `p.adjust.method`: wird eine α-Adjustierung (wie bei `p.adjust`, vgl. 15.5.1) nach einer der folgenden Methoden, unter Angabe des Namens angegeben: "`none`" (keine Korrektur), "`holm`", "`hochberg`", "`bonferroni`", "`BH`" (Benjamini-Hochberg), "`BY`" (Benjamini-Yekutieli), "`fdr`" (Benjamini-Hochberg), "`hommel`" .
- Einige Verfahren, z.B. die Kruskal-Wallis- und die Friedman-Anova erlauben wahlweise neben dem χ^2-Test auch einen F-Test (Iman & Davenport) zur Signifikanzprüfung. Die Standardmethode ist "`Chisquare`", alternativ "`Fdist`" oder "`Tukey`" (vgl. Kapitel 12.3).

Hinweise:

- In der Tabelle sind die parametrischen Tests gekennzeichnet, alle anderen sind nichtparametrisch.
- Alle Funktionen für Mittelwertvergleiche bieten über `p.adjust.method die Möglichkeit` einer α-Adjustierung, die auch meistens sinnvoll eingestellt ist. Da bei manchen Tests die Voreinstellung auf "`holm`" vorhanden ist, sollte man sich eventuell eine bessere Korrektur aussuchen, z.B. "`BH`" oder "`BY`". Bei Methoden, die keine Adjustierung erfordern, sollte gegebenenfalls `p.adjust.method="none"` gewählt werden.
- Soll eine andere α-Korrektur vorgenommen werden, so ist zunächst `p.adjust.method="none"` zu wählen. Anschließend können die p-Werte über das Attribut $p.value gespeichert und weiterverarbeitet werden (vgl. Beispiel R-4.2).
- Bei Vergleichen mit einer Kontrollgruppe ist dieses immer die 1. Gruppe. Gegebenenfalls muss eine Umkodierung vorgenommen werden (vgl. u.a. Beispiel).

Beispiel R-1.2a:

Wie im Beispiel R-1.1 soll für die Beispieldaten 2 (`mydata2`, vgl. Anhang) mittels des Newman-Keuls-Test der Effekt des Faktors `drugs` im Detail untersucht werden. Die Varianzen können als homogen angesehen werden. Das Ergebnis der Varianzanalyse enthält Tabelle 4-2 (siehe Anhang).

```
snkTest(x~drugs, mydata2)
```

```
        Pairwise comparisons using SNK test

data: x by drugs

  1      2      3
2 0.1655 -      -
3 0.1239 0.5004 -
4 0.0022 0.0274 0.0580

P value adjustment method: step down
alternative hypothesis: two.sided
```

Auf den ersten Blick lässt der Hinweis am Ende der Ausgabe vermuten, dass zusätzlich eine α-Adjustierung vorgenommen worden ist. Letztlich ist es allerdings nur ein Hinweis auf die Technik des SNK-Verfahrens. Dennoch muss festgestellt werden, dass diese Ergebnisse sich nicht mit denen anderer Funktionen für das SNK-Verfahren decken.

Beispiel R-1.2b:

Für den Datansatz `mydata10` (vgl. Anhang) sei angenommen, die Gruppe 6 sei die Kontrollgruppe, die mit allen anderen verglichen werden soll. Die abhängige Variable ist metrisch und kann als annähernd normalverteilt angenommen werden, so dass Dunnett's Test angewandt werden kann. Allerdings muss die Kontrollgruppe die erste sein, so dass die 6 z.B. in die 0 umkodiert werden muss, was in R nicht ganz so einfach ist, zumal für Faktoren. Dazu muss erst der Faktor "Gruppe" in eine numerische Variable gewandelt, die Umkodierung vorgenommen, zurück in einen Faktor transformiert und schließlich dem Dataframe wieder zugefügt werden:

```
iGruppe<-as.numeric(mydata10$Gruppe)
for (i in (1:langth(iGruppe)) if (iGruppe==6) iGruppe<-0
mydata10$Gruppe<-factor(iGruppe)

dunnettTest(x~Gruppe, mydata10)
```

```
        Pairwise comparisons using Dunnett's-test for multiple
                          comparisons with one control

data: x by Gruppe

  0
1 0.02375
2 0.45864
3 0.00038
4 0.96926
5 0.02922

P value adjustment method: single-step
alternative hypothesis: two.sided
```

15. 1. 3 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. 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.

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`	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 entsprechenden Vergleich mit einer Kontrollgruppe.

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-1.3a:

Zunächst soll ein Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen gerechnet werden. Dazu werden die Beispieldaten 10 (`mydata10`, vgl. Anhang) 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-1.3b:

Hier wird ein Beispiel für den Fall unabhängiger Stichproben mit inhomogenen Varianzen gerechnet. Dazu werden die Beispieldaten 3 (`mydata3`) verwendet (vgl. Anhang). 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 Lüpsen, 2019). 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:

```
library(sandwich)
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. 4 stats: TukeyHSD

Tukeys HSD-Test (für unabhängige 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-1.4:

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Anhang). Das Ergebnis der Varianzanalyse enthält Tabelle 4-2. 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. 5 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-1.5:

Hier wird das o.a. Beispiel 4 für den Fall unabhängiger Stichproben mit inhomogenen Varianzen wieder aufgegriffen (Beispieldaten 3 (`mydata3`), (vgl. Anhang). 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. 6 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-1.6:

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Anhang). 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. 7 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 α

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-1.7:

Es wird wieder das Beispiel für den Fall unabhängiger Stichproben mit homogenen Varianzen unter Verwendung der Beispieldaten 2 (`mydata2`) verwendet (vgl. Anhang). 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. 8 mutoss

Das Paket `mutoss` (derzeit immer noch einem „experimentellen“ 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 df_{Fehler}, 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-1.8:

Für die Beispieldaten 10 (`mydata10`, s. Anhang) 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. 1. 9 npmv

Das Paket `npmv` bietet nichtparametrische multivariate Vergleiche an: zum einen eine Varianzanalyse und zum anderen die Methode von Peritz (vgl. Kapitel 5.3) zum Vergleich der einzelnen Mittelwerte. Die Methode ist geeignet für metrische, ordinale und dichotome Merkmale. Näheres ist der Literatur zu entnehmen, die in der Beschreibung des Pakets aufgeführt ist. Selbstverständlich können mit `npmv` auch univariate Vergleiche durchgeführt werden. Der Aufruf der Varianzanalyse (in der einfachsten Variante) bzw. der Mittelwertvergleiche:

```
nonpartest(abh.Variablen ~ Faktor, Dataframe, permtest=T/F, permreps=n)

ssnonpartest(abh.Variablen ~ Faktor, Dataframe, alpha=.05,
             test=c(1,0,0,0),factors.and.variables=T)
```

- die abhängigen Variablen werden durch “|“ getrennt.
- bei `permtest` wird angegeben, ob eine exakte Berechnung der Wahrscheinlichkeiten über Permutationen erfolgen soll.
- mit `permreps` wird gegebenenfalls die Anzahl der Permutationen (z.B. 1000) festgelegt.
- mit `test` kann einer der 4 multivariaten Tests durch die Angabe von 0/1 in den 4 Positionen für das Verfahren festgelegt werden, wobei die folgenden Tests zur Verfügung stehen:

  (1) Anova F-Test
  (2) Lawley Hotellin
  (3) Bartlett-Nanda-Pilla
  (4) Wilks Lambda
- mit `factors.and.variables=T` wird gesteuert, ob das Verfahren Variablen und Faktoren berücksichtigen soll.

Beispiel R-1.9:

Es werden die Beispieldaten 10 (`mydata10,` vgl. Anhang) verwendet mit einem Vergleich der 6 Gruppen des Faktors Gruppe. Zunächst die Varianzanalyse, danach das Verfahren von Peritz, bei dem der F-Test verwendet werden soll:

```
nonpartest(x~Gruppe,mydata10,permreps=1000)
ssnonpartest(x~Gruppe,mydata10,alpha=.05,test=c(1,0,0,0),
             factors.and.variables=T)
```

```
                                    Test Statistic df1   df2   P-val
ANOVA type test p-value                     16.249 5.000 28.4211  0
McKeon approx. for the Lawley Hotelling Test 16.779 5.000 29.0000  0
Muller approx. Bartlett-Nanda-Pillai Test    16.300 5.147 29.0000  0
Wilks Lambda                                 16.779 5.000 29.0000  0
```

Die globale Hypothese wird mit allen 4 o.a. multivariaten Tests überprüft, die alle einen p-Wert ergeben, der quasi bei 0 liegt.

```
The ANOVA type statistic will be used in the following test
The Global Hypothesis is significant, subset algorithm will continue

~Performing the Subset Algorithm based on Factor levels~
The Hypothesis of equality between factor levels  1 2 3 4 5 6 rejectd
The Hypothesis of equality between factor levels  2 3 4 5 6   rejectd
......                                            .........
The Hypothesis of equality between factor levels  3 4 5 6 is rejected
The Hypothesis of equality between factor levels  2 4 5 6 is rejected
The Hypothesis of equality between factor levels  2 3 5 6 is rejected
......                                            .........
The Hypothesis of equality between factor levels  3 5 6 is rejected
The Hypothesis of equality between factor levels  3 4 6 is rejected
The Hypothesis of equality between factor levels  3 4 5 is rejected
......                                            .........
The Hypothesis of equality between factor levels  1 2 4 is rejected
The Hypothesis of equality between factor levels  1 2 3 is rejected
The Hypothesis of equality between factor levels  3 6 is rejected
The Hypothesis of equality between factor levels  3 5 is rejected
The Hypothesis of equality between factor levels  3 4 is rejected
The Hypothesis of equality between factor levels  2 5 is rejected
The Hypothesis of equality between factor levels  2 3 is rejected
The Hypothesis of equality between factor levels  1 5 is rejected
The Hypothesis of equality between factor levels  1 4 is rejected
All appropriate subsets using factor levels have been checked using
a closed multiple testing procedure, which controls the maximum over-
all type I error rate at alpha= 0.05
```

Es werden schrittweise alle Hypothesen überprüft, ausgehend von der globalen Hypothese der Gleichheit aller 6 Mittelwerte, bis schließlich jeweils 2 Mittelwerte auf Gleichheit getestet werden. Und letztere ergeben das Endergebnis, d.h. es unterscheiden sich (1-4), (1-5), (2-3), (2-5), (3-4), (3-5) sowie (3-6).

15. 1. 10 rgabriel

Dieses Paket enthält nur das Verfahren von Gabriel (vgl. Kapitel 8.3), wozu 2 Funktionen gehören: `rgabriel` zur Berechnung der Konfidenzintervalle sowie `gabriel.plot` zur grafischen Darstellung, die zum Verständnis des Ergebnisses erforderlich ist. Genau genommen berechnet `rgabriel` allerdings nur die halbe Intervallbreite. Da ja letztlich geprüft werden muss, welche Konfidenzintervalle sich überschneiden und welche nicht, müssen diese für alle Gruppen berechnet werden. Dazu werden zusätzlich die Gruppenmittelwerte bestimmt und damit die Intervalle errechnet. Die Benutzung dieser Funktionen soll an einem Beispiel gezeigt werden.

Beispiel R-1.10:

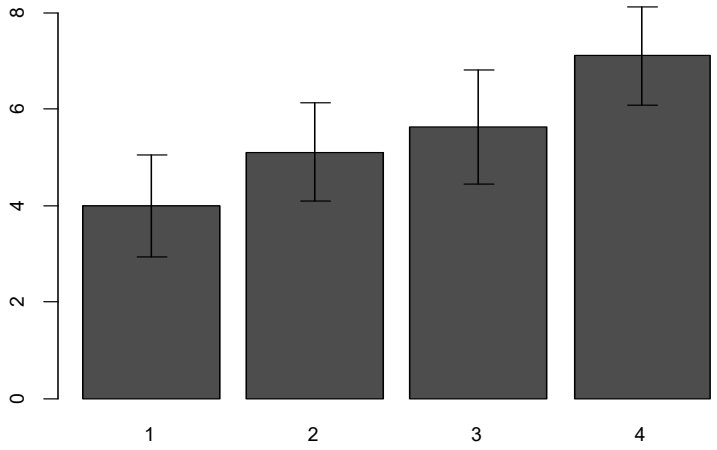
Wiederum soll dazu der Datensatz `mydata2` (s. Anhang) benutzt werden:

```
with(mydata2,rgabriel(x, drugs,a=0.05))->ci
with(mydata2,tapply(x,drugs,mean))->m        # Gruppenmittelwerte
m-ci                                         # unteres CI-Limit
m+ci                                         # oberes CI-Limit
with(mydata2,gabriel.plot(x, drugs,upper=ci))
```

Die mit `rgabriel` errechneten Konfidenzintervalle werden über den Vektor ci zum einen für die Erstellung der Konfidenzintervalle verwendet, zum anderen der Plotfunktion `gabriel.plot` übergeben

```
> m-ci
       1        2        3        4
2.937494 4.092983 4.440801 6.092983
> m+ci
       1        2        3        4
5.062506 6.129239 6.809199 8.129239
```

Hieraus kann abgelesen werden, dass lediglich die CIs der Gruppen 1 und 4 sich nicht überschneiden. Dasselbe kann man auch der Grafik entnehmen.



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`, s. Anhang) verwendet. Für den Faktor `Zeit` wurde dort ein signifikanter Haupteffekt ($p<0{,}001$) nachgewiesen (vgl. Tabelle 6-1). 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 6-3

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 (Lüpsen, 2019) 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`, s. Anhang) 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 α-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`, s.Anhang) verwendet. Für Mittelwertvergleiche im Anschluss an die Kruskal-Wallis-Varianzanalyse wird die α-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 (1981) vorgeschlagen wird. Vgl. auch Kapitel 14.6. 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

Das Paket `nparcomp` ist speziell für nichtparametrische multiple Mittelwertvergleiche konzipiert. Das mathematische Modell entspricht dem von Akritas & Brunner (vgl. Brunner & Muntel, 2002). Es bietet u.a. die folgenden Verfahren (im Wesentlichen für unabhängige Stichproben):

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

Der Aufruf aller Funktionen ist weitgehend identisch:

```
Funktion(Anova-Modell, Dataframe, type="Kontraste", alpha="Wert",
         asy.method="...")
```

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

Folgende Hinweise:

- Es werden nur 1-faktorielle Analysen durchgeführt.
- Über `type` wird der Typ der Kontraste festgelegt. 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. Näheres ist der Paket-Dokumentation zu entnehmen.
- 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*)
- Bei dem Vergleich mit einer Kontrollgruppe von Gao wird standardmäßig die erste Gruppe

als Kontrollgruppe angesehen. Über `control="..."` kann aber auch über die Angabe des Labels eine andere Gruppe angegeben werden.

- Über `asy.method` kann die Methode für die Berechnungen der asymptotischen Wahrscheinlichkeiten festgelegt werden: "`fisher`" (Standard) für eine Fisher-Transformation, `"t.test"` für t-Test unter Berücksichtigung heterogener Varianzen sowie `"normal"` für multivaraite Normalverteilung.

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 Lüpsen, 2019). 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 Statistiken, 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 PMCMRplus

Das Paket `PMCMRplus` bietet eine große Anzahl von nichtparametrischen Verfahren. Es wurde bereits in Kapitel 15.1.2 ausführlich behandelt.

Beispiel R-3.3:

Zunächst soll ein Beispiel für den Fall unabhängiger Stichproben gerechnet werden. Dazu werden die Beispieldaten 10 (`mydata10`, s. Anhang) verwendet. Es soll der Einfluss des Faktors `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 studentized range Verteilung gewählt wird.

```
kruskalTest(x~Gruppe,mydata10)
kwAllPairsNemenyiTest(x~Gruppe,mydata10,dist="Tukey")
```

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 Tukey-Kramer-Nemenyi all-pairs test with
Tukey-Dist approximation
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.

Alternativ kann der Tukey-Test nichtparametrisch angewandt werden, solange die n_i nicht stark variieren. Dazu sollten nicht die rangtransformierten Daten benutzt werden, die letztlich gleichverteilt und damit für die klassischen parametrischen Verfahren ungeeignet sind (vgl. Kapitel 13), sondern die in normal scores transformierten. Hierdurch erhält man deutlich mehr sigifikante Unterschiede:

```
nscore <- qnorm(rank(mydata10$x)/(length(mydata10$x)+1))
aov10r <- aov(nscore~Gruppe,mydata10)
TukeyHSD(aov10r)
```

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

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

$Gruppe
          diff          lwr         upr     p adj
2-1  0.5177810 -0.363984117  1.39954622 0.4873155
3-1 -0.5814003 -1.506203448  0.34340277 0.4128526
4-1  1.0681244  0.186359236  1.94988957 0.0106879
5-1  1.7985973  0.948907146  2.64828750 0.0000065
6-1  0.9268438  0.002040727  1.85164694 0.0492476
3-2 -1.0991814 -2.023984497 -0.17437828 0.0127184
4-2  0.5503434 -0.331421813  1.43210852 0.4207568
5-2  1.2808163  0.431126097  2.13050645 0.0010062
6-2  0.4090628 -0.515740323  1.33386589 0.7561004
4-3  1.6495247  0.724721633  2.57432785 0.0001019
5-3  2.3799977  1.485724562  3.27427076 0.0000001
6-3  1.5082442  0.542318830  2.47416952 0.0006448
5-4  0.7304729 -0.119217257  1.58016309 0.1242592
6-4 -0.1412806 -1.066083676  0.78352254 0.9969802
6-5 -0.8717535 -1.766026585  0.02251961 0.0593437
```

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* (2014) beschrieben sind, u.a. auch die folgenden für unabhängige Stichproben:

Funktionsname	Verfahren
pNDWol	Nemenyi Test zum Vergleich mit einer Kontrollgruppe
pSDCFlig	Dwass, Steel, Critchlow, Fligner paarweise Vergleiche
pFligPoli	Fligner-Policello Vergleich von zwei Stichproben
pHaySton	Hayter-Stone
pJCK	Jonckheere-Terpstra 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 . Die p-Werte sind in folgender Reihe angeordnet: (1-2), (1-3),.., (1-6), (2-3), (2-4),..., (4-6), (5-6). Im Vergleich zu den Ergebnissen des o.a. Beispiels R-3.3 ist zu entnehmen, 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`, s. Anhang) 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 Varianzen 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 0.0674325        NA        NA
3 0.0959041   0.50232        NA
4 0.0647352   0.33780    0.6548
```

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`, s. Anhang) verwendet. Ausgangsbasis ist wieder der in Kapitel 5.3.1 (Lüpsen, 2019) 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 PMCMRplus

Das Paket PMCMRplus war bereits in Kapitel 15.1.2 vorgestellt worden. Hier geht es nun um die Funktionen für Mittelwertvergleiche von Messwiederholungsfaktoren.

Der Aufruf der einzelnen Funktionen ist weitgehend identisch, und zwar für 1-faktorielle Analysen mit Messwiederholungen, basierend auf dem *nicht* umstrukturierten Datensatz:

```
Funktion (Dataframe[,abh.Variablen], p.adjust.method="Name")
```

Einige, in der o.a. Tabelle mit "x" gekennzeichneten Funktionen, erfordern entweder die Wandlung von `Dataframe[...]` in eine Matrix:

```
Funktion (as.matrix(Dataframe[,abh.Variablen]), p.adjust.method="Name")
```

oder einen anderen Aufruf, basierend auf dem umstrukturierten Datensatz:

```
with (Dataframe, Funktion (abh.Variable, groups=Faktor, blocks=Id,
      p.adjust.method="Name")
```

wobei

- *Funktion*: eine der u.a. Funktionsnamen
- *Faktor*: der Name des zu untersuchenden Messwiederholungsfaktor
- *abh.Variablen* entweder die lfd Nummern oder die Namen der einzelnen Messwiederholungsvariablen sind, die `c(..,..,..)` eingeschlossen sein müssen
- *abh.Variable*: der Name der abhängigen Variablen (Kriteriumsvariablen)
- *Id*: die Fall- bzw. Versuchspersonenkennung
- `p.adjust.method`: eine α-Adjustierung (wie bei `p.adjust`, vgl. 15.5.1) nach einer der folgenden Methoden, unter Angabe des Namens: "`none`" (keine Korrektur), "`holm`", "`hochberg`", "`bonferroni`", "`BH`" (Benjamini-Hochberg), "`BY`" (Benjamini-Yekutieli), "`fdr`" (Benjamini-Hochberg), "`hommel`" .

Hinweise:

- In der o.a. Tabelle sind die Tests gekennzeichnet, die eine α-Korrektur erfordern. Diese sind in allen Funktionen vorgesehen und meistens sinnvoll eingestellt. Da bei manchen Tests die Voreinstellung auf "`holm`" vorhanden ist, sollte man sich eventuell eine bessere Korrektur aussuchen, z.B. "`BH`" oder "`BY`". Bei Methoden, die keine Adjustierung erfordern, sollte gegebenenfalls `p.adjust.method="none"` gewählt werden.
- Bei Vergleichen mit einer Kontrollgruppe ist dieses immer die 1. Gruppe. Gegebenenfalls muss eine Umkodierung vorgenommen werden.

U.a. werden die folgenden, ausschließlich nichtparametrischen Verfahren für Analysen mit Messwiederholungen angeboten:

Funktionsname	*α–Adj*	*Aufruf*	*Test*
`frdAllPairsConoverTest`			Conover & Iman (paarweise)
`frdManyOneDemsarTest`	x		Demsar (mit Kontrollgruppe)
`durbinTest`		x	Durbin Anova
`durbinAllPairsTest`	x	x	Durbin (paarweise)
`friedmanTest`		x	Friedman Anova
`frdAllPairsExactTest`	x		exakter Test (paarweise)
`frdManyOneExactTest`			exakter Test (mit Kontrollgruppe)
`frdAllPairsNemenyiTest`			Nemenyi (paarweise)
`frdManyOneNemenyiTest`			Nemenyi (mit Kontrollgruppe)
`quadeAllPairsTest`	x		Quade (paarweise)
`frdAllPairsSiegelTest`	x		Siegel (paarweise)
`skillingsMackTest`		x	Skillings & Mack Anova

Beispiel R-4.2a:

Zunächst wird eine Friedman-Varianzanzanalyse mit dem Beispieldatensatz 4 (`winer518`, vgl. Anhang) durchgeführt, wobei der Aufruf entweder über

```
friedmanTest(as.matrix(winer518[,c(3,4,5)]))
```

oder

```
with(winer518t,friedmanTest(score,groups=Zeit,blocks=Vpn))
```

erfolgen kann:

```
        Friedman rank sum test

data:  score , Zeit and Vpn
Friedman chi-squared = 9.5556, df = 2, p-value = 0.008415
```

Anschließend werden die 3 Stufen mit dem exakten Test paarweise verglichen:

```
frdAllPairsExactTest(winer518[,c(3,4,5)])
```

```
Pairwise comparisons using Eisinga, Heskes, Pelzer & Te Grotenhuis
all-pairs test with exact p-values for a two-way balanced complete
block design

data: winer518[, c(3,4,5)]

   t1    t2
t2 0.320 -
t3 0.011 0.187

P value adjustment method: holm
```

Beispiel R-4.2b:

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.

```
friedmanTest (as.matrix(mydata9))
frdAllPairsNemenyiTest (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 ($p<0.05$).

Soll eine andere Adjustierung vorgenommen werden, z.B. die von Jianjun Li (vgl. Kapitel 2.2.4), das in mutoss (vgl. 15.5.2) enthaltene Verfahren von Blanchard-Roquain, so wird zunächst beim Aufruf die Adjustierung "none" gewählt und anschließend das Ergebnis-Attribut $p.value gespeichert. Da dies eine Matrix enthält, muss diese in einen Vektor gewandelt sowie die NAs (die Werte obergalb der Diagonalen) eliminiert werden. Hier die Anweisungen mit einer Li-Korrektur im Anschluss:

```
frdAllPairsNemenyiTest (as.matrix(mydata9), p.adjust.method="none")->erg
pvalues<-as.vector(erg$p.value)
pvalues <- pvalues[!is.na(pvalues)]
pmax<-max(pvalues)

pvalues/(pvalues+1-pmax)
```

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` , s. Anhang) 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. 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. Beispiel dazu in R-4.2 sowie R-6.1.

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 (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
two.stage	schrittweises Varfahren zur FDR-Kontrolle
twostageBR	Blanchard-Roquain 2-stage adaptive step-up procedure
BlaRoq	Blanchard-Roquain step-up procedure (für abhängige Vergleiche)
indepBR	Blanchard-Roquain 1-stage adaptive step-up procedure (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. 5. 3 Mediana

Das Paket Mediana bietet die Funktion `AdjustPvalues` u.a. mit diversen α-Adjustierungen nach der Gatekeeping-Methode unter Verwendung des Closure-Prinzips von Peritz und Marcus (vgl.Kapitel 2.6) sowie nach der Methode der parametrischen multiplen Tests:

Prozedurname	*Parameter*	*erforderl*	*Adjustierungsverfahren*
`BonferroniAdj`	`weight`		Bonferroni
`HolmAdj`	`weight`		Holm
`HochbergAdj`	`weight`		Hochberg
`HommelAdj`	`weight`		Hommel
`FixedSeqAdj`			Westfall & Krishen
`DunnettAdj`	`n`	x	Single-step Dunnett procedure
`StepDownDunnettAdj`	`n`	x	Step-down Dunnett procedure
`ChainAdj`	`weight, transition`	x	chain procedures
`FallbackAdj`	`weight`	x	Wiens & Dmitrienko
`NormalParamAdj`	`corr, weight`	x	Parametric multiple testing procedure
`Parallel GatekeepingAdj`	`family, proc, gamma`	x	Parallel gatekeeping procedures
`MultipleSequence GatekeepingAdj`	`family, proc, gamma`	x	Multiple-sequence gate-keeping procedures
`Mixture GatekeepingAdj`	`family, proc, gamma, serial, parallel`	x	Mixture-based gatekeeping procedures

Der Aufruf erfolgt über

```
AdjustPvalues(pvalues, "Prozedurname",
  par=parameters(Name=Spezifikation,...))
```

wobei u.a. folgende Namen für die Parameter vorgesehen sind:

- `weight:` Vektor mit Hypothesengewichten
- `n`: Anzahl der Versuchspersonen (Fälle)
- `corr`: Korrelationsmatrix der multivariaten Normalverteilung
- `family`: Vektor der Hypothesennummern in einer Familie
- proc: Vektor der Namen für die α-Adjustierungen (ein Name aus den ersten 9 der o.a. Prozeduren pro Familie)
- `gamma`: truncation parameter für die Mittelung mit der Bonferroni-Adjustierung (ein Wert pro Familie)

Die Ausgabe besteht lediglich aus den adjustierten p-Werten. Nachfolgend nur ein kleines einfaches Beispiel. Das Paket Mediana bietet sehr viel mehr Möglichkeiten. Einiges ist den Vignetten dieses Pakets auf "cran" des R-Projekts zu entnehmen.

Beispiel R-5.3:

Es soll hier an das Beispiel aus der Kurzbeschreibung der Gatekeeping-Methode in Kapitel 2.6.3 angeknüpft werden, bei der 4 Parameter A, B, C und D an einer Versuchs- sowie einer Kontrollgruppe erhoben wurden. Die 4 entsprechenden t-Tests haben folgende p-Werte ergeben: 0.0482, 0.0217, 0.0391, 0.0125. Zunächst werden die beiden Hypothesenfamilien definiert, die Methoden zur α-Adjustierung mit den dazugehörigen truncate-Parametern festgelegt. Schließlich werden sowohl die serielle (`MultipleSequenceGatekeepingAdj`) als auch die parallele Gatekeeping-Methode (`ParallelGatekeepingAdj`) angewandt:

```
pvalues<-c(0.0482,0.0217,0.0391,0.0125)
family = families(family1 = c(2, 4),family1 = c(1, 3))
component.procedure = families(family1 ="HommelAdj",family2 ="HolmAdj")
gamma = families(family1 = 1.0,family2 = 0.8)

AdjustPvalues(pvalues, proc = "MultipleSequenceGatekeepingAdj",
  par = parameters(family=family,
                   proc=component.procedure,
                   gamma=gamma))
```

```
[1] 0.0482 0.0217 0.0482 0.0217
```

```
AdjustPvalues(pvalues, proc = "ParallelGatekeepingAdj",
  par = parameters(family=family,
                   proc=component.procedure,
                   gamma=gamma))
```

```
[1] 0.0782 0.0217 0.0782 0.0217
```

Soll die parametrische multiple Testprozedur angewandt werden, so sind noch zusätzlich die Korrelationen der 4 Parameter A, B, C und D in Form einer Matrix anzugeben. Diese müsste vorher aus den Rohdaten ermittelt werden:

```
...
```

```
corr.mat<-matrix(c(1.0, 0.6, 0.5, 0.8,
                   0.6, 1.0, 0.7, 0.9,
                   0.5, 0.7, 1.0, 0.7,
                   0.8, 0.9, 0.7, 1.0),nrow=4,ncol=4)
AdjustPvalues(pvalues, proc = "NormalParamAdj",
              par = parameters(family=family, proc=component.procedure,
                    corr=corr.mat, gamma=gamma))
```

```
[1] 0.11244289 0.05503732 0.09363895 0.03301657
```

15. 5. 4 multcomp

Das Paket `multcomp` wurde bereits in 15.1.2 sowie 15.2.2 vorgestellt. Es bietet eine Reihe von α-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. 6 Paarweise 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.

Diese Funktion ist auch insofern interessant, als damit bekannte Methoden angewandt werden können, weil diese nur einen Welch t-Test mit einet bestimmten α-Adjustierung beinhalten: z.B. Tamhane T2 (mit Sidak-Adjustierung) und Simes (mit Benjamini & Hochberg-Adjustierung).

Beispiel R-6.1a:

Für die Beispieldaten 3 (`mydata3`, s. Anhang), einem Versuchsplan mit zwei unabhängigen Faktoren (Gruppe und Dosis) aber heterogenen Varianzen (vgl. Anhang) 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
```

Obwohl die Varianzanalyse für den Faktor Dosis einen signifikanten Einfluss ermittelt hatte, ist hiernach 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.

Möchte man eine andere Adjustierung wählen, als die in `p.adjust` verfügbaren, z.B. die Adjustierung von Jianjun Li (vgl. Kapitel 2.2.4), so muss man die Ergebnismatrix (`..$p.value`) zunächst in einen Vektor wandeln, die fehlenden Werte oben rechts der Diagonalen entfernen, bevor man die Korrektur darauf anwenden kann:

```
with(mydata3,pairwise.t.test (x, dosis,"none", pool.sd=F))->ergt
ptvalues<-as.vector(ergt$p.value)
ptvalues <- ptvalues[!is.na(ptvalues)]
pmax<-max(ptvalues)
ptvalues/(ptvalues+1-pmax)
```

```
[1] 0.0979316 0.1012691 0.0338364 0.4126323 0.2008572 0.6300298
```

so erhält man eine Signfikanz für den Vergleich 1-4 (p=`0.0338364`) und damit ein mit der Anova konformes Ergebnis.

Beispiel R-6.1b:

Die Beispieldaten 9 (`mydata9`, s. Anhang) 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 Lüpsen, 2019). 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 Lüpsen, 2019). 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 beim Aufruf `p.adjust.method="none"` festzulegen und die p-Werte über das Ergebnis-Attribut $p.value zu speichern. Diese werden allerdings als Matrix übergeben, so dass diese in einen Vektor gewandelt sowie die NAs entfernt werden müssen. Dazu das folgende Beispiel (vgl. auch die Beispiele R-6.1a und R-4.2b):

Beispiel R-6.2b:

Die gleiche Aufgabe mit denselben Daten wie im vorangegangenen Beispiel. Es werden die p-Werte im Vektor “pvalues“ gespeichert und bereinigt. Anschließend können darauf die p-Wert-Adjustierungen, z.B. `indepBR`, angewandt werden. Zuvor werden noch einmal die rohen p-Werte ausgegeben:

```
with (mydata9t, pairwise.wilcox.test
            (score, Bedingung, paired=T, p.adjust.method="none"))->erg
pvalues<-as.vector(erg$p.value)
pvalues <- pvalues[!is.na(pvalues)]
indepBR (pvalues, 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 α=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 in den Beispielen R-6.1a und R-4.2b gezeigt wurde.

15. 6. 3 ggstatsplot: pairwise_p

pairwise_p vereinigt gewissermaßen die beiden zuvor genannten `pairwise.t.test` und `pairwise.wilcox.test`. Diese Funktion führt paarweise parametrische und nichtparametrische Vergleiche durch, sowohl für homegene und inhomogene Varianzen, als auch für unabhängige und abhängige Stichproben. Die Wahl wird beim Aufruf angegeben:

```
pairwise_p(Dataframe, Faktor, abh.Variable, paired = T/F,
      var.equal = T/F, type = "parametric"/"nonparametric",
      p.adjust.method ="Adj.Methode")
```

Je nach Auswahl werden folgende Tests durchgeführt:

Design	*Testverfahren*
parametrisch, unabh. Stichproben, homogene Varianzen	t-Test
parametrisch, unabh. Stichproben, inhomogene Varianzen	Games-Howell
parametrisch, abh. Stichproben	t-Test
nichtparametrisch, unabh. Stichproben	Dwass-Steel-Crichtlow-Fligner
nichtparametrisch, abh. Stichproben	Durbin-Conover

Als Adjustierungsmethoden werden die der Funktion p.adjust (vgl. 15.5.1) angeboten. Zu beachten ist hierbei, dass das Verfahren von Games-Howell bereits Mehrfachvergleiche über die Verwendung der studentized range-Verteilung berücksichtigt, und daher keine α-Adjustierung mehr erfordert.

Beispiel R-6.3:

Hier wird die gleiche Aufgabe wie in Beispiel R-6.1a gelöst. Basis ist der Datensatz mydata3. Die Daten werden als annähernd normalverteilt, aber mit heterogenen Varianzen angenommen:

```
pairwise_p(mydata3,dosis,x,paired = F,var.equal = F,
           p.adjust.method ="none", type = "parametric")
```

```
Note: The parametric pairwise multiple comparisons test used- Games-
Howell test. Adjustment method for p-values: none

group1 group2  mean.    conf. conf.   se     t.value   df p.value sig
               diff     low   high

<chr>  <chr>  <dbl>     <dbl>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>

1 1      2    1.33      -0.41     3.08 0.387   2.43   8.22   0.145 ns
2 1      3    2.58      -0.803    5.97 0.781   2.34   9.86   0.154 ns
3 1      4    3.28      -0.055    6.61 0.784   2.95  11.1    0.054 ns
4 2      3    1.25      -2.05     4.55 0.729   1.21   8.02   0.637 ns
5 2      4    1.94      -1.28     5.17 0.733   1.88   9.15   0.302 ns
6 3      4    0.694     -3.38     4.77 0.999   0.492 14.9    0.96  ns
```

15. 6. 4 asht: tukeyWelsch

Die Funktion `tukeyWelsch` führt paarweise parametrische und nichtparametrische Vergleiche in ähnlicher Art durch wie die 3 zuvor aufgeführten Funktionen, jedoch wird die Adjustierung über das schrittweise Verfahren von Ryan & Welsch (vgl. Kapitel 5) vorgenommen. Die dazu verwandten Tests sind der (parametrische) t-Test bzw. der (nichtparametrische) Mann-Whitney U-Test. Im nichtparametrischen Fall resultiert daraus „in etwa“ das Verfahren von Campbell & Skillings (vgl. Kpaitel 12.8). Allerdings werden ungleiche Varianzen oder Messwiederholungsfaktoren nicht berücksichtigt. Der Aufuf:

```
with(Dataframe,tukeyWelsch(abh.Variable, Faktor, method="..."))
```

wobei über `method` quasi zwischen parametrisch, “aov“ für Anova mit F-Test oder “sr“ für studentized range-Test, und nichtparametrisch, “kw“ für Kruskal-Wallis-Anova gewählt werden kann. Standardmäßig wird die komplette Ausgabe mit allen Zwischenschritten angezeigt. Das wesentliche Ergebnis, die p-Werte der paarweisen Vergleiche, enthält das Ergebnis-Attribut `..$pairwise.pvalues`.

Beispiel R-6.4b:

Es wird wieder der Beipieldatensatz10 (mydata10, s.Anhang) benutzt und die Mittelwerte der Variablen x für die 6 Gruppen vergleichen:

```
with(mydata10,tukeyWelsch(x,Gruppe,method="kw"))$pairwise.pvalues
```

```
$pairwise.pvalues
        1-2         1-3         1-4         1-5         1-6
0.329755508 0.254787987 0.041093964 0.007336375 0.082802109
        2-3         2-4         2-5         2-6         3-4
0.038035851 0.406847447 0.013108566 0.463935731 0.017710525
        3-5         3-6         4-5         4-6         5-6
0.012175846 0.025610324 0.121505633 0.975752790 0.067754938
```

15. 6. 5 multcomp

Das Paket `multcomp` wurde bereits in 15.1.2 sowie 15.2.2 vorgestellt. Es aber auch elementare paarweise Vergleiche, z.B. über Kontraste, mit einer Reihe von α-Adjustierungen, die in anderen Paketen nicht zu finden sind.

15. 7 Andere Hilfsmittel - Verfahren von Ryan und Welsch

Die in Kapitel 5 beschriebenen und gut bewerteten Verfahren von Ryan und Welsch sind allgemeine Ansätze zur Korrektur der α bzw. p-Werte, die auf beliebige Tests für die Hypothese *k* gleicher Mittelwerte anwendbar ist. Leider gibt es in R bislang nur wenige Funktionen, die diese universellen Methoden zum multiplen Mittelwertvergleich verwenden:

- REGW (agricolae): Verwendung des Anova F-Tests,
- regwq (mutoss): Verwendung der studentized range-Statistik,
- tukeyWelsch (asht): wahlweise Verwendung des Anova F-Tests, der studentiezd range-Statistik oder der Kruskal Wallis-Anova,

jedoch alle nur für unabhängige Stichproben mit homogenen Varianzen. Die Anwendung „mit der Hand“ ist zwar etwas mühselig, soll aber hier einmal demonstriert werden.

Beispiel R-6.7:

Der Beispieldatensatz mydata3 hat heterogene Varianzen bzgl. Faktor dosis. Die Anova-Tabelle auf Basis des Tests von Brown & Forsythe (s. Anhang) zeigt für dosis einen signifikanten Einfluss, der nun aufgeschlüsselt werden soll. Dazu wird der Test von Alexander & Govern verwendet, weil dieser zusammen mit dem Test von James das beste Verfahren ist.

Dazu wird das schrittweise Verfahren von Ryan benutzt. Es wird α=0.05 gewählt. Die Mittelwerte sind bereits der Größe nach geordnet:

```
with(mydata3,tapply(x,dosis,mean))
       1        2        3        4
5.166667 6.500000 7.750000 8.444444
```

Nach dem Verfahren von Ryan wird zunächst die globale Hypothese $\mu_{(1)} = ... = \mu_{(4)}$ getestet:

```
ag.test(x~dosis,mydata3)
```

```
  Alexander-Govern Test
------------------------------------------------------------
  statistic  : 8.782242
  parameter  : 3
  p.value    : 0.03233067

  Result     : Difference is statistically significant.
```

Da diese Hypothese (p=0.032) abgelehnt wird, werden im nächsten Schritt $\mu_{(1)} = ... = \mu_{(3)}$ sowie $\mu_{(2)} = ... = \mu_{(4)}$ geprüft, wozu die entsprechenden Teildatensätze gebildet werden müssen:

```
subset(mydata3,dosis!=4)->h123
subset(mydata3,dosis!=1)->h234
ag.test(x~dosis,h123)
ag.test(x~dosis,h234)
```

mit den Ergebnissen

```
  p.value    : 0.05499496
  p.value    : 0.1427337
```

In diesem Schritt werden noch die p-Werte mit dem unkorrigierten α=0.05 verglichen. Da beide darüber liegen, ist eigentlich das Verfahren hier zu Ende. Um dieses aber besser demonstrieren zu können, wird einmal angenommen, dass die erste Hypothese (trotz p=0.055) abgelehnt wird. Dann werden im nächsten Schritt $\mu_{(1)} = \mu_{(2)}$ und $\mu_{(2)} = \mu_{(3)}$ getestet:

```
subset(h123,dosis!=3)->h12
subset(h123,dosis!=1)->h23
ag.test(x~dosis,h12)
ag.test(x~dosis,h23)
```

mit den Ergebnissen

```
  p.value    : 0.0410063
  p.value    : 0.2616119
```

Für diesen Schritt ist nun folgende α-Korrektur erforderlich: $\alpha' = 1 - (1 - \alpha)^{r/k}$. Bei r=2 und k=4 ist α'=0.025, so dass beide p-Werte als nicht signifikant anzusehen sind. Das Ergebnis ist

dann:

- $\mu_{(2)} = \mu_{(3)} = \mu_{(4)}$ wird angenommen,
- $\mu_{(1)} = \mu_{(2)} = \mu_{(3)}$ wird abgelehnt, aber $\mu_{(1)} = \mu_{(2)}$ wie auch $\mu_{(2)} = \mu_{(3)}$ angenommen,

Somit bleibt übrig: sowohl $\mu_{(1)}$ und $\mu_{(3)}$ als auch $\mu_{(1)}$ und $\mu_{(4)}$ unterscheiden sich.

15. 8 Weitere Pakete und Funktionen

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

- `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.
- `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.
- `multxpert`
  Dieses bietet fast dieselben Verfahren wie `Mediana`.
- `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.
- `stats`: `quade.test`
  Der Test von *Quade* (vgl. Kapitel 12.9) ist sowohl im Basis-Paket `stats`, als auch in `PMCMRplus` enthalten.

15. 9 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. 9. 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. 9. 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
}
```

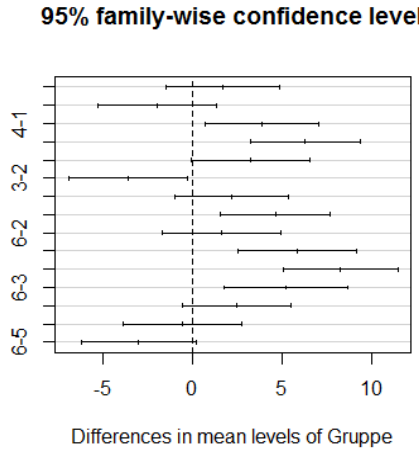
15. 10 Grafische Darstellungen

Eine Form der grafischen Darstellung bei multiplen Mittelwertvergleichen ist die der Konfidenzintervalle. Üblicherweise werden dazu multiple Boxplots oder Error Bar Charts gewählt. Al-

ternativ können aber auch Konfidenzintervalle für die Mittelwertdifferenz aufgestellt werden. 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.

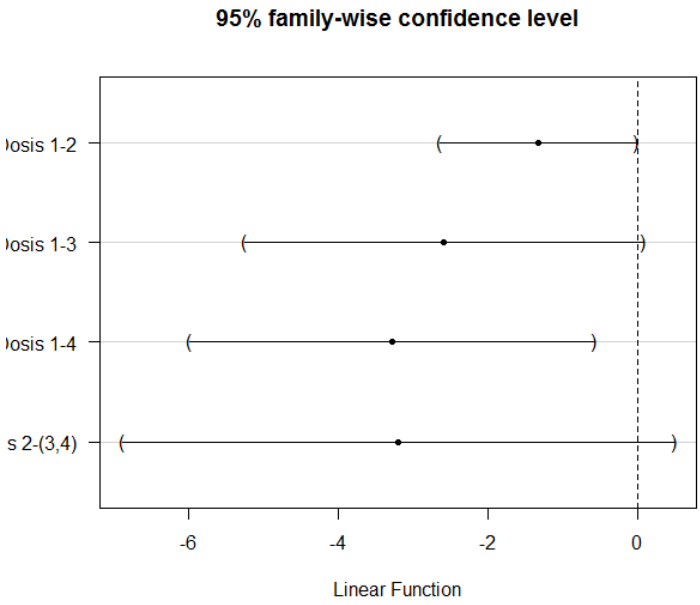
```
result <- TukeyHSD(aov(x~Gruppe,mydata10))
plot(result)
```



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:

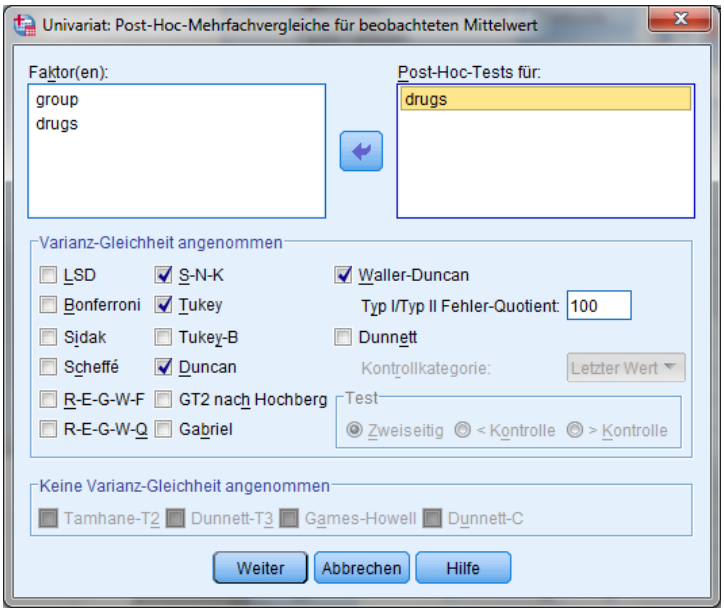
für homogene Varianzen		
	paarw. Vergl.	homog. Untergr.
Fisher LSD (ungeschützt)	x	
Fisher LSD mit Bonferroni-Adj.	x	x
Fisher LSD mit Sidak-Adj.		x
Tukey HSD		x
Tukey B		x
Newman-Keuls S-N-K		x
Duncan		x
R-E-G-W-F	x	x
R-E-G-W-Q	x	x
Hochberg GT2		x
Gabriel	x	x
Waller-Duncan	x	
Scheffé	x	
Dunnett	x	

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 von Lüpsen (2019).

Beispiel S-1:

Zunächst soll ein Beispiel für den Fall homogener Varianzen gerechnet werden. Dazu werden die Beispieldaten 2 (`mydata2`, s. Anhang) verwendet. 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 α, da sich ja die Mittelwerte einer Untergruppe nicht mehr signifikant unterschieden. Ein Wert nahe α 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`, s. Anhang) verwendet. 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 Lüpsen, 2019) 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 (1989) 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)

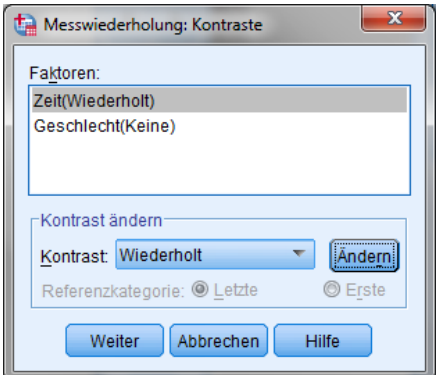
Beispiel S-3:

Dazu werden die Beispieldaten 5 (`winer518`, s. Anhang) verwendet. 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), 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 (Lüpsen, 2019) 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).

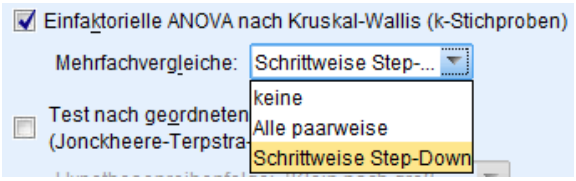
Darüber hinaus kann es sinnvoll sein, einen parametrsichen Test auf die transformierte abhängige Variable anzuwenden, wenn z.B. vorher eine entsprechende Varianzanalyse, z.B. nach der RT-, INT- oder ART-Methode, durchgeührt worden war (vgl. Kapitel 14). Alle drei Verfahren werden kurz vorgestellt, wenn auch die Bonferroni-Adjustierung unbedingt vermieden werden sollte.

Beispiel S-5:

Es werden wieder die Beispieldaten 2 (`mydata2`, s. Anhang) verwendet. 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).
```

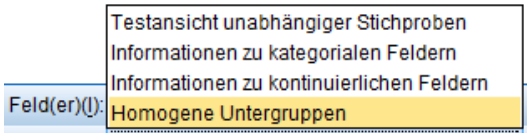
Erfolgt die Testauswahl über die Eingabemasken, so ist wie folgt auszuwählen:



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 über-

prüft. 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

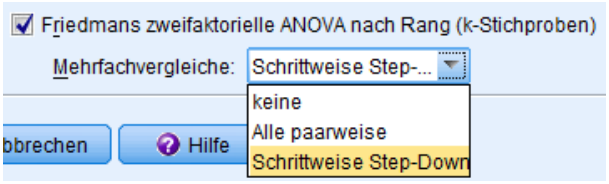
Jede Zeile testet die Nullhypothese, dass die Verteilungen von Stichprobe 1 und Stichprobe 2 gleich sind.
Asymptotische Signifikanzen (2-seitige Tests) werden angezeigt. Das Signifikanzniveau ist ,05.

Beispiel S-6:

Es werden wieder die Beispieldaten 4 (`winer518`, s. Anhang) verwendet, die Messwiederholungen auf dem Faktor Zeit aufweisen. Und zwar wird der Faktor `Zeit` untersucht, zunächst über einen globalen Test mittels Friedman-Test, zusätzlich mit Mittelwertvergleichen über das Verfahren von Campbell und Skillings. Die SPSS-Syntax dafür:

```
NPTests
  /Related Test (t1 t2 t3) Friedman (compare=stepwise) .
```

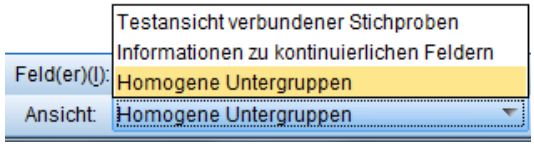
Erfolgt die Testauswahl über die Eingabemasken, so ist wie folgt auszuwählen:



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.

In dem Varianzanalyse-Skript (Lüpsen, 2019, Kapitel 4.3.7) war mit diesen Daten eine nicht-parametrsiche Varianzanalyse nach der INT-Methode durchgeführt worden. Dazu wurden folgende Transformationen vorgenommen:

```
Aggregate
   /outfile=* mode=addvariables
   /break= /nc=NU(x).
Rank Variables=x / rank into Rx.
compute nsx=Idf.normal(Rx/(nc+1),0,1).
execute.
```

Die entsprechende Anova-Tabelle 4-13 ist im Anhang zu finden. Es sollen nun für die Variable drugs Mittelwertvergleiche nach den Methoden von Tukey, SNK und Ryan-Welsch (mittels studentized range statistic) durchgeführt werden:

```
UNIANOVA nsx BY group drugs
  /METHOD=SSTYPE(3)
  /POSTHOC=drugs(SNK TUKEY QREGW)
  /DESIGN=group drugs group*drugs.
```

Multiple Comparisons

Abhängige Variable: nsx							
	(I)	(J)	Mittlere Differenz (I-J)	Standardfehler	Sig.	95%-Konfidenzintervall	
						Untergrenze	Obergrenze
Tukey-HSD	1	2	-,5760	,30864	,268	-1,4250	,2729
		3	-,8212	,31696	,070	-1,6930	,0507
		4	-1,4996*	,30864	,000	-2,3485	-,6506
	2	1	,5760	,30864	,268	-,2729	1,4250
		3	-,2451	,29759	,843	-1,0637	,5734
		4	-,9235*	,28870	,018	-1,7176	-,1294
	3	1	,8212	,31696	,070	-,0507	1,6930
		2	,2451	,29759	,843	-,5734	1,0637
		4	-,6784	,29759	,130	-1,4970	,1402
	4	1	1,4996*	,30864	,000	,6506	2,3485
		2	,9235*	,28870	,018	,1294	1,7176
		3	,6784	,29759	,130	-,1402	1,4970

nsx					
		N	Untergruppe		
			1	2	3
Student-Newman-Keuls[a,b,c]	1	7	-,7658		
	2	9	-,1897	-,1897	
	3	8		,0554	
	4	9			,7338
	Sig.		,069	,426	1,000
Ryan-Einot-Gabriel-Welsch-Bereich[c,d]	1	7	-,7658		
	2	9	-,1897	-,1897	
	3	8		,0554	,0554
	4	9			,7338
	Sig.		,173	,676	,071

16. 4 α-Adjustierungen

Beispiel S-7:

Noch einmal zurück zu dem Beispieldatensatz 3 (`winer518`). In den Beipielen S-1 und S-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 α=0,05. Da dieser sogar kleiner als $\alpha' = 0{,}05/3 = 0{,}0167$ ist, dem nach Bonferroni korrigierten α, 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`, s. Anhang) 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 α-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 α' 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 α' 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.

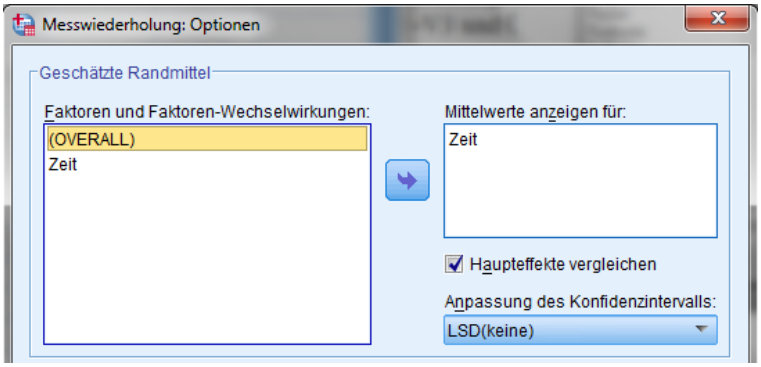
Bei *Li* ist schnell zu sehen, dass ebenfalls die Vergleiche 1 bis 4 signifikant sind.

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.

Beispiel S-9:

Man kann sich im o.a. Beispiel S-8 den „Umweg“ über die paarweisen t-Tests unter Umständen sparen, wenn man statt dessen bei der Varianzanalyse unter „Optionen“ im Abschnitt „Geschätzte Randmittel“ „Haupteffekte vergleichen“ für den zu analysierenden Faktor aktiviert



In dem Fall werden die Gruppen mit Fishers LSD verglichen und für jeden Vergleich ein p-Wert ausgegeben. Voraussetzung ist allerdings Varianzhomogenität, die allerdings bei diesen Beispieldaten (`mydata9`) nicht gegeben ist. Dennoch hier die Ausgabe:

Paarweise Vergleiche						
(I)Zeit	(J)Zeit	Mittlere Differenz (I-J)	Standardfehler	Sig.[b]	95% Konfidenzintervall für die Differenz[b]	
					Untergrenze	Obergrenze
1	2	-1,400*	,581	,039	-2,715	-,085
	3	-2,200*	,442	,001	-3,200	-1,200
	4	-4,100*	1,178	,007	-6,765	-1,435
2	1	1,400*	,581	,039	,085	2,715
	3	-,800	,593	,210	-2,140	,540
	4	-2,700*	1,136	,041	-5,269	-,131
3	1	2,200*	,442	,001	1,200	3,200
	2	,800	,593	,210	-,540	2,140
	4	-1,900	,912	,067	-3,964	,164
4	1	4,100*	1,178	,007	1,435	6,765
	2	2,700*	1,136	,041	,131	5,269
	3	1,900	,912	,067	-,164	3,964

Die p-Werte liegen etwas höher als die der t-Tests vom vorigen Beispiel. Das Prozedere der α-Adjustierung ist dasselbe wie in S-8. Zwar können bei der Aktivierung der paarweisen Vergleiche Adjustierungen nach Bonferroni oder Sidak gewählt werden, aber, wie schon häufig erwähnt, sollten diese nach Möglichkeit wegen ihres konservativen Verhaltens vermieden werden.

Anhang

1 Beispieldatensätze

Weitestgehend werden die Datensätze aus dem Skript über Varianzanalysen (Lüpsen, 2019) verwendet, zumal die dort gefundenen Anova-Ergebnisse hier mittels multipler Mittelwertvergleiche weiter untersucht werden. Bei Versuchsplänen mit Messwiederholungen müssen für Analysen in R die Daten umstrukturiert werden. Die dafür erforderlichen Anweisungen sind dem Kapitel 5 des o.a. Skripts zu entnehmen.

Alle Datensätze können von folgender Webseite heruntergeladen werden, wo größtenteils im txt-, R- (RData) und SPSS-Format (.por bzw. .sav) vorliegen:

http://www.uni-koeln.de/~luepsen/daten/

1. 1 Beispieldaten 2 (mydata2):

2 Patientengruppen (Faktor A: Kontrollgruppe und Behandlungsgruppe) werden jeweils in 4 Gruppen eingeteilt, die dann jeweils ein Medikament (Faktor B: drug 1, 2, 3 oder 4) erhalten. Die Zellenbestzungszahlen sind in diesem Datensatz ungleich. Die abhängige Variable ist eine Beurteilung auf einer Skala von 1 bis 9, also ordinal.

group	drug 1	drug 2	drug 3	drug 4
Kontrolle	4 5 5 6	5 6 6 7 7	5 6 7 7	5 6 6 7 9
Behandlung	2 3 3	3 3 4 5	3 4 5 8	6 7 9 9

In R wie auch in SPSS werden hierfür die Variablennamen `group`, `drug` und `x` verwendet. In R müssen group und drug als „factor“ deklariert sein. In R hat der Dataframe den Namen `mydata2`.

Anova-Ergebnisse: parametrische Varianzanalyse

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	12,024	1	12,024	7,346	,012
drugs	46,560	3	15,520	9,483	,000
group * drugs	17,932	3	5,977	3,652	,026
Fehler	40,917	25	1,637		

Tabelle 4-2 bzw. 4-4

Anova-Ergebnisse: RT-Varianzanalyse

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	364,168	1	364,168	8,400	,008
drugs	1157,722	3	385,907	8,902	,000
group * drugs	464,611	3	154,870	3,572	,028
Fehler	1083,787	25	43,352		

Tabelle 4-6 bzw. 4-8

Anova-Ergebnisse: INT-Varianzanalyse

```
Model:
nsx ~ group * drugs
            Df Sum of Sq     RSS      AIC F value    Pr(>F)
<none>                    9.3768 -25.5229
group        1    2.9521 12.3289 -18.4905  7.8708 0.0095852 **
drugs        3   10.6917 20.0684  -6.4128  9.5019 0.0002289 ***
group:drugs  3    4.1290 13.5058 -19.4817  3.6696 0.0256032 *
```

Tabelle 4-13

Ergebnisse der Varianzhomogenitäts-Tests:

group: p=,314 , drugs: p=,004, group*drugs: ,951

1. 2 Beispieldaten 3 (mydata3):

Bei diesem Datensatz werden wieder 2 Patientengruppen (Faktor A: Kontrollgruppe und Behandlungsgruppe) jeweils in 4 Gruppen eingeteilt, die dann jeweils ein Medikament in 4 verschieden hohen Dosierungen (Faktor B: dosis 1, 2, 3 oder 4) erhalten. Die Zellenbestzungszahlen sind in diesem Datensatz ungleich. Die abhängige Variable ist eine Beurteilung der Reaktion auf einer Skala von 1 bis 20. Durch Abbruch der Therapie kommt es hier zu unterschiedlichen n_i. Das Skalenniveau ist dasselbe wie im ersten Beispiel, also quasi metrisch, wenn auch streng genommen als Beurteilung ordinal.

gruppe	dosis 1	dosis 2	dosis 3	dosis 4
Kontrolle	4 5 7	5 6 7 6 7 8	4 6 8 9	5 6 7 9 10
Behandlung	4 5 6	6 6 7 7	5 7 11 12	5 9 11 14

In R wie auch in SPSS werden hierfür die Variablennamen `gruppe`, `dosis` und `x` verwendet. In R müssen gruppe und dosis vom Typ „factor“ deklariert sein. In R hat der Dataframe den Namen `mydata3`.

Anova-Ergebnisse: semiparametrische Varianzanalyse (Welch & James-Verfahren)

```
             Df Df.err  Sum Sq Mean Sq F value  Pr(>F)
gruppe        1 21.326   9.116  9.1162  1.4458 0.24239
dosis         3 18.618  45.922 15.3074  3.2354 0.04581 *
gruppe:dosis  3 12.422  11.072  3.6908  0.7499 0.54246
Residuals    25        120.617  4.8247
```

Ergebnisse des Varianzhomogenitäts-Tests:

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value   Pr(>F)
group  3  4.9647 0.006675 **
      29
```

1. 3 Beispieldaten 4 (winer518):

Der folgende Datensatz ist dem Buch von B.J.Winer (1991, S. 518) entnommen. Die Einstellung zu einem Thema wurde von Männer und Frauen dreimal im Abstand von mehreren Monaten auf einer ordinalen Skala von 1 - 9 (negativ - positiv) erfasst:

Geschlecht	Versuchsperson	t1	t2	t3
Männer	1	4	7	2
	2	3	5	1
	3	7	9	6
	4	6	6	2
	5	5	5	1
Frauen	6	8	2	5
	7	4	1	1
	8	6	3	4
	9	9	5	2
	10	7	1	1

In R muss `Geschlecht` vom Typ „factor“ deklariert sein, ebenso die für die Umstrukturierung zu ergänzende Fallkennzeichnung, etwa `Vpn`. In R hat der Dataframe den Namen `winer518`.

Anova-Ergebnisse:

```
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
```

Tabelle 6-1 bzw. 6-3

```
$ANOVA
           Effect DFn DFd        SSn      SSd          F            p
2      Geschlecht   1   8   53.33333 701.8333  0.6079316 4.580116e-01
3            Zeit   2  16  698.60000 249.9667 22.3581811 2.325487e-05
4 Geschlecht:Zeit   2  16  501.26667 249.9667 16.0426724 1.502651e-04
```

Tabelle 6-3 bzw. 6-4

1. 4 Beispieldaten 6 (winer568):

Der folgende Datensatz ist dem Buch von B.J.Winer (1991, S. 568) entnommen. Hierbei handelt es sich um ein Lernexperiment, bei dem in 4 aufeinanderfolgenden Versuchen (Faktor Zeit) jeweils ein Score von 0 bis 20 erzielt werden konnte. Die 12 Versuchspersonen sind bzgl. 2 Kriterien A bzw. B (Faktoren A und B) in jeweils 2 Gruppen eingeteilt worden.:

A	B	Versuchsperson	V1	V2	V3	V4
A1	B1	1	1	6	5	7
		2	0	6	7	9
		3	3	8	8	9
	B2	4	2	7	12	15
		5	1	6	8	9
		6	3	7	10	11
A2	B1	7	1	2	7	12
		8	1	1	4	10
		9	1	1	4	8
	B2	10	2	2	8	12
		11	3	2	10	15
		12	2	2	7	13

In R hat der Dataframe den Namen `winer568`.

Anova-Ergebnisse:

```
    Effect DFn DFd           F         p p<.05        ges
2        A   1   8   3.3160388 0.1060896       0.22755888
3        B   1   8   8.1885856 0.0211004     * 0.42112020
5     Zeit   3  24 235.4228709 0.0000000     * 0.89487936
4      A:B   1   8   0.1732461 0.6881851       0.01515789
6   A:Zeit   3  24  25.8348420 0.0000001     * 0.48298681
7   B:Zeit   3  24   4.8246813 0.0090990     * 0.14854504
8 A:B:Zeit   3  24   0.9709958 0.4226642       0.03392018
```

Tabelle 6-7

```
$`Mauchly's Test for Sphericity`
    Effect         W           p p<.05
5     Zeit 0.1007174 0.009413644     *
```

1. 5 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

Anova-Ergebnisse:

```
     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     *

$`Sphericity Corrections`
     Effect      GGe       p[GG] p[GG]<.05      HFe       p[HF]
p[HF]<.05
2 Bedingung 0.539389 0.003722797         * 0.640386 0.001978256
```

1. 6 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	

Anova-Ergebnisse:

```
           Df Sum Sq Mean Sq F value   Pr(>F)
Gruppe      5 253.30   50.66   15.73 1.67e-07 ***
Residuals  29  93.38    3.22
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.3176 0.8983
```

Literaturhinweise

Abundis, Marisela (2001): Multiple Comparison Procedures in Factorial Designs using the Alogned Rank Transformation, Thesis, Texas Tech University

Akritas M.G., Arnold S.F. & Brunner E. (1997): *Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs,* Journal of the American Statistical Association, Volume 92, Issue 437 , pages 258-265

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.

Bernhardson, Clemens S. (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

Blakesley R.E. , Mazumdar S. , Dew M.A. , Houck P. R. , Tang G. , Reynolds C. F. and Butters, M.A. (2009): *Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research*, Neuropsychology, Mar 2009; 23(2), pp 255–264

Blanchard, Gilles & Roquain, Etienne (2009): *Adaptive False Discovery Rate Control under Independence and Dependence,* Journal of Machine Learning Research 10, pp 2837-2871

Bohlert, Thorsten (2016): The Pairwise Multiple Comparison of MeanRanks Package (PMCMR), http://CRAN.R-project.org/package=PMCMR.

Boik, R.J. (1981): A priori Tests in Repeated Measures Designs: Effects of Nonsphericity. Psychmetrika, 46 (3), pp 241-255.

Braun, Henry I. & Tukey, John W. (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.

Brown, R.A. & Forsythe, A.B. (1974): *The small sample behaviour of some statistics which test the equality of several means*, Technometrics 16, 129-302

Brunner, E. & Munzel, U. (*2002): Nichtparametrische Datenanalyse - unverbundene Stichproben*, Springer, ISBN 3-540-43375-9

Carmer, S.G. & Walker, W.M. (1985): *Pairwise Multiple Comparisons of Treatment Means in Agronomic Research*, Journal of Agronimic Education, Vol. 14

Carmer, S.G. & Swanson, M.R. (1973): *An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods*. Journal of the American Statistical Association, Vol. 68, No. 341, pp.66-74.

Chuang-Stein, Christy & Tong, Donald M. (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

Cochran, W.G. (1950): *The comparison of percentages in matched samples*. Biometrika 3

Conover, W. J. & Iman, R. L. (1981): *Rank transformations as a bridge between parametric and nonparametric statistics*. American Statistician 35 (3): 124–129.

Conover, W.J. & Iman, R.L. (1979): *On multiple comparison procedures.* CIC-14 Report Collection, UCLA, https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-07677-MS

Cribbie, R.A. (2000): *Improved Pairwise Multiple Comparison Procedures*, thesis at the University of Manitoba. https://mspace.lib.umanitoba.ca/bitstream/handle/1993/19559/Cribbie_Improved_pairwise.pdf

Dallal, Gerard E. (2001/2012): *Multiple Comparison Procedures* aus *The Little Handbook of Statistical Practice* http://www.jerrydallal.com/LHSP/mc.htm

Day, R.W. & Quinn, G.P. (1989): Comparisons of Treatments after an Analysis of Variance in Ecology, Ecological Monographs, Vol. 59, No. 4, pp 433-463

Dayton, C. Mitchell (1998): *Information Criteria for the Paired-ComparisonsProblem*, The American Statistician, 52:2, pp. 144-151.

Dijkstra, J. B. (1987). Analysis of means in some non-standard situations. Technische Universiteit, Eindhoven DOI: 10.6100/IR272914.

Dmitrienko, A, Tamhane, A.C. (2007): Gatekeeping procedures with clinical trial applications. Pharmaceutical Statististics, 6, pp. 171–180.

Dmitrienko, A, Tamhane, A.C., Bretz, F. (Herausgeber) (2010): *Multiple Testing Problems in Pharmaceutical Statistics*, Chapman & Hall, Boca Raton

Dmitrienko, A, Bretz, F., Westfall, P.H., Troendle, J., Wiens, B.L., Tamhane, A.C., Hsu, J.C. (2012): *Multiple Testing Methodology*, https://web.njit.edu/~wguo/Math654_2012/DTB_Chapter2.pdf

Donoghue, John R. (1998): Implementing Shaffer‘s Multiple Comparison Procedure for Large Number of Groups, Educational Testing Services, Princeton, RR-98-47

Durbin, J. (1951): *Incomplete Blocks in Ranking Experiments*, British Journal of Mathematical and Statistical Psychology, 4 (2), pp. 85-90.

Einot, I. & Gabriel, K.R. (1975): *A Study of Powers of Several Methods of Multiple Comparisons*, Journal of the American Statistical Association, Vol 70, Number 361

Freedman, David A. (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

Gabriel, K.R. (1978): *A simple Method of Multiple Comparisons of Means*, Journal of the American Statistical Association, Vol 73, Number 364

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

Garcia S., Fernandez A., Luengo J. & Herrera F (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

Ge Yongchao, Dudoit Sandrine and Speed Terence P. (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

Genovese, Christopher R. (2003): *A Tutorial on False Discovery Control*, http://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf

Gonzalez, Richard (2009): *Contrasts and Post Hoc tests (Lecture Notes)*, University of Michigan, Ann Arbor, http://www-personal.umich.edu/~gonzo/coursenotes/file3.pdf

Heyse, J. and Rom, D. (1988), *Adjusting for Multiplicity of Statistical Tests in the Analysis of Carcinogenicity Studies*, Biometrical Journal, 30, pp 883–896.

Hollander M, Wolfe D A, Chicken E (2014): *Nonparametric Statistical Methods*, Wiley

Hsu, Jason (1999): *Multiple Comparisons: Theory and Methods*, Chapman & Hall

Kirk, R.E. (1995), *Experimental Design* (Jrd ed.), Pacific Grove, CA:Brooks/Cole

Keselman, H.J. (1994): *Stepwise and Simultaneous Multiple Comparison Procedures of Repeated Measures' Means.* Journal of Educational Statistics, Vol. 19 (2), pp. 127-162.

Keselman, H. J. & Lix, L.M. (1995): *Improved Repeated Measures Stepwise Multiple Comparison Procedures.* Journal of Educational and Behavioral Statistics, Vol 20, No. 1, pp 83-99

Konietschke F., Hothorn L.A., Brunner E. (2012): *Rank-based multiple test procedures and simultaneous confidence intervals*, Electronic Journal of Statistics Vol. 6, 738–759

Korhonen, M.P. (1982): On the Performance of Some Multiple Comparison Procedures with Unequal Variances. Scandinavian Jounal of Statistics, 9, pp 241-247.

Kromrey, JD & La Rocca, MA (1995): *Power and Type I Error Rates of New Pairwise Multiple Comparison Procedures under Heterogeneous Variances*. The Journal of Experimental Education, Vol. 63, No. 4, pp. 343-362.

Li, Jianjun (2007): *A two-step rejection procedure for testing multiple hypotheses*, Journal of Statistical Planning and Inference 138, pp 1521 – 1527

Liu, W. (1997): *On Step-up Tests for Comparing Several Treatments*: Statistica Sinica 7, pp 957-972.

Liu, L., Lee, C.C., Peng, J. (2002): *Max–min multiple comparison procedure for isotonic dose–response curves*. Journal of Statistical Planning and Inference, 107, pp 133–141.

Lunney, G.H. (1970): *Using Analysis of Variance with a dichtomous dependent variable: an empirical study*. Journal of Educational Measurement Volume 7, Issue 4

Luepsen, Haiko (2017):Comparison of nonparametric analysis of variance methods: A Vote for van der Waerden. Communications in Statistics - Simulation and Computation, Volume 30, pp 1-30, DOI: 10.1080/03610918.2017.1353613

Lüpsen, Haiko (2019): R Funktionen zur Varianzanalyse: http://www.uni-koeln.de/~luepsen/R/

Lüpsen, Haiko (2020): 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/texte/nonpar-anova.pdf

Marascuilo, Leonard A. & McSweeney, Maryellen (1977): *Nonparametric and distribution-free methods for the social sciences,* Brooks/Cole Pub. Co.

Marcus R, Peritz E, Gabriel K (1976). *On closed testing procedures with special reference to ordered analysis of variance*. Biometrika 63, pp 655–660

Maxwell, Scott E. (1989): *Pairwise Multiple Comparisons in Repeated Measures Designs*, Journal of Educational Statistics, Vol. 5, No. 3, pp. 269-287

Munzel, Ullrich & Tamhane, Ajit C. (2002): *Nonparametric Multiple Comparisons in Repeated Measures Designs for Data with Ties*, Biometrical Journal 44 Nr. 6, pp 762–779

Nashimoto, K & Wright, F.T. (2005): *Multiple comparison procedures for detectingdi(erences in simply ordered means*, Computational Statistics & Data Analysis 48, pp 291 – 306

Nashimoto, K & Wright, F.T. (2007): *Nonparametric multiple-comparison methods for simply ordered medianss*, Computational Statistics & Data Analysis 51, pp 5068 – 5076.

Neuhäuser, Markus (2007): An exact two-sample test based on thebaumgartner-weiss-schindler statistic and a modification of lepage's test. Communications inStatistics - Theory and Methods, 29:1, pp 67-78.

R: http://cran.r-project.org/web/packages/agricolae/vignettes/tutorial.pdf

Rafter J.A., Abell Martha L. , Braselton James P. (2002): *Multiple Comparison Methods for Means*, SIAM Review, Vol. 44, No. 2, pp 259-278

Ramseyer, Gary & Tcheng Tse-Kia (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

Rao, C.V. & Swarupchand, U. (2009): *Multiple Comparison Procedures - a Note and a Bibliography* , Journal of Statistics, Volume 16, 2009, pp. 66-109

Reilly, Cavan (2013): *Multiple comparison procedures*, University of Minesota, http://www.biostat.umn.edu/~cavanr/asfChap4-2.pdf

Rudolph, P.E. (1988): *Robustness of Multiple Comparison Procedures: Treatment versus Control,* Biometrical Journal 50, 1, 41-45

Scott, A. J. & Knott, M. (1974): *A Cluster Analysis Method for Grouping Means in the Analysis of Variance*, Biometrics, Vol. 30, No. 3, pp. 507-512

Seaman, M.A., Levin, J.R., Serlin, C.S. (1991): New Developments in Pairwise Multiple Comparisons: Some Powerful and Practicable Procedures. Psychological Bulletin, Vol. 110, No. 3, pp. 577-586.

Shaffer, Juliet P. (1995): *Multiple Hypothesis Testing*, Annual Reviews Psychology, 46, pp. 561-584

Simes, R.J. (1986): An improved Bonferroni procedure for multiple tests of significance, Biometrika, Vol 73, pp 751-754

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

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

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

Saville, D. J. (1990): Multiple Comparison Procedures: The Practical Solution. The American Statistician, Vol. 44, No. 2 (May, 1990), pp. 174-180.

Smith, W.C. & Han, C.P. (1981): *Error rate for testing a contrast after a significant F test*. Communications in Statistics - Simulation and Computation, 10:6, pp 545-556.

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

Tamhane, A.C. (1979): *A Comparison of Procedures for Multiple Comparisons of Means With unequal Variances,* Journal of the American Statistical Association, Vol 74, Number 365

UCLA: Contrast Coding Systems for categorical variables: http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

UCLA: *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

UCLA: Institute for Digital Research and Education, UCLA: *R Library: Contrast Coding Systems for categorical variables:* http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

U.S.Federal Department of Health (2017): *Multiple Endpoints in Clinical Trials - Guidance for Industry*. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf

Welsch, Roy E. (1977): *Stepwise Multiple Comparison Procedures.* Journal of the American Statistical Association, Vol. 72, No. 359, pp.566-575.

Westfall, P. H. & Young, S. S. (1993): *Resampling based Multiple Testing: Examples and Methods for p-value Adjustment*, New York, Wiley

Wikipedia: http://en.wikipedia.org/wiki/Multiple_comparisons_problem

Wikipedia: http://en.wikipedia.org/wiki/Post-hoc_analysis

Wikipedia: http://en.wikipedia.org/wiki/Bonferroni_correction

Wikipedia: http://en.wikipedia.org/wiki/Holm-Bonferroni_method

Wikipedia: http://en.wikipedia.org/wiki/Resampling_(statistics)

Wikipedia: http://en.wikipedia.org/wiki/False_discovery_rate

Wikipedia: http://en.wikipedia.org/wiki/Heteroscedasticity-consistent_standard_errors

Wikipedia: http://en.wikipedia.org/wiki/Van_der_Waerden_test

Wilcox, Rand R. (2013): *New Statistical Procedures for the Social Sciences: Modern Solutions To Basic Problems*, Psychology Press, Lawrence Erlbaum Assoc

Winer, B.J. et.al. (1991): *Statistical Principles in Experimental Design*, Wiley, New York

Xie, Changchun (2012): *Weighted multiple testing correction for correlated tests*. Statistics in Medicine, 31, pp 341–352.

Zolman, James F. (1993): *Experimental Design and Statistical Inference*, Oxford University Press