



**Varianzanalysen -  
Prüfen der Voraussetzungen und  
nichtparametrische Methoden  
sowie  
praktische Anwendungen mit R und SPSS**

Version 4.3  
(12.05.2021)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: [Luepsen@Uni-Koeln.de](mailto: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 nicht-parametrischen statistischen Verfahren besteht, insbesondere im Bereich Varianzanalyse. Immerhin sind die dazu zählenden Verfahren, vom t-Test bis zur mehrfaktoriellen Analyse mit Messwiederholungen, die am meisten verwendeten der Statistik. Umso erstaunlicher ist es, dass in den großen Statistiksystemen, insbesondere in SPSS, außer den alt bekannten 1-faktoriellen Klassikern Kruskal-Wallis- und Friedman-Tests keine nichtparametrischen Varianzanalysen angeboten werden. An Methoden mangelt es nicht, wie die nachfolgenden Kapitel sowie die angeführte Literatur zu diesem Thema zeigen.

Immerhin kann man mit mehr oder weniger Aufwand einige dieser Verfahren auch in SPSS durchführen, da sich manche auf die klassische Varianzanalyse zurückführen lassen. Solche Verfahren stehen daher im Vordergrund. Mit S bzw. R lassen sich naturgemäß alle Methoden programmieren. Auch da zeigen sich erstaunlicherweise noch Lücken im Angebot. Daher sind im Anhang selbst erstellte R-Funktionen zu diesem Thema angeführt, die als Bibliothek zum Download bereitgestellt werden.

Da sich zwangsläufig vor Durchführung der Varianzanalyse die Frage stellt: In wieweit sind die Voraussetzungen für die parametrische Analyse erfüllt und wie robust sind die Verfahren, werden diese Fragen auch ausführlich behandelt. Manchmal reichen auch robuste Varianten der „klassischen“ Varianzanalyse, die hier natürlich auch vorgestellt werden.

Dieses waren die Themen meiner Kurse. In den entsprechenden Kursunterlagen waren die Antworten bzw. Lösungen zu den o.a. Fragen und Methoden nur skizziert. Da ich im WWW keine vergleichbare Zusammenstellung gefunden hatte, entschloss ich mich 2012 dazu, die Kursunterlagen beider Kurse (SPSS und R) zu einem Skript „auszubauen“, das als Anleitung benutzt werden kann.

## Einige Jahre später

Nach dem Lesen von weit über 600 Veröffentlichungen zu nichtparametrischen Varianzanalysen habe ich meine Einstellung zur Anwendung dieser Verfahren allerdings ändern müssen: Während allgemein der Glaube herrscht, dass nichtparametrische Analysen eigentlich immer anwendbar seien, insbesondere wenn irgendwelche Voraussetzungen nicht erfüllt sind, so musste ich mich von dieser Annahme verabschieden, was auch deutlich in die letzten Versionen des Skripts eingeflossen ist.

Bei der Vorstellung der Verfahren in diesem Skript interessierte es mich zunehmend, wie gut oder wie schlecht diese unter diversen Bedingungen abschneiden bzw. welche Unterschiede es gibt. Da es nur wenig Literatur über Vergleiche der Verfahren gibt, insbesondere nur unter sehr „einfachen“ Bedingungen, hatte ich mich Ende 2014 entschlossen, selbst mittels Monte Carlo-Simulationen die hier vorgestellten Verfahren zu vergleichen. Ein erster Teil, unabhängige Stichproben betreffend, ist inzwischen abgeschlossen und in der Zeitschrift *Communications in Statistics - Simulation and Computation* veröffentlicht. Artikel mit weiteren Ergebnissen sind derzeit in Arbeit und sollen in Kürze ebenfalls veröffentlicht werden. Insbesondere haben die Ergebnisse o.a. Glauben ebenso deutlich widerlegt.

## Umfang und Lesehinweise

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.

Ursprünglich war geplant, das Thema „multiple Mittelwertvergleiche und  $\alpha$ -Adjustierungen“ ebenfalls in diesem Skript zu behandeln. Allerdings merkte ich schnell bei der Sichtung der Verfahren und der aktuellen Literatur, dass dies ein eigenes „Thema“ sein muss. Dementsprechend gibt es inzwischen dazu ein eigenes Skript, das an gleicher Stelle wie dieses abrufbar ist und das auf das vorliegende Bezug nimmt.

Zu jedem Versuchsplan und zu jeder Methode gibt es nach einer kurzen Beschreibung des Verfahrens jeweils ein ausführliches Beispiel. Dieses wird dann einmal mit R sowie - soweit möglich - einmal mit SPSS durchgerechnet. Die Ergebnistabellen aus R und SPSS sind zum Teil verkürzt wiedergegeben, d.h. Teile, die nicht zum Verständnis erforderlich sind, fehlen hier.

## Benutzung von R innerhalb SPSS

Seit 2010 gibt es eine Schnittstelle zwischen SPSS und R, die es ermöglicht, aus SPSS heraus R zu benutzen, insbesondere Funktionen mit speziellen statistischen Methoden auf die in SPSS verfügbaren Daten anzuwenden. Diese Schnittstelle verbindet gewissermaßen den Bedienungs-komfort und die leichte Handhabung von SPSS mit dem quasi unermesslichen Spektrum von statistischen Methoden in R. Diese Schnittstelle ist in den letzten Jahren komfortabel ausgebaut worden, so dass für deren Benutzung nicht mehr allzu viel Programmierkenntnis erforderlich ist. Eine Einführung in Benutzung inklusive einer elementaren Einführung in R mit den wichtigsten Anweisungen bietet das Dokument „*R als Ergänzung zu SPSS*“ von B. Baltes-Götz von der Universität Trier (<https://www.uni-trier.de/fileadmin/urt/doku/r4spss/R4SPSS.pdf>).

## Literatur

Am Ende ist eine umfangreiche Literaturliste zu finden. Da ich häufiger gebeten worden war, für die eine oder andere Aussage eine Quelle zu nennen, habe ich mich entschlossen, einen großen Teil der benutzten Literatur hier anzuführen. Fast ausnahmslos können die Zeitschriften-Artikel über [scholar.google.de](http://scholar.google.de) aus dem Internet heruntergeladen werden. Die angeführten Bücher sollten in den meisten Universitätsbibliotheken verfügbar sein.

## Disclaimer

Dieser Text wurde nach bestem Wissen erstellt. Ich bin sicher, dass darin noch einige sachliche Mängel oder Fehler enthalten sind. Für entsprechende Hinweise bin ich dankbar. Eine kurze Mail an [luepsen@uni-koeln.de](mailto:luepsen@uni-koeln.de) genügt.

## Weitere Texte

Weitere Texte, wie z.B. *Multiple Mittelwertvergleiche - parametrisch und nichtparametrisch - sowie alpha-Adjustierungen mit praktischen Anwendungen mit R und SPSS* oder *Checking the Homogeneity of Covariance Matrices: some practical aspects*, sowie die meisten der eigenen Veröffentlichungen sind online verfügbar unter:

<http://www.uni-koeln.de/~luepsen/statistik/texte/>

## Historie

- Version 4.3 (12.5.2021): Ergänzungen im Bereich Analysen mit Messwiederholungen, u.a. Huynhs GA und IGA Approximation und das modified Brown-Forsythe Verfahren, beides extrem robuste Verfahren bei inhomogenen Kovarianzmatrizen und fehlender Sphrizität, ergänzt um entsprechende Anwendungsbeispiele.
- Version 4.2 (22.11.2020): wesentliche Ergänzungen im Bereich multivariater Verfahren für Analysen mit Messwiederholungen (u.a. spatial ranks Methoden).
- Version 4.1 (22.10.2020): Korrektur von Fehlern zum Verhalten multivariater Tests.
- Version 4.0 (20.8.2020): Komplette Überarbeitung, Ergänzungen im Bereich Analysen mit Messwiederholungen, insbesondere Split-Plot-Designs, neue eigene R-Funktionen.
- Version 3.2 (28.4.2019): Ergänzung um Beispiele mit etwas „problematischeren“ Datensätzen, diverse Korrekturen sowie eine generelle Überarbeitung.
- Version 3.1 (5.8.2018): Korrekturen an den Puri & Sen-Verfahren.
- Version 3.0 (11.6.2018): Berücksichtigung neuerer Ergebnisse zur Analyse dichotomer Kriteriumsvariablen, GLM-Verfahren und simple effect-Analysen.
- Version 2.4 (20.7.2017): Ausführlichere Behandlung des Falls heterogener Varianzen.
- Version 2.3 (8.2.2017): Hinzunahme GEE und GLMM-Verfahren.
- Version 2.2 (25.11.2016): Hinzunahme logistische Regression mit Messwiederholungen.
- Version 2.1 (30.9.2016): Hinzunahme des multivariaten Tests von Hotelling-Lawley.
- Version 2.0 (29.6.2016): Komplette Überarbeitung des Skripts. Vorstellung zahlreicher neuerer Verfahren, z.B. ART+INT, sowie neuer R-Pakete .

# Inhaltsverzeichnis

1.	Allgemeines zur nichtparametrischen Statistik	1
1. 1	Wichtige Begriffe	1
1. 1. 1	Fehler 1. und 2. Art	1
1. 1. 2	Effizienz eines Tests	2
1. 1. 3	konservative und liberale Tests	2
1. 1. 4	starke und schwache Tests	2
1. 1. 5	robuste Tests	2
1. 1. 6	Pairing	3
1. 2	Methoden für metrische Merkmale	3
1. 3	Methoden für ordinale Merkmale	3
1. 4	Methoden für dichotome Merkmale	3
1. 5	Methoden für nominale Merkmale	4
1. 6	Prüfung auf Normalverteilung	4
1. 7	Prüfung von Voraussetzungen	6
2.	Nichtparametrische Varianzanalysen - Übersicht der Methoden	8
2. 1	Kruskal-Wallis und Friedman (KWF)	9
2. 2	Rank transform Tests (RT)	10
2. 3	Inverse normal transform (INT)	10
2. 4	Aligned rank transform (ART)	11
2. 5	Kombination von Aligned rank transform und Inverse normal transform (INT+ART)	12
2. 6	Puri & Sen-Tests	13
2. 7	van der Waerden	14
2. 8	Bredenkamp Tests - bifaktorieller H-Test	15
2. 9	Akritas, Arnold & Brunner ATS Tests	15
2. 10	Weitere Varianzanalysen für unabhängige Stichproben	16
2. 10. 1	Wilcoxon analysis (WA)	16
2. 10. 2	Gao & Alvo	16
2. 11	Weitere Varianzanalysen für abhängige Stichproben	16
2. 11. 1	Quade	16
2. 11. 2	Skillings & Mack	17
2. 11. 3	Spatial Signs und Spatial Ranks Methoden	17
2. 12	Weitere Varianzanalysen für gemischte Versuchspläne	17
2. 12. 1	Multivariate Tests: Hotelling-Lawley, Pillai und Wilks	18
2. 12. 2	Koch	18
2. 12. 3	Beasley & Zumbo	18
2. 12. 4	Agresti & Pendergast und Akritas & Arnold	18
2. 13	Varianzanalysen für heterogene Varianzen	18
2. 13. 1	Welch und Fligner-Policello	19
2. 13. 2	James 2nd order und Alexander & Govern	19
2. 13. 3	Welch & James	19
2. 13. 4	Brown & Forsythe	19
2. 13. 5	Brunner, Dette und Munk	20

2. 13. 6	Adjustierungen der Freiheitsgrade: $\epsilon$ , GA und IGA	20
2. 13. 7	Box-Korrektur	20
2. 14	Logistische Regression	21
2. 15	GEE und GLMM	21
2. 16	Alternative Rangberechnungen	22
2. 16. 1	Pseudo-Ränge	23
2. 16. 2	Spatial Ranks - multivariate Ränge	23
2. 17	Voraussetzungen	24
2. 17. 1	Versuchspläne ohne Messwiederholungen	25
2. 17. 2	Versuchspläne mit Messwiederholungen	25
2. 18	Vergleiche	26
2. 18. 1	Versuchspläne ohne Messwiederholungen	26
2. 18. 2	Versuchspläne mit Messwiederholungen	27
2. 19	Entscheidungshilfen zur Auswahl	28
2. 19. 1	Warnungen	28
2. 19. 2	Versuchspläne ohne Messwiederholungen	29
2. 19. 3	Versuchspläne mit Messwiederholungen	29
3.	Funktionen zur Varianzanalyse in R und SPSS	31
3. 1	Funktionen in R	31
3. 2	Funktionen in SPSS	33
3. 3	Fehler bei der Rangberechnung	34
3. 4	Fehlende Werte	34
3. 5	Beschränkungen	35
4.	Unabhängige Stichproben	37
4. 1	Voraussetzungen der parametrischen Varianzanalyse	38
4. 2	Die 1-faktorielle Varianzanalyse	43
4. 2. 1	Kruskal-Wallis-Test	43
4. 2. 2	Varianzanalysen für inhomogene Varianzen	44
4. 2. 3	Verfahren für nichtnormalverteilte Variablen	46
4. 2. 4	Weitere Verfahren	46
4. 3	Die 2-faktorielle Varianzanalyse	47
4. 3. 1	Anmerkungen zur 2-faktoriellen Varianzanalyse	47
4. 3. 1. 1	Balancierte und nichtbalancierte Versuchspläne	47
4. 3. 1. 2	Die Interaktion	47
4. 3. 1. 3	Reduzierung des statistischen Fehlers	49
4. 3. 1. 4	Interpretation der Ergebnisse	49
4. 3. 2	Das parametrische Verfahren und Prüfung der Voraussetzungen	50
4. 3. 3	Varianzanalysen für inhomogene Varianzen	55
4. 3. 3. 1	Verfahren von Box, Brown & Forsythe sowie Welch & James	57
4. 3. 3. 2	BDM-Test	58
4. 3. 3. 3	Variablentransformationen	58
4. 3. 4	Rank transform-Tests (RT)	59
4. 3. 5	Puri & Sen-Test (Verallgemeinerte Kruskal-Wallis-Analysen)	60
4. 3. 6	Aligned rank transform (ART und ART+INT)	64
4. 3. 7	normal scores- (INT-) und van der Waerden-Tests	68
4. 3. 8	ATS-Tests von Akritas, Arnold & Brunner	71
4. 3. 9	Bredenkamp Tests	72

4.4	Nichtparametrische Verfahren zur mehrfaktoriellen Varianzanalyse	73
5.	Abhängige Stichproben - Messwiederholungen	74
5.1	Datenstruktur	75
5.1.1	Besonderheiten bei R und SPSS	75
5.1.2	Umstrukturierungen in R	77
5.2	Voraussetzungen der parametrischen Varianzanalyse	80
5.3	Die 1-faktorielle Varianzanalyse	84
5.3.1	Parametrischer Test und Prüfung der Voraussetzung	84
5.3.2	Friedman-Test	89
5.3.3	rank transform (RT) und normal scores (INT)	91
5.3.4	Puri & Sen-Test	93
5.3.5	van der Waerden	95
5.3.6	ATS-Tests von Akritas, Arnold & Brunner	97
5.3.7	Quade-Test	98
5.3.8	Skillings-Mack-Test	98
5.3.9	Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Agresti-Pendergast	99
5.3.10	Multivariate Tests: Spatial Signs und Spatial Ranks Methode	102
5.4	Die 2-faktorielle Varianzanalyse	104
5.4.1	Das parametrische Verfahren und Prüfung der Voraussetzungen	104
5.4.2	Rank transform-Tests (RT) und normal scores -Tests (INT)	108
5.4.3	Puri & Sen-Test	112
5.4.4	Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF) und van der Waerden-Test	115
5.4.5	Aligned rank transform (ART und ART+INT)	118
5.4.6	ATS-Tests von Akritas, Arnold & Brunner	123
5.4.7	Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Akritas & Arnold	125
5.4.8	Bredenkamp Tests	126
6.	Gemischte Versuchspläne	127
6.1	Voraussetzungen der parametrischen Varianzanalyse	127
6.2	Parametrische Varianzanalyse und Prüfung der Voraussetzungen	132
6.3	Rank transform-Tests (RT) und normal scores-Tests (INT)	140
6.4	Puri & Sen-Test	144
6.4.1	Ein Gruppierungs- und ein Messwiederholungsfaktor	146
6.4.2	Ein Gruppierungs- und zwei Messwiederholungsfaktoren	147
6.5	Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF)	149
6.5.1	Ein Gruppierungs- und ein Messwiederholungsfaktor	149
6.5.2	Ein Gruppierungs- und zwei Messwiederholungsfaktoren	153
6.5.3	Zwei Gruppierungs- und ein Messwiederholungsfaktoren	155
6.6	van der Waerden-Tests	156
6.6.1	Ein Gruppierungs- und ein Messwiederholungsfaktor	156
6.6.2	Zwei Gruppierungs- und ein Messwiederholungsfaktor	159
6.6.3	Ein Gruppierungs- und zwei Messwiederholungsfaktoren	162
6.7	Aligned rank transform (ART und ART+INT)	164
6.7.1	Ein Gruppierungs- und ein Messwiederholungsfaktor	165
6.7.2	Ein Gruppierungs- und zwei Messwiederholungsfaktoren	168
6.7.3	Zwei Gruppierungs- und ein Messwiederholungsfaktor	171
6.8	ATS-Tests von Akritas, Arnold & Brunner	177

6. 9	Bredenkamp Tests	179
6. 9. 1	Ein Gruppierungs- und ein Messwiederholungsfaktor	179
6. 9. 2	Zwei Gruppierungs- und ein Messwiederholungsfaktor	181
6. 10	Verfahren ohne Sphäritäts-Voraussetzungen	183
6. 10. 1	Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und nichtparametrisch	183
6. 10. 2	Multivariate Analysen: Spatial Signs und Spatial Ranks Methode	186
6. 10. 3	Welch & James	188
6. 10. 4	Koch	190
6. 10. 5	GEE	190
6. 10. 6	GLMM	193
6. 10. 7	GA- und IGA-Approximationen von Huynh	196
6. 10. 8	modifizierter Brown-Forsythe-Test	196
7.	Analysen für dichotome Merkmale	198
7. 1	Anwendung der Verfahren für metrische Merkmale	199
7. 1. 1	Unabhängige Stichproben	200
7. 1. 2	Gemischte Versuchspläne	201
7. 2	Anwendung der Verfahren für ordinale Merkmale	203
8.	Logistische Regression	204
8. 1	dichotome abhängige Variablen	204
8. 2	ordinale abhängige Variablen	207
8. 3	dichotome abhängige Variablen und Messwiederholungen	212
8. 4	ordinale abhängige Variablen und Messwiederholungen	216
9.	Mittelwertvergleiche, Kontraste und Kodierungen	218
9. 1	Grundlagen	218
9. 2	Standard-Kontraste	220
9. 3	Auswahl der Kontraste	222
9. 4	nichtparametrische Kontraste für die RT-, ART- und Puri & Sen-Verfahren	223
9. 5	universelles Verfahren für Kontraste	227
9. 6	Kontraste bei logistischen Regressionen	28
9. 7	Kontraste für Messwiederholungen und Interaktionen	228
9. 8	Zusammenfassen von Kontrasten	232
10.	Simple effects - einfache Effekte	234
10. 1	Unabhängige Stichproben	234
10. 2	Gemischte Versuchspläne	237
11.	Beispiele mit problematischen Datensätzen	241
11. 1	Extrem heterogene Varianzen	241
11. 2	lognormal verteilte abhängige Variable	243
11. 3	negative pairing	245
11. 4	Gemischter Versuchsplan mit Varianzheterogenitäten	248
11. 5	Gemischter Versuchsplan: Prüfung der Voraussetzungen	251



Anhang		258
1.	Umstrukturieren von Messwiederholungen in SPSS	258
1. 1	Umstrukturieren von Messwiederholungen in Fälle	258
1. 1. 1	ein Faktor und eine Analyse-Variable	258
1. 1. 2	mehrere Faktoren und eine Analyse-Variablen	261
1. 1. 3	ein Faktor und mehrere Analyse-Variablen	264
1. 2	Umstrukturieren von Fälle in Messwiederholungen	268
2.	Spezielle robuste F-Tests und andere Statistiken	271
2. 1	Box-Korrektur für heterogene Varianzen	271
2. 2	Brown-Forsythe F-Test für inhomogene Varianzen	271
2. 3	Box-Andersen F-Test für nichtnormalverteilte Variablen	272
2. 4	Box-Cox-Transformationen	272
2. 5	Fishers combined probability test	272
2. 6	Levene-Test auf Gleichheit von Kovarianzmatrizen	273
2. 7	Wilcox-Test auf Gleichheit von Varianzen bei Messwiederholungen	273
3.	R-Funktionen	274
3. 1	box.f: Box-F-Test für inhomogene Varianzen	274
3. 2	bf.f: Brown & Forsythe-F-Test für inhomogene Varianzen	274
3. 3	mbf.f: modified Brown & Forsythe-F-test für inhomogene Varianzen in gemischten Versuchsplänen (split-plot)	274
3. 4	box.andersen.f: F-Test für nichtnormalverteilte Variablen	275
3. 5	check.covar: Test auf Homogenität von Kovarianzmatrizen	275
3. 6	check.corr: Test auf Homogenität von Korrelationsmatrizen	276
3. 7	check.sphere: Test auf Spherizität	276
3. 8	ats.2 und ats.3: 2- bzw. 3-faktorielle Varianzanalyse	277
3. 9	np.anova: nichtparametrische Varianzanalyse mittels des KWF-Verfahrens und der von Puri & Sen und van der Waerden	277
3. 10	art1.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Gruppierungsfaktoren)	278
3. 11	art2.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Messwiederholungsfaktoren)	278
3. 12	art3.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (für gemischte Versuchspläne)	279
3. 13	wj.anova: Welch-James-Varianzanalyse für heterogene Varianzen (nur Gruppierungsfaktoren)	279
3. 14	wj.spanova: Welch-James-Varianzanalyse für heterogene Varianzen (für gemischte Versuchspläne)	280
3. 15	koch.anova: nichtparametrische Varianzanalyse für gemischte Versuchspläne nach dem Verfahren von G.Koch	280
3. 16	iga und iga.anova: general approximation test (GA) und improved general approximation test (IGA) von H.Huynh	280
3. 17	ap.anova: nichtparametrische Varianzanalyse für Messwiederholungen und split-plot Versuchspläne von Agresti & Pendergast	281
3. 18	simple.effects: parametrische Analyse von simple effects	282
3. 19	gee.anova: Anova-like tests for GEE and GLMM models	282

**Datensätze**

Beispieldaten 1 (mydata1):	37
Beispieldaten 2 (mydata2):	37
Beispieldaten 3 (mydata3):	37
Beispieldaten 4 (winer518):	74
Beispieldaten 5 (mydata5):	74
Beispieldaten 6 (winer568):	75
Beispieldaten 7 (irish):	198
Beispieldaten 8 (koch):	198
Beispieldaten (industrial waste):	241
Beispieldaten (lognormal):	243
Beispieldaten 11:	245
Beispieldaten 12:	248
Beispieldaten 13:	251

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

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

# 1. Allgemeines zur nichtparametrischen Statistik

*Parametrischen* statistischen Verfahren ([http://de.wikipedia.org/wiki/Parametrische\\_Statistik](http://de.wikipedia.org/wiki/Parametrische_Statistik)) liegt in der Regel ein mathematisches Modell zugrunde, das auf einer Verteilungsannahme beruht, häufig der Normalverteilung. Dabei müssen nicht unbedingt die Merkmale selbst der Verteilung folgen, häufig sind es auch abgeleitete Größen wie z.B. die Residuen. Die im Modell angenommene Verteilung hat Parameter (z.B. Mittelwert  $\mu$  und Standardabweichung  $\sigma$  bei der Normalverteilung), über die sich dann die Parameter des Modells bestimmen lassen. Bei den *nichtparametrischen* Verfahren, auch *verteilungsfreie* Verfahren genannt, wird in der Regel keine solche Verteilung angenommen.

Parametrische Verfahren werden meistens angewandt, wenn die abhängige Variable metrisch ist und zusätzliche Verteilungsvoraussetzungen, wie Normalverteilung der Residuen, erfüllt sind. Häufig kommen zusätzliche Voraussetzungen hinzu, wie z.B. Homogenität der Varianzen oder Unabhängigkeit der Beobachtungen. So z.B. bei der Varianz- oder Regressionsanalyse. Ist eine der Voraussetzungen nicht erfüllt, versucht man, äquivalente nichtparametrische Verfahren anzuwenden, sofern vorhanden. Letztere haben gegenüber den parametrischen meistens eine geringere (asymptotische) Effizienz - mehr dazu im nächsten Kapitel, in der Regel zwischen 63.7% ( $2/\pi$ ), z.B. beim Vorzeichen- und Mediantest, und 95,5% ( $3/\pi$ ), so beim Mann-Whitney U- und Kruskal-Wallis H-Test, falls alle Voraussetzungen erfüllt sind. Die Effizienz nichtparametrischer Tests kann allerdings auch umgekehrt über 100% , sogar beliebig hoch, liegen, wenn die Verteilungsvoraussetzungen nicht erfüllt sind. D.h. je weniger die Voraussetzungen eines parametrischen Tests erfüllt sind, desto eher kann zu einem nichtparametrischen Test geraten werden.

Vielfach werden Vorbehalte gegen nichtparametrische Verfahren geltend gemacht, weil bei diesen nicht alle Informationen der Daten ausgeschöpft würden. Dieses mag zwar gelegentlich der Fall sein, z.B. beim Median-Test als nichtparametrische Varianzanalyse, gilt aber nicht allgemein und insbesondere nicht für die hier besprochenen Methoden. So hat z.B. Sawilowsky (1990) in seiner Zusammenstellung auch diesen allgemeinen Punkt betrachtet. Demnach schneiden die (hier aufgeführten) nichtparametrischen Verfahren fast genau so gut ab, wie die parametrische Varianzanalyse. Und insbesondere wenn die Voraussetzung der Normalverteilung nicht gegeben ist, sind die nichtparametrischen überlegen. Dennoch können auch diese in manchen Fällen, z.B. bei ungleichen Varianzen, ebenso schlecht, oder sogar noch schlechter abschneiden.

In Abhängigkeit vom Skalenniveau der abhängigen Variablen unterscheidet man die Verfahren. Vorab jedoch einige wichtige Begriffe, die für die Beurteilung von statistischen Tests von Bedeutung sind.

## 1. 1 Wichtige Begriffe

### 1. 1. 1 Fehler 1. und 2. Art

Wenn eine Hypothese  $H_0$  , z.B. gleiche Mittelwerte, vorliegt und diese mit einem Test überprüft werden soll, gibt man in der Regel eine Irrtumswahrscheinlichkeit  $\alpha$  vor. Dieses ist der *Fehler 1. Art*. Er bedeutet, dass z.B. bei einer Vorgabe  $\alpha=0,05$  in 5 von 100 Fällen  $H_0$  abgelehnt wird, obwohl  $H_0$  richtig ist. Dagegen bezeichnet man mit *Fehler 2. Art* die Wahrscheinlichkeit, dass  $H_0$  angenommen wird, obwohl  $H_0$  falsch ist. Diese Wahrscheinlichkeit wird mit  $\beta$  bezeichnet und  $1-\beta$  heißt die Teststärke oder Power.  $\beta$  ist zunächst unbekannt, kann aber für zahlreiche Tests bei Vorgabe einiger Daten, wie z.B.  $n$  oder der Effektgröße, errechnet werden.

### 1. 1. 2 Effizienz eines Tests

Die (*asymptotische*) *relative Effizienz* (ARE) eines nichtparametrischen Tests A in Bezug auf einen parametrischen Test B (zur Prüfung derselben Hypothese) ist definiert als (das Grenzwertverhältnis für große  $n$ )  $n_B/n_A$ , den Quotienten der erforderlichen Stichprobenumfänge ( $n_A$  für Test A und  $n_B$  für Test B) zur Erlangung desselben Wertes für  $\beta$ , bei einem beliebigen (aber festen)  $\alpha$  und unter der Annahme, dass die Voraussetzungen des parametrischen Tests erfüllt sind. (Dieser Grenzwert ist allerdings unabhängig von  $a$ .) D.h. eine Effizienz eines nichtparametrischen Tests A von 95% oder 67 % gegenüber einem parametrischen Test B bedeutet, dass z.B. bei gleichen Mittelwertunterschieden der nichtparametrische Test eine ca. 5%  $((100-95)/95)$  bzw. 50%  $((100-67)/67)$  größere Stichprobe erfordert, um dieselbe Signifikanz zu erreichen. Dies schließt nicht aus, dass ein nichtparametrischer Test eine höhere Effizienz als der entsprechende parametrische haben kann, wenn die Voraussetzungen für den parametrischen nicht erfüllt sind. So hat z.B. der Test von van der Waerden (vgl. Kapitel 2.6) für nichtnormalverteilte Variablen eine Effizienz größer als 1. Eine höhere Effizienz bedeutet immer auch eine größere Teststärke  $1-\beta$ .

Die Idee der asymptotischen relativen Effizienz ist folgende: Mit größer werdendem  $n$  wird auch der kleinste (Mittelwert-) Unterschied bei jedem Test einmal signifikant. Ein Test, der bis zu diesem Punkt ein kleineres  $n$  benötigt als ein anderer, kann als effizienter angesehen werden, da er mit einer kleineren Stichprobe auskommt.

### 1. 1. 3 konservative und liberale Tests

Ein Test reagiert *konservativ*, wenn die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art kleiner als das vorgegebene  $\alpha$  ist. D.h. wenn z.B. bei einem  $\alpha=0.05$  die Anzahl der irrtümlich abgelehnten Nullhypothesen unter 5% liegt. Entsprechend reagiert ein Test *liberal*, wenn die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art das vorgegebene  $\alpha$  überschreiten kann. D.h. wenn z.B. bei einem  $\alpha=0.05$  die Anzahl der irrtümlich abgelehnten Nullhypothesen nicht konsequent unter 5% liegt.

Ein Test A ist *konservativer* (*liberaler*) als ein Test B, wenn die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art für A kleiner (größer) als für B ist. So ist z.B. bei den multiplen Mittelwertvergleichen der Newman-Keuls-Test ein liberaler Test, und der Newman-Keuls-Test ist liberaler als der Tukey-Test. Umgekehrt ist der Tukey-Test konservativer als der Newman-Keuls-Test. Konservative Tests sind in der Regel schwächer als liberale Tests.

### 1. 1. 4 starke und schwache Tests

Ein Test A ist *stärker* (*schwächer*) als ein Test B, wenn bei gleichem  $\alpha$  und  $n$  die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art bei Test A größer (kleiner) ist als bei Test B. D.h. bei Test A ist es leichter (schwieriger), einen Unterschied nachzuweisen als bei Test B.

### 1. 1. 5 robuste Tests

Ein Test wird als *robust* bezeichnet, wenn auch bei (moderaten) Verletzungen der Voraussetzungen die Ergebnisse noch korrekt sind. Das beinhaltet zweierlei: Zum einen wird die Rate für den Fehler 1. Art  $\alpha$  eingehalten, d.h. bei z.B.  $\alpha=0.05$  sind auch nur 5 von 100 Ergebnissen zufällig signifikant. Zum anderen verändert sich die Wahrscheinlichkeit für einen Fehler 2. Art  $\beta$  nicht drastisch, d.h. auch bei verletzten Voraussetzungen kann man noch signifikante Resultate erhalten.

### 1. 1. 6 Pairing

Der Begriff des *Pairing* (*Paarung*) spielt in der Varianzanalyse eine bedeutende Rolle. Er bezeichnet im Falle mehrerer Gruppen die Relation zwischen ungleichen Stichprobenumfängen  $n_i$  und ein einem anderen Parameter, der zwischen den Gruppen variiert, meistens den Varianzen, aber auch Korrelationen. Haben die großen Stichproben auch die größeren Parameterwerte, z.B. Varianzen, spricht man von *positive pairing*, haben dagegen die großen Stichproben die kleineren Parameterwerte, z.B. Varianzen, so pricht man von *negative pairing*.

## 1. 2 Methoden für metrische Merkmale

Bei diesen werden die Werte der Variablen in Ränge umgerechnet (vgl. [http://de.wikipedia.org/wiki/Rang\\_\(Statistik\)](http://de.wikipedia.org/wiki/Rang_(Statistik))). Auf diese werden dann die klassischen parametrischen Verfahren angewandt. So ist z.B. der Spearman-Rangkorrelationskoeffizient nichts anderes als der Pearson-Produkt-Moment-Korrelationskoeffizient der Ränge. Lediglich die Signifikanztests sind dann nicht mehr korrekt. Die *korrekten Signifikanzen* errechnen sich mit Mitteln der Kombinatorik, allerdings nur für kleine  $n$  (etwa  $<20$ ) oder es werden *asymptotische Signifikanztests* angeboten, die nur für große  $n$  ( $n>20$ ) gültig sind. In SPSS wird beides angeboten. Es konnte allerdings gezeigt werden, dass die Anwendung der klassischen parametrischen Verfahren auf die rangtransformierten Daten (ohne Anpassung der Signifikanztests) zu i.a. gültigen Ergebnissen führt. Und dies sogar bei Verfahren, die sonst als sehr sensitiv bzgl. der Verletzungen von Voraussetzungen gelten, so z.B. multiple Mittelwertvergleiche und Diskriminanzanalyse, klassischen parametrischen Verfahren (vgl. dazu Conover & Iman, 1981.)

## 1. 3 Methoden für ordinale Merkmale

Die oben erwähnten Verfahren für metrische Verfahren setzen voraus, dass eine Variable keine gleichen Werte hat. Durch sog. *Bindungskorrekturen* werden diese Verfahren allerdings auch anwendbar für ordinale Variablen, bei denen typischerweise Werte mehrfach vorkommen und dieser Tatsache bei der Rangberechnung durch die sog. *Bindungen* Rechnung getragen wird. Inzwischen sind in allen diesen Verfahren Bindungskorrekturen eingebaut.

In den letzten Jahren sind auch zunehmend Modelle für ordinale Merkmale entwickelt worden, denen die *relativen Effekte* zugrunde liegen, u.a. von Akritas, Arnold und Brunner (2013). Die daraus resultierenden Verfahren haben eine vergleichsweise hohe Effizienz, z.B. im Gegensatz zum Median-Test, der auch ohne Rangtransformationen metrischer Variablen auskommt. Mehr dazu in Kapitel 2.8.

## 1. 4 Methoden für dichotome Merkmale

Dichotome Variablen könnte man einfach unter die nominalen Variablen subsummieren. Sie spielen aber eine Sonderrolle: Zum einen gestalten sich viele Formeln und mathematische Verfahren einfacher, wenn ein Merkmal nur zwei Ausprägungen hat. Zum anderen haben viele Simulationen gezeigt, dass man dichotome Variablen bei größeren Fallzahlen vielfach genauso handhaben kann wie metrische Variablen. So z.B. bei der Varianzanalyse. Hinzu kommt, dass man dichotome Variablen als Extremfall einer ordinalen Variablen betrachten kann und somit die dafür konzipierten Verfahren anwenden kann. Tatsächlich sind Verfahren für dichotome Variablen häufig identisch mit den äquivalenten für ordinale Variablen, z.B. der Phi-Koeffizient (Abhängigkeitsmaß) als Spezialfall des Spearman-Korrelationskoeffizienten oder Cochrans Q-Test als Spezialfall von Friedmans Varianzanalyse (vgl. Cochran, 1950 und Lunney, 1970).

## 1. 5 Methoden für nominale Merkmale

Hier sind die polychotomen Merkmale angesprochen, also solche mit drei oder mehr Ausprägungen. Für solche Variablen gibt es vergleichsweise wenig statistische Methoden. Hinzu kommt, dass diese nicht immer trivial anzuwenden und die Ergebnisse nicht immer leicht verständlich sind. Entsprechende Methoden werden hier nicht vorgestellt.

## 1. 6 Prüfung auf Normalverteilung

Die Normalverteilung spielt eine bedeutende Rolle bei der Entscheidung für oder gegen parametrische Verfahren. Insbesondere bei metrischen abhängigen Variablen wird i.a. eine Prüfung auf Normalverteilung vorgenommen, und zwar der Residuen  $e$ , die Bestandteil jedes varianzanalytischen Modells sind, z.B.

$$x_{ijm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijm} \quad (i=1,\dots,I, j=1,\dots,J \text{ und } m=1,\dots,n_{ij})$$

Im einfachen Fall der Analyse ohne Messwiederholungen ist die Normalverteilung der Residuen äquivalent mit der Normalverteilung der abhängigen Variablen in jeder Zelle, allerdings auf keinen Fall mit der Normalverteilung der abhängigen Variablen insgesamt. (Letzteres würde ja selten der Fall sein, da das untersuchte Merkmal für die einzelnen Zellen unterschiedliche Mittelwerte haben wird, die zu mehreren unterschiedlichen Gipfeln in der Gesamtverteilung führen würden.) Wollte man die abhängige Variable zellenweise auf Normalverteilung prüfen - wie es z.B. beim t-Test häufig gemacht wird - so müsste man eine Reihe von Prüfungen vornehmen, wo für jede von diesen nur ein geringes  $n$  zur Verfügung stünde, manchmal vielleicht weniger als 5. Damit lässt sich eine Normalverteilung weder beweisen noch widerlegen, egal mit welchem Verfahren. Das gleiche gilt natürlich auch, wenn man zellenweise die Residuen auf Normalverteilung überprüfen wollte.

Daher ist es erforderlich, alle Residuen  $e_{ijm}$  zusammen auf Normalverteilung zu überprüfen, denn dadurch kumulieren sich die  $n_{ij}$  zu einem brauchbaren  $n$ . Als Methoden gibt es sowohl Tests, u.a. der Shapiro-Wilk- oder der klassische Kolmogorov-Smirnov-Test, als auch Grafiken, u.a. Histogramme oder *normal probability Plots*.

Bei den Tests steckt man in einem Dilemma: Zum einen ist die Normalverteilungsvoraussetzung eher für kleinere Stichproben relevant als für größere, da bei großem  $n$  nach dem *zentralen Grenzwertsatz* ohnehin die Test-Statistiken die erforderlichen Verteilungsvoraussetzungen erfüllen. Zum anderen sprechen statistische Tests bei kleinem  $n$  nicht an, d.h. die Nullhypothese muss angenommen und eine Abweichung von der Normalverteilung kann nicht nachgewiesen werden.

Daher empfiehlt es sich, die Normalverteilung visuell über Grafiken zu überprüfen. *Normal probability Plots* sind insbesondere für Unerfahrene schwerer interpretierbar (siehe unten), so dass letztlich Histogramme das Verfahren der Wahl sind. Um nicht zu irreführenden Ergebnissen zu kommen, muss allerdings die Intervallzahl auf die Anzahl Beobachtungen  $n$  abgestimmt sein. Eine einfache aber dennoch sehr gute Faustregel ist

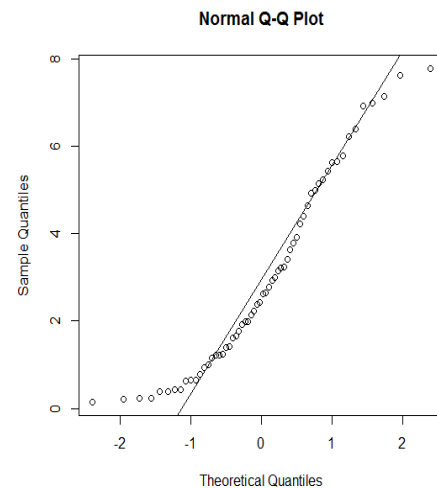
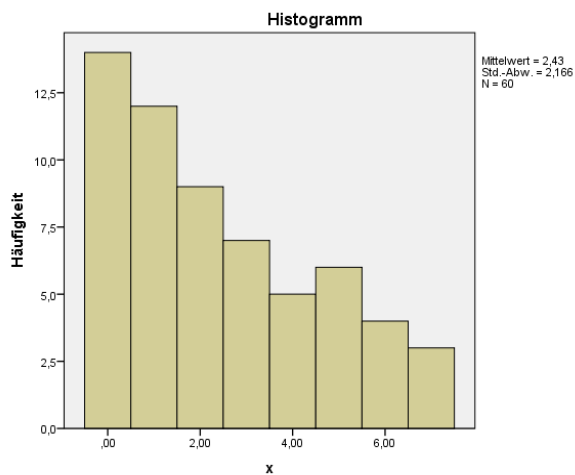
$$\text{Anzahl Intervalle} \sim \sqrt{n}$$

Aber auch dabei ist Vorsicht geboten, insbesondere wenn wie in SPSS gnadenlos die gewünschte Intervallzahl produziert wird: Bei diskreten (also nicht-stetigen) Merkmalen sollten alle Intervalle dieselbe Anzahl von Merkmalsausprägungen, also dieselbe Intervallbreite haben. Andernfalls zeigt das Histogramm ein verzerrtes Verteilungsbild. In R wird bei

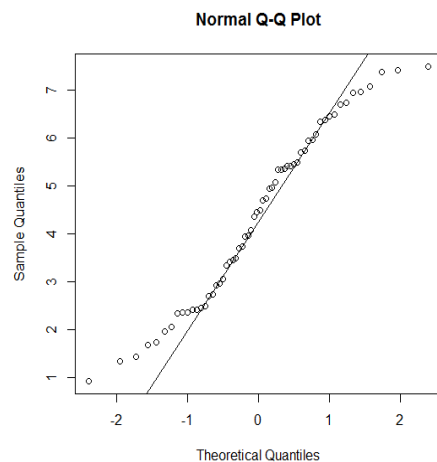
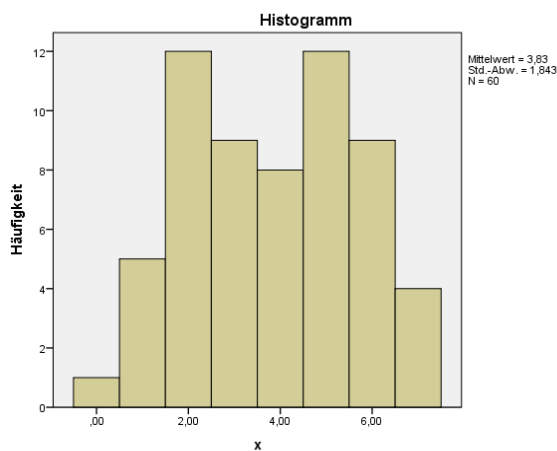
`hist(x,breaks=k,...)` diese Regel automatisch beachtet. In SPSS sollte die Intervallzahl anstatt über „Anzahl der Intervalle“ besser über die „Intervallbreite“ gesteuert werden.

Bei den *normal probability Plots*, oder allgemein bei den *Quantile-Quantile-Plots*, kurz *Q-Q-Plots* genannt (vgl. [http://en.wikipedia.org/wiki/Normal\\_probability\\_plot](http://en.wikipedia.org/wiki/Normal_probability_plot)), wird die empirische (kumulative) Verteilung mit der theoretischen, hier der Normalverteilung, verglichen. Üblicherweise ist die empirische Stichprobenverteilung  $y$  und die theoretische  $x$ . Leider ist das bei SPSS genau umgekehrt. Dabei wird zu jedem beobachteten Wert das Quantil  $y$  ermittelt und mit dem Quantil  $x$  der Vergleichsverteilung als Punkt eingezeichnet. Im Idealfall liegen also die Punkte auf einer Geraden. Im Gegensatz zu den Histogrammen sind diese Grafiken unabhängig von Intervalleinteilungen, die möglicherweise ein Bild „verzerren“ können.

Aber sowohl die Interpretation von Histogrammen auch der Q-Q-Plots bedarf ein wenig Erfahrung. Die wichtigsten Kennzeichen einer Normalverteilung sind Symmetrie und Eingipfligkeit. Nachfolgend werden einige typische Verteilungsformen aufgezeigt, die zum Teil nicht mehr als normal eingestuft werden können. Das Ergebnis des Shapiro-Wilk-Tests, alle basierend auf einem  $n=60$ , wird zur Verdeutlichung ebenfalls angeben:

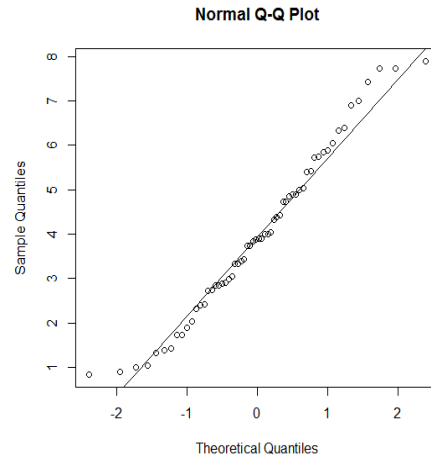
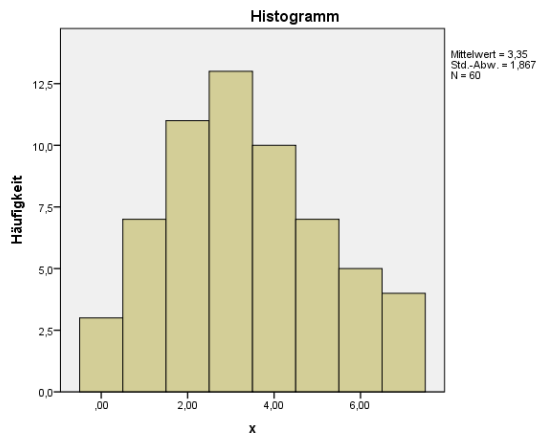


*stark rechtsschiefe Verteilung ( $W=0.894 - p=0.001$ )*

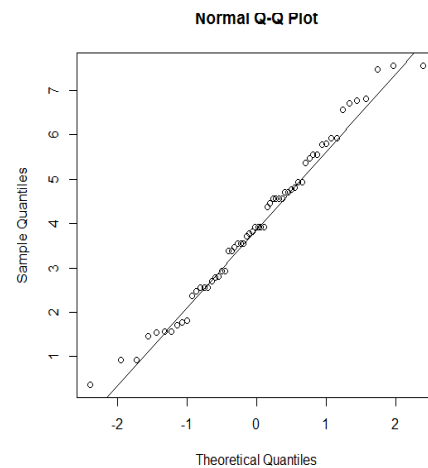
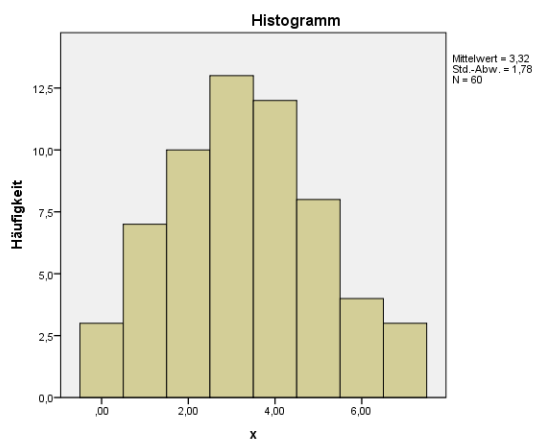


*zweigipflige Verteilung ( $W=0.944 - p=0.008$ )*

Während die beiden ersten Beispiele eher krasse Fälle von nichtnormalverteilten Werten darstellen, wird manch einem kaum ein Unterschied zwischen den letzten beiden Histogrammen auffallen, die immerhin unterschiedliche Resultate aufweisen. Das rechte ist deutlich symmetrischer und daher eher als normalverteilt zu akzeptieren.



*leicht rechtsschiefe Verteilung ( $W=0.955 - p=0.027$ )*



*(fast) normale Verteilung ( $W=0.962 - p=0.056$ )*

Die parametrischen multivariaten Tests, die auch für univariate Analysen mit Messwiederholungen eingesetzt werden, verlangen eine *multivariate* Normalverteilung der Residuen. In R werden dazu mehrere Pakete angeboten. So bietet z.B. das Paket `mvnNormalTest` die Tests von Fattorini, Mardia, Henze-Zirkler, Bowman & Shenton, Shapiro-Wilk sowie Zhou-Shao, das Paket `mvn` die Tests von Mardia, Henze-Zirkler, Royston sowie Doornik-Hansen. Beispiele dazu sind in Kapitel 5.3.9 und 11.5 zu finden.

## 1. 7 Prüfung von Voraussetzungen

Eine Warnung soll dieses Kapitel beenden. Am Thema „Prüfung von Voraussetzungen“ scheidet sich nämlich die Gemüter. Es wird nicht uneingeschränkt empfohlen, generell alle Voraussetzungen der parametrischen Anova zu prüfen. Der Grund: Zum einen sind die Prüfverfahren selbst unzuverlässig, d.h. sie können sowohl eine Abweichung von einer Voraussetzung anzeigen, obwohl diese gar nicht gegeben ist, als auch umgekehrt. Zum anderen haben diese Prüfverfahren wiederum Voraussetzungen, die nicht selten schärfer sind



als die des eigentlichen Verfahrens, also hier der Varianzanalyse. Dagegen kann man sich, zumindest in beschränktem Maße, auf die Robustheit der Varianzanalyse verlassen. Vor diesem Hintergrund hatte Box (1953) den inzwischen vielfach zitierten Satz geschrieben:

*To make a preliminary test on variances is rather like putting to sea in a row boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!*

Diese Problematik wird z.B. von Erceg-Hurn & Mirosevich (2008) behandelt, wo auch einige Beispiele dazu zu finden sind.

## 2. Nichtparametrische Varianzanalysen - Übersicht der Methoden

Nichtparametrische Varianzanalysen werden in der Regel angewandt, wenn die Voraussetzungen für die parametrische Analyse nicht gegeben sind, d.h. wenn die abhängige Variable entweder metrisch ist aber die Voraussetzungen „Normalverteilung der Residuen“ sowie „Varianzhomogenität“ nicht ausreichend erfüllt sind, oder aber wenn die abhängige Variable ordinales oder dichotomes Skalenniveau hat. Allerdings kann die Varianzanalyse als robustes Verfahren i.a. einige Abweichungen von den idealen Voraussetzungen vertragen. (Mehr dazu in den Kapiteln 4.1, 5.2. und 6.1.) Darüber hinaus gibt es auch *semiparametrische* Verfahren, eine Mischform aus parametrischem und nichtparametrischem Modell, z.B. wenn an die Verteilung der abhängigen Variablen keine Bedingungen gestellt werden, aber eine Form der Varianzhomogenität vorausgesetzt wird. Während beim parametrischen Modell die abhängige Variable genau ein Verteilungsmodell annimmt, können beim nichtparametrischen Ansatz quasi beliebige Verteilungsformen auftreten. Und so ist es nicht verwunderlich, dass man praktisch für jedes Verfahren eine Verteilungsform für die abhängige Variable finden kann, so dass die Ergebnisse unbefriedigend sind: von der Verletzung des  $\alpha$ -Risikos bis zu übermäßig konservativen Tests. Dies haben zahlreiche Simulationen gezeigt. Zu bedenken ist, dass die nichtparametrischen Verfahren in der Regel asymptotische Tests verwenden, also etwa für  $n > 20$  (mit  $n$  Gesamtzahl der Beobachtungen). Es gibt zwar für einige Verfahren sog. *exakte Tests* für kleine  $n$ , die aber hier nicht berücksichtigt werden.

Andererseits sind viele geneigt, „voreilig“ eine nichtparametrische anstatt der klassischen Varianzanalyse durchzuführen, z. B. weil das Skalenniveau der abhängigen Variablen ordinal ist oder die Varianzen der einzelnen Zellen möglicherweise ungleich sind. Hiervor muss eindringlich gewarnt werden. So schrieb z.B. Zimmerman (1998) „*It came to be widely believed that nonparametric methods always protect the desired significance level of statistical tests, even under extreme violation of those assumptions*“. So es gibt z.B. zahlreiche Studien, die belegen, dass nichtparametrische, insbesondere rangbasierte Verfahren nicht mit schiefen Verteilungen umgehen können, die auch nur leicht inhomogene Varianzen haben (vgl. z.B. G. Vallejo et al., 2010, Keselman et al., 1995 and Tomarken & Serlin, 1986). Dabei sind Varianzquotienten  $\max(\text{var})/\min(\text{var})$  von etwa 2 gemeint, was als normal anzusehen ist. Also:

*Nichtparametrische Verfahren sind kein Allheilmittel für den Fall, dass irgendwelche Voraussetzungen nicht erfüllt sind. Für diese Art von Varianzanalysen müssen ebenso wie bei der parametrischen Voraussetzungen beachtet werden.*

Neben den hier im Vordergrund stehenden „echten“ nichtparametrischen Verfahren darf nicht vergessen werden, dass es auch eine Reihe von *robusten Tests* gibt, z.B. für den Fall inhomogener Varianzen, die vorzugsweise dann angewandt werden können und sollten, wenn die abhängige Variable metrisch ist, aber keine Varianzhomogenität vorliegt. Die Methoden werden in späteren Kapiteln vorgestellt. Darüber hinaus gehören auch in diesen Kontext varianzanalytische Methoden für dichotome Merkmale, worauf später in Kapitel 7 kurz eingegangen wird.

Die wichtigsten Methoden werden im Folgenden kurz vorgestellt. Salazar-Alvarez et al. (2014) geben einen guten Überblick der nichtparametrischen Methoden zur mehrfaktoriellen Varianzanalyse. Eine leicht verständliche Einführung in diese Methoden bieten Erceg-Hurn & Mirosevich (2008). Seit 1990 sind eine Vielzahl von neuen Methoden zur nichtparametrischen Datenanalyse entwickelt worden, zum Teil mit neuen nichtparametrischen Modellen wie solchen mit „*relativen Effekten*“ (vgl. 2.9) oder „*spatial signs and ranks*“, räumlichen Vor-

zeichen und Rängen (vgl. 2.16). Hiervon können nur die „wichtigsten“ hier erwähnt werden. Dabei stehen solche im Vordergrund, die sich leicht mit Standardsoftware wie SPSS durchführen lassen. Trivialerweise lassen sich alle Verfahren in R (und natürlich S-Plus) realisieren.

Entscheidend für die Beurteilung eines Verfahrens ist das Verhalten hinsichtlich der Fehler 1. Art (*Irrtumswahrscheinlichkeit*  $\alpha$ ) und 2. Art ( $\beta$ , aber meistens über die *Power*  $1-\beta$  beurteilt). Dabei geht es um die Frage, in wie weit das vorgegebene  $\alpha$  eingehalten wird, bzw. in wie weit ein vorhandener Effekt nachgewiesen werden kann. Beide Fehler sind nicht unabhängig voneinander: Ein in einer bestimmten Situation, etwa bei inhomogenen Varianzen, liberaler Test wird auf der einen Seite das  $\alpha$ -Risiko verletzen, aber auf der anderen Seite in derselben Situation eine große Power zeigen. Umgekehrt wird ein konservativer Test meistens weniger irrtümlich falsche Signifikanzen ausweisen, dafür aber seltener einen tatsächlich vorhandenen Effekt nachweisen. Ein und derselbe Test kann in der einen Situation liberal, in einer anderen Situation konservativ reagieren.

Sofern nicht anders erläutert seien im Folgenden  $I$  die Anzahl der Gruppen, bzw.  $J$  die Anzahl der Messwiederholungen,  $N$ ,  $n_i$  bzw.  $n_{ij}$  die Anzahl der Merkmalsträger (Versuchspersonen) insgesamt, pro Gruppe bzw. Zelle, sowie  $x_{ijm}$  die beobachteten Werte mit  $m=1, \dots, n_i$ , und  $i=1, \dots, I$  sowie  $j=1, \dots, J$ .

## 2. 1 Kruskal-Wallis und Friedman (KWF)

Die klassischen nichtparametrischen Varianzanalysen sind die 1-faktoriellen Analysen mit den Tests von Kruskal & Wallis im Fall von unabhängigen Stichproben sowie dem von Friedman im Fall von abhängigen Stichproben (Messwiederholungen). Diese sind in (fast) allen gängigen Lehrbüchern ausführlich beschrieben. Beim Kruskal & Wallis-Test werden die  $x_{im}$  über alle Gruppen hinweg in Ränge  $R_m$  ( $m=1, \dots, n$ ), sog. *Wilcoxon-Ränge*, transformiert und daraus eine  $\chi^2$ -verteilte Testgröße errechnet, über die die Gleichheit der Mittelwerte geprüft wird. Beim Friedman-Test werden für jeden Merkmalsträger  $i$  die  $x_{im}$  in Ränge  $R_{jm}$  ( $j=1, \dots, J$ ), sog. *Friedman-Ränge*, transformiert und daraus eine  $\chi^2$ -verteilte Testgröße errechnet, über die die Gleichheit der Mittelwerte geprüft wird.

Die asymptotische Effizienz des Kruskal-Wallis-Tests (*K-W-Test*) liegt bei 0.955, die des Friedman-Tests bei  $0.955 \cdot J/(J+1)$ , also z.B. 0.64 (für  $J=2$ ) und 0.87 (für  $J=10$ ), wobei  $J$  die Anzahl der Gruppen (Versuchsbedingungen) ist. D.h. für große Stichproben ist der K-W-Test kaum schlechter als die parametrische Varianzanalyse.

Vielfach ist zu lesen, dass der Kruskal-Wallis-Test nicht nur auf Mittelwertunterschiede der zu vergleichenden Stichproben, sondern verschiedentlich auch auf Unterschiede der Streuung und Schiefe anspricht (vgl. Wilcox, 2003). Andere Autoren teilen dagegen nicht diese Bedenken (vgl. Marascuilo & McSweeney, 1977). Vargha & Delaney (1998) haben dieses Problem ausführlich untersucht und kommen zu dem Schluss, dass ein geringes Risiko besteht, dass der Test im Falle inhomogener Varianzen das  $\alpha$ -Risiko leicht verletzt, also auch darauf anspricht. Daher wird auch vielfach die gleiche Verteilungsform in allen Gruppen gefordert. Eine robuste Variante dieses Tests wurde von Brunner, Dette und Munk (vgl. Kapitel 2.13) entwickelt.

Der Friedman-Test hat im Vergleich zum K-W-Test eine geringe Effizienz. Iman und Davenport (1976) haben den  $\chi^2$ -Wert des Friedman-Tests in einen F-Wert transformiert:

$$F = \frac{(n-1)\chi^2}{n(J-1) - \chi^2} \quad (2-1)$$

mit  $J-1$  Zähler-FG und  $(J-1)(n-1)$  Nenner-FG, wobei  $n$  die Anzahl der Merkmalsträger ist. Er hat F-Wert deutlich bessere Eigenschaften und verleiht dem Friedman-Test eine etwas höhere Teststärke (siehe Iman und Davenport, 1976). Allerdings zeigen Harwell & Serlin (1984) das Gegenteil. Die Anwendung dieser Korrektur erübrigt sich selbstverständlich, wenn der  $\chi^2$ -Wert bereits als signifikant ausgewiesen worden ist. Wie auch die Beispiele in den Kapiteln 5 und 6 zeigen, sollte man von dieser Korrektur nicht zu viel erwarten. Eigene Simulationen haben gezeigt, dass lediglich für  $n \leq 10$  die Teststärke etwas besser ist.

Es sei noch erwähnt, dass es eine analoge Umrechnung des  $\chi^2$ -Werts des Kruskal-Wallis-Tests in einen F-Wert von Iman und Davenport gibt (vgl. Conover & Iman, 1981). Die ist dann allerdings mit dem F-Test des RT-Verfahrens (siehe nächstes Kapitel) identisch. Boos & Brownie (1995) haben beide Umrechnungen des  $\chi^2$ -Werts in einen F-Wert näher untersucht. Sie zitieren Studien, wonach die F-Werte für kleine und mittlere  $n$  vorteilhafter sind.

Lüpsen (2020b) hat ein Verfahren (*KWF*) entwickelt, das bei einer 1-faktoriellen Analyse sowohl den Kruskal-Wallis als auch den Friedman-Test als Spezialfall enthält und sich auf die „klassische“ Varianzanalyse mittels einer Variablentransformation zurückführen lässt. Damit sind diese Verfahren auch auf mehrfaktorielle Versuchspläne anwendbar. Der Algorithmus wird in Kapitel 6.4.2 erläutert.

## 2. 2 Rank transform Tests (RT)

Dies sind klassische Anova-F-Tests angewandt auf Rangdaten. D.h. alle Werte der abhängigen Variablen, über Gruppen und Messwiederholungen hinweg, werden in Ränge  $1, \dots, n \cdot I \cdot J$  umgerechnet, bevor dann eine parametrische Varianzanalyse mit F-Tests durchgeführt wird. Das Verfahren wurde 1981 von Conover & Iman (1981) vorgeschlagen und galt lange als eine brauchbare Lösung, bis in den 90er Jahren Simulationen einige Schwächen aufzeigten. So wird u.a. eine Verletzung des  $\alpha$ -Risikos für den Test der Interaktion berichtet, wenn zugleich signifikante Haupteffekte bestehen (vgl. u.a. Toothaker and De Newman, 1994). Der Grund dafür: die Additivität der Haupt- und Interaktionseffekte, d.h. die Unabhängigkeit der Tests, bleibt bei der Rangtransformation nicht erhalten (vgl. Beasley & Zumbo, 2009). Auf der anderen Seite konnten Hora und Iman (1984) sowohl theoretisch als auch durch Simulationen zeigen, dass zum einen die Tests der Haupteffekte von Gruppierungsfaktoren in jedem Fall asymptotisch, d.h. für größere  $n$ , valide sind, d.h. dass das Risiko für den Fehler 1. Art konsequent eingehalten wird, und zum anderen diese Tests stärker sind als die klassischen Tests von Kruskal-Wallis und Friedman oder auch als der von Quade. Analoge Resultate für die Messwiederholungsfaktoren zeigten Thompson & Ammann (1990).

Der Reiz dieser Methode liegt in der Einfachheit. Sie ist auch empfehlenswert, solange nicht eine Interaktion als signifikant ausgewiesen wird und zugleich mindestens ein Haupteffekt signifikant ist.

## 2. 3 Inverse normal transform (INT)

Eine Verbesserung der o.a. RT-Methode bringt die *inverse Normalverteilungs-Transformation* (*inverse normal transform*, INT). Bei dieser werden die oben erzeugten gleichverteilten RT-Werte  $R_i$  in (standard-) normalverteilte Scores umgerechnet:

$$\Phi^{-1}(R_i/(n+1)) \quad (2-2)$$

wobei  $\Phi$  die Standardnormalverteilung und  $n$  die Anzahl aller Werte insgesamt ist. (Diese Division durch  $n+1$  ist erforderlich, um den Wertebereich  $1 \dots n$  in das Intervall  $0 \dots 1$  zu transfor-

mieren.) Wie bei der o.a. RT-Methode werden dann für die transformierten Werte (*normal scores*) die klassischen F-Tests durchgeführt. Von dieser Transformation gibt es mehrere Varianten, die sich im Wesentlichen auf eine Formel zurückführen lassen:

$$\Phi^{-1}((R_i - c)/(n + 1 - 2c)) \quad (2 - 3)$$

Die o.a. zuerst aufgeführte, vielfach als *normal score test* bezeichnete Variante, erhält man z.B. über  $c=0$ . Huang (2007) hat mittels Simulationen gezeigt, dass bei Verwendung dieser Methode (im Gegensatz zur RT-Methode) das  $\alpha$ -Risiko auch für die Interaktionen nicht verletzt wird. Zu einem ähnlichen Ergebnis kommen Mansouri und Chang (1995). Unbestritten ist die vergleichsweise hohe Teststärke. Eine ausführliche Darstellung dieser Methoden ist bei Beasley, Erickson & Allison (2009) zu finden. Allerdings zeigen Letztere Beispiele auf, bei denen dennoch das  $\alpha$ -Risiko leicht verletzt wird.

Das INT-Verfahren geht u.a. auf van der Waerden in den 50er-Jahren zurück (vgl. Kapitel 2.7). Es ist zuletzt durch die Analyse von Gendaten wieder aktuell und beliebt geworden, da es auf der einen Seite ähnlich leicht wie das RT-Verfahren zu rechnen ist und auf der anderen Seite die falsch signifikanten Testergebnisse weitgehend vermeidet und zudem noch eine hohe Effizienz hat.

## 2. 4 Aligned rank transform (ART)

Eine andere Methode, die bei der o.a. RT-Methode möglichen fälschlich signifikanten Interaktionen zu vermeiden, wenn zugleich signifikante Haupteffekte vorliegen, bieten die *aligned rank transforms* oder auch *aligned rank tests* (ART). Das Verfahren ist anwendbar sowohl für Haupt- als auch für Interaktionseffekte. Es werden hierbei zunächst die Daten bzgl. der „störenden“ Effekte, z.B. der Haupteffekte im Fall der Analyse einer Interaktion, bereinigt, sog. *alignment*. Hierzu gibt es zwei Methoden, eine einfache und eine etwas aufwändigere, die jedoch zu demselben Ergebnis führen.

- Der *naive approach* (ART1): Zunächst werden von der Kriteriumsvariablen die „störenden“ Effekte subtrahiert, z.B. die Haupteffekte der Faktoren, die an der untersuchten Interaktion beteiligt sind. Für den Test der Interaktion wird also anstatt  $x$  die Variable  $x_{ijm} - \alpha_i - \beta_j$  untersucht, oder mit den Werten der Stichprobe:

$$x'_{ijm} = x_{ijm} - \bar{a}_i - \bar{b}_j + 2\bar{x} \quad (2 - 4)$$

wobei  $\bar{a}_i$ ,  $\bar{b}_j$ ,  $\bar{x}$  die Gruppenmittelwerte bzgl. der Faktoren A und B bzw. der Gesamtmittelwert sind.

- Der *standard approach* (ART2): Zunächst wird eine komplette Varianzanalyse der Kriteriumsvariablen (mit allen Effekten) durchgeführt. Zu den daraus resultierenden Residuen  $e_m$  wird der untersuchte Effekt addiert, z.B. der Interaktionseffekt, als Differenz von Zellen- und Gruppenmittelwerten. Für den Test der Interaktion wird also anstatt  $x$  die Variable

$$x'_{ijm} = e_m + (\bar{a}\bar{b}_{ij} - \bar{a}_i - \bar{b}_j + 2\bar{x}) \quad (2 - 5)$$

untersucht, wobei  $e_m$  die Residuen des kompletten varianzanalytischen Modells,  $\bar{a}_i$ ,  $\bar{b}_j$ ,  $\bar{a}\bar{b}_{ij}$ ,  $\bar{x}$  die Mittelwerte der Faktoren A und B bzw. der Gesamtmittelwert sind.

Die Ergebnisvariable wird anschließend in Ränge umgerechnet und dann wie bei dem RT-Verfahren weiter analysiert, um die Interaktion zu testen.

Im Fall von gemischten Versuchsplänen, sog. *split-plot designs*, ist das alignment wegen der zu berücksichtigenden Personeneffekte etwas komplizierter. Das Verfahren ist z.B. bei Lei et al. (2004) beschrieben, kann aber auch den Beispielen in Kapitel 6.7 entnommen werden.

Dieses Verfahren wird daher auch mit RAA (*ranking after alignment*) bezeichnet. Das Verfahren geht auf Hodges & Lehmann (1962) zurück und wurde von Higgins & Tashtoush (1994) populär gemacht. Neben den beiden o.a. Methoden gibt es inzwischen noch eine Vielzahl weiterer Varianten von ART. So wurden u.a. von Peterson (2002) Alignments (Korrekturen) mittels robuster Mittelwerte wie Median oder getrimmer Mittelwerte anstatt des arithmetischen Mittels vorgeschlagen. Diverse Untersuchungen zeigten jedoch, dass diese Varianten eher schlechtere als bessere Ergebnisse aufweisen (vgl. z.B. Toothaker & De Newman, 1994).

Für die Datentransformation wird ein spezielles Programm (*ARTool*) angeboten (vgl. Wobbrock, 2011), das Microsoft .NET 2.0 Framework voraussetzt. Die transformierten Daten können dann mit einem Standardprogramm wie SPSS analysiert werden.

Das ART-Verfahren kann aber auch mit ein wenig Aufwand ohne Zusatzsoftware in R oder SPSS angewandt werden, wie die Beispiele in den nachfolgenden Kapiteln demonstrieren. Für R gibt es auch das Paket `ARTOOL`, allerdings nicht für Designs mit Messwiederholungen. (Ein weiteres Paket, `ART`, ist wegen zum Teil falscher Resultate nicht zu empfehlen.) Im Wesentlichen müssen Aggregatdaten wie Mittelwerte ermittelt werden, die in die Berechnungen einfließen. Es sei ausdrücklich darauf hingewiesen, dass der Aufwand des ART- gegenüber dem RT-Verfahren nicht generell erforderlich ist, um falsch signifikante Ergebnisse zu vermeiden. Lediglich in dem Fall, dass eine Interaktion als signifikant ausgewiesen wird und zugleich mindestens ein Haupteffekt signifikant ist, sollte für die untersuchte Variable das ART-Verfahren angewandt werden. Dennoch werden bei den Beispielen in diesem Skript meistens auch Alignments für die Haupteffekte durchgeführt, allerdings nur zu Demonstrationszwecken.

Das ART-Verfahren kann aber nicht empfohlen werden, da es eine Reihe von Situationen gibt, in denen es das  $\alpha$ -Risiko krass verletzt, so u.a. in den Fällen

- heterogener Varianzen (s. z.B. Leys & Schumann, 2010, und Carletti & Clautriaux, 2005),
- stark schiefer Verteilungen wie der Exponential-Verteilung (vgl. Lüpsen, 2016b),
- diskreter abhängiger Variablen, insbesondere bei größeren  $n$  (vgl. Lüpsen, 2016a),
- von Tests der Haupteffekte bei größeren  $n$  (vgl. Lüpsen, 2016c).

Gerade der dritte Punkt ist gravierend, da somit die Anwendung bei ordinalen Variablen ausscheidet, insbesondere bei einer geringeren Anzahl von Ausprägungen, etwa  $<10$ . Bei zahlreichen Untersuchungen schneidet das ART-Verfahren relativ gut ab. Das liegt zum Teil aber daran, dass meistens die o.a. kritischen Punkte unberücksichtigt blieben.

## 2. 5 Kombination von Aligned rank transform und Inverse normal transform (INT+ART)

Mansouri & Chang (1995) schlugen eine Kombination aus den beiden vorigen Verfahren vor: Zuerst die Transformation der Werte nach dem ART-Verfahren, dann die Umrechnung der erhaltenen Ränge in normal scores nach dem INT-Verfahren. Hierbei ist es sinnvoll, alle Tests, also auch für die Haupteffekte, nach dieser Methode durchzuführen. So wie die Transformation in normal scores die teilweise zu hohe Fehlerrate 1. Art für die RT-Methode abmildert, so verkleinert auch hier die Transformation in normal scores die häufig zu hohen Fehlerraten der ART-Methode. Dies berichten u.a. Carletti & Clautriaux (2005) sowie Lüpsen (2016c). Die Anwendung der INT-Transformation führt übrigens auch zu einer deutlichen Vergrößerung der Power.

Daher gilt die Empfehlung: Wenn die ART-Technik angewandt werden soll, dann auf jeden Fall zusammen mit der INT-Transformation. Dennoch kann Letztere nicht im Fall diskreter

abhängiger Variablen helfen. Und bei stark schiefen Verteilungen bringt sie zwar deutliche Verbesserungen der Fehlerrate, aber leider bleiben Situationen, wo das  $\alpha$ -Risiko verletzt wird, nämlich beim Test von Haupteffekten im Fall von ungleichen Zellenbesetzungszahlen für  $n > 20$  (vgl. Lüpsen, 2016c).

## 2. 6 Puri & Sen-Tests

Bei den Puri & Sen-Tests werden ebenfalls alle Werte wie beim Kruskal & Wallis-Test oder beim o.a. RT-Verfahren zunächst in Ränge umgerechnet, bevor dann eine klassische Varianzanalyse durchgeführt wird. Allerdings wird dann anstatt des F-Tests ein  $\chi^2$ -Test durchgeführt, auch *L statistic* genannt. Bei Versuchsplänen ohne Messwiederholungen sind dies Verallgemeinerungen des Kruskal & Wallis-Tests. Die Testgröße errechnet sich im Fall von Versuchsplänen ohne Messwiederholungen als

$$\chi^2 = \frac{SS_{\text{Effekt}}}{MS_{\text{total}}} \quad (2 - 6a)$$

bzw. für Gruppierungsfaktoren im Fall von Versuchsplänen mit Messwiederholungen als

$$\chi^2 = \frac{SS_{\text{Effekt}}}{MS_{\text{zwischen}}} \quad (2 - 6b)$$

bzw. im Fall von Messwiederholungsfaktoren als

$$\chi^2 = \frac{SS_{\text{Effekt}}}{(SS_X + SS_{\text{Fehler}})/(df_X + df_{\text{Fehler}})} \quad (2 - 7)$$

wobei

- $SS_{\text{Effekt}}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,
- $MS_{\text{total}}$  die Gesamtvarianz (Mean Square),
- $MS_{\text{zwischen}}$  die Varianz (Mean Square) zwischen den Versuchspersonen,
- $SS_X$  die Summe der Streuungsquadratsummen (Sum of Squares) aller Haupt- und Interaktionseffekte, die denselben Fehlerterm haben wie der zu testende Effekt,
- $MS_{\text{innerhalb}}$  die Varianz (Mean Square) innerhalb der Versuchspersonen und
- $SS_{\text{Fehler}}$  die Streuungsquadratsumme des zum getesteten Effekt gehörenden Fehlers ist.

Alle  $SS$  und  $MS$  können aus den üblichen Anova-Tabelle abgelesen werden. Falls nur ein Messwiederholungsfaktor vorliegt, entspricht der Nenner in 2-7 der Varianz  $MS_{\text{innerhalb}}$ . Als Freiheitsgrade für den  $\chi^2$ -Test nimmt man die Zählerfreiheitsgrade aus der Varianzanalyse. Für die Haupteffekte ergibt dies die Tests der bekannten nichtparametrischen Anova von Kruskal-Wallis bzw. Friedman.

Diese Methode gilt als relativ konservativ, insbesondere für mehrfaktorielle Versuchspläne. Dies ist aus der o.a. Berechnung leicht zu erklären: Bei den F-Tests der parametrischen Varianzanalyse reduziert die Streuung der anderen Faktoren die Fehlervarianz und vergrößert somit den F-Wert (vgl. dazu Kapitel 4.3.1.3). Hier gilt dies nicht: Die Streuung der anderen Faktoren verkleinert nicht die Gesamtstreuung  $MS_{\text{total}}$  bzw.  $MS_{\text{innerhalb}}$ , die im Nenner steht. Somit hat dieser Test weniger Power als solche, die über den F-Test geprüft werden, und dies umso stärker wie andere Effekte signifikant sind. Auf der anderen Seite gibt es auch hier Situationen, in denen das  $\alpha$ -Risiko verletzt wird, obwohl die Methode als konservativ gilt, nämlich solche mit heterogenen Varianzen. Dafür kann dieses Verfahren aber bedenkenlos auf ordinale Merkmale angewandt werden. Positiv ist noch zu bemerken, dass im Fall von Messwiederholungen nicht die sonst kritische Sphärität der Kovarianzmatrizen bzw. deren Homogenität gefordert wird,

da hier  $\chi^2$ -Tests anstatt F-Tests verwendet werden.

Der Ansatz geht in die 60er Jahre zurück auf Bennett (1968), wurde von Scheirer, Ray & Hare (1976) sowie Shirley (1981) erweitert und schließlich von Puri & Sen (1985) systematisch dargestellt. Eine gut verständliche Beschreibung bieten Thomas et al. (1999). Das Verfahren ist in der Literatur auch als *L-Statistik* von Puri & Sen bekannt. Die 1-faktorielle Analyse ist auch bei Winer (1991) nachzulesen. Diese Tests werden im Folgenden mit Puri & Sen-Tests bezeichnet.

Die Umsetzung in R bzw. SPSS ist natürlich nicht ganz so simpel wie bei den RT- und INT-Verfahren. In der Regel genügt die Erzeugung einer neuen rangtransformierten Variablen. Allerdings müssen bei Messwiederholungen die Ränge nach dem Friedman-Verfahren, also fallweise, ermittelt werden, was insbesondere bei SPSS zusätzlichen Aufwand erfordert, nämlich die Umstrukturierung der Datenmatrix. Bei zwei oder mehr Messwiederholungsfaktoren wird der Aufwand allerdings größer. Hinzu kommt die Durchführung der  $\chi^2$ -Tests, die insbesondere in SPSS mit dem Taschenrechner erfolgen muss.

## 2. 7 van der Waerden

Die Methode von van der Waerden (vgl. Wikipedia) vereinigt gewissermaßen die INT-Methode sowie die Verfahren von Kruskal & Wallis und Friedman bzw. das o.a. Puri & Sen-Verfahren. Van der Waerden hat das Verfahren als 1-faktorielle Varianzanalyse für unabhängige Stichproben entwickelt: Zunächst werden wie bei der INT-Methode die normal scores errechnet. Auf diese wird dann der  $\chi^2$ -Test wie beim Kruskal-Wallis-Test angewandt, z.B. in der Berechnung wie bei (2-4). Mansouri und Chang (1995) haben das Verfahren auf mehrfaktorielle Versuchspläne verallgemeinert. Dieses funktioniert dann so, dass die Puri & Sen-Tests auf die normal scores anstatt der Ränge angewandt werden. Marascuilo und McSweeney (1977) haben analog einen Test für abhängige Stichproben konstruiert, allerdings nur für einen Messwiederholungsfaktor. Lüpsen (2020b) hat das Verfahren für gemischte Versuchspläne erweitert. Allgemein werden die Effektttests mittels  $\chi^2$ -Tests wie beim Puri & Sen-Verfahren errechnet, lediglich mit anders transformierten  $y$ -Werten.

Der Test hat bei 1-faktoriellen Varianzanalysen für unabhängige Stichproben eine asymptotische Effizienz von 1, ist also der parametrischen Varianzanalyse ebenbürtig, und ist im Fall, dass die Voraussetzungen der klassischen Varianzanalyse nicht erfüllt sind, ihr deutlich überlegen (vgl. Sheskin (2004), der sich auf Conover sowie Marascuilo & McSweeney (1977) bezieht). Bedingt durch das rechnerische Vorgehen leidet zunächst einmal das Verfahren an demselben konservativen Verhalten wie die Puri & Sen-Methode. Allerdings ist es auch wieder die Anwendung der INT-Transformation, die zum einen die erhöhten Fehlerraten bei heterogenen Varianzen abschwächt und zum anderen dem Test eine deutlich höhere Teststärke verleiht, insbesondere bei nicht allzu kleinen  $n > 10$ . Positiv ist noch zu bemerken, dass im Fall von Messwiederholungen nicht die sonst kritische Sphärizität der Kovarianzmatrizen bzw. deren Homogenität gefordert wird, da hier  $\chi^2$ -Tests anstatt F-Tests verwendet werden. Dadurch ist der v.d.Waerden-Test das empfehlenswerteste nichtparametrische Verfahren, zumindest bei unabhängigen Stichproben, während bei gemischten Versuchsplänen die Power geringer eher durchschnittlich ausfällt, wie Lüpsen (2020b) gezeigt hat.

Der Aufwand ist exakt derselbe wie bei den im vorigen Kapitel skizzierten verallgemeinerten Kruskal-Wallis- und Friedman-Analysen mit Puri & Sen-Tests.



## 2. 8 Bredenkamp Tests - bifaktorieller H-Test

Schon 1974 entwickelte Bredenkamp (1974) eine Verallgemeinerung der Tests von Kruskal-Wallis und Friedman auf 2-faktorielle Analysen. Die Idee dazu stammte von Lemmer & Stoker (1967) und wird mit *bifactorial H-test* bezeichnet. Und zwar wird unter Ausnutzung der Additivität der  $\chi^2$ -Werte die Interaktion auf mehrere 1-faktorielle Analysen zurückgeführt. Dazu wird ein einfaktorieller Test über alle Zellen hinweg durchgeführt und anschließend vom resultierenden  $\chi^2$ -Wert die  $\chi^2$ -Werte der Haupteffekte subtrahiert. Die Methode lässt sich auch auf mehr als zwei Faktoren erweitern und ohne Aufwand mit Standardsoftware durchführen. Diese Methode ist mit dem Puri & Sen-Verfahren identisch, ist allerdings zum einen nur für balancierte Versuchspläne gültig, und zum anderen bei mehrfacher Messwiederholung falsch. Die Tests sind auch ausführlich beschrieben bei Lienert (1981, S. 1024 ff).

## 2. 9 Akritas, Arnold & Brunner ATS Tests

Akritas, Arnold und Brunner stellen ein anderes Modell mit beliebigen Verteilungen vor, das nicht einfach auf einer Umrechnung der Werte in Ränge basiert (vgl. Akritas, Arnold & Brunner, 1997), gut verständlich dargestellt von Brunner & Munzel (2013).

Ein Begriff, der bei diesem Verfahren eine wichtige Rolle spielt, ist der *relative Effekt*. Er dient zur Unterscheidung zwischen zwei Verteilungen, etwa den Zufallsvariablen  $X_1$  und  $X_2$ . Der relative Effekt von  $X_1$  zu  $X_2$  ist definiert als die Wahrscheinlichkeit, dass  $X_1$  kleinere Werte annimmt als  $X_2$ , d.h.  $p^+ = P(X_1 < X_2) + P(X_1 = X_2)/2$ , unter Berücksichtigung des Falles  $P(X_1 = X_2) > 0$ . Dabei hat  $X_2$  eine stochastische Tendenz zu größeren Werten als  $X_1$ , falls  $p^+ < 1/2$  und eine stochastische Tendenz zu kleineren Werten, falls  $p^+ > 1/2$  ist. Detaillierte Ausführungen hierzu sind bei E. Brunner & U. Munzel (2002) zu finden.

Trotz des anderen Ansatzes mit beliebigen Verteilungen resultieren dann doch im Wesentlichen ähnliche F-Quotienten wie bei Rank transform Tests. Allerdings werden sehr viel differenziertere Freiheitsgrade verwendet. Wegen der Ähnlichkeit zu den F-Tests der Anova werden sie ATS (*Anova type statistic*) genannt. Parallel zu den ATS bieten die Autoren auch eine weitere  $\chi^2$ -verteilte Statistik WTS (*Wald type statistic*) an, die aber hier nicht berücksichtigt wird, da die ATS bessere Eigenschaften aufweist. Letztlich werden dabei die relativen (Behandlungs) Effekte  $p_i$ , anstatt Mittelwerte, verglichen, mit

$$p_i = (\bar{R}_i - 0,5)/n \quad (\text{mit } \bar{R}_i = \text{mittlerer Rang und } n = \sum n_i)$$

Dieser Ansatz wird von Munzel & Brunner (2000) auf multivariate Analysen, von Brunner, Munzel & Puri (1999) auf Analysen mit Messwiederholungen sowie von Akritas & Brunner (2003) auf Kovarianzanalysen erweitert. Bei letzteren sind sogar fehlende Werte erlaubt und es gibt Lösungen sowohl für den Fall homogener Varianzen-Kovarianzen (*compound symmetry*) als auch für den allgemeinen Fall. Diese Tests sind ausdrücklich auch für ordinale und dichotome abhängige Variablen anwendbar. Es sei darauf aufmerksam gemacht, dass es zwei Varianten des ATS gibt: eine semiparametrische (vgl. Formel 5 in Brunner et al, 1997) und eine nichtparametrische (vgl. Formel 14 in Brunner et al, 1997). Darüber hinaus gibt es noch eine Variante von Friedrich et al. (2017).

Die Autoren attestieren ihnen eine vergleichsweise hohe Effizienz sowie die exakte Einhaltung des  $\alpha$ -Niveaus. Negativ wird vermerkt, dass die Tests nicht nur auf Mittelwertunterschiede, sondern auch auf andere Verteilungsunterschiede, insbesondere Streuungsunterschiede ansprechen und somit doch nicht konsequent den Fehler 1. Art unter Kontrolle hält. Richter &

Payton (2003) kommen bei einem Vergleich mit dem F-Test zu dem Ergebnis, dass die ATS sehr konservativ reagiert. Allerdings schnitt die ATS-Methode bei einem Vergleich mit den anderen hier vorgestellten Verfahren vergleichsweise schlecht ab (vgl. Lüpsen, 2016c). Zum einen hat es dieselben Schwächen bei ungleichen Varianzen wie das RT-Verfahren, was die Aussage im vorigen Satz bestätigt, zum anderen hat es in den meisten Situationen die geringste Power der hier besprochenen Methoden. Lediglich in einem Fall ist die ATS-Methode unschlagbar: Bei Versuchsplänen mit ungleichen  $n_i$  und ungleichen Varianzen  $s_i^2$ , wenn kleine  $n_i$  mit großen  $s_i^2$  gepaart sind.

Für die ATS- und WTS-Verfahren gibt es R-Pakete: `GFD` und `rankFD` (semiparametrisch) bzw. `BDM` (nichtparametrisch) für unabhängige Stichproben sowie `nparLD` und `MANOVA.RM` für Messwiederholungen. In SPSS sind diese Tests wegen der umfangreichen Matrizenrechnungen nicht durchführbar.

## 2. 10 Weitere Varianzanalysen für unabhängige Stichproben

An dieser Stelle werden noch zwei Tests erwähnt, für die entsprechende Funktionen zur Anwendung in R über Cran bereitgestellt werden. Da beide jedoch außerordentlich liberal reagieren (vgl. Lüpsen, 2016c), werden sie hier nicht näher vorgestellt. Und von einer Benutzung wird abgeraten.

### 2. 10. 1 Wilcoxon analysis (WA)

Hettmansperger and McKean (2011) haben eine nichtparametrische Regression, *Wilcoxon Analysis* (WA), entwickelt, bei der die Ränge der Residuen die zentrale Rolle spielen und somit der Einfluss von Ausreißern reduziert wird. Trivialerweise lässt sich der Ansatz auf die Varianzanalyse anwenden. Eine Erweiterung dieser Methode ist die *weighted Wilcoxon technique* (WW), bei der auch die  $x$ -Variablen in Ränge transformiert werden. Dieses Verfahren zählt zu den semiparametrischen, da es auf den Parametern der linearen Regression basiert.

Es gibt das R-Paket `rfit` zur Anwendung dieser Methode in R (vgl. Kloke & McKean, 2012). In einem Vergleich von Lüpsen (2016) zeigte sich allerdings, dass das  $\alpha$ -Risiko selbst bei einem Modell ohne Effekte krass überschritten wird. Diese Methode wird daher hier nicht behandelt.

### 2. 10. 2 Gao & Alvo

Gao & Alvo (2005) haben einen Test für die Interaktion in 2-faktoriellen Versuchsplänen (ohne Messwiederholungen) entwickelt. Es wird ihm zwar eine hohe Power attestiert, allerdings zu Lasten der Kontrolle des Fehlers 1. Art. Der Test steht in der Funktion `interaction.test` aus dem Paket `StatMethRank` zur Verfügung.

## 2. 11 Weitere Varianzanalysen für abhängige Stichproben

### 2. 11. 1 Quade

Der Test von *Quade* (vgl. Wilcox et al., 2013) ist ein globaler Test auf Gleichheit der Mittelwerte bei Messwiederholungen, ähnlich dem Friedman-Test. Er liegt bislang nur als 1-faktorielle Analyse vor.

Die Idee ist folgende: Bei der Rangbildung  $R_{ij}$  für die Friedman-Analyse, bei der pro Fall/Merkmalsträger  $m$  ( $m=1, \dots, n$ ) die Werte  $j=1, \dots, J$  vergeben werden, ist nur eine geringe Differenzierung zwischen den  $J$  Gruppen (Messwiederholungen) möglich. Daher wird eine

Fallgewichtung  $Q_m$  eingeführt, die Fälle mit einem größeren Wertespektrum bevorzugt.  $Q_m$  errechnet sich aus der Spannweite  $D_m$  der Werte eines Falls (Differenz von Maximum und Minimum der  $x_{mj}$ ), die dann in Ränge umgerechnet wird. Aus beiden Rängen  $R_{mj}$  und  $Q_m$  zusammen wird dann das Produkt  $W_{mj} = Q_m * R_{mj}$  errechnet. Zum Vergleich zweier Gruppen werden schließlich die Rangsummen von  $W_{mm}$  verwendet:

$$T_j = \left( \sum_{m=1}^n W_{mj} \right) / (n(n+1)/2)$$

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

Der Quade-Test hat für  $J < 6$  eine größere Teststärke als der Friedman-Test und ist daher diesem überlegen (vgl. u.a. Wikipedia). Auf der anderen Seite wird er nicht für ordinal-skalierte Variablen empfohlen. Dieser Test ist in R als `quade.test` sowie im Paket `PMCMRplus` verfügbar.

### 2. 11. 2 Skillings & Mack

Der Test von Skillings & Mack ist ebenfalls eine Alternative zum Friedman-Test, also für abhängige Stichproben (Messwiederholungen), allerdings für den Fall von fehlenden Werten. Er ist anschaulich beschrieben von Chatfield und Mander (2009). Auch dieses Verfahren liegt bislang nur als 1-faktorielle Analyse vor.

Liegen weder fehlende Werte noch Bindungen vor, so liefern die Tests von Skillings & Mack und von Friedman dieselben Resultate. Im Fall von vielen Bindungen und/oder kleinen Fallzahlen ist dieser Test dem von Friedman leicht überlegen.

Dieser Test ist als Funktion `Skimack` im Paket `skillings.Mack` sowie als Funktion `skillingsMackTest` im Paket `PMCMRplus` verfügbar. An dieser Stelle sei darauf hingewiesen, dass das in Kapitel 2.8 erwähnte Verfahren von Akritas, Arnold und Brunner in der Version des R-Pakets `nparLD` auch fehlende Werte zulässt.

### 2. 11. 3 Spatial Signs und Spatial Ranks Methoden

In den letzten Jahren wurden nichtparametrische Verfahren für multivariate Daten entwickelt, die auf *Spatial Signs* und *Spatial Ranks* sowie *Spatial Symmetrized Signs* und *Spatial Sign Ranks* basieren. Dies sind zunächst einmal Verallgemeinerungen von Vorzeichen und Rängen, die im 1-dimensionalen Raum zum üblichen Werkzeug gehören, auf den mehrdimensionalen Raum. Ein paar Erläuterungen dazu sind in Abschnitt 2.16.2 zu finden.

Diese Methoden finden Anwendung zum einen bei Varianzanalysen mit Messwiederholungen, in gleicher Weise wie andere multivariate Verfahren, etwa das unten vorgestellte von Hotelling-Lawley, mit dem Vorteil ohne Spherizität auszukommen, und zum anderen zum Test auf Spherizität. Die für Varianzanalysen mit Messwiederholungen verfügbaren R-Funktionen sind zusammen mit einigen Performance-Hinweisen in Kapitel 5.3.10 aufgeführt.

## 2. 12 Weitere Varianzanalysen für gemischte Versuchspläne

Eine entscheidende Voraussetzung bei Versuchsplänen mit Messwiederholungen ist die Spherizität (vgl. Kapitel 5.2). Insbesondere für gemischte Versuchspläne, also solchen mit sowohl Gruppierungs- als auch Messwiederholungsfaktoren, gibt es jedoch Ansätze, diese zu umgehen. Auf der anderen Seite gibt es einige Methoden, speziell für den Fall ungleicher Varianzen bzw. fehlender Spherizität, die im nächsten Abschnitt aufgeführt werden.

### 2. 12. 1      **Multivariate Tests: Hotelling-Lawley, Pillai und Wilks**

Neben der „klassischen“ parametrischen Varianzanalyse, die die o.a. Sphärität voraussetzt, gibt es noch ein anderes parametrisches Verfahren, das auf der multivariaten Varianzanalyse basiert. Allerdings erfordert dieses eine multivariate Normalverteilung der Residuen. Dies ist zum einen deutlich mehr als die Normalverteilung aller Residuen, zum anderen auch nur aufwändig zu überprüfen.

Hierbei werden zunächst für die  $J$  Messwiederholungen  $x_1, \dots, x_J$  einer Variablen  $x$   $J-1$  Differenzen  $d_1=x_2-x_1$ ,  $d_2=x_3-x_2$ , ... errechnet. Der Ausgangshypothese entspricht dann, dass alle diese  $d_j$  gleich 0 sind. Dies wird über eine multivariate Varianzanalyse geprüft, z.B. mit den Tests von Wilks, Hotelling-Lawley oder Pillai, wobei letzterer eher konservativ, die beiden anderen eher liberal reagieren. Pillais Test ist insbesondere bei nichtnormalen Daten und ungleichen Kovarianzmatrizen zu empfehlen. Der Test von Roy ist dagegen nicht zu empfehlen (vgl Olson, 1976).

### 2. 12. 2      **Koch**

Das Verfahren von Koch (1969) basiert auf dem oben erwähnten Ansatz einer multivariaten Varianzanalyse (vgl. Kapitel 5.2). Dieses wird auf Rangdaten übertragen. Eine R-Funktion wird vom Autor angeboten (vgl. Anhang 3). Bei Monte Carlo-Simulationen schneidet dieses in der Regel sehr gut ab. Vgl dazu Tandon & Moeschberger (1989), Ernst & Kepner (1993) sowie Lüpsen (2020b).

### 2. 12. 3      **Beasley & Zumbo**

Beasley & Zumbo (2009) haben eine Reihe von Tests für die Interaktion bei gemischten Versuchsplänen zusammengestellt. Neben einigen Verfahren, die relativ aufwändig zu programmieren sind, sind auch die Interaktion aus dem Puri & Sen- sowie aus dem ART-Verfahren angeführt. Deren Fazit: I.a. ist die ART-Prozedur den anderen vorzuziehen.

### 2. 12. 4      **Agresti & Pendergast und Akritas & Arnold**

Agresti & Pendergast (1986) haben ein nichtparametrisches Äquivalent zum o.a. multivariaten Test entwickelt. Letztlich werden dabei nur sämtliche  $y$ -Werte zusammen in Ränge transformiert - ähnlich wie bei der o.a. Methode von Puri & Sen - und anschließend o.a. multivariater Test durchgeführt. Hierfür gibt es eine  $\chi^2$ -verteilte und eine F-verteilte Prüfgröße, wobei letztere i.a. vorgezogen wird (siehe z.B. Beasley, 2002). Beasley hat auch einen entsprechenden Test für die Interaktion entwickelt (Beasley, 2002).

Harwell & Serlin (1995) stellen neben dem Test von Agresti & Pendergast noch weitere vor, die auf dem multivariaten Test von Hotelling & Lawley basieren, u.a. einen von Akritas & Arnold (1994), der lediglich die Anwendung des multivariaten Tests auf die wie beim RT-Verfahren rangtransformierten Daten beinhaltet. Die Autoren schlagen zwar einen  $\chi^2$ -Test vor, ebenso ist allerdings der F-Test von Hotelling & Lawley möglich. Wenn auch in der Berechnung nur ein geringer Unterschied besteht, haben eigene Simulationen gezeigt, dass dieser wesentlich besser hinsichtlich Kontrolle der Fehlerrate wie auch der Power abschneidet als der von Agresti & Pendergast.

## 2. 13      **Varianzanalysen für heterogene Varianzen**

Da es inzwischen zahlreiche varianzanalytische Tests für den Fall ungleicher Varianzen gibt, allerdings fast ausschließlich für unabhängige Stichproben, können nur die „wichtigsten“ Ver-

fahren hier aufgeführt werden. In R gibt es dazu insbesondere die Pakete `onewaytests` und `doex`, die eine Reihe Funktionen für weitere Methoden enthalten.

### 2. 13. 1 Welch und Fligner-Policello

Das wohl bekannteste Verfahren stammt von Welch. Er entwickelte einen Zweistichproben- t-Test für ungleiche Varianzen (vgl. Wikipedia). Diesen gibt es auch in einer Version für  $K$  Gruppen (unabhängige Stichproben), der sowohl in R (Funktion `oneway.test`) als auch in SPSS (Prozedur `Oneway`) verfügbar ist.

An dieser Stelle sollte auch der Test von *Fligner-Policello* erwähnt werden. Dieser ist in gleicher Weise die „Rangversion“ des Welch-Tests wie der U-Test von Mann-Whitney die „Rangversion“ des t-Tests ist. Diesen Test gibt es allerdings nur für den 2-Stichproben-Vergleich. Er bietet sich an, wenn ein Mittelwertunterschied getestet werden soll, aber möglicherweise zugleich ungleiche Streuungen vorliegen, weil in solchen Fällen der U-Test auch auf ungleiche Streuungen ansprechen kann. Dieser Test ist in R als Funktion `fp.test` im Paket `RVAideMemoire` vorhanden. Es sei darauf aufmerksam gemacht, dass der *Fligner-Killeen*-Test keinen Mittelwertvergleich sondern einen Test auf homogene Varianzen beinhaltet.

### 2. 13. 2 James 2nd order und Alexander & Govern

Allgemein als beste Tests - hinsichtlich des Fehlers 1. Art sowie der Power - im Fall von inhomogenen Varianzen gelten der von James (1951), genannt 2nd order (wegen der Verwendung einer Taylorreihe 2. Ordnung), sowie der von Alexander & Govern (1994). Die Teststatistik des James-Test folgt leider keiner gängigen Verteilung, weswegen diese mühsam approximiert werden muss. Der Test galt lange als „unberechenbar“. Alexander & Govern haben eine Vereinfachung dieses Tests entwickelt, die aber als fast genauso gut einzustufen ist. Beide Tests gibt es leider nur in einer 1-faktoriellen Version, allerdings auch als SAS-Macro sowie als R-Funktionen `james.test` bzw. `ag.test` im Package `onewaytests`.

### 2. 13. 3 Welch & James

Ein weiterer Versuch, den o.a. Test von James berechenbar zu machen, beinhaltet der Test von Welch & James, und zwar in einer Version von Johansen. Er ist beschrieben von Algina & Olejnik (1984), auch für 2-faktorielle Versuchspläne, erfordert allerdings einigen Programmieraufwand. Eine Variante für gemischte Versuchspläne wurde von Keselman, Carriere & Lix (1993) vorgestellt. Diverse Simulationen (vgl. Algina, 1994) haben jedoch gezeigt, dass letztere, je nach Anzahl  $J$  der Messwiederholungen, nur für sehr große  $n_i$  ( $>50$ ) die Fehlerrate unter Kontrolle hat. Derzeit sind diese Verfahren in den Standardprogrammen nicht verfügbar. Für R werden jedoch beide Varianten als Funktionen vom Autor angeboten (vgl. Anhang 3). Für `split-plot designs` gibt es darüber hinaus die Funktion `welchADF.test` im Paket `welchADF`.

### 2. 13. 4 Brown & Forsythe

Brown & Forsythe (1974) haben einen F-Test für heterogene Varianzen entwickelt für 1- und 2-faktorielle Varianzanalysen (vgl. auch Anhang 2.2), allerdings nur für Gruppierungsfaktoren. Dieser wurde von Mehrotra (1997) verbessert und wird vielfach mit *modified Brown Forsythe* Test bezeichnet. Für 1-faktorielle Analysen ist er als Funktion `bf.test` im Paket `onewaytests`, in der verbesserten Version als Funktion `MBF` im Paket `doex`, sowie in SPSS (Prozedur `Oneway`) verfügbar. Für R wird die Funktion `bf.f` für 2-faktorielle Varianzanalysen vom Autor angeboten (vgl. Anhang 3).

Es gibt eine Erweiterung dieses Verfahrens für gemischte Versuchspläne, die sowohl heterogene Varianzen des Gruppierungsfaktors wie auch des Messwiederholungsfaktors berücksichtigt (vgl. z.B. Vallejo et al., 2004), auch modifizierter Brown & Forsythe-Test (MBF) genannt. Hierzu wird die Funktion `mbf.f` vom Autor zur Verfügung gestellt.

### 2. 13. 5 Brunner, Dette und Munk

Im Zusammenhang mit der Analyse von Kruskal und Wallis wurde oben der Test von Brunner, Dette und Munk (BDM-Test) erwähnt. Er bietet sich an, wenn die Streuungen der Gruppen als unterschiedlich anzusehen sind, da letztlich alle o.a. Methoden auf inhomogene Varianzen reagieren können. Das Verfahren ähnelt dem o.a. von Akritas, Arnold und Brunner, was nicht verwunderlich ist, da dieselben Autoren beteiligt sind, ist aber konservativer. Die Durchführung des Tests ist relativ komplex, da er wie die ATS auf komplexer Matrix-Algebra basiert. Das Verfahren gibt es in einer parametrischen und einer nichtparametrischen Version, z.B. für ordinale Merkmale, und ist von Brunner et al (1997) sowie von Wilcox (2012 und 2013) beschrieben worden. R bietet dafür folgende Pakete: `GFD` für die parametrische Variante (mehr-faktoriell) sowie `asbio` für die nichtparametrische Variante als 1- und 2-faktorielle Varianzanalyse.

Ein anderer Test von *Rust & Fligner* ist ebenfalls in den o.a. Büchern von Wilcox beschrieben. Dieser wird allerdings gegenüber dem oben erwähnten BDM-Test als weniger empfehlenswert angesehen, insbesondere da er keine Bindungen erlaubt.

### 2. 13. 6 Adjustierungen der Freiheitsgrade: $\epsilon$ , GA und IGA

Für den Fall ungleicher Varianzen des Messwiederholungsfaktors, insbesondere fehlender Sphärität, haben Huynh & Feldt sowie Greenhouse & Geisser eine Adjustierung des F-Tests vorgeschlagen, derart dass Zähler- und Nennerfreiheitsgrade des F-Tests mit  $\epsilon'$  multipliziert werden, wobei  $\epsilon' \leq 1$  auf Box  $\epsilon$ , dem Grad der Heterogenität, basiert, Je nach Größe von  $\epsilon$  wird dadurch der F-Test konservativer. Dieses ist heutzutage das gängige Verfahren, um ungleiche Varianzen des Messwiederholungsfaktors zu berücksichtigen. Beide Adjustierungen werden in R und SPSS standardmäßig angeboten. (Mehr dazu in den Kapiteln 5.2 und 6.1.)

Dieses Verfahren hat allerdings Schwächen in gemischten Versuchsplänen, insbesondere bei inhomogenen Kovarianzmatrizen, wenn z.B. die Sphärität in den einzelnen Gruppen unterschiedlich ausfällt. Huynh (1978) hat dazu eine Alternative GA (*general approximation*) für die Adjustierung entwickelt, die nicht auf Box  $\epsilon$  basiert, sondern lediglich auf den Varianzen und Kovarianzen. Er hat diese weiter verbessert zu IGA (*improved general approximation*). Hierbei werden nicht nur die Freiheitsgrade sondern auch der F-Wert gemäß der Heterogenität verkleinert, Wie bei o.a.  $\epsilon'$  führt dieses zu einem konservativerem F-Test. Entsprechende R-Funktionen (`iga` und `iga.anova`) werden vom Autor angeboten.

### 2. 13. 7 Box-Korrektur

An dieser Stelle kann auch eine Korrektur der Freiheitsgrade erwähnt werden, die von Box entwickelt wurde (vgl. Winer, 1991). Über solche Korrekturen wird üblicherweise Varianzhomogenitäten Rechnung getragen. Diese Box-Korrektur ist allerdings als vergleichsweise konservativ einzustufen. Eine entsprechende R-Funktion ist im Anhang 2 zu finden.

## 2. 14 Logistische Regression

Neben der bekannten logistischen Regression für dichotome Kriteriumsvariablen gibt es auch eine für ordinale Variablen. Unter dem Aspekt, dass die parametrische Varianzanalyse ein Spezialfall der linearen Regression ist, bei der die nominalen Prädiktoren passend kodiert werden, ist es einleuchtend, dass dasselbe Vorgehen auch bei der dichotomen und ordinalen logistischen Regression zu einer Varianzanalyse für dichotome bzw. ordinale Kriteriumsvariablen führt. Unter praktischen Aspekten müssen allerdings drei Einschränkungen gemacht werden:

- Erstens ist eine relativ hohe Fallzahl erforderlich,
- zweitens führt das Iterationsverfahren der Maximum-Likelihood-Schätzung nicht immer zum Erfolg, d.h. verschiedentlich gibt es kein Ergebnis, und
- drittens sollte die abhängige Variable nicht zu viele Ausprägungen haben (unter 10).

Das eigentliche Ergebnis der logistischen Regression besteht aus Schätzungen der Modell-Parameter und der dazugehörigen Tests auf Verschiedenheit von 0. Hat ein Faktor mehr als 2 Ausprägungen, so müssen diese Tests für jeden Effekt zu einem varianzanalytischen Test (*anova-like test*) zusammengefasst werden, was je nach Programm nicht automatisch erfolgt. Methoden dazu sind in 9.8 aufgeführt. Im Gegensatz zu den zuvor aufgeführten Verfahren, die alle primär für metrische Kriteriumsvariablen konzipiert, allerdings auch für ordinale Variablen anwendbar sind, ist die ordinale logistische Regression eine Methode, die speziell auf ordinale Merkmale zugeschnitten ist. Die Anwendung ist allerdings nicht so ganz so einfach wie die der übrigen Verfahren. Dank der u.a. Methoden GEE und GLMM ist die logistische Regression auch auf Versuchspläne mit Messwiederholungen anwendbar.

## 2. 15 GEE und GLMM

In den 90er Jahren wurden zwei neue Schätzmethoden speziell für Messwiederholungen entwickelt: GEE (*Generalized Estimating Equations*) sowie die GLMM (*Generalized Linear Mixed Models*), für die mittlerweile zahlreiche Programme bzw. Funktionen, insbesondere in R, verfügbar sind. GEE ist eine Weiterentwicklung des *Marginal Probability Model*, und letztlich sind beide Verallgemeinerungen der *Generalized Linear Models* (GLM) auf Daten mit Messwiederholungen bzw. korrelierende Daten und daher für gemischte Versuchspläne geeignet. Typisch für diese Verfahren sind die *Cluster*, die jeweils sämtliche Messwiederholungen einer Erhebungseinheit, z.B. Versuchsperson, enthalten. Beide Verfahren sind sowohl für metrische, ordinale und dichotome abhängige Variablen einsetzbar. Dies ist möglich über die Spezifikation einer Link-Funktion, die üblicherweise die Werte "gaussian" (metrisch/normalverteilt), "poisson" (Häufigkeiten) und "binomial" (dichotom) annehmen kann. (Einigermaßen) verständliche Einführungen in diese Verfahren bieten u.a. Baltés (2016) und Weyer (2008).

Insbesondere GEE hat im Vergleich zur parametrischen Varianzanalyse und zu GLMM schwächere Voraussetzungen, u.a. keine Normalverteilung der Residuen und keine Varianzhomogenitäten. Auf der anderen Seite muss eine Struktur für die Korrelationsmatrix der Messwiederholungen angegeben werden (vgl. auch Abschnitt 5.2.). Gängige Strukturen für die Korrelationen  $r_{ij}$  sind:

- *exchangeable*: alle  $r_{ij}$  ( $i \neq j$ ) sind gleich,
- *independence*: alle  $r_{ij}$  ( $i \neq j$ ) sind 0,
- *unspecified / unstructured*: alle  $r_{ij}$  ( $i \neq j$ ) sind beliebig,
- *autogressive*: die  $r_{ij}$  ( $i \neq j$ ) errechnen sich als  $r^{i-j}$  ( $i > j$ )

*independence* ist unrealistisch, da Messwiederholungen üblicherweise korrelieren, und *unspecified* ist unpraktibel wegen des sehr hohen Schätzaufwands. *exchangeable* entspricht der *compound symmetry* (vgl. Abschnitt 5.2.) und ist der realistischste Fall neben *autogressive*, bei dem die Korrelationen mit größerem Abstand der Messwiederholungen abnehmen. Wenn auch die Korrelationsstruktur angegeben werden muss, hat sie in der Praxis wenig Einfluss auf das Ergebnis. GLMM erfordern keine entsprechende Spezifikation.

Beide Methoden basieren auf asymptotischer Statistik, d.h. erfordern sehr große Stichproben. Wünschenswert ist ein  $n > 100$ . Dies gilt insbesondere für GLMM, für das Maximum Likelihood-Schätzung verwendet wird, während GEE-Modelle mittels kleinster Quadrat-Schätzung gelöst werden. Während die mit GEE erzielten Schätzungen (Ergebnisse) insbesondere für kleinere  $n$  als zuverlässiger gelten, erlauben die GLMM auch Versuchspläne mit fehlenden Werten auf den Messwiederholungen, ohne dass entsprechende Fälle eliminiert werden müssen.

Wie bei logistischen Regression (s.o.) besteht zunächst einmal das Ergebnis aus der Schätzung der Modell-Parameter und der dazugehörigen Tests auf Verschiedenheit von 0. Hat ein Faktor mehr als 2 Ausprägungen, so müssen diese Tests für jeden Effekt zu einem varianzanalytischen Test (*anova-like test*) zusammengefasst werden, was bei GEE und GLMM in der Regel nicht automatisch erfolgt. Methoden dazu sind in 9.8 aufgeführt. Basis für diese varianzanalytischen Tests sind neben den Parameterschätzungen die Kovarianzmatrizen der Parameterschätzungen. Insbesondere für GEE gibt es hierzu eine Vielzahl von Methoden, wobei der *sandwich estimator* von Liang & Zeger das Standard-Verfahren ist. Eine Übersicht geben Wang et al. (2016). Nicht viel besser sieht es bei GLMM aus, wozu Li et al. (2016) eine Reihe von Methoden zusammengestellt haben.

Abschließend einige verunsichernde Warnungen, die zum einen auf eigenen Erfahrungen, basierend auf Simulationen, beruhen (vgl. Lüpsen, 2018), zum anderen auf Erfahrungen anderer Autoren, die dort zitiert werden, und die letztlich von der Verwendung von GEE und GLMM abraten:

- Insbesondere bei kleineren Stichproben ( $n < 100$ ) kann sehr häufig kein Ergebnis gefunden werden oder es kommt zu Abbrüchen. Die Ausfallrate kann je nach Methode und Daten bis zu 90% (!) betragen.
- Sowohl für GEE als auch für GLMM gibt es mehrere mathematische Verfahren, die nicht einheitliche Resultate liefern. Der fortgeschrittene Leser sei auf Ziegler et al. (1998) für eine Übersicht der GEE-Verfahren bzw. Tuerlinckx (2006) für die GLMM-Verfahren verwiesen.
- Die Wahl der oben erwähnten verschiedenen Parameterschätzungen der Kovarianzmatrizen kann zu recht unterschiedlichen Ergebnissen führen.
- Dieselbe Methode kann bei verschiedenen Programmen oder auch R-Funktionen zu deutlich unterschiedlichen Ergebnissen führen.

GEE wie auch GLMM sind sowohl in R als auch in SPSS verfügbar. In R werden dazu zahlreiche Pakete angeboten, wovon einige der darin enthaltenen Funktionen in den Kapiteln 6.10, 8.3 und 8.4 vorgestellt werden. SPSS bietet dazu die Prozeduren GENLIN (GEE) und GENLINMIXED (GLMM).

## 2. 16 Alternative Rangberechnungen

Normalerweise werden die Ränge für eine Variable  $x$  ermittelt, indem die Werte von  $x$  sortiert und anschließend diesen die Werte  $1, \dots, N$  zugeordnet werden. Aber es gibt auch andere Rangbe-



rechnungen, zum Beispiel unter Berücksichtigung der Stichprobenumfänge  $n_i$ . Und wie werden Ränge von multivariaten Variablen (Vektoren) errechnet?

### 2. 16. 1 Pseudo-Ränge

Die übliche Rangberechnung im Zusammenhang mit der Varianzanalyse kann im Fall von ungleichen  $n_i$  zu paradoxen und widersprüchlichen Ergebnissen führen. Z.B. kann im Fall von 3 Gruppen das Ergebnis sein:  $\mu_1 < \mu_2$ ,  $\mu_2 < \mu_3$  und  $\mu_3 < \mu_1$ , während von den ersten beiden Vergleichen  $\mu_1 < \mu_3$  zu erwarten gewesen wäre. Beispiele dazu sind u.a. bei Happ et al. (2020) zu finden. Eine Lösung dafür bieten die Pseudo-Ränge, bei denen die Stichprobenumfänge  $n_i$  berücksichtigt werden. Dazu kurz die Berechnungen.

Zunächst einmal die „normalen“ Ränge  $R(x)$ . Diese können nicht nur durch Sortieren der Werte, sondern auch „arithmetisch“ berechnet werden:

$$R_{im}(x) = \frac{1}{2} + \sum_k^I \sum_l^{n_k} c(x_{im} - x_{kl}) \text{ für den Rang von Objekt } m \text{ in Gruppe } i.$$

Hierbei ist  $c(\cdot)$  eine Funktion mit den Werten 0, 1/2, 1, je nachdem ob das Argument  $<0$ , 0 oder  $>0$  ist. Die Berechnung des pseudo-Rangs  $\tilde{R}(x)$  erfolgt ähnlich:

$$\tilde{R}_{im}(x) = \frac{1}{2} + \frac{N}{I} \sum_k^I \sum_l^{n_k} c(x_{im} - x_{kl}) / n_k$$

Hierdurch wird die stärkere Gewichtung größerer Stufen bzw. Zellen bei der Berechnung eliminiert, indem  $n_i$  durch  $I/N$  ersetzt wird, ähnlich der Methode der ungewichteten Mittel (siehe Abschnitt 4.3.1.1).

Für R gibt es eine Funktion `pseudorank` im gleichnamigen Paket. Dort wird lediglich der Kruskal-Wallis-Test mit Benutzung der pseudo-Ränge angeboten. Allerdings bietet die Funktion `np.anova` (vgl. Anhang 3.8) ebenfalls eine entsprechende Option. Es sei darauf aufmerksam gemacht, dass die Berechnung der Pseudo-Ränge sehr rechenintensiv ist.

### 2. 16. 2 Spatial Ranks - multivariate Ränge

Die Frage nach multivariaten Rängen stellt sich spätestens bei multivariaten nichtparametrischen Varianzanalysen. Letztere sind zwar hier nicht das Hauptthema, doch wie in den Abschnitten 2.11 und 2.12 angedeutet, können multivariate Verfahren für Analysen mit Messwiederholungen eingesetzt werden.

*Spatial signs* und *spatial ranks* sind zunächst einmal Verallgemeinerungen von Vorzeichen und Rängen, die im 1-dimensionalen Raum zum üblichen Werkzeug gehören, auf den mehrdimensionalen Raum. Die Definitionen sind nicht einfach zu verstehen. Leider gibt es dazu kaum Einführungen. Einzig die Beschreibung des R-Pakets `SpatialNP` (Sirkiä et al., 2007) bietet ein paar verständliche Seiten zum Einstieg. Hier wenigstens die Definitionen:

Das Vorzeichen (*spatial sign*)  $U(y)$  eines Vektors  $y=(y_1, \dots, y_J)$  der Dimension  $J$  errechnet sich als

$$U(y) = y / \|y\|$$

wobei  $\|y\|$  die Norm von  $y$  ist, d.h.  $\|y\| = \sqrt{\left(\sum_j y_j^2\right)}$ .  $U(y)$  ist also selbst wieder ein Vektor, der

auf dem Einheitskreis bzw. -kugel liegt, in Analogie zum univariaten Vorzeichen  $\text{sgn}(x)=x/|x|$  mit Werten 1 und -1, bzw. 0 für die 0. Dieses multivariate Vorzeichen ist z.B. wichtig bei der

Rangberechnung zur Entscheidung, ob ein Vektor  $y^{(1)}$  kleiner oder größer als ein anderer Vektor  $y^{(2)}$  ist. Ist nun  $Y$  eine Matrix, bestehend aus  $n$   $J$ -dimensionalen Vektoren  $y^{(i)}$  ( $i=1, \dots, n$ ), so errechnet sich der *multivariate Rang* (*spatial rank*) von  $y$   $R(y, Y)$  bezogen auf  $Y$

$$R(y, Y) = \frac{1}{n} \sum_i^n U(y - y^{(i)})$$

und der *multivariate Rang mit Vorzeichen* (*spatial signed rank*) von  $y$   $Q(y, Y)$

$$Q(y, Y) = \frac{1}{2n} \sum_i^n \{U(y - y^{(i)}) + U(y + y^{(i)})\}$$

Eine Analogie zur univariaten Rangberechnung ist erkennbar, wenn man die Formel für  $R(y, Y)$  mit der im vorigen Abschnitt angeführten Berechnung von  $R$  vergleicht: In beiden Fällen werden die Vorzeichen aufsummiert, die sich für den Vergleich eines Wertes bzw. Vektors mit allen übrigen ergeben. Die Rangvektoren  $R$  und  $Q$  liegen alle innerhalb des Einheitskreises bzw. -kugel, und zwar umso näher am Rand desto größer der Rang des Vektors zu bewerten ist.  $Q$  und  $R$  weisen beide ungefähr in die Richtung des Vektors. Sie fallen zusammen, wenn der Datensatz symmetrisch zum Nullpunkt ist.

Während das multivariate Vorzeichen in der Literatur einheitlich definiert ist, gibt es für die multivariaten Ränge mehrere Modelle, die zu unterschiedlichen Werten führen. So ist hier noch die Oja-Methode zu erwähnen (vgl. Fischer et al., 2020). Diese Methoden der Rangberechnung sind bereits in vielen Verfahren zum Einsatz gekommen. Hier zu erwähnen die multivariate Varianzanalyse (vgl. 5.3.10) und Tests auf Spherizität (vgl. 5.2).

## 2. 17 Voraussetzungen

Die meisten oben vorgestellten Verfahren basieren auf einer Rangtransformation und sind in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, *nicht* jedoch für Variablen mit *beliebigen* Eigenschaften. D.h. hat die untransformierte Variable  $x$  ungleiche Varianzen, so kann das auch noch für die rangtransformierte Variable  $R(x)$  gelten. Das gilt insbesondere für die RT-, ART-, und INT-Verfahren. U.a. haben Beasley & Zumbo (2009) im Falle der ART-Prozedur darauf hingewiesen. Durch die Rangtransformation werden Verteilungsdeformationen bestenfalls abgemildert, nicht aber beseitigt. So ist es sinnvoll, gegebenenfalls auch  $R(x)$  auf Varianzhomogenität zu überprüfen und gegebenenfalls einen der in Kapitel 4.3.3 vorgestellten Tests zu benutzen. Verschiedentlich wird auch beim Kruskal-Wallis-Test darauf hingewiesen, dass dieser auch auf inhomogene Varianzen anspricht (vgl. Wilcox, 2003), was zwangsläufig dann auch für die Puri & Sen-Tests gilt. Der Grund ist letztlich der Folgende: Fast allen nichtparametrischen Tests liegt die folgende Hypothese  $H_0$  zugrunde: die Verteilungsfunktion  $F_i(x)$  von  $x$  ist in allen  $I$  Gruppen dieselbe:

$$F_1(x) = F_2(x) = \dots = F_I(x)$$

Wird  $H_0$  abgelehnt, so kann man daraus eigentlich nur dann auf ungleiche Mittelwerte schließen, wenn die Verteilungen gleiche Streuungen - streng genommen auch die gleiche Verteilungsform - haben. Daher findet man auch häufig bei Tests die zusätzliche Voraussetzung: die Verteilungsfunktionen  $F_i(x)$  haben die gleiche Form, was gleiche Streuungen impliziert.

Eine generelle Empfehlung: Wird eine Voraussetzung mit einem Test geprüft, z.B. Normalverteilung oder Varianzhomogenität, und gibt es auf der anderen Seite ein alternatives Verfahren für den Fall, dass die Voraussetzung nicht erfüllt ist, so sollte man sich schon bei einem schwach signifikanten Testergebnis (etwa  $\alpha=0.10$ ) für die alternative Methode entscheiden.

### 2. 17. 1      **Versuchspläne ohne Messwiederholungen**

Im Fall von unabhängigen Stichproben empfiehlt sich zum Test der Varianzhomogenität von  $R(x)$  der *Levene-Test*, da dieser (im Gegensatz zum klassischen F-Test, zum  $F_{\max}$ -Test oder zum Bartlett-Test) robust gegen Abweichungen von der Normalverteilung und auch für ordinale Variablen anwendbar ist. Allerdings gibt es wenig Alternativen für den Fall, dass sich die Varianzen auch nach der Rangtransformation als inhomogen erweisen. Einige werden in den Kapiteln 4.2.2 und 4.3.3 vorgestellt, sind aber fast ausschließlich nur in R verfügbar. Eine allgemeine Möglichkeit besteht in der Box-Korrektur der Freiheitsgrade (vgl. Anhang 2.1), die im Fall der RT-, ART-, und INT-Verfahren angewendet werden kann. Leider ist die Box-Korrektur weder in R noch in SPSS standardmäßig verfügbar. Allerdings wird dafür eine Funktion, `box.f`, vom Autor bereitgestellt (vgl. Anhang 3).

### 2. 17. 2      **Versuchspläne mit Messwiederholungen**

Im Fall von abhängigen Stichproben (Messwiederholungen) wird man meistens den Mauchly-Test (vgl. Kapitel 5.2) zur Prüfung der Varianzhomogenität benutzen - genau genommen zur Prüfung der Sphärizität, was etwas mehr beinhaltet. Dieser hat allerdings nicht annähernd die robuste Eigenschaften eines Levene-Tests. Es gibt zwar einen entsprechenden Test für Rangdaten von *Hallin und Paindaveine* (2006), der aber noch nicht in den Softwaresystemen verfügbar ist. Darüber hinaus gibt es z.B. einen nichtparametrischen Test, der auf räumlichen Vorzeichen beruht (vgl. Kapitel 5.3.1). Alternativ wird in Kapitel 5 ein einfacher robuster Test auf Gleichheit der Varianzen vorgestellt. Beasley und Zumbo (2009) propagieren, bei den F-Tests einfach eine der Korrekturen der Freiheitsgrade von Huynh-Feldt oder Greenhouse-Geisser vorzunehmen (siehe Kapitel 5.2), ohne das Ergebnis des Mauchly-Tests zu berücksichtigen. Darüber hinaus werden in Kapitel 6.10 mehrere Verfahren vorgestellt, die keine Homogenitätsvoraussetzungen haben, z.B. die multivariate Varianzanalyse sowie einige andere, die allerdings zum Teil nur für R bereitstehen.

Auf der anderen Seite kann geschlossen werden: Erfüllen die nichttransformierten Daten die Voraussetzung der Varianzhomogenität, so gilt diese auch für die rangtransformierten Daten, so dass gegebenenfalls eine Überprüfung dafür entfallen kann.

Einige nichtparametrische Methoden, wie das Puri & Sen-, das verallgemeinerte Kruskal-Wallis-Friedman (KWF), das van der Waerden- und das Koch-Verfahren verwenden  $\chi^2$ - anstatt F-Tests. Harwell & Serlin (1994) weisen (in Berufung auf E.L. Lehmann) darauf hin, dass der Friedman-Test, und damit auch der van der Waerden-Test, formal sowohl gleiche Varianzen als auch gleiche Kovarianzen verlangt, wenn dies auch von fast allen anderen Autoren ignoriert wird. Jedoch gilt i.a. als Voraussetzung, dass die Verteilungen in allen Gruppen die gleiche Form (*equal shapes*) haben. Das impliziert insbesondere gleiche Streuungen. Allerdings haben Simulationen gezeigt (z.B. Lüpsen, 2020a, und Lüpsen, 2020b), dass insbesondere die oben angeführten Methoden deutlich robuster gegen Varianzheterogenitäten sind als der parametrische F-Test, auch im Fall von Messwiederholungen. Da die Methode von Koch auf der multivariaten Varianzanalyse basiert, entfällt dafür die Voraussetzung der Sphärizität, insbesondere also der Varianzhomogenität. Dennoch ist es empfehlenswert, auch bei den Methoden, die auf  $\chi^2$ - anstatt F-Tests beruhen, die Gleichheit der Varianzen für die Messwiederholungsvariablen zu überprüfen, zumindest im Fall von signifikanten Ergebnissen, da die Tests im Fall von heterogenen Varianzen mit Verletzung des  $\alpha$ -Risikos reagieren. Allerdings verlangen die drei o.a. Verfahren Gleichheit der Kovarianzmatrizen, wenn diese auch relativ robust gegen Verletzungen dieser Art sind, zanz besonders das KWF-Verfahren (vgl. Lüpsen, 2020a).

## 2. 18 Vergleiche

### 2. 18. 1 Versuchspläne ohne Messwiederholungen

An dieser Stelle sollte ein Vergleich von Lüpsen (2016c) erwähnt werden, in dem fast alle hier vorgestellten Verfahren für die 2-faktorielle Varianzanalyse ohne Messwiederholungen verglichen werden, für 14 verschiedene Verteilungen, für homogene und heterogene Varianzen sowie für diverse Modelle. Er favorisiert die Methode von van der Waerden wegen seiner Robustheit hinsichtlich heterogener Varianzen und wegen der relativ hohen Power, insbesondere für große Zellenbesetzungszahlen  $n > 10$ . Dagegen halten das ART-, ART+INT-, RT- und ATS-Verfahren, manchmal auch das INT- und Puri & Sen-Verfahren, in Fällen ungleicher Varianzen häufig das  $\alpha$ -Risiko nicht unter Kontrolle, gelegentlich selbst bei gleichen  $n_{ij}$ . Dabei steigen die Fehlerraten deutlich über 0,10 (bei  $\alpha=0,05$ ) für  $n_{ij} > 50$ . Auf die Tücken des ART wurde bereits in Abschnitt 2.4 hingewiesen. Für kleine  $n$ , etwa 10 und kleiner, ist auch die ART+INT-Methode eine gute Wahl, die allerdings für  $n > 20$  häufig den Fehler 1. Art nicht unter Kontrolle halten kann. Einzig bei negativem Pairing (siehe Abschnitt 1.1.6) wird die ATS-Methode empfohlen. In dieser Publikation sind auch Hinweise auf zahlreiche Vergleiche anderer Autoren zu finden.

Die RT-, ART- und Puri & Sen-Methoden werden von Sawilowsky (1990) und Toothaker & De Newman (1994) mit dem F-Test verglichen. Deren Ergebnis: Der Puri & Sen-Test hält zwar den Fehler 1. Art unter Kontrolle, ist aber recht konservativ, wenn andere Effekte vorhanden sind. Für diesen Fall schlagen sie die ART-Prozedur vor. Da aber alle untersuchten Verfahren in irgendwelchen Situationen zu liberal reagieren, geben sie keine generelle Empfehlung aus.

Mansouri und Chang (1995) vergleichen die INT-Verfahren (normal scores und van der Waerden) u.a. mit dem RT-Verfahren. Sie kommen zu dem Schluss, dass die Transformation in normal scores sowohl beim RT- wie auch beim ART-Verfahren, woraus die INT- bzw. ART+INT-Verfahren resultieren, durchweg eine Verbesserung bringen: zum einen bzgl. der Power, zum anderen bzgl. des Fehlers 1. Art bei heterogenen Varianzen.

Ein Vergleich des ATS-Tests mit anderen Methoden wird von Hahn, Konietzschke und Salmaso (2013) geboten. Sie bestätigen, dass die ATS-Methode zwar generell das  $\alpha$ -Risiko unter Kontrolle hält, jedoch auf Kosten der vergleichsweise sehr geringen Teststärke. Allerdings werden nur kleinere  $n_{ij} < 15$  berücksichtigt. Zu ähnlichen Schlüssen gelangen auch Richter & Payton (2003).

Die ART-Methode wird zwar von vielen Autoren recht positiv beurteilt, u.a. von Lei et al. (2004) und Mansouri et al. (2004), die aber meistens kritische Fälle wie ungleiche Varianzen nicht berücksichtigen.

Eine Übersicht fast aller Verfahren mit einem Vergleich der Fehlerraten 1. Art und der Power auf Basis verschiedener Simulationen bietet Danaba (2009), wenn auch diese Arbeit wegen typografischer Mängel nicht ganz einfach zu lesen ist. Sein Fazit: RT, INT, Puri & Sen sowie ATS verhalten sich robust gegen Verletzungen der Voraussetzungen und haben eine Power, die der des F-Tests überlegen ist, ausgenommen im Fall der Exponential-Verteilung. Dagegen fällt das ART-Verfahren bei dem Vergleich durch. Leider berücksichtigt er nicht heterogene Varianzen.

Zum Schluss noch zum Vergleich der Verfahren für heterogene Varianzen. Generell gilt, dass die (hier vorgestellten) nichtparametrischen Verfahren die durch ungleiche Varianzen verursachten Probleme nicht lösen können (siehe z.B. Alexander & Govern, 1994, sowie Tomarken & Serlin, 1986). Leider beschränken sich die Vergleiche fast ausnahmslos auf die 1-

faktorielle Analyse. Diese zeigen, dass es kein Verfahren gibt, das universell empfehlenswert ist, da es für jedes Situationen gibt, in denen die Resultate unbefriedigend sind, insbesondere bei schiefen Verteilungen und im Fall von kleinen Zelhäufigkeiten, etwa  $n < 10$  (siehe z.B. Lix et al., 1996, sowie Schneider & Penfield, 1997). Dennoch werden die o.a. Methoden von James sowie von Alexander & Govern am besten eingestuft. Von den weiter verbreiteten Tests ist der von Brown & Forsythe gegenüber dem von Welch vorzuziehen (vgl. Jennifer J. Clinch et al., 1982). Hilfreich ist auch eine Studie von Dijkstra (1987), der die 1-faktorielle Varianzanalyse bei fehlenden Voraussetzungen untersucht hat und die van der Waerden-Methode favorisiert.

Abschließend sei noch darauf aufmerksam gemacht, dass in Kapitel 7.1 ausführlich über Vergleiche der Methoden für dichotome abhängige Variablen berichtet wird, sowohl für Versuchspläne ohne wie auch mit Messwiederholungen.

## 2. 18. 2      Versuchspläne mit Messwiederholungen

Vorab sei angemerkt, dass es zwar zahlreiche Monte Carlo-Simulationsstudien zu diesen Designs gibt, die Ergebnisse in der Regel allerdings recht "kompliziert" sind und vom Zusammenspiel mehrerer Faktoren wie Sphärität, Homogenität der Kovarianzmatrizen, gewählte Verteilungen, Anzahl der Messwiederholungen oder Größe der Stichproben abhängig und daher schwer wiederzugeben sind. Ein paar sollten dennoch erwähnt werden:

Lüpsen (2020b) verglich den F-Test mit der KWF-, van der Waerden-, Puri & Sen-, INT-, ART+INT-, ATS- und Koch-Methode. Der F-Test zusammen mit der Huynh-Feldt-Adjustierung ist in der Regel eine gute Wahl, insbesondere hinsichtlich der Teststärke. Dennoch bieten die nichtparametrischen Verfahren KWF, van der Waerden und Koch in vielen Fällen eine bessere Kontrolle des Fehlers 1. Art, und eine bessere Power bei nichtnormalen Verteilungen, während die ART+INT- und die Puri & Sen-Methoden häufig sich zu liberal verhalten. Sowohl der F-Test, mit und ohne Huynh-Feldt-Adjustierung, als auch die multivariaten Tests reagieren empfindlich auf heterogene Kovarianzmatrizen: beide mit stark erhöhtem Fehler 1. Art bei negativem Pairing der Varianzen, und mit geringerer Power bei positivem Pairing der Korrelationen (vgl. Lüpsen, 2020a). Er machte auch darauf aufmerksam, dass das KWF- sowie das van der Waerden-Verfahren als einzige das  $\alpha$ -Risiko auch bei heterogenen Kovarianzmatrizen unter Kontrolle halten und dadurch eine gute Wahl sind, zumal sie zwar keine Sphärität verlangen und somit lästige Prüfungen der Voraussetzungen entfallen, jedoch die Gleichheit der Varianzen kontrolliert werden sollte. Allerdings zeigte sich auch, dass die van der Waerden-Methode bei gemischten Versuchsplänen nicht die gleiche hervorragende Teststärke besitzt wie bei Analysen ohne Messwiederholungen. Komplett enttäuschend schnitt das ATS-Verfahren ab, da es für kleine bis mittelere  $n_i < 30$  stark überhöhte Fehlerraten zeigt, z.B. bis zu 0.15 für ein nominelles  $\alpha=0.05$  bei  $n_i=10$ , insbesondere für den Gruppierungsfaktor (vgl. auch Tian & Wilcox, 2007). Für das ART+INT-Verfahren gilt dieselbe Einschränkung bzgl. diskreter abhängiger Variablen wie bei Designs ohne Messwiederholungen.

In neueren Studien werden das Verfahren IGA von Huynh, der modifizierte Brown-Forsythe-Test (MBF) von Vallejo und das Verfahren von Welch & James als die robustesten angesehen. Hierfür gibt es allerdings noch wenig Vergleiche, zumal hierfür kaum Funktionen bereit stehen. Der Autor hat jedoch inzwischen die Funktionen `iga . anova`, `mbf . f` sowie `wj . spanova` für R erstellt, auf deren Basis derzeit in Vergleich erstellt wird. Erste, noch nicht veröffentlichte, Ergebnisse haben gezeigt, dass die IGA- und MBF-Methoden komplett das  $\alpha$ -Risiko einhalten, selbst in den unangenehmen Pairing-Situationen, und auf der anderen Seite eine vergleichsweise hervorragende Power aufweisen.

Harwell & Serlin (1994) verglichen den F-Test mit dem RT-, dem ART Verfahren sowie den Tests von Puri & Sen und Friedman bei 1-faktoriellen Analysen mit Messwiederholungen. Es zeigte sich, dass die nichtparametrischen Methoden ebenso unangenehm auf ungleiche Varianzen der Messwiederholungen reagieren wie der F-Test, insbesondere auch bei nichtnormalen Verteilungen. In einer weiteren Studie haben Harwell & Serlin (1994) auch den multivariaten Test von Hotelling-Lawley untersucht. Dieser schnitt in der Regel gut ab, auch bei nichtnormalen Verteilungen, während der Puri & Sen-Test zu konservativ und das RT-Verfahren häufig zu liberal war, insbesondere bei heterogenen Varianzen.

Beasley & Zumbo (2003) haben einen ähnlichen Vergleich durchgeführt und kamen zu dem Ergebnis, dass (a) fehlende Sphärizität zu liberalen Ergebnissen führt, (b) multivariate Tests erst ab  $n > 20$  zuverlässig sind und (c) der F-Test durchweg die größte Power besitzt. Zu dem gleichen Resultat gelangten auch Stiger et al. (1998), die jedoch ordinale Kriteriumsvariablen auf einer 4er-Skala untersucht hatten.

## 2. 19 Entscheidungshilfen zur Auswahl

Bei allen oben genannten positiven und negativen Eigenschaften der Verfahren ist es nicht leicht, das passende auszuwählen. Daher werden nachfolgend einige Kriterien aufgeführt, die natürlich voraussetzen, dass der Untersucher einige Kenntnisse über seine Daten besitzt. Generell kann jedoch gesagt werden, dass in den meisten Fällen der klassische F-Test eine durchaus gute Wahl ist. Daher sollte man, egal ob das zu analysierende Merkmal metrisches oder ordinale Skalenniveau hat, zunächst die Voraussetzungen prüfen und danach entscheiden, ob überhaupt in Anbetracht der Robustheit der Varianzanalyse ein nichtparametrisches Verfahren erforderlich ist. Die einfachste nichtparametrische Varianzanalyse ist die normal scores-Methode (INT), die eine relativ hohe Effizienz hat und gegenüber dem RT-Verfahren einige Bedenken ausräumen kann. Mit dem etwas aufwändigeren van der Waerden-Test ist man allerdings auf der sichereren Seite hinsichtlich der Kontrolle des Fehlers 1. Art, leider auf Kosten der Power, insbesondere bei kleinen Stichproben.

### 2. 19. 1 Warnungen

Im Fall von rechtsschiefen Verteilungen, insbesondere bei einer Lognormal- oder Exponentialverteilung sollte der parametrische F-Test angewandt werden. In solchen Fällen können bei allen rangbasierten Verfahren - und das sind die meisten nichtparametrischen Verfahren - die kleinsten Streuungsunterschiede schon zu falsch signifikanten Ergebnissen führen (vgl. Lüpsen, 2016b). Und im Fall einer Exponentialverteilung hält der F-Test das  $\alpha$ -Risiko komplett unter Kontrolle und hat zugleich die größte Power. Vgl. dazu Zimmerman (2004) sowie Carletti & Clautriaux (2005). Diese Verteilungsformen kommen in der Praxis häufig vor, typischerweise in der Medizin, z.B. Blutdruck, oder in der Wirtschaft, z.B. Verbrauchsdaten oder Einkommen.

Vielfach wird die ART-Methode favorisiert. Deren Anwendung sollte jedoch vermieden werden bei heterogenen Varianzen, bei rechtsschiefen Verteilungen, bei diskreten, insbesondere dichotomen, Variablen mit wenigen Ausprägungen (unter 8) und generell für den Test von Haupteffekten. Die negativen Eigenschaften werden zum Teil abgemildert durch die INT-Transformation, allerdings nicht bei diskreten Variablen. D.h. die ART+INT-Methode ist gegebenenfalls vorzuziehen (vgl. dazu Lüpsen, 2016a).

Es muss allerdings noch einmal an die Bedenken einiger Autoren hinsichtlich des Tests von Interaktionen erinnert werden (vgl. Abschnitt 2.2), die insbesondere die RT-, INT-, Puri & Sen-, KWF- und abgeschwächt auch für die van der Waerden- und Koch-Methode gelten. Diese ziel-

en darauf, dass der Test der Interaktion nicht zuverlässig ist und gelegentlich fälschlicherweise signifikant sein kann, insbesondere wenn einer der Haupteffekte von Null verschieden ist. Dieses Problem verschärft sich zwangsläufig bei 3- und mehrfaktoriellen Versuchsplänen. Wie man auch in solchen Fällen trotz Anwendung von rangbasierten Verfahren zum Ziel kommen kann, hat u.a. Koch (1970) an einer Auswertung eines 3-faktoriellen Split-Plot Designs demonstriert, indem zunächst diverse Einzelhypothesen aufgestellt und dann schrittweise getestet wurden. Leider haben die anderen Methoden wie ART und ATS, oder GEE und GLMM ebenfalls Schwächen, wie z.B. eine geringe Teststärke oder erhöhte Fehlerraten in anderen Situationen.

Eine andere Warnung betrifft die Rangberechnungen im Fall von Stichproben mit stark unterschiedlichen  $n_i$  (vgl. Abschnitt 2.16.1). Die Probleme können mit der Verwendung von Pseudo-Rängen umgangen werden. Doch zum einen gibt es bislang kaum Varianzanalyse-Programme, die diese unterstützen, und zum anderen gibt es noch wenig empirische Erfahrungen, da die Entwicklung der Pseudo-Ränge noch recht jung ist.

### 2. 19. 2 Versuchspläne ohne Messwiederholungen

Der parametrische F-Test kann problemlos angewandt werden, solange entweder gleiche Zellenbesetzungszahlen  $n_i$  oder gleiche Varianzen vorliegen. Lediglich die Verbindung von nichtbalancierten (ungleichen  $n_i$ ) Versuchsplänen mit heterogenen (ungleichen  $s_i^2$ ) Varianzen verlangt nach besonderen Methoden. Bei Versuchsplänen mit ungleichen  $n_i$  und ungleichen Varianzen  $s_i^2$  spielt die Paarung eine entscheidende Rolle. Der kritischste Fall liegt vor, wenn kleine  $n_i$  mit großen  $s_i^2$  gepaart sind (*negative pairing*). Hier sind die ATS-Methode sowie die in 2.13. aufgeführten Methoden die einzigen, die den Fehler 1. Art unter Kontrolle halten. Von allen anderen sind der Puri & Sen- und der van der Waerden-Test diejenigen, die noch am besten abschneiden, wenn auch sie das  $\alpha$ -Risiko verletzen, allerdings in Maßen. (Das KWF-Verfahren ist hier mit dem von Puri & Sen identisch.) Vergleichsweise harmlos ist dagegen der Fall, wenn kleine  $n_i$  mit kleinen  $s_i^2$  gepaart sind (*positive pairing*). Hier verletzen zwar keine Verfahren den Fehler 1. Art, reagieren allerdings konservativ, haben also nur noch eine geringe Teststärke. In diesem Fall sind wiederum die in 2.13. aufgeführten Methoden vorzuziehen. Der Fall, dass die (ungleichen)  $n_i$  und die (ungleichen)  $s_i^2$  unabhängig sind, wird der Normalfall sein. Bei heterogenen Varianzen „schwächelt“ nicht nur der F-Test sondern leider auch fast alle nichtparametrischen Tests. Einzig der v.d.Waerden-Test hält das  $\alpha$ -Risiko unter Kontrolle. Wem der Rechenaufwand zu groß ist, kann ersatzweise auch die INT-Methode wählen.

Abschließend noch ein paar Bemerkungen zu den Methoden für heterogene Varianzen, die im Fall eines positiven oder negativen pairings vorzuziehen sind. Mit Abstand am besten werden die Verfahren von James sowie Alexander & Govern bewertet, die allerdings nur in 1-faktoriellen Versionen bekannt sind. Für 2-faktorielle Analysen gibt es die Methoden von Welch & James (WJ), Brown & Forsythe (BF), Akritas & Brunner (ATS) sowie Brunner, Dette & Munk (BDM). ATS und BDM sind extrem konservativ, während BF leicht liberal reagiert. Der WJ ist vielleicht trotz Schwächen bei der Interaktion der empfehlenswerteste.

### 2. 19. 3 Versuchspläne mit Messwiederholungen

Auch im Fall von Messwiederholungen ist der F-Test durchweg zu empfehlen, aber er sollte generell mit der Huynh-Feldt-Adjustierung verwendet werden. In den Kapiteln 5.2 und 6.1 wird ausführlich über die recht umfangreichen Voraussetzungen des F-Tests berichtet und Alternativen erwähnt, was an dieser Stelle nicht wiederholt werden soll. Insbesondere inhomogene Kovarianzmatrizen führen bei den parametrischen Verfahren zu unzuverlässigen Ergebnissen, was

eine Prüfung dieser Voraussetzung mit einer anschließenden Wahl der passenden Methode nahe legt. Die sichersten Alternativen, insbesondere bei ungleichen Kovarianzmatrizen, sind die IGA-Adjustierung von Huynh-Feldt und der modifizierte Brown-Forsythe-Test (MBF). Eine einigermaßen sichere Alternative, die kaum Prüfungen von Voraussetzungen verlangt, ist das KWF-Verfahren. Allerdings muss als nachteilig erwähnt werden, dass dieses in einigen Fällen nur eine um ca. 20% geringere Power gegenüber dem F-Test erreicht, so u.a. bei kleineren Stichproben ( $n_i \leq 30$ ) sowie bei normal oder exponentiell verteilten Daten. Dagegen hat es eine etwas höhere Teststärke bei heterogenen Varianzen und bei diversen nichtnormalverteilten Daten.

Eine ausführliche Abhandlung über die Kriterien zur Auswahl eines Verfahrens im Fall von gemischten Versuchsplänen gibt es von Koch et al. (1980).



### 3. Funktionen zur Varianzanalyse in R und SPSS

Auch für die nichtparametrischen Varianzanalysen greift man fast immer auf die klassischen parametrischen Methoden zurück, um anschließend die Ergebnisse weiterzuverarbeiten. Daher nachfolgend ein Überblick über die Möglichkeiten in R (Version 3.5.3) und SPSS (Version 25).

#### 3. 1 Funktionen in R

Varianzanalysen sind in R nicht so problemlos durchzuführen, wie man erwarten sollte. Das hat im Wesentlichen zwei Gründe:

- Zum einen verwendet R für die in der Programmiersprache S vorgesehene Funktion `aov` die Berechnung der Streuungsquadrate vom Typ I, eine Methode, die zum einen problematisch ist und zum anderen von kaum einem anderen Programm benutzt wird (vgl. dazu das Kapitel 4.3.1.1). Weitere Hinweise hierzu bieten Scholer (2016) und Meyer (2008).
- Zum anderen müssen viele im Zusammenhang mit der Varianzanalyse erforderlichen Tests (z.B. Varianzhomogenitätstests oder multiple Mittelwertvergleiche) mühsam mit anderen Funktionen durchgeführt werden, was allerdings in R nicht unüblich ist.

Das hat dazu geführt, dass es inzwischen fast zahllose Funktionen zur Varianzanalyse in diversen hinzuzuladenden Paketen gibt. Von denen können hier nur wenige erwähnt werden.

Generell müssen die Faktoren, die unabhängigen Variablen, deren Einfluss getestet werden soll, vom Typ „factor“ sein, auch wenn sie nur zwei Stufen (Ausprägungen) haben. Darüber hinaus ist vielfach, insbesondere bei Messwiederholungen, eine numerische Fallkennung „subject“ erforderlich, die ebenfalls vom Typ „factor“ sein muss. Eine Anweisung sollte immer zu Beginn jeder Sitzung ausgeführt werden:

```
options (contrasts=c("contr.sum", "contr.poly"))
```

(vgl. Kapitel 9.4) um korrekte Ergebnisse zu erhalten.

- `aov`  

```
aov (abh.Variable ~ Faktor1*Faktor2*..., Dataframe)
```

für unabhängige Stichproben  

```
aov (abh.Variable ~ Faktor1*... + Error(subject/Faktor1*...), Dataframe)
```

für abhängige Stichproben  
`aov` berechnet Quadratsummen vom Typ I. Um solche vom Typ III zu erhalten, ist neben der o.a. `options`-Anweisung der folgende Schritt erforderlich:  
Wenn `model` das Ergebnis von `aov` enthält, dann werden die Quadratsummen vom Typ III mit Tests ausgegeben über  

```
drop1 (model, ~. , test="F")
```
- `lm`  

```
anova (lm (abh.Variable ~ Faktor1*Faktor2*..., Dataframe)
```

für unabhängige Stichproben  
Um Quadratsummen vom Typ III zu erhalten, sind dieselben Schritte wie bei `aov` erforderlich.  
Vorteil gegenüber `aov`: Die Ergebnisse, wie z.B. die Quadratsummen lassen sich weiterverarbeiten, was vielfach erforderlich ist.
- `lm` mit `Anova` (im Paket `car`)  

```
Anova (lm (abh.Variable ~ Faktor1*Faktor2*..., Dataframe), type="III")
```

Für direkte Berechnung der Quadratsummen vom Typ III und mit weiterverarbeitbaren Ergebnissen.

- `lm` mit `anova` für multivariate Tests bei Analysen mit Messwiederholungen  
`anova(lm(cbind(x2-x1,x3-x2,...)~Faktor1*Faktor2*...),  
test="Name")` mit `Name=Hotelling-Lawley, Pillai` oder `Wilks`  
wobei `x1,x2,x3,..` die Messwiederholungsvariablen sind.
- `ezANOVA` (im Paket `ez`)  
`azANOVA (Dataframe, .(abh.Variable), .(subject),  
between=.(Faktoren), within=.(Faktoren))`  
sowohl für Gruppierungsfaktoren (`between=.`)  
als auch für Messwiederholungsfaktoren (`within=.`)  
Bei Messwiederholungsfaktoren Ausgabe des Mauchly-Tests sowie der modifizierten Tests von Geisser & Greenhouse sowie von Huynh & Feldt, sonst Ausgabe des Levene-Tests. Berechnung der Quadratsummen vom Typ III möglich (`type=3`). Diese Funktion ist zwar einfach zu benutzen, hat aber zwei Schwächen: zum einen muss immer eine numerische Fallkennung `subject` angegeben werden, zum anderen meldet sie häufig fälschlicherweise Eingabefehler oder ungültige Variablenangaben.
- `aov_ez` (im Paket `afex`)  
`aov_ez (data=Dataframe, dv="abh.Variable", id="subject",  
between=c(Faktoren), within=c(Faktoren))`  
für parametrische Analysen mit und ohne Messwiederholungen.
- `rankFD` (im Paket `rankFD`)  
`rankFD (abh.Variable ~ Faktor1*Faktor2*..., Dataframe)`  
für unabhängige Stichproben nach dem ATS-Verfahren von Akritas, Arnold & Brunner.
- `nparLD` (im Paket `nparLD`)  
`nparLD(abh.Variable~Faktor1*Faktor2*...,Dataframe, subject)`  
für nichtparametrische Analysen mit Messwiederholungen nach dem ATS-Verfahren von Akritas, Arnold & Brunner.  
Es können auch Versuchspläne mit fehlenden Werten analysiert werden. Dafür stehen je nach Design die Funktionen `f1.ld.f1`, `f2.ld.f1`, `f1.ld.f2`, `ld.f1` und `ld.f2` zur Verfügung.
- `RM` (im Paket `MANOVA.RM`)  
für dieselben Anwendungen wie beim Paket `nparLD`, jedoch basierend auf der Resampling-Methode zur Berechnung der Irrtumswahrscheinlichkeiten.
- `oneway.test`  
`oneway.test (abh.Variable ~ Faktor, Dataframe)`  
für unabhängige Stichproben  
Robuste 1-faktorielle Varianzanalyse für inhomogene Varianzen nach dem Verfahren von Welch.
- `ag.test`, `james.test`, `bf.test`, `welch.test` (im Paket `onewaytests`)  
`...test (abh.Variable ~ Faktor, Dataframe)`  
für unabhängige Stichproben  
Robuste 1-faktorielle Varianzanalysen für inhomogene Varianzen nach den Verfahren von Alexander & Govern, James, Brown & Forsythe sowie Welch.
- `friedman.test`  
`friedman.test (Datenmatrix)`  
1-faktorielle nichtparametrische Varianzanalyse mit Messwiederholungen nach dem Verfahren von Friedman.

- `quade.test`  
`quade.test (Datenmatrix)`  
 1-faktorielle nichtparametrische Varianzanalyse mit Messwiederholungen nach dem Verfahren von Quade.
- `waerden.test` (im Paket `agricolae`)  
`waerden.test (abh.Variable, Faktor, group=F, console=T)`  
 1-faktorielle Varianzanalyse mit normal scores nach dem Verfahren von van der Waerden.
- BDM (im Paket `asbio`) und GFD (im Paket `GFD`)  
`BDM.test (abh.Variable, Faktor)`  
`BDM.2way (abh.Variable, Faktor1, Faktor2)`  
`GFD (abh.Variable ~ Faktor1*Faktor2*..., Dataframe)`  
 mehrfaktorielle robuste Varianzanalyse nach dem Verfahren von Brunner, Dette, Munk.
- `SkiMack` (im Paket `Skillings.Mack`)  
`SkiMack (as.matrix(Datenmatrix))`  
 1-faktorielle Varianzanalyse mit Messwiederholungen bei fehlenden Werten nach dem Verfahren von Skillings & Mack.

Darüber hinaus wird vom Autor eine Bibliothek mit überwiegend nichtparametrischen Varianzanalysen zur Verfügung gestellt. Für Details siehe Anhang 3.

## 3. 2 Funktionen in SPSS

Varianzanalysen sind mit SPSS vergleichsweise einfach durchzuführen. Generell ist zu beachten, dass gegebenenfalls vorher das Skalenniveau der analysierten Variablen auf „Skala“ gesetzt wird. Zur Verfügung stehen:

- `Oneway`  
`Oneway abh.Variable BY Faktor`  
 (Menü: Mittelwerte vergleichen -> einfaktorielle ANOVA)  
 1-faktorielle Analyse für unabhängige Stichproben.  
 Unter „Optionen“ kann der Levene-Test auf Gleichheit der Varianzen sowie die F-Tests von Welch und Brown & Forsythe im Falle von heterogenen Varianzen angefordert werden.
- `Unianova`  
`Unianova abh.Variable BY Faktor1 Faktor2...`  
 (Menü: Allgemeines lineares Modell -> Univariat)  
 mehrfaktorielle Analyse für unabhängige Stichproben.  
 Unter „Optionen“ kann der Levene-Test auf Gleichheit der Varianzen angefordert werden.  
 Unter „Modell“ kann die Methode zur Berechnung der Streuungsquadrate gewählt werden (Typ I, II oder III).
- `GLM`  
`GLM Messwiederholungsvariablen BY Faktor1 Faktor2...`  
`/WSFactor=... /WSDesign=... /Design=...`  
 (Menü: Allgemeines lineares Modell -> Messwiederholung)  
 mehrfaktorielle Analyse für unabhängige und abhängige Stichproben, mit multivariaten Tests bei Messwiederholungen.  
 Unter „Optionen“ kann der Levene-Test auf Gleichheit der Varianzen bzw. der Box-Test auf Gleichheit Kovarianzmatrizen angefordert werden.  
 Unter „Modell“ kann die Methode zur Berechnung der Streuungsquadrate gewählt werden (Typ I, II oder III).

Mauchly's Test auf Sphärität sowie der modifizierten Tests von Geisser & Greenhouse bzw. von Huynh & Feldt werden immer ausgegeben.

- `Nptests`

`/independent test (abh.Variable) group (Faktor) kruskal_wallis`

`/related test (Messwiederholungsvariablen) friedman`

(Menü: Nichtparametrische Verfahren -> k Stichproben)

1-faktorielle nichtparametrische Analyse für unabhängige Stichproben (Kruskal-Wallis-

Test) bzw. 1-faktorielle nichtparametrische Analyse für abhängige Stichproben (Friedman-Test).

### 3.3 Fehler bei der Rangberechnung

Gelegentlich werden die Ränge mit der Funktion `rank` sowohl in R als auch in SPSS falsch berechnet. Das hört sich schlimm an, hat aber einen einfachen Grund: Rundungsfehler. Solche Fehler treten natürlich nicht auf, wenn die eingelesenen Variablen in Ränge umgerechnet werden, sondern nur dann, wenn abgeleitete statistische Variablen, wie z.B. Residuen, oder selbst neu errechnete Variablen, wie z.B. Variablensummen und -mittelwerte, in Ränge transformiert werden. Ein Beispiel soll das illustrieren:

Angenommen, es werden aus einer Reihe von Variablen mit den Werten -1, 0, 1 mehrere Mittelwerte gebildet, die dann zu einem Gesamtscore zusammengefasst werden. Dabei resultieren für zwei Probanden die folgenden Teilmittelwerte  $1/3$  und  $-1/3$  sowie  $2/3$  bzw.  $-2/3$ , die natürlich nicht als Bruch sondern als Dezimalzahl gespeichert werden:

```
1: 0,6666667 - 0.3333333 - 0.3333333
```

```
2: - 0,6666667 + 0.3333333 + 0.3333333
```

Werden jetzt jeweils die Summen aus den drei Teilmittelwerten gebildet, erhält man:

```
1: 0.0000001
```

```
2: -0.0000001
```

Beide Summen müssten natürlich „theoretisch“ Null sein. Beim „normalen“ Rechnen macht diese Differenz von  $0.0000001$ , die durch Rundungsfehler entsteht, nichts aus, da sie verschwindend klein ist. Anders jedoch, wenn diese Summe in Ränge transformiert wird. Für die beiden o.a. Probanden sind die Summen nicht mehr gleich und erhalten dadurch verschiedene Ränge. Konkret wird dieses Problem häufiger bei den *aligned rank transform*-Tests auftreten (vgl. Kapitel 4.3.2.3), da dort von Residuen Mittelwerte subtrahiert und das Ergebnis in Ränge umgerechnet werden.

In R lässt sich dieses Problem lösen: Dort gibt es die Funktion `round(x, digits=...)`, über die ein Vektor `x` auf die vorgegebene Anzahl von Dezimalstellen gerundet werden kann. In der Regel sollte ein Wert `digits=6` ausreichend sein. `round` muss dann vor der Rangberechnung auf die zu transformierende Variable angewandt werden. Würde man diese Funktion auf die Summe des o.a. Beispiels anwenden, so wären die Summen für beide Probanden Null.

### 3.4 Fehlende Werte

Fehlende Werte (*missing values*), insbesondere der abhängigen Variablen (*Kriterium*), sollten i.a. keine Probleme bereiten, sondern automatisch statistisch sinnvoll von den Programmen behandelt werden. Das funktioniert auch weitgehend so. Allerdings ist dabei zu bedenken, dass bei Messwiederholungen, zumindest bei den hier behandelten Standardmethoden, keine

fehlenden Werte auftreten dürfen.

Bei SPSS werden bei Versuchsplänen mit Messwiederholungen Fälle mit fehlenden Werten automatisch eliminiert.

Dagegen empfiehlt es sich bei der Benutzung von R, im Fall von fehlenden Werten generell vor Durchführung der Varianzanalysen mit der Funktion `na.omit(Dataframe)` eine Teildatenmatrix der in der Analyse verwendeten Variablen (Faktoren und Kriterium) ohne fehlende Werte zu erzeugen. Dies ist ganz besonders in den folgenden Fällen ratsam:

- Die Funktion `ezANOVA` kann nicht mit fehlenden Werten umgehen, auch nicht bei Designs, die keine Messwiederholungen enthalten. Hier empfiehlt sich immer:

```
ezANOVA(na.omit(Dataframe), ...)
```

- Im Fall von fehlenden Werten bei Messwiederholungen müssen in jedem Fall, sowohl bei der Analyse mittels `aov` als auch mittels `ezANOVA` sowie anderer R-Funktionen wie `np.anova` vor der Umstrukturierung der Daten mittels `reshape` entsprechende Fälle (Versuchspersonen) komplett eliminiert werden.
- Bei den nichtparametrischen Analysen ist fast immer eine Rangtransformation erforderlich. Bei der Rangbildung mittels `rank(...)` erhalten standardmäßig (unsinnigerweise) auch fehlende Werte Ränge, nämlich die höchsten Ränge. Mittels des Parameters

```
rank(..., na.last="keep")
```

kann das vermieden werden.

### 3. 5 Beschränkungen

Lediglich die Standardfunktionen bzw. -prozeduren zur parametrischen Varianzanalyse erlauben quasi beliebig viele Gruppierungs- und Messwiederholungsfaktoren. Anders die Funktionen für spezielle nichtparametrische Methoden, bei diesen insbesondere für die Analyse von Versuchsplänen mit Messwiederholungen. Dies betrifft natürlich in erster Linie R, etwa Funktionen für die Methoden von Akritas, Arnold & Brunner (ATS), von Koch, von Welch & James oder die nichtparametrischen multivariaten Analysen für gemischte Versuchspläne. Diese erlauben nur einen Gruppierungs- und einen Messwiederholungsfaktor.

Liegt ein Design mit einem Messwiederholungsfaktor C, aber mehreren Gruppierungsfaktoren vor, etwa A und B, so kann man nacheinander für die Analyse einmal A und C und einmal B und C auswählen. Für die Interaktion A\*B kann ein äquivalentes Verfahren ohne Messwiederholungsfaktoren ausgewählt werden. Lediglich die Interaktion A\*B\*C kann nicht überprüft werden. Allerdings liegen in edr Praxis eher selten Hypothesen für 3-er Interaktionen vor.

Liegt dagegen ein Design mit einem Gruppierungsfaktor A, aber mehreren Messwiederholungsfaktoren vor, etwa C und D, so kann man zwar auch in diesem Fall nacheinander für die Analyse einmal A und C und einmal A und D auswählen. Allerdings ist das ohne ein wenig Vorarbeit nicht möglich, denn es müssen zunächst für jede Stufe des Faktors C die Summen (oder Mittelwerte) über die Stufen des Faktors D, bzw. für die zweite Analyse für jede Stufe des Faktors D die Summen (oder Mittelwerte) über die Stufen des Faktors C errechnet werden. Hier ist es in der Regel leider nicht möglich, die Interaktion der beiden Messwiederholungsfaktoren C und D zu testen.

In SPSS ist diese Summenbildung relativ einfach mittels `compute` und den Funktionen `sum` oder `mean`, z.B.

```
compute s1=sum(v11, v12, v13,...).
compute s2=sum(v21, v22, v23,...).
....
```

wobei anschließend die Analyse, z.B. ein Friedman-Test, über die Variablen `s1`, `s2`,.. erfolgt.

In R muss unterschieden werden, ob die Funktion die Daten im breiten Format oder im langen Format benötigt. Genauer dazu in Kapitel 5.1. Zunächst der Fall für Daten im breiten Format, d.h. alle Werte einer Erhebungseinheit liegen in einer Zeile vor. Hierzu werden (wie oben bei SPSS) die Funktionen `rowSums` oder `rowMeans` verwendet, z.B.

```
within(Dataframe, {S1<-rowSums(v11, v12,...) ;
                    S2<-rowSums(v21, v22,...) ; ...})->Dataframe
```

wobei anschließend die Analyse, z.B. ein multivariater Test, über die Variablen `s1`, `s2`,.. erfolgt.

Wenn die Daten im langen Format vorliegen müssen, d.h. die Werte jeder Messwiederholung liegen als getrennte Zeile vor, ist als Folge der vorangegangenen Umstrukturierung ohnehin eine Fallkennung vorhanden, die hier als `Vpn` angenommen wird. Soll nun für Faktor C die Summen `Sy` der abhängigen Variablen `y` über Faktor D berechnet werden, so lautet die Anweisung:

```
within(Dataframe, Sy<-ave(y, Vpn, C, FUN=sum) )->Dataframe.C
```

Allerdings wird die Summe `Sy` für jede Messwiederholung von D angefügt. D.h. wenn z.B. D 4 Stufen hat, dann hat dieser Dataframe den eigentlich gewünschten genau 4-mal. Also muss einer ausgewählt werden, z.B. über

```
subset(Dataframe.C, D==1) ->Dataframe.C1
```

Die abhängige Variable für die anschließende Analyse ist dann `Sy`. Ein Beispiel dazu gibt es in Abschnitt 11.5.

## 4. Unabhängige Stichproben

Es wird im Folgenden angenommen, dass die Werte einer abhängigen Variablen  $x$  für  $I$  Gruppen mit Stichprobenumfängen  $n_i$  vorliegen. Üblicherweise werden die Gruppen, und damit die Stichproben, über eine Variable, die Gruppierungsvariable definiert. Diese wird i.a. *Gruppierungsfaktor* genannt, im Gegensatz zu den Messwiederholungsfaktoren. Bei mehrfaktoriellen Analysen entsprechend über mehrere Gruppierungsvariablen.

### Beispieldaten 1 (mydata1):

Im Folgenden wird ein Datensatz verwendet, bei dem 2 Patientengruppen (Faktor A: Schizophrene und Depressive, je 9 Personen) jeweils in 3 Gruppen zu 3 Personen eingeteilt werden, die dann jeweils ein Medikament (Faktor B: drugs 1, 2 oder 3) erhalten. Alle Zellen haben daher dieselbe Anzahl Versuchspersonen ( $n=3$ ). Die abhängige Variable ist eine Beurteilung auf einer Skala von 0 bis 19, also quasi metrisch, wenn auch streng genommen als Beurteilung ordinal.

patients	drug 1	drug 2	drug 3
Schizophrene	8 4 0	10 8 6	8 6 4
Depressive	16 12 8	6 4 2	17 14 11

In R wie auch in SPSS werden hierfür die Variablennamen `patients`, `drugs` und `x` verwendet. In R müssen `patients` und `drug` vom Typ „factor“ deklariert sein. In R hat der Dataframe den Namen `mydata1`.

### Beispieldaten 2 (mydata2):

Im Weiteren wird ein Datensatz verwendet, bei dem 2 Patientengruppen (Faktor A: Kontrollgruppe und Behandlungsgruppe) jeweils in 4 Gruppen eingeteilt werden, die dann jeweils ein Medikament (Faktor B: drug 1, 2, 3 oder 4) erhalten. Die Zellenbestutzungszahlen 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` vom Typ „factor“ deklariert sein. In R hat der Dataframe den Namen `mydata2`.

### Beispieldaten 3 (mydata3):

Darüber hinaus wird ein Datensatz verwendet, bei dem wieder 2 Patientengruppen (Faktor A: Kontrollgruppe und Behandlungsgruppe) jeweils in 4 Gruppen eingeteilt werden, die dann jeweils ein Medikament in 4 verschiedenen hohen Dosierungen (Faktor B: dosis 1, 2, 3 oder 4) erhalten. Die Zellenbestutzungszahlen 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`.

## 4.1 Voraussetzungen der parametrischen Varianzanalyse

Vom t-Test her kennt man zwei Voraussetzungen: Erstens müssen die Beobachtungen der abhängigen Variablen  $x$  in beiden Gruppen normalverteilt sein und zweitens müssen die Varianzen beider Gruppen homogen (statistisch gleich) sein. Dies lässt sich noch problemlos von zwei auf beliebig viele  $I$  Gruppen verallgemeinern. (Mit  $I$  wird im Folgenden die Anzahl von Stufen/Gruppen eines unspezifizierten Faktors bezeichnet.) Doch insbesondere die Normalverteilungsvoraussetzung kann auch anders formuliert werden: Die Residuen  $e_{ij}$  müssen normalverteilt sein, wobei sich die Residuen aus dem varianzanalytischen Modell ergeben, hier für den 1-faktoriellen Fall eines Faktors  $A$  mit  $I$  Stufen/Gruppen:

$$x_{im} = \mu + \alpha_i + e_{im} \quad (i=1, \dots, I \text{ und } m=1, \dots, n_i) \quad (4-1)$$

wobei  $\alpha_i = \mu_i - \mu$  die Abweichungen des Gruppenmittelwertes  $\mu_i$  vom Gesamtmittel  $\mu$  sind, der *Effekt* von Faktor  $A$  mit  $I$  Stufen (Gruppen).  $n_i$  bzw.  $n_{ij}$  seien die Zellsensungszahlen und  $N$  deren Summe. Die Aufgabe der Varianzanalyse ist die Prüfung gleicher Gruppenmittelwerte, d.h.  $H_0$  lautet:

$$\mu_1 = \mu_2 = \dots = \mu_I$$

was in der Terminologie des o.a. Modells äquivalent ist zu:

$$\alpha_1 = \alpha_2 = \dots = \alpha_I = 0$$

Es sei darauf aufmerksam gemacht, dass die nichtparametrischen Tests eine andere Nullhypothese  $H_0$  haben: die Verteilungsfunktion  $F_i(x)$  von  $x$  ist in allen Gruppen dieselbe:

$$F_1(x) = F_2(x) = \dots = F_I(x)$$

woraus zu folgern ist, dass bei Ablehnung von  $H_0$  sich nicht notwendigerweise die Mittelwerte  $\mu_i$  unterscheiden, sondern möglicherweise andere Verteilungsparameter wie Varianz oder Schiefe.

Das Modell der 2- oder mehrfaktoriellen Analyse unterscheidet sich kaum von dem 1-faktoriellen, da dieses auch nur eine einzige Residuenvariable  $e_{ijm}$  enthält. Dabei sei  $B$  der zweite Faktor, mit  $J$  Stufen (Gruppen) sowie den Effekten  $\beta_j$ :

$$x_{ijm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijm} \quad (i=1, \dots, I, j=1, \dots, J \text{ und } m=1, \dots, n_{ij}) \quad (4-2)$$

(Auf die Interaktion  $\alpha\beta_{ij}$  wird in Kapitel 4.3.1.2 kurz eingegangen.) Logisch sind zwar beide Bedingungen (Normalität innerhalb jeder Gruppe und Normalität der Residuen) identisch, doch in der Praxis ist es sinnvoll, die Gesamtheit der Residuen zu überprüfen. Weitere Erläuterungen zur Prüfung auf Normalverteilung sind in Kapitel 1.6 zu finden.

Die Varianzhomogenität ( $\sigma_1^2 = \dots = \sigma_I^2$ ) wird zweckmäßigerweise mit dem *Levene-Test* überprüft, da dieser (im Gegensatz zum klassischen F-Test, zum  $F_{\max}$ -Test oder zum Bartlett-Test) robust gegen Abweichungen von der Normalverteilung ist. Einen kompletten Vergleich von 14 Tests zur Prüfung der Varianzhomogenität bietet Shuqiang Zhang (1998). Es ist zu beachten,



dass bei mehrfaktoriellen Analysen für jeden Effektttest andere Varianzen relevant sind. So sind z.B. bei einer 2-faktoriellen Analyse 1. die Varianzen der Gruppen von Faktor A, 2. die Varianzen der Gruppen von Faktor B, und 3. die Zellvarianzen (für die Interaktion A\*B) auf Homogenität zu überprüfen.

Doch was, wenn eine der Voraussetzungen nicht erfüllt ist? Muss dann direkt zur nichtparametrischen Varianzanalyse gegriffen werden? Nein! Die Varianzanalyse ist ein sehr robustes statistisches Testverfahren (vgl. Kapitel 1.1). Hierzu gibt es zahlreiche Untersuchungen, insbesondere solche, die das Verhalten von  $\beta$  (Wahrscheinlichkeit für einen Fehler 2. Art) zum Inhalt haben. Brauchbare Übersichten findet man u.a. bei Field (2009), Bortz (1984) und Ito (1980).

Zunächst einmal zwei generelle positive Aussagen:

- Je größer die Stichproben, desto weniger sind die Voraussetzungen noch relevant. Insbesondere ist nach dem *zentralen Grenzwertsatz* die Normalverteilungsvoraussetzung nur für kleinere Stichproben ( $n_i < 50$ ) relevant (siehe auch weiter unten).
- Bei annähernd gleichgroßen Stichprobenumfängen  $n_i$  (siehe unten) wirken sich weder nicht-normalverteilte Residuen noch (nicht allzu) inhomogene Varianzen störend aus.

Vielfach ist entscheidend, ob die Stichprobenumfänge  $n_i$  gleich oder ungleich sind (vgl. nächster Absatz). Manchmal möchte man gerne ein Auge zudrücken und *annähernd gleiche* Stichprobenumfänge als gleich ansehen, wie dies auch in diesem Text verschiedentlich formuliert wird. Das ist z.B. möglich, wenn die ursprünglich als gleich geplanten  $n_i$  etwa durch fehlende Messungen ungleich geworden sind. Akzeptabel sind z.B. Abweichungen der  $n_i$  um maximal 10-15%,

### Ungleiche Stichprobenumfänge $n_i$ :

Ungleiche  $n_i$  verursachen zunächst einmal bei der Varianzanalyse keine Probleme. Doch sind die Stichprobenumfänge verschieden, gibt es verschiedene Berechnungs- bzw. Lösungsmethoden. Standardmäßig wird die Varianzanalyse auf die Regressionsanalyse, d.h. lineare Modelle zurückgeführt. Dies führt zu *Methoden gewichteter Mittelwerte* (*weighted means solution*). So wird in o.a. Modell (4.1) der Gesamtmittelwert  $\mu$  als gewichtetes Mittel der Gruppenmittelwerte  $\mu_i$  angenommen:

$$\mu = (n_1\mu_1 + \dots + n_I\mu_I) / N$$

d.h. die Gruppen fließen entsprechend ihres Umfangs in die Gesamtberechnungen ein, was sinnvoll ist, wenn die  $I$  Stichproben sich entsprechend den Teilen der Grundgesamtheit verhalten. Darüber hinaus gibt es noch für diese Berechnungsmethode verschiedene Lösungen zur Ermittlung der sog. Streuungsquadrate und damit des Tests eines Effekts, die mit Typ I, II und III bezeichnet werden, deren Unterschiede allerdings erst bei mehrfaktoriellen Analysen zum Tragen kommen. Mehr dazu in Abschnitt 4.3.1.1. Bei diesen Methoden können allerdings Zusammenhänge zwischen den Faktoren und der abhängigen Variablen  $x$  zu "unsauberen" Testergebnissen führen. Dennoch ist dies das üblicherweise verwendete Verfahren, insbesondere das vom Typ III.

Anders die Methode der *ungewichteten Mittelwerte* (*unweighted means solution*). Hier wird der Gesamtmittelwert aus den Gruppenmittelwerten, bzw. im Fall mehrfaktorieller Analysen die Gruppenmittelwerte aus den Zellenmittelwerten, als ungewichtete Mittel errechnet:

$$\mu = (\mu_1 + \dots + \mu_I) / I$$

wodurch die o.a. Probleme gelöst werden. Hierbei wird praktisch eine Analyse mit gleichen Zellenbesetzungszahlen gerechnet, bei der ein einheitliches  $\tilde{n}$  angenommen wird, das als harmonisches Mittel der  $n_i$  berechnet wird (vgl. Winer, 1991). D.h. alle Zellen bzw. Stufen des Designs haben dasselbe Gewicht, während bei den o.a. Methoden sich das Gewicht proportional zu den  $n_i$  verhält. Diese Methode ist allerdings nur dann sinnvoll, wenn die  $n_i$  annähernd gleich sind (siehe o.a. Erläuterung). Erstaunlicherweise werden hierfür in R und SPSS praktisch keine Funktionen angeboten. Lediglich die in Abschnitt 2.9 vorgestellten ATS-Methoden basieren auf der Methode der ungewichteten Mittelwerte. Darüber hinaus haben Brunner et al. (2020) sog. *Pseudo-Ränge* entwickelt (vgl. Abschnitt 2.16.1), eine alternative Methode der Rangberechnung, die im Fall von rangbasierten Varianzanalysen den gleichen Effekt haben.

Wie nachfolgend ausgeführt, haben allerdings ungleiche Stichprobenumfänge einen erheblichen Einfluss im Fall von Heterogenitäten, wie z.B. ungleichen Varianzen oder ungleichen Innergruppenkorrelationen.

### Zur Voraussetzung der Normalverteilung:

(Details sind bei Wilcox (2005), Osborne (2008) sowie Lindman (1974) nachzulesen.)

Wenn im Folgenden von gleichen  $n_i$  die Rede ist, dann ist nicht notwendigerweise die exakte Gleichheit aller  $n_i$  gefordert. Leichte Abweichungen infolge fehlender Angaben sind erlaubt.

- Moderate Abweichungen von der Normalverteilung, z.B. eine Schiefe, führen schlimmstenfalls zu einer leichten Vergrößerung von  $\beta$ . D.h. gegebenenfalls können Unterschiede nicht nachgewiesen werden. Oder positiv ausgedrückt: Signifikante Unterschiede können als gesichert gelten.
- Schmalgipflige, steile Verteilungen, d.h. mit negativem Exzess ([http://de.wikipedia.org/wiki/Wölbung\\_\(Statistik\)](http://de.wikipedia.org/wiki/Wölbung_(Statistik))), machen den F-Test konservativer. Breitgipflige Verteilungen machen dagegen den Test liberaler, können aber auch das  $\alpha$ -Risiko vergrößern, allerdings nur in einem sehr geringen Maß (vgl. Ito, 1980).
- Drastische Abweichungen von der Normalverteilung können zu unbrauchbaren Ergebnissen führen, insbesondere wenn die Stichprobenumfänge  $n_i$  verschieden sind. (Der F-Test kann in solchen Fällen sowohl zu liberal als auch zu konservativ reagieren).
- Box & Andersen (1955) haben einen F-Test entwickelt, der die Abweichung von der Normalverteilung durch eine Korrektur der Freiheitsgrade kompensiert (vgl. Anhang 2.3). Eine entsprechende R-Funktion ist im Anhang 3 zu finden.

### Zur Voraussetzung der Varianzhomogenität:

Dies ist die gravierendere Voraussetzung, verlangt einige Kenntnisse über die Daten und gegebenenfalls besondere Analysemethoden. Viele der hier angeführten Faustregeln sind bei Blanca et al. (2017) zu finden, die sich allerdings nur auf 1-faktorielle Designs beziehen und annähernd normalverteilte Daten voraussetzen.

- Der störende Einfluss inhomogener Varianzen ist umso stärker, je größer die Streuung der Varianzen ist, wie Box (1954) bewiesen hat. Eine gute Abschätzung hierfür bietet der Variationskoeffizient  $c$  der Varianzen, also die Standardabweichung der Varianzen dividiert durch den Mittelwert der Varianzen.

$$c = \sqrt{\frac{1}{I} \sum_i (s_i^2 - \bar{s}^2)^2} / \bar{s}^2 \quad \text{wobei} \quad \bar{s}^2 = \frac{1}{I} \sum_i s_i^2$$

So wirkt sich z.B. die Folge von Varianzen 1:2:3:4 weniger störend aus als die Folge 1:1:1:4.

Zwar ist für beide Folgen der Varianzquotient 4, aber die erste hat einen Variationskoeffizient von 0,52 und die zweite einen Variationskoeffizient von 0,86.

- Gleiche Varianzen  $s_i^2$  führen auch bei ungleichen  $n_i$  zu keinerlei Beeinträchtigungen.
- Entgegen den Aussagen in den meisten Lehrbüchern können ungleiche Varianzen auch bei gleichen  $n_i$  zu einer Erhöhung der Fehlerrate 1. Art führen. Dies wurde bereits von Box (1954) bewiesen wie auch durch viele Simulationsstudien bestätigt (vgl. z.B. Dijkstra, 1987). Allerdings gilt das im Wesentlichen für stark unterschiedliche Varianzen (ab etwa  $\max(s_i^2)/\min(s_i^2) > 4$ ) und sollte in der Praxis ignoriert werden (vgl. Blanca et al., 2017).
- Bei ungleichen  $n_i$  gilt: Haben die großen Stichproben auch die größeren Varianzen (*positive pairing*), reagiert der F-Test konservativ. Haben dagegen die großen Stichproben die kleineren Varianzen (*negative pairing*), reagiert der F-Test liberal (vgl. u.a. Luepsen, 2017, Delacre et al., 2019, und Dijkstra, 1987). Diese Regel gilt in abgeschwächter Form auch für die anderen Tests. Die Stärke des Zusammenhangs von  $s_i^2$  und  $n_i$  (*pairing*) wird üblicherweise über die Korrelation der beiden Größen gemessen, wenn auch Box (1954) hier einen *bias ratio* definiert hat. Beim (unangenehmen) *negative pairing* gilt: Die Ergebnisse sind gültig, solange der Varianzquotient  $< 2$  und  $|r| < 0.5$  ist.
- Mit zunehmender Varianzheterogenität nimmt auch die Fehlerrate 1. Art zu.
- Moderate Abweichungen von der Varianzhomogenität führen ebenfalls schlimmstenfalls zu einer leichten Vergrößerung von  $\beta$ . Allerdings gilt auch hier, dass die Stichprobenumfänge  $n_i$  nicht zu stark divergieren dürfen. Als Faustregel gilt:  $\max(s_i^2)/\min(s_i^2) < 3$  und  $\max(n_i^2)/\min(n_i^2) < 4$ .
- Der gesamte Stichprobenumfang  $N$  spielt bei der Varianzhomogenität keine Rolle.
- Je größer die Gruppenzahl  $I$ , desto robuster ist der F-Test bzgl. Varianzheterogenität.
- Beim parametrischen F-Test sowie bei den rangbasierten nichtparametrischen Tests ist die Interaktion stärker von Varianzinhomogenitäten betroffen als die Haupteffekte.
- Bei der Anwendung rangbasierter nichtparametrischer Verfahren bleibt in der Regel eine Varianzinhomogenität der Rohwerte erhalten (vgl. dazu Fan, 2006 sowie Beasley, 2002), d.h. die Anwendung von nichtparametrischen Varianzanalysen anstatt des parametrischen F-Tests löst nicht das Problem ungleicher Varianzen.
- Im Fall von schiefen Verteilungen können ungleiche Varianzen auch bei den in Kapitel 2.13 aufgeführten Verfahren für heterogene Varianzen zu erhöhten Fehlerraten führen (vgl. Lix et al., 1996 sowie Schneider & Penfield, 1997).
- Korrelieren im Falle inhomogener Varianzen die Zellenmittelwerte mit den -varianzen, nehmen also mit steigenden Zellenmittelwerten auch die Zellvarianzen zu, wird eine Datentransformation der Kriteriumsvariablen  $x$  empfohlen: gute Chancen bieten die einfachen Funktionen  $\sqrt{x}$  und  $\log(x)$ . Die Box-Cox-Transformationen (vgl. Online Statistics Education) perfektionieren diese Idee. Auf der anderen Seite warnen Feng et al. (2014) insbesondere vor der log-Transformation.
- In Kapitel 2.13 sind verschiedene Verfahren speziell für den Fall ungleicher Varianzen aufgeführt, so z.B. die Tests von Welch sowie von Brown & Forsythe. Diese sind für 1-faktorielle Varianzanalysen auch in SPSS enthalten. In R gibt es diese auch 2-faktoriell sowie eine Reihe weiterer Methoden, die robust gegen heterogene Varianzen sind.
- Darüber hinaus hat Box (1954) eine Korrektur (genauer gesagt: Reduzierung) der Freiheitsgrade für den F-Test entwickelt, der die Heterogenität der Varianzen berücksichtigt. Diese erfordert zwar ein wenig Programmieraufwand, ist aber in R realisierbar. Näheres dazu bei

Winer (1991, S. 109, sowie im Anhang 2.1.)

Details sind bei Glass et al. (1972) sowie Osborne (2008) nachzulesen. Eine gute Übersicht, insbesondere der robusten parametrischen Verfahren, ist bei Fan (2006) zu finden. Eine hilfreiche Zusammenstellung der Auswirkungen der Verletzungen von Voraussetzungen sowie alternativer Methoden bietet Dijkstra (1987). Speziell der Einfluss inhomogener Varianzen wird von Lix et al. (1996) ausführlich behandelt, die auch die o.a. Ergebnisse von Box (1954) ausführlich wiedergeben. Blanca et al. (2017) gibt eine Reihe praktischer Empfehlungen.

Neben den beiden o.a. Voraussetzungen gibt es allerdings noch eine dritte: die Unabhängigkeit der Beobachtungen. Diese lässt sich allerdings kaum „testen“, sondern setzt eher eine saubere Versuchsplanung voraus. Dies ist allerdings nicht Thema dieses Skripts. Dennoch ein kleines Beispiel hierzu. Hat man einen Faktor wie „Geschlecht“, so wird man diesen normalerweise als Gruppierungsfaktor mit unabhängigen Stichproben ansehen. Das ist nicht immer so. Werden z.B. Vater und Mutter eines behinderten Kindes unabhängig voneinander befragt, wie sie damit umgehen bzw. welche Auswirkungen dies auf das Zusammenleben hat, so sind die Antworten beider Elternteile nicht mehr unabhängig, da sich diese auf dasselbe Kind beziehen. Hier ist das „Geschlecht“ des antwortenden Elternteils als Messwiederholungsfaktor zu behandeln.

Noch eine Erläuterung zum *zentralen Grenzwertsatz*, nach dem für größere  $n$  insbesondere die Normalverteilungsvoraussetzung vernachlässigt werden kann. Dieser Satz beinhaltet, dass der Mittelwert der Beobachtungen  $x_{im}$  asymptotisch, d.h. für größere  $n$ , normalverteilt ist, egal welche Verteilung die  $x_{im}$  haben. Dies lässt sich leicht verifizieren anhand des Würfel-experiments. Die Augenzahl (1,...,6) ist gleichverteilt. Würfelt man  $n$ -mal und berechnet den Mittelwert der jeweils erzielten Augenzahl, so kann man beobachten, wie diese gegen 3.5 konvergiert. Und ein Histogramm der Mittelwerte zeigt, wie sich der Mittelwert mit zunehmendem  $n$  der Normalverteilung nähert. Bei der Varianzanalyse werden im Wesentlichen Summen und Mittelwerte berechnet, die für große  $n$  normalverteilt sind, daraus Quadrate, die  $\chi^2$ -verteilt sind, und deren Quotienten F-verteilt sind. Je besser die Mittelwerte nun an die Normalverteilung herankommen, desto exakter ist anschließend der F-Test.

### Fazit und generelle Empfehlungen:

- In jedem Fall ist es ratsam, vor Durchführung einer Varianzanalyse sich ein Bild von den  $n_i$  und  $s_i$  zu machen, da diese am stärksten die Auswahl des Verfahrens beeinflussen.
- Einige Autoren raten davon ab, die Varianzhomogenität mit einem Test zu überprüfen, da diese meistens stärkere Voraussetzungen haben als der F-Test selbst (vgl. Blanca, 2017).
- Ist die abhängige Variable metrisch, die Stichprobenumfänge  $n_i$  nicht stark unterschiedlich, die Abweichungen von der Normalverteilung der Residuen wie auch von der Varianzhomogenität moderat, so kann die parametrische Varianzanalyse durchgeführt und die Ergebnisse ohne Einschränkung interpretiert werden. Vgl. dazu auch Kapitel 2.17.
- Im Fall von pairing  $|r| > 0.5$ , insbesondere bei  $|r| > 0.8$ , ist es ratsam, Verfahren für heterogene Varianzen (vgl. Kapitel 2.13) anzuwenden.
- Bei ungleichen  $n_i$  besteht praktisch immer ein (wenn auch kleiner) Zusammenhang zwischen den  $n_i$  und  $s_i$  (*pairing*). Dieser kann auch bei ungleichen, aber nicht signifikant verschiedenen Varianzen zu verfälschten Ergebnissen führen. Daher ist in solchen Fällen eine Varianzanalyse für heterogene Varianzen sofern möglich vorzuziehen.

## Empfehlungen

- Bei gleichen  $n_i$  kann im Normalfall selbst bei ungleichen Varianzen und nichtnormalen Verteilungen bedenkenlos der parametrische F-Test angewandt werden.
- Bei ungleichen  $n_i$  sollte die Varianzhomogenität und ein pairing geprüft werden. Liegt kein Zusammenhang zwischen den  $n_i$  und  $s_i$  vor, kann entweder der von der Waerden-Test (insbesondere bei  $n_i > 10$ ) oder das ART+INT-Verfahren ( $n_i \leq 10$ ) verwendet werden (vgl. Lüpssen, 2016c). Im Falle von pairing wird eines der Verfahren für heterogene Varianzen (siehe unten sowie Kapitel 2.13) empfohlen.

Für die Situation ungleicher Varianzen stehen als 1-faktorielle Analysen sowohl in R als auch in SPSS eine Auswahl von Verfahren zur Verfügung, u.a. die von Welch sowie von Brown & Forsythe, wenn auch die Methoden von James sowie Alexander & Govern vorzuziehen sind. Für mehrfaktorielle Analysen stehen derzeit nur in R Verfahren zur Verfügung, die heterogene Varianzen berücksichtigen: neben dem oben genannten Verfahren von Brown & Forsythe insbesondere die mehrfaktorielle robuste Varianzanalyse von Welch & James sowie die von Brunner, Dette, Munk. Letztere hält zwar den Fehler 1. Art besser unter Kontrolle, gilt allerdings als extrem konservativ (vgl. Richter & Payton, 2003). Daher werden insbesondere SPSS-Benutzer geneigt sein, nach Möglichkeit die parametrische Analyse durchzuführen oder „notfalls“ eines der in den folgenden Kapiteln vorgestellten Verfahren, die sich mit relativ wenig Mühe auch in SPSS durchführen lassen. Dazu sind, je nach Größe der  $n_i$ , die beiden o.a. Methoden von v.d.Waerden und ART+INT noch die am besten geeigneten.

Beispiele zur Prüfung der Voraussetzungen in R bzw. SPSS werden in den nachfolgenden Kapiteln, u.a. 4.3.2 vorgestellt.

## 4. 2 Die 1-faktorielle Varianzanalyse

### 4. 2. 1 Kruskal-Wallis-Test

Eine 1-faktorielle nichtparametrische Varianzanalyse erfolgt üblicherweise über den *Kruskal-Wallis-H-Test*, einer Verallgemeinerung des *Mann-Whitney-U-Tests* von zwei auf beliebig viele Gruppen. Die Logik sieht so aus, dass alle Werte in Ränge transformiert werden, so dass letztlich anstatt der Mittelwerte die mittlere Rangsummen verglichen werden. Für den Test wird ein Wert  $H$  errechnet, der  $\chi^2$ -verteilt ist mit  $(I-1)$  Freiheitsgraden.

Derselbe Test lässt sich auch über eine 1-faktorielle klassische Varianzanalyse der Ränge der abhängigen Variablen durchführen. Dies wird in Abschnitt 4.3.5 ausführlich beschrieben.

#### mit R:

Sollen für den o.a. Datensatz 1 die Reaktionen bzgl. der 3 Medikamente (Faktor `drugs`) verglichen werden, lautet die Anweisung:

```
mydata1 <- within(mydata1, drugs<-factor(drugs))
kruskal.test (x, drugs)
```

mit der Ausgabe

```
Kruskal-Wallis rank sum test

data:  x and drugs
Kruskal-Wallis chi-squared = 2.023, df = 2, p-value = 0.3637
```

was zunächst einmal indiziert, dass die Reaktionen auf die 3 Medikamente sich nicht signifikant unterscheiden.

### mit SPSS:

```
Nptests
/independent test (x) group (drugs) kruskal_wallis (compare=pairwise).
```

mit folgender Ausgabe:

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilung von ist über Kategorien von gleich.	Kruskal-Wallis-Test unabhängiger Stichproben	,364	Nullhypothese behalten.

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.

Gesamtanzahl	18
Teststatistik	2,023
Freiheitsgrade	2
Asymptotische Sig. (zweiseitiger Test)	,364

#### 4.2.2 Varianzanalysen für inhomogene Varianzen

Varianzhomogenität ist nicht nur eine der Voraussetzungen für die „normale“ parametrische Varianzanalyse, sondern ebenso für nichtparametrische Analysen. Zumal durch die meistens angewandten Rangtransformationen sich Streuungsunterschiede nicht notwendigerweise auflösen. Daher ist im Fall von inhomogenen Varianzen der Griff zu nichtparametrischen Verfahren nicht unbedingt hilfreich, zumal es eine Reihe Methoden gibt, die mit diesem Fall gut umgehen können. Die bekanntesten sind die Tests von Welch bzw. von Brown & Forsythe, wovon letzterer der neuere und bessere ist. Allerdings sollte er nicht mit dem gleichnamigen Test zur Prüfung der Varianzhomogenität verwechselt werden. Trivialerweise dürfen diese Tests natürlich auch angewandt werden, wenn die Varianzen homogen sind. Im Falle von exakt gleichen Varianzen sind die F-Werte dieser Tests mit dem „normalen“ F-Test identisch, so dass es durchaus angebracht ist, diese Tests immer als 1-faktorielle Varianzanalyse zu benutzen. Beide Tests sind in R (Welch standardmäßig bzw. Brown & Forsythe im Paket `onewaytests`) und SPSS verfügbar. Allerdings gelten die Tests von Alexander & Govern sowie von James als die besseren, sind aber nur in R (Paket `onewaytests`) vorhanden. Darüber hinaus werden eine Reihe weiterer Tests im R-Paket `doex` angeboten. Anzumerken ist noch, dass die nichtganzzahligen Freiheitsgrade typisch für solche Tests sind, die keine Varianzhomogenität voraussetzen. Weitere Tests für ungleiche Varianzen mit Beispielen folgen in Kapitel 4.3.3.

Für das nachfolgende Beispiel wird der Beispieldatensatz 3 benutzt und dort einfaktoriel der Faktor `dosis` untersucht.

### mit R:

Zunächst die Prüfung der Varianzhomogenität mittels des Levene-Tests:

```
leveneTest(x~dosis, center=mean, data=mydata3)
```

```

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

```

Infolge der stark signifikanten Inhomogenität ist anstatt des normalen F-Tests ein dafür geeigneter robuster F-Test zu wählen. Der Welch-Test ist durchführbar über die Funktion `oneway.test`, der Brown & Forsythe-Test über die Funktion `bf.test`. Für die Variable `x` aus dem Beispieldatensatz 3 mit dem Faktor `dosis` lauten die Anweisungen:

```

oneway.test(x~dosis, mydata3)
library(onewaytests)
bf.test(mydata3$x, mydata3$dosis)

```

```

      One-way analysis of means (not assuming equal variances)

data:  x and dosis
F = 3.8789, num df = 3.000, denom df = 13.308, p-value = 0.03433

      Brown-Forsythe Test
data:  y vs group
F = 3.2177, num df = 3.000, denom df = 18.618, p-value = 0.04655

```

Die beiden p-Werte mit 0,034 bzw. 0,047 belegen, dass die Dosis eine Wirkung zeigt. Abschließend noch die Tests von Alexander & Govern (`ag.test`) sowie von James (`james.test`):

```

library(onewaytests)
ag.test(mydata3$x, mydata3$dosis)
james.test(mydata3$x, mydata3$dosis)

```

```

      Alexander-Govern Test

data:  y vs group
X-squared = 8.7822, df = 3, p-value = 0.03233

      James's Second-Order Test

data:  y vs group
Jtest = 12.803, CriticalValue = 11.323

```

Beim James-Test wird die Testgröße (`Jtest`) zusammen mit dem kritischen Wert ausgegeben. Der Vergleich zeigt, dass dieser Test auch eine Signifikanz anzeigt.

### mit SPSS:

Beide Tests sind durchführbar über `Oneway` (Menü: Mittelwerte vergleichen -> Einfaktorielle Anova). Allerdings müssen die robusten Tests über die „Optionen“ angefordert werden. Für die Variable `x` aus dem Beispieldatensatz 3 mit dem Faktor `dosis` lautet die Syntax:

```

Oneway x by dosis
  /statistics homogeneity brownforsythe welch.

```

In der Ausgabe erscheint nach dem Test auf Homogenität der Varianzen zunächst das Ergebnis für homogene Varianzen:

Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
4,965	3	29	,007

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	45,672	3	15,224	3,130	,041
Innerhalb der Gruppen	141,056	29	4,864		
Gesamt	186,727	32			

danach die Tests für beliebige Varianzen, die hier sogar eine größere Signifikanz als der „normale“ F-Test zeigen, was häufig vorkommt, wenn Voraussetzungen des „normalen“ Tests nicht erfüllt sind.

Robuste Testverfahren zur Prüfung auf Gleichheit der Mittelwerte				
	Statistik <sup>a</sup>	df1	df2	Sig.
Welch-Test	3,879	3	13,308	,034
Brown-Forsythe	3,218	3	18,618	,047

#### 4. 2. 3 Verfahren für nichtnormalverteilte Variablen

Wegen der großen Robustheit der Varianzanalyse hinsichtlich Abweichungen der Residuen von der Normalverteilung gibt es nur wenige Verfahren speziell für metrische nichtnormalverteilte abhängige Variablen. Auf zwei soll hier kurz eingegangen werden.

Zum einen gibt es einen modifizierten F-Test von Box & Andersen (1955) (vgl. auch Anhang 2.3), bei dem sich die Abweichung von der Normalverteilung in der Korrektur der Freiheitsgrade widerspiegelt, wie dies üblicherweise auch bei den entsprechenden modifizierten F-Tests für heterogene Varianzen der Fall ist. Dieses Verfahren macht z.B. Sinn bei extrem schiefen Verteilungen. Eine entsprechende R-Funktion ist im Anhang 3 zu finden.

Erceg-Hurn & Miroseovich (2008) erinnern an die Methode der *Winsorisierung* (oder auch *Trimmen*), die relativ selten angewandt wird, weil sie den Verdacht der Datenmanipulation aufkommen lässt, die aber statistisch durchaus sinnvoll ist. Hierbei werden ein fester Prozentsatz der größten und kleinsten Werte einer Variablen durch die nächstkleinere bzw. durch die nächstgrößere ersetzt. Häufig ersetzt man jeweils 5% der Werte, bei kleineren Stichproben auch jeweils 10% , am oberen Ende durch den nächstkleineren Wert sowie 5% bzw. 10% der Werte am unteren Ende durch den nächstgrößeren Wert. Dieses Verfahren ist sinnvoll insbesondere beim Vorliegen von Ausreißern. R bietet dazu die Funktion `winsorize` im Paket `DescTools`.

#### 4. 2. 4 Weitere Verfahren

Die nachfolgend für die 2-faktorielle Varianzanalyse beschriebenen Rank transform Tests (RT), normal scores-Test (INT) und van der Waerden-Tests sind ebenso als 1-faktorielle Analyse einsetzbar. Dagegen macht das ART-Verfahren nur im mehrfaktoriellen Design Sinn. Die ATS von Akritsis & Co ist als 1-faktorielle Analyse nicht bekannt.



### 4.3 Die 2-faktorielle Varianzanalyse

Bevor die einzelnen Methoden, von der parametrischen Analyse inklusive Prüfung der Voraussetzungen bis zu den verschiedenen nichtparametrischen Methoden, im Detail besprochen werden, sollen zunächst noch ein paar grundlegende Eigenschaften der mehrfaktoriellen Varianzanalyse erwähnt werden. Leser, die schon Erfahrungen auf dem Gebiet der Anova haben, werden damit schon vertraut sein.

#### 4.3.1 Anmerkungen zur 2-faktoriellen Varianzanalyse

##### 4.3.1.1 Balancierte und nichtbalancierte Versuchspläne

Man unterscheidet zwischen *balancierten* (engl. *balanced*) und *nichtbalancierten* (engl. *unbalanced*) Versuchsplänen bzw. Zellenbesetzungszahlen. Bei balancierten Versuchsplänen sind die Zellenbestetzungszahlen zeilenweise oder spaltenweise proportional zueinander, z.B. bei einem Versuchsplan mit den Faktoren A (4 Stufen) und B (3 Stufen)

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	10	12	16
A <sub>2</sub>	15	18	24
A <sub>3</sub>	20	24	32
A <sub>4</sub>	10	12	16

In diesem Beispiel sind die Zellenbesetzungszahlen der 2. bzw. 3. Spalte das 1,2-fache bzw 1,6-fache der 1. Spalte. Umgekehrt kann man auch erkennen, dass die Zellenbesetzungszahlen der 2. bzw. 3. Zeile das 1,5-fache bzw. das 2-fache der ersten Zeile sind.

Versuchspläne mit gleichen Zellenbesetzungszahlen sind natürlich immer balanciert. Solche, bei denen die o.a. Proportionalität nicht zutrifft, sind nichtbalanciert. Diese Unterscheidung ist insofern relevant, als dass die Lösung für die 2- und mehrfaktorielle Varianzanalyse, d.h. die Berechnung der durch die einzelnen Faktoren bzw. Effekte erklärten Streuungen, bei nichtbalancierten Versuchsplänen nicht mehr eindeutig ist. Es gibt mehrere Schätzmethoden: Typ I, Typ II und Typ III, von denen die *Resgressionsmethode der kleinsten Quadrate* (LS), auch mit *Schätzungen vom Typ III* bezeichnet, die gebräuchlichste und unproblematischste ist. Bei ihr werden alle Effekte gleichermaßen berücksichtigt. Dagegen werden bei der Schätzmethode vom Typ I die Effekte in der Reihenfolge ihrer Spezifikation (in der Anweisung) ermittelt, so dass die zuerst aufgeführten Effekte vergleichsweise stärker ausfallen. Bei der Schätzmethode vom Typ II werden die Haupteffekte um die anderen Haupteffekte bereinigt. Eine ausführliche Erläuterung der Unterschiede bieten Smith & Cribbie (2014).

Darüber hinaus sei noch an die in Abschnitt 4.1 erwähnte Methode der ungewichteten Mittel (*unweighted means solution*) erinnert.

##### 4.3.1.2 Die Interaktion

Soll der Einfluss zweier Einflussfaktoren A und B auf eine abhängige Variable x untersucht werden, so bringen zwei 1-faktorielle Varianzanalysen der Faktoren A und B nur die halbe Wahrheit hervor, mitunter sogar irreführende Ergebnisse. Neben den sog. *Haupteffekten* der Faktoren A und B, dem Einfluss von A bzw. B ohne Berücksichtigung des jeweils anderen Fak-

tors, gibt es einen sog. *Interaktionseffekt*  $A*B$ , auch *Wechselwirkung* genannt. Dieser zeigt an, ob der Einfluss von A von B abhängig ist, und umgekehrt, ob der Einfluss von B von A abhängig ist. So kann es durchaus vorkommen, dass die Haupteffekte A und B nicht signifikant sind, dafür aber  $A*B$ . Dies besagt, dass ein Einfluss von A vorhanden ist, der je nach Gruppe (Stufe) des Faktors B unterschiedlich ausfällt, und umgekehrt, dass ein Einfluss von B vorhanden ist, der je nach Gruppe (Stufe) des Faktors A unterschiedlich ausfällt. In der Praxis heißt das, dass häufig der Einfluss eines Faktors erst dadurch zu Tage tritt, dass dieser in Zusammenhang mit einer anderen Einflussgröße analysiert wird.

Im mathematischen Modell für die 2-faktorielle Varianzanalyse

$$x_{ijm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijm} \quad (4-3)$$

erscheint die Interaktion  $\alpha\beta_{ij}$  als eine weitere erklärende Komponente von  $x$ , neben den Anteilen  $\alpha_i$ , den durch Faktor A erklärten Abweichungen ( $\mu - \mu_{Ai}$ ), sowie den  $\beta_j$ , den durch Faktor B erklärten Abweichungen ( $\mu - \mu_{Bj}$ ). Während die Haupteffekte für A und B die Hypothesen

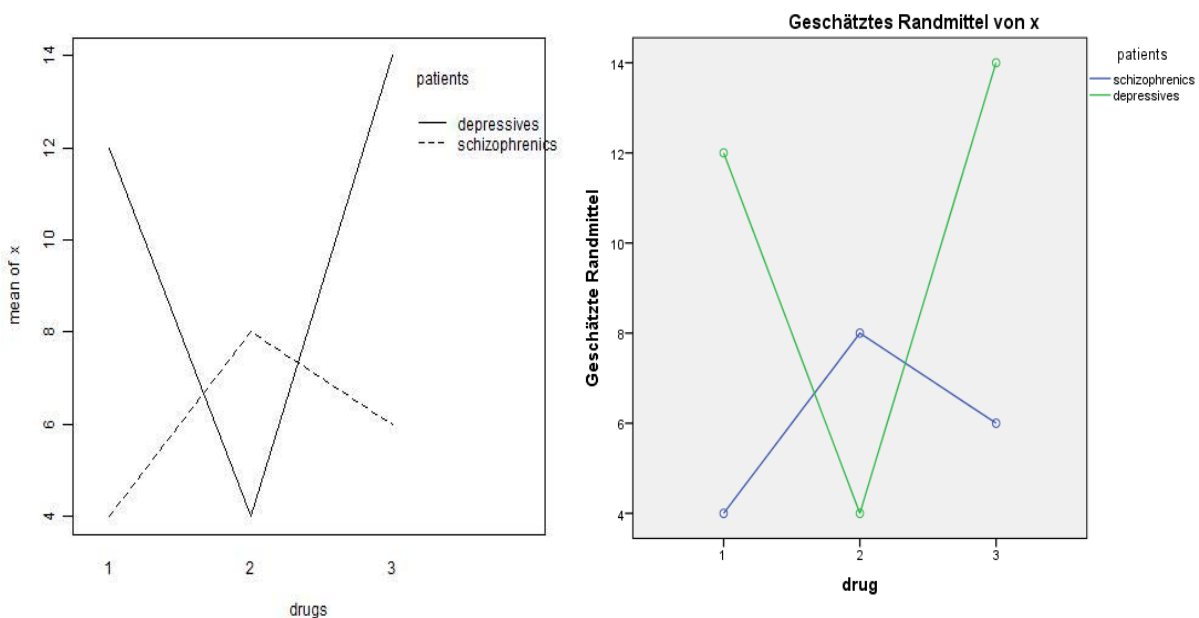
$$H_A: \alpha_i = 0 \text{ für } i=1,\dots,I \text{ (entspricht } \mu_{A1} = \mu_{A2} = \dots = \mu_{AI})$$

$$H_B: \beta_j = 0 \text{ für } j=1,\dots,J \text{ (entspricht } \mu_{B1} = \mu_{B2} = \dots = \mu_{BJ})$$

testen, wird über die Interaktion  $A*B$  die folgende Hypothese geprüft:

$$H_{AB}: \alpha\beta_{ij} = 0 \text{ für } i=1,\dots,I \text{ und } j=1,\dots,J$$

d.h. sowohl die durch A erklärten Abweichungen  $\alpha_i$  sind für alle Stufen von B gleich groß als auch die durch B erklärten Abweichungen  $\beta_j$  sind für alle Stufen von A gleich groß.



*Interaktionsplot für den o.a. Datensatz: links mit R erstellt, rechts mit SPSS*

Dies lässt sich grafisch durch einen sog. *Interaktionsplot* (in SPSS *Profilplot* genannt) veranschaulichen. Dort werden Mittelwertlinien des Faktors A getrennt für die Stufen des Faktors B gezeichnet. Ein nicht paralleler Verlauf der Kurven deutet auf eine signifikante Interaktion hin. Dies kann zum einen sein: Der Einfluss von A ist unterschiedlich stark für die Gruppen von B, oder der Einfluss von A ist für die Gruppen von B gegensätzlich. Bei der 2-faktoriellen Varianz-

analyse lassen sich zwei solcher Plots erstellen: einmal erscheinen die Stufen von A auf der x-Achse und die Stufen von B als verschiedene Linien und einmal erscheinen die Stufen von B auf der x-Achse und die Stufen von A als Linien. Welches nun der aussagekräftigere Plot ist, muss individuell entschieden werden.

#### mit R

```
interaction.plot (factor1, factor2, x)
```

wobei die Variablen *factor1*, *factor2* vom Typ „factor“ sein müssen.

#### mit SPSS

In SPSS ist der Interaktionsplot erhältlich über die parametrische Varianzanalyse (Analysieren -> Allg. lineare Modell -> univariat -> Diagramme)

### 4.3.1.3 Reduzierung des statistischen Fehlers

Die folgenden Ausführungen gelten in erster Linie für die parametrische Varianzanalyse sowie für die anderen Verfahren, bei denen die klassische Aufspaltung der Gesamtstreuung in Effekt- und Residuenstreuung vorgenommen wird. Das sind neben den robusten Verfahren für heterogene Varianzen in erster Linie die oben erwähnten Rank transform Tests (RT, INT, ART und ART+INT). Ferner gilt das Folgende ausschließlich für Versuchspläne mit Gruppierungsfaktoren und bei gemischten Versuchsplänen für die Tests der Messwiederholungsfaktoren.

Neben der Analyse der Wechselwirkung bringt die 2-faktorielle Analyse einen weiteren Gewinn gegenüber zwei 1-faktoriellen Analysen: Durch die Hinzunahme eines weiteren Einflussfaktors kann ein weiterer Anteil der Streuung von x erklärt werden. Die statistischen Tests der Faktoren erfolgen über F-Tests mit einem F-Wert, bei dem im Nenner die Residuenstreuung, die Reststreuung, erscheint. Wird letztere nun reduziert, vergrößert sich der F-Wert und damit verkleinert sich der daraus errechnete p-Wert, was eine höhere Signifikanz bedeutet.

Ausnahme: Falls ein hinzugenommener Faktors keinen Einfluss hat, auch nicht über die Interaktion, und keine zusätzliche Streuung erklärt, sollte dieser weggelassen werden. Denn der Haupteffekt sowie die Interaktion des hinzugenommenen Faktors beanspruchen Freiheitsgrade, die von denen der Residuenstreuung abgezogen werden. Und dadurch fallen die Tests für die anderen Effekte schlechter aus. Ob ein Faktor nun Teil eines Anova-Modells sein sollte oder nicht, muss der Untersuchende aufgrund der vorliegenden Hypothesen entscheiden.

Was hier für die Interaktion der 2-faktoriellen Varianzanalyse gesagt wurde, gilt analog für höhere Interaktionen bei der 3- und mehrfaktoriellen Analyse. Mit einem Unterschied: 3-fach und höhere Interaktionen sind zum einen sehr schwer zu interpretieren, sind aber (zum Glück) in der Praxis seltener signifikant. Daher werden diese in der Regel nicht in die Modelle einbezogen.

### 4.3.1.4 Interpretation der Ergebnisse

Zunächst einmal besteht das Ergebnis einer 2-faktoriellen Varianzanalyse aus 3 Testergebnissen: für die Haupteffekte A und B sowie für die Interaktion A\*B. Um diese richtig zu interpretieren, ist es wichtig, zuerst mit dem Interaktionseffekt zu beginnen.

Ist die Interaktion nicht signifikant, reduziert sich das o.a. Modell (4-3) auf

$$x_{ijm} = \mu + \alpha_i + \beta_j + e_{ijm}$$

d.h. der Effekt von A  $\alpha_i$  ist derselbe für alle Stufen j von B. Gleichmaßen ist der Effekt von

$B\beta_i$  derselbe für alle Stufen  $i$  von A. Ist z.B. A=Geschlecht und B=Behandlung, dann hieße das: der Unterschied zwischen Männern und Frauen ist für alle Behandlungsstufen gleich groß. Ob nun A und B einen Einfluss haben, zeigen die Tests für die Haupteffekte A und B an. Ist ein Haupeffekt signifikant und hat der Faktor mehr als 2 Stufen, so kann man über multiple Mittelwertvergleiche detailliert prüfen, zwischen welchen Stufen Unterschiede bestehen. Dies ist ausführlich in einem anderen Skript „*Multiple Mittelwertvergleiche - parametrisch und nicht-parametrisch*“ (Lüpsen, 2014) beschrieben.

Ist die Interaktion allerdings signifikant, so sind die o.a. Schlüsse falsch. Denn die Interaktion besagt dann, dass sowohl der Effekt von Faktor A für die einzelnen Stufen von Faktor B unterschiedlich ausfällt als auch der Effekt von Faktor B für die einzelnen Stufen von Faktor A. So könnte in obigem Beispiel entweder die Differenz zwischen Männern und Frauen für eine Behandlungsstufe einmal positiv, für eine andere dagegen negativ ausfallen, oder diese Differenz kann in den einzelnen Behandlungsstufen unterschiedlich hoch ausfallen. Damit erübrigt sich auch eine Interpretation der Haupteffekte A und B. In diesem Fall ist die Analyse der sog. *simple effects* (*einfache Effekte*) erforderlich (im Gegensatz zu den „normalen“ *overall effects*, die die eingangs angeführten Ergebnisse liefern). Mehr dazu in Kapitel 10.

### 4.3.2 Das parametrische Verfahren und Prüfung der Voraussetzungen

Zum Vergleich seien die Ergebnisse für die parametrische Analyse vorangestellt sowie die Tests auf Normalverteilung und Homogenität der Varianzen, und zwar zunächst für die Beispieldaten 1 mit einem balancierten Versuchsplan. Anschließend folgt jeweils die Analyse für die Beispieldaten 2 mit einem unbalancierten Design:

#### mit R:

Da hier ein balancierter Versuchsplan ausgewertet wird, kann die in Kapitel 3.1 angeführte `drop1`-Anweisung entfallen. Anweisungen und Ergebnis:

```
mydata1 <- within(mydata1, {drugs<-factor(drugs);
                        patients<-factor(patients)})
aov1 <- aov(x~patients*drugs, mydata)
summary(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
patients	1	72	72.00	8.151	0.01449	*
drugs	2	48	24.00	2.717	0.10634	
patients:drugs	2	144	72.00	8.151	0.00581	**
Residuals	12	106	8.83			

Tabelle 4-1

Zur Prüfung der Normalverteilung der Residuen können diese aus dem Anova-Ergebnis über `aov1$residuals` gewonnen werden. Der Shapiro-Wilk-Test und der Levene-Test zur Prüfung der Homogenität der Varianzen können über folgende Anweisungen erfolgen:

```
library(car)
shapiro.test(aov1$residuals)
leveneTest(x~patients*drugs, data=mydata1, center=mean)
leveneTest(x~patients, data=mydata1, center=mean)
leveneTest(x~drugs, data=mydata1, center=mean)
```

mit folgender Ausgabe:

```
Shapiro-Wilk normality test
data:  aov1$residuals
W = 0.9372, p-value = 0.2592

Levene's Test for Homogeneity of Variance (center = mean)
  Df F value Pr(>F)
group 5  0.4377  0.814
     12
```

```
Levene's Test for Homogeneity of Variance (center = mean)
  Df F value  Pr(>F)
group 1  4.6921 0.04575 *
     16

Levene's Test for Homogeneity of Variance (center = mean)
  Df F value Pr(>F)
group 2  1.1321 0.3484
     15
```

Welche Schlüsse hieraus zu ziehen sind, ist weiter unten erläutert.

Nachfolgend nun die Varianzanalyse für die Beispieldaten 2. Da es sich dabei nicht um einen balancierten Versuchsplan handelt, weichen die erforderlichen Kommandos von den oben aufgeführten etwas ab.

```
options (contrasts=c("contr.sum", "contr.poly"))
mydata2 <- within(mydata2, {drugs<-factor(drugs);
                        group<-factor(group) })
aov2 <- aov(x~group*drugs, mydata2)
drop1(aov2, ~. , test="F")
```

mit dem Ergebnis:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			40.917	23.096			
group	1	12.024	52.940	29.598	7.3464	0.0119656	*
drugs	3	46.560	87.477	42.171	9.4827	0.0002319	***
group:drugs	3	17.932	58.848	29.089	3.6521	0.0260399	*

*Tabelle 4-2*

Auch hier werden die mit den 3 Tests korrespondierenden Varianzen auf Homogenität überprüft:

```
leveneTest(x~group*drugs, mydata2)
leveneTest(x~group, mydata2)
leveneTest(x~drugs, mydata2)
```

```

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 7  0.8608 0.5498
    25
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1  5.6149 0.02422 *
    31
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  0.1115 0.9527
    29

```

Welche Schlüsse hieraus zu ziehen sind, ist weiter unten erläutert.

An dieser Stelle soll noch auf eine weitere Funktion zur parametrischen Varianzanalyse hingewiesen werden: `aov_ez` im Paket `afex`. Ein Vorteil: unbalancierte Versuchspläne müssen nicht gesondert behandelt werden. Ein Nachteil: auch bei Analysen ohne Messwiederholungen muss eine Fallkennung vorhanden sein. Ein Beispiel mit dem Datensatz `mydata2`, wobei eine Kennung (`Vpn`) mit den Werten 1,...,33 zuvor noch angefügt wird:

```

library(afex)
mydata2 <- cbind(mydata2, Vpn=1:33)
aov_ez(data=mydata2, id="Vpn", dv="x", between=c("group", "drugs"))

```

```

Anova Table (Type 3 tests)

Response: x
      Effect      df  MSE      F ges p.value
1      group 1, 25 1.64    7.35 * .227    .012
2      drugs 3, 25 1.64   9.48 *** .532    <.001
3 group:drugs 3, 25 1.64    3.65 * .305    .026

```

Alternativer Aufruf:

```
aov_car(x~group*drugs+Error(Vpn), data=mydata2)
```

Die Residuen erhält man über `residuals(...)` angewandt auf das Ergebnisobjekt.

### mit SPSS:

Die Prüfung der Voraussetzungen, d.h. die Analyse der Residuen sowie der Varianzhomogenität, sollte schon bei der Durchführung der Varianzanalyse berücksichtigt werden, indem sowohl unter „Speichern“ die Residuen (z.B. „standardisiert“) als zusätzliche Variable angefordert werden und unter „Optionen“ der Homogenitätstest angefordert wird. Allerdings werden bei `Unianova` nur die Varianzen auf Gleichheit geprüft, die für die Interaktion relevant sind. Die Prüfung für die beiden Haupteffekte muss zusätzlich angefordert werden, z.B. mittels `Oneway`. Die Syntax dafür und die Varianzanalysetabelle:

```

Unianova x by patients drugs
  /save = zresid
  /print = homogeneity
  /design = patients drugs patients*drugs.
Oneway x by patients
  /statistics homogeneity.

```

```
Oneway x by drugs
  /statistics homogeneity.
```

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	264,000 <sup>a</sup>	5	52,800	5,977	,005
Konstanter Term	1152,000	1	1152,000	130,415	,000
patients	72,000	1	72,000	8,151	,014
drugs	48,000	2	24,000	2,717	,106
patients * drugs	144,000	2	72,000	8,151	,006
Fehler	106,000	12	8,833		
Gesamt	1522,000	18			
Korrigierte Gesamtvariation	370,000	17			

Tabelle 4-3

mit der Prüfung der Varianzhomogenität bzgl. der Interaktion:

Levene-Test auf Gleichheit der Fehlervarianzen <sup>a</sup>			
F	df1	df2	Sig.
,438	5	12	,814

sowie der Prüfung der Varianzhomogenität bzgl. der beiden Haupteffekte:

Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
4,692	1	16	,046

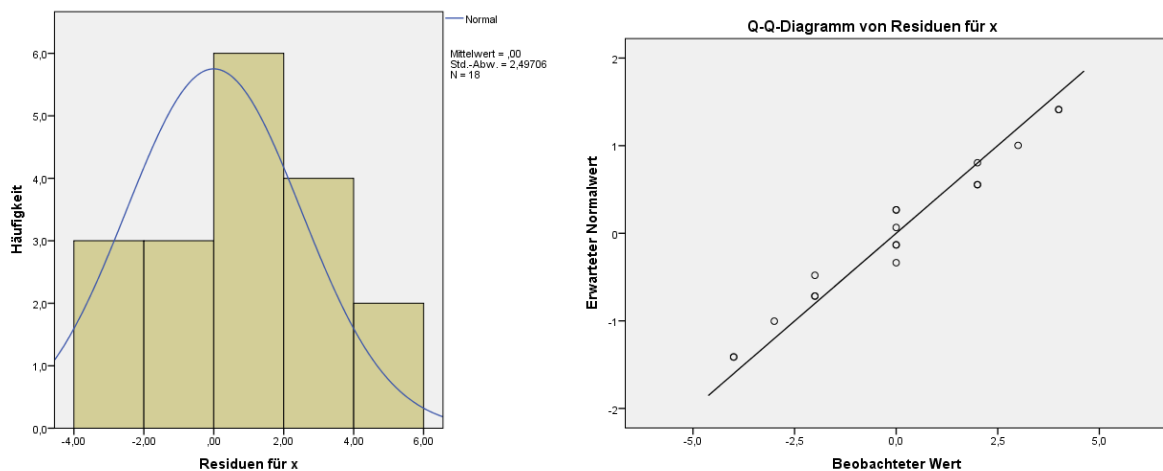
Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
1,132	2	15	,348

Welche Schlüsse hieraus zu ziehen sind, ist weiter unten erläutert.

Die Prüfung der Residuen auf Normalverteilung muss anschließend gesondert vorgenommen werden. Z.B. grafisch mittels eines Histogramms der in der Varianzanalyse erzeugten Residuenvariablen (RES\_1) oder mittels des Shapiro-Wilks-Tests. Beides zusammen kann man über das Menü „Deskriptive Statistiken -> Explorative Datenanalyse“ erzeugen. Die SPSS-Syntax dazu:

```
Examine variables=RES_1
  /plot histogram npplot.
```

Zur besseren Interpretation des Histogramms sollte allerdings die Intervallzahl auf ca.  $\sqrt{n}$  geändert werde, d.h. in diesem Fall bei  $n=18$  auf maximal 5 Intervalle. Der Zusatz npplot führt zu einem *normal probability plot* oder *Q-Q-Diagramm* (vgl. auch Kapitel 1.6). Beide zeigen keine deutlichen Abweichungen von der Normalverteilung.



Histogramm und normal probability plot für die Residuen aus dem Datensatz mydata1.

Standardmäßig werden auch zwei Tests auf Normalverteilung ausgegeben: der klassische Kolmogorov-Smirnov- und der etwas modernere Shapiro-Wilk-Test, die hier ebenfalls keine Abweichungen von der Normalverteilung anzeigen:

Tests auf Normalverteilung						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Residuen für x	,167	18	,200*	,937	18	,259

Nachfolgend nun noch die Varianzanalyse für die Beispieldaten 2:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	69,265 <sup>a</sup>	7	9,895	6,046	,000
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 4-4

sowie die Ergebnisse der 3 Tests auf Varianzhomogenität:

Levene-Test auf Gleichheit der Fehlervarianzen <sup>a</sup>			
F	df1	df2	Sig.
1,251	7	25	,314

Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
9,489	1	31	,004

Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
,115	3	29	,951



Zu den Schlüssen, die aus der Prüfung der Voraussetzungen zu ziehen sind: Für die Beispieldaten 1 (mydata1) ist zwar die Voraussetzung der normalverteilten Residuen erfüllt, allerdings die der Homogenität der Varianzen nur teilweise. Während der erste Test (für die Interaktion) mit  $p=0.814$  und der dritte Test (für Faktor `drugs`) mit  $p=0.3484$  nicht signifikant sind, zeigt der zweite Test (für Faktor `patients`) mit  $p=0.04575$  eine leichte Varianzheterogenität an. Da aber einerseits die  $n_i$  gleich sind und andererseits das Varianzverhältnis 28/9 bei 3 liegt, ist eine spezielle Analyse für ungleiche Varianzen nicht erforderlich, d.h. die Ergebnisse der parametrischen Analyse können als gültig angesehen werden. Bei den Beispieldaten 2 liegen ähnliche Ergebnisse vor: Während der erste Test (für die Interaktion) mit  $p=0.5498$  und der dritte Test (für Faktor `drugs`) mit  $p=0.9527$  nicht signifikant sind, zeigt der zweite Test (für Faktor `group`) mit  $p=0.02422$  Varianzheterogenität an. Nur, hier sind die  $n_i$  ungleich. Daher muss der Test für Faktor `group` mit einem Verfahren durchgeführt werden, das robust gegen ungleiche Varianzen ist, im einfachsten Fall dem robusten t-Test in der Version von Welch. Dieser liefert ein  $p=0.107$ . Ein Nachteil: Wie in 4.3.1.3 erläutert kann durch den 1-faktoriellen Test Effizienz gegenüber einer 2-faktoriellen Varianzanalyse verloren gehen. Für R gibt es die Alternative, anstatt des t-Tests die 2-faktoriellen Verfahren von Brown & Forsythe oder Welch & James anzuwenden (siehe nächster Abschnitt).

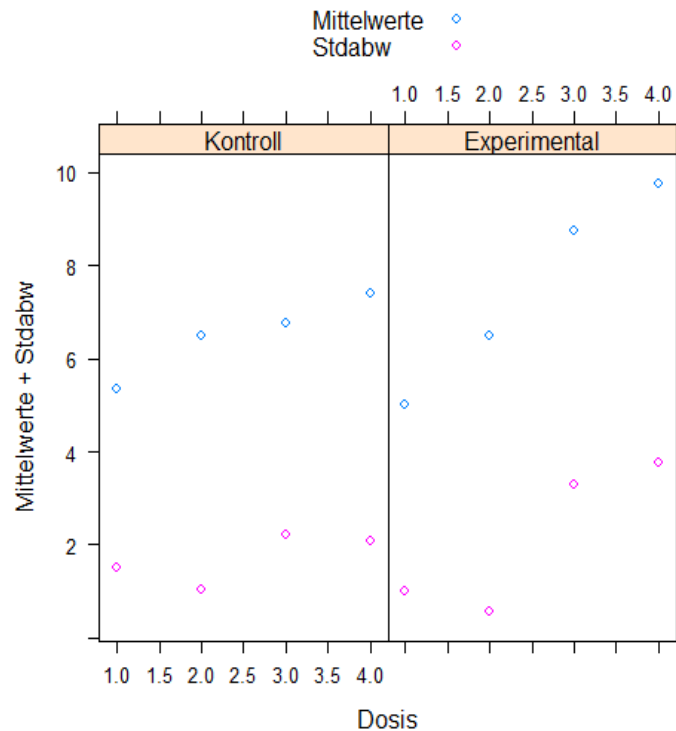
An dieser Stelle soll noch einmal auf die Ausführungen des Kapitels 4.3.1.3 zurückgekommen werden. Dort war darauf hingewiesen worden, dass durch die Hinzunahme eines Faktors häufig der statistische Fehler reduziert werden kann und Effekte erst bei mehrfaktoriellen Analysen als signifikant nachgewiesen werden können. Aus Tabelle 4-3 (Beispieldaten 1) konnten signifikante Effekte für den Faktor `patients` ( $p=0,014$ ) sowie für die Interaktion ( $p=0,006$ ) abgelesen werden. Würde man nur 1-faktorielle Analysen durchführen, so erhielte man keine Signifikanzen, abgesehen davon, dass Interaktionen ohnehin nur mehrfaktoriell erkennbar sind. Hier die Ergebnisse mit SPSS:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
patients	72,000	1	72,000	3,866	,067
Fehler	298,000	16	18,625		

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
drugs	48,000	2	24,000	1,118	,353
Fehler	322,000	15	21,467		

### 4.3.3 Varianzanalysen für inhomogene Varianzen

Für mehrfaktorielle Versuchspläne gibt es leider nur wenige robuste F-Tests speziell für heterogene Varianzen. In Kapitel 2.13 waren einige Verfahren vorgestellt worden, von denen allerdings keines in SPSS verfügbar ist. Der dort vorgestellte Test von Brown & Forsythe (vgl. auch Anhang 2.2) ist vermutlich der bekannteste, wenn auch die Verbesserung durch Mehrotra nicht so verbreitet ist, während der Test von Welch & James weitgehend unbekannt ist. An dieser Stelle sollte auch der Test von Brunner, Dette und Munk, auch *BDM-Test* genannt, erwähnt werden. Eigentlich ist er ein nichtparametrischer Test und als Alternative zum Kruskal-Wallis-Test für den Fall stark inhomogener Varianzen gedacht.



Für den Datensatz 3 zeigt die obige Grafik, dass bei diesem tatsächlich die Varianzen mit den Mittelwerten ansteigen. Der Levene-Test auf Varianzhomogenität zeigt übrigens mit einem  $p=0,012$  einen relativ starken Unterschied der Zellvarianzen. Und da zugleich die Zellenbesetzungszahlen stark schwanken, von 3 bis 6, ist hier eine besondere Behandlung erforderlich.

Verschiedentlich werden für den Fall inhomogener Varianzen auch die Rangtransformation empfohlen, also Anwendung des RT-Verfahrens. Wie in Kapitel 2.12 dargelegt, kann diese Methode zum „Erfolg“ führen, muss es aber nicht. Auf ein Beispiel soll an dieser Stelle verzichtet werden, da dieses Verfahren ohnehin in den nachfolgenden Kapiteln ausführlich behandelt wird. Allerdings sei hier erwähnt, dass für den hier benutzten Datensatz 3 die Homogenität der Varianzen durch die Rangtransformation hergestellt werden kann. Nachfolgend die Ergebnisse (p-Werte) des Levene-Tests ohne und mit Rangtransformation sowie mit einer normal score-Transformation (siehe Kapitel 2.3):

Effekt	ohne Transformation	mit Rangtransformation	mit normal score-Transf.
gruppe	0.018854	0.39388	0.33077
dosis	0.0066747	0.054783	0.17853
gruppe*dosis	0.011643	0.36508	0.53687

Die o.a. robusten F-Tests sowie der BDM-Test werden mit R gezeigt, während in SPSS Varianzanalysen mit transformierten Daten durchgeführt werden.

In Kapitel 4.1 war darauf hingewiesen worden, dass im Fall ungleicher  $n_i$  und  $s_i^2$  ein pairing überprüft werden sollte. Hierzu müssen die Zellvarianzen  $s_i^2$  und die Zellenbesetzungszahlen  $n_i$  berechnet und miteinander korreliert werden. Für den Datensatz 3 wird dies durchgeführt.

**mit R**

```
si <- with(hetero, tapply(x, list(gruppe, dosis), sd))
ni <- with(hetero, table(gruppe, dosis))
cor(as.vector(si), as.vector(ni))

[1] -0.03766469
```

**mit SPSS**

```
Dataset Declare temp.
Aggregate
  /outfile='temp'
  /break=Gruppe Dosis
  /si=sd(x)
  /ni=NU(x).
compute si=si**2.
Correlations
  /variables=si ni.

-0.038
```

Hieraus ergibt sich also, dass kein pairing, also kein Zusammenhang zwischen den  $n_i$  und  $s_i^2$  besteht und daher keine speziellen Verfahren für heterogene Varianzen anzuwenden sind.

**4.3.3.1 Verfahren von Box, Brown & Forsythe sowie Welch & James****mit R**

Zunächst einmal werden für den o.a. Datensatz 2-faktorielle Varianzanalysen gerechnet, und zwar mit den oben erwähnten F-Tests von Box, Brown & Forsythe in der Version von Mehrotra sowie von Welch & James mit Hilfe der im Anhang 3 aufgelisteten Funktionen `box.f`, `bf.f` bzw. `wj.anova`, wobei zu beachten ist, dass die Syntax für `wj.anova` von den anderen abweicht:

```
box.f(x~gruppe*dosis, mydata3)
bf.f(x~gruppe*dosis, mydata3)
wj.anova(mydata3, "x", "gruppe", "dosis")
```

In der Anova-Tabelle des Box-Tests werden in den Spalten `Eps1` und `Eps2` die Korrekturfaktoren wiedergegeben, mit denen die Zähler- bzw. Nenner-Freiheitsgrade des F-Tests multipliziert werden und dann `Df1` bzw. `Df2` ergeben:

	Eps1	Eps2	Df1	Df2	Sum Sq	Mean Sq	F value	Pr(>F)
gruppe	1.000	0.794	1.00	19.85	9.12	9.116	1.8895	0.1846
dosis	0.708	0.618	2.12	15.45	45.92	15.307	3.1727	0.0677
gruppe:dosis	0.553	0.514	1.66	12.85	11.07	3.691	0.7650	0.4622
Residuals			25.00		120.62	4.825		

In der Anova-Tabelle der Tests von Brown & Forsythe wird neben den Zählerfreiheitsgraden des F-Tests (`Df`) noch die Nenner-Freiheitsgarde (`Df.err`) ausgewiesen:

	Df	Df.err	Sum Sq	Mean Sq	F value	Pr(>F)
gruppe	1.000	3.3822	9.116	9.1162	1.4458	0.3066
dosis	2.135	1.3117	45.922	15.3074	3.2354	0.3137
gruppe:dosis	3.000	12.4217	11.072	3.6908	0.7499	0.5425
Residuals	25.000		120.617	4.8247		

Im Gegensatz zu den anderen Verfahren basiert der Test von Welch & James auf der  $\chi^2$ -Verteilung. Die beiden Faktoren werden in der Tabelle einfach mit „A“ (gruppe) und „B“ (dosis) ausgegeben:

	Chi Sq	df	P(Chi>value)
A	1.653269	1	0.22050000
B	11.738904	3	0.05450455
A:B	2.667716	3	0.53750000

Wie zu sehen ist, differieren die Resultate kaum. Für die Ergebnisse der Varianzanalyse mit der transformierten Variable  $x$  sei auf den Abschnitt „SPSS“ verwiesen. Im Kapitel 4.3.9 werden alle Ergebnisse für diesen Datensatz, auch die von nichtparametrischen Verfahren, gegenübergestellt.

### 4.3.3.2 BDM-Test

mit R:

Der BDM-Test in der nichtparametrischen Version ist im Paket `asbio` u.a. als Funktion `BDM.2way` für eine 2-faktorielle Varianzanalyse enthalten. Nachfolgend ein Beispiel mit demselben oben benutzten Datensatz:

```
library(asbio)
with(mydata3, BDM.2way(x, gruppe, dosis))
```

Two way Brunner-Dette-Munk test				
	df1	df2	F*	P(F > F*)
X1	1.000000	14.05996	0.4143377	0.53013638
X2	2.786237	14.05996	2.9306761	0.07310691
X1:X2	2.786237	14.05996	0.3190448	0.79777127

In der Ausgabe werden mit `x1` und `x2` die beiden Faktoren bezeichnet, hier also Gruppe (`x1`) und Dosis (`x2`). Das Testergebnis zeigt, dass der BDM-Test noch konservativer reagiert als die beiden vorher durchgeführten Tests für heterone Varianzen.

### 4.3.3.3 Variablentransformationen

Alternativ wird verschiedentlich als Abhilfe empfohlen, die Kriteriumsvariable  $x$  zu transformieren. Genannt werden die Transformationen  $\sqrt{x}$  für Häufigkeiten,  $\log(x)$  für rechtsschiefe Verteilungen und  $\arcsin(x)$  für Proportionen. Allerdings bieten solche Transformationen keine Garantie, dass für die transformierte Variable Varianzhomogenität erreicht wird.

mit SPSS

Bei einer Transformation  $\sqrt{x}$  erhält man bei der Überprüfung der Varianzhomogenität immerhin noch einen p-Wert von 0,051, was allerdings akzeptabel wäre. Doch bei einer Transformation  $\log(x)$  verbessert sich das Ergebnis auf  $p=0,170$ . Die entsprechende Varianzanalyse für die Variable  $\ln x = \ln(x)$  ergibt:

Abhängige Variable: $\ln x$					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	1,072 <sup>a</sup>	7	,153	1,743	,144

Konstanter Term	112,524	1	112,524	1280,659	,000
Gruppe	,097	1	,097	1,104	,303
Dosis	,854	3	,285	3,241	,039
Gruppe*Dosis	,139	3	,046	,526	,669
Fehler	2,197	25	,088		
Gesamt	123,359	33			
Korrigierte Gesamtvariation	3,269	32			

so dass hier die log-Transformation wirklich zum Erfolg geführt hat, da zum einen die Varianzen „stabilisiert“ worden sind und zum anderen der Dosis-Effekt signifikant ist.

#### 4.3.4 Rank transform-Tests (RT)

Bei den einfachen Rank transform Tests (RT) wird lediglich vor der Durchführung der parametrischen Varianzanalyse die abhängige Variable in Ränge transformiert. Die statistischen Tests bleiben unverändert. Dieses Verfahren von Conover & Iman (1981) ist in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, nicht jedoch für Variablen mit heterogenen Verteilungen. D.h. hat die untransformierte Variable  $x$  ungleiche Varianzen, so kann das auch noch für die rangtransformierte Variable  $R(x)$  gelten. Daher ist es sinnvoll, auch  $R(x)$  auf Varianzhomogenität zu überprüfen und gegebenenfalls entweder einen der Tests in Kapitel 4.3.3 oder eine weniger empfindliche Methode zu benutzen, z.B. das INT-Verfahren oder den v.d.Waerden-Test, die in den folgenden Kapiteln vorgestellt werden. Für die beiden nachfolgend benutzten Datensätze erübrigt sich dies allerdings, da in Kapitel 4.3.2 für diese keine Varianzhomogenitäten nachgewiesen worden waren.

Das Verfahren wird an den Datensätzen 1 und 2 (`mydata1` und `mydata2`) demonstriert.

##### mit R:

Für das o.a. erste Beispiel (Daten `mydata1`) sind die Anweisung wie folgt zu modifizieren:

```
mydata1 <- within(mydata1, {drugs<-factor(drugs);
                           patients<-factor(patients); rx<-rank(x)})
aov1r <- aov(rx~patients*drugs, mydata1)
summary(aov1r)
```

mit dem Ergebnis:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
patients	1	72.00	72.00	6.680	0.02389	*
drugs	2	56.58	28.29	2.625	0.11333	
patients:drugs	2	217.58	108.79	10.094	0.00268	**
Residuals	12	129.33	10.78			

Tabelle 4-5

Für das o.a. zweite Beispiel lauten die Anweisungen:

```
mydata2 <- within(mydata2, {drugs<-factor(drugs);
                             group<-factor(group); rx<-rank(x)})
aov2r <- aov(rx~group*drugs, mydata2)
drop1(aov2r, ~., test="F")
```

mit der Ausgabe:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			1083.8	131.23			
group	1	364.17	1448.0	138.79	8.4003	0.0076982	**
drugs	3	1157.72	2241.5	149.21	8.9018	0.0003464	***
group:drugs	3	464.61	1548.4	137.00	3.5724	0.0281287	*

Tabelle 4-6

**mit SPSS:**

Zunächst muss über das Menü „Transformieren -> Rangfolge bilden“ bzw. über die Syntax

Rank variables=x (A) /rank into Rx.

x in Ränge transformiert werden, woraus die neue Variable R<sub>x</sub> resultiert. Die Varianzanalyse für R<sub>x</sub>:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	346,167 <sup>a</sup>	5	69,233	6,424	,004
patients	72,000	1	72,000	6,680	,024
drugs	56,583	2	28,292	2,625	,113
patients * drugs	217,583	2	108,792	10,094	,003
Fehler	129,333	12	10,778		

Tabelle 4-7

Für das o.a. zweite Beispiel:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	1820,713 <sup>a</sup>	7	260,102	6,000	,000
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 4-8

Wie ein Vergleich mit den Ergebnissen der parametrischen Varianzanalyse (vgl. Kapitel 4.3.2) zeigt, weichen die Ergebnisse des Rank transform Tests nur geringfügig ab.

#### 4.3.5 Puri & Sen-Test (Verallgemeinerte Kruskal-Wallis-Analysen)

Diese Verfahren gehen gegenüber dem RT-Verfahren einen Schritt weiter: Es werden zwar auch die Werte der abhängigen Variable in Ränge transformiert, jedoch nicht die F-Tests verwendet, sondern aus den Streuungsquadratsummen (SS, Sum of Sq) werden  $\chi^2$ -Tests konstruiert. Diese sind als Verallgemeinerung des Kruskal-Wallis-H-Test anzusehen, da diese im 1-faktoriellen Fall mit letzterem identisch sind.

Die  $\chi^2$ -Werte haben den Aufbau (vgl. Formel 2-6):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{MS_{\text{total}}}$$

wobei  $SS_{Effekt}$  die Streuungsquadratsumme (SS, Sum of Squares) des zu testenden Effektes (A, B oder A\*B) ist und  $MS_{total}$  die Gesamtvarianz (MS, Mean Square). Sie haben die gleichen Freiheitsgrade wie der Zähler des entsprechenden F-Tests.

Da bei der Errechnung der Testgröße nicht die Reduzierung des Fehlers durch andere im Versuchsplan berücksichtigte Faktoren eingeht, hat er zwangsläufig eine geringere Effizienz wie z.B. der o.a. Rank transform Test, der die in Kapitel 4.3.1.3 erwähnte Fehlerreduzierung durch mehrfaktorielle Designs ausnutzt, oder der unten aufgeführte ART.

Natürlich könnte man die o.a.  $\chi^2$ -Werte mit dem Taschenrechner ausrechnen und mit den kritischen Werten in den klassischen Tafelwerken vergleichen. Z.B. für den Test von Faktor `patients` (aus dem ersten Datensatz `mydata1`) errechnet man zunächst  $MS_{total} = 27,94$ . In SPSS ist dieser Wert aus der Zeile `Korrigierte Gesamtvariation` zu entnehmen (vgl. Tabelle 4-7:  $475,500/17$ ), während in R die SS und df aufzusummieren sind (vgl. Tabelle 4-5:  $(72,0 + 56,58 + 217,58 + 129,33) / (1 + 2 + 2 + 12)$ ). Anschließend die Testgröße:

$$\chi^2_{patients} = \frac{72,0}{27,94} = 2,58$$

Da der kritische Wert bei 1 Fg bei einem  $\alpha=0.05$  3,84 beträgt, bestätigt der errechnete  $\chi^2$ -Wert, dass die Patientengruppen keinen signifikanten Einfluss haben.

Mit SPSS ist man auch darauf beschränkt. Mit R lassen sich allerdings diese Schritte auch „programmieren“. Nachfolgend wird das Verfahren mit R an den Beispieldaten 1 (`mydata1`) und 2 (`mydata2`) demonstriert, mit SPSS nur am ersten Datensatz.

#### mit R:

An dieser Stelle sollen die Berechnungen mit der Funktion `av` durchgeführt werden. Die alternative Verwendung von `ezANOVA` wird in Kapitel 5 gezeigt.

Die o.a. Anova-Tabelle 4-5 `aov1r` für das erste Beispiel wird nun weiterverarbeitet.

- Als erstes ist das Objekt `aov1r` mithilfe der Funktion `anova` zu wandeln, damit die Werte in einer Matrix einzeln ansprechbar sind.
- Zunächst muss  $MS_{total}$  als Summe der Sum Sq-Spalte (2. Spalte) dividiert durch die Summe der df-Spalte (1. Spalte) berechnet werden.
- Anschließend wird die 2. Spalte durch die  $MS_{total}$  dividiert.
- Errechnen der p-Werte mit der Funktion `pchisq` unter Verwendung der Freiheitsgrade der F-Werte in der 1. Spalte.
- Zum Schluss wird aus den Berechnungen ein Dataframe erstellt, für den die Effektnamen (Zeilennamen) von `aov1x` übernommen werden.

D.h. die oben in Kapitel 4.3.4 angeführten R-Kommandos sind zu ergänzen um:

```
aov1x <- anova(aov1r)
mstotal <- sum(aov1x[,2])/sum(aov1x[,1])
chisq <- aov1x[,2]/mstotal
df <- aov1x[,1]
pvalues <- 1-pchisq(chisq,df)
aov1y <- data.frame(chisq,df,pvalues)
```

```
row.names(aov1y) <- row.names(aov1x)
aov1y[1:3,]
```

Die daraus resultierende Ausgabe:

	chisq	df	pvalues
patients	2.574132	1	0.10862364
drugs	2.022958	2	0.36368065
patients:drugs	7.779005	2	0.02045552

Tabelle 4-9

Ein Vergleich mit den Tabellen 4-1 und 4-5 zeigt, dass in diesem Fall nicht alle Signifikanzen der parametrischen bzw. der Rank transform Tests mit den Puri & Sen-Tests reproduziert werden können. Anzumerken ist noch, dass das Testergebnis für den Faktor `drugs` (wie vorher bereits darauf hingewiesen) identisch ist mit dem Kruskal-Wallis H-Test, der 1-faktoriellen Analyse (vgl. Kapitel 4.2.1).

Für das o.a. zweite Beispiel sind auch hier wegen des unbalancierten Versuchsplans ein paar zusätzliche Schritte erforderlich. Insbesondere werden mit `drop1` die Streuungsquadrate vom Typ III ermittelt. Die Berechnung von  $MS_{total}$  erfolgt wie oben aus der ursprünglichen Varianzanalyse `aov2r` durch Summation der Streuungsquadratsummen `aov2x[, 2]` und Residuen `aov2x[, 1]`. `aov2r` muss wie im vorigen Beispiel mit `anova` in ein verarbeitbares Format gebracht werden. Zu beachten ist, dass die Ausgabe von `drop1`, auf `aov2s` gespeichert, eine redundante 1. Zeile enthält (vgl. Tabelle 4-6).

```
mydata2 <- within(mydata2, {drugs<-factor(drugs);
                           group<-factor(group); rx<-rank(x)})
aov2r <- aov(rx~group*drugs, mydata2)
aov2s <- drop1(aov2r, ~. , test="F")
aov2x <- anova(aov2r)
mstotal <- sum(aov2x[, 2]) / sum(aov2x[, 1])
chisq <- aov2s[, 2] / mstotal
df <- aov2s[, 1]
pvalues <- 1-pchisq(chisq, df)
aov2y <- data.frame(chisq, df, pvalues)
row.names(aov2y) <- row.names(aov2s)
aov2y[2:4,]
```

mit der Ausgabe:

	chisq	df	pvalues
group	4.012175	1	0.045172850
drugs	12.755071	3	0.005197361
group:drugs	5.118797	3	0.163302051

Tabelle 4-10

Ein Vergleich mit den Tabellen 4-2 und 4-6 zeigt, dass auch in diesem Fall nicht alle Signifikanzen der parametrischen bzw. der Rank transform Tests mit den Puri & Sen-Tests reproduziert werden können.

Alternativ können die Puri & Sen-Tests auch mit der Funktion `np.anova` (vgl. Anhang 3.8) durchgeführt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Nachfolgend die Ein- und Ausgabe:



```
np.anova(x~group*drugs, mydata2)
```

```
generalized Kruskal-Wallis/Friedman tests
      Df Sum Sq Chi Sq Pr(>Chi)
group    1  189.00  2.0823  0.149017
drugs     3 1157.72 12.7551  0.005197
group:drugs 3  464.61  5.1188  0.163302
Residuals 25 1083.79
```

Die Funktion `np.anova` erlaubt auch die Verwendung von Pseudo-Rängen anstatt der “normalen“ Ränge (vgl. Abschnitt 2.16). Dazu ist der Parameter `pseudo=T` anzugeben, wodurch die Signifikanzen allerdings (wegen der geringen Streuung der  $n_i$ ) nur geringfügig abgeschwächt werden

```
np.anova(x~group*drugs, mydata2, pseudo=T)
```

```
generalized Kruskal-Wallis/Friedman tests
      Df Sum Sq Chi Sq Pr(>Chi)
group    1  185.36  1.9876  0.158590
drugs     3 1182.84 12.6835  0.005373
group:drugs 3  476.98  5.1146  0.163594
Residuals 25 1113.87
```

### mit SPSS:

Ausgangsbasis ist die Anova-Tabelle 4-7. Zunächst muss die Gesamtvarianz  $MS_{total}$  in SPSS *korrigierte Gesamtvariation* bezeichnet, berechnet werden, da nur die Quadratsumme und Freiheitsgrade ausgegeben werden, nicht aber das Mittel der Quadrate (Mean Square):

$$MS_{total} = \frac{475}{17} = 27,94$$

Anschließend werden für jeden Effekt die  $\chi^2$ -Werte errechnet:

$$\chi^2_{patients} = \frac{72}{27,94} = 2,58 \quad df_{patients} = 1$$

$$\chi^2_{drugs} = \frac{56,68}{27,94} = 2,03 \quad df_{drugs} = 2$$

$$\chi^2_{Interaktion} = \frac{217,53}{27,94} = 7,78 \quad df_{Interaktion} = 2$$

Die 5%-Schranken für die  $\chi^2$ -Verteilung liegen bei 3,8 für  $df=1$  bzw. 6,0 für  $df=2$ . Somit liegt nur ein signifikanter Interaktionseffekt vor. Ein Vergleich mit den Tabellen 4-3 und 4-7 zeigt, dass in diesem Fall nicht alle Signifikanzen der parametrischen bzw. der Rank transform Tests mit den Puri & Sen-Tests reproduziert werden können.

Auf die Berechnung für das zweite Beispiel kann hier verzichtet werden, da in SPSS nicht zwischen balancierten und unbalancierten Versuchsplänen unterschieden werden muss.

### 4.3.6 Aligned rank transform (ART und ART+INT)

Verschiedene Studien, u.a. von Sawilowsky, S., Blair, R. C., & Higgins, J. J. (1989), haben gezeigt, dass für den Test der Interaktion, insbesondere nach dem o.a. Rank transform-Verfahren, der Fehler 1. Art nicht immer korrekt eingehalten wird, d.h. dass mehr Interaktionen zufällig signifikant sind, als es das vorgegebene  $\alpha$  zulässt. Als Ursache wird angesehen, dass der Test der Interaktion nicht von den Tests der beiden Haupteffekte unabhängig ist. Als Lösung wird propagiert, zunächst ein komplettes Modell zu analysieren, anschließend für dessen Residuen die beiden Haupteffekte herauszupartialisieren, dann diese bereinigten Residuen in Ränge umzurechnen, um schließlich wiederum ein normales Modell mit Interaktion zu rechnen. Die Streuungsquadrate für die Haupteffekte sollten dann bei diesem Modell bei Null liegen. Die Haupteffekte sind dann aus der Analyse des ersten Modells zu entnehmen. Beim zweiten Modell interessiert dann lediglich der Test für die Interaktion. Im Folgenden werden auch zur Demonstration ART-Tests der Haupteffekte durchgeführt, wenn das auch nicht erforderlich und wie in Kapitel 2.4 erwähnt nicht angebracht ist.

Die Schritte im Einzelnen:

- Durchführung einer (normalen) Anova mit Haupt- und Interaktionseffekten.
- Speichern der Residuen ( $e_m$ ),
- Eliminieren des zu untersuchenden Effekts aus den Residuen:

$$\text{Interaktionseffekt: } e_m + (\bar{ab}_{ij} - \bar{a}_i - \bar{b}_j + 2\bar{x})$$

$$\text{Haupteffekte: } e_m + (\bar{a}_i + \bar{b}_j - \bar{x})$$

bzw. wenn beide Haupteffekte separat getestet werden sollen:

$$\text{Haupteffekt A: } e_m + \bar{a}_i$$

$$\text{Haupteffekt B: } e_m + \bar{b}_j$$

bzw. im Fall einer 3-faktoriellen Varianzanalyse für die 3-fach-Interaktion:

$$\text{Interaktionseffekt: } e_m + (\bar{abc}_{ijl} - \bar{ab}_{ij} - \bar{ac}_{il} - \bar{bc}_{jl} + \bar{a}_i + \bar{b}_j + \bar{c}_l)$$

- Umrechnung der bereinigten Residuen in Ränge.
- Durchführung einer normalen Anova mit Haupt- und Interaktionseffekten mit den Rängen, aus der dann der untersuchte Effekt abgelesen werden kann.

Es wird empfohlen (siehe Mansouri & Chang, 1995 sowie Carletti & Clautriaux, 2005) anschließend die Ränge in normal scores (vgl. Kapitel 2.3) umzurechnen, um einerseits etwaige falsche Signifikanzen abzuschwächen und andererseits eine größere Power zu erhalten.

Es soll nun im Folgenden für den Beispieldatensatz 2 überprüft werden, ob die oben ausgewiesene Signifikanz der Interaktion garantiert ist.

#### mit R:

Zunächst die Durchführung des Verfahrens „per Hand“, d.h. das Alignment, also die Umrechnung der Werte wird elementar vorgenommen.

Dazu wird als erstes für  $x$  die klassische Anova errechnet ( $aov3$ ) und daraus die Residuen extrahiert. Zu den Residuen werden dann einmal zur Ermittlung der Interaktion dieser Effekt addiert ( $r_{ab}$ ) sowie einmal zur Ermittlung des Haupteffekte die entsprechende Effekt addiert ( $r_a$  und  $r_b$ ). Anschließend werden die bereinigten Residuen in Ränge transformiert ( $r_{abr}$  bzw.  $r_{ar}$ ). Zur Überprüfung der Interaktion bzw. der Haupteffekte wird jeweils ein

komplettes Modell mit diesen Residuenrängen analysiert. Gemäß den Anmerkungen in Kapitel 3.3 zu Fehlern bei der Rangberechnung empfiehlt es sich, vorher die bereinigten Residuen mittels `round` auf 7 Dezimalstellen zu runden.

```
mydata2 <- within(mydata2, {drugs<-factor(drugs); group<-factor(group)})
aov3 <- aov(x~group*drugs, mydata2)
rab <- aov3$residuals
ra <- rab

# Zellenmittelwerte
mij <- ave(mydata2[,3], mydata2[,1], mydata2[,2], FUN=mean)
ai <- ave(mydata2[,3], mydata2[,1], FUN=mean) # Effekte Faktor A
bj <- ave(mydata2[,3], mydata2[,2], FUN=mean) # Effekte Faktor B
mm <- mean(mydata2[,3]) # Gesamtmittel

# Bereinigung der Residuen
rab <- rab + (mij - ai - bj + 2*mm) # Interaktion
ra <- ra + (ai + bj - mm) # Haupteffekte
rabr <- rank(round(rab, digits=7)) # Runden und
rar <- rank(round(ra, digits=7)) # Umrechnung in Ränge
aov3ab <- aov(rabr~group*drugs, mydata2) # Anova Interaktion
drop1(aov3ab, ~. , test="F") # Ergebnis Interaktionseffekt
aov3a <- aov(rar~group*drugs, mydata2) # Anova Haupteffekte
drop1(aov3a, ~. , test="F") # Ergebnis Haupteffekte
```

mit den Ergebnissen für den Interaktionseffekt:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			2094.9	152.97		
group	1	15.16	2110.1	151.21	0.1809	0.67423
drugs	3	2.48	2097.4	147.01	0.0099	0.99862
group:drugs	3	876.49	2971.4	158.51	3.4866	0.03058 *

sowie für die Haupteffekte:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			1223.2	135.22		
group	1	363.94	1587.1	141.81	7.4385	0.0115045 *
drugs	3	1407.31	2630.5	154.49	9.5879	0.0002159 ***
group:drugs	3	4.14	1227.3	129.33	0.0282	0.9934370

Tabelle 4-11

Vergleicht man diese Ergebnisse mit dem Ergebnis der Rank transform Tests von  $x$  (vgl. Tabelle 4-6), können sowohl die Interaktion als auch die Haupteffekte als gesichert angesehen werden. Es sei noch angemerkt, dass die beiden o.a. Ergebnisse für die Interaktion sowie die Haupteffekte ohne die Rundung mittels `round` leicht von den obigen abweichen.

Seit Anfang 2015 wird das Paket `ARTool` für R zur Verfügung gestellt, mit dessen Hilfe die Umrechnung der Werte bequem vorgenommen werden kann. Dazu dient die Funktion `art`, die u.a. unter `$aligned.ranks` die Ränge der umgerechneten Werte für alle Effekte als Dataframe enthält. Die beiden Argumente der Funktion sind mit denen von `aov` identisch. Doch Vorsicht: die Namen der Variablen sind die Namen der Effekte, in diesem Beispiel also `group`, `drugs` und `group:drugs`, also in der Regel mit den Faktornamen identisch und sollten daher umbenannt werden.

Nachfolgend die Durchführung des ART-Verfahrens zur Ermittlung des bereinigten Tests für die Interaktion. `mydata2a` ist das Ergebnis von `art`, das mit dem Ausgangsdatensatz mittels `cbind` zusammengeführt wird. Dabei erhalten die Variablennamen durch die Angabe `aligned=` das Präfix `aligned`, z.B. `aligned.drugs`, werden aber anschließend umbenannt.

```
library(ARTool)
mydata2a <- art(x~group*drugs,mydata2)$aligned.ranks
mydata2x <- cbind(mydata2,aligned=mydata2a)
names(mydata2x)[4:6] <- c("a.g","a.d","a.gd")
drop1(aov(a.gd~group*drugs,mydata2x), ~. , test="F")
```

a.gd ~ group * drugs	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			2094.9	152.97		
group	1	15.16	2110.1	151.21	0.1809	0.67423
drugs	3	2.48	2097.4	147.01	0.0099	0.99862
group:drugs	3	876.49	2971.4	158.51	3.4866	0.03058 *

Für die Umrechnung in normal scores, d.h. Anwendung des ART+INT-Verfahrens, sind zusätzlich zu den zuletzt angeführten noch die folgenden Anweisungen erforderlich, zunächst mit `n.gd` für den Interaktionseffekt, danach mit `n.g` und `n.d` für die beiden Haupteffekte:

```
nc <- dim(mydata2)[1]
n.gd <- qnorm(mydata2x$a.gd/(nc+1))
drop1(aov(n.gd~group*drugs,mydata2x), ~. , test="F")
n.g <- qnorm(mydata2x$a.g/(nc+1))
drop1(aov(n.g~group*drugs,mydata2x), ~. , test="F")
n.d <- qnorm(mydata2x$a.d/(nc+1))
drop1(aov(n.d~group*drugs,mydata2x), ~. , test="F")
```

Hier lediglich die Ausgabe für den Test der Interaktion:

n.gd ~ group * drugs	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			19.020	-2.1839		
group	1	0.8859	19.906	-2.6815	1.1645	0.29084
drugs	3	0.0384	19.058	-8.1174	0.0168	0.99695
group:drugs	3	7.8930	26.913	3.2711	3.4582	0.03144 *

Alternativ kann das ART+INT-Verfahren auch bequem über die Funktion `art1.anova` (vgl. Anhang 3) durchgeführt werden. Diese Funktion dient primär dem ART-Verfahren (alternativ zu der o.a. Funktion `art` des Pakets `ARTool`), doch über den Parameter `INT=T` wird nach der Rangbildung noch die Transformation in normal scores vorgenommen:

```
art1.anova(x~group*drugs,mydata2,INT=T)
```

#### mit SPSS:

- Zunächst wird für `x` die klassische Anova (`Unianova`) errechnet und dabei die Residuen gespeichert.
- Dann müssen mittels `Aggregate` die Effekte als Mittelwerte für die Gruppen ermittelt werden: `mij` für die Interaktion, `ai` für Faktor `group` und `bj` für Faktor `drugs`. Diese werden in die Arbeitsdatei eingefügt.

- Zu den Residuen werden dann einmal zur Ermittlung der Interaktion dieser Effekt addiert ( $r_{ab}$ ) sowie einmal zur Ermittlung der Haupteffekte deren Effekte addiert ( $r_a$  und  $r_b$ ).
- Anschließend werden die bereinigten Residuen in Ränge transformiert ( $r_{abr}$  bzw.  $r_{ar}$ ).
- Zur Überprüfung der Interaktion bzw. der Haupteffekte wird jeweils ein komplettes Modell mit diesen Residuenrängen analysiert.

```

Unianova  x by group drugs
  /save=resid (rab)
  /design=group drugs group*drugs.

Compute ra=rab.

Aggregate
  /outfile=* mode=addvariables
  /break=group drugs  /mij=mean(x) .
Aggregate
  /outfile=* mode=addvariables
  /break=group  /ai=mean(x) .
Aggregate
  /outfile=* mode=addvariables
  /break=drugs  /bj=mean(x) .

Aggregate
  /outfile=* mode=addvariables
  /break=      /mm=mean(x) .

Compute rab=rab + (mij - ai - bj + 2*mm) .
Compute ra =ra  + (ai + bj - mm) .

Rank variables=ra rab (A)
  /rank into rar rabr.

Unianova  rabr by group drugs
  /design=group drugs group*drugs.
Unianova  rar by group drugs
  /design=group drugs group*drugs.

```

mit den Ergebnissen für den Interaktionseffekt:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	10,592	1	10,592	,131	,721
drugs	40,762	3	13,587	,167	,917
group * drugs	938,767	3	312,922	3,856	,021
Fehler	2028,817	25	81,153		

sowie für die Haupteffekte:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	319,765	1	319,765	5,638	,026
drugs	1267,690	3	422,563	7,450	,001
group * drugs	8,802	3	2,934	,052	,984
Fehler	1418,017	25	56,721		

Tabelle 4-12

Vergleicht man diese Ergebnisse mit dem Ergebnis der Rank transform Tests von  $x$  (vgl. Tabelle 4-8), können sowohl die Interaktion als auch die Haupteffekte als gesichert angesehen werden.

Für die Umrechnung in normal scores, d.h. Anwendung des ART+INT-Verfahrens, sind noch zusätzlich die folgenden Anweisungen erforderlich:

```
Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(x) .

Compute nsar=Idf.normal(rar/(nc+1),0,1) .
Compute nsabr=Idf.normal(rabr/(nc+1),0,1) .

Unianova nsabr by group drugs
  /design=group drugs group*drugs .
Unianova nsar by group drugs
  /design=group drugs group*drugs .
```

mit den Ergebnissen für den Interaktionseffekt:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	,009	1	,009	,011	,916
drugs	,038	3	,013	,017	,997
group * drugs	7,893	3	2,631	3,458	,031
Fehler	19,020	25	,761		

sowie für die Haupteffekte:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
group	3,309	1	3,309	7,403	,012
drugs	12,785	3	4,262	9,535	,000
group * drugs	,075	3	,025	,056	,982
Fehler	11,173	25	,447		

#### 4.3.7 normal scores- (INT-) und van der Waerden-Tests

Bei der einfachen *inverse normal transformation* (INT) wird lediglich vor der Durchführung der parametrischen Varianzanalyse zunächst die abhängige Variable  $x$  in Ränge  $R(x)$  transformiert und anschließend über die inverse Normalverteilung in normal scores ungerechnet:

$$nscore_i = \Phi^{-1}(R(x_i)/(N+1))$$

wobei  $N$  die Anzahl der Werte ist und  $\Phi^{-1}$  die Umkehrfunktion der Normalverteilung. Die statistischen Tests bleiben unverändert. Dieses Verfahren ist wie beim o.a. RT-Verfahren in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, nicht jedoch für Variablen mit beliebigen Eigenschaften. D.h. hat die untransformierte Variable  $x$  ungleiche Varianzen, so kann das auch noch für die transformierte Variable  $nscore$  gelten. So kann es sinnvoll sein, gegebenenfalls auch  $nscore$  auf Varianzhomogenität zu überprüfen und gegebenenfalls einen der Tests in Kapitel 4.3.3 oder den anschließend vorgestellten van der Waerden-Test zu benutzen. In den nachfolgenden Beispielen wird darauf verzichtet, da bereits die nichttransformierten Daten homogen sind.

Bei dem Verfahren von *van der Waerden* werden anstatt der „klassischen“ F-Tests die  $\chi^2$ -Tests des Kruskal-Wallis-Tests bzw. wie bei der o.a. Puri & Sen-Methode gerechnet. Die  $\chi^2$ -Werte haben den Aufbau (vgl. Formel 2-6a):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{MS_{\text{total}}}$$

wobei  $SS_{\text{Effekt}}$  die Streuungsquadratsumme (SS, Sum of Squares) des zu testenden Effektes (A, B oder A\*B) ist und  $MS_{\text{total}}$  die Gesamtvarianz (MS, Mean Square). Sie haben die gleichen Freiheitsgrade wie der Zähler des entsprechenden F-Tests. (Vgl. auch Kapitel 4.3.5.)

Im folgenden Beispiel wird der zuletzt benutzte Datensatz `mydata2` verwendet.

#### mit R:

Wegen des nichtbalancierten Versuchsplans müssen zunächst mittels `option` die Standard-Kontraste zugewiesen werden sowie nach der Anova mit `aov` mittels `drop1` Quadratsummen vom Typ III errechnet werden. `nc` enthält die Anzahl der Merkmalsträger, die bei der Umrechnung in normal scores einfließt.

```
options (contrasts=c("contr.sum", "contr.poly"))
nc      <- dim(mydata2) [1]
Rx      <- rank(x)
nsx     <- qnorm(Rx/(nc+1))
aov2ns  <- aov(nsx~group*drugs, mydata2)
aov2ns1 <- drop1(aov2ns, ~. , test="F")
```

Diese Anweisungen dienen zunächst für die Analyse der normal scores (INT-Verfahren) mit folgendem Ergebnis:

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

Für die Durchführung der *van der Waerden*-Tests sind noch zusätzlich die folgenden Anweisungen erforderlich, um die  $\chi^2$ -Tests durchzuführen (vgl. auch das Beispiel in Kapitel 4.3.5):

```
aov2ns  <- anova(aov2ns)
mstotal <- sum(aov2ns[,2])/sum(aov2ns[,1])
chisq   <- aov2ns[,2]/mstotal
df      <- aov2ns[,1]
pvalues <- 1-pchisq(chisq,df)
aov2vdw <- data.frame(chisq,df,pvalues=round(pvalues,digits=5))
row.names(aov2vdw) <- row.names(aov2ns1)
aov2vdw[2:4,]
```

	chisq	df	pvalues
group	3.710002	1	0.05409
drugs	13.436428	3	0.00378
group:drugs	5.189060	3	0.15847

Ein Vergleich mit den Tabellen 4-6 und 4-13 zeigt, dass in diesem Fall nicht alle Signifikanzen der Rank transform Tests bzw. des einfachen normal scores-Tests mit den van der Waerden-Tests reproduziert werden können.

Alternativ kann das van der Waerden-Verfahren auch mit der Funktion `np.anova` (vgl. Anhang 3.8) durchgeführt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `anova`. Über den Zusatz `method=1` werden anstatt Puri & Sen-Tests van der Waerden-Tests durchgeführt. Nachfolgend die Ein- und Ausgabe:

```
np.anova(x~group*drugs, mydata2, method=1)
```

	Df	Sum Sq	Chi Sq	Pr(>Chi)
group	1	2.9521	3.7100	0.054087
drugs	3	10.6917	13.4364	0.003782
group:drugs	3	4.1290	5.1891	0.158465
Residuals	25	9.3768		

#### mit SPSS:

Die Rang-Transformation sowie die Umrechnung in normal scores (Ergebnisvariable `nsx`) werden zweckmäßigerweise über das Syntax-Fenster vorgenommen. Das für die Umrechnung erforderliche  $n$  (Anzahl der Fälle, Variable `nc`) wird über `Aggregate` ermittelt:

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

Abhängige Variable: nsx					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	16,086 <sup>a</sup>	7	2,298	6,127	,000
Konstanter Term	,163	1	,163	,435	,516
group	2,952	1	2,952	7,871	,010
drugs	10,692	3	3,564	9,502	,000
group * drugs	4,129	3	1,376	3,670	,026
Fehler	9,377	25	,375		
Gesamt	25,463	33			
Korrigierte Gesamtvariation	25,463	32			

Tabelle 4-14

Für den van der Waerden-Test müssen wie beim Puri & Sen-Test (Kapitel 4.3.5)  $\chi^2$ -Werte errechnet werden. Zunächst muss die Gesamtvarianz  $MS_{total}$ , in SPSS *korrigierte Gesamt-*



*variation* bezeichnet, berechnet werden, da nur die Quadratsumme und Freiheitsgrade ausgegeben werden, nicht aber das Mittel der Quadrate (Mean Square):

$$MS_{total} = \frac{25,46}{32} = 0,796$$

Anschließend werden für jeden Effekt die  $\chi^2$ -Werte errechnet:

$$\chi^2_{group} = \frac{2,95}{0,796} = 3,71 \quad df_{patients} = 1$$

$$\chi^2_{drugs} = \frac{10,69}{0,796} = 13,43 \quad df_{drugs} = 3$$

$$\chi^2_{Interaktion} = \frac{4,13}{0,796} = 5,19 \quad df_{Interaktion} = 3$$

Die 5%-Schranken für die  $\chi^2$ -Verteilung liegen bei 3,8 für  $df=1$  bzw. 7,8 für  $df=3$ . Somit liegt nur ein signifikanter Haupteffekt (Faktor *drugs*) vor. Ein Vergleich mit den Tabellen 4-8 und 4-14 zeigt, dass in diesem Fall nicht alle Signifikanzen der Rank transform Tests bzw. des einfachen normal scores-Tests mit den van der Waerden-Tests reproduziert werden können.

#### 4.3.8 ATS-Tests von Akritas, Arnold & Brunner

Im Gegensatz zum RT-, dem INT- oder dem ART-Verfahren sind, steckt in den Tests von Akritas, Arnold & Brunner sehr viel mehr Mathematik. Die Berechnung ist vergleichsweise kompliziert, so dass sie in SPSS nicht möglich ist und in R einen erheblichen Programmieraufwand erfordert. Sie ist allerdings sehr übersichtlich dokumentiert in dem Buch von Edgar Brunner und Ullrich Munzel (2013). Entsprechende R-Funktionen wurden in Kapitel 3.1 vorgestellt.

##### mit R:

Das Verfahren soll am 2. Datensatz demonstriert werden. Dazu wird die Funktion `rankFD` benutzt. Alternativ kann die Funktion `GFD` aus dem Paket `GFD` angewandt werden, wenn die Vermutung inhomogener Varianzen besteht. Dafür müssen allerdings konservativere Ergebnisse in Kauf genommen werden. Nachfolgend die Anweisungen für beide Funktionen sowie die Ausgabe für `rankFD`:

```
library(rankFD)
rankFD(x~group*drugs, mydata2) $ANOVA.Type.Statistic
GFD(x~group*drugs, mydata2, nperm=1) $ATS
```

	Statistic	df1	df2	p-Value
group	10.11375	1.000000	13.40492	0.006998751
drugs	10.11411	2.279963	13.40492	0.001631986
group:drugs	3.71235	2.279963	13.40492	0.047601610

die zeigt, dass mit diesem Verfahren alle drei Signifikanzen der Rank transform Tests (vgl. Tabelle 4-6) und des ART (vgl. Tabelle 4-11) reproduziert werden können.

### 4.3.9 Bredenkamp Tests

Der Test von Bredenkamp ist ein Spezialfall der Tests von Puri & Sen, und zwar ausschließlich für balancierte Versuchspläne. Er bedient sich des H-Tests von Kruskal-Wallis, also nur der 1-faktoriellen Varianzanalyse, und ist sehr einfach durchzuführen. Dabei ist zu bedenken, dass dieses Verfahren noch aus einer Zeit stammt, bevor die (inzwischen vielen) neueren Verfahren zur nichtparametrischen Varianzanalyse publiziert waren.

Das Verfahren beruht auf der Additivität des  $\chi^2$ -Tests. Die Tests für die Haupteffekte A und B werden wie gewohnt mit dem H-Test durchgeführt. Anschließend wird ein H-Test über alle Zellen hinweg gerechnet. Von diesem  $\chi^2$ -Wert werden die Werte aus den H-Tests für Faktor A und Faktor B subtrahiert. Das Ergebnis ist der  $\chi^2$ -Wert für die Interaktion. Analog werden die Freiheitsgrade ermittelt. Vgl. obige Tabelle.

H-Testwerte ( $\chi^2$ -Werte)	Freiheitsgrade
$\chi^2_{AB}$	$IJ-1$
- $\chi^2_A$	$I-1$
- $\chi^2_B$	$J-1$
$\chi^2_{AB} - \chi^2_A - \chi^2_B$	$(I-1)(J-1)$

Da das Prozedere mit R und SPSS gleichermaßen abläuft, soll hier nur mit SPSS ein Beispiel durchgerechnet werden.

#### mit SPSS:

Es wird hier eine 2-faktorielle Varianzanalyse für das erste Datenbeispiel durchgeführt, das einen balancierten Versuchsplan beinhaltet. Zunächst müssen die Zellen für den ersten H-Test einmal durchnummeriert werden:

Durchnummerierung der Zellen:= (Patients - 1)\*#Drugs + Drugs

Anschließend werden H-Tests für die Zellen, für Faktor `patients` und Faktor `drug` gerechnet. Die SPSS-Syntax hierfür:

```
Compute zelle=(patients-1)*3 + drugs.
Nptests /independent test (x) group (zelle) kruskal_wallis.
Nptests /independent test (x) group (patients) kruskal_wallis.
Nptests /independent test (x) group (drugs) kruskal_wallis.
```

Die SPSS-Ergebnisse sind in folgender Tabelle zusammengefasst:.

Effekt	H-Testwerte ( $\chi^2$ -Werte)	Freiheitsgrade	Signifikanz
Zellen	12,376	5	
patients	2,574	1	n.s. (< 3,8)
drugs	2,023	2	n.s. (< 3,8)
patients*drugs	7,779	2	s. (> 6,0)

Tabelle 4-15

Diese Ergebnisse decken sich mit denen aus den Tests von Puri & Sen (vgl. Tabelle 4-9).

#### **4. 4 Nichtparametrische Verfahren zur mehrfaktoriellen Varianzanalyse**

Die in 4.3. vorgestellten Verfahren lassen sich alle ohne Weiteres auf drei und mehr Faktoren erweitern. Lediglich für die in 4.3.3 vorgestellten Verfahren für ungleiche Varianzen liegen nur 2-faktorielle Lösungen vor.

## 5. Abhängige Stichproben - Messwiederholungen

Es wird im Folgenden davon ausgegangen, dass ein Merkmal  $x$   $J$ -mal (unter verschiedenen Bedingungen) erhoben wurde, so dass Variablen  $x_1, \dots, x_J$  vorliegen, deren Mittelwerte verglichen werden sollen. Z.B. können von dem Merkmal "Herzfrequenz" (HF) mehrere Messungen vorliegen, z.B. der Ruhewert, der Wert direkt nach Beendigung des Joggens sowie Werte 10 und 20 Minuten nach Beendigung, also insgesamt 4 Werte. Die Struktur kann aber auch hier mehrfaktoriell sein, wenn z.B. o.a. HF-Messungen einmal ohne Einnahme eines Medikaments und einmal mit Einnahme vorgenommen worden sind.

### Beispieldaten 4 (winer518):

Der folgende Datensatz ist dem Buch von B.J.Winer (1991, S. 518) entnommen. Die Einstellung zu einem Thema wurde von Männern 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`.

### Beispieldaten 5 (mydata5):

Im folgenden Datensatz geht es um die Reaktionsfähigkeit in Abhängigkeit von der Einnahme von 2 verschiedenen Medikamenten. 8 Personen, 4 Männer und 4 Frauen, müssen 3 verschiedene Aufgaben (1, 2, 3) lösen, einmal ohne Einnahme eines Präparats (Kontrollmessung K) sowie je einmal nach Einnahme von Medikament A bzw. B (A, B). Das Kriterium ist die Fehlerzahl, mit der eine Aufgabe bearbeitet wurde. Dieses ist zwar eigentlich metrisch, wegen des kleinen Wertebereichs aber eher ordinal zu handhaben.

Geschlecht	Versuchsperson	Kontrolle K			Medikament A			Medikament B		
		Aufgabe			Aufgabe			Aufgabe		
		1	2	3	1	2	3	1	2	3
Männer	1	3	3	1	4	4	2	5	4	3
	2	2	0	0	3	2	2	4	3	3
	3	5	4	3	5	3	3	6	3	4
	4	3	5	2	4	4	3	4	4	4
Frauen	5	2	2	1	2	2	2	5	2	3
	6	4	1	0	3	2	1	5	2	2
	7	3	2	1	3	2	1	4	3	2
	8	1	3	0	5	2	1	6	3	3

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 `mydata5`, in dem die 9 Messwiederholungsvariablen die Namen `v1`, . . . , `v9` haben.

### 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`.

## 5.1 Datenstruktur

### 5.1.1 Besonderheiten bei R und SPSS

In der Regel liegen die Daten in Form einer Datenmatrix vor, bei der die Zeilen den Erhebungseinheiten (Fällen) entsprechen, also typischerweise Versuchspersonen, und die Spalten den erhobenen Merkmalen (Variablen). Liegen z.B. von der Variablen HF die oben aufgeführten 4 Werte vor, so sind diese normalerweise als 4 Variablen (z.B. `HF_Ruhe`, `HF_0`, `HF_10` und `HF_20`), also 4 Spalten, in der Datenmatrix zu finden. Bei den meisten Statistikprogrammen, so auch bei SPSS, werden dann zum Vergleich der Messwiederholungen diese Variablen angegeben.

Nicht so bei R. Hier werden die Messwiederholungen von Variablen i.d.Regel nicht als Spalten, sondern als Zeilen in der Datenmatrix wiederholt. Dies erfordert zwei zusätzliche Kennungen:

- eine Kennzeichnung der Erhebungseinheit, üblicherweise Fall- oder Versuchspersonennummer, sowie
- eine Kennung der Messwiederholung, ähnlich einem Gruppierungsfaktor.

Für die statistischen Funktionen ist es ganz wichtig, dass beide Variablen vom Typ „factor“ sind, insbesondere da die Funktionen auch fehlerfrei durchlaufen, wenn diese Deklaration vergessen wurde. Nur: Die Ergebnisse sind dann falsch. Variablen, die nicht mehrfach gemessen wurden, wie z.B. `Geschlecht`, bleiben dann in den Wiederholungszeilen für die Messwiederholungen konstant.

Zum Wandeln der Datenstruktur, um Versuchspläne mit Messwiederholungen in R analysieren zu können, wird die Funktion `reshape` benutzt, mit der sowohl Messwiederholungen in Fälle ("langes" Format, Parameter `direction=long`), mit ein wenig Aufwand auch für mehrfaktorielle Designs, gewandelt werden können, als auch umgekehrt Fälle in Messwiederholungen ("breites" Format, Parameter `direction=wide`). Die Funktion soll hier nicht näher beschrieben werden, da die "Standard"-Benutzung aus den Beispielen hervorgeht. Auf eines muss allerdings aufmerksam gemacht werden: Anders als in der u.a. schematischen Darstellung werden bei der Wandlung in das lange Format zuerst alle Erhebungseinheiten für die erste Messwiederholung generiert, danach alle Erhebungseinheiten für die zweite Messwiederholung usw..

Allerdings ist eine solche Umstrukturierung verschiedentlich auch bei SPSS erforderlich, und zwar zur Berechnung der Ränge. SPSS bietet nur eine Funktion zur Berechnung von Rängen, und zwar für eine Variable über alle Fälle, also spaltenweise. Bei Messwiederholungen ist allerdings auch die zeilenweise Rangberechnung erforderlich. Daher müssen die Messwiederholungen wie oben skizziert in mehrere Zeilen umgewandelt werden. SPSS bietet dazu Verfahren an. Diese sind ausführlich im Anhang 1 beschrieben.

Der erforderliche Umwandlungsprozess soll an zwei Beispielen veranschaulicht werden. Zunächst einmal an dem einfachen Fall eines Mermals HF, das zu 4 Zeitpunkten beobachtet worden ist (siehe oben): zuerst die Ausgangsbasis, darunter die erforderliche Struktur mit den zusätzlichen Variablen `vpn` (Fallkennzeichnung) und `zeit` (Kennzeichnung der Messwiederholung):

Sex	Alter	...	HF_R	HF_0	HF_10	HF_20
1	51	...	70	91	82	76
2	64	...	78	102	87	79
...	...	...	...	...	...	...

V <sub>pn</sub>	Sex	Alter	...	Zeit	HF
1	1	51	...	1	70
1	1	51	...	2	91
1	1	51	...	3	82
1	1	51	...	4	76
2	2	64	...	1	78
2	2	64	...	2	102
...	...	...	...	...	...

Nachfolgend der etwas kompliziertere Fall von zwei Merkmalen, systolischer und diastolischer Blutdruck (Sys.. bzw. Dia..), die zum einen zu 3 Zeitpunkten (..1, ..2, ..3) und zum anderen ohne und mit einer Medikamentendosierung (..o, ..m) gemessen worden sind. Auch hier sind 3 neue Variablen erforderlich: `vpn` (Fallkennzeichnung), `Dosis` (Messwiederholung Dosierung) und `zeit` (Messwiederholung Zeit). Zunächst die Ausgangsstruktur:

Sex	Alter	Sys1o	Dia1o	Sys2o	Dia2o	Sys3o	Dia3o	Sys1m	Dia1m	Sys2m	Dia2m	Sys3m	Dia3m
2	51	100	71	112	76	121	85	102	69	114	72	118	80
1	64	105	82	116	88	125	93	109	85	114	88	120	93
...	...	...	...	...	...	...	...	...	...	...	...	...	...

und hier die Daten nach der Umstrukturierung:

Vpn	Sex	Alter	Dosis	Zeit	Sys	Dia
1	2	51	1	1	100	71
1	2	51	1	2	112	76
1	2	51	1	3	121	85
1	2	51	2	1	102	69
1	2	51	2	2	114	72
1	2	51	2	3	118	80
...	...	...	...		...	

### 5. 1. 2 Umstrukturierungen in R

Nachfolgend wird gezeigt, wie die drei o.a. Datensätze in R die erforderliche Struktur für Messwiederholungen erhalten. Hierzu dient die Funktion `reshape`. Gelegentlich ist es auch erforderlich, einen Datensatz im langen Format wieder zurück in das breite Format zu transformieren. Beispiele dazu sind in Kapitel 5.3.9 und für den Fall zweier Messwiederholungsfaktoren in Kapitel 11.5 zu finden.

#### Beispieldaten 4 (`winer518`):

- Zunächst erhält der Dataframe `winer518` eine Fallkennzeichnung, hier `Vpn` genannt. Dieser Schritt kann natürlich entfallen, wenn der Datensatz bereits eine Fallkennung besitzt.
- Geschlecht und `Vpn` müssen als „factor“ deklariert werden.
- Mittels der Funktion `reshape` bekommt der Dataframe die für Messwiederholungen erforderliche Struktur, wobei die abhängige Variable den Namen `score` und der Faktor den Namen `Zeit` erhalten.
- Das Ergebnis wird `winer518t` benannt.
- `Zeit` muss als „factor“ deklariert werden.

```
Vpn      <- 1:10
winer518 <- cbind(Vpn,winer518)
winer518 <- within(winer518,
                   {Geschlecht<-factor(Geschlecht); Vpn<-factor(Vpn)})
winer518t<- reshape(winer518, direction="long", timevar="Zeit",
                   v.names="score", varying=c("t1","t2","t3"), idvar="Vpn")
winer518t<- within(winer518t, Zeit<-factor(Zeit))
```

Der erzeugte Dataframe `winer518t` hat dann folgende Gestalt:

	Vpn	Geschlecht	Zeit	score
1.1	1	1	1	4
2.1	2	1	1	3
3.1	3	1	1	7
4.1	4	1	1	6
5.1	5	1	1	5
6.1	6	2	1	8
7.1	7	2	1	4
8.1	8	2	1	6
9.1	9	2	1	9
....	..	..	..	..

### Beispieldaten 5 (mydata5):

Zunächst einmal muss der Dataframe `mydata5` eine Fallkennung (`Vpn`) erhalten. Während `mydata5` zwei Messwiederholungsfaktoren beinhaltet, kann `reshape` nur einen verarbeiten. Die Funktion muss daher zweimal aufgerufen werden:

- Beim ersten `reshape`-Aufruf werden die Stufen des Faktors `Medikament` in Zeilen gewandelt, während die Stufen des Faktors `Aufgaben` als Variablen behandelt werden. Die umzustrukturierenden Variablen `v1`, ..., `v9` können einfach durch die lfd Nummer, hier 3:11 angegeben werden. Die neuen abhängigen Variablen werden `a1`, `a2`, `a3` genannt. Der erzeugte Dataframe erhält den Namen `mydata5a`.
- Beim zweiten `reshape`-Aufruf wird dann der Faktor `Aufgaben` umstrukturiert. Allerdings darf dann `Vpn` nicht mehr als ID-Variable spezifiziert werden, da die `Vpn`-Werte nach dem ersten Aufruf von `reshape` mehrfach vorkommen und daher nicht zur Identifikation herangezogen werden können. Es wird aber eine neue ID-Variable `id` angefügt, die verwendet werden kann. Die neue abhängige Variable wird `Fehler` genannt. Über den Parameter `times=1:3` werden die Werte des Faktors (`Medikament` bzw. `Aufgabe`) festgelegt. Der erzeugte Dataframe erhält den Namen `mydata5b`.
- Abschließend müssen noch die beiden Variablen `Medikament` und `Aufgabe` vom Typ „factor“ deklariert werden. Der erzeugte Dataframe erhält den Namen `mydata5t`.

```
Vpn <- 1:8
mydata5 <- cbind(Vpn, mydata5)
names(mydata5)[2] <- "Geschlecht"
mydata5 <- within(mydata5,
  {Vpn <- factor(Vpn); Geschlecht <- factor(Geschlecht)})
mydata5a <- reshape(mydata5, direction="long", varying=3:11, idvar="Vpn",
  timevar="Medikament", times=1:3, v.names=c("a1", "a2", "a3"))
mydata5b <- reshape(mydata5a, direction="long",
  varying=c("a1", "a2", "a3"), idvar="id",
  timevar="Aufgabe", times=1:3, v.names="Fehler")
mydata5t <- within(mydata5b, {Medikament <- factor(Medikament);
  Aufgabe <- factor(Aufgabe)})
```

Nach dem ersten Aufruf von `reshape` hat der Dataframe folgende Struktur:

	Vpn	Geschlecht	Medikament	a1	a2	a3
1.1	1	1	1	3	3	1
2.1	2	1	1	2	0	0
3.1	3	1	1	5	4	3
4.1	4	1	1	3	5	2



```

5.1  5          2          1  2  2  1
6.1  6          2          1  4  1  0
7.1  7          2          1  3  2  1
8.1  8          2          1  1  3  0
1.2  1          1          2  4  4  2
2.2  2          1          2  3  2  2
...  .          .          .  .  .  .

```

und nach dem zweiten Aufruf von `reshape` :

```

      Vpn Geschlecht Medikament Aufgabe Fehler id
1.1    1           1           1         1       3  1
2.1    2           1           1         1       2  2
3.1    3           1           1         1       5  3
4.1    4           1           1         1       3  4
5.1    5           2           1         1       2  5
6.1    6           2           1         1       4  6
7.1    7           2           1         1       3  7
8.1    8           2           1         1       1  8
9.1    1           1           2         1       4  9
10.1   2           1           2         1       3 10

```

### Beispieldaten 6 (winer568):

Da `winer568` nur einen Messwiederholungsfaktor beinhaltet, erfolgt die Umstrukturierung ähnlich wie oben gezeigt für `winer518`:

```

Vpn      <- 1:12
winer568 <- cbind(Vpn,winer568)
winer568t <- reshape(winer568, direction="long", timevar="Zeit",
                    v.names="x", varying=c("V1","V2","V3","V4"), idvar="Vpn")
winer568t <- within(winer568t, {A<-factor(A); B<-factor(B);
                    Zeit<-factor(Zeit); Vpn<-factor(Vpn) })

```

Der erzeugte Dataframe `winer568t` hat dann folgende Gestalt:

```

      A B Zeit x Vpn
1.1  1 1   1  1  1
2.1  1 1   1  0  2
3.1  1 1   1  3  3
4.1  1 2   1  2  4
5.1  1 2   1  1  5
6.1  1 2   1  3  6
7.1  2 1   1  1  7
8.1  2 1   1  1  8
9.1  2 1   1  1  9
10.1 2 2   1  2 10
11.1 2 2   1  3 11
12.1 2 2   1  2 12
.....

```

### Hinweis:

Soll ein Datensatz zuerst in das lange Format, danach wieder in das breite Format zurücktransformiert werden, z.B. um alle Werte einer Variablen, Fälle mit den Messwiederholungen, in Ränge zu wandeln, so ist das bequem ohne die Funktion `reshape` möglich: Es werden die Messwiederholungsvariablen aus dem Dataframe "herausgenommen" und in eine Matrix, danach in einen Vektor gewandelt, der dann in Ränge transformiert werden kann. Nach der Zu-

rückwandlung eine eine Matrix sind allerdings Variablennamen etc. verloren. Nachfolgend ein Beispiel mit dem o.a. Datensatz

```
matrix(rank(as.vector(as.matrix(winer568[,3:6]))),12,4)
```

## 5.2 Voraussetzungen der parametrischen Varianzanalyse

Hier geht es zunächst einmal um Versuchspläne, die ausschließlich abhängige Stichproben beinhalten, also ohne Gruppierungsfaktoren. Für die 1-faktorielle Varianzanalyse lautet das Modell dann für einen Faktor C mit  $J$  Messwiederholungen/Stufen - nachfolgend wird gelegentlich auch wieder die Anzahl mit  $K$  bezeichnet:

$$x_{jm} = \mu + \gamma_j + \pi_m + e_{jm} \quad (j=1, \dots, J \quad m=1, \dots, n) \quad (5-1)$$

wobei  $n$  die Anzahl der Merkmalsträger/Versuchspersonen ist. Gegenüber dem entsprechenden Modell ohne Messwiederholungen (vgl. Kapitel 4.1) gibt es einen personenspezifischen Effekt:  $\pi_m$ . Die Aufgabe der Varianzanalyse ist die Prüfung gleicher Gruppenmittelwerte, d.h.  $H_0$  lautet wie in Kapitel 4.1:

$$\mu_1 = \mu_2 = \dots = \mu_J$$

was in der Terminologie des o.a. Modells äquivalent ist zu:

$$\gamma_1 = \gamma_2 = \dots = \gamma_J = 0$$

Es sei darauf aufmerksam gemacht, dass die nichtparametrischen Tests eine andere Nullhypothese  $H_0$  haben: die Verteilungsfunktion  $F_i(x)$  von  $x$  ist in allen Gruppen dieselbe:

$$F_1(x) = F_2(x) = \dots = F_J(x)$$

woraus zu folgern ist, dass bei Ablehnung von  $H_0$  nicht notwendigerweise sich die Mittelwerte  $\mu_i$  unterscheiden, sondern möglicherweise andere Verteilungsparameter wie Varianz oder Schiefe.

Die Voraussetzungen betreffen wiederum die Normalverteilung der Residuen und die Varianzhomogenität. Schaut man in die Lehrbücher, so wird dort kaum das Thema Normalverteilung behandelt, sondern im Wesentlichen die Varianzhomogenität, da die, im Gegensatz zur Analyse ohne Messwiederholungen, eine sehr viel größere Bedeutung hat.

### Normalverteilung

Doch zunächst zur Normalverteilung. Wenn auch das mathematische Modell auf einer multivariaten Normalverteilung der  $J$  Variablen basiert, so genügt es in der Praxis, die Residuen zu überprüfen. Bei Varianzanalysen mit Messwiederholungen gibt es mehrere Residuen, denn jeder Test eines Wiederholungsfaktors hat seine eigene Fehlervarianz (und die ihr zugrunde liegenden Residuen), über die die Effektvarianz beurteilt wird. Diese Streuungen müssen alle aus normalverteilten Grundgesamtheiten kommen. Dazu sind zwei Tests erforderlich:

Zum einen müssen die Residuen  $e_{im}$  geprüft werden. Hier genügt es nicht, die Abweichungen vom Zellenmittelwert  $\bar{x}_j$  zu betrachten, vielmehr müssen die personenspezifischen Abweichungen ebenfalls berücksichtigt werden. Dazu muss von den Abweichungen  $x_{jm} - \bar{x}_j$  noch  $\pi_m$  subtrahiert werden. (Werden diese nicht subtrahiert, können extreme Werte der  $\pi_m$  zu Abweichungen der Residuen von der Normalverteilung führen.) Auch hier entspricht es weder der Forderung, noch ist es praktikabel, die Voraussetzung für jede Messwiederholung bzw. Zelle separat zu überprüfen. Vielmehr sollte man alle Residuen zu einer Variablen zusammenfassen und analysieren. Auf die Ermittlung der Residuen wird in Kapitel 5.3.1 näher eingegangen. Al-

ternativ können auch die Residuen auf eine multivariate Normalverteilung überprüft werden. Ein Beispiel dazu folgt in Abschnitt 5.3.9. Mehr zur allgemeinen Überprüfung auf Normalverteilung im Kapitel 1.6.

Zum anderen müssen die personenspezifischen Abweichungen  $\pi_m$  (*Personeneffekt*) auf Normalverteilung überprüft werden. Diese errechnen sich als Mittelwerte aller Messwiederholungen einer Versuchsperson, wovon noch der Mittelwert  $\bar{\pi}$  abzuziehen ist. Für die Überprüfung kann allerdings der letzte Schritt entfallen, da er für die Verteilungsform nicht relevant ist.

Ergeben beide Tests keine Abweichungen, so können alle Residuen als normalverteilt angenommen werden, da sich diese aus den beiden o.a. Residuen zusammensetzen. Zur Prüfung kann wieder zum einen der Shapiro-Wilks-Test, zum anderen grafische Verfahren herangezogen werden. Durch die bei R erforderliche Umstrukturierung der Daten, ist es dort bequem, eine globale Residuen-Variable zu bestimmen und zu untersuchen. Bei SPSS bedarf es dazu etwas mehr Aufwand. Mehr dazu im Kapitel 5.3.1.

### Varianzhomogenität - Sphärität

Dazu kommt wieder die Voraussetzung der Varianzhomogenität, in diesem Fall der  $J$  zu vergleichenden Variablen:  $\sigma_1^2 = \dots = \sigma_J^2$ . (Allerdings nur für den Fall  $J > 2$ . Denn im Fall  $J = 2$  kann zum Vergleich der beiden Variablen einfach deren Differenz verwendet werden.) Da diese Voraussetzung einerseits sehr essentiell ist, zum anderen die Prüfung nicht trivial ist, soll ihr hier mehr Raum eingeräumt werden. Diese umfasst allerdings hier mehr als die Gleichheit der Varianzen, sondern stellt auch Anforderungen an die Kovarianzen. D.h. die Kovarianzmatrix  $\Sigma$  ( $\sigma_{ij}$ ,  $i, j=1, \dots, J$ ) der Variablen  $x_j$  muss besondere Eigenschaften haben.

Damit die F-Tests der Varianzanalyse korrekt sind, muss  $\Sigma$  *zirkulär* (*circular*) sein, das entspricht in etwa der Bedingung

$$\sigma_{x_1 - x_2}^2 = \sigma_{x_1 - x_3}^2 = \sigma_{x_2 - x_3}^2 = \dots$$

d.h. die Varianzen von allen Differenzen je zweier Variablen sind gleich. Diese Bedingung ist nicht leicht nachzuvollziehen. Es gibt aber noch eine andere Bedingung, die *Compound Symmetry*, die zwar mehr als gefordert beinhaltet, aber „verständlicher“ ist. Bei dieser wird gefordert, dass zum einen alle  $J$  Varianzen gleich sind, und zum anderen die Kovarianzen - und damit die Korrelationen, was aber äquivalent ist - je zweier (verschiedener) Variablen gleich sind. Die Prüfung von  $\Sigma$  auf Zirkularität (*circularity*) erfolgt über eine andere Matrix-Eigenschaft: *Sphärität*, die gegeben ist, wenn alle diagonalen Werte (hier Varianzen) gleich und alle übrigen (hier Kovarianzen) null sind, mathematisch  $\Sigma = \sigma I$  (mit einer Einheitsmatrix  $I$ ). Und zwar ist  $\Sigma$  zirkulär, wenn mittels einer orthonormalen Kontrastmatrix  $C$  folgende Bedingung erfüllt werden kann:

$$C\Sigma C' = \sigma I$$

Dies ist ausführlich, gut verständlich und mit Beispielen in Winer (1991, S. 246 ff) beschrieben. Diese Bedingung der Sphärität wird für jeden der Tests der Messwiederholungsfaktoren gefordert. Liegt also z.B. ein Design mit zwei Messwiederholungsfaktoren C und D vor, so ist ein entsprechender Test für die Effekte von C, D und C\*D durchzuführen.

Die Prüfung der Sphärität ist von großer Bedeutung, da zahllose Studien gezeigt haben (z.B. Boik, 1981), dass eine Verletzung dieser Voraussetzung die F-Tests der Varianzanalyse stark beeinträchtigt. D.h. zum einen erhöht sich das Risiko des Fehlers 1. Art dramatisch, zum anderen nimmt die Teststärke deutlich ab. Zur Prüfung der Sphärität wird allgemein der *Mauchly-*

*Test* verwendet (siehe z.B. Winer, S. 255), so auch standardmäßig in R und SPSS. Dieser Test hat allerdings im Vergleich zu einigen anderen Tests Nachteile: Zum einen reagiert er empfindlich auf Abweichungen von der multivariaten (!) Normalverteilung der  $J$  zu vergleichenden Variablen, und zum anderen hat er eine geringe Power, insbesondere für kleinere bis normale Stichproben. So hat z.B. Boik (1981) nachgewiesen, dass bei einem  $n=40$  die Power unter 0.2 liegt, d.h. eine vorhandene Abweichung von der Sphärität nur in weniger als 20 von 100 Fällen nachgewiesen werden kann. Daher ist er nur zweite Wahl. Es gibt zwar zahlreiche andere Testverfahren, doch leider sind Vergleiche rar. Insbesondere basieren fast alle ebenfalls auf der multivariaten Normalverteilung. Zu erwähnen sind folgende, für die zumindest in R Funktionen zur Verfügung stehen, z.B. in der Funktion `check.sphere` (siehe Anhang 3.6):

- John's V-Test und Nagao's Test (beide beschrieben in Wang & Yao, 2013), wobei Nagao für John's Test eine exaktere p-Wert-Berechnung durchführt, sowie eine Variante von Li & Yao (2016), die auf der etwas allgemeineren elliptischen Verteilung basiert und über den Exzess (Kurtosis) die Abweichung von der Normalverteilung berücksichtigt und somit einen weiteren Anwendungsbereich abdeckt,
- der Likelihood Ratio-Test, auf dem letztlich der Mauchly-Test beruht, sowie eine Variante von Muirhead & Waternaud (1980), die auf der etwas allgemeineren elliptischen Verteilung basiert (s.o.),
- ein Test auf Compound Symmetry, beschrieben im Buch von Winer (S. 517).
- Darüber hinaus gibt es relativ neue nichtparametrische Verfahren, die auf räumlichen Vorzeichen und Rängen basieren. Die Methode ist nicht leicht zu verstehen (siehe z.B. Sirkiä, 2007, sowie Zou et al., 2014). Immerhin gibt es R-Funktionen zum Test auf Sphärität in den Paketen `spatialNP` und `MNM`.

Aber insbesondere die drei zuerst genannten Verfahren sowie das von Sirkiä neigen - im Gegensatz zu Mauchly's Test - zum anderen Extrem: bei größerem  $n$  (etwa ab 50) reagieren sie sehr liberal. Statistische Vergleiche der Verfahren, die eine Auswahl erleichtern könnten, gibt es praktisch nicht. Nimmt man als Maßstab, wie gut sie die Abweichung von  $\epsilon$  vom Idealwert 1 (für Sphärität) messen, z.B. über die Korrelation von  $\epsilon$  mit dem p-Wert, so schneiden, je nach zugrunde liegender Verteilung von  $x$ , die Verfahren von Nagao, Li & Yao sowie Muirhead & Waternaud am besten ab. Dagegen zeigen die neueren, auf räumlichen Vorzeichen basierenden Tests kaum einen Zusammenhang mit  $\epsilon$  auf. Mehr dazu bei Lüpsen (Luepsen, 2020c). Allgemein gilt John's Test, und damit auch die o.a. Varianten, als der Empfehlenswerteste. Es sei noch darauf hingewiesen, dass für diese Tests die Anzahl der Beobachtungen  $n$  größer als die Anzahl der Messwiederholungen sein muss. Die Prüfung beider Voraussetzungen in R bzw. SPSS wird in Kapitel 5.3.1 beschrieben.

Der Test auf Sphärität dient in erster Linie dem Test auf Gleichheit der Varianzen der Messwiederholungsvariablen. Huynh & Feldt (1970) haben gezeigt, dass Sphärität nicht unbedingt eine Gleichheit der Korrelationen der Messwiederholungsvariablen  $\rho_{x_1x_2} = \rho_{x_2x_3} = \rho_{x_3x_4} = \dots$  impliziert wie etwa bei der compound symmetry gefordert. Eine Ungleichheit der Korrelationen führt allerdings (bei allen Methoden) zu leicht erhöhten Typ I Fehlerraten, wie z.B. Garcia et al. (2010) gezeigt haben. Eigene Simulationen haben jedoch gezeigt, dass alle o.a. Sphäritätstests auch auf Ungleichheit der Korrelationen ansprechen. Dabei schneidet erwartungsgemäß der o.a. Test auf compound symmetry am besten ab, auch für kleine  $n$ , wobei gegebenenfalls vorher die Messwiederholungsvariablen auf Varianz 1 zu transformieren sind, damit der Test nicht auf heterogene Varianzen reagiert.

## Alternativen

Auch hier stellt sich die Frage: Was ist zu tun, wenn eine der Voraussetzungen nicht erfüllt ist? Die in Kapitel 4.1 angeführte Robustheit der Verfahren hinsichtlich Abweichungen von der Normalverteilung gilt hier ganz besonders, da keine unterschiedlichen  $n_i$  vorliegen. Abweichungen von der Varianzhomogenität, hier von der Sphärität, sind dagegen gravierender, können aber statistisch aufgefangen werden. Ein Maß für die Abweichung ist das von Box errechnete  $\epsilon$  (mit  $1/(J-1) \leq \epsilon \leq 1$ ), und  $\epsilon=1$  im Fall perfekter Sphärität). Sowohl *Geisser & Greenhouse* als auch *Huynh & Feldt* haben auf diesem  $\epsilon$  basierende modifizierte F-Tests entwickelt, die auch bei Abweichungen von der Sphärität angewandt werden können. (Greenhouse-Geisser verwenden das Box  $\epsilon$ , daher häufig mit GG bezeichnet, während das  $\epsilon$  von Huynh & Feldt etwas modifiziert wurde, größer als das Box  $\epsilon$  ist, und meist mit HF bezeichnet wird.) Hierbei werden (wie häufig in der Statistik, z.B. bei der Welch-Approximation für den klassischen t-Test) die Zähler- und Nenner-Freiheitsgrade des F-Tests entsprechend der Abweichung  $\epsilon$  verkleinert. Der F-Wert selbst bleibt davon unberührt. Als Konsequenz daraus reagiert der F-Test konservativer, je stärker die Abweichung  $\epsilon$  ist. Von diesen beiden alternativen Tests ist der von Geisser & Greenhouse der konservativere. In SPSS (GLM Messwiederholungen) werden sowohl der Mauchly-Test als auch beide modifizierten F-Tests automatisch ausgegeben. In R gibt es Funktionen, die den Mauchly-Test wie auch die F-Tests von Geisser & Greenhouse sowie von Huynh & Feldt ausgeben, u.a. `ezANOVA` in dem Paket `ez`.

Beasley (2002) hat in einer umfangreichen Studie gezeigt, dass zum einen das Aligned Rank Transform (ART) Verfahren auch bei Daten, die weder normalverteilt sind noch die Sphärität erfüllen, sowohl der Fehler 1. Art  $\alpha$  eingehalten wird, als auch der Fehler 2. Art unter Kontrolle bleibt. Auf der anderen Seite wird darauf hingewiesen, dass bei einer „einfachen“ Rangtransformation Verteilungseigenschaften meist erhalten bleiben, wenn auch in abgeschwächter Form. (Hierauf wird auch von Fan (2006) aufmerksam gemacht.) D.h. dass z.B. bei Anwendung des Rank transform Tests (RT, ART und INT) bei Varianzanalysen mit Messwiederholungen eine Korrektur der Freiheitsgrade nach Huynh-Feldt oder Greenhouse-Geisser angebracht ist, wie dies von Beasley und Zumbo (2009) propagiert wird. Das Ergebnis des Mauchly-Tests auf Sphärität interessiert in dem Zusammenhang nicht, da dessen Voraussetzungen ohnehin kaum erfüllt sein werden. Das Verhalten der Kovarianzmatrizen, um die es ja bei der Sphärität geht, ist von Bryan (2009) ausführlich im Zusammenhang mit Rangtransformationen untersucht worden, ist aber zu umfangreich, um hier kurz wiedergegeben zu werden.

Häufiger wird auch vorgeschlagen, zum Test eines Messwiederholungsfaktors anstatt der klassischen univariaten Tests einen multivariaten Test (vgl. Abschnitt 2.12) zu verwenden. Der Vorteil: Dieser verlangt nicht die Voraussetzung der Varianzhomogenität (Sphärität). Der Nachteil: Es wird eine multivariate Normalverteilung gefordert. D.h. der univariate F-Test ist generell robuster gegenüber Verletzungen der Normalverteilung als ein multivariater Test (vgl. Lei et al., 2004). Darüber hinaus hat die Fehlervarianz des multivariaten Tests weniger Freiheitsgrade als die des univariaten Tests, was letztlich zu einer schwächeren Teststärke führt. Dieses Prozedere ist auch ausführlich bei Beasley & Zumbo (2009) beschrieben. SPSS gibt übrigens bei Analysen mit Messwiederholungen immer zuerst die Ergebnisse der multivariaten Varianzanalyse aus. Auf dieser Methode basiert das in Kapitel 2.12 erwähnte Verfahren von Koch, der diese multivariate Analyse auf Rangdaten überträgt und daraus  $\chi^2$ -Tests konstruiert.

Die oben angeführten Methoden sind im Wesentlichen für metrische abhängige Variablen und als Alternative zum F-Test gedacht. Einige der in diesem Kapitel vorgestellten Verfahren verwenden  $\chi^2$ -Tests und sind gut geeignet für ordinale Variablen, so z.B. der Puri & Sen-Test, das verallgemeinerte Kruskal-Wallis-Friedman-Verfahren (KWF), das verallgemeinerte van der

Waerden- und das Koch-Verfahren. Wie schon in Kapitel 2.16 erwähnt, verlangen diese zwar keine Sphärität, jedoch etwa gleiche Verteilungsformen, d.h. u.a. gleiche Varianzen der Messwiederholungsvariablen. Dies mittels eines der o.a. Sphäritäts-Tests zu überprüfen, wäre zu viel. Besser ist ein multipler nichtparametrischer Test auf Gleichheit abhängiger Varianzen, z.B. ein Levene-like Test. D.h. es werden für jede Variable die absoluten Differenzen vom Median berechnet, die dann mit einer Varianzanalyse verglichen werden. Dieser und einige andere sind bei R.R. Wilcox (1989) beschrieben, wobei Wilcox entweder den Quade-Test (vgl. Abschnitt 2.11.1) oder die Friedman-Anova favorisiert. Beispiele dazu kommen in den Kapiteln 5.3.4 und 5.4.3. Paarweise Varianzvergleiche sind im R-Paket `PairedData` enthalten. Damit kann man entweder alle  $J$  Streuungen paarweise oder die kleinste gegen die größte vergleichen. Allerdings gibt es im Fall heterogener Varianzen keine alternativen Verfahren.

Der Vollständigkeit wegen sei noch erwähnt, dass es auch Modelle für Anovas mit Messwiederholungen gibt, die andere Strukturen der Varianz-Kovarianzmatrix als die o.a. Sphärität voraussetzen, so z.B. autoregressive und unstrukturierte. R bietet dafür auch mit der Funktion `gls` im Paket `nlme` Lösungen.

Gute Erläuterungen der Voraussetzungen zu Varianzanalysen bieten der Klassiker B.J. Winer (1991) und R.N. Cardinal (2004). Beide gehen jedoch nicht auf Details zur Überprüfung der Normalverteilung ein.

## 5.3 Die 1-faktorielle Varianzanalyse

### 5.3.1 Parametrischer Test und Prüfung der Voraussetzung

An den Beispieldaten 4, allerdings hier ohne Berücksichtigung der Gruppenstruktur, soll ein Vergleich der Einstellung zu den 3 Zeitpunkten mittels einer parametrischen Varianzanalyse durchgeführt und die Prüfung der Voraussetzungen, Varianzhomogenität und Normalverteilung der Residuen, demonstriert werden.

Zur Berechnung der Residuen gibt es folgende Möglichkeit: Der oder die Messwiederholungsfaktoren C, D,.. werden als Gruppierungsfaktoren gehandhabt. Dazu muss der Datensatz umstrukturiert werden, indem die Messwiederholungen in Fälle gewandelt werden. (Dies ist in R ohnehin für Analysen mit Messwiederholungen erforderlich.) Dann wird folgendes Modell (*ohne* Messwiederholungen) analysiert:

$$C + V_{pn} \quad \text{bzw.} \quad C * D + V_{pn}$$

wobei  $V_{pn}$  die Fallkennung, z.B. Versuchspersonennummer, ist. Die Residuen dieses Modells sind die Residuen des Modells mit Messwiederholungen auf C (und D).

Dies ist zwar prinzipiell auch bei SPSS möglich, verursacht aber wegen der erforderlichen Umstrukturierung etwas Aufwand. SPSS gibt allerdings für jede Messwiederholungsvariable  $x_j$  andere Residuen aus:  $e'_{jm} = x_{jm} - \gamma_j$ . Aus dem Modell 5-1 ergibt sich für diese  $e'_{jm} = \pi_m + e_{jm}$ , d.h. um die Residuen  $e_{jm}$  zu erhalten, müssen von den  $e'_{jm}$  die  $\pi_m$  subtrahiert werden. Die erforderlichen Schritte sind dann:

- Speichern der Residuen:  $e'_{jm}$ ,
- Ermitteln des Personeneffekts  $\pi_m$  aus  $p_m = \left( \sum_j x_{jm} \right) / I$  und  $\pi_m = (p_m - \bar{p})$ ,
- und schließlich  $e_{jm} = e'_{jm} - \pi_m$ .

(Die Subtraktion von  $\bar{p}$  von  $p_m$  zur Ermittlung von  $\pi_m$  kann entfallen, da sie für die Beurteilung der Residuen  $e_{jm}$  ohne Bedeutung ist.)

Für größere  $n$  ( $n > 20$ ) können diese  $e_{jm}$  für  $j=1, \dots, J$  auf Normalverteilung überprüft werden. Die  $J$  Testergebnisse, etwa die p-Werte  $p_1, \dots, p_J$  können z.B. über *Fishers combined probability test* (vgl. Anhang 2.5) zu einem Testergebnis zusammengefasst werden.

Für kleinere  $n$  müssten die  $J$  Variablen zu einer mit  $n \cdot J$  Werten zusammengefasst werden, entweder per copy & paste oder wieder mittels der aufwändigen Umstrukturierung. Dann sollte aber besser der erste oben beschriebene Weg gewählt werden.

### mit R:

Ausgangsbasis ist der in 5.1.2 erstellte Dataframe `winer518t`. Die Anova wird mit der Standardfunktion `aov` durchgeführt, wobei durch den Modellterm `Error(Vpn/Zeit)` die Messwiederholungen auf dem Faktor `Zeit` gekennzeichnet werden:

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

mit dem Ergebnis:

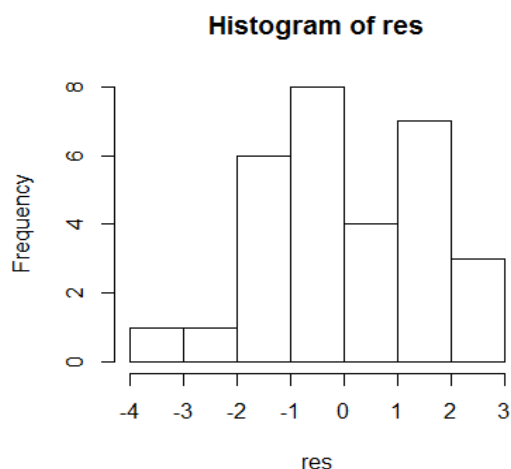
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	59.87	6.652		
Error: Vpn:Zeit					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zeit	2	58.07	29.033	7.926	0.0034 **
Residuals	18	65.93	3.663		

Tabelle 5-1

Für die Prüfung der Voraussetzungen bietet das Ergebnisobjekt `aov1` keine Möglichkeiten. Zunächst einmal zu den Residuen  $e_{jm}$ . Diese lassen sich, wie oben erläutert, bequem als Residuen eines Anova-Modells ohne Messwiederholungen ermitteln:

```
aov2<-aov(score~Zeit+Vpn, winer518t)
res<-aov2$residuals
hist(res)
shapiro.test(res)
```

mit folgenden Ergebnissen für die Tests auf Normalverteilung:



Shapiro-Wilk normality test  
 data: res  
 W = 0.9695, p-value = 0.5255

Das Histogramm zeigt mit einer leichten Linksschiefe eine geringe Abweichung von der Normalverteilung, die allerdings nicht als bösartig angesehen werden muss. Diese resultiert

zum Teil auch aus der zu großen Intervallzahl. Dahingegen weist der Shapiro-Test keine Abweichung aus.

Zur Überprüfung der Normalverteilung der versuchspersonenspezifischen Abweichungen  $\pi_m$  müssen diese ebenfalls erst ermittelt werden. Dazu muss man auf den ursprünglichen Dataframe `winer518` zurückgreifen und die Summen oder Mittelwerte der Variablen `t1`, `t2` und `t3` berechnen. Diese können dann wie üblich überprüft werden. Die Ergebnisse werden hier wegen der zu geringen Fallzahl ( $n=10$ ) nicht wiedergegeben.

```
pm <- (winer518$t1 + winer518$t2 + winer518$t3)/3
hist(pm)
shapiro.test(pm)
```

Zur Überprüfung der Varianzhomogenität, in diesem Fall also der Spherizität, gibt es, wie oben aufgeführt, mehrere Tests, u.a. den bekannten Mauchly-Test. Hierzu bieten sich mehrere Funktionen an. Zunächst die Funktion `mauchly.test`. Diese verlangt zwar als Eingabe ein `lm`- oder ein `SSD`-Objekt, doch der Aufruf ist einfach, solange man sich keine Gedanken über deren Bedeutung macht. Basis ist hier der ursprüngliche Datensatz im `wide`-Format:

```
mauchly.test(lm(as.matrix(winer518[,3:5])~1), X=~1)
```

```
data: SSD matrix from lm(formula = as.matrix(winer518[, 3:5]) ~ 1)
W = 0.64415, p-value = 0.1722
```

Das Ergebnis ( $p \sim 0.172$ ) zeigt keine Abweichung von der Spherizität an.

Vom Autor wird die Funktion `check.sphere` angeboten (siehe Angang 3), die neben dem Mauchly-Test noch die anderen im vorigen Abschnitt aufgeführten Verfahren durchführt. Die Funktion kann wahlweise auf den Datensatz im `wide`-Format (`winer518`) oder im `long`-Format (`winer518t`) angewandt werden:

```
check.sphere(winer518[,3:5]) # wide Format
with(winer518t, check.sphere(score, trial=Zeit, id=Vpn)) # long Format
```

```
$results
      statistic df    p value
Mauchly test    3.5185467  2 0.17216993
Likelihood Ratio 3.9583650  5 0.55542585
John's test chisq 6.2422936  2 0.04410656
John's test beta  0.1779233  5 0.41777756
John-Nagao      3.5584657  5 0.59354348
Muirhead & Waternaud 6.9438006  5 0.22485359
compound symmetry 3.6785211  4 0.45125761

$Box.epsilon
[1] 0.7375466
```

Wie oben angedeutet weichen die Testergebnisse deutlich voneinander ab. Dabei gelten John's und Nagao's Test als die empfindlichsten. Eine Empfehlung kann leider nicht gegeben werden (vgl. Lüpsen, 2020c).

Weiterhin gibt es das Paket `SpatialNP`, das sich neuen nichtparametrischen auf räumlichen Vorzeichen und Rängen basierenden Verfahren widmet, worin mit der Funktion `sr.sphere.test` ein Test auf Spherizität enthalten ist. Auch dieses Verfahren reagiert



empfindlich auf Abweichungen von der Sphärität. Die Funktion verlangt die Datenmatrix im wide-Format. Ein- und Ausgabe:

```
library(SpatialNP)
sr.sphere.test(winer518[,3:5])
```

```
Test of sphericity using spatial signs
Q.2 = 34.671, df = 5, p-value = 1.75e-06
alternative hypothesis: true shape is not equal to diag(3)
```

Praktischer ist die Benutzung der Funktion `ezANOVA` aus dem Paket `ez`, bei der Mauchly's Test bei Varianzanalysen im Fall von Messwiederholungen automatisch ausgegeben wird:

```
library(ez)
ezANOVA(winer518t, score, Vpn, within=Zeit)
```

```
$ANOVA
  Effect DFn DFd      F      p p<.05      ges
2   Zeit    2  18 7.926188 0.003397427 * 0.3158086

$`Mauchly's Test for Sphericity`
  Effect      W      p p<.05
2   Zeit 0.6441534 0.1721699

$`Sphericity Corrections`
  Effect      GGe      p[GG] p[GG]<.05      HFe      p[HF] p[HF]<.05
2   Zeit 0.7375466 0.00859794 * 0.8472485 0.005821856 *
```

Tabelle 5-2

Die ersten Zeilen enthalten die schon oben erzielte Varianzanalyse (vgl. Tabelle 5-1). Anzu-merken ist, dass darin „ges“ die *generalized effect size* (Effektgröße  $\eta^2$ ) ist. Darunter das Ergebnis des Tests von Mauchly ( $p \sim 0.172$ ), das keine Signifikanz und somit Varianzhomogenität zeigt. Die letzten Zeilen bieten für den Fall heterogener Varianzen die beiden alternativen Signifikanzberechnungen für die Varianzanalyse von Geisser & Greenhouse (GG) sowie Huynh & Feldt (HF), jeweils mit dem Zusatz „e“ für den Korrekturfaktor der Freiheitsgrade  $\epsilon$  bzw. dem Zusatz „p“ für die Irrtumswahrscheinlichkeit. Der unter GGe angegebene Korrekturfaktor 0.7375 ist das eigentliche Box  $\epsilon$ .

Auch die bereits in Kapitel 4.3.2 vorgestellte Funktion `aov_ez` aus dem Paket `aFex` kann bequem Versuchspläne mit Messwiederholungen analysieren, die wie `aov` und `ezANOVA` den transformierten Dataframe, hier `winer518t`, verlangt. Ein Beispiel dazu ist in Kapitel 6.2 zu finden.

Was die Entscheidung über die Annahme der Sphärität anbetrifft, so ist der Benutzer hier in Anbetracht der recht unterschiedlichen Ergebnisse für die verschiedenen Tests vorerst auf sich alleine gelassen, da derzeit noch keine Vergleiche der verschiedenen Verfahren vorliegen.

### mit SPSS:

Varianzanalysen mit Messwiederholungen erhält man in SPSS über das Menü „Allgemeines lineares Modell -> Messwiederholung“.

Die Anweisungen für den Beispieldatensatz 4 mit Speicherung der Residuen lauten:

```
GLM t1 t2 t3
  /wsfactor=Zeit 3 polynomial
  /save=resid
  /wsdesign=Zeit.
```

Die Ausgabe umfasst u.a. die zunächst interessierende Varianzanalyse in folgender Tabelle:

Tests der Innersubjekteffekte						
Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	58,067	2	29,033	7,926	,003
	Greenhouse-Geisser	58,067	1,475	39,365	7,926	,009
	Huynh-Feldt	58,067	1,694	34,268	7,926	,006
	Untergrenze	58,067	1,000	58,067	7,926	,020
Fehler(Zeit)	Sphärizität angenommen	65,933	18	3,663		
	Greenhouse-Geisser	65,933	13,276	4,966		
	Huynh-Feldt	65,933	15,250	4,323		
	Untergrenze	65,933	9,000	7,326		

Tabelle 5-3

Die „normale“ Signifikanzüberprüfung für den Faktor *zeit* ist in der Zeile „Sphärität angenommen“ abzulesen. Die beiden Zeilen „Greenhouse-Geisser“ und „Huynh-Feldt“ bieten alternative Tests für den Fall, dass die Voraussetzung der Sphärität, also der Varianzhomogenität, nicht erfüllt ist. Den Mauchly-Test zur Überprüfung dieser Voraussetzung enthält die folgende Tabelle:

Mauchly-Test auf Sphärizität							
Innersubjekteffekt	Mauchly- W	Approximiertes Chi-Quadrat	df	Sig.	Epsilon		
					Greenhouse -Geisser	Huynh -Feldt	Untergrenze
Zeit	,644	3,519	2	,172	,738	,847	,500

aus der hervorgeht ( $p \sim 0.172$ ), dass die Varianzhomogenität erfüllt ist. Die rechten Spalten „Epsilon“ enthalten den Korrekturfaktor der Freiheitsgrade  $\epsilon$  für den entsprechenden Test, der in der o.a. Varianzanalysetabelle zur Berechnung der Signifikanzen verwendet wird.

Die Überprüfung der Residuen auf Normalverteilung bei Messwiederholungen ist in SPSS mit etwas Aufwand verbunden. Zum einen gibt es die am Anfang dieses Kapitels beschriebene Möglichkeit über ein varianzanalytisches Modell ohne Messwiederholungen, was aber eine Umstrukturierung des Datensatzes erfordert. Ein Beispiel dazu folgt in Kapitel 6.2. Zum anderen kann man auf den Residuen  $e'_{im}$  aufbauen, die SPSS bei Messwiederholungsmodellen ausgibt. Dies soll hier kurz gezeigt werden.

Es wird für jede Messwiederholungsvariable ( $t_1, t_2, t_3$ ) eine Residuenvariable erzeugt: *Res\_1, Res\_2, Res\_3*. Von diesen muss nun zunächst der Personeneffekt  $\pi_m$  abgezogen werden, der allerdings vorher noch berechnet werden muss. Nachfolgend die Kommandos hierfür, wobei im zweiten Schritt der Mittelwert von  $\pi$  errechnet wird - hier einfach über *Descriptive* und Einsetzen des Wertes 4.27, alternativ über *Aggregate*. Allerdings ist, wie oben bemerkt, die Subtraktion des Mittelwert von  $\pi$  nicht erforderlich.

```
Compute Pi=Mean(t1,t2,t3).
Descriptive Variables=Pi.
```

```

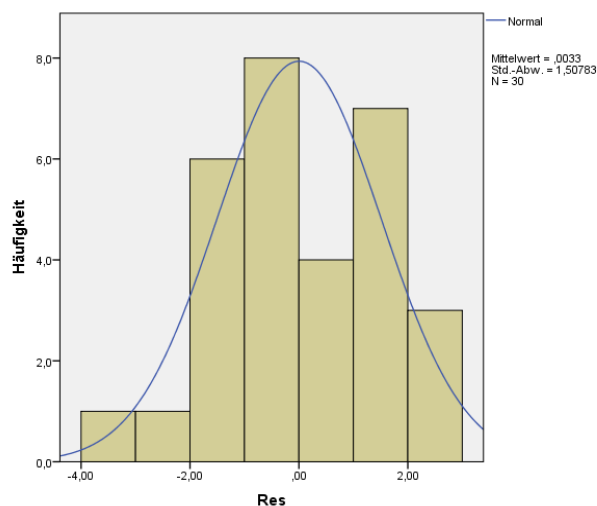
Compute R1 = Res_1 - (Pi-4.27) .
Compute R2 = Res_2 - (Pi-4.27) .
Compute R3 = Res_3 - (Pi-4.27) .

```

Bei größeren Stichproben könnte jede dieser Variablen separat auf Normalverteilung überprüft werden, nicht aber bei kleineren wie hier  $n=10$ . Weder ein Histogramm noch ein Test können hier ein klares Bild geben. Zwei der Möglichkeiten, die Residuenvariablen zu einer einzigen zusammenzufassen, sollen hier kurz skizziert werden.

Zum einen können im Dateneditor über copy & paste sämtliche Residuenvariablen (hier: R1, R2 und R3) zu einer zusammengefügt werden. Dies dürfte, insbesondere bei nicht zu großen Datensätzen, der einfachste Weg sein.

Alternativ wird der Datensatz umstrukturiert, so dass die Messwiederholungen zu Fällen werden, hier also die Variablen R1, R2 und R3 zu einer Variablen Res, deren Werte sich jeweils auf 3 Fälle verteilen. Die Vorgehensweise ist ausführlich im Anhang 1 beschrieben. Die Variable Res kann nun über ein Histogramm oder über den Shapiro-Wilk-Test (erhältlich über das Menü „Deskriptive Statistiken -> Explorative Datenanalyse“ und dort bei „Diagramme“ „Normalverteilungsdiagramm mit Tests“ aktivieren) auf Normalverteilung überprüft werden.



Tests auf Normalverteilung						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Res	,120	30	,200*	,969	30	,526

Das Histogramm zeigt mit einer leichten Linksschiefe eine geringe Abweichung von der Normalverteilung, die allerdings nicht als bössartig angesehen werden muss. Diese resultiert zum Teil auch aus der zu großen Intervallzahl. Dahingegen weist der Shapiro-Test keine Abweichung aus.

### 5.3.2 Friedman-Test

Der Friedman-Test ist das nichtparametrische Pendant zur 1-faktoriellen Varianzanalyse mit Messwiederholungen. (Er wird gelegentlich auch irreführend als 2-faktorielle Varianzanalyse bezeichnet, da rein formal neben dem betrachteten Messwiederholungsfaktor noch der „Faktor“ Vpn in die Rechnung einfließt.) Der Algorithmus sieht so aus, dass zunächst innerhalb jeder

Vpn die Werte in Ränge 1,...,J (mit J Anzahl der Stufen), sog. *Friedman-Ränge*, transformiert werden, danach mit diesen wie gewohnt weitergerechnet wird, aber zum Schluss anstatt eines F-Tests ein  $\chi^2$ -Test durchgeführt wird (vgl. auch Kapitel 2.5 und 5.3.3). An den Beispieldaten 4, hier ohne Berücksichtigung der Gruppenstruktur, soll die Berechnung gezeigt werden.

### mit R:

Die Funktion `friedman.test` kann auf zwei verschiedene Arten benutzt werden:

- zum einen mittels Eingabe der zu analysierenden Datenmatrix (Dataframe `winer518`), allerdings nicht vom Typ „data.frame“, sondern vom Typ „matrix“ (Umwandlung z.B. über `as.matrix`), wobei die Daten die ursprüngliche Struktur haben müssen.
- zum anderen mittels Angabe eines Modells wie bei der Funktion `aoV`, wobei die Daten wie für `aoV` umstrukturiert sein müssen (Dataframe `winer518t` aus Kapitel 5.1.2),

Variante 1:

```
friedman.test (as.matrix(winer518[,3:5]))
```

Variante 2:

```
friedman.test (score~Zeit | Vpn, data=winer518t)
```

Die Ausgabe ist bei beiden natürlich identisch:

```
Friedman rank sum test
Friedman chi-squared = 9.5556, df = 2, p-value = 0.008415
```

### mit SPSS:

Hier muss beachtet werden, dass gegebenenfalls vorher das Skalenniveau der analysierten Variablen auf „Skala“ gesetzt wird. Die Syntax für den Friedman-Test sowie die Ausgabe:

```
Nptests /related test(t1 t2 t3) friedman(compare=pairwise).
```

#### Übersicht über Hypothesentest

Gesamtanzahl	10	Nullhypothese	Test	Sig.	Entscheidung
Teststatistik	9,556	1 Die Verteilungen von , and sind gleich.	Friedmans Zweifach-Rangvarianzanalyse verbundener Stichproben	,008	Nullhypothese ablehnen.
Freiheitsgrade	2				
Asymptotische Sig. (zweiseitiger Test)	,008	Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau ist ,05.			

Das Ergebnis ist zwar signifikant. Dennoch soll hier kurz noch die Iman & Davenport-Korrektur gezeigt werden (vgl. Formel 2-1 in Kapitel 2.1):

$$F = \frac{(10 - 1) \cdot 9,5556}{10 \cdot (3 - 1) - 9,5556} = 8,308$$

Dieser F-Wert hat 2 Zähler-FG und 20 Nenner-FG. Der entsprechende p-Wert: 0.00236, der tatsächlich etwas kleiner ausfällt als der p-Wert des Friedman-Tests.

### 5.3.3 rank transform (RT) und normal scores (INT)

Bei der einfachen *rank transform* (RT)-Analyse wird lediglich vor der Durchführung der parametrischen Varianzanalyse zunächst die abhängige Variable  $x$  über alle Messwiederholungen hinweg in Ränge  $R(x)$  transformiert. Beim einfachen *inverse normal transformation* (INT) werden anschließend zusätzlich die Ränge  $R(x_m)$  über die inverse Normalverteilung in normal scores umgerechnet:

$$n\text{score}_m = \Phi^{-1}(R(x_m)/(M+1))$$

wobei  $M$  die Anzahl aller Werte ist, also  $n \cdot J$  (mit  $n$  Anzahl der Merkmalsträger und  $J$  Anzahl der Messwiederholungen), und  $\Phi^{-1}$  die Umkehrfunktion der Normalverteilung. Die statistischen Tests bleiben unverändert. Beide Verfahren sind in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, nicht jedoch für Variablen mit beliebigen Eigenschaften. D.h. hat die untransformierte Variable  $x$  ungleiche Varianzen, so kann das auch noch für die transformierten Variablen  $R(x)$  und  $n\text{score}$  gelten. So ist es sinnvoll, auch  $R(x)$  bzw.  $n\text{score}$  auf Sphärizität zu überprüfen. Da allerdings fast alle zur Verfügung stehenden Tests selbst u.a. Normalverteilung voraussetzen, wären dessen Ergebnisse unter Vorbehalt zu interpretieren sind. Beasley und Zumbo (2009) propagieren daher, bei den F-Tests einfach eine der Korrekturen der Freiheitsgrade von Huynh-Feldt oder Greenhouse-Geisser vorzunehmen, ohne das Ergebnis des Sphärizität-Tests, z.B. Mauchly-Test, zu berücksichtigen. Das INT-Verfahren soll am Datensatz des Beispiels 4 für den Faktor `Zeit` demonstriert werden.

#### mit R:

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer518t`. Zunächst wird die Kriteriumsvariable `score` in Ränge (`rscore`) transformiert, anschließend diese in normal scores umgerechnet, wobei die Anzahl der Fälle `nc` einfließt. Die Varianzanalyse wird mit `ezANOVA` (Paket `ez`) durchgeführt, um neben dem Test von Mauchly auf Varianzhomogenität (Sphärizität) die adjustierten Signifikanztests von Geisser-Greenhouse und Huynh-Feldt zu erhalten:

```
library(ez)
nc      <- dim(winer518t)[1]
winer518t <- within(winer518t, rscore<-rank(score))
winer518t <- within(winer518t, nscore<-qnorm(rscore/(nc+1)))
ezANOVA(winer518t, nscore, Vpn, within=Zeit)
```

```
$ANOVA
  Effect DFn DFd      F      p p<.05      ges
2   Zeit    2  18 8.570491 0.002427323 * 0.309934

$`Mauchly's Test for Sphericity`
  Effect      W      p p<.05
2   Zeit 0.5617469 0.09957784

$`Sphericity Corrections`
  Effect      GGe      p [GG]      HFe      p [HF]
2   Zeit 0.6952879 0.007838653 0.7823034 0.00559601
```

Oben die Ergebnisse für das normal score (INT)-Verfahren. Danach ist die Varianzhomogenität zwar erfüllt ( $p \sim 0.100$ ). Dennoch liest man zweckmäßigerweise das Ergebnis für den Zeit-Effekt nicht im oberen ANOVA-Teil ( $p=0.0024$ ), sondern im unteren unter *Sphericity Corrections* ( $p[\text{HF}]$ ) ab ( $p=0.0056$ ) ab.

**mit SPSS:**

Wie in Kapitel 5.3.3 sind die folgenden Schritte erforderlich, um die Werte über die Messwiederholungen hinweg in Ränge transformieren zu können:

- Zunächst müssen für den Datensatz über das Menü „Daten -> Umstrukturieren“ die Messwiederholungen in Fälle transformiert werden (siehe dazu im Anhang 1.1.1).
- Die Variable `score` wird dann über das Menü „Transformieren -> Rangfolge bilden“ in Ränge umgerechnet. Ergibt Variable `Rscore`.
- Diese Variable `Rscore` wird nun in normal scores umgerechnet. Dazu muss noch vorab über `Aggregate` die Anzahl der Werte `nc` ermittelt werden, da die Ränge durch  $(n+1)$  dividiert werden. Die Ergebnisvariable wird `nscore` genannt.
- Danach muss der Datensatz wieder zurück in das „normale“ Format mit Messwiederholungen transformiert werden (vgl. 1.2). Dabei werden aus `nscore` wieder 3 Variablen `nscore.1`, `nscore.2`, `nscore.3`.
- Abschließend wird dann eine Varianzanalyse mit Messwiederholungen (Menü: „Allgemeines lineares Modell -> Messwiederholung“) für `nscore.1`, ... gerechnet.

Nachfolgend die Syntax für diese Schritte:

```
Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.

Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(score).
Rank Variables=score / rank into Rscore.
compute nscore=Idf.normal(Rscore/(nc+1),0,1).

Sort cases by Vpn Zeit.
casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

GLM nscore.1 nscore.2 nscore.3
  /WSfactor = Zeit 3 Polynomial
  /WSdesign Zeit
```

Nachfolgend zunächst der Test auf Varianzhomogenität, der zwar mit  $p=0,100$  gerade noch akzeptabel ist, aber ohnehin keine Rolle spielen sollte. Denn zweckmäßigerweise sollten die Ergebnisse für die Varianzanalyse (in der zweiten Tabelle) ohnehin einer der Zeilen mit den adjustierten Testergebnissen, z.B. Huynh-Feldt, entnommen werden.

Mauchly-Test auf Sphärität <sup>a</sup>						
Innersubjekt- effekt	Mauchly-W	Approximiertes Chi-Quadrat	df	Sig.	Epsilon <sup>b</sup>	
					Greenhouse- Geisser	Huynh-Feldt
Zeit	,562	4,614	2	,100	,695	,782

Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	6,909	2	3,454	8,570	,002
	Greenhouse-Geisser	6,909	1,391	4,968	8,570	,008
	Huynh-Feldt	6,909	1,565	4,416	8,570	,006
	Untergrenze	6,909	1,000	6,909	8,570	,017
Fehler(Zeit)	Sphärizität angenommen	7,255	18	,403		
	Greenhouse-Geisser	7,255	12,515	,580		
	Huynh-Feldt	7,255	14,081	,515		
	Untergrenze	7,255	9,000	,806		

Danach ist der Zeit-Effekt mit  $p=0,006$  signifikant.

### 5.3.4 Puri & Sen-Test

Bei dem Puri & Sen-Test werden wie bei der o.a. RT-Methode die beobachteten Werte über alle Erhebungseinheiten und Messwiederholungen hinweg in Ränge  $1, \dots, n \cdot J$  transformiert.

Folgende Schritte sind durchzuführen:

- Alle  $n \cdot J$  Werte werden in Ränge  $(1, \dots, n \cdot J)$  transformiert.
- Mit den Rängen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle wird folgender  $\chi^2$ -Test aufgestellt (vgl. Formel 2-7):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{(SS_{\text{Effekt}} + SS_{\text{Fehler}}) / (df_{\text{Effekt}} + df_{\text{Fehler}})}$$

wobei  $SS_{\text{Effekt}}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes (A),  $SS_{\text{Fehler}}$  die Sum of Squares des Fehlers ist sowie  $df$  die entsprechenden Freiheitsgrade.

- Der  $\chi^2$ -Wert ist dann in den Tabellen für den  $\chi^2$ -Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade ( $df_{\text{Effekt}}$ ) des entsprechenden F-Tests sind.
- Schließlich kann noch die Iman & Davenport-Korrektur (Formel 2-1) angewandt werden, falls der  $\chi^2$ -Test nicht signifikant war.

Die Überprüfung der Sphärizität entfällt, da kein F-Test, sondern ein  $\chi^2$ -Test durchgeführt wird. Statt dessen wird die Gleichheit der Varianzen mit einem von Wilcox (1989) vorgeschlagenen Levene-like-Test überprüft. Die dazu erforderlichen Schritte:

- Für jede Variable  $x_1, \dots, x_J$  wird der Median  $m_j$  ermittelt.
- Berechnen der absoluten Differenz  $d_j = \text{abs}(x_j - m_j)$
- Durchführung eines Friedman-Tests über die  $d_j$ .

Die Schritte sollen am Datensatz des Beispiels 4 demonstriert werden.

#### mit R:

Basis ist der oben in Kapitel 5.1.2 erstellte Dataframe `winer518t`. Zunächst werden die Werte `score` in Ränge `rscore` transformiert, für dann mittels `aov` eine Varianzanalyse durchgeführt wird:

```
winer518t <- within(winer518t, rscore<-rank(score))
summary(aov(rscore~Zeit+Error(Vpn/Zeit),winer518t))
```

```
Error: Vpn
      Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9  755.2   83.91

Error: Vpn:Zeit
      Df Sum Sq Mean Sq F value Pr(>F)
Zeit    2  698.6   349.3   8.369 0.00269 **
Residuals 18  751.2    41.7
```

Hieraus sind abzulesen:  $SS_{Effekt} = 698.6$  sowie  $SS_{Fehler} = 751.2$ . Daraus ergibt sich die Testgröße (L statistic):

$$\chi^2 = \frac{698,6}{(698,6 + 751,2)/(2 + 18)} = 9,64$$

die bei 2 FG auf dem 1%-Niveau signifikant ist.

Der Puri & Sen-Test kann auch mit der Funktion `np.anova` (vgl. Anhang 3.8) durchgeführt werden. Ein Beispiel ist in Abschnitt 5.4.3 zu finden.

Zur Überprüfung der Varianzen mit dem Levene-like-Test: Hier liefert die Funktion `ave` die Mediane für jede Erhebungseinheit:

```
winer518t <- within(winer518t,
  diff <- abs(score-ave(score,Zeit,FUN=median)))
with(winer518t,friedman.test(diff~Zeit | Vpn, data=winer518t))
```

```
Friedman rank sum test

data:  diff and Zeit and Vpn
Friedman chi-squared = 1.1875, df = 2, p-value = 0.5523
```

Hiernach können die Varianzen der Variablen  $x_1, \dots, x_J$  als gleich angesehen werden.

### mit SPSS:

Die Anweisungen sind weitgehend identisch mit denen des RT-Verfahrens im vorigen Abschnitt, lediglich sind mit GLM die Variablen `rscore.1, ..., rscore.3` anstatt `nscore.1, ..., nscore.3` zu analysieren, mit folgendem Ergebnis:

**Tests der Innersubjekteffekte**

Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	698,600	2	349,300	8,369	,003
Fehler(Zeit)	Sphärizität angenommen	751,233	18	41,735		

Hieraus sind abzulesen:  $SS_{Effekt} = 698.6$  sowie  $SS_{Fehler} = 751.2$ . Daraus ergibt sich die Testgröße (L statistic):

$$\chi^2 = \frac{698,6}{(698,6 + 751,2)/(2 + 18)} = 9,64$$

die bei 2 FG auf dem 1%-Niveau signifikant ist.



Zur Überprüfung der Varianzen mit dem Levene-like-Test: Hier werden zunächst mit `aggregate` die Mediane errechnet, in einer `do repeat`-Schleife die absoluten Differenzen ermittelt und schließlich über `Nptests` der Friedman-Test ausgeführt.

```
Aggregate
/outfile=* mode=addvariables
  /m1=median(t1)
  /m2=median(t2)
  /m3=median(t3) .
do repeat t=t1 to t3 / m=m1 to m3 / d=d1 to d3.
compute d=abs(t-m) .
end repeat.
Nptests /related test(d1 d2 d3) friedman(compare=pairwise) .
```

mit dem Ergebnis, dass die Varianzen als gleich anzusehen sind:

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilungen von d1, d2 and d3 sind gleich.	Friedmans Zweifach- Rangvarianzanalyse verbundener Stichproben	,552	Nullhypothese behalten.

### 5.3.5 van der Waerden

Bei dem Verfahren von *van der Waerden* werden ähnlich dem o.a. Puri & Sen-Test anstatt der „klassischen“ F-Tests die  $\chi^2$ -Tests wie beim Friedman-Tests gerechnet. Allerdings wird eine andere Transformation in Ränge vorgenommen als beim o.a. INT-Verfahren: Wie beim Friedman-Verfahren werden die Ränge 1,...,J fallweise vergeben (Friedman-Ränge).

Folgende Schritte sind durchzuführen:

- Für jede Erhebungseinheit (Versuchsperson) werden die Werte in Ränge (1,...,J) transformiert.
- Die Ränge werden in normal scores umgerechnet (vgl. Formel 2-2):  
 $nscore_m = \Phi^{-1}(R(x_m)/(J+1))$
- Mit diesen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle wird folgender  $\chi^2$ -Test aufgestellt (vgl. Formel 2-7):

$$\chi^2 = \frac{SS_{Effekt}}{(SS_{Effekt} + SS_{Fehler}) / (df_{Effekt} + df_{Fehler})}$$

wobei  $SS_{Effekt}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,  $SS_{Fehler}$  die Streuungsquadratsumme des Fehlers ist sowie  $df$  die entsprechenden Freiheitsgrade.

- Der  $\chi^2$ -Wert ist dann in den Tabellen für den  $\chi^2$ -Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade ( $df_{Effekt}$ ) des entsprechenden F-Tests sind.

Die Schritte sollen am Datensatz des Beispiels 4 für den Faktor `zeit` demonstriert werden. Die Überprüfung der Sphärität kann entfallen, da hier  $\chi^2$ - anstatt F-Tests durchgeführt werden. Statt dessen sollte ein Test auf Gleichheit der Varianzen, wie im vorigen Abschnitt gezeigt, durchgeführt werden.

**mit R:**

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer518t`. Zuerst wird mittels der Funktion `ave` die Variable `score` für jeden Wert von `Vpn` in Ränge `rscore` transformiert, diese anschließend in normal scores `nscore` umgerechnet. Der Dataframe wird um diese Variablen ergänzt. Für `nscore` wird dann eine Varianzanalyse durchgeführt:

```
rscore <- ave(winer518t$score, winer518t$Vpn, FUN=rank)
nscore <- qnorm(rscore/4) # Division durch J+1
winer518t <- cbind(winer518t, rscore, nscore)
summary(aov(nscore~Zeit+Error(Vpn/Zeit), winer518t))
```

Error: Vpn:Zeit					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zeit	2	3.847	1.9237	8.163	0.003 **
Residuals	18	4.242	0.2357		

Der  $\chi^2$ -Wert des van der Waerden-Tests, der 2 FG hat, errechnet sich nun leicht per Hand:

$$\chi^2 = 3.847 / ((3.847 + 4.242)/(2+18)) = 3.85 / 0.4045 = 9.52$$

Dieser ist auf dem 1%-Niveau signifikant (kritischer Wert: 9.2).

Der van der Waerden-Test kann auch mit der Funktion `np.anova` (vgl. Anhang 3.8) durchgeführt werden:

```
np.anova(score~Zeit+Error(Vpn/Zeit), winer518t, method=1)
```

generalized van der Waerden tests					
	Df	Sum Sq	Chisq	Pr(>Chi)	
Zeit	2	3.8475	9.5124	0.0085982	
Residuals Zeit	18	4.2419			

**mit SPSS:**

Wie im Kapitel 5.3.3 sind die folgenden Schritte erforderlich, um fallweise die Werte in Ränge transformieren zu können:

- Zunächst müssen für den Datensatz über das Menü „Daten -> Umstrukturieren“ die Messwiederholungen in Fälle transformiert werden (siehe dazu im Anhang 1.1.1).
- Die Variable `score` wird dann über das Menü „Transformieren -> Rangfolge bilden“ in Ränge umgerechnet, wobei im Feld „Sortieren nach“ die Variable `Vpn` eingetragen werden muss, damit die Rangbildung pro `Vpn` vorgenommen wird. Ergibt Variable `Rscore`.
- Diese Variable `Rscore` wird nun in normal scores umgerechnet. Dabei werden die Ränge durch  $(J+1)$ , hier also 4, dividiert. Die Ergebnisvariable wird `nscore` genannt.
- Danach muss der Datensatz wieder zurück in das „normale“ Format mit Messwiederholungen transformiert werden (vgl. 1.2). Dabei werden aus `nscore` wieder 3 Variablen `nscore.1`, `nscore.2`, `nscore.3`.
- Abschließend wird dann eine Varianzanalyse mit Messwiederholungen (Menü: „Allgemeines lineares Modell -> Messwiederholung“) für `nscore.1`, ... gerechnet.

Nachfolgend die Syntax für diese Schritte sowie die Anova-Tabelle:

```

Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.

Rank Variables=score by Vpn / rank into Rscore.
compute nscore=Idf.normal(Rscore/4,0,1).

Sort cases by Vpn Zeit.

Casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

GLM nscore.1 nscore.2 nscore.3
  /WSfactor =Zeit 3 Polynomial
  /WSdesign Zeit

```

Da hier nur die Quadratsummen interessieren, nicht aber die verschiedenen Testergebnisse in Abhängigkeit von der Sphärizität, wir hier nur jeweils die 1. Zeile wiedergeben:

Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	3,847	2	1,924	8,163	,003
Fehler(Zeit)	Sphärizität angenommen	4,242	18	,236		

Hieraus (Spalten „Quadratsumme“ bzw. „df“) wird der  $\chi^2$ -Wert des van der Waerden-Tests errechnet, der 2 FG hat:

$$\chi^2 = 3.847 / ((3.847 + 4.242)/(2+18)) = 3.85 / 0.4045 = 9.52$$

Dieser ist auf dem 1%-Niveau signifikant (kritischer Wert: 9.2).

### 5.3.6 ATS-Tests von Akritas, Arnold & Brunner

Den von Akritas, Arnold und Brunner entwickelten ATS-Test gibt es auch für Varianzanalysen mit Messwiederholungen. In R stehen dazu Funktionen in zwei Paketen zur Verfügung: u.a. `nparLD` im Paket `nparLD` mit näherungsweise Tests sowie `RM` im Paket `MANOVA.RM`, das die Fehlerwahrscheinlichkeiten auf Basis der Resampling-Methode errechnet, dafür allerdings sehr rechenintensiv ist. Eigene Simulationen zeigten allerdings, dass `RM` keinen Gewinn gegenüber `nparLD` bringt. Ein Beispiel zu dem Programm ist in Kapitel 6.8 zu finden. In SPSS gibt es derzeit keine Möglichkeit zur Anwendung dieses Verfahrens.

#### mit R:

Die 1-faktorielle Analyse mittels `nparLD` soll am Datensatz des Beispiels 4 gezeigt werden. Ausgangsbasis ist wieder der in 5.1.2 erstellte Dataframe `winer518t`. Die Analyse kann mittels zwei Funktionen erfolgen:

- `nparLD` ist eine universelle Funktion für alle verarbeitbaren Designs.
- `ld.f1` erlaubt fehlende Werte bei den Messwiederholungen, gibt einen Mittelwertplot aus sowie eine Reihe weiterer, hier allerdings nicht interessierender Statistiken aus.

Beide geben sowohl die WTS als auch die interessantere ATS aus. Die Ausgabe unterscheidet sich nicht hinsichtlich dieser Statistiken. Nachfolgend zunächst die Eingabe für beide Varianten, wobei zu beachten ist, dass bei `nparLD` trotz Angabe des Dataframes die Variablennamen nicht automatisch gefunden werden. Daher muss bei beiden Funktionen entweder jeder Variablenname zusammen mit dem Dataframe-Namen in der üblichen Form, z.B. `winer518t$score` angegeben werden oder mit `with(Dataframe,...)` ausgeführt werden:

```
library(nparLD)
with(winer518t, nparLD(score~Zeit, winer518t, Vpn))
with(winer518t, ld.f1(score, Zeit, Vpn, time.name="Zeit"))
```

Bei `ld.f1` muss die Variable zweimal angegeben werden: zum einen zur Identifikation des Faktors, zum anderen in “...” als Name des Faktors für die Ausgabe. Nachfolgend die Ausgabe von `nparLD`, die die Signifikanz des Friedman-Tests bestätigt:

```
Call:
score ~ Zeit

Wald-Type Statistic (WTS):
  Statistic df      p-value
Zeit  43.42399  2 3.720494e-10

ANOVA-Type Statistic (ATS):
  Statistic      df      p-value
Zeit  8.369437 1.433543 0.001127567
```

### 5.3.7 Quade-Test

Das Verfahren von *Quade* war in Kapitel 2.10.2 skizziert worden. An den Beispieldaten 4, allerdings hier ohne Berücksichtigung der Gruppenstruktur, soll die Berechnung gezeigt werden. R bietet dazu die Funktion `quade.test`.

**mit R:**

Nachfolgend die Ein- und Ausgabe. Eine Umstrukturierung ist wie bei der Friedman-Analyse nicht erforderlich:

```
quade.test(as.matrix(winer518[,3:5]))
```

```
Quade test
data:  as.matrix(winer518[, 3:5])
Quade F = 6.2019, num df = 2, denom df = 18, p-value = 0.008935
```

Das Ergebnis bestätigt allerdings nicht, dass der Quade-Test bei kleinerer Anzahl von Messwiederholungen stärker ist als der Friedman-Test ( $p=0,0084$ ).

### 5.3.8 Skillings-Mack-Test

Das Verfahren von *Skillings & Mack* war in Kapitel 2.10.3 erwähnt worden. An den Beispieldaten 4, allerdings hier ohne Berücksichtigung der Gruppenstruktur, soll die Berechnung gezeigt werden. R bietet dazu die Funktion `skiMack` im Paket `Skillings.Mack`.

**mit R:**

Nachfolgend die Ein- und Ausgabe (auszugsweise). Eine Umstrukturierung ist wie bei der Friedman-Analyse nicht erforderlich:

```
library(.Mack)
SkiMack(as.matrix(winer518[,3:5]))
```

```
Skillings-Mack Statistic = 13.545455 , p-value = 0.139438
Note: the p-value is based on the chi-squared distribution with df=9
```

Dass dieser Test hier schlechter als der Friedman-Test abschneidet, ist höchstwahrscheinlich den Bindungen zuzuschreiben.

### 5.3.9 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Agresti-Pendergast

Die parametrischen multivariates Tests wurden in Abschnitt 2.12 kurz vorgestellt und bei der Besprechung der Voraussetzungen in Kapitel 5.2 als parametrische Alternative zum F-Test erwähnt. Sie setzen allerdings eine multivariate Normalverteilung der Residuen voraus, die wesentlich mehr beinhaltet als die univariate Normalverteilung aller Residuen. Es gibt zur Überprüfung einige Verfahren, u.a. die bekannten von K.V. Mardia (vgl. Ito, 1980) und Shapiro-Wilk. In R wird hierfür u.a. die Pakete `mvnortest` und `MVN` bereitgestellt. Ersatzweise muss man sich auf die univariate Überprüfung beschränken und die einzelnen Ergebnisse mit dem Test von Fisher (vgl. Anhang 2.5) zusammenfassen. Dies soll aber hier nicht vorgestellt werden.

Es gibt aber auch nichtparametrische Versionen dieses Tests von Agresti & Pendergast (1986) bzw. Akritas & Arnold (1994). Hierbei werden zunächst wie beim RT-Verfahren sämtliche Werte in Ränge  $1, \dots, n \cdot J$  transformiert. Anschließend wird ein multivariater Test, z.B. der Hotelling-Lawley-Test durchgeführt. Für beide Verfahren gibt es jeweils zwei Varianten, einmal wird der F-Test der multivariaten Version übernommen, zum anderen gibt es auch einen  $\chi^2$ -Test. Ersterer wird i.a. vorgezogen, nicht nur aus Bequemlichkeit. Leider gibt es kaum Vergleiche dieser nichtparametrischen Tests mit anderen, woraus zu entnehmen ist, wann der Test von Agresti & Pendergast bzw. von Akritas & Arnold empfehlenswert ist. Eigene Simulationen haben gezeigt, dass Letzterer zumindest den Fehler 1. Art besser unter Kontrolle hält als der von Agresti & Pendergast. Die Verfahren zum Test des Messwiederholungseffekts werden anhand des Datensatzes `winer568` mit den Variablen  $v_1, \dots, v_4$  vorgestellt.

**mit R:**

Die multivariaten Tests werden u.a. über zwei Standardfunktionen angeboten, `manova` sowie `lm` für allgemeine lineare Modelle. In diesem Fall ist `lm` einfacher anzuwenden. In jedem Fall ist die Berechnung der Differenzen der 4 Messwiederholungsvariablen  $v_1, \dots, v_4$  erforderlich:  $v_4 - v_3$ ,  $v_3 - v_2$  und  $v_2 - v_1$ . Dieses kann implizit im Aufruf der Funktion erfolgen, wobei allerdings in jedem Fall diese Variablen zu einer Matrix zusammengefasst werden müssen, z.B. mittels `cbind`. Die Struktur der Datenmatrix muss hier die „normale“, also untransformierte sein. Nachfolgend Eingabe und Ausgabe bei Wahl des Tests von Pillai, wonach der Faktor Zeit einen signifikanten Einfluss hat:

```
with(winer568, anova(lm(cbind(V4-V3, V3-V2, V2-V1) ~ 1),
  test="Pillai"))
```

## Analysis of Variance Table

	Df	Pillai approx	F num	Df den	Df	Pr(>F)
(Intercept)	1	35.051	105.15	3	9	2.522e-07 ***
Residuals	11					

Für den Durchführung des Tests von Akritas & Arnold müssen die Werte der Variablen  $v_1, \dots, v_4$  in Ränge transformiert werden. Da diese sowohl über die Erhebungseinheiten als auch über die Variablen ermittelt werden, ist zunächst die Erstellung des Datensatzes im langen Format erforderlich: `winer568t` (vgl. Abschnitt 5.1.2). Anschließend werden die Ränge ermittelt und der Datensatz zurück in das breite Format gewandelt:

```
winer568t <- within(winer568t, Rx<-rank(x))
winer568y <- reshape(winer568t, direction="wide", timevar="Zeit",
  v.names=c("x", "Rx"), times=1:4, idvar="Vpn")
```

Die Ränge von  $v_1, \dots, v_4$  haben im Ergebnis-Dataframe die Namen `Rx.1, \dots, Rx.4`. Schließlich der Test auf Basis des Tests von Hotelling-Lawley:

```
with(winer568y, anova(lm(cbind(Rx.4-Rx.3, Rx.3-Rx.2, Rx.2-Rx.1)~1),
  test="Hotelling-Lawley"))
```

## Analysis of Variance Table

	Df	Hotelling-Lawley	approx F num	Df den	Df	Pr(>F)
(Intercept)	1	34.854	104.56	3	9	2.584e-07 ***
Residuals	11					

Da bei der multivariaten Analyse die Daten nicht als Dataframe, sondern als Matrix verlangt werden, kann man auch die Transformation von `winer568` in das lange Format, die Rangbildung und die Rücktransformation ins breite Format in einer Anweisung erledigen:

```
mat568 <- matrix(rank(as.vector(as.matrix(winer568[,3:6]))), 12, 4)
```

(Hierbei wird mit 12 die Anzahl der Erhebungseinheiten und mit 4 die Anzahl der Variablen angegeben.) Ebenso elegant lassen sich danach die Differenzen beim Aufruf von `lm` ermitteln:

```
anova(lm((mat568[,2:4]-mat568[,1:3])~1))
```

Für den Test von Agresti & Pendergast wird die Funktion `ap.anova` zur Verfügung gestellt (vgl. Anhang 3). Hier Aufruf und Ausgabe:

```
ap.anova(winer568t, "x", "Vpn", "Zeit")
```

	chisq	df	p value	F	df num	df denom	p value
main effect	313.6847	3	0	104.5616	3	33	0

was sich mit den o.a. Ergebnissen deckt.

Zum Abschluss noch die Überprüfung auf multivariate Normalverteilung. Da ja nicht die Variablen  $v_1, \dots, v_4$  selbst, sondern die Residuen des Anova-Modells überprüft werden sollen, müssen diese zunächst ermittelt werden. Die Vorgehensweise war in Kapitel 5.3.1 erläutert worden. Basierend auf dem Datensatz `winer568t` wird für die Variable `x` ein Modell ohne Messwiederholungen aufgestellt, daraus die Residuen (Variable `res`)

gespeichert und dem Datensatz zugefügt und zurück in das breite Format transformiert. Dadurch erhält man für 4 Variablen die zugehörigen Residuen (`res.1,...,res.4`), die dann auf multivariate Normalverteilung überprüft werden, hier mittels des Tests von Mardia (Funktion `mardia`). Die univariaten Tests von Shapiro-Wilk werden automatisch mit ausgegeben:

```
aov1      <- aov(x~Zeit+Vpn, winer568t)
res       <- aov1$residuals
winer568t <- cbind(winer568t,res)
winer568x <- reshape(winer568t,direction="wide",
                    timevar="Zeit",v.names=c("x","res"),times=1:4,idvar="Vpn")

library(mvnormalTest)
mardia(winer568x[,c("res.1","res.2","res.3","res.4")])
```

	Test	Statistic	p-value	Result
1	Skewness	-223.8753	1	YES
2	Kurtosis	21.6719	0	NO
3	MV Normality	<NA>	<NA>	NO

\$uv.shapiro				
	W	p-value	UV.Normality	
res.1	0.9508	0.6485	Yes	
res.2	0.9417	0.5202	Yes	
res.3	0.9701	0.9115	Yes	
res.4	0.9127	0.2309	Yes	

Das Ergebnis besagt, dass zwar die Schiefe nicht aus dem Rahmen fällt, jedoch der Exzess, und damit keine multivariate Normalverteilung gegeben ist. Und das trotz der kleinen Fallzahl von  $N=12$ . Zu beachten ist, dass dennoch die 4 univariaten Tests auf Normalverteilung negativ sind, d.h. univariate Normalverteilungen angenommen werden können.

### mit SPSS:

Multivariate Tests werden in SPSS bei Varianzanalysen mit Messwiederholungen immer automatisch als erstes Ergebnis (zusätzlich zur normalen parametrischen Analyse) ausgegeben. Eine Bildung der Differenzen oder ähnliches ist hier nicht erforderlich. Nachfolgend Eingabe und Ausgabe, wonach der Faktor Zeit, unabhängig von der Wahl des multivariaten Tests, einen signifikanten Einfluss hat:

```
GLM V1 V2 V3 V4
  /WSfactor=Zeit 4 Polynomial
  /WSdesign=Zeit
```

Multivariate Tests						
Effekt		Wert	F	Hypothese df	Fehler df	Sig.
Zeit	Pillai-Spur	,972	105,152	3,000	9,000	,000
	Wilks-Lambda	,028	105,152	3,000	9,000	,000
	Hotelling-Spur	35,051	105,152	3,000	9,000	,000
	Größte charakteristische	35,051	105,152	3,000	9,000	,000
	Wurzel nach Roy					

Für den Durchführung des Tests von Akritas & Arnold müssen die Werte der Variablen  $v_1, \dots, v_4$  in Ränge transformiert werden. Da diese sowohl über die Erhebungseinheiten als

auch über die Variablen ermittelt werden, ist zunächst die Erstellung des Datensatzes im langen Format erforderlich (vgl. Abschnitt 5.3.3).

```

Varstocases
  /Id=Vpn
  /Make score from v1 v2 v3 v4
  /index=Zeit(4)
  /keep=A B
  /null=keep.

Rank Variables=score / rank into Rscore.
Sort cases by Vpn Zeit.

casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

GLM Rscore.1 to Rscore.4
  /WSfactor=Zeit 4 Polynomial
  /WSdesign=Zeit.

```

Multivariate Tests						
Effekt		Wert	F	Hypothese df	Fehler df	Sig.
	Pillai-Spur	,972	104,562	3,000	9,000	,000
	Wilks-Lambda	,028	104,562	3,000	9,000	,000
Zeit	Hotelling-Spur	34,854	104,562	3,000	9,000	,000
	Größte charakteristische	34,854	104,562	3,000	9,000	,000
	Wurzel nach Roy					

Das Ergebnis des nichtparametrischen Tests weicht nur unbedeutend vom parametrischen ab.

### 5.3.10 Multivariate Tests: Spatial Signs und Spatial Ranks Methode

In den Abschnitten 2.11.3 und 2.16 waren kurz neuere multivariate Methoden erwähnt worden, die auf räumlichen Vorzeichen und Rängen beruhen, u.a. für multivariate Varianzanalysen. Diese verlangen keinerlei Varianz-Homogenitäten. Im vorigen Abschnitt war gezeigt worden, wie diese für eine 1-faktorielle Analyse mit Messwiederholungen eingesetzt werden kann. In R gibt es dazu mehrere Funktionen bzw. Pakets:

- Funktion `sr.loc.test` im SpatialNP (Sirkiä et al., 2007),
- Funktion `mv.1sample.test` im Paket MNM (Nordhausen & Oja, 2011),
- Funktion `ojasampleTest` im Paket OjaNP (Fischer et al., 2020),
- Funktion `HP.loc.test` im Paket ICSNP (Hallin & Paindaveine, 2002),
- Funktion `rank.ctest` im Paket ICSNP (Puri & Sen, 1971),

die alle in gleicher Weise benutzt werden können. Der Unterschied besteht im Wesentlichen in den unterschiedlichen Modellen für die multivariaten Ränge, die in der o.a. Literatur beschrieben werden. Vergleiche gibt es derzeit in der Literatur noch keine. Eigene Simulationen ergaben jedoch folgendes Bild: Die Kontrolle des Fehlers 1. Art ist bei allen Funktionen für alle Verteilungsformen relativ gut. Lediglich die Funktion `sr.loc.test` verhält sich für kleine  $n_i$  relativ liberal. Und alle Funktionen verletzen im Fall von negativ pairing (vgl. 1.1.6) das  $\alpha$ -Ri-



siko fast so stark wie die parametrischen Methoden. Die Power liegt ca. 10-15% unter der des F-Tests, was als gut zu bewerten ist. Bei der Funktion `oja1sampleTest` muss vor dem hohen Bedarf an Rechenleistung gewarnt werden, wenn mehr als 4 Messwiederholungen vorliegen.

### mit R:

Diese Verfahren werden anhand des Datensatzes `winer568` mit den Variablen  $v_1, \dots, v_4$  vorgestellt. Die Struktur der Datenmatrix muss hier, wie oben beim parametrischen Verfahren, die „normale“, also untransformierte sein. Ein- und Ausgabe:

```
library(SpatialNP)
with(winer568, sr.loc.test(cbind(V4-V3,V3-V2,V2-V1),score="rank"))
```

```
One sample location test using spatial signed ranks

data: cbind(V4 - V3, V3 - V2, V2 - V1)
Q.2 = 19.436, df = 3, p-value = 0.0002221
alternative hypothesis: true location is not equal to c(0,0,0)
```

```
library(MNM)
with(winer568, mv.1sample.test(cbind(V4-V3,V3-V2,V2-V1),score="rank"))
```

```
One sample spatial signed-rank test using outer standardiza-
tion

data: cbind(V4 - V3, V3 - V2, V2 - V1)
Q.2 = 11.26, df = 3, p-value = 0.0104
alternative hypothesis: true location is not equal to c(0,0,0)
```

```
library(OjaNP)
with(winer568, oja1sampleTest(cbind(V4-V3,V3-V2,V2-V1),score="rank"))
```

```
OJA 1 SAMPLE SIGNED RANK TEST

data: cbind(V4 - V3, V3 - V2, V2 - V1)
Q.R = 10.841, df = 3, p-value = 0.01262
alternative hypothesis: true location is not equal to c(0,0,0)
```

Alle 3 Ergebnisse  $p=0.0002$ ,  $p=0.0104$  bzw.  $p=0.0126$  decken sich mit den oben erzielten.

Im Paket ICSNP stehen 2 Funktionen zur Auswahl: `HP.loc.test` und `rank.ctest`, wobei letztere nicht nur multivariate Ränge sondern auch normal scores erlaubt (Vorsicht: Parameter scores anstatt score):

```
library(ICSNP)
with(winer568, HP.loc.test(cbind(V4-V3,V3-V2,V2-V1),score="rank"))
```

```
TYLER ANGLES RANK TEST

data: cbind(V4 - V3, V3 - V2, V2 - V1)
Q.W = 7.3304, df = 3, p-value = 0.06208
alternative hypothesis: true location is not equal to c(0,0,0)
```

Bei dieser Methode kann der Messwiederholungseffekt nicht nachgewiesen werden.

```
with(winer568, rank.ctest(cbind(V4-V3, V3-V2, V2-V1), scores="rank"))
with(winer568, rank.ctest(cbind(V4-V3, V3-V2, V2-V1), scores="normal"))
```

```

Marginal One Sample Signed Rank Test

data: cbind(V4 - V3, V3 - V2, V2 - V1)
T = 11.443, df = 3, p-value = 0.009554
alternative hypothesis: true location is not equal to c(0,0,0)

Marginal One Sample Normal Scores Test

data: cbind(V4 - V3, V3 - V2, V2 - V1)
T = 11.348, df = 3, p-value = 0.009986
alternative hypothesis: true location is not equal to c(0,0,0)

```

## 5.4 Die 2-faktorielle Varianzanalyse

Mit der 2-faktoriellen Varianzanalyse mit Messwiederholungen ist hier ein Design ohne Gruppierungsfaktoren, ausschließlich mit zwei Messwiederholungsfaktoren gemeint, hier mit C und D bezeichnet, jeweils mit *I* bzw. *J* Stufen. Sie unterscheidet sich allerdings gegenüber den Analysen ohne Messwiederholungen sowie der 1-faktoriellen Analyse mit Messwiederholungen dahingehend, dass sie mehrere Fehlerstreuungen hat, und zwar eine für jeden Effekt: C, D sowie C\*D. Auch hier nimmt man für die Durchführung nichtparametrischer Analysen in der Regel den Umweg über die parametrische Analyse. Anzumerken ist noch, dass der Friedman-Test häufig irreführend als 2-faktorielle Analyse bezeichnet wird.

Während für die Analysen mit R ohnehin die Datenmatrix umstrukturiert werden muss und für die nichtparametrischen Tests kein gesonderter Aufwand entsteht, muss zur Rangberechnung an dieser Stelle auch in SPSS eine solche Umstrukturierung vorgenommen werden.

### 5.4.1 Das parametrische Verfahren und Prüfung der Voraussetzungen

Auch hier soll zunächst einmal zum Vergleich die parametrische Varianzanalyse durchgeführt werden, und zwar anhand der Beispieldaten 5 (`mydata5`) für den Vergleich der Reaktionen in Abhängigkeit von drei Medikamenten bzw. drei Aufgaben, jedoch ohne Berücksichtigung der Gruppeneinteilung in Männer und Frauen.

Im Gegensatz zum Datensatz 4 (`winer518`) aus dem letzten Kapitel zeigt hier Mauchlys Test signifikante Abweichungen von der Sphärizität. Für jeden der drei Tests C, D und C\*D (im Beispiel: `Medikament`, `Aufgabe` und Wechselwirkung) wird die dafür relevante Sphärizität überprüft. Da sowohl für `Medikament` als auch für die Wechselwirkung Mauchlys Test signifikant ist, sollten anstatt des „normalen“ F-Tests die Approximationen von Geisser & Greenhouse oder von Huynh & Feldt verwendet werden. Entscheidet man sich für letztere, so erhält man aus den Tabellen 5-5 (R) bzw. 5-6 (SPSS) für den Medikamenten-Effekt einen p-Wert, der nur geringfügig über dem „normalen“ liegt. Für den Interaktionseffekt bedeutet dies jedoch den Verlust der Signifikanz, da der p-Wert des „normalen“ Tests 0,023 beträgt gegenüber einem  $p=0,058$  für die Huynh & Feldt-Approximation.

#### mit R:

Ausgangsbasis ist der in 5.1.2 erstellte Dataframe `mydata5t`. Die Varianzanalyse mit doppelten Messwiederholungen wird nun zunächst wieder mit `aov` durchgeführt, wobei jetzt zwei Messwiederholungsfaktoren zu berücksichtigen sind. Beide sind für den Error-Term

als eingebettet in `Vpn` zu deklarieren, wobei die Klammern dringend erforderlich sind:

```
aov1 <- summary(aov (Fehler ~ Medikament*Aufgabe
                    + Error(Vpn/(Medikament*Aufgabe)), mydata5t))
```

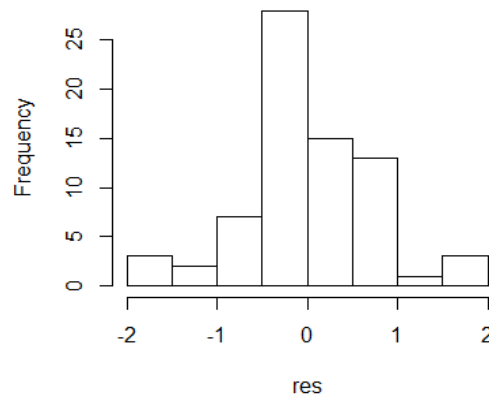
Die Ausgabe (nachfolgende Tabelle 5-4) wirkt auf den ersten Blick etwas unübersichtlich, da jeder Effekt einen eigenen Fehlerterm (Residuals) besitzt. Das Ergebnis: Sowohl zwischen den beiden Medikamenten bzw. der Kontrollmessung als auch zwischen den drei Aufgaben bestehen hinsichtlich der Bearbeitung der Aufgaben (Fehlerzahl) signifikante Unterschiede. Hinzu kommt eine signifikante Wechselwirkung beider Faktoren. Auf Details der Interpretation soll hier nicht eingegangen werden.

Error: Vpn							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
Residuals	7	32.65	4.665				
Error: Vpn:Medikament							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
Medikament	2	27.444	13.722	20.83	6.37e-05	***	
Residuals	14	9.222	0.659				
Error: Vpn:Aufgabe							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
Aufgabe	2	40.78	20.389	20.55	6.83e-05	***	
Residuals	14	13.89	0.992				
Error: Vpn:Medikament:Aufgabe							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
Medikament:Aufgabe	4	6.056	1.5139	3.361	0.0229	*	
Residuals	28	12.611	0.4504				

Tabelle 5-4

Die Prüfung der Voraussetzungen erfolgt wie bei der 1-faktoriellen Analyse (vgl. Kapitel 5.3.1). Die Residuen erhält man über folgendes Anova-Modell, das auch auf dem zuletzt erstellten Dataframe `mydata5t` aufsetzt. Diese können dann wie üblich betrachtet werden:

```
aov2 <- aov (Fehler ~ Medikament*Aufgabe + Vpn, mydata5t)
res <- aov2$residuals
hist(res)
```



Die versuchspersonenspezifische Abweichungen  $\pi_m$  basieren auf dem ursprünglichen Dataframe `mydata5`. Für den Test auf Normalverteilung genügt es, die Personenmittelwerte der 9 abhängigen Variablen zu betrachten, die bequem mittels `rowMeans` errechnet werden

können. Auf die Ausgabe wird hier verzichtet:

```
hist(rowMeans(mydata5[,3:11]))
```

Die Varianzhomogenität bzw. Sphärität wird wieder mit der Funktion `ezANOVA` des Pakets `ez` geprüft. Die Spezifikation des Modells ist damit deutlich einfacher:

```
library(ez)
ezANOVA(mydata5t, Fehler, Vpn, within=. (Medikament, Aufgabe))
```

Das Ergebnis, das hinsichtlich der Tests auf Sphärität bereits oben interpretiert wurde:

	Effect	DFn	DFd	F	p	p<.05	ges
2	Medikament	2	14	20.831325	6.367208e-05	*	0.28641832
3	Aufgabe	2	14	20.552000	6.833046e-05	*	0.37358443
4	Medikament:Aufgabe	4	28	3.361233	2.286928e-02	*	0.08135846

\$`Mauchly's Test for Sphericity`							
	Effect	W		p	p<.05		
2	Medikament	0.35012339	0.04292036		*		
3	Aufgabe	0.86860800	0.65534724				
4	Medikament:Aufgabe	0.02042957	0.01630533		*		

\$`Sphericity Corrections`							
	Effect	GGe	p [GG]	HFe		p [HF]	
p [HF] < .05							
2	Medikament	0.6061059	0.0011688272	0.6649945		7.533244e-04	
3	Aufgabe	0.8838670	0.0001589182	1.1602880		6.833046e-05	
4	Medikament:Aufgabe	0.4258173	0.0752372276	0.5487419		5.794030e-02	

Tabelle 5-5

Die Überprüfung der Sphärität mittels der Funktion `check.sphere` ist dagegen etwas aufwändiger. Überprüft man "einfach" die 9 Variablen  $v_1, \dots, v_9$ , so entspricht das dem Test der Interaktion auf Sphärität:

```
check.sphere(mydata5[,2:10])
```

\$results				
	statistic	df	p	value
Mauchly test	172.5402410	35.0000	0.00000000	
Likelihood Ratio	292.7955605	44.0000	0.00000000	
John's test chisq	59.4141552	35.0000	0.00614969	
John's test beta	0.1194727	38.2963	0.93171346	
John-Nagao	30.5850153	44.0000	0.94430360	
Muirhead & Waternaud	568.7818268	44.0000	0.00000000	
compound symmetry	279.5765959	43.0000	0.00000000	

Für die Überprüfung nur eines Faktors, z.B. Medikament, müssen die Summen über die Stufen des anderen Faktors, hier Aufgabe, gebildet und anschließend getestet werden. Die Summen erhalten die Namen  $m_1, m_2, m_3$  und haben im Dataframe die Positionen 11-13:

```
within(mydata5, {m1<-v1+v2+v3; m2<-v4+v5+v6; m3<-v7+v8+v9}) ->mydata5
check.sphere(mydata5[,11:13])
```

\$results			
	statistic	df	p value
Mauchly test	6.2968179	2	0.042920360
Likelihood Ratio	7.3462876	5	0.196133771
John's test chisq	11.8825805	2	0.002628636
John's test beta	0.3249383	5	0.178109538
John-Nagao	5.1990129	5	0.371141004
Muirhead & Waternaud	16.7902744	5	0.004915211
compound symmetry	10.8858369	4	0.027877459

**mit SPSS:**

Die Spezifikation für die Syntax (mit Speicherung der 9 Residuenvariablen) ist relativ einfach:

```
GLM v1 v2 v3 v4 v5 v6 v7 v8 v9
  /wsfactor=Medikament 3 polynomial Aufgabe 3 polynomial
  /save=resid
  /wsdesign=Medikament Aufgabe Medikament*Aufgabe.
```

Mit folgenden relevanten Tabellen: des Mauchly-Tests und der (auf 2 Seiten verteilte) Anova-Tabelle:

Mauchly-Test auf Sphärizität							
Innersubjekteffekt	Mauchly-W	Approximiertes Chi-Quadrat	df	Sig.	Epsilon <sup>b</sup>		
					Greenhouse -Geisser	Huynh -Feldt	Unter grenze
Medikament	,350	6,297	2	,043	,606	,665	,500
Aufgabe	,869	,845	2	,655	,884	1,000	,500
Medikament * Aufgabe	,020	21,075	9	,016	,426	,549	,250

Tabelle 5-7

Tests der Innersubjekteffekte						
Quelle		Quadrat- summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angenommen	27,444	2	13,722	20,831	,000
	Greenhouse-Geisser	27,444	1,212	22,640	20,831	,001
	Huynh-Feldt	27,444	1,330	20,635	20,831	,001
	Untergrenze	27,444	1,000	27,444	20,831	,003
Fehler (Medikament)	Sphärizität angenommen	9,222	14	,659		
	Greenhouse-Geisser	9,222	8,485	1,087		
	Huynh-Feldt	9,222	9,310	,991		
	Untergrenze	9,222	7,000	1,317		
Aufgabe	Sphärizität angenommen	40,778	2	20,389	20,552	,000
	Greenhouse-Geisser	40,778	1,768	23,068	20,552	,000
	Huynh-Feldt	40,778	2,000	20,389	20,552	,000
	Untergrenze	40,778	1,000	40,778	20,552	,003
Fehler (Aufgabe)	Sphärizität angenommen	13,889	14	,992		
	Greenhouse-Geisser	13,889	12,374	1,122		
	Huynh-Feldt	13,889	14,000	,992		
	Untergrenze	13,889	7,000	1,984		

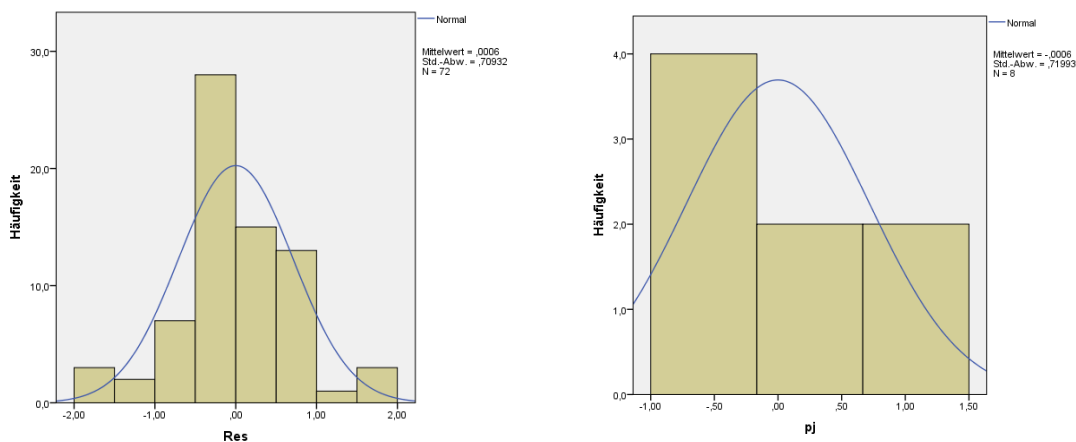
Medikament * Aufgabe	Sphärizität angenommen	6,056	4	1,514	3,361	,023
	Greenhouse-Geisser	6,056	1,703	3,555	3,361	,075
	Huynh-Feldt	6,056	2,195	2,759	3,361	,058
	Untergrenze	6,056	1,000	6,056	3,361	,109
Fehler (Medikmt*Aufgabe)	Sphärizität angenommen	12,611	28	,450		
	Greenhouse-Geisser	12,611	11,923	1,058		
	Huynh-Feldt	12,611	15,365	,821		
	Untergrenze	12,611	7,000	1,802		

Tabelle 5-6

Das Ergebnis des Mauchly-Tests und dessen Konsequenzen wurden bereits am Anfang dieses Kapitels erörtert. Werden die 9 Residuenvariablen zu einer zusammengefasst, erhält man für die Überprüfung auf Normalverteilung ein Ergebnis, das keine bedeutsamen Abweichungen erkennen lässt:

Tests auf Normalverteilung						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Residuen	,130	72	,004	,968	72	,063

Besser ist es aber, wie in Kapitel 5.3.1. demonstriert, vorher von den Residuen den Versuchspersoneneffekt abzuziehen. Der Shapiro-Wilk-Test ergibt dann ein  $p=0,173$ . Unten links das dazugehörige Histogramm, unten rechts das Histogramm für die  $\pi_m$ , das allerdings bei  $n=8$  kaum Aussagefähigkeit hat und daher i.a. entfallen kann:



#### 5. 4. 2 Rank transform-Tests (RT) und normal scores -Tests (INT)

Bei den einfachen Rank transform Tests wird lediglich vor der Durchführung der parametrischen Varianzanalyse die abhängige Variable über alle Werte (Fälle und Messwiederholungen) hinweg in Ränge transformiert. Die statistischen Tests bleiben unverändert. Dieses Verfahren von Conover & Iman (1981) ist in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, nicht jedoch für Variablen mit beliebigen Eigenschaften. Insofern sollte die Möglichkeit genutzt werden, für die rangtransformierten Daten anstatt des normalen F-Tests die Variante von Huynh & Feldt anzuwenden, um eventuellen Varianzhomogenitäten zu begegnen.

Das INT-Verfahren unterscheidet sich von dem RT-Verfahren nur marginal: Die Ränge  $R(x_{mk})$  werden noch in normal scores (INT) umgerechnet:

$$nscore_{mk} = \Phi^{-1}(R(x_{mk})/(M+1))$$

wobei  $M$  die Anzahl aller Werte ist, also  $n \cdot I \cdot J$  (mit  $n$  Anzahl der Merkmalsträger und  $I$  und  $J$  Anzahl der Messwiederholungen der Faktoren C und D), sowie  $\Phi^{-1}$  die Umkehrfunktion der Normalverteilung.

Bei den Ergebnissen zeigt sich, dass durch die INT-Transformation - im Gegensatz zur RT-Transformation - die Varianzhomogenität nicht beseitigt werden konnte. Aber auf der anderen Seite sind die Ergebnisse qualitativ dieselben, wenn man bei der normal scores-Analyse die Huynh-Feldt-adjustierten F-Tests wählt.

Da die Ausgabe bei beiden Verfahren dieselbe ist, werden die Ergebnistabellen (etwas verkürzt) lediglich einmal in der (leichter lesbaren) Version von SPSS wiedergegeben.

#### mit R:

Um die Sphärität prüfen zu können bzw. die adjustierten F-Tests zu erhalten, wird die Varianzanalyse mit `ezANOVA` (Paket `ez`) durchgeführt. Ausgehend vom in Kapitel 5.1.2 erstellten Dataframe `mydata5t` sind folgende Anweisungen erforderlich:

```
library(ez)
RFehler <- rank(mydata5t$Fehler)
mydata5t <- cbind(mydata5t, RFehler)
ezANOVA(mydata5t, RFehler, Vpn, within=. (Medikament, Aufgabe))
```

Da alle drei Mauchly-Tests nicht signifikant sind, kann die Anova-Tabelle (`$ANOVA`) herangezogen werden, deren Ergebnisse zum Teil (Medikament und Interaktion) sogar besser sind, als bei der „rein parametrischen“ unter Verwendung der Huynh & Feldt-Approximationen (vgl. auch Tabelle 5-5 in Abschnitt 5.4.1).

Für die Berechnung der normal scores sowie deren Varianzanalyse sind die o.a. Anweisungen zu ergänzen - die Ergebnisse folgen weiter unten:

```
nc <- dim(mydata5t)[1]
mydata5t <- within(mydata5t, nsFehler<-qnorm(RFehler/(nc+1)))
ezANOVA(mydata5t, nsFehler, Vpn, within=. (Medikament, Aufgabe))
```

#### mit SPSS:

- Zunächst müssen für den Datensatz über das Menü „Daten -> Umstrukturieren“ die Messwiederholungen in Fälle transformiert werden (siehe dazu Anhang 1.1.2).
- Die Variable `Fehler` wird dann über das Menü „Transformieren -> Rangfolge bilden“ in Ränge umgerechnet.
- Danach muss der Datensatz wieder zurück in das „normale“ Format mit Messwiederholungen transformiert werden (vgl. Anhang 1.2).
- Abschließend wird dann eine Varianzanalyse mit Messwiederholungen (Menü: „Allgemeines lineares Modell -> Messwiederholung“) für die Variablen `RFehler.1.1`, `RFehler.1.2`, ..., `RFehler.3.3` gerechnet, die bei der Umstrukturierung gebildet werden:

Die Syntax für den ersten Schritt der Umstrukturierung, der Rangbildung bzw. des zweiten Schritts der Umstrukturierung in der SPSS-Syntax:

```

Varstocases
  /Id=Vpn
  /make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
  /index=Medikament(3) Aufgabe(3)
  /keep=Geschlecht
  /null=keep.

Rank variables=Fehler (A)
  /rank into RFehler.

Sort cases by Vpn Medikament Aufgabe.
Casestovars
  /Id=Vpn
  /index=Medikament Aufgabe
  /groupby=variable.

GLM RFehler.1.1 RFehler.1.2 RFehler.1.3 RFehler.2.1 RFehler.2.2
  RFehler.2.3 RFehler.3.1 RFehler.3.2 RFehler.3.3
  /WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
  /WSdesign=Medikament Aufgabe Medikament*Aufgabe.

```

Für die Errechnung der normal scores muss die Rank-Anweisung durch die folgenden ersetzt werden:

```

Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(Fehler).
Rank Variables=Fehler / rank into RFehler.
compute nFehler=Idf.normal(RFehler/nc,0,1).

```

Und in den GLM-Anweisungen ist entsprechend RFehler... durch nFehler zu ersetzen.

Hier nun die Ergebnisse in der Version von SPSS:

Zunächst für das RT-Verfahren, und zwar der Mauchly-Test:

Mauchly-Test auf Sphärizität							
Innersubjekteffekt	Mauchly W	Approx. Chi-Quadrat	df	Sig.	Epsilon		
					Greenhouse -Geisser	Huynh Feldt	Unter grenze
Medikament	,470	4,524	2	,104	,654	,743	,500
Aufgabe	,922	,485	2	,785	,928	1,000	,500
Medikament * Aufgabe	,070	14,377	9	,125	,490	,679	,250

sowie das Ergebnis für die Varianzanalyse auf Basis der Rangtransformation, bei dem wegen der für alle drei Tests gegebenen Sphärizität die jeweils erste Zeile genommen werden kann, wenn auch empfohlen wird, generell die Resultate der Huynh-Feldt-adjustierten Tests zu verwenden. Lediglich für die Interaktion fallen die Ergebnisse auf dem 5%-Niveau unterschiedlich aus: ,046 bei Annahme der Sphärizität, 074 andernfalls. Die Ergebnisse sind zum Teil (Medikament und Interaktionen) sogar besser sind, als bei der „rein parametrischen“ unter Verwendung der Huynh-Feldt-Approximationen (vgl. Tabelle 5-6):



Tests der Innersubjekteffekte						
Quelle		Quadrat summe vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	5419,083	2	2709,542	21,880	,000
	Greenhouse-Geisser	5419,083	1,308	4144,310	21,880	,001
	Huynh-Feldt	5419,083	1,486	3645,959	21,880	,000
	Untergrenze	5419,083	1,000	5419,083	21,880	,002
Fehler (Medikament)	Sphärizität angen.	1733,750	14	123,839		
	Greenhouse-Geisser	1733,750	9,153	189,415		
	Huynh-Feldt	1733,750	10,404	166,638		
	Untergrenze	1733,750	7,000	247,679		
Aufgabe	Sphärizität angen.	8037,750	2	4018,875	18,529	,000
	Greenhouse-Geisser	8037,750	1,856	4330,863	18,529	,000
	Huynh-Feldt	8037,750	2,000	4018,875	18,529	,000
	Untergrenze	8037,750	1,000	8037,750	18,529	,004
Fehler (Aufgabe)	Sphärizität angen.	3036,583	14	216,899		
	Greenhouse-Geisser	3036,583	12,991	233,737		
	Huynh-Feldt	3036,583	14,000	216,899		
	Untergrenze	3036,583	7,000	433,798		
Medikament * Aufgabe	Sphärizität angen.	1099,667	4	274,917	2,774	,046
	Greenhouse-Geisser	1099,667	1,962	560,571	2,774	,098
	Huynh-Feldt	1099,667	2,718	404,605	2,774	,074
	Untergrenze	1099,667	1,000	1099,667	2,774	,140
Fehler (Medikament*Aufgabe)	Sphärizität angen.	2774,500	28	99,089		
	Greenhouse-Geisser	2774,500	13,732	202,049		
	Huynh-Feldt	2774,500	19,025	145,833		
	Untergrenze	2774,500	7,000	396,357		

Tabelle 5-9

Nun das Ergebnis für das normal score (INT)-Verfahren, zunächst der Mauchly-Test:

Mauchly-Test auf Sphärizität <sup>a</sup>						
Innersubjekteffekt	Mauchly-W	Approx.. Chi- Quadrat	df	Sig.	Epsilon <sup>b</sup>	
					Greenhouse -Geisser	Huynh-Feldt
Medikament	,350	6,297	2	,043	,606	,665
Aufgabe	,869	,845	2	,655	,884	1,000
Medikament * Aufgabe	,020	21,075	9	,016	,426	,549

der zeigt, dass lediglich für den Effekt *Aufgabe* durch die Transformation die Varianzheterogenität beseitigt werden konnte. Abgesehen davon empfehlen Beasley & Zumbo (2009) ohnehin, in jedem Fall die adjustierten F-Tests, z.B. den von Huynh-Feldt, zu verwenden. Nachfolgend die (um die Fehlerterme) verkürzte Anova-Tabelle:

Tests der Innersubjekteffekte						
Quelle		Quadrats. vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	27,444	2	13,722	20,831	,000
	Greenhouse-Geisser	27,444	1,212	22,640	20,831	,001
	Huynh-Feldt	27,444	1,330	20,635	20,831	,001
	Untergrenze	27,444	1,000	27,444	20,831	,003
Aufgabe	Sphärizität angen.	40,778	2	20,389	20,552	,000
	Greenhouse-Geisser	40,778	1,768	23,068	20,552	,000
	Huynh-Feldt	40,778	2,000	20,389	20,552	,000
	Untergrenze	40,778	1,000	40,778	20,552	,003
Medikament * Aufgabe	Sphärizität angen.	6,056	4	1,514	3,361	,023
	Greenhouse-Geisser	6,056	1,703	3,555	3,361	,075
	Huynh-Feldt	6,056	2,195	2,759	3,361	,058
	Untergrenze	6,056	1,000	6,056	3,361	,109

### 5. 4. 3 Puri & Sen-Test

Bei diesem Verfahren werden die Werte über alle Merkmalsträger und alle Messwiederholungen hinweg wie beim o.a. RT-Verfahren in Ränge 1,...,n\*I\*J (I\*J=Anzahl der gesamten Messwiederholungen) transformiert und damit eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt. Für jeden der 3 Effekttests sind folgende Schritte durchzuführen, wobei zu beachten ist, dass, wie in 5.4 eingangs erwähnt, die Fehler/Residuenstreuung  $SS_{Fehler}$  für jeden Effekt eine andere ist:

- Auf Basis der Anova-Tabelle werden folgende  $\chi^2$ -Tests aufgestellt (vgl. Formel 2-7):

$$\chi^2 = \frac{SS_{Effekt}}{(SS_{Effekt} + SS_{Fehler}) / (df_{Effekt} + df_{Fehler})}$$

wobei  $SS_{Effekt}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes (C, D oder C\*D) ist,  $SS_{Fehler}$  die Streuungsquadratsumme des zum Effekt gehörenden Fehlers ist sowie  $df$  die entsprechenden Freiheitsgrade.

- Die  $\chi^2$ -Werte sind dann in den Tafeln für den  $\chi^2$ -Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade ( $df_{Effekt}$ ) des entsprechenden F-Tests sind.
- Die  $\chi^2$ -Werte sollten alternativ gemäß Iman & Davenport (vgl. Formel 2-1b) in F-Werte umgerechnet werden, in diesem Fall entspricht dies:

$$F = \frac{(n-1)\chi^2}{df1 + df2 - \chi^2}$$

wobei  $df1$  die Zähler- und  $df2$  die Nennerfreiheitsgrade des entsprechenden F-Tests sind.

Wie schon im 1-faktoriellen Fall wird empfohlen, die Gleichheit der Varianzen zu überprüfen, und zwar mit dem im Abschnitt 5.3.4 vorgestellten Levene-like-Test, anstatt eines Tests auf Sphärizität. Auch hier muss, streng genommen, die Homogenität der Varianzen für beide Haupteffekte und die Interaktion getestet werden. Für letztere ist aus den beiden Faktoren ein neuer zu bilden, hier  $Int$  genannt, der alle Kombinationen der Faktorstufen beinhaltet:

$Int = (C-1) * \#C + D$ , wobei  $\#C$  die Anzahl der Stufen von C sind.

Die Schritte sollen am Datensatz des Beispiels 5 demonstriert werden.

**mit R:**

Zunächst wird die elementare Berechnung, anschließend eine R-Funktion hierfür vorgestellt. Diese Berechnung wird wieder mit der Funktion `ezANOVA` (Paket `ez`) durchgeführt. Dieses Mal aus folgendem Grund: Bei Analysen mit Messwiederholungen ist das Ergebnisobjekt von `aov` vom Typ „aovlist“ (anstatt vom Typ „aov“). Diese sind aber äußerst kompliziert aufgebaut, so dass eine Weiterverarbeitung von Ergebnissen wie die „Sum of Sq“ und „Df“ einen erheblichen Programmieraufwand erfordert, wohingegen die Anova-Tabelle von `ezANOVA` ein simpler Dataframe ist.

Ausgehend vom in 5.1.2 erstellten Dataframe `mydata5t` werden zunächst mittels der Funktionen `ave` und `rank` pro `Vpn` die Fehlerwerte in Ränge umgerechnet und an den Dataframe angehängt. Beim Aufruf von `ezANOVA` werden mittels des Parameters `detailed` die „Sum of Sq“ sowie die „Df“ ausgegeben, die für die weiteren Berechnungen benötigt werden. Vom Ergebnis interessiert nur die Komponente `ANOVA` mit der entsprechenden Tabelle, wobei die letzten Spalten, u.a. mit den p-Werten, hier nicht wiedergegeben werden:

```
mydata5t <- within(mydata5t, RFehler<- rank(Fehler))
aov2r <- ezANOVA(mydata5t, RFehler, Vpn, within=. (Medikament, Aufgabe),
  detailed=T)
aov2ra <- aov2r$ANOVA
aov2ra
```

	Effect	DFn	DFd	SSn	SSd	F
1	(Intercept)	1	7	95922.000	7589.667	88.469498
2	Medikament	2	14	5419.083	1733.750	21.879500
3	Aufgabe	2	14	8037.750	3036.583	18.528802
4	Medikament:Aufgabe	4	28	1099.667	2774.500	2.774434

Tabelle 5-10

Die Spalten `SSn` und `SSd` (4. und 5. Spalte) enthalten die  $SS_{\text{Effekt}}$  bzw. den dazugehörigen Fehlerterm  $SS_{\text{Fehler}}$ , die Spalten `DFn` und `DFd` (2. und 3. Spalte) die entsprechenden Freiheitsgrade. Mit folgenden Anweisungen lassen sich die  $\chi^2$ -Werte berechnen und auf Signifikanz überprüfen:

```
denom <- (aov2ra[, 4]+aov2ra[, 5]) / (aov2ra[, 2]+aov2ra[, 3])
chisq <- aov2ra[, 4] / denom
df <- aov2ra[, 2]
pvalue <- 1-pchisq(chisq, df)
data.frame(Effekt=aov2ra[, 1], Chisq=chisq, Df=df,
  Pvalue=round(pvalue, digits=7))
```

	Effekt	Chisq	DF	Pvalue
1	(Intercept)	7.413425	1	0.0064739
2	Medikament	12.121817	2	0.0023323
3	Aufgabe	11.612798	2	0.0030082
4	Medikament:Aufgabe	9.083072	4	0.0590563

Alternativ kann auch die Funktion `np.anova` (vgl. Anhang 3.8) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`mydata5t`). Eingabe und Ausgabe:

```
np.anova(Fehler~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
  mydata5t, method=2)
```

Puri & Sen tests		Df	Sum Sq	Chisq	Pr(>Chi)	Pr(>F)
Medikament		2	5419.1	12.1218	0.002332	
Residuals	Medikament	14	1733.7			
Aufgabe		2	8037.8	11.6128	0.003008	
Residuals	Aufgabe	14	3036.6			
Medikament:Aufgabe		4	1099.7	9.0831	0.059056	
Residuals	Medikament:Aufgabe	28	2774.5			

Zum Test auf Gleichheit der Varianzen: Zunächst muss der Faktor `Int` aus den Faktoren `Medikament` und `Aufgabe` gebildet werden, anschließend die absoluten Differenzen für jeden Faktor. Der Friedman-Test kann hier nicht ohne Weiteres mit der Funktion `friedman.test` durchgeführt werden, bestenfalls für den Faktor `Int`. Alternativ kann die oben bereits vorgestellte Funktion `np.anova` benutzt werden.

```
mydata5t <- within(mydata5t,
  Int<- (as.integer(Medikament)-1)*3 + as.integer(Aufgabe))
mydata5t <- within(mydata5t, Int<-factor(Int))
mydata5t <- within(mydata5t,
  diff.M<-abs(Fehler-ave(Fehler,Medikament,FUN=median)))
mydata5t <- within(mydata5t,
  diff.A<-abs(Fehler-ave(Fehler,Aufgabe,FUN=median)))
mydata5t <- within(mydata5t,
  diff.I<-abs(Fehler-ave(Fehler,Int,FUN=median)))
np.anova(diff~Medikament+Error(Vpn/Medikament), mydata5t)
np.anova(diff~Aufgabe+Error(Vpn/Aufgabe), mydata5t)
np.anova(diff~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
  mydata5t)
```

Nachfolgend die Ausgabe für die Interaktion, die die Varianz-Gleichheit bestätigt:

generalized Kruskal-Wallis/Friedman tests including Iman & Davenport F-tests						
	Df	Sum Sq	Chisq	Pr(>Chi)	F value	Pr(>F)
Int	8	44.63	6.7358	0.56538	0.8234	0.5855
Residuals	56	379.37				

### mit SPSS:

Die Puri & Sen-Tests bauen auf der RT-Analyse (siehe vorigen Abschnitt, Tabelle 5-9) auf. Da hier  $\chi^2$ -Tests anstatt F-Tests verwendet werden, spielt die Sphärität keine Rolle, so dass in der o.a. Tabelle nur die Zeilen „Sphärität angen.“ relevant sind.

Die  $\chi^2$ -Werte müssen nun „mit der Hand“ aus den Werten der o.a. Tabelle (Spalten „Quadratsumme“ und „df“) berechnet werden:

$$\chi_{\text{Medikament}}^2 = \frac{5419,1}{(5419,1 + 1733,8)/(2 + 14)} = 12,12 \quad df_{\text{Medikament}} = 2$$

$$\chi_{\text{Aufgabe}}^2 = \frac{8057,8}{(8057,8 + 3036,6)/(2 + 14)} = 11,61 \quad df_{\text{Aufgabe}} = 2$$

$$\chi_{\text{Interaktion}}^2 = \frac{1099,7}{(1099,7 + 2774,5)/(4 + 28)} = 9,08 \quad df_{\text{Interaktion}} = 4$$

Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 6,0 bzw. 9,2 (df=2) und bei 9,5 bzw. 13,3 (df=4). Somit sind alle Effekte signifikant.

Die Durchführung der o.a. Tests auf Gleichheit der Varianzen lässt sich in SPSS leider nicht mit vertretbarem Aufwand durchführen. Lediglich die Gleichheit der  $J=9$  Varianzen, die für die Interaktion relevant ist, ist bequem wie in Abschnitt 5.3.4 zu testen:

```
Aggregate
  /outfile=* mode=addvariables
  /m1=median(v1) /m2=median(v2) /m3=median(v3) /m4=median(v4)
  /m5=median(v5) /m6=median(v6) /m7=median(v7) /m8=median(v8)
  /m9=median(v9) .
do repeat v=v1 to v9 / m=m1 to m9 / d=d1 to d9 .
compute d=abs(v-m) .
end repeat .
Nptests /related test(d1 d2 d3 d4 d5 d6 d7 d8 d9)
  friedman(compare=pairwise) .
```

mit dem Testergebnis  $p=0.655$ , so dass die Gleichheit angenommen werden kann.

#### 5.4.4 Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF) und van der Waerden-Test

Das KWF-Verfahren ermöglicht einen mehrfaktoriellen Friedman-Test. Dazu werden für jede Erhebungseinheit ( $V_{pn}$ ) Ränge für die einzelnen Messwiederholungen vergeben (*Friedman-Ränge*), hier die Ränge  $1, \dots, I \cdot J$ , und damit eine parametrische Varianzanalyse durchgeführt. Es werden nicht die F-Tests verwendet, sondern aus den Streuungsquadratsummen ( $SS$ , Sum of Sq) werden  $\chi^2$ -Tests konstruiert:

Für die Effekte (Haupteffekte und Interaktionen) mit Messwiederholungsfaktoren z.B. C, D, B\*C,... (vgl. Formel 2-7):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{(SS_{\text{Effekt}} + SS_{\text{Fehler}}) / (df_{\text{Effekt}} + df_{\text{Fehler}})}$$

wobei

- $SS_{\text{Effekt}}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,
- $SS_{\text{Fehler}}$  die Streuungsquadratsumme des zum Effekt gehörenden Fehlers ist sowie
- $df_{\text{Effekt}}$  und  $df_{\text{Fehler}}$  die entsprechenden Freiheitsgrade.

Die  $\chi^2$ -Werte sind dann in den Tafeln für den  $\chi^2$ -Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade ( $df_{\text{Effekt}}$ ) des entsprechenden F-Tests sind.

Der van der Waerden-Test unterscheidet sich von dem o.a. KWF-Verfahren nur dadurch, dass die für jede Erhebungseinheit ( $V_{pn}$ )  $m$  ermittelten Ränge  $R(x_{mk})$  ( $k=1, \dots, I \cdot J$ ) noch zusätzlich in normal scores umgerechnet werden:

$$n\text{score}_{mk} = \Phi^{-1}(R(x_{mk}) / (IJ + 1))$$

Die Überprüfung der Sphärität kann entfallen, da hier  $\chi^2$ - anstatt F-Tests durchgeführt werden. Statt dessen sollte ein Test auf Gleichheit der Varianzen, wie im vorigen Abschnitt gezeigt, durchgeführt werden.

Die Schritte sollen am Datensatz des Beispiels 5 demonstriert werden.

**mit R:**

Ein Beispiel für die elementare Berechnung ist in Kapitel 6 zu finden. Der Einfachheit halber wird hier die R-Funktion `np.anova` verwendet. Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `mydata5t`. Der Aufruf für das KFW-Verfahren:

```
np.anova(Fehler~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
         mydata5t,method=0)
```

```
generalized Kruskal-Wallis/Friedman tests including Iman & Davenport
F-tests
```

	Df	Sum Sq	Chisq	Pr(>Chi)	F value	Pr(>F)
Medikament	2	111.062	12.7536	0.0017006	27.4996	1.416e-05
Residuals Medikament	14	28.271				
Aufgabe	2	150.583	11.1889	0.0037185	16.2793	0.0002223
Residuals Aufgabe	14	64.750				
Medikament:Aufgabe	4	26.417	11.5273	0.0212356	3.9414	0.0116312
Residuals Medik:Aufg	28	46.917				

und für den van der Waerden-Test:

```
np.anova(Fehler~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
         mydata5t,method=1)
```

```
generalized van der Waerden tests
```

	Df	Sum Sq	Chisq	Pr(>Chi)
Medikament	2	10.1452	12.944	0.0015459
Residuals Medikament	14	2.3949		
Aufgabe	2	13.7037	11.537	0.0031238
Residuals Aufgabe	14	5.3005		
Medikament:Aufgabe	4	2.7001	13.454	0.0092566
Residuals Medikament:Aufgabe	28	3.7218		

Hiernach sind wie bei der parametrischen Analyse alle drei Effekte bei beiden Methoden signifikant.

**mit SPSS:**

Zuerst muss eine Umstrukturierung des Datensatzes erfolgen, um die Friedman-Ränge für die Variablen `V1`,...,`V9` berechnen zu können. Anschließend die Rangberechnung (`Ry`) und für den van der Waerden-Test zusätzlich die Transformation in normal scores (`nscore`). Danach wird die ursprüngliche Datenstruktur wieder hergestellt, wobei die abhängige Variable `Ry` die Namen `Ry.1`,...,`Ry.9` bzw. `nscore` die Namen `nscore.1`,...,`nscore.9` erhält. Schließlich die Varianzanalyse mit doppelter Messwiederholung.

```
Varstocases
  /Id=Vpn
  /Make Fehler from v1 to v9
  /index=Medikament(3) Aufgabe(3)
  /keep=Geschlecht
  /null=keep.

Rank Variables=Fehler by Vpn / rank into RFehler.
compute nscore=Idf.normal(RFehler/10,0,1).

Sort cases by Vpn Medikament Aufgabe.
Casestovars
```

```

/Id=Vpn
/index=Medikament Aufgabe
/groupby=variable.

```

### Zunächst die parametrische Varianzanalyse für das KWF-Verfahren

```

GLM Rfehler.1.1 to Rfehler.3.3
/WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
/WSdesign=Medikament Aufgabe Medikament*Aufgabe.

```

Tests der Innersubjekteffekte

Quelle		Quadratsum. vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angenommen	111,063	2	55,531	27,500	,000
Fehler(Medikament)	Sphärizität angenommen	28,271	14	2,019		
Aufgabe	Sphärizität angenommen	150,583	2	75,292	16,279	,000
Fehler(Aufgabe)	Sphärizität angenommen	64,750	14	4,625		
Medikament * Aufgabe	Sphärizität angenommen	26,417	4	6,604	3,941	,012
Fehler(Medi*Aufgabe)	Sphärizität angenommen	46,917	28	1,676		

In obiger Anova-Tabelle sind für jeden Effekt jeweils nur die Zeile wiedergegeben, da die Quadratsummen für alle Adjustierungsversionen identisch sind. Auf Basis der o.a. Quadratsummen werden nun die folgenden Tests errechnet:

$$\chi^2_{\text{Medikament}} = \frac{111,063}{(111,063 + 28,271)/(2 + 14)} = 12,75 \quad df_{\text{Medikament}} = 2$$

$$\chi^2_{\text{Aufgabe}} = \frac{150,58}{(150,58 + 64,75)/(2 + 14)} = 11,19 \quad df_{\text{Aufgabe}} = 2$$

$$\chi^2_{\text{Interaktion}} = \frac{28,41}{(28,41 + 46,92)/(4 + 28)} = 11,52 \quad df_{\text{Interaktion}} = 4$$

### Und nun die parametrische Varianzanalyse für das van der Waerden-Verfahren

```

GLM nscore.1.1 to nscore.3.3
/WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
/WSdesign=Medikament Aufgabe Medikament*Aufgabe.

```

Tests der Innersubjekteffekte

Quelle		Quadratsum. vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angenommen	10,145	2	5,073	29,653	,000
Fehler(Medikament)	Sphärizität angenommen	2,395	14	,171		
Aufgabe	Sphärizität angenommen	13,704	2	6,852	18,097	,000
Fehler(Aufgabe)	Sphärizität angenommen	5,301	14	,379		
Medikament * Aufgabe	Sphärizität angenommen	2,700	4	,675	5,078	,003
Fehler(Medi*Aufgabe)	Sphärizität angenommen	3,722	28	,133		

Aus diesen Quadratsummen lassen sich nun die  $\chi^2$ -Tests wie folgt ermitteln:

$$\chi_{\text{Medikament}}^2 = \frac{10,145}{(10,145 + 2,395)/(2 + 14)} = 12,94 \quad df_{\text{Medikament}} = 2$$

$$\chi_{\text{Aufgabe}}^2 = \frac{13,704}{(13,704 + 5,3)/(2 + 14)} = 11,54 \quad df_{\text{Aufgabe}} = 2$$

$$\chi_{\text{Interaktion}}^2 = \frac{2,7}{(2,7 + 3,722)/(4 + 28)} = 13,45 \quad df_{\text{Interaktion}} = 4$$

#### 5.4.5 Aligned rank transform (ART und ART+INT)

Das Prinzip des Aligned rank transform-Tests wurde oben bereits erläutert (vgl. Kapitel 4.3.6). Die Schritte noch einmal im Einzelnen:

- Durchführung einer (normalen) Anova mit Haupt- und Interaktionseffekten.
- Speichern der Residuen ( $e_m$ ),
- Eliminieren des zu untersuchenden Effekts aus den Residuen:  
 Interaktionseffekt:  $e_m + (\bar{a}\bar{b}_{ij} - \bar{a}_i - \bar{b}_j + 2\bar{x})$   
 Haupteffekte:  $e_m + (\bar{a}_i + \bar{b}_j - \bar{x})$
- Umrechnung der bereinigten Residuen in Ränge.
- Durchführung einer normalen Anova mit Haupt- und Interaktionseffekten mit den Rängen, aus der dann der untersuchte Effekt abgelesen werden kann.

Es sei noch einmal darauf aufmerksam gemacht, dass die ART-Tests für die beiden Haupteffekte statistisch nicht erforderlich sind und sogar falsch signifikante Ergebnisse bringen können.

Dieses Verfahren stellt in erster Linie eine Verbesserung des o.a. Rank transform Tests da, um die Haupt- und Interaktionseffekte sauber zu trennen (vgl. Kapitel 4.3.6). Es ist also in erster Linie für metrische Variablen gedacht, die die Normalverteilungs-Voraussetzung nicht erfüllen, nicht jedoch für Variablen mit beliebigen Eigenschaften. Insofern sollte die Möglichkeit genutzt werden, die rangtransformierten Daten mittels des Mauchly-Tests auf Varianzhomogenität bzw. Sphärität zu überprüfen, um dann gegebenenfalls anstatt des normalen F-Tests die Variante von Huynh & Feldt anzuwenden. Oder alternativ ohne Beachtung des Mauchly-Tests die adjustierten F-Tests z.B. von Huynh & Feldt verwenden. Nach Beasley (2002) spielt bei dieser ART-Methode die Sphärität keine Rolle, so dass ein Blick auf den Mauchly-Test entfallen kann und in der Anova-Tabelle ausschließlich der „normale“ F-Test von Bedeutung ist.

Es wird empfohlen (siehe Mansouri & Chang, 1995 sowie Carletti & Clautriaux, 2005) anschließend die Ränge in normal scores (vgl. Kapitel 2.3) umzurechnen (ART+INT-Verfahren), um einerseits etwaige falsche Signifikanzen abzuschwächen und andererseits eine größere Power zu erhalten.

Es soll nun im Folgenden für den Beispieldatensatz 5 überprüft werden, ob die oben ausgewiesene Signifikanz der Interaktion garantiert ist.

#### mit R:

Zunächst wird die elementare Berechnung, anschließend eine R-Funktion hierfür vorgestellt. Ausgehend vom in Kapitel 5.1.2 erstellten Dataframe `mydata5t` werden zunächst

- die Residuen der Varianzanalyse mit den Faktoren `Medikament` und `Aufgabe` ermittelt (vgl. dazu 5.3.1),



- die Effekte  $ma$  des Faktors `Medikamente` bzw.  $mb$  des Faktors `Aufgaben` berechnet,
- die Zellenmittelwerte  $mab$  sowie den Gesamtmittelwert  $mm$ ,
- in der Variablen  $rabr$  die Residuen um die Haupteffekte bereinigt und in Ränge transformiert,
- in der Variablen  $rar$  die Residuen um den Interaktionseffekt bereinigt und in Ränge transformiert.
- Anschließend werden Varianzanalysen für  $rabr$  zum Test des Interaktionseffekts durchgeführt:

```
aov3r <- aov(Fehler~Medikament*Aufgabe + Vpn, mydata5t)
mydata5s <- cbind(mydata5t, resid=aov3r$residuals)
mydata5s <- within(mydata5s,
  { ma <- ave(Fehler,Medikament,FUN=mean);
    mb <- ave(Fehler,Aufgabe,FUN=mean);
    mab<- ave(Fehler,Medikament,Aufgabe, FUN=mean);
    mm <- mean(Fehler)})
mydata5s <- within(mydata5s,
  { rabr<- rank(round(resid-mab+ma+mb-mm,digits=7));
    rar <- rank(round(resid-ma-mb+2*mm,digits=7))})
aov3rab <- aov(rabr~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
  mydata5s)
summary(aov3rab)
aov3ra <- aov(rar~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
  mydata5s)
summary(aov3ra)
```

Nachfolgend zunächst die Ergebnisse der Anova zum Test des Interaktionseffekts, dessen Signifikanz ( $p=0.017$ ) danach bestätigt ist:

Error: Vpn						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Residuals	7	518.6	74.08			
Error: Vpn:Medikament						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Medikament	2	10	5.0	0.011	0.989	
Residuals	14	6215	443.9			
Error: Vpn:Aufgabe						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Aufgabe	2	32	16.0	0.019	0.981	
Residuals	14	11491	820.8			
Error: Vpn:Medikament:Aufgabe						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Medikament:Aufgabe	4	4363	1090.8	3.617	0.0169	*
Residuals	28	8443	301.5			

Tabelle 5-12

sowie der Ergebnisse für  $rar$  zum Test der Haupteffekte, die beide signifikant sind:

```

Error: Vpn
      Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7      52    7.429

Error: Vpn:Medikament
      Df Sum Sq Mean Sq F value    Pr(>F)
Medikament  2   8042    4021   25.11 2.34e-05 ***
Residuals  14   2242     160

Error: Vpn:Aufgabe
      Df Sum Sq Mean Sq F value    Pr(>F)
Aufgabe    2  12830    6415   23.58 3.29e-05 ***
Residuals  14   3808     272

Error: Vpn:Medikament:Aufgabe
      Df Sum Sq Mean Sq F value Pr(>F)
Medikament:Aufgabe  4    185    46.32    0.33  0.855
Residuals          28   3931   140.40

```

Tabelle 5-13

Schließlich noch die Alternative mit der R-Funktion `art2.anova` (vgl. Anhang 3.10). Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz `mydata5t`. Eingabe und Ausgabe:

```
art2.anova(Fehler~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
mydata5t)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Medikament	2	27.4	13.72	20.8313	6.367e-05 ***
Residuals	14	9.2	0.66		
Aufgabe	2	40.8	20.39	20.5520	6.833e-05 ***
Residuals	14	13.9	0.99		
Medikament:Aufgabe	4	4363.0	1090.76	3.6173	0.01692 *
Residuals	28	8443.0	301.54		

Zur Anwendung des ART+INT-Verfahrens müssen die nach dem ART-Verfahren errechneten Ränge in normal scores (vgl. Kapitel 2.3) transformiert werden. Zunächst mittels der zuerst angeführten elementaren Berechnung. Dazu ist *vor* Durchführung der Varianzanalyse noch die Ermittlung des  $N$  ( $nc$ ) sowie die Transformation mittels der inversen Normalverteilung erforderlich, hier allerdings nur für die Prüfung der Interaktion vorgestellt:

```
nc<-dim(mydata5s)[1]
nsabr <- qnorm(mydata5s$rabr/(nc+1))
aov3rab <- aov(nsabr~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
mydata5s)
summary(aov3rab)
```

```

....
Error: Vpn:Medikament:Aufgabe
      Df Sum Sq Mean Sq F value Pr(>F)
Medikament:Aufgabe  4   9.038   2.2594    3.231 0.0267 *
Residuals          28  19.579   0.6992

```

Das Testergebnis für den Interaktionseffekt ist in der o.a. Tabelle, die genauso aufgebaut ist wie Tabelle 5-13, unter `Vpn:Medikament:Aufgabe` abzulesen.

Einfacher ist dies mittels der o.a. Funktion `art2.anova` über den zusätzlichen Parameter `INT` möglich, wobei auf die Ausgabe hier verzichtet wird:

```
art2.anova(Fehler~Medikament*Aufgabe+Error(Vpn/(Medikament*Aufgabe)),
          mydata5t, INT=T)
```

### mit SPSS:

Wie beim Rank Transform-Test (vgl. Kapitel 5.4.2) muss zunächst der Datensatz umstrukturiert werden, wobei die Messwiederholungen in Fälle gewandelt werden.

```
Varstocases
  /Id=Vpn
  /make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
  /index=Medikament(3) Aufgabe(3)
  /keep=Geschlecht
  /null=keep.
```

Mit diesem Datensatz wird zur Ermittlung der Residuen des Modells mit den Faktoren `Medikament` und `Aufgaben` eine Varianzanalyse (ohne Messwiederholungen, dafür mit dem Faktor `Vpn` der Versuchspersonenkennung) gerechnet (im Menü „Modell“ „Anpassen“ wählen, die Interaktion von `Medikament` und `Aufgaben` für die rechte Seite auswählen sowie den Haupteffekt `Vpn`):

```
Unianova Fehler by Medikament Aufgabe Vpn
  /save=resid
  /design=Aufgabe*Medikament Vpn.
```

Über `Aggregate` werden nun die Mittelwerte für `Medikament` ( $a_i$ ), `Aufgaben` ( $b_j$ ), Zellen ( $m_{ij}$ ) und gesamt ( $mm$ ) berechnet, um die Effekte von den Residuen abzuziehen und das Ergebnis in Ränge umzurechnen:

- $r_{ab}$  bzw. die Ränge  $r_{abr}$  zum Test der Interaktion
- $r_a$  bzw.  $r_{ar}$  zum Test der Haupteffekte

```
Aggregate
  /outfile=* mode=addvariables
  /break=Medikament Aufgabe /mij=mean(Fehler).
Aggregate
  /outfile=* mode=addvariables
  /break=Medikament /ai=mean(Fehler).
Aggregate
  /outfile=* mode=addvariables
  /break=Aufgabe /bj=mean(Fehler).
Aggregate
  /outfile=* mode=addvariables
  /break= /mm=mean(Fehler).
Compute rab = res_1 + (mij - ai - bj + 2*mm).
Compute ra = res_1 + (ai + bj - mm).
Rank variables=ra rab (A)
  /rank into rar rabr.
execute.
```

Anschließend wird der Datensatz wieder in die ursprüngliche Form transformiert:

```
Sort cases by Vpn Medikament Aufgabe.
Casestovars /Id=Vpn
  /index=Medikament Aufgabe
  /groupby=variable.
```

Schließlich wird dann für  $rabr$ , die im umstrukturierten Datensatz die Namen  $rabr.1.1$ ,  $rabr.1.2$ , ... hat, bzw.  $rar$ , eine Varianzanalyse mit Messwiederholungen mit den Faktoren **Medikament** und **Aufgaben** gerechnet:

```
GLM rabr.1.1 rabr.1.2 rabr.1.3 rabr.2.1 rabr.2.2 rabr.2.3
    rabr.3.1 rabr.3.2 rabr.3.3
  /wsfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
  /wsdesign=Medikament Aufgabe Medikament*Aufgabe.
```

Nachfolgend die Ergebnisse für den Test der Interaktion (ohne Wiedergabe der Fehlerterme). Nach Beasley (2002) spielt bei dieser ART-Methode die Sphärizität keine Rolle, so dass ein Blick auf den Mauchly-Test entfallen kann und in der Anova-Tabelle ausschließlich die Zeile „Sphärizität angenommen“ von Bedeutung ist:

Tests der Innersubjekteffekte						
Quelle		Quadrat- summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	9,146	2	4,573	,010	,990
Aufgabe	Sphärizität angen.	30,896	2	15,448	,019	,981
Medikament * Aufgabe	Sphärizität angen.	4313,458	4	1078,365	3,573	,018

Tabelle 5-14

bzw. die Anova-Tabelle für den Test der Haupteffekte:

Tests der Innersubjekteffekte						
Quelle		Quadrat- summe vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	8042,312	2	4021,156	25,113	,000
Fehler(Medikament)	Sphärizität angen.	2241,687	14	160,121		
Aufgabe	Sphärizität angen.	12830,333	2	6415,167	23,584	,000
Fehler(Aufgabe)	Sphärizität angen.	3808,167	14	272,012		
Medikament * Aufgabe	Sphärizität angen.	185,292	4	46,323	,330	,855
Fehler(Medikament* Aufgabe)	Sphärizität angen.	3931,208	28	140,400		

Tabelle 5-15

Für die Umrechnung in normal scores, d.h. Anwendung des ART+INT-Verfahrens, müssen noch *vor* der Rücktransformation der Datenmatrix die folgenden Anweisungen zur Berechnung der Fallzahl ( $nc$ ) und der INT-Transformation eingefügt werden:

```
Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(Fehler).
compute nsar =Idf.normal(rar / (nc+1), 0, 1).
compute nsabr=Idf.normal(rabr/ (nc+1), 0, 1).
execute.
```

Nachdem die Datenmatrix wieder die normale Struktur hat, erfolgt die Varianzanalyse (hier nur für die Interaktion) über:

```
GLM nsabr.1.1 nsabr.1.2 nsabr.1.3 nsabr.2.1 nsabr.2.2 nsabr.2.3
    nsabr.3.1 nsabr.3.2 nsabr.3.3
  /wsfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
  /wsdesign=Medikament Aufgabe Medikament*Aufgabe.
```

Bei der Ausgabe interessieren auch hier wieder nur die Zeilen „Sphärizität angenommen“:

Tests der Innersubjekteffekte						
Quelle		Quadrat- summe vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	,039	2	,019	,019	,981
Fehler(Medikament)	Sphärizität angen.	14,413	14	1,030		
Aufgabe	Sphärizität angen.	,002	2	,001	,001	,999
Fehler(Aufgabe)	Sphärizität angen.	22,520	14	1,609		
Medikament * Aufgabe	Sphärizität angen.	9,421	4	2,355	3,596	,017
Fehler (Medikament*Aufgabe)	Sphärizität angen.	18,341	28	,655		

#### 5. 4. 6 ATS-Tests von Akritas, Arnold & Brunner

Den von Akritas et al. entwickelten ATS-Test gibt es auch für mehrfaktorielle Varianzanalysen mit Messwiederholungen. Während in R dazu die Pakete `nparLD` und `MANOVA.RM` zur Verfügung stehen, gibt es in SPSS derzeit keine Möglichkeit zur Anwendung dieses Verfahrens.

##### mit R:

Die 2-faktorielle Analyse mittels `nparLD` soll am Datensatz des Beispiels 5 gezeigt werden. Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `mydata5t`. Die Analyse kann mittels zwei Funktionen erfolgen:

- `nparLD` ist eine universelle Funktion für alle verarbeitbaren Designs.
- `ld.f2` erlaubt fehlende Werte bei den Messwiederholungen, gibt einen Mittelwertplot aus sowie eine Reihe weiterer, hier allerdings nicht interessierende Statistiken aus.

Beide geben sowohl die WTS als auch die ATS aus. Die Ausgabe unterscheidet sich nicht hinsichtlich dieser Statistiken. Nachfolgend zunächst die Eingabe für beide Varianten, wobei zu beachten ist, dass bei der Funktion `nparLD` trotz Angabe des Dataframes die Variablennamen nicht automatisch gefunden werden. Daher muss entweder jeder Variablenname zusammen mit dem Dataframe-Namen in der üblichen Form, z.B. `mydata5t$Fehler` angegeben werden oder mit `with(Dataframe, ...)` ausgeführt werden.

```
attach(mydata5t)
with(mydata5t, nparLD(Fehler~Medikament*Aufgabe, mydata5t, mydata5t$Vpn))
with(mydata5t, ld.f2(score, Medikament, Aufgabe, Vpn,
  time1.name="Medikament", time2.name="Aufgabe")) -> ano
round(ano$ANOVA.test, 4)
```

Bei `ld.f2` müssen die Faktoren zweimal angegeben werden: zum einen zur Identifikation des Faktors, zum anderen in „...“ als Name des Faktors für die Ausgabe.

Nachfolgend die Ausgabe von `nparLD`:

```

Call:
lm(Fehler ~ Medikament * Aufgabe)

Wald-Type Statistic (WTS):
      Statistic df      p-value
Medikament    44.43367  2 2.245694e-10
Aufgabe       43.50097  2 3.580012e-10
Medikament:Aufgabe 12.38836  4 1.468530e-02

ANOVA-Type Statistic (ATS):
      Statistic      df p-value
Medikament    21.8795  1.3076  0.0000
Aufgabe       18.5288  1.8559  0.0000
Medikament:Aufgabe  2.7744  1.9617  0.0635

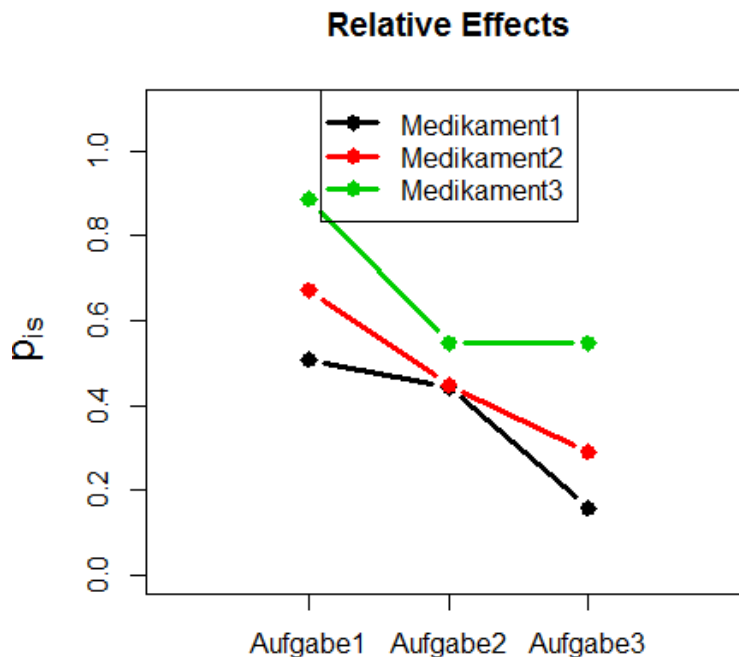
```

Tabelle 5-16

Bei der Ausgabe von `ld.f2` gibt es die Möglichkeit, einzelne Teile auszugeben, etwa die ATS- (Anova-) Tabelle (`..$ANOVA.test`) oder die WTS- (Wald-Test-) Tabelle (`..$wald.test`). Dies hat den Vorteil, dass man über die Funktion `round` die Zahlendarstellung der Art `xxxe-nn` ändern kann.

	Statistic	df	p-value
Medikament	21.8795	1.3076	0.0000
Aufgabe	18.5288	1.8559	0.0000
Medikament:Aufgabe	2.7744	1.9617	0.0635

`ld.f2` gibt noch zusätzlich einen Interaktionsplot aus (siehe nächste Seite), allerdings der relativen Effekte (vgl. Kapitel 2.8) anstatt der Mittelwerte, da sich ja die Hypothesen auf erstere beziehen:



### 5.4.7 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Akritas & Arnold

Die meisten der in Abschnitt 5.3.9 vorgestellten multivariaten Methoden können auch im Fall zweier Messwiederholungsfaktoren angewandt werden. Soll das nichtparametrische Verfahren von Akritas & Arnold benutzt werden, so sind sämtliche Werte in Ränge  $1, \dots, n \cdot I \cdot J$  zu transformieren, bevor die parametrische multivariate Varianzanalyse, wie unten gezeigt, angewandt wird. Als Beispiel dient wieder der Datensatz `mydata5`.

#### mit R:

Will man lediglich die beiden Haupteffekte Medikament und Aufgabe testen, so müssen zunächst die Summen über den jeweilig anderen Faktor ermittelt werden: `m1`, `m2`, `m3` bzw. `a1`, `a2`, `a3`. Danach werden wie beim 1-faktoriellen Fall die Differenzen errechnet und mittel `lm` und `anova` überprüft: Als Test wird die Methode von Wilks gewählt.

```
rowSums(mydata5[,c("v1", "v2", "v3")]) -> m1
rowSums(mydata5[,c("v4", "v5", "v6")]) -> m2
rowSums(mydata5[,c("v7", "v8", "v9")]) -> m3
rowSums(mydata5[,c("v1", "v4", "v7")]) -> a1
rowSums(mydata5[,c("v2", "v5", "v8")]) -> a2
rowSums(mydata5[,c("v3", "v6", "v9")]) -> a3
cbind(mydata5, m1, m2, m3, s1, s2, s3) -> mydata5

anova(lm(cbind(s2-s1, s3-s2) ~ 1, mydata5), test="Wilks")
```

Analysis of Variance Table						
	Df	Wilks	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.11026	24.209	2	6	0.00134 **
Residuals	7					

```
anova(lm(cbind(m2-m1, m3-m2) ~ 1, mydata5), test="Wilks")
```

Analysis of Variance Table						
	Df	Wilks	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.133	19.557	2	6	0.002352 **
Residuals	7					

Das Testergebnis ist jeweils in der Zeile `(Intercept)` abzulesen. Für die Interaktion `Medikament*Aufgabe` lässt sich leider auf diesem Weg wegen der Differenzenbildung kein eindeutiger Test durchführen. Dazu ist ein anderer Weg zu beschreiten mittels der Funktion `Anova` aus dem `car`-Paket, die allerdings ein wenig Vorarbeit erfordert,

- Medikament, der Faktor mit den langsamer laufenden Indizes, wird als Faktor mit 3 Stufen, aber auch mit 3 Wiederholungen definiert,
- Aufgabe, der Faktor mit den schneller laufenden Indizes, wird als "normaler" Faktor mit 3 Stufen definiert,
- aus beiden wird ein Dataframe `idf` erstellt,
- wie oben wird mittels `lm` ein multivariates lineares Modell `mod` gerechnet,
- das dann mittels `Anova` als Varianzanalyse mit doppelter Messwiederholung analysiert wird:

```

Medikament <- factor(rep(c("Kontrolle", "Med A", "Med B"), each=3),
                    levels=c("Kontrolle", "Med A", "Med B"))
idf <- data.frame(Medikament, Aufgabe)
mod <- lm(cbind(v1, v2, v3, v4, v5, v6, v7, v8, v9) ~ 1, data=mydata5)
Anova(mod, idata=idf, idesign=~Medikament*Aufgabe, type=3, test="Wilks")

```

Type III Repeated Measures MANOVA Tests: Wilks test statistic						
	Df	test stat	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.053972	122.698	1	7	1.086e-05 ***
medikament	1	0.110258	24.209	2	6	0.001340 **
Aufgabe	1	0.132996	19.557	2	6	0.002352 **
medikament:Aufgabe	1	0.230606	3.336	4	4	0.135011

### mit SPSS:

In Abschnitt 5.4.1 war bereits der Datensatz `mydata5` mit SPSS analysiert worden. Dort zeigte sich auch, dass die Voraussetzung der Spherizität nicht erfüllt ist. Daher ist es sinnvoll, die Daten multivariat auszuwerten, eine Alternative zur oben verwendeten Huynh-Feldt-Korrektur. Die nachfolgende Tabelle wird automatisch mit den in 5.4.1 aufgelisteten Anweisungen erzeugt (und war oben nicht wiedergegeben).

Effekt		Wert	F	Hypot df	Fehler df	Sig.
Medikament	Pillai-Spur	,890	24,209	2,000	6,000	,001
	Wilks-Lambda	,110	24,209	2,000	6,000	,001
	Hotelling-Spur	8,070	24,209	2,000	6,000	,001
Aufgabe	Pillai-Spur	,867	19,557	2,000	6,000	,002
	Wilks-Lambda	,133	19,557	2,000	6,000	,002
	Hotelling-Spur	6,519	19,557	2,000	6,000	,002
Medikament *	Pillai-Spur	,769	3,336 <sup>b</sup>	4,000	4,000	,135
	Wilks-Lambda	,231	3,336 <sup>b</sup>	4,000	4,000	,135
Aufgabe	Hotelling-Spur	3,336	3,336 <sup>b</sup>	4,000	4,000	,135

### 5.4.8 Bredenkamp Tests

Bredenkamp (vgl. Lienert, 1987, S. 1024 ff und Bredenkamp, 1974) hat für den Versuchsplan mit Messwiederholungen auf zwei Faktoren auch eine Lösung vorgeschlagen, die sich aber nicht mit dem verallgemeinerten Friedman-Test (vgl. Kapitel 5.4.3) deckt. Das Verfahren erfolgt analog zu dem für unabhängige Stichproben (vgl. Kapitel 4.3.8):

- Für den Test von Faktor A wird für jede Stufe von A die Summe der abhängigen Variablen über die Stufen von Faktor B ermittelt. Anschließend wird ein Friedman-Test über diese Summen durchgeführt.
- Der Test von Faktor B erfolgt analog zu dem o.a. Test für Faktor A.
- Für den Test der Interaktion wird zunächst ein Friedman-Test über alle Messwiederholungen durchgeführt. Von dem resultierenden  $\chi^2$ -Wert werden die beiden  $\chi^2$ -Werte von den Tests von A und B subtrahiert. Analog werden die Freiheitsgrade subtrahiert. Das Ergebnis wird schließlich in der Tabelle der  $\chi^2$ -Verteilung überprüft.

Leider gibt es zu diesem Verfahren keine vergleichenden Tests.



## 6. Gemischte Versuchspläne

Unter *gemischten Versuchsplänen*, auch *Split Plot Designs* genannt, versteht man in der Regel solche, in denen sowohl Messwiederholungsfaktoren als auch Gruppierungsfaktoren enthalten sind. So wird im Folgenden davon ausgegangen, dass ein Merkmal  $x$   $J$ -mal (unter verschiedenen Bedingungen) erhoben wurde, so dass Variablen  $x_1, \dots, x_J$  vorliegen, deren Mittelwerte verglichen werden sollen. Die Struktur kann aber auch hier mehrfaktoriell sein. Die Ausgangssituation ist also ähnlich wie in Kapitel 5. Hier kommt allerdings hinzu, dass die Beobachtungseinheiten, z.B. Versuchspersonen, Gruppen zugeordnet sind, deren Einfluss ebenfalls getestet werden soll.

Der einfachste Fall der in diesem Abschnitt betrachteten Versuchspläne beinhaltet jeweils einen Gruppierungsfaktor sowie einen Messwiederholungsfaktor. Allerdings unterscheiden sich mehrfaktorielle Designs, etwa mit zwei oder mehr Gruppierungsfaktoren oder mehreren Messwiederholungsfaktoren, nicht grundsätzlich von dem hier behandelten einfachen Fall. Verschiedentlich wird auf die Ausdehnung auf mehr als zwei Faktoren kurz eingegangen. Für den Fall zweier Messwiederholungsfaktoren sind zum Teil die Ergebnisse des letzten Kapitels 5 hier anzuwenden. Beispiele für 3-faktorielle Versuchspläne bieten die Datensätze 5, mit zwei Messwiederholungsfaktoren und einem Gruppierungsfaktor, sowie 6, mit einem Messwiederholungsfaktor und zwei Gruppierungsfaktoren, die zu Beginn des Kapitels 5 vorgestellt wurden. An die Datenstruktur werden dieselben Anforderungen gestellt wie in Kapitel 5.1 beschrieben.

Im Folgenden wird weitgehend der einfache 2-faktorielle Fall behandelt. Ein entsprechender Datensatz bieten die Beispieldaten 4 (winer518).

### 6. 1 Voraussetzungen der parametrischen Varianzanalyse

Hier geht es um Versuchspläne, die sowohl abhängige als auch unabhängige Stichproben beinhalten. Für den einfachsten Fall einer 2-faktoriellen Varianzanalyse mit einem Gruppierungsfaktor A (mit  $I$  Gruppen) und einem Messwiederholungsfaktor C (mit  $J$  Wiederholungen) lautet das Modell dann:

$$x_{ijm} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \pi_m + \varepsilon_{ijm} \quad (i=1, \dots, I, j=1, \dots, J \text{ und } m=1, \dots, n_i) \quad (6-1)$$

Auch hier gibt es einen personenspezifischen Effekt:  $\pi_m$ . Die Hypothesen sind dieselben wie in Kapitel 4 und 5. Die Voraussetzungen betreffen wiederum die Normalverteilung der Residuen und die Varianzhomogenität. Und hier kumulieren sich jetzt die Voraussetzungen der Analysen ohne Messwiederholungen (siehe Kapitel 4.1) sowie der Analysen mit Messwiederholungen (siehe Kapitel 5.2), die hier allerdings zum Teil etwas abgewandelt werden. Dazu kommen allerdings noch weitere, auf die nachfolgend näher eingegangen wird.

#### Normalverteilung

Doch zunächst wieder zur Normalverteilung der Residuen sowie der Personeneffekte  $\pi_m$ : Hier sind dieselben Schritte erforderlich wie in Kapitel 5.2 beschrieben.

#### Varianzhomogenität - Sphärität

Zur Varianzhomogenität hinsichtlich der Messwiederholungen: Auch hier ist wie in 5.2 beschrieben ein Test auf Sphärität (für alle Messwiederholungsfaktoren und deren Interaktionen) durchzuführen. Doch hier gibt es noch eine weitere Hürde: Die Kovarianzmatrizen der einzelnen Gruppen (für  $i=1, \dots, I$ ) müssen (statistisch) gleich sein. Näheres dazu weiter unten.

Die Verfahren zur Überprüfung der Spherizität wurden in Kapitel 5.2 vorgestellt. Die Anwendung erfolgt allerdings nicht auf die „normale“ Kovarianzmatrix, sondern auf die „gepoolte“. Diese errechnet sich ähnlich der normalen, wobei jedoch die Abweichungen nicht vom Gesamtmittelwert  $\bar{x}$ , sondern von den jeweiligen Gruppenmittelwerten  $\bar{x}_i$  verwendet werden. Allerdings gibt es auch eine Version des Mauchly-Tests von Mendoza (1980), der nicht die Homogenität der Kovarianzmatrizen verlangt. Er ist auch in der Funktion `check.sphere` enthalten.

### Alternativen - fehlende Spherizität

Für den Fall von Verletzungen der Spherizität kann wieder die Approximation von Huynh & Feldt (alternativ von Geisser & Greenhouse) empfohlen werden. Doch ein Hinweis: Hier, im Fall von Gruppierungsfaktoren, ist der meistens benutzte Korrekturfaktor  $\epsilon$  von Huynh & Feldt nicht korrekt, insbesondere etwas zu groß. Lecoutre (1991) hat ein korrektes  $\epsilon$  angegeben. (Im Zähler des  $\epsilon$  von Huynh & Feldt ist  $N$  durch  $(N-I+1)$  zu ersetzen. Doch leider ist dies bislang weder in R noch in SPSS berücksichtigt worden. Der Fehler ist in der Regel vernachlässigbar, denn er macht sich erst bei größerer Gruppenanzahl bemerkbar.

Andere Verfahren, die keine Spherizität voraussetzen, sind:

- den in Kapitel 2.12 erwähnten und in Kapitel 5.2 kurz vorgestellten multivariaten Test (z.B. von *Hotellings-Lawley*, *Wilks* oder *Pillai*) zum Test des Messwiederholungseffekts, wobei die Interaktion von Messwiederholungsfaktor mit Gruppierungsfaktor sich als Haupteffekt des Gruppierungsfaktors angewandt auf die Differenzen errechnet,
- die auf diesem multivariaten Test basierende nichtparametrischen Verfahren von Agresti & Pendergast sowie von Akritas & Arnold (Kapitel 5.3.9 und 6.10.1),
- die auf sphärischen Vorzeichen und Rängen basierende multivariate Varianzanalyse von Sirkiä (vgl. Kapitel 6.10.2),
- die in Kapitel 2.12 erwähnte Varianzanalyse von Koch, die den oben erwähnten multivariaten Test zum Test des Messwiederholungseffekts auf Rangdaten überträgt und damit das Problem der Spherizität umgeht,
- die Kapitel 2.15 vorgestellten *Generalized Linear Models* (GEE und GLMM),
- die in Kapitel 2.13 vorgestellten Methoden für Adjustierungen der Freiheitsgrade GA und IGA, quasi Verbesserungen der in Kapitel 5.2 beschriebenen  $\epsilon$ -Adjustierung von Huynh & Feldt,
- der in Kapitel 2.13 vorgestellte modifizierte Test von Brown & Forsythe.

### Varianzhomogenität - Kovarianzmatrizen

Zur Varianzhomogenität bzgl. der Gruppen. Diese entspricht der Varianzhomogenität aus Kapitel 4.1. Sie beinhaltet hier die statistische Gleichheit der Kovarianzmatrizen, die erforderlich ist, um sie zu einer Matrix (der o.a. „gepoolten“ Kovarianzmatrix) zusammenfassen zu können. Dies ist eine generelle Voraussetzung für die gemischten Versuchspläne, nicht nur, um die Spherizität überprüfen zu können. Und zwar verlangen sowohl die „normalen“ univariaten F-Tests, auch mit den unten vorgestellten Korrekturen, als auch die in 2.12 vorgestellten und in 6.10 besprochenen multivariaten Tests die Gleichheit der Kovarianzmatrizen (vgl. z.B. Kesselman et al., 1993). Simulationsstudien haben gezeigt, dass zwar, ähnlich den Varianzanalysen ohne Messwiederholungen, die Varianzhomogenität bei gleichen Zellenbestimmungszahlen  $n_i$  sich nicht gravierend auswirkt, jedoch bei ungleichen  $n_i$  der Fehler 1. Art für die Tests des Messwiederholungsfaktors nicht mehr unter Kontrolle gehalten werden kann. Die Fehlerraten

steigen dann, je nach Grad der Varianzheterogenität, deutlich über 10%, insbesondere für die Interaktion und bei negativem pairing (vgl. 4.1), sowie bei nichtnormalen Verteilungen. Und dieser Fehler tritt sogar noch stärker bei den multivariaten Tests zutage, die keine Sphärität voraussetzen.

Lüpsen (2020a) hat eigene Simulationen zu den Auswirkungen von heterogenen Kovarianzmatrizen durchgeführt. Hierbei wurde besonders berücksichtigt, dass eine Kovarianz ein Produkt aus Korrelation und zwei Varianzen ist, d.h. dass heterogene Kovarianzmatrizen sowohl durch heterogene Korrelationsmatrizen als auch durch heterogene Varianzen verursacht werden können. Dabei haben sich zwar die o.a. Ergebnisse von Algina (1994) und Keselman et al. (1993) bestätigt, aber es zeigte sich auch, dass zum einen der Fehler 1. Art und Teststärke bei multivariaten Tests auch bei gleichen  $n_i$  betroffen ist, und zum anderen dass sich das Verhältnis von  $n_i$  zu den Korrelationen  $r_i$  der Messwiederholungen innerhalb der Gruppen erheblich auswirkt. Im Detail die Auswirkungen auf die Interaktion, zunächst beim parametrischen F-Test:

- Heterogene Varianzen beim Gruppierungsfaktor, und damit heterogene Kovarianzmatrizen, wirken sich bei gleichen Stichprobenumfängen  $n_i$  nur in geringem Maße aus, im Fall ungleicher  $n_i$  jedoch stark mit Fehlerraten 0.07-0.09 (bei nominellem  $\alpha=0.05$ ).
- Bei ungleichen  $n_i$  und hinsichtlich der Streuung ungleichen Kovarianzmatrizen ist das Pairing (Korrelation zwischen  $n_i$  und Gruppenvarianzen  $s_{ij}^2, j=1, \dots, J$ ) wesentlich: Liegt kein Pairing vor, ist auch der F-Test nicht stark beeinträchtigt. Bei negativem Pairing steigt die Fehlerrate (bei nominellem  $\alpha=0.05$ ) schnell auf 0.10 ( $\text{corr}(n_i, s_{ij}^2)=-0.3$ ) bis 0.24 ( $\text{corr}(n_i, s_{ij}^2)=-0.7$ ). Bei positivem Pairing sinkt die Power des F-Tests.
- Bei ungleichen  $n_i$  und ungleichen Korrelationen  $r_i$  innerhalb der Gruppen ist ebenfalls das Pairing (Korrelation zwischen  $n_i$  und Gruppenkorrelationen  $r_i$ ) wesentlich: Liegt kein Pairing vor, ist auch der F-Test nicht beeinträchtigt. Bei positivem Pairing steigt die Fehlerrate (bei nominellem  $\alpha=0.05$ ) schnell auf 0.20 ( $\text{corr}(n_i, r_i)=0.8$ ). Bei negativem Pairing sinkt die Power des F-Tests.
- Bei ungleichen  $n_i$ , hinsichtlich der Streuung ungleichen Kovarianzmatrizen und ungleichen Korrelationen  $r_i$  innerhalb der Gruppen, wenn also die beiden zuvor genannten Fälle gleichzeitig auftreten, addieren sich die durch das Pairing (Korrelation zwischen  $n_i$  und Gruppenkorrelationen  $r_i$  sowie Korrelation zwischen  $n_i$  Gruppenvarianzen  $s_{ij}^2$ ) auftretenden Effekte. Ist z.B. die Korrelation zwischen  $n_i$  und  $s_{ij}^2$  negativ und die Korrelation zwischen  $n_i$  und Gruppenkorrelationen  $r_i$  positiv, so steigt beim F-Test die Fehlerrate auf 0.40 und höher, während im umgekehrten Fall die Power des F-Tests verschwindend gering ist.

Nun die Auswirkungen bei multivariaten Tests (Pillai, Wilks oder Hotelling-Lawley), die allerdings im Großen und Ganzen die gleichen sind wie oben beim F-Test aufgeführt, jedoch mit folgenden Besonderheiten:

- Unterschiede zwischen den Tests wirken sich nur auf die Interaktion aus: der Test von Hotelling-Lawley ist relativ liberal, der von Pillai eher konservativ, was sich dort insbesondere für kleinere  $n_i \leq 20$  zeigt.
- Gelegentlich treten auch beim Test des Haupteffekts von B erhöhte Fehlerraten 1. Art auf, mit Werten bis zu 0.09, jedoch nur für größere Designs und kleine  $n_i \leq 20$ .
- Bei negativem Pairing von  $n_i$  und Gruppenvarianzen  $s_{ij}^2 (j=1, \dots, J)$  sind die Fehlerraten 1. Art etwas höher als beim F-Test, insbesondere für den Test von Hotelling-Lawley, mit Werten bis 0.12 (bei nominellem  $\alpha=0.05$ ) für  $\text{corr}(n_i, s_{ij}^2)=-0.3$  bzw. bis 0.25 für  $\text{corr}(n_i, s_{ij}^2) = -0.7$ . Bei positivem Pairing sinkt wie beim F-Test die Power der multivariaten Tests.

- Hinsichtlich des Einflusses ungleicher Korrelationen  $r_i$  innerhalb der Gruppen ist kein Unterschied zum Verhalten des F-Tests erkennbar.

Zur Prüfung wird in der Regel der *Box-M-Test* empfohlen, doch dieser setzt, ähnlich wie der Mauchly-Test, multivariate Normalverteilung der Messwiederholungsvariablen voraus. Das ist wesentlich mehr, als für die eigentliche Varianzanalyse gefordert wird. D.h. Ergebnisse dieses Voraussetzungstests sind mit besonderer Vorsicht zu betrachten. SPSS gibt bei Messwiederholungen den Box-Test aus, und für R gibt es auch eine entsprechende Funktion, u.a. vom Autor (vgl. Anhang 3.4). Doch es gibt auch eine Reihe alternativer Tests auf Homogenität der Kovarianzmatrizen, die allerdings vielfach ebenso auf der multivariaten Normalverteilung basieren. Einige davon werden nachfolgend aufgeführt, zumal dazu vom Autor eine R-Funktion `check.covar` angeboten wird (vgl. Anhang 3):

- Box M-Test, basierend auf der multivariaten Normalverteilung,
- Schott's  $T_1$ , eine Verbesserung des M-Tests, basierend auf der multivar. Normalverteilung,
- Schott's  $T_2$ , für elliptische Verteilungen, mit Berücksichtigung des Exzesses als Maß für die Abweichung von der Normalverteilung, wobei die Verteilungen für alle Variablen gleich sein müssen,
- Schott's  $T_3$ , für elliptische Verteilungen, ebenfalls mit Berücksichtigung des Exzesses, wobei die Verteilungen für die Variablen verschieden sein dürfen,
- multivariater Levene-Test, der lediglich ordinales Skalenniveau voraussetzt, sowie ein
- multivariater Dispersionstest.

Bis auf die letzten beiden sind alle Tests sowie einige andere in Hallin & Paindaveine (2008) beschrieben. Anzumerken ist noch, dass es sich bei elliptischen Verteilungen um einen größeren Bereich von Verteilungen handelt, die die Normalverteilung als Spezialfall enthalten. Die entsprechenden Verfahren haben also einen größeren Anwendungsbereich. Einen umfassenden Vergleich der Methoden mit praktischen Tipps zur Vorgehensweise gibt Lüpsen (2020a), wo auch Verweise auf die wenigen anderen Vergleiche zu finden sind. Eine generelle Empfehlung kann nicht ausgesprochen werden. Dennoch erscheint die Anwendung des Box-Test sinnvoll bei symmetrischen Verteilungen und Schotts  $T_2$  bei rechtsschiefen Verteilungen. Da aber i.d.Regel wenig über die zugrunde liegende Verteilung bekannt ist, sollte ein robuster Test vorgezogen werden: der Levene-Test oder der Dispersionstest. Beide können allerdings nicht ungleiche Korrelationsmatrizen erkennen. Doch im Fall von ungleichen Korrelationskoeffizienten  $r_i$  für die Kovarianzmatrizen der einzelnen Gruppen verfügen auch die anderen Methoden nur über eine geringe Teststärke, d.h. im Fall eines nicht-signifikanten Ergebnisses des Homogenitätstests ist es nicht klar, ob wegen Gleichheit der Kovarianzmatrizen oder wegen Ungleichheit der Korrelationen. Daher wird dringend geraten bei einem nicht-signifikanten Ergebnis noch zusätzlich einen Test auf Homogenität der Korrelationsmatrizen durchzuführen. Hierfür bietet der Autor eine R-Funktion `check.corr` mit 4 verschiedenen Tests an (vgl. Anhang 3.5), wobei der von Larntz & Perlman (1991) der empfehlenswerteste ist.

Insbesondere bei nichtparametrischen Tests, die einen  $\chi^2$ - anstatt eines F-Tests benutzen, ist nicht unbedingt die Sphärität erforderlich. Es genügt auch die Gleichheit der Varianzen der Messwiederholungsvariablen, wobei die meisten, wie das KWF-, van der Waerden und Koch-Verfahren sogar relativ robust gegen solche Heterogenitäten sind. Ähnliches scheint auch für die nichtparametrischen Tests zu gelten, die auf multivariaten Rängen basieren (vgl. Abschnitte 2.16 und 5.3.10). Hierzu war in Kapitel 5.2 ein Levene-like-Test (vgl. R.R. Wilcox, 1989) vorgestellt worden. Vgl. dazu auch das umfangreiche Beispiel in Kapitel 11.5.

Es gibt aber auch noch einen “Notbehelf“ zur Überprüfung der Homogenität der Kovarianzmatrizen: statt dessen wird die Homogenität der Fehler- oder Residuen-Varianzen geprüft. Hierbei werden allerdings ungleiche Korrelationsmatrizen nicht erkannt. Man kann sich das folgendermaßen vorstellen: Es wird für jede Gruppe des/der Gruppierungsfaktoren eine Varianzanalyse für den/die Messwiederholungsfaktoren gerechnet. Dann hat jeder Test eines Effektes (der Messwiederholungen) einen „eigenen“ Error-Term. Jeder dieser Fehlerterme muss nun über die Gruppen hinweg statistisch gleich sein. Das zu überprüfen ist ein mühseliges Unterfangen, zumal diese Fehlerterme von den Programmen nicht gruppenweise ausgewiesen werden. Es sei denn, man rechnet wirklich für jede Gruppe eine Varianzanalyse und vergleicht die Ergebnisse. Ähnlich wie bei der Analyse der Residuen kann man sich damit behelfen, dass für jede Messwiederholungsvariable ein Test auf Homogenität der Varianzen durchgeführt wird, z.B. mit dem schon mehrfach erwähnten Levene-Test. So macht es auch SPSS. Die damit geprüfte Homogenitätseigenschaft ist zwar notwendig, aber nicht hinreichend. D.h. statistisch gleiche Kovarianzmatrizen implizieren die o.a. Varianzhomogenität, aber nicht umgekehrt.

### Alternativen - heterogene Kovarianzmatrizen

Bezüglich des Problems heterogener Kovarianzmatrizen gibt es vergleichsweise wenig Alternativen, insbesondere kaum Standard-Verfahren in R bzw. SPSS. Zu erwähnen sind hier:

- Die in Kapitel 2.13 erwähnte Analyse für heterogene Varianzen von Welch & James. Hierfür wird vom Autor eine R-Funktion bereitgestellt (siehe Anhang 3) und am Ende dieses Kapitels in einem Beispiel vorgestellt. Doch Vorsicht: Keselmann et al. (1993), sowie eigene Simulationen haben gezeigt, dass für diesen Test etwa  $\min(n_i) > 5 \cdot J$  sein muss, insbesondere für den Test der Interaktion, um zuverlässige Ergebnisse zu erhalten. Algina (1994) hat auf Basis eigener Untersuchungen diese Bedingung noch weiter verschärft.
- Darüber hinaus ist noch eine multivariate Version des in 4.3.3 vorgestellten Tests von Brown & Forsythe anzuführen (siehe z.B. Algina, 1994, und Vallejo et al., 2004), wofür in R eine Funktion zur Verfügung steht (siehe Anhang 3).
- Algina (1994) favorisiert modifizierte F-Tests zur Kompensierung solcher Kovarianzheterogenitäten, wie etwa in Kapitel 4.2.2 oder 4.3.3 vorgestellt. Huynh (Huynh, 1987, sowie Algina, 1994) hat für diesen Fall eine *general approximate procedure* (GA) sowie eine *improved general approximate procedure* (IGA) entwickelt, die zusätzlich Abweichungen von der Sphärität kompensieren kann. Sie hat gegenüber dem o.a. Verfahren von Welch & James den Vorteile, keine Bedingungen an die  $n_i$  zu stellen. Eine R-Funktion steht zur Verfügung (siehe Anhang 3)..
- Lüpsen (2020a) hat gezeigt, dass der verallgemeinerte Kruskal-Wallis-Friedman-Test (KWF) das  $\alpha$ -Risiko sowohl bei heterogenen Kovarianzmatrizen als auch bei heterogenen Korrelationsmatrizen unter Kontrolle hält. Das Verfahren verlangt keine Sphärität, lediglich gleiche Varianzen, ist allerdings gegen ungleiche Varianzen relativ robust, was ein weiterer Vorteil ist. Lediglich die Teststärke kommt bei kleinen  $n_i$  und bei symmetrischen Verteilungen nicht an die des F-Tests heran und liegt ca. 20% darunter.

Es sei hier ausdrücklich darauf aufmerksam gemacht, dass die in Kapitel 5 erwähnten Adjustierungen der Freiheitsgrade von Huynh & Feldt bzw. Geisser & Greenhouse zwar zur Kompensierung von Nicht-Sphärität, jedoch nicht zur Abmilderung der Heterogenität der Kovarianzmatrizen verwendet werden können. Dies zeigte sich deutlich in Simulationsstudien (z.B. Lüpsen, 2020a).

Wie schon erwähnt, befreien nichtparametrische Verfahren in der Regel nicht von der Überprüfung der Homogenitätsvoraussetzung, da die Rangtransformationen in der Regel Hetero-

genitäten bestenfalls abschwächen. Andererseits basieren fast alle Prüfverfahren auf der Normalverteilung. Daher sind besonders bei ordinalen Kriteriumsvariablen nur der Test auf Sphärität von Sirkiä bzw. der Levene-Test auf Gleichheit der Kovarianzmatrizen zu empfehlen.

## 6.2 Parametrische Varianzanalyse und Prüfung der Voraussetzungen

Auch hier soll zunächst einmal zum Vergleich die parametrische Varianzanalyse durchgeführt und die Prüfung der Voraussetzungen gezeigt werden. Das Prozedere wie auch die Ergebnisse sind zum Teil zwangsläufig mit denen aus Kapitel 5.3.1 identisch. Dieses wird noch einmal für den Fall gemischter Versuchspläne erläutert.

Zur Berechnung der Residuen gibt es folgende Möglichkeit: Der oder die Messwiederholungsfaktoren C, D,... werden als Gruppierungsfaktoren gehandhabt. Dazu muss der Datensatz umstrukturiert werden, indem die Messwiederholungen in Fälle gewandelt werden. (Dies ist in R ohnehin für Analysen mit Messwiederholungen erforderlich.) Dann wird folgendes Modell *ohne* Messwiederholungen analysiert:

$$A * C * D + V_{pn} \quad (6-2)$$

wobei  $V_{pn}$  die Fallkennung, z.B. Versuchspersonennummer, ist. Die Residuen dieses Modells sind die Residuen des Modells mit dem Gruppierungsfaktor A sowie mit Messwiederholungen auf C (und D). Dies gilt auch analog für mehrere Gruppierungsfaktoren A, B,...

Dies ist zwar prinzipiell auch bei SPSS möglich, verursacht aber wegen der erforderlichen Umstrukturierung etwas Aufwand. SPSS gibt allerdings für jede Messwiederholungsvariable  $x_j$  andere Residuen aus:  $e'_{ijm} = x_{ijm} - \alpha\gamma_{ij} - \alpha_i - \gamma_j$ . Aus dem Modell 6-1 ergibt sich für diese  $e'_{ijm} = \pi_m + e_{ijm}$ , d.h. um die Residuen  $e_{ijm}$  zu erhalten, müssen von den  $e'_{ijm}$  die  $\pi_m$  subtrahiert werden. Die Subtraktion von  $\bar{p}$  von  $p_m$  zur Ermittlung von  $\pi_m$  kann entfallen, da sie für die Beurteilung der Residuen  $e_{ijm}$  ohne Bedeutung ist. Die erforderlichen Schritte sind dann:

- Speichern der Residuen:  $e'_{ijm}$ ,
- Ermitteln des Personeneffekts  $\pi_m$  aus

$$p_m = \left( \sum_j x_{jm} \right) / J \quad \text{und}$$

$a_i$  = Mittelwert der  $p_m$  für Gruppe  $i$  :

$$\pi_m = (p_m - \bar{p} - a_i),$$

- und schließlich  $e_{ijm} = e'_{ijm} - \pi_m$ .

Wie bei dieser Art der Residuen-Ermittlung diese gehandhabt und beurteilt werden können, wurde bereits in Kapitel 5.3.1 erläutert. Wie man sieht, ist dieses Verfahren relativ aufwändig, insbesondere wenn das Design mehrere Gruppierungsfaktoren enthält. Insofern empfiehlt es sich, das oben skizzierte Verfahren 6-2 anzuwenden.

Wenn man in den nachfolgenden Beispielrechnungen das Ergebnis des Mauchly-Tests hier mit dem aus 5.3.1 vergleicht, mögen die unterschiedliche Ergebnisse irritieren, da ja eigentlich die Gruppenstruktur nicht in den Test einfließen sollte. Tut sie aber doch. Denn hier werden im Gegensatz zum Modell ohne Gruppierungsfaktoren *gepoolte* Kovarianzmatrizen errechnet. D.h. die Berechnung erfolgt quasi gruppenweise, bevor die Matrizen zusammengefasst werden. Der Unterschied kann u.a. durch die verschiedenen Gruppenmittelwerte verursacht werden. Hierher rührt auch die in 6.1 erwähnte Voraussetzung der Homogenität der Kovarianzmatrizen.

**mit R:**

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer518t`. Zunächst wird die Anova zum Vergleich mit der Standardfunktion `aov` durchgeführt, wenn das auch i.a. nicht sinnvoll ist, weil die Funktion `ezANOVA` zugleich den Mauchly-Test durchführt (siehe unten). Dabei werden durch den Modellterm `Error (Vpn/Zeit)` die Messwiederholungen auf dem Faktor `Zeit` gekennzeichnet:

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

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

Zunächst einmal zur Prüfung der Residuen  $\varepsilon_{ijm}$  auf Normalverteilung. Diese lassen sich, wie oben erläutert, bequem als Residuen eines Anova-Modells ohne Messwiederholungen ermitteln:

```
aov2<-aov(score~Geschlecht*Zeit+Vpn, winer518t)
res<-aov2$residuals
hist(res)
shapiro.test(res)
```

Da die Ergebnisse denen aus Kapitel 5.3.1 weitgehend ähnlich sind, wird auf die Wiedergabe hier verzichtet. Die Überprüfung der Normalverteilung der versuchspersonenspezifischen Abweichungen  $\pi_m$  ist dieselbe wie in Kapitel 5.3.1. Eine Überprüfung auf multivariate Normalverteilung der Residuen war in Kapitel 5.3.9 gezeigt worden.

Die Überprüfung der Spharizität kann z.B. mittels des Mauchly-Tests und der Funktion `ezANOVA` des Pakets `ez` vorgenommen:

```
library(ez)
ezANOVA(winer518t, score, Vpn, within=Zeit, between=Geschlecht)
```

\$ANOVA							
	Effect	DFn	DFd	F	p	p<.05	ges
2	Geschlecht	1	8	0.4716981	5.116202e-01		0.04118616
3	Zeit	2	16	22.0506329	2.522847e-05	*	0.42800983
4	Geschlecht:Zeit	2	16	17.0379747	1.086241e-04	*	0.36635819
\$`Mauchly's Test for Sphericity`							
	Effect	W	p	p<.05			
3	Zeit	0.9306201	0.7775055				
4	Geschlecht:Zeit	0.9306201	0.7775055				

\$`Sphericity Corrections`					
	Effect	GGe	p [GG]	HFe	p [HF]
3	Zeit	0.9351214	4.280809e-05	1.209851	2.522847e-05
4	Geschlecht:Zeit	0.9351214	1.683544e-04	1.209851	1.086241e-04

Tabelle 6-2

Der Aufbau der Tabelle 6-2 wurde bereits kurz in 5.3.1 erläutert. Die Anova-Tabelle ist natürlich mit der in Tabelle 6-1 identisch. Da der Mauchly-Test keine Signifikanz zeigt, werden die Ergebnisse aus der ersten Tabelle (ANOVA) verwendet.

Die in 5.3.1 vorgestellte Funktion `check.sphere` bietet zum einen alternative Tests auf Sphärität, aber auch einen multisample Sphäritäts-Test nach Mauchly, der nicht die Homogenität der Kovarianzmatrizen verlangt. Hier werden wieder beide Aufrufvarianten gezeigt:

```
check.sphere(winer518[,3:5],winer518[,2])           # wide Format
with(winer518t,check.sphere(score,                 # long Format
  groups=Geschlecht,trial=Zeit,id=Vpn))
```

Mauchly test	0.5752	2	0.7500
multisample Mauchly	23.3189	11	0.0159
Likelihood Ratio	0.6471	5	0.9857
John's test	27.4049	2	0.0000
John's test (using F)	2.3054	5	0.0761
John-Nagao	27.4049	5	0.0000
Muirhead & Waternaud	0.3297	5	0.9970
compound symmetry	0.5659	4	0.9668

Der multisample Mauchly ist auch ohne Gleichheit der Kovarianzmatrizen gültig, während die übrigen Tests die Homogenität voraussetzen, die als nächstes geprüft wird.

Für den Test auf Homogenität der Kovarianzmatrizen gibt es zwar eine Funktion `boxM` für den Box M-Test im Paket `biotools`. Dennoch sei auf die Funktion `check.covar` im Anhang 3 verwiesen, die neben dem Box-Test auch die anderen im vorigen Abschnitt aufgeführten Tests durchführt. Diese kann wahlweise auf den Datensatz im wide-Format (`winer518`) oder im long-Format (`winer518t`) angewandt werden:

```
check.covar(winer518[,3:5],winer518$Geschlecht)   # wide Format
with(winer518t,check.covar(score,                 # long Format
  groups=Geschlecht,trial=Zeit,id=Vpn))
```

	Chisquare	df	p value
LR	7.5870	6	0.2699
Box M	6.6386	6	0.3556
Schott T1	3.8944	6	0.6910
Schott T2	7.0812	6	0.3134
Schott T3	12.1771	6	0.0581
Levene	0.4722	6	0.8007

Der empfehlenswerteste Test ist der Levene-Test. Der p-Wert 0.8007 indiziert die Homogenität der Kovarianzmatrizen, die, wie vorher erwähnt, auch Voraussetzung der Varianzanalyse ist. Wie im vorigen Abschnitt erwähnt ist es angebracht, zusätzlich die Homogenität der Korrelationsmatrizen zu überprüfen, wenn der vorherige Test nicht signifikant war. Dazu gibt es die Funktion `check.corr` (vgl. Anhang 3):



```
check.corr(winer518[,3:5],winer518[,2])
```

	statistic	df	p value
Jennrich test (chisquare)	3.5653	3	0.3124
Levene correlation (F)	0.3019	3	0.8234
Larntz & Perlman (chisq)	2.5860	3	0.2898
Box M (chisq)	4.4388	3	0.2178

Da auch dieser Test negativ ausfällt, wobei das Ergebnis des Tests von Larntz & Perlman das zuverlässigste ist, kann die Homogenität der Kovarianzmatrizen angenommen werden.

Alternativ die Überprüfung der Gleichheit der Fehlervarianzen: Hier werden der Einfachheit halber für die drei Messwiederholungsvariablen (Variablenindizes 3,4,5) jeweils die Gruppenvarianzen mit dem Levene-Test überprüft. Auch hier wird der ursprüngliche Dataframe `winer518` benutzt. In diesem Fall liegt nur ein Gruppierungsfaktor vor. Somit lassen sich alle Variablen mittels `apply` in einem Funktionsaufruf überprüfen:

```
library(car)
apply(winer518[,3:5], 2 ,leveneTest,win_518$Geschlecht)
```

```
$t1
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1      0.1 0.7599
      8

$t2
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1      0      1
      8

$t3
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1     0.05 0.8287
      8
```

Da keines der Ergebnisse signifikant ist, kann die Varianzhomogenität angenommen werden.

Auch die bereits in Kapitel 4.3.2 vorgestellte Funktion `aov_ez` aus dem Paket `afex` kann bequem Versuchspläne mit Messwiederholungen analysieren, die wie `aov` und `ezANOVA` den transformierten Dataframe, hier `winer518t`, verlangt. Aus dem folgenden Beispiel lässt sich die Syntax gut ablesen. Standardmäßig wird für die Tests, die Messwiederholungsfaktoren enthalten, die Greenhouse-Geisser-Adjustierung der Freiheitsgrade vorgenommen. Über die Zusatzfunktion `nice` kann alternativ `HN` (für die Huynh-Feldt-Adjustierung) oder `none` (für keine Adjustierung) angefordert werden. Somit ist es sinnvoll, das primäre Ergebnis der Funktion zu speichern (hier: `aov1`), um später die Zusatzfunktionen anwenden zu können :

```
library(afex)
aov1 <- aov_ez(data=winer518t, dv="score",
              between="Geschlecht", within="Zeit", id="Vpn")
nice(aov1, correction="HF")
```

```
Anova Table (Type 3 tests)

Response: score
      Effect          df    MSE          F    ges p.value
1    Geschlecht         1, 8  7.07         0.47  .041    .512
2          Zeit  1.87, 14.96  1.41  22.05 ***  .428    <.001
3 Geschlecht:Zeit  1.87, 14.96  1.41  17.04 ***  .366    <.001
Sphericity correction method: HF
```

Auch für Designs mit Messwiederholungen gibt es eine an aov angelehnte Syntax:

```
aov_car(score~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t)
```

Anzumerken ist, dass darin „ges“ wie oben die *generalized effect size* (Effektgröße  $\eta^2$ ) ist. Darüber hinaus gibt es die Zusatzfunktionen `test_levene` für den Test auf Homogenität der Gruppen-Varianzen sowie `test_sphericity` für den Mauchly-Test auf Sphärität:

```
test_levene(aov1)
test_sphericity(aov1)
```

```
Levene's Test for Homogeneity of Variance (center = center)
      Df F value Pr(>F)
group  1  0.0024 0.9624
      8
```

```
          Test statistic p-value
Zeit          0.93062 0.77751
Geschlecht:Zeit  0.93062 0.77751
Warnmeldung:
In summary.Anova.mlm(object$Anova, multivariate = FALSE) :
  HF eps > 1 treated as 1
```

Eine Warnung: `residuals(...)`, angewandt auf ein Ergebnisobjekt, gibt bei gemischten Designs nicht die korrekten Residuen aus.

### mit SPSS:

Varianzanalysen mit Messwiederholungen erhält man in SPSS über das Menü „Allgemeines lineares Modell -> Messwiederholung“. Die Syntax für den Beispieldatensatz 4 (winer518) mit Ausgabe der Homogenitätstests lautet:

```
GLM t1 t2 t3 by Geschlecht
  /wsfactor=Zeit 3 polynomial
  /print homogeneity
  /wsdesign=Zeit
  /design=Geschlecht.
```

mit folgender Ausgabe des Mauchly-Tests, der Anova-Tabelle für die Messwiederholungseffekte (Innersubjekteffekte) und der Anova-Tabelle für den Gruppierungsfaktor (Zwischensubjekteffekte), wobei der Mauchly-Test keine Inhomogenitäten zeigt, so dass die Ergebnisse der Zeile „Sphärität angenommen“ verwendet werden können:

Mauchly-Test auf Sphärizität <sup>a</sup>							
Innersubjekt- effekt	Mauchly-W	Approx. Chi-Quadrat	df	Sig.	Epsilon <sup>b</sup>		
					Greenhouse-Geisser	Huynh-Feldt	
Zeit	,931	,503	2	,778	,935	1,000	

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	546,133	1	546,133	77,283	,000
Geschlecht	3,333	1	3,333	,472	,512
Fehler	56,533	8	7,067		

Tests der Innersubjekteffekte						
Quelle		Quadrat- summe vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	58,067	2	29,033	22,051	,000
	Greenhouse-Geisser	58,067	1,870	31,048	22,051	,000
	Huynh-Feldt	58,067	2,000	29,033	22,051	,000
Zeit * Geschlecht	Sphärizität angen.	44,867	2	22,433	17,038	,000
	Greenhouse-Geisser	44,867	1,870	23,990	17,038	,000
	Huynh-Feldt	44,867	2,000	22,433	17,038	,000
Fehler(Zeit)	Sphärizität angen.	21,067	16	1,317		
	Greenhouse-Geisser	21,067	14,962	1,408		
	Huynh-Feldt	21,067	16,000	1,317		

Tabelle 6-3

Darüberhinaus werden über den Zusatz `/print homogeneity` der Box-M-Test zur Überprüfung der Gleichheit der Kovarianzmatrizen für die beiden Gruppen sowie für alle 3 Variablen ein Levene-Test auf Gleichheit der Zellenvarianzen ausgegeben:

Box-Test auf Gleichheit der Kovarianzmatrizen	
Box-M-Test	7,587
F	,734
df1	6
df2	463,698
Sig.	,622

Der Box-Test zeigt keine Ungleichheit der Varianzen, so dass eine Voraussetzung für die Durchführung des Mauchly-Tests gegeben ist, wenn ihm auch nicht allzu viel Bedeutung beigemessen werden sollte.

Levene-Test auf Gleichheit der Fehlervarianzen <sup>a</sup>				
	F	df1	df2	Sig.
t1	,159	1	8	,700
t2	,000	1	8	1,000
t3	,015	1	8	,905

Da alle drei Tests nicht signifikant sind, kann auch die Homogenität der Fehlervarianzen angenommen werden.

Im vorigen Abschnitt war darauf hingewiesen worden, dass der Box-Test nicht empfehlenswert ist. Dort war statt dessen der Levene-Test auf Gleichheit der Kovarianzmatrizen empfohlen worden. SPSS bietet diesen nicht an, aber er kann mit etwas Aufwand durchgeführt werden. Wie unten zu sehen ist das lediglich für eine kleinere Anzahl von Messwiederholungen praktikabel. Entsprechend der Formel im Anhang 2.6. sind folgende Schritte erforderlich:

- Berechnung der Mediane  $m_1, \dots, m_3$  für die Variablen  $t_1, \dots, t_3$  getrennt für die Gruppenvariable Geschlecht mittels aggregate,
- Berechnung von  $(y_i - \text{median}_j)(y_j - \text{median}_j)$  für  $i=1, \dots, 3$  und  $j=i, \dots, 3$ , wodurch  $3 \cdot 4/2 = 6$  neue Variablen  $c_1, \dots, c_6$  erzeugt werden.
- Transformation von  $c$ :  $c = \text{sign}(c) \cdot \sqrt{\text{abs}(c)}$
- Anwendung einer multivariaten Varianzanalyse (manova) auf die  $c$ -Variablen, z.B. der Wilks-Lambda.

Die Schwierigkeit liegt bei SPSS darin, dass keine geschachtelten Schleifen (do repeat) möglich sind, um den o.a. 2. Schritt elegant durchzuführen. D.h. eine Schleifenanweisung muss entsprechend der Variablenzahl (hier 3) wiederholt werden. Die Anweisungen:

```
aggregate
  /Outfile=* Mode=Addvariables
  /Break=Geschlecht
  /m1=median(t1)
  /m2=median(t2)
  /m3=median(t3) .

do repeat v=t1 to t3 / m=m1 to m3 /c=c1 to c3.
compute c=(v-m)*(t1-m1) .
end repeat.

do repeat v=t2 to t3 / m=m2 to m3 /c=c4 to c5.
compute c=(v-m)*(t2-m2) .
end repeat.

compute c6=(t3-m3)*(t3-m3) .
execute.

do repeat c=c1 to c6.
compute vorz=1.
if (c lt 0) vorz=-1.
compute c=vorz*sqrt(vorz*c) .
end repeat.
execute.

GLM c1 to c6 BY Geschlecht
  /DESIGN= Geschlecht.
```

mit folgender Ausgabe der Varianzanalyse:

Multivariate Tests

Effekt		Wert	F	Hypothese df	Fehler df	Sig.
Konstanter Term	Pillai-Spur	,875	3,500	6,000	3,000	,166
	Wilks-Lambda	,125	3,500	6,000	3,000	,166
	Hotelling-Spur	6,999	3,500	6,000	3,000	,166
	Größte charakteristische Wurzel nach Roy	6,999	3,500	6,000	3,000	,166
	Pillai-Spur	,486	,472	6,000	3,000	,801
Geschlecht	Wilks-Lambda	,514	,472	6,000	3,000	,801
	Hotelling-Spur	,944	,472	6,000	3,000	,801
	Größte charakteristische Wurzel nach Roy	,944	,472	6,000	3,000	,801
	Pillai-Spur	,486	,472	6,000	3,000	,801
	Wilks-Lambda	,514	,472	6,000	3,000	,801

Das Ergebnis für den Levene-Test ist in der letzten Spalte und der Rubrik Geschlecht abzulesen:  $p=.801$  (egal für welchen Test), d.h. Gleichheit der Kovarianzmatrizen ist gegeben. Alternativ kann die Berechnung der Kreuzprodukte auch in einer DO REPEAT-Schleife anstatt drei erfolgen:

```
do repeat v=t1 t2 t3 t2 t3 t3 /
          m=m1 m2 m3 m2 m3 m3 /
          w=t1 t1 t1 t2 t2 t3 /
          n= m1 m1 m1 m2 m2 m3 /
          c=c1 to c6.
compute c=(v-m)*(w-n).
end repeat.
```

Bleibt noch die Überprüfung der Residuen auf Normalverteilung. Dazu wird das am Eingang dieses Kapitels genannte Modell ohne Messwiederholungen 6-2 gerechnet. Zunächst muss der Datensatz umstrukturiert werden, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben. Die Syntax hierfür lautet:

```
Varstocases
  /id=Vpn
  /make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.
```

Die ersten Fälle des umstrukturierten Datensatzes sehen etwa folgendermaßen aus:

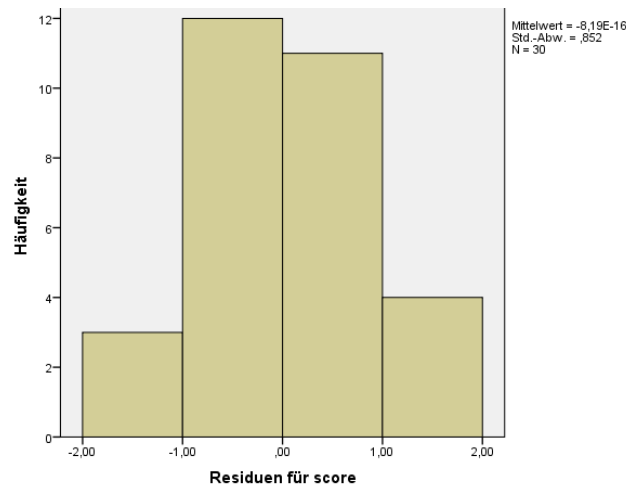
	Vpn	Geschlecht	Zeit	score
1	1	1	1	4
2	1	1	2	7
3	1	1	3	2
4	2	1	1	3
5	2	1	2	5
6	2	1	3	1
7	3	1	1	7
8	3	1	2	9

Für diesen Datensatz wird nun eine Varianzanalyse mit den Faktoren Vpn, Geschlecht und

Zeit gerechnet, wobei das Modell angepasst werden muss: Anstatt des gesättigten Modells sind neben den Haupteffekten die Interaktion `Geschlecht*Zeit` auszuwählen. Ferner müssen die Residuen gespeichert werden, die anschließend den Namen `RES_1` haben. Schließlich werden diese dann in `Examine` (Explorative Datenanalyse) mittels Shapiro-Test und Histogramm auf Normalverteilung überprüft. Die Anweisungen hierfür:

```
Unianova score BY Geschlecht Zeit Vpn
  /save=resid
  /design=Geschlecht Zeit Geschlecht*Zeit Vpn.

Examine variables=RES_1
  /plot histogram.
```



Das automatisch erzeugte Histogramm basiert zunächst auf 11 Intervallen, was bei einem  $n$  von 30 keinen Sinn macht. Möglich wären hier 4, 5 oder 6 Intervalle (vgl. Kapitel 1.6), so dass eine Nachbereitung mit dem Grafikeditor erforderlich ist und o.a. Abbildung erzeugt.

Tests auf Normalverteilung						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
RES_1	,126	30	,200*	,959	30	,288

Auf Basis dieser Ergebnisse kann man die Residuen als normalverteilt annehmen.

### 6.3 Rank transform-Tests (RT) und normal scores-Tests (INT)

Bei dem *Rank transform Test* werden die Werte der abhängigen Variablen über alle Messwiederholungen und Gruppen hinweg in Ränge gewandelt, um mit diesen dann eine „normale“ parametrische Varianzanalyse zu rechnen.

Bei dem *normal score- bzw. inverse normal transform-Verfahren* (INT) werden lediglich die Werte der abhängigen Variablen  $x$  über alle Messwiederholungen und Gruppen hinweg zunächst wie beim RT-Verfahren in Ränge  $R(x)$  gewandelt und anschließend in normal scores umgerechnet:

$$nscore_m = \Phi^{-1}(R(x_m)/(M+1))$$

wobei  $M$  die Anzahl aller Werte ist, also das Produkt aus Messwiederholungen und Anzahl der Merkmalsträger  $n$ , also z.B.  $n*J$ . Mit diesen scores wird dann eine „normale“ parametrische Varianzanalyse gerechnet. In beiden Fällen sollte man den Mauchly-Test durchführen, oder di-

rekt die korrigierten F-Tests von Huynh & Feldt benutzen, auch wenn die Spherizität gegeben ist. Beide Verfahren soll wieder am Beispieldatensatz 4 demonstriert werden. Die Ergebnisse zeigen, dass dieses Verfahren „besser“ abschneidet als das einfachere Rank transform (RT).

### mit R:

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer518t`. Für die RT-Methode wird zunächst die Variable `score` in Ränge umgerechnet, anschließend die Anova mit der `ezANOVA` durchgeführt, um die Ergebnisse des Mauchly-Tests zu erhalten:

```
winer518t <- within(winer518t, Rscore<-rank(score))
ezANOVA(winer518t, Rscore, Vpn, within=Zeit, between=Geschlecht, detailed=T)
```

\$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
\$`Mauchly's Test for Sphericity`							
	Effect	W	p	p < .05			
3	Zeit	0.9861432	0.9523355				
4	Geschlecht:Zeit	0.9861432	0.9523355				
\$`Sphericity Corrections`							
	Effect	GGe	p [GG]	p [GG] < .05	HFe	p [HF]	
3	Zeit	0.9863326	2.602008e-05	*	1.306878	2.325487e-05	
4	Geschlecht:Zeit	0.9863326	1.640879e-04	*	1.306878	1.502651e-04	

Tabelle 6-3

Obwohl der Mauchly-Test keine Signifikanzen zeigt, wird empfohlen, die korrigierten F-Tests von Hynh-Feldt zu benutzen. Dessen Ergebnisse weichen nicht nennenswert von denen der o.a. parametrischen Analyse (Tabelle 6-2) ab. Die Voraussetzung der Normalverteilung braucht hier nicht geprüft werden.

Die Berechnung der normal scores erfolgt genauso wie Kapitel 5.3.4. Die Varianzanalyse enthält allerdings hier zusätzlich den Test des Faktor `Geschlecht`. Die Analyse wird mit `ezANOVA` (Paket `ez`) durchgeführt. Die Varianzhomogenität (Spherizität) war schon in Kapitel 5.3.4 bestätigt worden:

```
library(ez)
ezANOVA(winer518t, nscore, Vpn, within=Zeit, between=Geschlecht)
```

	Effect	DFn	DFd	F	p	p < .05	ges
2	Geschlecht	1	8	0.4589120	5.172406e-01		0.04306605538
3	Zeit	2	16	26.1823940	9.001193e-06	*	0.41354670288
4	Geschlecht:Zeit	2	16	19.4945215	5.137485e-05	*	0.34428051545

### mit SPSS:

Ausgangspunkt ist hier der im Kapitel 5.3.3 umstrukturierte Datensatz. Für die RT-Methode wird zunächst die Variable `score` in Ränge gewandelt und erhält den Namen `Rscore`, bevor der Datensatz dann wieder in die Ausgangsform zurücktransformiert wird (vgl. Anhang 1.2). Dabei wird `Rscore` für die 3 Zeitstufen zu `Rscore.1`, `Rscore.2`, `Rscore.3`,

Schließlich wird dann für diese Variablen wie im vorigen Kapitel die parametrische Varianzanalyse durchgeführt.

```
Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.
```

```
Rank variables=score (A)
  /Rank into Rscore.
```

```
Sort cases by Vpn Zeit.
```

```
Casestovars
  /Id=Vpn
  /Index=Zeit
  /Groupby=variable.
```

Hier der Datensatz nach der erneuten Umstrukturierung:

	Vpn	Geschlecht	score.1	score.2	score.3	Rscore.1	Rscore.2	Rscore.3
1	1	1	4	7	2	14,000	26,000	8,500
2	2	1	3	5	1	11,500	18,000	3,500
3	3	1	7	9	6	26,000	29,500	22,500
4	4	1	6	6	2	22,500	22,500	8,500
5	5	1	5	5	1	18,000	18,000	3,500
6	6	2	8	2	5	28,000	8,500	18,000
7	7	2	4	1	1	14,000	3,500	3,500
8	8	2	6	3	4	22,500	11,500	14,000
9	9	2	9	5	2	29,500	18,000	8,500
10	10	2	7	1	1	26,000	3,500	3,500

```
GLM Rscore.1 Rscore.2 Rscore.3 by Geschlecht
  /wsfactor=Zeit 3 polynomial
  /wsdesign=Zeit
  /design=Geschlecht.
```

Nachfolgend zunächst der Test auf Sphärität (Varianzhomogenität), danach die Ergebnisse der Varianzanalyse für den Effekt des Gruppierungsfaktors und zuletzt die Effekte des Messwiederholungsfaktors (Innersubjekteffekte). Bei diesen wird empfohlen, die Resultate aus der Zeile „Huynh-Feldt“ abzulesen, obwohl der entsprechende Mauchly-Test keine Signifikanzen aufweist.

Mauchly-Test auf Sphärität <sup>a</sup>						
Innersubjekteffekt	Mauchly W	Approx. Chi-Quadrat	df	Sig.	Epsilon <sup>b</sup>	
					Greenhouse-Geisser	Huynh-Feldt
Zeit	,986	,098	2	,952	,986	1,000



Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	7207,500	1	7207,500	82,156	,000
Geschlecht	53,333	1	53,333	,608	,458
Fehler	701,833	8	87,729		

Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	698,600	2	349,300	22,358	,000
	Greenhouse-Geisser	698,600	1,973	354,140	22,358	,000
	Huynh-Feldt	698,600	2,000	349,300	22,358	,000
	Untergrenze	698,600	1,000	698,600	22,358	,001
Zeit * Geschlecht	Sphärizität angen.	501,267	2	250,633	16,043	,000
	Greenhouse-Geisser	501,267	1,973	254,106	16,043	,000
	Huynh-Feldt	501,267	2,000	250,633	16,043	,000
	Untergrenze	501,267	1,000	501,267	16,043	,004
Fehler(Zeit)	Sphärizität angen.	249,967	16	15,623		
	Greenhouse-Geisser	249,967	15,781	15,839		
	Huynh-Feldt	249,967	16,000	15,623		
	Untergrenze	249,967	8,000	31,246		

Tabelle 6-4

Die Ergebnisse weichen nicht nennenswert von denen der o.a. parametrischen Analyse ab. Weitere Voraussetzungen brauchen hier nicht geprüft werden.

Die Durchführung der INT-Methode erfolgt im Wesentlichen genauso wie oben bei der RT-Methode. Lediglich die Ermittlung der Anzahl der Werte  $n_c$  und die Transformation der Ränge in normal scores kommt hinzu. Die Schritte im Einzelnen:

- Zunächst muss der Datensatz umstrukturiert werden, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben. Dabei wird die abhängige Variable `score` gebildet.
- Über Aggregate wird die Anzahl der Werte  $n_c$  ermittelt.
- Die Werte werden in Ränge umgerechnet.
- Über die inverse Normalverteilung (`Idf.normal`) werden die Ränge in normal scores umgerechnet.
- Der Datensatz wird zurück in die ursprüngliche Form transformiert. Daraus resultieren aus `n_score` die Variablen `n_score.1,...`
- Schließlich kann die parametrische Varianzanalyse auf die Variablen `n_score.1,...` angewandt werden.

Die Syntax hierfür sowie nachfolgend die Ausgabe der Anova-Tabellen. Die Varianzhomogenität (Sphärizität) war schon in Kapitel 5.3.4 bestätigt worden, so dass für die Messwiederholungseffekte nur die Zeilen „Sphärizität angenommen“ relevant sind.

```

Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.

Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(score).
Rank Variables=score / rank into Rscore.
compute nscore=Idf.normal(Rscore/(nc+1),0,1).
Sort cases by Vpn Zeit.

Casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

GLM nscore.1 nscore.2 nscore.3
  /wsfactor=Zeit 3 polynomial
  /wsdesign=Zeit
  /design=Geschlecht.

```

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Konstanter Term	,003	1	,003	,003	,955
Geschlecht	,441	1	,441	,459	,517
Fehler	7,686	8	,961		

Tests der Innersubjekteffekte						
Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	6,909	2	3,454	26,182	,000
Zeit * Geschlecht	Sphärizität angen.	5,144	2	2,572	19,495	,000
Fehler(Zeit)	Sphärizität angen.	2,111	16	,132		

## 6. 4 Puri & Sen-Test

Beim Puri & Sen-Test werden die beobachteten Werte wie beim o.a. RT-Verfahren über alle  $n$  Merkmalsträger und alle  $J$  Messwiederholungen hinweg in Ränge (Wilcoxon-Ränge) transformiert. Da die F-Tests hier nicht interessieren, ist auch eine Überprüfung der Sphärizität bei den Puri & Sen-Tests nicht erforderlich. Allerdings ist eine Überprüfung der Gleichheit der Varianzen der Messwiederholungsvariablen, wie sie in den Abschnitten 5.3.4 und 5.4.3 vorgenommen worden war, ratsam.

Da zum einen die Konstruktion der  $\chi^2$ -Tests beim Puri & Sen-Test nicht ganz trivial ist, und zum anderen diese  $\chi^2$ -Tests in gleicher Weise auch beim KWF- und van der Waerden-Verfahren ebenfalls von großer Wichtigkeit sind, wird hier etwas ausführlicher darauf eingegangen. Es sei zunächst noch einmal daran erinnert, dass bei der Analyse von Split-Plot-Designs zum einen alle Gruppierungsfaktoren und deren Interaktionen eine gemeinsame Fehlerstreuung haben, und zum anderen jeder Messwiederholungsfaktor sowie jede Interaktion von Messwiederholungsfaktoren jeweils eine eigene Fehlerstreuung haben. So hat z.B. ein Design mit nur 2 Mess-

wiederholungsfaktoren 3 Fehlerterme und eines mit 2 Gruppierungsfaktoren und 2 Messwiederholungsfaktoren 4 Fehlerterme (vgl. dazu die 2-faktorielle Varianzanalyse mit Messwiederholungen, Tabellen 5-4 und 5-6).

Folgende Schritte sind für eine Analysevariable  $x$  durchzuführen:

- Mit diesen Rängen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle werden folgende  $\chi^2$ -Tests aufgestellt:  
Für die Effekte ohne Messwiederholungsfaktoren, z.B. A, B, A\*B (vgl. Formel 2-6b):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{MS_{\text{zwischen}}}$$

und für die Effekte (Haupteffekte und Interaktionen) mit Messwiederholungsfaktoren z.B. C, D, A\*C, A\*D, B\*C, ...A\*B\*C,... (vgl. Formel 2-7):

$$\chi^2 = \frac{SS_{\text{Effekt}}}{(SS_X + SS_{\text{Fehler}}) / (df_X + df_{\text{Fehler}})}$$

wobei

- $SS_{\text{Effekt}}$  die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,
- $MS_{\text{zwischen}}$  die Varianz der gesamten Zwischensubjektstreuung (MS, Mean Square), die die Streuung aller Gruppierungsfaktoren und deren Interaktionen sowie der damit verbundenen Fehlerstreuung beinhaltet,
- $SS_{\text{Fehler}}$  die Streuungsquadratsumme des zum Effekt gehörenden Fehlers ist sowie
- $SS_X$  die Streuungsquadratsummen aller Effekte, die  $SS_{\text{Fehler}}$  als Fehlerterm haben, also insbesondere der zu testende Effekt  $SS_{\text{Effekt}}$  sowie Interaktionen mit allen Gruppierungsfaktoren,
- $df_X$  die entsprechenden Freiheitsgrade.

(Der Nenner der  $\chi^2$ -Tests für die Messwiederholungseffekte entspricht genau  $MS_{\text{innerhalb}}$ , also der Varianz innerhalb der Versuchspersonen.)

- Die  $\chi^2$ -Werte sind dann in den Tabellen für den  $\chi^2$ -Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade ( $df_{\text{Effekt}}$ ) des entsprechenden F-Tests sind.
- Die  $\chi^2$ -Werte für die Tests, die ausschließlich Messwiederholungsfaktoren beinhalten, also C, D, C\*D, sollten alternativ gemäß Iman & Davenport (vgl. Formel 2-1b) in F-Werte umgerechnet werden. In diesem Fall entspricht dies:

$$F = \frac{(N-1)\chi^2}{df_X + df_{\text{Fehler}} - \chi^2}$$

wobei  $df_X$  und  $df_{\text{Fehler}}$  die o.a. Freiheitsgrade sind.

Dazu ein paar Beispiele für die  $\chi^2$ -Tests, wobei letztlich nur die Nenner interessieren, da der Zähler einheitlich die Streuung des zu testenden Effekts  $SS_{\text{Effekt}}$  ist. Angenommen das Design enthält 2 Gruppierungsfaktoren A und B sowie 2 Messwiederholungsfaktoren C und D. Dann ist der Nenner

- für die Tests von A, B und AB:  $SS_A + SS_B + SS_{AB} + SS_{Fehler(A,B)}$
- für den Test von C:  $SS_C + SS_{AC} + SS_{BC} + SS_{ABC} + SS_{Fehler(C)}$
- für den Test von D:  $SS_D + SS_{AD} + SS_{BD} + SS_{ABD} + SS_{Fehler(D)}$
- für den Test von CD:  $SS_{CD} + SS_{ACD} + SS_{BCD} + SS_{ABCD} + SS_{Fehler(CD)}$

Analog summieren sich die Freiheitsgrade. Fehlt einer der Faktoren, z.B. B oder D, so entfallen oben alle Summanden, die diese Faktoren enthalten. Fehlt z.B. Faktor B, so sind  $SS_B$ ,  $SS_{AB}$ ,  $SS_{BC}$ ,  $SS_{BD}$ ,  $SS_{BCD}$ ,  $SS_{ABCD}$  gleich 0 zu setzen.

### 6.4.1 Ein Gruppierungs- und ein Messwiederholungsfaktor

Für ein 2-faktorielles Design werden die Schritte am Datensatz des Beispiels 4 (`winer518`) demonstriert.

#### mit R:

Zunächst die relativ einfache Durchführung des Puri & Sen-Tests: Dazu genügt es, die für die Berechnung der  $\chi^2$ -Werte erforderlichen Streuungsquadratsummen aus der Tabelle 6-3 (Kapitel 6.3) zu entnehmen: Die Spalten  $SS_n$  enthalten die  $SS_{Effekt}$ , die Spalten  $SS_d$  die  $SS_{Fehler}$  und natürlich  $DF_n$  und  $DF_d$  die dazugehörigen Freiheitsgrade. In diesem Fall ist es am einfachsten, die  $\chi^2$ -Werte daraus „mit der Hand“ auszurechnen:

$$MS_{zwischen} = \frac{53,3 + 701,8}{1 + 8} = 83,9$$

$$\chi_{Geschlecht}^2 = \frac{53,3}{83,9} = 0,635$$

$$\chi_{Zeit}^2 = \frac{698,6}{(698,6 + 501,3 + 249,9)/(2 + 2 + 16)} = 9,64$$

$$\chi_{Interaktion}^2 = \frac{501,3}{(698,6 + 501,3 + 249,9)/(2 + 2 + 16)} = 6,915$$

#### Ergebnisse 6-1

Die für die Tests erforderlichen Freiheitsgrade entsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse, d.h. der Spalte  $DF_n$ . Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 ( $df=1$ ) sowie bei 6,0 bzw. 9,2 ( $df=2$ ). Somit sind der Effekt `zeit` sowie die Interaktion signifikant.

Alternativ kann auch die Funktion `np.anova` (siehe Anhang 3.8) dazu benutzt werden:

```
np.anova(score~Geschlecht*Zeit+Error(Vpn/Zeit),winer518t,method=2)
```

#### mit SPSS:

Zunächst die relativ einfache Durchführung der klassischen Puri & Sen-Tests: Dazu genügt es, die für die Berechnung der  $\chi^2$ -Werte erforderlichen Streuungsquadratsummen aus der Tabelle 6-4 (Kapitel 6.3) zu entnehmen: Die Spalten `Quadratsummen` enthalten die  $SS_{Effekt}$  bzw.  $SS_{Fehler}$  und natürlich `df` die dazugehörigen Freiheitsgrade. Die  $\chi^2$ -Werte sind daraus „mit der Hand“ auszurechnen. Die Berechnung der  $\chi^2$ -Werte ist oben bei R unter `Ergebnisse 6-1` wiedergegeben.

### 6. 4. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren

Das Puri & Sen-Verfahren wird nun auf einen 3-faktoriellen Versuchsplan mit zwei Messwiederholungsfaktoren angewandt. Dazu wird der Beispieldatensatz 5 (`mydata5`) benutzt.

#### mit R:

Zunächst wird die elementare Berechnung gezeigt. Dazu wird der Dataframe `mydata5t` benutzt (vgl. auch Kapitel 5.4.3), die abhängige Variable `Fehler` in Ränge `RFehler` transformiert und damit eine parametrische Varianzanalyse mittels `ezANOVA` durchgeführt.

```
mydata5t<-within(mydata5t,RFehler<-rank(Fehler))
ezANOVA(mydata5t, RFehler, Vpn, within=.(Medikament,Aufgabe),
        between=Geschlecht, detailed=T)->ano
```

	Effect	DFn	DFd	SSn	SSd	F
	(Intercept)	1	6	95922.0000	3848.542	149.5454772
	Geschlecht	1	6	3741.1250	3848.542	5.8325340
	Medikament	2	12	5419.0833	1574.500	20.6506828
	Aufgabe	2	12	8037.7500	2513.000	19.1908078
	Geschlecht:Medikament	2	12	159.2500	1574.500	0.6068593
	Geschlecht:Aufgabe	2	12	523.5833	2513.000	1.2500995
	Medikament:Aufgabe	4	24	1099.6667	2585.333	2.5520887
	Geschlecht:Medikament:Aufgabe	4	24	189.1667	2585.333	0.4390150

Die für die Berechnung der  $\chi^2$ -Werte erforderlichen Streuungsquadratsummen sind aus dieser Tabelle zu entnehmen: Die Spalten `SSn` enthalten die  $SS_{Effekt}$ , die Spalten `SSd` die  $SS_{Fehler}$  und natürlich `DFn` und `DFd` die dazugehörigen Freiheitsgrade. Die  $\chi^2$ -Werte werden daraus „per Hand“ ausgerechnet (zu deren Aufbau vgl. den Anfang von Kapitel 6.4):

$$\chi_{Geschlecht}^2 = \frac{3741}{1084,2} = 3,45$$

$$MS_{innerhalb(Medikamente)} = \frac{5419,06 + 159,2 + 1574,5}{2 + 2 + 12} = 447,0$$

$$\chi_{Medikamente}^2 = \frac{5419,06}{447,0} = 12,12$$

$$\chi_{Medikamente \times Geschlecht}^2 = \frac{159,2}{447,0} = 0,36$$

$$MS_{innerhalb(Aufgabe)} = \frac{8037,75 + 523,6 + 2513}{2 + 2 + 12} = 692,1$$

$$\chi_{Aufgabe}^2 = \frac{8037,75}{692,1} = 11,62$$

$$\chi_{Aufgabe \times Geschlecht}^2 = \frac{523,6}{692,1} = 0,755$$

$$MS_{innerhalb(Interaktion)} = \frac{1099,7 + 189,17 + 2585,3}{4 + 4 + 24} = 121,06$$

$$\chi_{Interaktion}^2 = \frac{1099,7}{121,06} = 9,08$$

$$\chi_{Interaktion \times Geschlecht}^2 = \frac{189,17}{121,06} = 1,56$$

Die für die Tests erforderlichen Freiheitsgrade entsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse. Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 (df=1), bei 6,0 bzw. 9,2 (df=2) sowie bei 9,5 bzw. 13,3 (df=4). Somit sind lediglich die Effekte `Medikamente` sowie `Aufgabe` signifikant.

Alternativ kann auch die Funktion `np.anova` (siehe Anhang 3.8) dazu benutzt werden:

```
np.anova(Fehler~Geschlecht*Medikament*Aufgabe+
        Error(Vpn/(Medikament*Aufgabe)), mydata5t, method=2)
```

### mit SPSS:

Nach Transformation der abhängigen Variablen `Fehler` in Ränge `RFehler` wird damit eine parametrische Varianzanalyse durchgeführt.

```
Varstocases
/Id=Vpn
/Make Fehler from v1 to v9
/index=Medikament(3) Aufgabe(3)
/keep=Geschlecht
/null=keep.

Rank Variables=Fehler / rank into RFehler.
Sort cases by Vpn Medikament Aufgabe.

casestovars
/Id=Vpn
/index=Medikament Aufgabe
/groupby=variable.

GLM RFehler.1.1 RFehler.1.2 RFehler.1.3 RFehler.2.1 RFehler.2.2
    RFehler.2.3 RFehler.3.1 RFehler.3.2 RFehler.3.3 by Geschlecht
/WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
/WSdesign=Medikament Aufgabe Medikament*Aufgabe
/design=Geschlecht.
```

Tests der Innersubjekteffekte						
Quelle		Quadrat summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	5419,083	2	2709,542	20,651	,000
Medikament * Geschlecht	Sphärizität angen.	159,250	2	79,625	,607	,561
Fehler(Medikament)	Sphärizität angen.	1574,500	12	131,208		
Aufgabe	Sphärizität angen.	8037,750	2	4018,875	19,191	,000
Aufgabe * Geschlecht	Sphärizität angen.	523,583	2	261,792	1,250	,321
Fehler(Aufgabe)	Sphärizität angen.	2513,000	12	209,417		
Medikament * Aufgabe	Sphärizität angen.	1099,667	4	274,917	2,552	,065
Medik * Aufg * Geschl	Sphärizität angen.	189,167	4	47,292	,439	,779
Fehler (Medik*Aufg	Sphärizität angen.	2585,333	24	107,722		

Da hier anstatt des F-Tests der  $\chi^2$ -Test benutzt wird, spielt die Sphärizität keine Rolle, so dass die Ergebnisse aus der entsprechenden Zeile zu entnehmen sind, während die übrigen in o.a. Tabelle weggelassen wurden. Aus den Spalten „Quadratsumme“ und „df“ werden nun die  $\chi^2$ -Werte berechnet. Die Berechnung ist exakt dieselbe wie oben für R (siehe Ergebnisse für die oben errechneten  $\chi^2$ -Werte).

## 6. 5 Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF)

Die Rangtransformation werden beim KWF-Verfahren anders als beim Puri & Sen-Test vorgenommen: Zum einen erhält jede Erhebungseinheit (Vpn) einen (*Wilcoxon-*) Rang, zum anderen werden wie beim Friedman-Test pro Vpn Ränge für die einzelnen Messwiederholungen vergeben (*Friedman-Ränge*). Beide Ränge werden dann zu einem zusammengefasst. Darüber hinaus werden nicht die F-Tests verwendet, sondern wie beim Puri & Sen-Test aus den Streuungsquadratsummen (SS, Sum of Sq)  $\chi^2$ -Tests konstruiert. Die Tests der Haupteffekte (in den Beispielen z.B. Geschlecht und Zeit) sind mit denen von Kruskal-Wallis bzw. von Friedman identisch. Wie beim o.a. Puri & Sen-Test ist auch hier eine Überprüfung der Sphärizität nicht erforderlich, bestenfalls die der Varianzhomogenität. Allerdings hat sich gezeigt, dass dieses Verfahren relativ robust sowohl gegen Heteroskedastizität der Messwiederholungsvariablen als auch der Kovarianzmatrizen ist (vgl. Lüpsen, 2020a und Lüpsen, 2020b).

Folgende Schritte sind für eine Analysevariable  $x$  durchzuführen:

- Für die Analyse-Variablen  $x_1, \dots, x_J$  für jede Erhebungseinheit (Versuchsperson)  $m$  die Summe aller  $J$  Messwiederholungen ( $Sum$ ) errechnen
- Diese Summen  $Sum$  in Ränge ( $R_{Sum}$ )  $1, \dots, n$  umrechnen.
- Für jede Erhebungseinheit (Versuchsperson)  $m$  werden die Werte  $x_1, \dots, x_J$  in Friedman-Ränge ( $1, \dots, J$ ) transformiert und ergeben  $R_{x_{m1}}, \dots, R_{x_{mJ}}$ .
- Für jede Erhebungseinheit  $m$  und Messwiederholung  $j=1, \dots, J$  berechnen von  $(R_{Sum_m} - 1) * J + R_{x_{mj}}$
- Mit diesen Rängen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Schließlich werden dieselben  $\chi^2$ -Tests wie beim Puri & Sen-Verfahren durchgeführt (Details siehe oben Abschnitt 6.4).

### 6. 5. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor

Für ein 2-faktorielles Design werden die Schritte am Datensatz des Beispiels 4 (*winer518*) demonstriert.

**mit R:**

Zunächst wird die elementare Berechnung vorgestellt. Die ersten Schritte sind weitgehend dieselben wie in Kapitel 5.1.2. Zusätzlich sind am Anfang erforderlich:

- Die Summe der Variablen  $t_1, \dots, t_3$  errechnen und diese in Ränge ( $R_{sum}$ ) wandeln.

Nach der Umstrukturierung noch folgende Schritte:

- Die Messwiederholungsvariablen pro Vpn in Friedman-Ränge  $R_{score}$  umrechnen.
- Aus  $R_{sum}$  und  $R_{score}$  die zu analysierende Variable  $R_y$  bilden.

Schließlich wird die Anova mit `aov` oder `ezANOVA` durchgeführt. (Falls die  $\chi^2$ -Werte „mit der Hand“ ausgerechnet werden, empfiehlt sich die Verwendung von `aov`. Soll dagegen die Berechnung in R programmiert werden, ist `ezANOVA` vorzuziehen.)

```

Rsum      <- rank(rowSums(winer518[,3:5]))
Vpn       <- 1:10
winer518  <- cbind(Vpn, Rsum, winer518)
winer518  <- within(winer518,
                    {Geschlecht<-factor(Geschlecht); Vpn<-factor(Vpn)})
winer518t<- reshape(winer518,direction="long",timevar="Zeit",
                    v.names="score", varying=c("t1","t2","t3"),idvar="Vpn")
winer518t<- within(winer518t, Zeit<-factor(Zeit))
Rscore    <- ave(winer518t$score,winer518t$Vpn,FUN=rank)
Ry        <- (Rsum-1)*3 + Rscore
aov3      <- aov(Ry~Geschlecht*Zeit+Error(Vpn/Zeit),winer518t)
summary(aov3)

```

bzw. alternativ mit `ezANOVA`, wobei zu beachten ist, dass alle verwendeten Variablen Teil des angegebenen Dataframes sein müssen (während `aov` da weniger penibel ist und auch andere Variablen akzeptiert, sofern sie die passende Länge haben) und dass mit `detailed=T` die Streuungsquadratsummen ausgegeben werden:

```

winer518t <- cbind(winer518t,Rscore,Ry)
ezANOVA(winer518t,Ry,Vpn,within=Zeit,between=Geschlecht,detailed=T)

```

Error: Vpn					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	24.3	24.3	0.089	0.773
Residuals	8	2176.2	272.0		
Error: Vpn:Zeit					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zeit	2	8.6	4.300	34.4	1.61e-06 ***
Geschlecht:Zeit	2	7.4	3.700	29.6	4.20e-06 ***
Residuals	16	2.0	0.125		

Tabelle 6-5

In diesem Fall ist es am einfachsten, die  $\chi^2$ -Werte aus den Spalten „Sum Sq“ und „Df“ „mit der Hand“ auszurechnen:

$$MS_{\text{zwischen}} = \frac{24,3 + 2176,2}{1 + 8} = 244,5$$

$$\chi_{\text{Geschlecht}}^2 = \frac{24,3}{244,5} = 0,1$$

$$\chi_{\text{Zeit}}^2 = \frac{8,6}{(8,6 + 7,4 + 2,0)/(2 + 2 + 16)} = 9,56$$

$$\chi_{\text{Interaktion}}^2 = \frac{7,4}{(8,6 + 7,4 + 2,0)/(2 + 2 + 16)} = 8,22$$

Ergebnisse 6-2

Die für die Tests erforderlichen Freiheitsgrade entsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse. Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 (df=1) sowie bei 6,0 bzw. 9,2 (df=2). Somit sind der Effekt `zeit` sowie die Interaktion stark signifikant.



Die Ausgabe der Anova-Tabelle von ezANOVA (zum Vergleich):

\$ANOVA							
	Effect	DFn	DFd	SSn	SSd	F	p
1	(Intercept)	1	8	7207.5	2176.2	26.49572650	8.771197e-04
2	Geschlecht	1	8	24.3	2176.2	0.08933002	7.726474e-01
3	Zeit	2	16	8.6	2.0	34.40000000	1.606176e-06
4	Geschlecht:Zeit	2	16	7.4	2.0	29.60000000	4.199689e-06

Hier bezeichnen  $SS_n$  die Sum of Squares des jeweiligen Effekts und  $SS_d$  die Streuung des dazugehörenden Fehler- (Residuen) Terms. Bei diesem vergleichsweise einfachen Design sind die  $\chi^2$ -Werte für die Effekte der Gruppierungs- wie auch der Messwiederholungsfaktoren gleich aufgebaut. Liegen allerdings mehrere Gruppierungsfaktoren vor, ist das Prozedere etwas schwieriger, da bei  $MS_{zwischen}$  mehr als Effekt- und Residuenstreuung zu berücksichtigen sind. Dazu wird auf die nachfolgenden Kapitel verwiesen, da statt dessen die Verwendung der u.a. R-Funktion empfohlen wird.

Die Umrechnung der  $\chi^2$ -Werte in F-Werte gemäß Iman & Davenport erübrigt sich hier, da diese nur für den Effekt *Zeit* vorgenommen werden kann, was bereits früher gezeigt wurde.

Alternativ kann auch die Funktion `np.anova` (vgl. Anhang 3.8) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`winer518t`). Eingabe und Ausgabe:

```
np.anova(score~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t)
```

generalized Kruskal-Wallis/Friedman (Puri & Sen) tests including Iman & Davenport F-tests							
	Df	Sum Sq	Chisq	Pr(>Chi)	F value	Pr(>F)	
Geschlecht	1	24.3	0.1064	0.74424			
Residuals Btw.Vpn	8	2030.4					
Zeit	2	8.6	9.5556	0.00841	8.2340	0.003478	**
Geschlecht:Zeit	2	7.4	8.2222	0.01639	6.2830	0.009686	**
Residuals	16	2.0					

### mit SPSS:

Die ersten Schritte sind weitgehend dieselben wie in Kapitel 5.1.2. Zusätzlich sind am Anfang erforderlich:

- Errechnen der Summe der Messwiederholungsvariablen ( $sum$ ) Transformation in Ränge ( $R_{sum}$ ).
- Umstrukturieren des Datensatzes, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben.
- Pro  $V_{pn}$  aus den Werten von `score` die Ränge  $R_{score}$  errechnen.
- Aus  $R_{sum}$  und  $R_{score}$  die zu analysierende Variable  $R_y$  errechnen.
- Zurücktransformieren des Datensatzes, wobei aus  $R_y$  für die 3 Zeitpunkte die Variablen  $R_{y.1}$ ,  $R_{y.2}$ ,  $R_{y.3}$  entstehen.
- Durchführen der Varianzanalyse

Die hierfür erforderlichen SPSS-Anweisungen:

```
compute sum=t1+t2+t3.
rank variables=Sum (A)
  /rank into RSum.

Varstocases
  /Id=Vpn
  /make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht Sum RSum
  /null=keep.

rank variables=score(A) by Vpn
  /rank into RScore.

compute Ry=(RSum-1)*3 + RScore.

Casestovars
  /Id=Vpn
  /Index=Zeit
  /Groupby=variable.

GLM Ry.1 Ry.2 Ry.3 by Geschlecht
  /wsfactor=Zeit 3 polynomial
  /wsdesign=Zeit
  /design=Geschlecht.
```

Nachfolgend die Ergebnisse der Varianzanalyse, zunächst die Effekte des Messwiederholungsfaktors (Innersubjekteffekte), danach der Effekt des Gruppierungsfaktors (Zwischen-subjekteffekte). Da eine Prüfung der Sphärizität hier entfällt, interessieren in der Anova-Tabelle nur die Zeilen mit den unkorrigierten F-Tests.

Quelle		Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	8,600	2	4,300	34,400	,000
Zeit * Geschlecht	Sphärizität angen.	7,400	2	3,700	29,600	,000
Fehler(Zeit)	Sphärizität angen.	2,000	16	,125		

Tabelle 6-6a

Quelle	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Konstanter Term	7207,500	1	7207,500	26,496	,001
Geschlecht	24,300	1	24,300	,089	,773
Fehler	2176,200	8	272,025		

Tabelle 6-6b

Aus den Spalten „Quadratsumme“ und „df“ werden nun die  $\chi^2$ -Werte berechnet, zunächst aus Tabelle 6-6b für den Gruppierungsfaktor *Geschlecht*, danach aus Tabelle 6-6a für die Messwiederholungseffekte. Die Berechnung ist exakt dieselbe wie oben für R (siehe Ergebnisse 6-2).

Die Umrechnung der  $\chi^2$ -Werte in F-Werte gemäß Iman & Davenport erübrigt sich hier, da diese nur für den Effekt *Zeit* vorgenommen werden kann, was bereits früher gezeigt wurde.

### 6. 5. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren

Das oben beschriebene Verfahren wird nun auf einen 3-faktoriellen Versuchsplan mit zwei Messwiederholungsfaktoren angewandt. Dazu wird der Beispieldatensatz 5 (`mydata5`) benutzt.

#### mit R:

Dazu wird der Dataframe `mydata5t` benutzt (vgl. auch Kapitel 5.1.2 und 5.4.3). Die Durchführung der Analyse auf Basis der KWF-Rangtransformation wird wieder mit der o.a. Funktion `np.anova` gezeigt. Die elementare Berechnung ist zum einen aus dem vorigen Abschnitt ersichtlich, zum anderen die Bildung der  $\chi^2$ -Werte aus der Lösung mit SPSS.

```
np.anova(Fehler~Geschlecht*Medikament*Aufgabe+
         Error(Vpn/(Medikament*Aufgabe)),mydata5t)
```

mit folgender Ausgabe (ohne die Ergebnisse des Iman & Davenport-Tests):

generalized Kruskal-Wallis/Friedman (Puri & Sen) tests				
	Df	Sum Sq	Chisq	Pr(>Chi)
Geschlecht	1	5832.0	1.3494	0.24538
Residuals Btw.Vpn	6	24421.5		
Medikament	2	111.1	12.7536	0.0017
Geschl:Medikament	2	0.9	0.1029	0.94987
Residuals Medikament	12	27.4		
Aufgabe	2	150.6	11.1889	0.00372
Geschlecht:Aufgabe	2	2.6	0.1920	0.90849
Residuals Geschl:Aufgabe	12	62.2		
Medikament:Aufgabe	4	26.4	11.5273	0.02124
Geschlecht:Medikament:Aufgabe	4	1.7	0.7455	0.94561
Residuals Geschl:Medikament:Aufgabe	24	45.2		

#### mit SPSS:

Die Kommandos sind praktisch dieselben wie in Abschnitt 5.4.4, wo dasselbe Design, allerdings ohne Gruppierungsfaktor, analysiert wurde. Lediglich muss noch vorab die Summe der Variablen `v1, . . . , v9` berechnet und in Ränge `RSum` transformiert werden. Zusätzlich werden noch die Ränge `RSum` und `RFehler` miteinander verknüpft und ergeben `Ry`.

```
compute sum=sum(v1 to v9).
rank variables=Sum (A)
  /rank into RSum.

Varstocases
/Id=Vpn
/make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
/index=Medikament(3) Aufgabe(3)
/keep=Geschlecht RSum
/null=keep.

Rank variables=Fehler (A) by Vpn
/rank into RFehler.
compute Ry=(RSum-1)*9 + RFehler.

Sort cases by Vpn Medikament Aufgabe.

Casestovars
/Id=Vpn
/index=Medikament Aufgabe
/groupby=variable.
```

Schließlich die eigentliche Varianzanalyse:

```
GLM Ry.1.1 Ry.1.2 Ry.1.3 Ry.2.1 Ry.2.2 Ry.2.3 Ry.3.1 Ry.3.2 Ry.3.3
  by Geschlecht
  /WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
  /WSdesign=Medikament Aufgabe Medikament*Aufgabe
  /design=Geschlecht.
```

Nachfolgend die Ergebnisse der Varianzanalyse, zunächst die Effekte des Messwiederholungsfaktors (Innersubjekteffekte), danach der Effekt des Gruppierungsfaktors (Zwischen-subjekteffekte). Da eine Prüfung der Sphärizität hier entfällt, interessieren in der Anova-Tabelle nur die Zeilen mit den unkorrigierten F-Tests.

Quelle		Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	111,063	2	55,531	24,342	,000
Medikament * Geschlecht	Sphärizität angen.	,896	2	,448	,196	,824
Fehler(Medikament)	Sphärizität angen.	27,375	12	2,281		
Aufgabe	Sphärizität angen.	150,583	2	75,292	14,534	,001
Aufgabe * Geschlecht	Sphärizität angen.	2,583	2	1,292	,249	,783
Fehler(Aufgabe)	Sphärizität angen.	62,167	12	5,181		
Medikament * Aufgabe	Sphärizität angen.	26,417	4	6,604	3,506	,022
Medikam* Aufgabe* Geschl	Sphärizität angen.	1,708	4	,427	,227	,921
Fehler(Medikament*Aufgabe)	Sphärizität angen.	45,208	24	1,884		

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	95922,000	1	95922,000	23,567	,003
Geschlecht	5832,000	1	5832,000	1,433	,276
Fehler	24421,500	6	4070,250		

Die Berechnung der  $\chi^2$ -Werte (vgl. dazu deren Aufbau am Anfang von Kapitel 6.4):

$$MS_{\text{zwischen}} = \frac{5832 + 24421,5}{1 + 6} = 4321,93$$

$$\chi_{\text{Geschlecht}}^2 = \frac{24421,5}{4321,93} = 1,35$$

$$MS_{\text{innerhalb(Medikamente)}} = \frac{111,06 + 0,9 + 27,38}{2 + 2 + 12} = 8,709$$

$$\chi_{\text{Medikamente}}^2 = \frac{111,06}{8,709} = 12,75$$

$$\chi_{\text{Medikamente} \times \text{Geschlecht}}^2 = \frac{0,9}{8,709} = 0,10$$

$$MS_{\text{innerhalb(Aufgabe)}} = \frac{150,58 + 2,58 + 62,17}{2 + 2 + 12} = 13,458$$

$$\chi_{\text{Aufgabe}}^2 = \frac{150,58}{13,458} = 11,19$$

$$\chi_{\text{Aufgabe} \times \text{Geschlecht}}^2 = \frac{2,58}{13,458} = 0,19$$

$$MS_{\text{innerhalb(Interaktion)}} = \frac{26,42 + 1,71 + 45,21}{4 + 4 + 24} = 2,292$$

$$\chi_{\text{Interaktion}}^2 = \frac{26,42}{2,292} = 11,53$$

$$\chi_{\text{Interaktion} \times \text{Geschlecht}}^2 = \frac{1,71}{2,292} = 0,75$$

Ergebnisse 6-4

Die für die Signifikanzprüfung erforderlichen Freiheitsgrade sind der o.a. parametrischen Varianzanalyse zu entnehmen, also  $df=1$  für den Gruppeneffekt bzw.  $df=2$  für die einfachen Messwiederholungseffekte bzw.  $df=4$  für die Messwiederholungsinteraktion. Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 ( $df=1$ ), bei 6,0 bzw. 9,2 ( $df=2$ ) sowie bei 9,5 bzw. 13,3 ( $df=4$ ). Somit sind die Effekte „Medikamente“, „Aufgabe“ sowie die Interaktion stark signifikant.

### 6. 5. 3 Zwei Gruppierungs- und ein Messwiederholungsfaktoren

Das oben beschriebene Verfahren wird nun auf einen 3-faktoriellen Versuchsplan mit zwei Gruppierungsfaktoren angewandt. Dazu wird der Beispieldatensatz 6 (winer568) benutzt.

#### mit R:

Hier soll die Durchführung der Analyse lediglich wieder mit der o.a. Funktion `np.anova` gezeigt werden. Die elementare Berechnung ist aus den vorangegangenen Abschnitten, z.B. Kapitel 6.5.3, ersichtlich. Ausgangsbasis ist der in Kapitel 5.1.2 erstellte Dataframe

winer568t :

```
np.anova(x ~ A*B*Zeit+Error(Vpn/Zeit),winer568t, method=0)
```

generalized Kruskal-Wallis/Friedman tests including Iman & Davenport F-tests						
	Df	Sum Sq	Chisq	Pr(>Chi)	F value	Pr(>F)
A	1	1200.00	1.4680	0.22567		
B	1	4800.00	5.8719	0.01538		
A:B	1	48.00	0.0587	0.80853		
Residuals Btw.Vpn	8	2944.00				
Zeit	3	52.13	32.6348	0.00000	106.6744	5.185e-14
A:Zeit	3	2.63	1.6435	0.64957	0.5262	0.6685
B:Zeit	3	0.79	0.4957	0.91985	0.1536	0.9264
A:B:Zeit	3	0.29	0.1826	0.98035	0.0561	0.9821
Residuals Zeit	24	1.67				

#### mit SPSS:

Für die Durchführung der Analyse wird hier auf das Kapitel 6.6.2 verwiesen. Dort wird für diesen Versuchsplan das Verfahren von van der Waerden gezeigt, das hinsichtlich des Prozederes mit dem von Puri & Sen weitgehend identisch ist. Bei den Rechenvorgängen ist lediglich zu beachten, dass die Transformation in normal scores entfällt und die kombinierten Ränge sich über

```
compute Ry=(Rsum-1)*4 + Rscore.
```

errechnen. Die Bildung der  $\chi^2$ -Werte erfolgt bei beiden Verfahren nach demselben Prinzip.

## 6. 6 van der Waerden-Tests

Das Verfahren von van der Waerden verläuft zunächst ähnlich dem KWF-Verfahren (vgl. Kapitel 6.5). D.h. zum einen erhalten die einzelnen Erhebungseinheiten (Vpn) Ränge ( $R_{Sum}$ ) entsprechend der Summe der Messwiederholungen, und zum anderen werden die Werte der Messwiederholungen pro Fall analog dem Friedman-Test in Friedman-Ränge  $R_{xi}$  transformiert. Die Ränge werden jeweils in normal scores umgerechnet. Beide scores werden addiert. Schließlich werden die  $\chi^2$ -Tests wie beim Verfahren von Puri & Sen durchgeführt. Wie beim o.a. Puri & Sen-Test und KWF-Verfahren ist auch hier eine Überprüfung der Sphärizität nicht erforderlich, bestenfalls die der Varianzhomogenität. Allerdings hat sich gezeigt, dass dieses Verfahren relativ robust sowohl gegen Heteroskedastizität der Messwiederholungsvariablen als auch der Kovarianzmatrizen ist (vgl. Lüpsen, 2020a und Lüpsen, 2020b).

Folgende Schritte sind für eine Analysevariable  $x$  durchzuführen, wobei im Folgenden  $J$ =Anzahl der gesamten Messwiederholungen ist und die Anzahl der Analysevariablen im Beispiel 4 genau eine:

- Für die Analyse-Variable  $x$  (Variablen  $x_1, \dots, x_j$ ) pro Erhebungseinheit  $m$  die Summe aller  $J$  Messwiederholungen ( $Sum$ ) errechnen.
- Diese Summen  $Sum$  in Ränge ( $R_{Sum}$ )  $1, \dots, n$  umrechnen.
- Umrechnung der Rangsummen  $R_{Sum}$  in normal scores:  $nsum = \Phi^{-1}(R_{sum}/(n+1))$ , wobei  $n$  die Anzahl der Erhebungseinheiten ist.
- Für jede Erhebungseinheit (Versuchsperson)  $m$  werden die Werte  $x_{m1}, \dots, x_{mj}$  in Friedman-Ränge  $(1, \dots, J)$  transformiert und ergeben  $R_{x_{m1}}, \dots, R_{x_{mj}}$ .
- Umrechnung von  $R_{x_{mj}}$  in normal scores:  $nscore_{mj} = \Phi^{-1}(R_{x_{mj}}/(J+1))$ .
- Für jede Erhebungseinheit  $m$  und Messwiederholung  $j=1, \dots, J$   

$$nsx_{mj} = nsum_m + nscore_{mj}$$
 berechnen
- Mit diesen normal scores wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle werden  $\chi^2$ -Tests aufgestellt, exakt wie beim Puri & Sen-Verfahren (Details siehe oben Abschnitt 6.4).

### 6. 6. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor

Die Schritte sollen zunächst wiederum am Datensatz des Beispiels 4 demonstriert werden. Die Überprüfung der Sphärizität kann entfallen, da hier  $\chi^2$ - anstatt F-Tests durchgeführt werden

#### mit R:

Auch hier wieder zunächst die elementare Berechnung, anschließend unter Verwendung der R-Funktion `np.anova` für dieses Verfahren. Ausgangsbasis ist der Dataframe `winer518`. Die Schritte zur Erlangung der Anova-Tabelle, mit deren Hilfe die  $\chi^2$ -Tests errechnet werden können, sind weitgehend identisch mit denen aus Kapitel 6.4 und 6.5. Zusätzlich wird zunächst die Anzahl der Merkmalsträger `nc` ermittelt, mit deren Hilfe die normal scores `nsum` für die Merkmalsträger berechnet werden. Ebenso werden die normal scores `nscore` für die 3 Messwiederholungen berechnet. Die Summe aus beiden zusammen bilden die normal scores `nsx`, auf deren Basis die Varianzanalyse durchgeführt wird:

```

Rsum      <- rank(rowSums(winer518[,3:5]))
nc        <- dim(winer518)[1]
nsum      <- qnorm(Rsum/(nc+1))
Vpn       <- 1:10
winer518  <- cbind(winer518, Vpn, Rsum, nsum)
winer518  <- within(winer518,
                    {Geschlecht<-factor(Geschlecht); Vpn<-factor(Vpn)})
winer518t<- reshape(winer518, direction="long", timevar="Zeit",
                    v.names="score", varying=c("t1","t2","t3"), idvar="Vpn")
winer518t<- within(winer518t, Zeit<-factor(Zeit))
Rscore    <- ave(winer518t$score, winer518t$Vpn, FUN=rank)
nscore    <- qnorm(Rscore/4)
nsx       <- nsum + nscore
aov3      <- aov(nsx~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t)
summary(aov3)

```

Zunächst die Ausgabe der (parametrischen) Anova:

Error: Vpn						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Geschlecht	1	0.368	0.3681	0.165	0.695	
Residuals	8	17.833	2.2291			
Error: Vpn:Zeit						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Zeit	2	3.847	1.9237	33.81	1.8e-06	***
Geschlecht:Zeit	2	3.331	1.6657	29.27	4.5e-06	***
Residuals	16	0.910	0.0569			

Nun zur Berechnung der  $\chi^2$ -Werte:

Aus dem oberen Teil der Anova-Tabelle ist zu entnehmen:

$$MS_{zwischen} = \frac{0,368 + 17,833}{1 + 8} = 2,022$$

$$\chi_{Geschlecht}^2 = \frac{0,368}{2,022} = 0,182$$

Aus dem unteren Teil der Anova-Tabelle ist zu entnehmen:

$$\chi_{Zeit}^2 = \frac{3,847}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,847}{0,4044} = 9,513$$

$$\chi_{Interaktion}^2 = \frac{3,331}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,331}{0,4044} = 8,24$$

Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 (df=1) sowie bei 6,0 bzw. 9,2 (df=2). Somit sind der Effekt „Zeit“ sowie die Interaktion stark signifikant.

Alternativ kann auch die Funktion `np.anova` (vgl. Anhang 3.8) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Über den Parameter `method=1` wird das van der Waerden-Verfahren ausgewählt. Basis ist auch hierfür der umstrukturierte Datensatz (`winer518t`). Eingabe und Ausgabe:

```
np.anova(score~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t, method=1)
```

	Df	Sum Sq	Chisq	Pr(>Chi)
Geschlecht	1	0.3681	0.1944	0.65929
Residuals Btw.Vpn	8	16.6742		
Zeit	2	3.8475	9.5124	0.00860
Geschlecht:Zeit	2	3.3315	8.2367	0.01627
Residuals	16	0.9104		

**mit SPSS:**

Ausgangspunkt ist der Beispieldatensatz 4. Folgende Schritte sind erforderlich:

- Errechnen der Summe der Messwiederholungsvariablen (Sum)
- Transformation der Summe in Ränge (RSum).
- Ermitteln der Anzahl der Fälle (nc) mittels `Aggregate`.
- Umwandeln von RSum in normal scores (Variable nsum) mittels `Idf.normal`.
- Umstrukturieren des Datensatzes, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben. Daraus resultiert die abhängige Variable `score`.
- Pro Vpn aus den Werten von `score` die Ränge `Rscore` errechnen.
- Umrechnen in normal score `nscore` mittels `Idf.normal`.
- Aus `nsum` und `nscore` die zu analysierende Variable `nsx` als deren Summe errechnen.
- Zurücktransformieren des Datensatzes, wobei aus `nsx` für die 3 Zeitpunkte die Variablen `nsx.1`, `nsx.2`, `nsx.3` entstehen.
- Durchführen der Varianzanalyse für die Variablen `nsx.1`, `nsx.2`, `nsx.3`.
- Berechnung der  $\chi^2$ -Werte gemäß Formeln 2-6 bzw. 2-7.

Die hierfür erforderlichen SPSS-Anweisungen:

```
compute sum=t1+t2+t3.
rank variables=Sum (A)
  /rank into RSum.
Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(RSum).
compute nsum=Idf.normal(RSum/(nc+1),0,1).

Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht Sum RSum nsum nc
  /null=keep.

Rank Variables=score by Vpn / rank into Rscore.
compute nscore=Idf.normal(Rscore/4,0,1).
compute nsx=nsum+nscore.
Sort cases by Vpn Zeit.
```



```

Casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

GLM nsx.1 nsx.2 nsx.3
  /wsfactor=Zeit 3 polynomial
  /wsdesign=Zeit
  /design=Geschlecht.

```

Zunächst die Ausgabe der (parametrischen) Anova:

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	9,006E-005	1	9,006E-005	,000	,995
Geschlecht	,368	1	,368	,165	,695
Fehler	17,833	8	2,229		

Tabelle 6-9a

Tests der Innersubjekteffekte						
Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	3,847	2	1,924	33,808	,000
Zeit * Geschlecht	Sphärizität angen.	3,331	2	1,666	29,274	,000
Fehler(Zeit)	Sphärizität angen.	,910	16	,057		

Tabelle 6-9b

Aus Tabelle 6-9a ist zu entnehmen:

$$MS_{\text{zwischen}} = \frac{0,368 + 17,833}{1 + 8} = 2,022$$

$$\chi_{\text{Geschlecht}}^2 = \frac{0,368}{2,022} = 0,182$$

Aus Tabelle 6-9b ist zu entnehmen:

$$\chi_{\text{Zeit}}^2 = \frac{3,847}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,847}{0,4044} = 9,513$$

$$\chi_{\text{Interaktion}}^2 = \frac{3,331}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,331}{0,4044} = 8,24$$

Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 (df=1) sowie bei 6,0 bzw. 9,2 (df=2). Somit sind der Effekt „Zeit“ sowie die Interaktion stark signifikant.

### 6. 6. 2 Zwei Gruppierungs- und ein Messwiederholungsfaktor

Die Schritte sollen am Datensatz des Beispiels 6 (winer<sub>568</sub>) demonstriert werden. Die Überprüfung der Sphärizität kann wieder entfallen, da hier  $\chi^2$ - anstatt F-Tests durchgeführt werden.

Eine Bemerkung vorab zu den nachfolgenden Ergebnissen. Dort sind die Tests für die Interaktionen mit der Messwiederholung „Zeit“ mit p=0,64 (A\*Zeit) bzw. p=0,93 (B\*Zeit) weit entfernt von einem signifikanten Ergebnis. Dagegen wurden diese Effekte in der ART- wie auch in

der ART+INT-Analyse (Kapitel 6.5.3) als hochsignifikant ausgewiesen. Die gleichen signifikanten Ergebnisse erhielt man mit der parametrischen Analyse und dem RT-Verfahren. Der eklatante Unterschied der Puri & Sen- und der van der Waerden-Tests gegenüber den anderen Verfahren hinsichtlich der Interaktionen A\*Zeit und B\*Zeit ist auf die geringe Residuenstreuung der Messwiederholungseffekte zurückzuführen. Diese geht bei der dort vorgenommenen Rangbildung zum Teil verloren.

### mit R:

Hier soll die Durchführung der Analyse lediglich wieder mit der o.a. Funktion `np.anova` gezeigt werden. Die elementare Berechnung ist zum einen aus den vorangegangenen Kapiteln ersichtlich, zum anderen die Bildung der  $\chi^2$ -Werte aus der Lösung mit SPSS.

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer568t`. Die Varianzanalyse nach Puri & Sen wird mittels der Funktion `np.anova` durchgeführt:

```
np.anova(x ~ A*B*Zeit+Error(Vpn/Zeit),winer568t, method=1)
```

generalized van der Waerden tests				
	Df	Sum Sq	Chisq	Pr(>Chi)
A	1	4.7612	1.6841	0.19438
B	1	16.0715	5.6846	0.01711
A:B	1	0.1196	0.0423	0.83705
Residuals Btw.Vpn	8	10.1467		
Zeit	3	15.8957	32.6629	0.00000
A:Zeit	3	0.8279	1.7013	0.63665
B:Zeit	3	0.2285	0.4695	0.92554
A:B:Zeit	3	0.0794	0.1632	0.98330
Residuals Zeit	24	0.4882		

### mit SPSS:

Die elementaren Berechnungen sollen hier ausführlich gezeigt werden, da für dieses Design die Durchführung des Puri & Sen-Verfahrens nicht gezeigt worden war.

Folgende Schritte sind erforderlich:

- Errechnen der Summe der Messwiederholungsvariablen (`Sum`) und Transformation der Summe in Ränge (`RSum`).
- Ermitteln der Anzahl der Fälle (`nc`) mittels `Aggregate`.
- Umwandeln von `RSum` in normal scores (Variable `nsum`) mittels `Idf.normal`.
- Umstrukturieren des Datensatzes, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben. Daraus resultiert die abhängige Variable `score`.
- Pro `Vpn` aus den Werten von `score` die Ränge `Rscore` sowie die normal scores `nscore` mittels `Idf.normal` errechnen.
- Aus `nsum` und `nscore` die zu analysierende Variable `nsx` als deren Summe errechnen.
- Zurücktransformieren des Datensatzes, wobei aus `nsx` für die 3 Zeitpunkte die Variablen `nsx.1`, `nsx.2`, `nsx.3` entstehen.
- Schließlich die Varianzanalyse für die Variablen `nsx.1`, `nsx.2`, `nsx.3`.

```

compute sum=sum(v1 to v4).
rank variables=Sum (A)
/rank into RSum.

Aggregate
/outfile=* mode=addvariables
/break= /nc=NU(RSum) .
compute nsum=Idf.normal(RSum/(nc+1),0,1) .
execute.

Varstocases
/Id=Vpn
/make Score from v1 v2 v3 v4
/index=Zeit(4)
/keep=A B RSum nsum
/null=keep.

Rank variables=Score (A) by Vpn
/rank into RScore.
compute nscore=Idf.normal(Rscore/5,0,1) .
compute nsx=nsum+nscore.
execute.

Sort cases by Vpn Zeit.
Casestovars
/Id=Vpn
/index=Zeit
/groupby=variable.

GLM nsx.1 nsx.2 nsx.3 nsx.4 by A B
  /WSfactor=Zeit 4 Polynomial
  /WSdesign=Zeit
  /design=A B A*B.

```

Nachfolgend zunächst die Tabelle für die Tests der Gruppierungsvariablen A und B (Zwischensubjekteffekte), danach die Tabelle für alle Tests, bei denen die Messwiederholung Zeit involviert ist (Innersubjekteffekte). Da die Sphärizität nicht erforderlich ist, werden nur die entsprechenden Zeilen wiedergegeben:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	,001	1	,001	,001	,981
A	4,761	1	4,761	3,754	,089
B	16,071	1	16,071	12,671	,007
A * B	,120	1	,120	,094	,767
Fehler	10,147	8	1,268		

Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	15,896	3	5,299	260,499	,000
Zeit * A	Sphärizität angenommen	,828	3	,276	13,568	,000
Zeit * B	Sphärizität angenommen	,228	3	,076	3,744	,024
Zeit * A * B	Sphärizität angenommen	,079	3	,026	1,302	,297
Fehler(Zeit)	Sphärizität angenommen	,488	24	,020		

Nun zur Berechnung der  $\chi^2$ -Werte aus den o.a. Quadratsummen:

$$MS_{zwischen} = \frac{4,761 + 16,071 + 0,12 + 10,147}{1 + 1 + 1 + 8} = 2,827$$

$$\chi_A^2 = \frac{4,761}{2,827} = 1,68$$

$$\chi_B^2 = \frac{16,071}{2,827} = 5,68$$

$$\chi_{A \times B}^2 = \frac{0,12}{2,827} = 0,04$$

$$MS_{innerhalb} = \frac{15,9 + 0,83 + 0,23 + 0,08 + 0,49}{3 + 3 + 3 + 3 + 24} = 0,487$$

$$\chi_{Zeit}^2 = \frac{15,9}{0,487} = 32,65$$

$$\chi_{A \times Zeit}^2 = \frac{0,83}{0,487} = 1,70$$

$$\chi_{B \times Zeit}^2 = \frac{0,23}{0,487} = 0,47$$

$$\chi_{A \times B \times Zeit}^2 = \frac{0,08}{0,487} = 0,16$$

Die für die Signifikanzprüfung erforderlichen Freiheitsgrade sind der o.a. parametrischen Varianzanalyse zu entnehmen, also  $df=1$  für die Gruppeneffekte bzw.  $df=3$  für die Messwiederholungseffekte. Die 5%- bzw. 1%-Schranken der  $\chi^2$ -Verteilung liegen bei 3,8 bzw. 6,6 ( $df=1$ ) sowie bei 7,8 bzw. 11,34 ( $df=3$ ). Somit sind der Effekt „B“ schwach und „Zeit“ stark signifikant.

### 6. 6. 3 Ein Gruppierungs- und zwei Messwiederholungsfaktoren

Die Schritte sollen am Beispieldatensatz 5 (`mydata5`) demonstriert werden. Die Überprüfung der Sphärität kann wieder entfallen, da hier  $\chi^2$ - anstatt F-Tests durchgeführt werden.

**mit R:**

Dazu wird der Dataframe `mydata5t` benutzt (vgl. auch Kapitel 5.1.2 und 5.4.3). Die Durchführung der Analyse wird wieder mit der o.a. Funktion `np.anova` gezeigt. Die elementare Berechnung ist zum einen aus dem vorigen Kapitel ersichtlich:

```
np.anova(Fehler~Geschlecht*Medikament*Aufgabe+
         Error(Vpn/(Medikament*Aufgabe)),mydata5t, method=1)
```

mit folgender Ausgabe (ohne die Ergebnisse des Iman & Davenport-Tests):

generalized van der Waerden tests				
	Df	Sum Sq	Chisq	Pr(>Chi)
Geschlecht	1	6.380	1.0950	0.29536
Residuals Btw.Vpn	6	34.406		
Medikament	2	10.145	12.9443	0.00155
Geschlecht:Medikament	2	0.097	0.1240	0.93990
Residuals Medikament	12	2.298		
Aufgabe	2	13.704	11.5374	0.00312
Geschlecht:Aufgabe	2	0.237	0.1997	0.90498
Residuals Geschlecht:Aufgabe	12	5.063		
Medikament:Aufgabe	4	2.700	13.4543	0.00926
Geschlecht:Medikament:Aufgabe	4	0.133	0.6619	0.95595
Residuals Geschlecht:Medikament:Aufgabe	24	3.589		

**mit SPSS:**

Die Anweisungen sind nahezu identisch mit denen aus den Abschnitten 5.4.4 und 6.5.2, in denen das van der Waerden-Verfahren für doppelte Messwiederholungen bzw. das KWF-Verfahren auf diesen Datensatz angewandt worden war. Zusätzlich wird hier die Transformation der Ränge in normal scores vorgenommen, sowie eine andere Zusammenfassung der scores für die Erhebungseinheiten und für die Messwiederholungen.

Die Kommandos zur Ermittlung der Ränge `RSum` sind ähnlich wie die im vorigen Kapitel:

```
compute sum=sum(v1 to v9).
rank variables=Sum (A)
  /rank into RSum.
```

Nun die Kommandos zur Umstrukturierung, um damit anschließend die Friedman-Ränge `RFehler` zu berechnen, sowie die Wiederherstellung der ursprünglichen Datenstruktur mit denselben Kommandos wie in Kapitel 5.4.3:

```
Varstocases
/Id=Vpn
/make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
/index=Medikament(3) Aufgabe(3)
/keep=Geschlecht RSum
/null=keep.

Rank variables=Fehler (A) by Vpn
/rank into RFehler.
compute nFehler=Idf.normal(RFehler/10,0,1).
execute.

Aggregate
/outfile=* mode=addvariables
/break= /nc=NU(RSum).

compute nSum=Idf.normal(RSum/(nc+1),0,1).

Sort cases by Vpn Medikament Aufgabe.
Casestovars
/Id=Vpn
/index=Medikament Aufgabe
/groupby=variable.
```

Schließlich die eigentliche Varianzanalyse:

```
GLM nscore.1.1 to nscore.3.3 by Geschlecht
  /WSfactor=Medikament 3 Polynomial Aufgabe 3 Polynomial
  /WSdesign=Medikament Aufgabe Medikament*Aufgabe
  /design=Geschlecht.
```

Nachfolgend die Ergebnisse der Varianzanalyse, zunächst die Effekte des Messwiederholungsfaktors (Innersubjekteffekte), danach der Effekt des Gruppierungsfaktors (Zwischen-subjekteffekte). Da eine Prüfung der Sphärität hier entfällt, interessieren in der Anova-Tabelle nur die Zeilen mit den unkorrigierten F-Tests.

Tests der Innersubjekteffekte

Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärität angenommen	10,145	2	5,073	26,492	,000
Medikament * Geschlecht	Sphärität angenommen	,097	2	,049	,254	,780
Fehler(Medikament)	Sphärität angenommen	2,298	12	,191		
Aufgabe	Sphärität angenommen	13,704	2	6,852	16,239	,000
Aufgabe * Geschlecht	Sphärität angenommen	,237	2	,119	,281	,760
Fehler(Aufgabe)	Sphärität angenommen	5,063	12	,422		
Med * Aufg	Sphärität angenommen	2,700	4	,675	4,514	,007
Med * Aufg * Geschl	Sphärität angenommen	,133	4	,033	,222	,923
Fehler(Med*Aufg)	Sphärität angenommen	3,589	24	,150		

## 6. 7 Aligned rank transform (ART und ART+INT)

Das Prinzip des Aligned rank transform-Tests wurde oben bereits erläutert (vgl. Kapitel 4.3.6 und 5.4.4). Würde man jedoch dasselbe Verfahren auf ein gemischtes Design anwenden, so erhielte man „merkwürdige“ Signifikanzen. Der Grund: der Effekt des Gruppierungsfaktors  $\alpha_i$  lässt sich nicht vom Personeneffekt  $\pi_m$  trennen. Daher muss hier ein anderer Weg eingeschlagen werden (vgl. dazu Beasley, 2002). Da es letztlich nur um einen „sauberen“ Test für die Interaktion geht, genügt es, nur für diesen das ART-Verfahren anzuwenden. Die Haupteffekte werden über die o.a. Rank transform Tests (Kapitel 6.3) ermittelt. Aber der Aufwand zur Überprüfung der Interaktion lohnt auch nur dann, wenn der RT hierfür eine Signifikanz ergab, da letztlich mit dem ART nur der liberalere RT abgesichert wird.

Auf Folgendes sei noch aufmerksam gemacht: Beasley (2002) hat zwar auf die Vorzüge des ART im Fall von gemischten Modellen auch bei nichtsphärischen Kovarianzmatrizen und nichtnormalen Daten hingewiesen, dennoch haben Kowalchuk et al. (2003) gezeigt, dass dies nicht mehr gilt, wenn die Kovarianzmatrizen nicht mehr gleich (homogen) sind. Allerdings empfiehlt sich nicht, hier den Box-Test durchzuführen, um diese Voraussetzung zu überprüfen, da der Box-Test selbst sehr viel mehr voraussetzt, so u.a. multivariate Normalverteilung, so dass der Test in diesem Zusammenhang letztlich unbrauchbar wird. Alternativ kann einer der anderen in Kapitel 6.1 erwähnten Tests, z.B. der Levene-like Test, dazu benutzt werden.

Es wird hier an die Ausführungen in Kapitel 2.5 sowie an die Bemerkungen in Kapitel 5.4.4 erinnert, wonach empfohlen wird, nach der Berechnung der Ränge diese noch in normal score (vgl. Kapitel 2.3) umzurechnen.

Hier ist es erforderlich, den einfachen Fall der 2-faktoriellen Analyse und die beiden Fälle der 3-faktoriellen Analyse getrennt zu behandeln. Hieraus lassen sich dann auch Lösungen für höher-faktorielle Versuchspläne ableiten.

### 6.7.1 Ein Gruppierungs- und ein Messwiederholungsfaktor

Die Schritte im Einzelnen:

- Durchführung einer (normalen) Anova mit Haupt- und Interaktionseffekten für die Ränge  $R_x$  der Kriteriumsvariablen  $x$ . Hieraus werden nur die Haupteffekte verwendet.
- per *naive approach* (vgl. Formel 2-4): Eliminieren des Haupteffekts  $\gamma_j$  der Messwiederholungen sowie des Personeneffekts  $\pi_m$  aus der Kriteriumsvariablen  $x$ :

$$e_{jm} = x_{jm} - (\bar{p}_m + \bar{c}_j - \bar{x})$$

alternativ per *standard approach* (vgl. Formel 2-5): Berechnung der Residuen  $e_{jm}$  wie in Kapitel 6.2, anschließend Addition des „reinen“ Interaktionseffekts:

$$e_{jm} = e_{jm} + \bar{ac}_{ij} - (\bar{p}_m + \bar{c}_j - \bar{x})$$

wobei  $\bar{c}_j, \bar{ac}_{ij}$  die Mittelwerte von  $C$  bzw.  $AC$  und  $\bar{p}_m, \bar{x}$  die Personenmittelwerte bzw. das Gesamtmittel sind.

- Umrechnung der so errechneten Residuen  $e_{jm}$  in Ränge.
- Durchführung einer Anova mit Haupt- und Interaktionseffekten mit den Rängen, aus der dann der Interaktionseffekt abgelesen werden kann.

Als Beispiel soll nachfolgend wieder der bereits verwendete Datensatz 4 (`winer518`) dienen.

**mit R:**

Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer518t`. In Kapitel 6.3 wurde der Rank transform Test durchgeführt, aus dem die Haupteffekte abzulesen sind (Tabelle 6-3).

Für die Berechnung der Residuen  $e_{jm}$  wird hier der o.a. „naive approach“ gewählt. Dazu müssen zunächst die Effekte  $\gamma_j$  (mb) und  $\pi_m$  (mp) sowie der Gesamtmittelwert (mm) berechnet werden, um sie von der Kriteriumsvariablen `score` abzuziehen. Diese werden dann nach Rundung auf 6 Stellen in Ränge transformiert, um darauf die Varianzanalyse anzuwenden.

```
attach(winer518t)
mb <- tapply(score, Zeit, mean)
mp <- tapply(score, Vpn, mean)
mm <- mean(score)
ek <- score
n <- dim(winer518t)[1]
for (k in 1:n) {j=Zeit[k]; i=Vpn[k]
  ek[k] <- ek[k] - mb[j] - mp[i] + mm }
ek <- rank(round(ek, digits=6))
summary(aov(ek~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t))
```

Die Anova-Tabelle zeigt einen signifikanten Interaktionseffekt, während die anderen beiden Haupteffekte keine Bedeutung haben:

```

Error: Vpn
      Df Sum Sq Mean Sq F value Pr(>F)
Geschlecht  1  2.133  2.1333   2.265  0.171
Residuals   8  7.533  0.9417

Error: Vpn:Zeit
      Df Sum Sq Mean Sq F value Pr(>F)
Zeit      2    0.2    0.1   0.002  0.998
Geschlecht:Zeit  2 1550.9  775.4  18.132 7.72e-05 ***
Residuals  16  684.3   42.8

```

Alternativ kann auch die Funktion `art3.anova` (vgl. Anhang 3.11) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`winer518t`). Eingabe und Ausgabe:

```
art3.anova(score~Geschlecht*Zeit+Error(Vpn/Zeit), winer518t)
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
Geschlecht  1  53.33  53.33  0.6079  0.458
Zeit      2  698.60  349.30 22.3582 2.325e-05 ***
Geschlecht:Zeit  2 1550.87  775.43 18.1317 7.716e-05 ***

```

Der Unterschied für das Ergebnis der Haupteffekte im Vergleich zur vorigen Tabelle liegt darin begründet, dass bei der Funktion `art3.anova` für die Haupteffekte die Ergebnisse aus der Analyse mit dem RT-Verfahren eingesetzt werden.

Zur Anwendung des ART+INT-Verfahrens müssen die Ränge  $e_k$  in normal scores  $n_{sek}$  transformiert werden, wozu vor der Varianzanalyse noch einzufügen ist:

```
nsek<-qnorm(ek/(n+1))
```

mit folgender Ausgabe:

```

Error: Vpn
      Df Sum Sq Mean Sq F value Pr(>F)
Geschlecht  1 0.00004 0.000044  0.005  0.948
Residuals   8 0.07763 0.009703

Error: Vpn:Zeit
      Df Sum Sq Mean Sq F value Pr(>F)
Zeit      2  0.007  0.003  0.006 0.993613
Geschlecht:Zeit  2 16.044  8.022 15.609 0.000174 ***
Residuals  16  8.223  0.514

```

### mit SPSS:

Wie bei der Durchführung der Rank transform-Tests muss zunächst der Datensatz umstrukturiert werden, wobei die Messwiederholungen in Fälle gewandelt werden. Dies wurde bereits in Kapiteln 5.3.3 sowie 6.3 durchgeführt. Für die Berechnung der Residuen  $e_{jm}$  wird hier der o.a. „naive approach“ gewählt.

Über Aggregate werden nun die Mittelwerte für Personen (mp), Zeit (mb) und gesamt (mm) berechnet und in der Arbeitsdatei ergänzt, um die Effekte von den Werten der Kriteriumsvariablen `score` abzuziehen und das Ergebnis in Ränge umzurechnen:



```

Varstocases
  /Id=Vpn
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.

Aggregate
  /outfile=* mode=addvariables
  /break=Vpn      /mp=mean(score) .
Aggregate
  /outfile=* mode=addvariables
  /break=Zeit    /mb=mean(score) .
Aggregate
  /outfile=* mode=addvariables
  /break=        /mm=mean(score) .

Compute ek = score - (mp + mb - mm) .

Rank variables=ek (A) /rank into rek.
execute.

```

Anschließend wird der Datensatz wieder in die ursprüngliche Form transformiert:

```

Sort cases by Vpn Zeit.
Casestovars
  /Id=Vpn
  /index=Zeit
  /groupby=variable.

```

Schließlich wird dann für *rek*, die im umstrukturierten Datensatz die Namen *rek.1*, *rek.2*, ... hat, eine Varianzanalyse mit Messwiederholungen mit den Faktoren *Geschlecht* und *Zeit* gerechnet:

```

GLM rek.1 rek.2 rek.3 by Geschlecht
  /wsfactor=Zeit 3 Polynomial
  /wsdesign=Zeit
  /design=Geschlecht.

```

Nachfolgend die Anova-Tabelle der Variablen *rek.1* ... für den bereinigten Test der Interaktion, wobei nur die Zeilen „Sphärizität angen.“ relevant sind. Demnach ist die Signifikanz der Interaktion gesichert.

Tests der Innersubjekteffekte						
Quelle		Quadrat summe	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	,200	2	,100	,002	,998
Zeit * Geschlecht	Sphärizität angen.	1550,867	2	775,433	18,132	,000
Fehler(Zeit)	Sphärizität angen.	684,267	16	42,767		

Zur Anwendung des ART+INT-Verfahrens müssen die nach dem ART-Verfahren errechneten Ränge in normal scores (vgl. Kapitel 2.3) transformiert werden. Dazu ist vor der Rücktransformation der Datenmatrix noch die Ermittlung des  $N(n_c)$  sowie die Transformation mittels der inversen Normalverteilung erforderlich, hier allerdings nur für die Prüfung der Interaktion vorgestellt:

```

Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(score).
Compute nsek=Idf.normal(rek/(nc+1),0,1).
execute.

```

Nach der Rückwandlung in das „normale“ Datenformat resultieren daraus die normal scores `nsek.1`, `nsek.2`, `nsek.3` und bringen folgende Ergebnistabelle (nur für die Interaktion):

Quelle		Quadrat- summe	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	,007	2	,003	,006	,994
	Huynh-Feldt	,007	2,000	,003	,006	,994
Zeit * Geschlecht	Sphärizität angenommen	16,044	2	8,022	15,609	,000
	Huynh-Feldt	16,044	2,000	8,022	15,609	,000
Fehler(Zeit)	Sphärizität angenommen	8,223	16	,514		
	Huynh-Feldt	8,223	16,000	,514		

### 6. 7. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren

Der Gruppierungsfaktor wird mit A, die beiden Messwiederholungsfaktoren mit C und D bezeichnet. Die Schritte im Einzelnen:

- Durchführung einer (normalen) Anova mit Haupt- und Interaktionseffekten für die Ränge  $R_x$  der Kriteriumsvariablen  $x$  (vgl. Kapitel 5.4.2). Hieraus werden nur die Haupteffekte verwendet. Für die Haupteffekte der Messwiederholungsfaktoren C und D können allerdings auch die bereinigten Tests wie in Kapitel 5.4.4 errechnet werden.
- Die Interaktion C\*D, ein reiner Messwiederholungseffekt, wird mit der ART wie in Kapitel 5.4.4 ermittelt, wobei Faktor A außer Acht gelassen wird.
- Für die Interaktion A\*C, ein gemischter Interaktionseffekt, werden die Werte der Kriteriumsvariablen  $x$  über die Stufen von Faktor D gemittelt (oder summiert), um mit diesen Werten die ART wie im vorhergehenden Kapitel 6.5.1 durchzuführen.
- Für die Interaktion A\*D ist das Verfahren analog der Interaktion A\*C durchzuführen.

Ein bereinigter Test für die 3er Interaktion A\*C\*D ist kein entsprechendes Verfahren bekannt.

Als Beispiel soll nachfolgend der bereits verwendete Datensatz 5 dienen:

- Die Haupteffekte `Medikament` und `Aufgabe` wurden bereits mit dem Rank transform-Test in Kapitel 5.4.2 ermittelt. Dort ist es kein Problem, auch den Faktor `Geschlecht` miteinzubeziehen.
- Der Interaktionseffekt `Medikament*Aufgabe` wurde in Kapitel 5.4.4 ermittelt.
- Bleiben noch die Interaktionen `Geschlecht*Medikament` und `Geschlecht*Aufgabe`, von denen nur die erste hier behandelt wird, da das Verfahren für beide identisch ist.

#### mit R:

Ausgangsbasis ist der in Kapitel 5.1.2 erstellte und in 5.4.4 verwendete Dataframe `mydata5t`. Zunächst werden mittels `aggregate` die Summen von `Fehler` über die 3 Aufgabenstufen berechnet. Dabei entsteht ein neuer Dataframe (`mydata5s`) mit den Mittelwerten als Variable  $x$ .

```

      Vpn Geschlecht Medikament      x
1      1          1           1 2.3333333
2      2          1           1 0.6666667
3      3          1           1 4.0000000
4      4          1           1 3.3333333
5      5          2           1 1.6666667
6      6          2           1 1.6666667
7      7          2           1 2.0000000
8      8          2           1 1.3333333
9      1          1           2 3.3333333
10     2          1           2 2.3333333
..     ..          ..          ..

```

Für die Berechnung der Residuen  $e_{jm}$  (vgl. vorigen Abschnitt) müssen zunächst die Effekte  $\mu_j$  (mb) und  $\pi_m$  (mp) sowie der Gesamtmittelwert (mm) berechnet werden, um diese von der Kriteriumsvariablen  $x$  abzuziehen. Diese werden dann nach Rundung auf 6 Stellen in Ränge transformiert, um darauf die Varianzanalyse anzuwenden. Hierfür wird diesmal wieder `ezANOVA` verwendet, wobei zu beachten ist, dass alle verwendeten Variablen Teil des angegebenen Dataframes sein müssen. D.h. in diesem Fall muss die neu erzeugte Variable `ez` mit `cbind` angehängt werden.

```

library(ez)
mydata5s <- aggregate(mydata5t$Fehler,
                      mydata5t[,c("Vpn", "Geschlecht", "Medikament")], mean)
attach(mydata5s)
mb <- tapply(x, Medikament, mean)
mp <- tapply(x, Vpn, mean)
mm <- mean(x)
ek <- x
n <- dim(mydata5s)[1]
for (k in 1:n) {j=Medikament[k]; i=Vpn[k]
  ek[k] <- ek[k] - mb[j] - mp[i] + mm }
ek <- rank(round(ek, digits=6))
ezANOVA(cbind(mydata5s, ek), ek, Vpn,
        within=. (Medikament), between=. (Geschlecht))$ANOVA

```

Das Ergebnis für die Interaktion ist nicht signifikant. Hätte man sich diese Interaktion beim Rank transform-Test (RT) angeschaut, hätte man sich die Durchführung des ART hierfür sparen können.

	Effect	DFn	DFd	F	p	p < .05
2	Geschlecht	1	6	0.14555256	0.7159674	
3	Medikament	2	12	0.04571522	0.9554795	
4	Geschlecht:Medikament	2	12	0.62084221	0.5538958	

Zur Anwendung des ART+INT-Verfahrens müssen die Ränge `ek` in normal scores `nsek` transformiert werden, wozu vor der Varianzanalyse noch einzufügen ist:

```
nsek <- qnorm(ek / (n+1))
```

mit folgender Ausgabe:

	Effect	DFn	DFd	F	p	p < .05
2	Geschlecht	1	6	0.53076731	0.4937263	
3	Medikament	2	12	0.03085359	0.9696942	
4	Geschlecht:Medikament	2	12	0.50722075	0.6145175	

**mit SPSS:**

Wie bei der Durchführung der Rank transform-Tests muss zunächst der Datensatz umstrukturiert werden, wobei die Messwiederholungen in Fälle gewandelt werden. Dies wurde bereits in Kapitel 5.4.2 einmal durchgeführt und in 5.4.4 wieder verwendet. Zunächst werden mittels `aggregate` die Mittelwerte von `Fehler` über die 3 Aufgabenstufen berechnet. Die Syntax dafür sowie ein Ausschnitt der Ergebnismatrix (`mydata5s`):

```
Varstocases
  /Id=Vpn
  /make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
  /index=Medikament(3) Aufgabe(3)
  /keep=Geschlecht
  /null=keep.

Dataset Declare mydata5s.
Aggregate
  /outfile='mydata5s'
  /break=Vpn Geschlecht Medikament
  /MFehler=mean(Fehler) .
```

	id	Geschlecht	Medikament	MFehler
1	1	1	1	2,33
2	1	1	2	3,33
3	1	1	3	4,00
4	2	1	1	,67
5	2	1	2	2,33
6	2	1	3	3,33
7	3	1	1	4,00
8	3	1	2	3,67
9	3	1	3	4,33
10	1	1	1	3,33

Über `Aggregate` werden nun die Mittelwerte für Personen (`pi`), Zeit (`bj`) und gesamt (`mm`) berechnet, um die Effekte von den Werten der Kriteriumsvariablen `MFehler` abzuziehen und das Ergebnis in Ränge umzurechnen. Die Anweisungen hierfür sind weitgehend identisch mit denen des vorigen Abschnitts. Lediglich `score` ist durch `MFehler` zu ersetzen.

Anschließend wird der Datensatz mit den Anweisungen wie im vorigen Abschnitt wieder in die ursprüngliche Form transformiert.

```
Aggregate
  /outfile=* mode=addvariables
  /break=Vpn /pi=mean(MFehler) .
Aggregate
  /outfile=* mode=addvariables
  /break=Medikament /bj=mean(MFehler) .
Aggregate
  /outfile=* mode=addvariables
  /break= /mm=mean(MFehler) .

Compute ek = MFehler - (pi + bj - mm) .
Rank variables=ek (A) /rank into rek.
execute.

Sort cases by Vpn Medikament .
```

```

Casestovars
  /Id=Vpn
  /index=Medikament
  /groupby=variable.

```

Schließlich wird dann für `rek`, die im umstrukturierten Datensatz die Namen `rek.1`, `rek.2`, ... hat, eine Varianzanalyse mit Messwiederholungen mit den Faktoren `Geschlecht` und `Medikament` gerechnet (Anweisungen siehe voriger Abschnitt). Nachfolgend die Anova-Tabelle für den bereinigten Test der Interaktion, wobei nur die Zeilen „Sphärität angenommen“ relevant sind. Demnach liegt für die Interaktion keine Signifikanz vor.

Tests der Innersubjekteffekte						
Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärität angen.	7,750	2	3,875	,046	,955
Medikament * Geschlecht	Sphärität angen.	110,583	2	55,292	,656	,537
Fehler(Medikament)	Sphärität angen.	1012,167	12	84,347		

Zur Anwendung des ART+INT-Verfahrens müssen die nach dem ART-Verfahren errechneten Ränge in normal scores (vgl. Kapitel 2.3) transformiert werden. Dazu ist vor der Rücktransformation der Datenmatrix in das „normale“ Format noch die Ermittlung des  $N(n_c)$  sowie die Transformation mittels der inversen Normalverteilung erforderlich, hier allerdings nur für die Prüfung der Interaktion vorgestellt:

```

Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(MFehler) .
Compute nsek=Idf.normal(rek/(nc+1),0,1) .
execute.

```

mit folgenden Ergebnissen für die Interaktion:

Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärität angen.	,088	2	,044	,031	,970
	Huynh-Feldt	,088	1,357	,065	,031	,922
Medikament * Geschlecht	Sphärität angen.	1,511	2	,756	,532	,600
	Huynh-Feldt	1,511	1,357	1,114	,532	,540
Fehler(Medikament)	Sphärität angen.	17,032	12	1,419		
	Huynh-Feldt	17,032	8,140	2,092		

### 6.7.3 Zwei Gruppierungs- und ein Messwiederholungsfaktor

Die Gruppierungsfaktoren werden mit A und B, der Messwiederholungsfaktor mit C bezeichnet, die Effekte mit  $\alpha_i$ ,  $\beta_j$  bzw.  $\gamma_l$ . Die Schritte im Einzelnen:

- Durchführung einer (normalen) Anova mit Haupt- und Interaktionseffekten für die Ränge  $R_x$  der Kriteriumsvariablen  $x$  (vgl. Kapitel 5.4.2). Hieraus werden nur die Haupteffekte verwendet.
- Für die Interaktion  $A*B$ , ein Effekt ohne Messwiederholungen, werden die Werte der Kriteriumsvariablen  $x$  über die Stufen von Faktor C gemittelt (oder summiert), um mit diesen Werten die ART wie im Kapitel 4.3.6 durchzuführen.

- Für die Interaktionen A\*C und B\*C sind zunächst die Residuen  $e_m$  des kompletten Modells zu berechnen (vgl. Kapitel 6.2).
- Für die Interaktion A\*C sind zu den Residuen zunächst der Interaktionseffekt zu addieren und danach der Messwiederholungseffekt  $\gamma_l$  sowie der Personeneffekt  $\pi_m$  zu subtrahieren:

$$e_m(a) = e_m + \bar{ac}_{il} - (\bar{p}_m + \bar{c}_l - \bar{x})$$

- Für die Interaktion B\*C wird analog A\*C vorgegangen.

$$e_m(b) = e_m + \bar{bc}_{jl} - (\bar{p}_m + \bar{c}_l - \bar{x})$$

wobei  $\bar{c}_l$ ,  $\bar{ac}_{il}$ ,  $\bar{bc}_{jl}$  die Mittelwerte von C, AC bzw. BC und  $\bar{p}_m$ ,  $\bar{x}$  die Personenmittelwerte bzw. das Gesamtmittel sind.

- Umrechnung der so errechneten Residuen  $e_m(a)$  sowie  $e_m(b)$  in Ränge.
- Durchführung einer Anova mit Haupt- und Interaktionseffekten jeweils mit den Rängen  $R(e_m(a))$  bzw.  $R(e_m(b))$ , aus der dann der jeweilige Interaktionseffekt abgelesen werden kann.

Ein bereinigter Test für die 3er Interaktion A\*B\*C liegt kein entsprechendes Verfahren vor.

Das Verfahren soll am Datensatz 6 (`winer568`) demonstriert werden. Die Anova-Tabelle der 3-faktoriellen Varianzanalyse für Rx, aus der die Haupteffekte A, B, und Zeit abzulesen sind:

	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

D.h. die Haupteffekte B und Zeit sind signifikant, insbesondere aber auch die Interaktionen A\*Zeit sowie B\*Zeit, die nun mittels dem ART gesondert berechnet werden. Zur Demonstration soll allerdings auch die Interaktion A\*B untersucht werden, wenn dies auch nicht erforderlich ist.

### mit R:

Als Basis muss wieder der umstrukturierte Dataframe `winer568t` aus Kapitel 5.1.2 genommen werden. Damit werden für die Analyse der Interaktionen A\*C und B\*C die Residuen (`em`) des Modells  $A*B*C+V_{pn}$  ermittelt:

```
em <- aov(x~A*B*Zeit+Vpn, winer568t)$residuals
```

Anschließend werden die Effekte für die beiden untersuchten Interaktionen (`mac` bzw. `mbc`), die Zeit (`mc`) sowie den Personeneffekt `mv` ausgerechnet und gemäß o.a. Formel mit den Residuen `em` verrechnet, um schließlich für die bereinigten Werte für A\*Zeit (`ema`) und B\*Zeit (`emb`) eine Varianzanalyse durchzuführen:

```
attach(winer568t)
mc <- tapply(x, Zeit, mean)
mv <- tapply(x, Vpn, mean)
mac <- tapply(x, winer568t[, c("A", "Zeit")], mean)
```

```

mbc <- tapply(x, winer568t[, c("B", "Zeit")], mean)
mm <- mean(x)
n <- dim(winer568t)[1]
ema <- em
emb <- em
for (m in 1:n) {ia=A[m]; ib=B[m]; ic=Zeit[m]; vm=Vpn[m]
  ema[m] <- ema[m] + mac[ia,ic] - mc[ic] -mv[vm] + mm
  emb[m] <- emb[m] + mbc[ib,ic] - mc[ic] -mv[vm] + mm }
rema<-rank(round(ema,digits=7))
remb<-rank(round(emb,digits=7))
library(ez)
ezANOVA(cbind(winer568t, rema), rema, Vpn,
        between=. (A,B), within=. (Zeit))
ezANOVA(cbind(winer568t, remb), remb, Vpn,
        between=. (A,B), within=. (Zeit))

```

Bei der Varianzanalyse für `rema` (bereinigte Interaktion A\*Zeit) zeigt der Mauchly-Test auf Varianzhomogenität mit  $p=0,029$  eine signifikante Abweichung an. Aber unabhängig davon ist vorsichtshalber in der Anova-Ausgabe die Signifikanz im Teil `Sphericity Corrections` und dort unter „p[HF]“ (Huynh-Feldt-korrigiert) abzulesen, allerdings *ausschließlich* für die Interaktion A\*Zeit (auf die Tabelle für `ekb` wird hier verzichtet). Der p-Wert (0,00006) bestätigt den oben mit dem RT-Test errechneten Einfluss von A\*Zeit:

\$`Sphericity Corrections`					
	Effect	GGe	p [GG]	HFe	p [HF]
5	Zeit	0.4925664	0.9606032485	0.5774698	0.9751581
6	A:Zeit	0.4925664	0.0001875066	0.5774698	0.0000645
7	B:Zeit	0.4925664	0.7383084419	0.5774698	0.7730265
8	A:B:Zeit	0.4925664	0.8874259252	0.5774698	0.9150948

Nun zur Interaktion A\*B.

- Zunächst werden mittels `aggregate` die Summen von  $v_1, \dots, v_4$  über die 4 Zeitstufen berechnet. Dabei entsteht ein neuer Dataframe (`winer568s`) mit den Mittelwerten als Variable `x`.
- Wie in Kapitel 4.3.6 werden die Effekte `mab` (Interaktion), `ma` (Faktor A) sowie `mb` (Faktor B) errechnet.
- Ermittlung der Residuen `em` der Varianzanalyse des Modells A\*B,
- Addition bzw. Subtraktion der vorher errechneten Effekte von `em`,
- Durchführung der Varianzanalyse für `em` zur Kontrolle des Effekts A\*B:

```

winer568s <- aggregate(winer568t$x, winer568t[, c("Vpn", "A", "B")], mean)
attach(winer568s)
ma <- tapply(x, A, mean)
mb <- tapply(x, B, mean)
mab <- tapply(x, list(A, B), mean)
mm <- mean(x)
em <- aov(x~A*B, winer568s)$residuals
n <- dim(winer568s)[1]
for (m in 1:n) {ia=A[m]; ib=B[m]
  em[m] <- em[m] + mab[ia,ib] - ma[ia] - mb[ib] + mm }
rem <- rank(em)
summary(aov(rem~A*B, winer568s))

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	2.08	2.083	0.122	0.736
B	1	0.33	0.333	0.020	0.892
A:B	1	0.75	0.750	0.044	0.839
Residuals	8	136.33	17.042		

Alternativ ist auch hier - wie bereits in Kapitel 6.5.1 - das ART-Verfahren mit der Funktion `art3.anova` (vgl. Anhang 3.11) bequem durchführbar. Basis ist auch hier der umstrukturierte Datensatz `winer568t`. Nachfolgend Eingabe und Ausgabe:

```
art3.anova(x~A*B*Zeit+Error(Vpn/Zeit),winer568t)
```

	Df	Sum of Sq	F value	Pr(>F)
A	1	18.8	3.2609	0.108588
B	1	75.0	13.0435	0.006866 **
A:B	1	0.7	0.0440	0.839079
Zeit	3	6637.2	235.4229	< 2.2e-16 ***
A:Zeit	3	3528.8	22.7165	3.421e-07 ***
B:Zeit	3	1764.9	6.8443	0.001714 **

Auch hier müssen wieder zur Anwendung des ART+INT-Verfahrens im ersten Teil die Ränge `rema` und `remb` sowie im zweiten Teil die Ränge `rem` in normal scores `nsema` und `nsemb` bzw. `nsem` transformiert werden, wozu vor den Varianzanalysen noch jeweils einzufügen ist:

```
nsema<-qnorm(rema/(n+1))
nsemb<-qnorm(remb/(n+1))
```

bzw.

```
nsem<-qnorm(rem/(n+1))
```

Auf die Ausgabe wird hier verzichtet und auf die nachfolgenden SPSS-Ergebnisse verwiesen.

Alternativ kann die Analyse auf Basis des ART+INT-Verfahrens für alle Effekte auch bequem mittels der Funktion `art3.anova` durchgeführt werden:

```
art3.anova(x~A*B*Zeit+Error(Vpn/Zeit),winer568t,INT=T,main=T)
```

	Df	Sum of Sq	F value	Pr(>F)
A	1	1.0	1.5901	0.242840
B	1	2.3	5.4003	0.048626 *
A:B	1	0.1	0.0793	0.785435
Zeit	3	6637.2	235.4229	< 2.2e-16 ***
A:Zeit	3	17.0	25.3226	1.303e-07 ***
B:Zeit	3	8.7	7.7526	0.000864 ***

**mit SPSS:**

Zunächst muss wieder der Datensatz aus Beispiel 6 (`winer568`) wie in Kapitel 6.2 umstrukturiert werden, wobei `vpn` die `Vpn`-Kennzeichnung ist. Anschließend werden für die Analyse der Interaktionen `A*C` und `B*C` die Residuen (Variable `Res_1`) des Modells (ohne Messwiederholungen) `A*B*C+Vpn` ermittelt:



```

Varstocases
  /id=Vpn
  /make score from v1 v2 v3 v4
  /index=Zeit(4)
  /keep=A B
  /null=keep.

Unianova x by Vpn A B Zeit
  /Save=resid
  /design=A*B*Zeit Vpn.

```

Anschließend werden die Effekte für die beiden untersuchten Interaktionen ( $mac$  bzw.  $mbc$ ), die Zeit ( $mc$ ) sowie den Personeneffekt  $mv$  ausgerechnet, der Arbeitsdatei angehängt und gemäß o.a. Formel mit den Residuen  $em$  verrechnet, um schließlich für  $ema$  und  $emb$  eine Varianzanalyse durchzuführen:

```

Aggregate
  /outfile=* mode=addvariables
  /break=Vpn /mp=mean(score).
Aggregate
  /outfile=* mode=addvariables
  /break=Zeit /mc=mean(score).
Aggregate
  /outfile=* mode=addvariables
  /break=A Zeit /mac=mean(score).
Aggregate
  /outfile=* mode=addvariables
  /break=B Zeit /mbc=mean(score).
Aggregate
  /outfile=* mode=addvariables
  /break= /mm=mean(score).
Compute ema = res_1 + mac - (mp + mc - mm).
Compute emb = res_1 + mbc - (mp + mc - mm).
Rank variables=ema (A) /rank into rema.
Rank variables=emb (A) /rank into remb.
execute.

```

Nun wird wie Kapitel 6.3 der Datensatz in die ursprüngliche Form zurücktransformiert:

```

Casestovars
  /Id=Vpn
  /Index=Zeit
  /Groupby=variable.

```

Dabei werden aus den zu analysierenden Rängen von  $eka$  und  $ekb$  die Messwiederholungsvariablen  $rema.1, \dots, rema.4$  bzw.  $remb.1, \dots, remb.4$ . Bei der Varianzanalyse mit Messwiederholungen für  $rema$  zeigt der Mauchly-Test mit  $p=0,027$  eine signifikante Abweichung von der Varianzhomogenität. Vorsichtshalber sollte in jedem Fall die Signifikanz des Effekts in der Zeile „Huynh-Feldt“ abgelesen werden. Allerdings kann aus der Tabelle *ausschließlich* der Effekt  $A*Zeit$  entnommen werden. Der p-Wert ( $< 0,001$ ) bestätigt den oben mit dem RT-Test errechneten Einfluss von  $A*Zeit$ . (Auf die Ausgabe für  $remb$  wird hier verzichtet):

Tests der Innersubjekteffekte						
Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angen.	2,250	3	,750	,015	,998
	Huynh-Feldt	2,250	2,412	,933	,015	,993
Zeit * A	Sphärizität angen.	3552,083	3	1184,028	23,039	,000
	Huynh-Feldt	3552,083	2,412	1472,632	23,039	,000
Zeit * B	Sphärizität angen.	38,167	3	12,722	,248	,862
	Huynh-Feldt	38,167	2,412	15,823	,248	,821
Zeit * A * B	Sphärizität angen.	12,083	3	4,028	,078	,971
	Huynh-Feldt	12,083	2,412	5,010	,078	,950

Zur Anwendung des ART+INT-Verfahrens müssen die nach dem ART-Verfahren errechneten Ränge in normal scores (vgl. Kapitel 2.3) transformiert werden. Dazu ist vor der Rücktransformation der Datenmatrix in das „normale“ Format noch die Ermittlung des  $N(nc)$  sowie die Transformation mittels der inversen Normalverteilung erforderlich:

```
Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(score).
Compute nsema=Idf.normal(rema/(nc+1),0,1).
Compute nsemb=Idf.normal(remb/(nc+1),0,1).
execute.
```

mit folgenden Ergebnissen:

Quelle		Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zeit	Sphärizität angenommen	,018	3	,006	,028	,994
	Huynh-Feldt	,018	1,938	,010	,028	,970
Zeit * A	Sphärizität angenommen	17,131	3	5,710	25,584	,000
	Huynh-Feldt	17,131	1,938	8,838	25,584	,000
Zeit * B	Sphärizität angenommen	,011	3	,004	,016	,997
	Huynh-Feldt	,011	1,938	,006	,016	,982
Zeit * A * B	Sphärizität angenommen	,014	3	,005	,021	,996
	Huynh-Feldt	,014	1,938	,007	,021	,977

Nun zur Interaktion A\*B. Ausgangsbasis ist die oben im ersten Schritt erzeugte umstrukturierte Arbeitsdatei. Zunächst werden mittels `aggregate` die Summen von  $v_1, \dots, v_4$  über die 4 Zeitstufen berechnet. Dabei muss eine neue Datei mit den Mittelwerten als Variable `mx` angelegt werden.

```
Dataset Declare winer568s.
Aggregate      /outfile='winer568s'
  /break=Vpn A B /mx=MEAN(x).
```

Ermittlung der Residuen (Variable `RES_1`) der Varianzanalyse des Modells A\*B:

```
Unianova mx by A B
  /Save=resid
  /design=A*B.
```

Wie in Kapitel 4.3.6 werden die Effekte  $mab$  (Interaktion),  $ma$  (Faktor A) sowie  $mb$  (Faktor B) errechnet. Anschließend Addition bzw. Subtraktion der vorher errechneten Effekte von  $res\_1$ :

```
Aggregate      /outfile=*   mode=addvariables
               /break=A B   /mab=mean(mx) .
Aggregate      /outfile=*   mode=addvariables
               /break=A     /ma=mean(mx) .
Aggregate      /outfile=*   mode=addvariables
               /break=B     /mb=mean(mx) .
Aggregate      /outfile=*   mode=addvariables
               /break=      /mm=mean(mx) .

Compute em = res_1 + mab - (ma + mb - mm) .
Rank variables=ek (A) /rank into rem.
execute.
```

Durchführung der Varianzanalyse für  $em$  zur Kontrolle des Effekts  $A*B$ , wonach die Interaktion  $A*B$  nicht signifikant ist.

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
A	2,083	1	2,083	,122	,736
B	,333	1	,333	,020	,892
A * B	,750	1	,750	,044	,839
Fehler	136,333	8	17,042		

Für die Durchführung des ART+INT-Verfahrens müssen die oben im letzten Schritt errechneten Ränge  $rem$  in normal scores transformiert werden:

```
Aggregate
  /outfile=* mode=addvariables
  /break= /nc=NU(mx) .
Compute nsem=Idf.normal (rem/ (nc+1) , 0, 1) .

Unianova nsem by A B
  /design=A B A*B.
```

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
A	,085	1	,085	,092	,769
B	,004	1	,004	,004	,950
A * B	,073	1	,073	,079	,785
Fehler	7,366	8	,921		

## 6. 8 ATS-Tests von Akritas, Arnold & Brunner

Den von Akritas, Arnold und Brunner entwickelten ATS-Test gibt es auch für mehrfaktorielle Varianzanalysen mit gemischten Designs. Während in R dazu die Pakete `nparLD` und `MANOVA.RM` zur Verfügung stehen gibt es in SPSS derzeit keine Möglichkeit zur Anwendung dieses Verfahrens.

**mit R:**

Die 2-faktorielle Analyse soll ebenfalls am Datensatz des Beispiels 4 gezeigt werden. Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte umstrukturierte Dataframe `winer518t`. Zunächst mittels `nparLD`, das schon in Abschnitten 5.3.6 und 5.4.6 vorgestellt worden war. Die Analyse kann mittels zwei Funktionen erfolgen:

- `nparLD` ist eine universelle Funktion für alle verarbeitbaren Designs.
- `f1.ld.f1` erlaubt fehlende Werte bei den Messwiederholungen, gibt einen Mittelwertplot aus sowie eine Reihe weiterer, hier allerdings nicht interessierender Statistiken. (Darüber hinaus gibt es entsprechende Funktionen für 3-faktorielle Designs: `f2.ld.f1` für zwei Gruppierungs- und einen Messwiederholungsfaktor sowie `f1.ld.f2` für einen Gruppierungs- und zwei Messwiederholungsfaktoren.)

Beide geben sowohl die WTS als auch die ATS aus. Die Ausgabe unterscheidet sich nicht hinsichtlich der Wiedergabe dieser Statistiken. Nachfolgend zunächst die Eingabe für beide Varianten, wobei zu beachten ist, dass bei `nparLD` trotz Angabe des Dataframes die Variablennamen nicht automatisch gefunden werden. Daher muss bei beiden Funktionen entweder jeder Variablenname zusammen mit dem Dataframennamen in der üblichen Form, z.B. `winer518t$score` angegeben werden oder mit `with(winer518t, ...)` ausgeführt werden:

```
library(nparLD)
with(winer518t, nparLD(score~Geschlecht*Zeit, winer518t, Vpn))
with(winer518t, f1.ld.f1(score, Zeit, Geschlecht, Vpn,
  time.name="Zeit", group.name="Geschlecht")) -> ano
round(ano$ANOVA.test, 3)
```

Bei `f1.ld.f1` müssen die Faktoren zweimal angegeben werden: zum einen zur Identifikation des Faktors, zum anderen in "...“ als Name des Faktors für die Ausgabe. Diese Funktion gibt noch zusätzlich einen Interaktionsplot aus, allerdings der relativen Effekte (vgl. Kapitel 2.5) anstatt der Mittelwerte, da sich ja die Hypothesen auf erstere beziehen:

Die Ergebnisse von `nparLD`:

Wald-Type Statistic (WTS):			
	Statistic	df	p-value
Geschlecht	0.6079316	1	4.355677e-01
Zeit	40.2018842	2	1.863253e-09
Geschlecht:Zeit	36.3186594	2	1.298683e-08
ANOVA-Type Statistic (ATS):			
	Statistic	df	p-value
Geschlecht	0.6079316	1.000000	4.355677e-01
Zeit	22.3581811	1.972665	2.515147e-10
Geschlecht:Zeit	16.0426724	1.972665	1.281568e-07

Bei der Ausgabe von `f1.ld.f1` gibt es die Möglichkeit, einzelne Teile auszugeben, etwa die ATS- (Anova-) Tabelle (`..$ANOVA.test`) oder die WTS- (Wald-Test-) Tabelle (`..wald.test`). Dies hat den Vorteil, dass man über die Funktion `round` die Zahlendarstellung der Art `xxxe-nn` ändern kann.

	Statistic	df	p-value
Geschlecht	0.6079	1.0000	0.4356
Zeit	22.3582	1.9727	0.0000
Geschlecht:Zeit	16.0427	1.9727	0.0000

Nun zum Paket `MANOVA.RM`, das ebenfalls sowohl die WTS- als auch ATS-Tests durchführt. Allerdings basieren die Verfahren in `nparLD` und `MANOVA.RM` auf unterschiedlichen Modellen, wenngleich beide viele Gemeinsamkeiten haben. In `MANOVA.RM` werden die WTS-Tests mithilfe resampling-Verfahren errechnet, die zuverlässigere Ergebnisse liefern, dafür aber sehr rechenintensiv sind. Für dieses Verfahren gibt der Parameter `iter` die Anzahl der Resampling-Schritte an. Wird der WTS-Test nicht benötigt, kann dieser Parameter auf 1 gesetzt werden. Zur Analyse von repeated measures designs gibt es die Funktion `RM`. Diese beansprucht für sich, robust gegen heterogene Kovarianzmatrizen zu sein.

```
library(MANOVA.RM)
RM(score~Geschlecht*Zeit, winer518t, "Vpn", iter=1)
```

Wald-Type Statistic (WTS):				
	Test statistic	df	p-value	
Geschlecht	"0.472"	"1"	"0.492"	
Zeit	"35.621"	"2"	"<0.001"	
Geschlecht:Zeit	"46.241"	"2"	"<0.001"	
ANOVA-Type Statistic (ATS):				
	Test statistic	df1	df2	p-value
Geschlecht	"0.472"	"1"	"12.774"	"0.504"
Zeit	"22.051"	"1.87"	"Inf"	"<0.001"
Geschlecht:Zeit	"17.038"	"1.87"	"Inf"	"<0.001"

## 6.9 Bredenkamp Tests

Zunächst sei noch einmal darauf hingewiesen, dass die Tests von Bredenkamp (vgl. Lienert, 1987, S. 1024 ff und Bredenkamp, 1974) letztlich mit den Puri & Sen-Tests (vgl. u.a. Kapitel 6.4) identisch sind. Lediglich die Berechnung erfolgt auf einem anderen Weg. Insbesondere für SPSS-Benutzer können die Tests von Bredenkamp bei gemischten Versuchsplänen von Nutzen sein, da zum einen die ATS aus dem vorigen Kapitel nicht zur Verfügung stehen und zum anderen bei diesen Tests keine Umstrukturierungen der Daten erforderlich sind. Daher werden nachfolgend Beispiele nur mit SPSS gerechnet. Hinzu kommt, dass es in R erhebliche Schwierigkeiten bereitet, Friedman-Tests für Teildatensätze durchzuführen, was bei den Bredenkamp Tests erforderlich ist.

Für das Prozedere werden die einzelnen Versuchspläne unterschieden. Im ersten Fall der 2-faktoriellen Analyse wird das Grundprinzip gezeigt, in den 3-faktoriellen Analysen dann erweitert.

### 6.9.1 Ein Gruppierungs- und ein Messwiederholungsfaktor

Die drei Effekte (Gruppierungsfaktor A, Messwiederholungsfaktor B sowie die Interaktion) werden wie folgt überprüft:

- Haupteffekt A:  
pro Erhebungseinheit (z.B. Versuchsperson) wird die Summe aller Messwiederholungen errechnet. Hierauf wird dann der Kruskal-Wallis-H-Test angewandt.

- Haupteffekts B: ein Friedman-Test wird über die Messwiederholungen durchgeführt, wobei die Gruppeneinteilung durch den Faktor A ignoriert wird.
- Interaktion:

$\chi^2$ -Testwerte	Freiheitsgrade
$\chi^2_B(A_1)$	$J-1$
$+ \chi^2_B(A_2)$	$J-1$
$+ \dots$	$\dots$
$+ \chi^2_B(A_k)$	$J-1$
$- \chi^2_B$	$J-1$
$Summe(\chi^2_B(A_i)) - \chi^2_B$	$(I-1)(J-1)$

Unter Ausnutzung der Additivität der  $\chi^2$ -Werte wird für jede Stufe des Faktors A ein Friedman-Test für B errechnet, die resultierenden  $\chi^2$ -Werte aufsummiert, davon der  $\chi^2$ -Wert des Friedman-Tests des Haupteffekts B abgezogen und schließlich der Restwert anhand der Tabelle der  $\chi^2$ -Verteilung auf Signifikanz überprüft:

#### mit SPSS:

Es wird wieder der Datensatz 4 (winer518) benutzt. Zunächst muss das Skalenniveau der Variablen  $t_1, t_2, t_3$  auf „Skala“ gesetzt werden, anschließend deren Summe  $t_{sum}$  errechnet, damit der Kruskal-Wallis-Test zum Test des Geschlechtseffekts sowie der Friedman-Test für  $t_1, t_2, t_3$  zum Test des Zeiteffekts durchgeführt werden können. Danach wird wiederum der Friedman-Test durchgeführt, allerdings dann mittels `split file` für die beiden Geschlechtsgruppen getrennt.

```
compute tsum=t1+t2+t3.
Nptests
  /Independent test (tsum) group (Geschlecht) kruskal_wallis.
Nptests
  /Related test(t1 t2 t3) friedman.

Sort cases by Geschlecht.
Split File separate by Geschlecht.
Nptests
  /Related test(t1 t2 t3) friedman.
```

Die Ausgabe zeigt zunächst links den K-W-Test (für „Geschlecht“), rechts den Friedman-Test (für „Zeit“):

<b>Gesamtanzahl</b>	10
<b>Teststatistik</b>	,099
<b>Freiheitsgrade</b>	1
<b>Asymptotische Sig. (zweiseitiger Test)</b>	,753

<b>Gesamtanzahl</b>	10
<b>Teststatistik</b>	9,556
<b>Freiheitsgrade</b>	2
<b>Asymptotische Sig. (zweiseitiger Test)</b>	,008

Die Teststatistiken ( $\chi^2$ -Werte) für die beiden Friedman-Test zur Ermittlung der Interaktion:

$$\chi^2_{\text{Zeit}} (\text{Männer}) = 9,333 \quad (2 \text{ Fg})$$

$$\chi^2_{\text{Zeit}} (\text{Frauen}) = 8,444 \quad (2 \text{ Fg})$$

Zieht man von der Summe 17,777 (4 Fg) den o.a.  $\chi^2_{\text{Zeit}}$  (gesamt) mit dem Wert 9,556 (2 Fg) ab, so erhält man  $\chi^2_{\text{Interaktion}} = 8,222$  mit 2 Fg. Die 5%-Schranke für die  $\chi^2$ -Verteilung liegt bei 6,0 für  $df=2$ , so dass die Interaktion als signifikant angesehen werden kann. Die Ergebnisse decken sich mit denen aus Kapitel 6.4.

### 6.9.2 Zwei Gruppierungs- und ein Messwiederholungsfaktor

Das Prinzip aus dem vorigen Abschnitt wird nun auf drei Faktoren erweitert. Allerdings können die Bredenkamp Tests nur für balancierte Versuchspläne (vgl. Kapitel 4.3.1.1) angewandt werden. Im Folgenden werden die beiden Gruppierungsfaktoren mit A und B (mit Gruppenzahl  $I$  bzw.  $J$ ), der Messwiederholungsfaktor mit C (mit Gruppenzahl  $K$ ) bezeichnet. Die Effekte werden wie folgt überprüft:

- Haupteffekte A und B:  
pro Erhebungseinheit (z.B. Versuchsperson) wird die Summe aller Messwiederholungen errechnet. Hierauf wird dann jeweils für A und B der Kruskal-Wallis-H-Test angewandt.
- Interaktion A\*B: Zunächst wird ein H-Test über alle Zellen hinweg gerechnet. Von diesem  $\chi^2$ -Wert werden die Werte aus den H-Tests für Faktor A und Faktor B subtrahiert. Das Ergebnis ist der  $\chi^2$ -Wert für die Interaktion A\*B. Analog werden die Freiheitsgrade ermittelt.
- Haupteffekt C: ein Friedman-Test wird über die Messwiederholungen durchgeführt, wobei die Gruppeneinteilung durch die Faktoren A und B ignoriert wird.
- Interaktion A\*C: Unter Ausnutzung der Additivität der  $\chi^2$ -Werte wird für jede Stufe des Faktors A ein Friedman-Test für C errechnet, die resultierenden  $\chi^2$ -Werte aufsummiert, davon der  $\chi^2$ -Wert des Friedman-Tests des Haupteffekts C abgezogen und schließlich der Restwert anhand der Tabelle der  $\chi^2$ -Verteilung auf Signifikanz überprüft.
- Interaktion B\*C: analog Interaktion A\*C.
- Interaktion A\*B\*C: für jede der  $I*J$  Zellen von A\*B wird ein Friedman-Test für C errechnet, die resultierenden  $\chi^2$ -Werte aufsummiert, davon die  $\chi^2$ -Werte des Haupteffekts C, der Interaktion A\*C sowie der Interaktion B\*C abgezogen und schließlich der Restwert anhand der Tabelle der  $\chi^2$ -Verteilung auf Signifikanz überprüft. Die dafür erforderlichen Freiheitsgrade errechnen sich analog zu den  $\chi^2$ -Werten.

#### mit SPSS:

Die Berechnungen sollen am Datensatz 6 (winer568) demonstriert werden. Zunächst muss das Skalenniveau der Variablen  $v_1, \dots, v_4$  auf „Skala“ gesetzt werden, anschließend deren Summe  $v_{\text{sum}}$  errechnet, womit zwei H-Tests zur Prüfung der Effekte A und B durchgeführt werden. Für  $v_1, \dots, v_4$  wird eine Friedman-Analyse zum Test des Zeiteffekts gerechnet. Hier empfiehlt es sich, die „alten“ Anweisungen für die nichtparametrischen Tests (Npar Tests) zu verwenden, da bei diesen die Ergebnisse „direkt“ im Ausgabefenster angezeigt werden und nicht erst über ein Doppelklick in einem separaten Fenster erscheinen.

```
compute vsum=v1+v2+v3+v4.
Npar tests
  /K-W = vsum by A.
```

```

Npar tests
  /K-W = vsum by B.
Npar tests
  /Friedman = v1 v2 v3 v4.

```

Nachfolgend werden nur die  $\chi^2$ -Werte protokolliert:

$$\chi^2_A = 1,468 \quad (1 \text{ Fg})$$

$$\chi^2_B = 5,872 \quad (1 \text{ Fg})$$

$$\chi^2_{\text{Zeit}} = 32,635 \quad (3 \text{ Fg})$$

Für die Interaktion A\*B wird zunächst eine Zellennummer  $Z_{\text{elle}}$  errechnet, für die Gruppierung dann ein H-Test bzgl.  $v_{\text{sum}}$  errechnet. Anschließend von der resultierenden Teststatistik die beiden o.a. Statistiken für A und B subtrahiert:

```

compute Zelle=(a-1)*2+b.
Npar tests  /K-W = vsum by Zelle.

```

$$\chi^2_{\text{Zellen}} = 7,399 \quad (3 \text{ Fg})$$

$$\chi^2_{A*B} = \chi^2_{\text{Zellen}} - \chi^2_A - \chi^2_B = 7,399 - 1,468 - 5,872 = 0,059 \quad (1 \text{ Fg})$$

Für die Interaktion A\*Zeit werden jeweils Friedman-Tests für die zwei Stufen von A errechnet, die resultierenden  $\chi^2$ -Werte addiert und davon der oben errechnete Wert  $\chi^2_{\text{Zeit}}$  subtrahiert:

```

Sort cases  by A.
Split File separate by A.
Npar tests
  /Friedman = v1 v2 v3 v4.

```

$$\chi^2_{\text{Zeit}}(A_1) = 16,932 \quad (3 \text{ Fg})$$

$$\chi^2_{\text{Zeit}}(A_2) = 17,357 \quad (3 \text{ Fg})$$

$$\chi^2_{A*\text{Zeit}} = \chi^2_{\text{Zeit}}(A_1) + \chi^2_{\text{Zeit}}(A_2) - \chi^2_{\text{Zeit}} = 16,932 + 17,357 - 32,635 = 1,654 \quad (3 \text{ Fg})$$

Analog erhält man für die Interaktion B\*Zeit:

$$\chi^2_{\text{Zeit}}(B_1) = 16,158$$

$$\chi^2_{\text{Zeit}}(B_2) = 16,966$$

$$\chi^2_{B*\text{Zeit}} = \chi^2_{\text{Zeit}}(B_1) + \chi^2_{\text{Zeit}}(B_2) - \chi^2_{\text{Zeit}} = 16,158 + 16,966 - 32,635 = 0,489 \quad (3 \text{ Fg})$$

Für die Interaktion A\*B\*Zeit werden zuerst für alle 4 Zellen von A\*B jeweils ein Friedman-Test für den Faktor Zeit gerechnet und die resultierenden Teststatistiken ( $\chi^2$ -Werte) addiert. Davon werden dann der oben errechnete Wert  $\chi^2_{\text{Zeit}}$  sowie die  $\chi^2$ -Werte der Interaktionen A\*Zeit und B\*Zeit subtrahiert:

```

Sort cases  by Zelle.
Split File separate by Zelle.
Npar tests
  /Friedman = v1 v2 v3 v4.

```

$$\chi^2_{\text{Zeit}}(A_1B_1) = 8,379 \quad (3 \text{ Fg})$$

$$\chi^2_{\text{Zeit}}(A_1B_2) = 9,000 \quad (3 \text{ Fg})$$



$$\chi^2_{\text{Zeit}}(A_2B_1) = 8,786 \quad (3 \text{ Fg})$$

$$\chi^2_{\text{Zeit}}(A_2B_2) = 8,786 \quad (3 \text{ Fg})$$

$$\begin{aligned} \chi^2_{A*B*Zeit} &= \chi^2_{\text{Zeit}}(A_1B_1) + \chi^2_{\text{Zeit}}(A_1B_2) + \chi^2_{\text{Zeit}}(A_2B_1) + \chi^2_{\text{Zeit}}(A_2B_2) \\ &\quad - \chi^2_{\text{Zeit}} - \chi^2_{A*Zeit} - \chi^2_{B*Zeit} \\ &= 8,379 + 9,000 + 8,786 + 8,786 - 1,654 - 0,489 - 32,635 \\ &= 0,172 \quad (3 \text{ Fg}) \end{aligned}$$

Die p-Werte für die drei Haupteffekte werden in SPSS ja ausgegeben. Die  $\chi^2$ -Werte für die Interaktionen müssen mit den tabellierten kritischen Werten verglichen werden. Die 5%-Schranke für die  $\chi^2$ -Verteilung liegt bei 3,8 bzw. 9,0 für  $df=1$  bzw.  $df=3$ , so dass keine Interaktion als signifikant angesehen werden kann. Vergleicht man diese Ergebnisse mit denen der ART-Methode (Kapitel 6.5.3), so zeigt sich deutlich, dass bei diesem Verfahren die Tests der Interaktionen relativ konservativ ausfallen.

## 6. 10 Verfahren ohne Sphäritäts-Voraussetzungen

Hierunter fallen zum einen die in Kapitel 5.2 kurz vorgestellten multivariaten Tests, die darauf basierenden nichtparametrischen Verfahren von Koch, Akritas & Arnold sowie von Agresti & Pendergast, ein komplett verteilungsfreies, auf sphärischen Rängen basierendes Verfahren sowie das Verfahren für nichthomogene Varianzen von Welch & James. Die multivariate Tests (z.B. Pillai und Hotelling-Lawley) waren bereits in Kapitel 5.3.9 für die 1-faktorielle Analyse vorgestellt worden. Bei gemischten Versuchsplänen wird allerdings dennoch die Homogenität der Kovarianzmatrizen, allerdings der Differenzen, gefordert, nicht jedoch die Sphärität. Darüber hinaus gehören die Methoden GEE und GLMM in diese Kategorie.

Die meisten der genannten Verfahren werden in der Literatur lediglich für 2-faktorielle gemischte Versuchspläne beschrieben. Gegebenenfalls kann man sich bei 3- oder mehrfaktoriellen Designs damit behelfen, jeweils einen Gruppierungs- und einen Messwiederholungsfaktor auszuwählen und das Verfahren darauf anzuwenden, da Hypothesen für 3er-Interaktionen eher seltener vorliegen. Bei der Auswahl eines von mehreren Messwiederholungsfaktoren müssen vorher die Summen über den/die anderen Messwiederholungsfaktoren gebildet und das ausgewählte Verfahren darauf angewandt werden. Einige der Verfahren basieren auf umfangreichen Matrizenrechnungen und sind daher mit SPSS nicht durchführbar. Für die Anwendung in R werden vom Autor entsprechende Funktionen bereitgestellt (vgl. Anhang 3). Lediglich der multivariate Test (u.a. Hotelling-Lawley) und das nichtparametrische Pendant können in SPSS durchgeführt werden. Alle Verfahren werden anhand des Datensatzes `winer568` vorgestellt.

### 6. 10. 1 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und nichtparametrisch

Nachfolgend werden die parametrischen multivariaten Methoden vorgestellt. Auch hier können alternativ die nichtparametrischen Pendants, die Tests von Agresti & Pendergast sowie von Akritas & Arnold, benutzt werden, die in Abschnitt 5.3.9 in einem Beispiel vorgestellt worden waren. Bei der Besprechung der Voraussetzungen in Kapitel 5.2 sowie in 5.3.9 wurde bereits darauf hingewiesen, dass der Test eine multivariate Normalverteilung der Residuen voraussetzt, und wie dies ersatzweise überprüft werden kann. Da die multivariaten Tests nur die Messwiederholungseffekte betreffen, wird für die Tests der Gruppierungsfaktoren der "normale" univariate parametrische F-Test durchgeführt. Als Datensätze dienen wieder `winer568` und `mydata5`. Die numerische Abweichung des Ergebnisses für den Faktor Zeit mit dem entsprechenden Ergebnis in Kapitel 5.3.9 erklärt sich durch die Hinzunahme des Gruppierungsfaktors A.

**mit R:**

Die multivariaten Tests werden u.a. über zwei Standardfunktionen angeboten, `manova` sowie `lm` für allgemeine lineare Modelle. In diesem Fall ist `lm` einfacher anzuwenden. Zu beachten ist, dass bei einer mehrfaktoriellen Analyse, wie im nachfolgenden Beispiel mit einem Gruppierungs- und einem Messwiederholungsfaktor, je nach Vorgehensweise mit einem multivariaten Verfahren nur die Effekte getestet werden, die den Messwiederholungsfaktor enthalten, also diesen sowie dessen Interaktionen mit dem oder den Gruppierungsfaktoren. Die Tests der Gruppierungshaupteffekte sind ja nicht von der Sphärität betroffen und können daher bei Bedarf mit der „normalen“ parametrischen Varianzanalyse geprüft werden.

Zunächst wird an das Beispiel aus Abschnitt 5.3.9 angeknüpft, bei dem die Berechnung der Differenzen der 4 Messwiederholungsvariablen  $v_1, \dots, v_4$  erforderlich ist:  $v_4 - v_3$ ,  $v_3 - v_2$  und  $v_2 - v_1$ . Dieses kann implizit im Aufruf der Funktion erfolgen, wobei allerdings in jedem Fall diese Variablen zu einer Matrix zusammengefasst werden müssen, z.B. mittels `cbind`. Die Struktur der Datenmatrix muss hier die „normale“, also untransformierte sein. Nachfolgend die Ein- und Ausgabe, hier wieder unter Verwendung des Tests von Pillai:

```
with(winer568, anova(lm(cbind(V4-V3, V3-V2, V2-V1) ~ A),
  test="Pillai"))
```

Analysis of Variance Table								
	Df	Pillai	approx F	num Df	den Df	Pr(>F)		
(Intercept)	1	40.514	108.039	3	8	8.205e-07	***	
A	1	30.118	80.316	3	8	2.590e-06	***	
Residuals	10							

In der Zeile „Intercept“ wird der Test für den Faktor Zeit ausgegeben, der bereits in Kapitel 5.3.9 überprüft worden war. In der Zeile A ist das Ergebnis für die Interaktion A\*Zeit abzulesen. Beide Effekte sind signifikant.

Soll der Haupteffekt A getestet werden, müsste der Datensatz wegen der Messwiederholungen umstrukturiert werden und eine Varianzanalyse wie in Kapitel 6.2 beschrieben durchgeführt werden. Es geht aber auch einfacher: Die Summe der Messwiederholungsvariablen wird errechnet und damit eine Varianzanalyse ohne Messwiederholungen (siehe Kapitel 4.3.2) durchgeführt. Der Test ist identisch mit dem „normalen“ univariaten parametrischen F-Test:

```
within(winer568, Vsum<-V1+V2+V3+V4) ->winer568
summary(aov(Vsum~A, winer568))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	75.0	75.00	2.045	0.183
Residuals	10	366.7	36.67		

Die Methode, die Differenzen mittels `anova` und `lm` zu analysieren, funktioniert nicht mehr bei mehr als einem Messwiederholungsfaktor. Dann ist das Verfahren anzuwenden, das in Abschnitt 5.4.7 vorgestellt worden ist und das zudem den Vorteil hat, zugleich die Gruppierungseffekte zu testen und die Ergebnisse in einer übersichtlichen Tabelle zu präsentieren. Hier wird das dort angeführte Beispiel mit dem Datensatz `mydata5` ergänzt um den Gruppierungsfaktor `Geschlecht`, der nun beim Aufruf von `lm` berücksichtigt werden muss:

```

Medikament <- factor(rep(c("Kontrolle", "Med A", "Med B"), each=3),
                    levels=c("Kontrolle", "Med A", "Med B"))
idf <- data.frame(Medikament, Aufgabe)
mod <- lm(cbind(v1, v2, v3, v4, v5, v6, v7, v8, v9) ~ Geschlecht, data=mydata5)
Anova(mod, idata=idf, idesign=~Medikament*Aufgabe, type=3, test="Wilks")

```

Type III Repeated Measures MANOVA Tests: Wilks test statistic								
	Df	test	stat	approx	F	num Df	den Df	Pr(>F)
(Intercept)	1	0.03263	177.881	1	6	1.099e-05	**	
Geschlecht	1	0.59124	4.148	1	6	0.0878433	.	
Medikament	1	0.03599	66.961	2	5	0.0002458	**	
Geschlecht:Medikament	1	0.25373	7.353	2	5	0.0324291	*	
Aufgabe	1	0.11266	19.690	2	5	0.0042605	**	
Geschlecht:Aufgabe	1	0.79134	0.659	2	5	0.5570588		
Medikament:Aufgabe	1	0.22409	2.597	4	3	0.2295496		
Geschl:Medikam:Aufgabe	1	0.54893	0.616	4	3	0.6818768		

Abschließend noch einige Anmerkungen zu den nichtparametrischen Varianten. Für das Verfahren von Agresi & Pendergast war Abschnitt 5.3.9 die Funktion `ap.anova` vorgestellt worden. Diese kann auch bei gemischten Versuchsplänen eingesetzt werden, allerdings jeweils auf einen Gruppierungs- und einen Messwiederholungsfaktor beschränkt. Der Test des Gruppierungsfaktors A muss wie oben beschrieben gesondert mittels des nicht-transformierten Datensatzes `winer568` vorgenommen werden: zunächst Bildung der Summe (oder Mittelwerte)  $V_{sum}$  von  $V_1, \dots, V_4$ , danach z.B. ein Kruskal-Wallis-Test:

```

ap.anova(winer568t, "x", "Vpn", "Zeit", "A")
within(winer568, Vsum<-V1+V2+V3+V4) ->winer568
with(winer568, kruskal.test(rmeans, A))

```

	chisq	df	p value	F	df num	df denom	p value
main effect	313.68468	3	0.000000	104.561560	3	33	0.0000
interaction	11.46859	3	0.009444	2.548576	3	8	0.1290

Kruskal-Wallis chi-squared = 2.6005, df = 1, p-value = 0.1068

Zur Durchführung des Verfahrens von Akritas & Arnold muss zunächst, wie in Abschnitt 5.3.9 beschrieben, die abhängige Variable, bestehend aus  $v_1, \dots, v_4$ , in Ränge umgerechnet werden:

```
mat568 <- matrix(rank(as.vector(as.matrix(winer568[, 3:6]))), 12, 4)
```

Wie dort beschrieben können die Differenzen  $v_4 - v_3$ ,  $v_3 - v_2$  und  $v_2 - v_1$  einfach durch die Differenz aus der Matrix der Variablen  $v_2, v_3, v_4$  und der Matrix der Variablen  $v_1, v_2, v_3$  gebildet werden, die als Argument in `lm` eingesetzt wird:

```
anova(lm((mat568[, 2:4] - mat568[, 1:3]) ~ winer568$A), test="Hotelling-Lawley")
```

	Df	Hotelling-Lawley	approx	F	num Df	den Df	Pr(>F)
(Intercept)	1	38.325	102.199	3	8	1.019e-06	***
winer568\$A	1	21.582	57.551	3	8	9.295e-06	***
Residuals	10						

**mit SPSS:**

Zunächst wird an das Beispiel in Kapitel 5.3.9 angeknüpft. Lediglich wird zusätzlich der bzw. die Gruppierungsfaktoren angegeben:

```
GLM V1 V2 V3 V4 by A
  /WSfactor=Zeit 4 Polynomial
  /WSdesign=Zeit
  /design=A.
```

Multivariate Tests						
Effekt		Wert	F	Hypothese df	Fehler df	Sig.
Zeit	Pillai-Spur	,976	108,039 <sup>b</sup>	3,000	8,000	,000
	Wilks-Lambda	,024	108,039 <sup>b</sup>	3,000	8,000	,000
	Hotelling-Spur	40,514	108,039 <sup>b</sup>	3,000	8,000	,000
	Größte charakteristische Wurzel nach Roy	40,514	108,039 <sup>b</sup>	3,000	8,000	,000
Zeit * A	Pillai-Spur	,968	80,316 <sup>b</sup>	3,000	8,000	,000
	Wilks-Lambda	,032	80,316 <sup>b</sup>	3,000	8,000	,000
	Hotelling-Spur	30,118	80,316 <sup>b</sup>	3,000	8,000	,000
	Größte charakteristische Wurzel nach Roy	30,118	80,316 <sup>b</sup>	3,000	8,000	,000

In den Zeilen sind, je nach Wahl des multivariaten Tests, die Ergebnisse für den Haupteffekt Zeit bzw. für die Interaktion A\*Zeit abzulesen. Beide Effekte sind signifikant. Am Ende wird noch die Tabelle der Tests für die Gruppierungsfaktoren ausgegeben:

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	1752,083	1	1752,083	191,136	,000
A	18,750	1	18,750	2,045	,183
Fehler	91,667	10	9,167		

Für ein Beispiel mit dem Datensatz mydata5 wird an das aus Abschnitt 5.4.1 angeknüpft und der Gruppierungsfaktor Geschlecht in in der GLM-Anweisung hinter den abhängigen Variablen angefügt:

```
GLM v1 v2 v3 v4 v5 v6 v7 v8 v9 by Geschlecht
  /wsfactor=Medikament 3 polynomial Aufgabe 3 polynomial
  /wsdesign=Medikament Aufgabe Medikament*Aufgabe.
```

Auf die Ausgabe wird hier verzichtet und auf Abschnitt 5.4.1 verwiesen.

## 6. 10. 2 Multivariate Analysen: Spatial Signs und Spatial Ranks Methode

In Kapitel 2.11.3 waren kurz neuere multivariate verteilungsfreie Methoden erwähnt worden, die auf räumlichen Vorzeichen und Rängen beruhen, u.a. für multivariate Varianzanalysen. Im Abschnitt 5.3.10 waren die dafür verfügbaren Funktionen vorgestellt und deren Einsatz für eine 1-faktorielle Analyse mit Messwiederholungen gezeigt worden. Allerdings können mit den multivariaten Verfahren bei gemischten Versuchsplänen nur die Effekte getestet werden, die den Messwiederholungsfaktor enthalten. Bei den hier vorgestellten Funktionen werden zwar der Haupteffekt des Messwiederholungsfaktors und die Interaktion mit derselben Funktion ge-

testet, doch in getrennten Schritten. Der erste Schritt war in Abschnitt 5.3.10 beschrieben worden, der Test des Interaktionseffekts wird nachfolgend gezeigt. Für den Test des Gruppierungsfaktors muss ein anderes nichtparametrisches Verfahren gewählt werden.

### mit R:

Das Prozedere wird wieder anhand des Datensatzes `winer568` mit den Variablen  $V_1, \dots, V_4$  vorgestellt, zunächst mit der Funktion `sr.loc.test` des Pakets `SpatialNP`. Für den Test des Messwiederholungsfaktors war folgende Anweisung erforderlich (vgl. Abschnitt 5.3.10):

```
with(winer568, sr.loc.test(cbind(V4-V3,V3-V2,V2-V1),score="rank"))
```

Der Test der Interaktion wird nun gesondert angefordert durch Angabe des Gruppierungsfaktors mittels des Parameters `g`,

```
library(SpatialNP)
within(winer568,A <- factor(A))->winer568
with(winer568, sr.loc.test(cbind(V4-V3,V3-V2,V2-V1),g=A,score="rank"))
```

```
Several samples location test using spatial ranks

data: cbind(V4 - V3, V3 - V2, V2 - V1) by factor(A)
Q.2 = 11.487, df = 3, p-value = 0.009363
true difference between group locations is not equal to c(0,0,0)
```

Bei Wahl des Pakets `MNM` muss die Funktion `mv.Csample.test` anstatt `mv.Csample.test` genommen werden, hier ohne den Parameter `g`:

```
library(MNM)
within(winer568,A <- factor(A))->winer568
with(winer568,mv.Csample.test(cbind(V4-V3,V3-V2,V2-V1),A,score="rank"))
```

```
Several samples location test using spatial ranks

data: cbind(V4 - V3, V3 - V2, V2 - V1) by factor(A)
Q.2 = 11.22, df = 3, p-value = 0.01059
true location difference between some groups is not equal to c(0,0,0)
```

Bei Wahl des Pakets `OjaNP` muss die Funktion `ojaCsampleTest` anstatt `ojasampleTest` genommen und ein Modell mit „~“ spezifiziert werden:

```
library(OjaNP)
within(winer568,A <- factor(A))->winer568
with(winer568, ojaCsampleTest(cbind(V4-V3,V3-V2,V2-V1)~A,score="rank"))
```

```
OJA C SAMPLE RANK TEST

data: cbind(V4 - V3, V3 - V2, V2 - V1) by factor(A)
Q.R = 10.08, df = 3, p-value = 0.0179
true location difference is not equal to c(0,0,0)
```

Im Paket `ICSNP` bietet die Funktion `rank.ctest` einen Test der Interaktion, bei der wahlweise Ränge (`scores="rank"`) oder normal scores (`scores="normal"`) verwendet werden können:

```
library(ICSNP)
within(winer568,A <- factor(A))->winer568
with(winer568, rank.ctest(cbind(V4-V3,V3-V2,V2-V1)~A,scores="rank"))
```

Marginal Two Sample Rank Sum Test

```
data: cbind(V4 - V3, V3 - V2, V2 - V1) by A
T = 10.919, df = 3, p-value = 0.01217
true location difference is not equal to c(0,0,0)
```

Auch hier decken sich die Ergebnisse mit den oben erzielten. Möchte man den Haupteffekt A testen, so muss man auf andere Methoden zurückgreifen, z.B. auf den Kruskal-Wallis-Test (Kapitel 4.2.1). Dazu müssen zunächst für jede Erhebungseinheit Summe oder Mittelwert errechnet werden (hier mittels der Funktion `rowMeans`), diese mit `cbind` mit dem Dataframe verbunden werden, bevor mit `kruskal.test` der Test durchgeführt werden kann.

```
with(cbind(winer568, MW=rowMeans(winer568[,3:6])),kruskal.test(MW,A))
```

Kruskal-Wallis rank sum test

```
data: MW and factor(A)
Kruskal-Wallis chi-squared = 1.468, df = 1, p-value = 0.2257
```

### 6. 10. 3 Welch & James

Das Verfahren von Welch & James kann als semiparametrisch angesehen werden, ähnlich den Mittelwertvergleichen für inhomogene Varianzen. Es setzt weder Sphärität der Kovarianzmatrix noch deren Homogenität über die einzelnen Gruppen voraus. Damit ist es unproblematischer anzuwenden als die parametrischen Varianzanalysen unter Verwendung der  $\epsilon$ -Korrekturen. Keselman, Carriere & Lix haben sich intensiv mit dem Verfahren von Welch & James auseinandergesetzt, u.a. in dem eingangs erwähnten Artikel (1993). Das Verfahren datiert zwar aus den 50er Jahren ist aber erst 1980 von Johansen in einer praktikablen Version präsentiert worden. In verschiedenen Artikeln schneidet es bei Vergleichen relativ gut ab. Allerdings mit einer Einschränkung: Insbesondere für den Test der Interaktion sind hinreichend große  $n_i$  erforderlich, da bei zu kleinen  $n_i$  der Test liberal reagiert, Keselman et al. (1993) empfehlen  $n_i > 4*(J-1)$ , wobei  $J$  die Anzahl der Messwiederholungen ist.

#### mit R:

Das Verfahren wird auf den Beispieldatensatz 6 (`winer568`) angewandt, der zwei Gruppierungsfaktoren A und B enthält. Hier soll die Varianzanalyse für die Faktoren Zeit (Messwiederholung) und A durchgeführt werden. Es sei darauf aufmerksam gemacht, dass die o.a. Bedingung für die  $n_i$  hier nicht erfüllt ist, da  $n_i = 6$  kleiner als  $4*(4-1) = 12$  ist. Zunächst wird mit der Funktion `ezANOVA` angezeigt, dass die Sphärität nicht erfüllt ist ( $p < 0.01$ ). Dazu dient wieder die umstrukturierte Version `winer568t`.

```
ezANOVA(winer568t,x,Vpn,between=.(A),within=.(Zeit))
```

```
$`Mauchly's Test for Sphericity`
  Effect      W      p p<.05
3   Zeit 0.05875131 0.0001770111 *
4 A:Zeit 0.05875131 0.0001770111 *
```

Alternativ kann dies auch mit der Funktion `check.sphere` (vgl. Anhang 3.6) erfolgen:

```
check.sphere(winer568[,3:6])
```

	statistic	df	p value
Mauchly test	21.2038442	5.000000	7.412765e-04
Likelihood Ratio	23.9906351	9.000000	4.316141e-03
John's test chisq	32.5455437	5.000000	4.631992e-06
John's test beta	0.3170646	8.769231	4.459785e-03
John-Nagao	17.1214872	9.000000	2.661820e-02
Muirhead & Waternaud	45.6028296	9.000000	7.135443e-07
compound symmetry	33.5292718	8.000000	4.943008e-05

Die Funktion `wj.spanova` (vgl. Anhang 3) führt die Varianzanalyse nach dem Verfahren von Welch & James aus, gibt allerdings keinen Test für den Test des Gruppierungsfaktors aus. Dazu muss die abhängige Variable zunächst über die Messwiederholungen addiert oder gemittelt, z.B. mit Hilfe der Funktion `rowMeans`, und die Summe dann mit dem Welch & James-Verfahren für unabhängige Stichproben getestet werden. Hierbei ist allerdings der ursprüngliche Dataframe `winer568` zu verwenden. Zu beachten ist, dass bei den Aufrufen von `wj.spanova` und `wj.anova` die Variablennamen in " " gesetzt werden müssen.

```
wj.spanova(winer568t, "x", "A", "Zeit", "Vpn")
V <- rowMeans(winer568[,c("V1", "V2", "V3", "V4")])
winer568 <- cbind(winer568, V)
wj.anova(winer568, "V", "A")
```

Hier die Ausgabe zunächst von `wj.spanova`, wonach beide Effekte stark signifikant sind, danach von `wj.anova`.

	F value	df num	df denom	p value
Zeit	115.87041	3	8.055823	5.790882e-07
A:Zeit	86.13801	3	8.055823	1.847051e-06

	Chi Sq	df	P(Chi>value)
A	2.045455	1	0.2225

Inzwischen gibt es auch eine entsprechende R-Funktion auf cran: `welchADF.test` im Paket `welchADF`, die sogar den Vorteil, ebenfalls den Welch-James-Test für den Gruppierungsfaktor durchzuführen:

```
library(welchADF)
welchADF.test(formula=winer518t, response="score",
              between.s="Geschlecht", within.s="Zeit", subject="Vpn")
```

mit folgender Ausgabe:

Welch-James Approximate DF Test (Least squares means & variances)			
Omnibus test(s) of effect and/or interactions			
	Geschlecht	Zeit	Geschlecht : Zeit
Pr(>WJ)	0.5121347	0.004583745	0.002406693

### 6. 10. 4 Koch

Koch hat diverse nichtparametrische Verfahren für gemischte Versuchspläne entwickelt (vgl. Koch, 1993). Eines davon entspricht einer Übertragung des multivariaten Ansatzes des Messwiederholungsmodells (vgl. Kapitel 5.2), das zwar keine Spherizität, dafür aber multivariate Normalverteilung voraussetzt, auf rangtransformierte Daten. Damit entfallen auch hier die entsprechenden Prüfungen von Voraussetzungen.

#### mit R:

Das Verfahren wird wieder auf den Beispieldatensatz 6 (`winer568`) angewandt, der zwei Gruppierungsfaktoren A und B enthält. Hier soll die Varianzanalyse für die Faktoren Zeit (Messwiederholung) und A durchgeführt werden, für die, wie im vorigen Abschnitt gezeigt wurde, die Spherizität nicht erfüllt ist. Dazu dient ausnahmsweise die untransformierte Version `winer568`. Beim Aufruf der Funktion `koch.anova` (vgl. Anhang 3) werden aus dem Dataframe zwei Parameter übergeben: zum einen die abhängigen Variablen (die Variablen 3 bis 6), zum anderen die Gruppierungsvariable (Variable A):

```
koch.anova(winer568[,3:6],winer568$A)
```

	chisquare	df	p value
A	1.467972	1	0.225666013
B	12.000000	3	0.007383161
A:B	10.285442	3	0.016289293

Bei der Ausgabe ist zu beachten, dass die Faktoren einfach mit „A“ und „B“ bezeichnet werden, d.h. in diesem Beispiel entspricht „A“ wirklich dem Faktornamen, und „B“ entspricht dem Faktor Zeit (Messwiederholungen).

### 6. 10. 5 GEE

In Kapitel 2.15 war darauf hingewiesen worden, dass die GEE-Methode deutlich schwächere Voraussetzungen hat als die parametrische Varianzanalyse, die Anwendung allerdings problematisch ist, insbesondere wenn die Fallzahl nicht hinreichend groß ist. Wenn also das Programm mit einer Fehlermeldung abbricht, so kann man einen weiteren Versuch ohne Interaktionen starten, weil dadurch die zu schätzende Parameteranzahl deutlich reduziert wird. Das Verfahren soll hier wiederum am Datensatz des Beispiels 4 demonstriert werden.

#### mit R:

In R gibt es u.a. die folgenden Funktionen für Analyse mit Messwiederholungen mittels der GEE-Methode:

- `gee` (Paket `gee`)
- `geeglm` (Paket `geepack`)
- `geem` (Paket `geem`)
- `gee` (Paket `drgee`)
- `MGEE` (Paket `PGEE`)

Die Eingabe ist bei allen Funktionen weitgehend identisch. Leider werden von allen nur die Kontrast-Koeffizienten mit Tests ausgegeben, aber keine Anova-Tabelle. Gegebenenfalls muss man aus diesen wie weiter unten in Kapitel 9.8 beschrieben für einen Faktor einen Gesamttest mit der Hand ausrechnen. Hier soll `gee` (Paket `gee`) vorgestellt werden. Als



Basis dient wieder der umstrukturierte Datensatz `winer518t` (vgl. Abschnitt 5.1.2), in dem Geschlecht, Zeit und Vpn als Faktor deklariert sein müssen. Ein- und Ausgabe:

```
library(gee)
erg <- gee(score~Geschlecht*Zeit,id=Vpn,family=gaussian,data=winer518t)
summary(erg)
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	4.26667	0.32829	12.99643	0.29364	14.53045
Geschlecht1	0.33333	0.32829	1.01535	0.29364	1.13519
Zeit.L	-2.40416	0.56862	-4.22804	0.52535	-4.57625
Zeit.Q	-0.16330	0.56862	-0.28718	0.49125	-0.33241
Geschlecht1:Zeit.L	0.56569	0.56862	0.99483	0.52535	1.07676
Geschlecht1:Zeit.Q	-2.04124	0.56862	-3.58979	0.49125	-4.15514

Die Ausgabe enthält für beide Faktoren und die Interaktion lediglich Tests für die einzelnen Kontraste. Für den Messwiederholungsfaktor (hier `Zeit`) wird ein Test auf linearen (`Zeit.L`) bzw. quadratischen Trend (`Zeit.Q`) ausgegeben. Aus den z-Werten der Kontraste ergibt sich:

- Geschlecht: kein signifikanter Haupteffekt ( $z=1.13519$ ).
- Zeit:  $\chi^2 = (-4.57625)^2 + (-0.33241)^2 = 21.05$  ist signifikant bei  $FG=2$ .
- Interaktion:  $\chi^2 = (1.07676)^2 + (4.15514)^2 = 18.42$  ist signifikant bei  $FG=2$ .

Hiernach besteht ein Unterschied zwischen den Zeitpunkten, der für Männer und Frauen unterschiedlich ausfällt.

Alternativ kann auch der Wald-Test mittels der Funktion `gee . anova` (vgl. Anhang 3) durchgeführt werden. Diese erwartet als Eingabe:

- die Koeffizienten: `erg$coefficients`
- die Kovarianzmatrix: `erg$"robust.variance"`
- die Freiheitsgrade für die 3 Tests (Geschlecht, Zeit, Geschlecht\*Zeit): 1, 2, 2
- die Anzahl der Fälle  $n$ : 10

```
gee.anova(erg.gee$coefficients,erg.gee$"robust.variance",c(1,2,2),n=10)
```

	df	Chi	P.Chi	F	P.F	nerror	err.invert
1	1	1.289	0.256	1.289	0.286	0	0
2	2	14.131	0.001	7.066	0.017	0	0
3	2	19.060	0.000	9.530	0.008	0	0

Die 3 Ergebniszeilen entsprechen den Tests für die Effekte Geschlecht, Zeit und Geschlecht\*Zeit. Während der  $\chi^2$ -Test für größere Stichproben konzipiert ist, sollte der F-Test bei kleineren  $n$  angewandt werden. Zahlreiche Funktionen für die GEE-Modelle, so auch das hier benutzte `gee`, erlauben neben `family=gaussian` für metrische Daten alternativ auch `family=poisson`, eigentlich für den Fall, dass die abhängige Variable Häufigkeiten repräsentiert, der aber auch für den Fall ordinaler Variablen angewandt werden kann. Auf ein Beispiel dafür soll hier verzichtet werden.

**mit SPSS:**

SPSS bietet für die Analyse mit Messwiederholungen mittels der GEE-Methode die Prozedur GENLIN an. SPSS erwartet hier ausnahmsweise die Daten nicht in der „normalen“ Struktur (alle Werte pro Fall in einer Zeile), sondern in der für R typischen Form, in der die Werte jeder Messwiederholung in einer separaten Zeile angeordnet sein müssen, verbunden mit einer Fallidentifikation, hier *Vpn*, einer Variablen für den Messwiederholungsfaktor, hier *Zeit*, sowie einem Namen für die abhängige Variable, hier *score*. Die Umstrukturierung wird im Anhang 1.1 beschrieben. Nachfolgend zunächst die Eingabe:

```
GENLIN score BY Geschlecht Zeit
/MODEL Geschlecht Zeit Geschlecht*Zeit
  DISTRIBUTION=NORMAL LINK=Identity
/REPEATED SUBJECT=Vpn CORRTYPE = EXCHANGEABLE
/EMMEANS TABLES = Zeit
  compare = Zeit
  contrast=repeated
/EMMEANS TABLES = Geschlecht
  compare = Geschlecht
  contrast=pairwise.
```

Mittels der beiden EMMEANS-Befehle werden Einzelvergleiche durchgeführt und ein Gesamttest für den Faktor ausgegeben. Für den Messwiederholungsfaktor empfiehlt sich häufig die Option von „repeated“-Kontrasten (siehe Kapitel 9), für den Gruppierungsfaktor wäre in diesem Fall der Befehl entbehrlich, da er nur 2 Gruppen hat. Nachfolgend zunächst der wesentliche Teil der Standardausgabe (Parameterschätzung mit Tests), danach die jeweilige Ausgabe der beiden EMMEANS-Befehle (Mittelwertvergleiche und Gesamttest) für Geschlecht und Zeit.

Parameter	Regressions koeffizient B	Standard Fehler	95% Wald- Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi- Quadrat	df	Sig.
(Konstanter Term)	2,600	,7266	1,176	4,024	12,803	1	,000
[ =1]	-,200	1,1027	-2,361	1,961	,033	1	,856
[ =2]	0 <sup>a</sup>	.	.	.	.	.	.
[Zeit=1]	4,200	,8672	2,500	5,900	23,457	1	,000
[Zeit=2]	-,200	,8672	-1,900	1,500	,053	1	,818
[Zeit=3]	0 <sup>a</sup>	.	.	.	.	.	.
[ =1] * [Zeit=1]	-1,600	1,0198	-3,599	,399	2,462	1	,117
[ =1] * [Zeit=2]	4,200	,9121	2,412	5,988	21,202	1	,000
[ =1] * [Zeit=3]	0 <sup>a</sup>	.	.	.	.	.	.
[ =2] * [Zeit=1]	0 <sup>a</sup>	.	.	.	.	.	.
[ =2] * [Zeit=2]	0 <sup>a</sup>	.	.	.	.	.	.
[ =2] * [Zeit=3]	0 <sup>a</sup>	.	.	.	.	.	.
(Skala)	3,233						

Mittelwertvergleiche und Gesamttest) für Geschlecht:

Paarweise Vergleiche							
(I)	(J)	Mittlere Differenz (I- J)	Standard Fehler	df	Sig.	95% Wald-Konfidenzintervall für die Differenz	
						Unterer Wert	Oberer Wert
1	2	,67	,868	1	,443	-1,03	2,37
2	1	-,67	,868	1	,443	-2,37	1,03

Gesamttestergebnisse		
Wald-Chi- Quadrat	df	Sig.
,590	1	,443

Mittelwertvergleiche und Gesamttest) für Zeit

Individuelle Testergebnisse					
Zeit Wiederholter Kontrast	Kontrastschätz- er	Standard Fehler	Wald-Chi- Quadrat	df	Sig.
Niveau 1 vs. Niveau 2	1,50	,405	13,720	1	,000
Niveau 2 vs. Niveau 3	1,90	,456	17,356	1	,000

Gesamttestergebnisse		
Wald-Chi- Quadrat	df	Sig.
44,526	2	,000

Aus der ersten Tabelle Gesamttestergebnisse ist zu entnehmen, dass der Haupteffekt von Geschlecht nicht signifikant ist ( $p=,443$ ), während die entsprechende Tabelle für Zeit einen signifikanten Effekt ( $p<0.001$ ) anzeigt. Das Ergebnis für die Interaktion ist der Tabelle der Parameterschätzungen zu entnehmen: Aus den dort ausgewiesenen  $\chi^2$ -Werten (2,462 und 21,202) wird die Summe 23,664 errechnet mit mittels der  $\chi^2$ -Verteilung auf Signifikanz überprüft, wobei als Freiheitsgrade die Summe der entsprechenden Freiheitsgrade aus der Tabelle zu nehmen sind (hier also 1+1). Bei 5% Irrtumswkt beträgt der kritische Wert 5,99, so dass eine signifikante Interaktion nachgewiesen ist.

### 6. 10. 6 GLMM

In Kapitel 2.15 war darauf hingewiesen worden, dass die GLMM-Methode deutlich schwächere Voraussetzungen hat als die parametrische Varianzanalyse, die Anwendung allerdings problematisch ist, insbesondere wenn die Fallzahl nicht hinreichend groß ist. Ein Vorteil von GLMM ist, Datensätze mit fehlenden Werten verarbeiten zu können. Das Verfahren soll hier wiederum am Datensatz des Beispiels 4 demonstriert werden.

**mit R:**

In R gibt es u.a. die folgenden Funktionen für Analyse mit Messwiederholungen mittels der GLMM-Methode:

- `lmer` (Paket `lme4`)
- `glmmML` (Paket `glmmML`)
- `glmmPQL` (Paket `MASS`)
- `mixed` (Paket `afex`)

Hier soll die Funktion `lmer` vorgestellt werden, die zum einen am häufigsten empfohlen wird, und die zum anderen die Möglichkeit bietet, mittels der Funktion `Anova` (Paket `car`) für die Effekte varianzanalytische Tests auszugeben, u.a. den in Kapitel 9.8 erwähnten Typ II Wald-Test. Hier hilft die Funktion `nlminb` des Optimierungspaket `optimx` die in 2.15 beschriebene Schwierigkeiten beim Finden einer Lösung zu reduzieren. Als Basis dient wieder der umstrukturierte Datensatz `winer518t` (vgl. Abschnitt 5.1.2), in dem Geschlecht, Zeit und `Vpn` als Faktor deklariert sein müssen. Ein- und Ausgabe:

```
library(lme4)
library(optimx)
library(car)
erg <- lmer(score~Geschlecht*Zeit+(1|Vpn), data=winer518t,
+         control=lmerControl(optimizer="optimx",
+         optCtrl=list(method="nlminb")))
Anova(erg)
```

```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: score
           Chisq Df Pr(>Chisq)
Geschlecht    0.4717  1    0.4922
Zeit          44.1013  2  2.652e-10 ***
Geschlecht:Zeit 34.0759  2  3.986e-08 ***
```

Auch für GLMM gibt es die Möglichkeit, ein Modell für Häufigkeiten oder ordinale abhängige Variablen zu analysieren. Dazu ist lediglich die Funktion `glmer` anstatt `lmer` aufzurufen, die diverse Verteilungsfamilien erlaubt, sowie die Optimierungsoptionen anzupassen:

```
erg.glm <- glmer(score~Geschlecht*Zeit+(1|Vpn), data=winer518t,
+              family=poisson,
+              control=glmerControl(optimizer="optimx", calc.derivs = F,
+              optCtrl=list(method="nlminb")))
Anova(erg)
```

```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: score
           Chisq Df Pr(>Chisq)
Geschlecht    0.3422  1  0.558569
Zeit          13.3985  2  0.001232 **
Geschlecht:Zeit  9.2821  2  0.009648 **
```

Abschließend noch eine Bemerkung zur o.a. Funktion `mixed`. Diese basiert auf der o.a. Funktion `lmer` und soll deren Benutzung erleichtern. Doch leider zeigte sich, dass die Ergebnisse von `lmer` in `mixed` nicht korrekt weiterverarbeitet werden.

### mit SPSS:

In SPSS gibt es zwar die Prozedur `GENLINMIXED` für die GLMM-Methode, doch sie bricht häufig mit Fehlermeldungen ab, manchmal mit der Meldung, dass bei der Schätzung eine Matrix nicht „positiv definit“ ist, vielfach auch mit nicht näher spezifizierten Meldungen, sowohl bei den hier verwendeten kleinen Datensätzen als auch bei den meisten größeren mit einem  $n > 100$ . Daher wird hier nur die Syntax aufgeführt. SPSS erwartet hier

wie bei der GEE-Methode die Daten nicht in der „normalen“ Struktur (alle Werte pro Fall in einer Zeile), sondern in der Form, in der die Werte jeder Messwiederholung in einer separaten Zeile angeordnet sein müssen (vgl. vorigen Abschnitt zu GEE).

```

GENLINMIXED
/ DATA_STRUCTURE SUBJECTS=Vpn
  REPEATED_MEASURES = Zeit
  GROUPING = Geschlecht
/ FIELDS TARGET=score
/ TARGET_OPTIONS DISTRIBUTION=normal LINK=identity
/ FIXED EFFECTS=Geschlecht
  USE_INTERCEPT=TRUE
/ RANDOM EFFECTS=Zeit Geschlecht*Zeit
  USE_INTERCEPT=TRUE
  SUBJECTS=Vpn
  COVARIANCE_TYPE = COMPOUND_SYMMETRY
/ BUILD_OPTIONS MAX_ITERATIONS = 500
/ EMMEANS TABLES=Geschlecht
  COMPARE=Geschlecht.

```

Allerdings ist die Ausgabe ohnehin recht dürftig und bietet keine varianzanalytischen Tests für die zu testenden Effekte. Die einzig interessanten Teile werden hier wiedergegeben:

Zufälliger Effekt	Schätzung	Standardfehler	Z	Sig.	95% Konfidenzintervall	
					Unterer	Oberer
Var(Konstanter Term)	1,725	1,441	1,197	,231	0,336	8,869
Var(Zeit)	0,149	1,842	0,081	,936	0,000	5.184.973.019,340
Var(Geschlecht*Zeit)	17,000 <sup>a</sup>					

Kovarianzstruktur: Varianzkomponenten  
 Subjektspezifikation: Vpn  
<sup>a</sup>Der Parameter ist redundant.

### Modellzusammenfassung

Ziel:

Wahrscheinlichkeitsverteilung	Normal
Verknüpfungsfunktion	Identität
Informationskriterium — Akaike (korrigiert)	155,556
Bayes	157,546

Informationskriterien beruhen auf der -2 Log-Likelihood (127,556) und dienen zum Modellvergleich. Modelle mit kleineren Werten für Informationskriterien passen besser.

In der oberen Tabelle sollten eigentlich in der Spalte “Sig.“ Tests für die 3 Effekte (Geschlecht, Zeit und Interaktion) erscheinen. Hier fehlt der für die Interaktion. Die Abweichung zu den o.a. Ergebnissen von R ist dadurch zu erklären, dass der Algorithmus in SPSS keine Lösung finden konnte. Die untere Tabelle gibt Testgrößen, allerdings ohne Signifikanzüberprüfung, aus, die eine Aussage über das Gesamtmodell ermöglichen.

### 6. 10. 7 GA- und IGA-Approximationen von Huynh

Huynh (1978) hat die bekannte  $\epsilon$ -Adjustierung für die Freiheitsgrade des F-Tests verbessert: die GA-Approximation zur Kompensierung fehlender Sphärität sowie die IGA-Approximation, die zusätzlich heterogene Kovarianzmatrizen berücksichtigt, also insbesondere ungleiche Varianzen für die Gruppen von Faktor A. Daher ist letztere i.a. vorzuziehen. Bei beiden Methoden werden durch die Adjustierung sowohl der F-Wert als auch Zähler- und Nenner-Freiheitsgrade des F-Tests verkleinert, so dass die Teststärke je nach Ausmaß der Varianz- und Kovarianzheterogenitäten sich reduziert. Funktionen hierfür sind nur in R verfügbar.

#### mit R:

Vom Autor werden dafür die Funktionen `iga` und `iga.anova` angeboten. `iga` berechnet lediglich die Korrekturfaktoren, während `iga.anova` eine komplette Varianzanalyse durchführt und daher i.a. vorzuziehen ist. Dennoch ist auch `iga` für die Berechnungen erforderlich. Da der Gruppierungsfaktor hiervon nicht berührt ist, wird für diesen auch kein Test durchgeführt. Dies soll wieder am Beispieldatensatz 6 (`winer568`, vgl. Abschnitt 5.1.2) demonstriert werden, der zwei Gruppierungsfaktoren A und B enthält, wovon der erste in der Analyse verwendet wird. Die Daten können sowohl im breiten als auch im langen Format eingegeben werden. Zunächst die Version mit breitem Format.

```
iga.anova(winer568[,3:6],winer568[,1])
```

Hyunh IGA (improved general approximation)							
	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
W	1.4723	572.92	190.972	120.192	2.373e-09	***	
groups:W	1.3306	72.92	24.306	15.297	0.0006992	***	
Residuals	14.9948	47.67	1.589				

In der Ausgabe wird der Messwiederholungsfaktor mit “w“ bezeichnet, die Interaktion mit “groups:w“. Die Eingabe der Daten im langen Format sieht folgendermaßen aus:

```
with(winer568t, iga.anova(x,A,trial=Zeit,id=Vpn))
```

### 6. 10. 8 modifizierter Brown-Forsythe-Test

Vallejo et al. (2004) haben die bekannte Varianzanalyse-Methode für heterogene Varianzen auf split-plot-Designs verallgemeinert. Diese berücksichtigt sowohl fehlende Sphärität als auch heterogene Kovarianzmatrizen. Eine entsprechende Funktionen ist nur in R verfügbar.

#### mit R:

Vom Autor wird dafür die Funktion `mbf.f` angeboten, die eine komplette Varianzanalyse durchführt. Dies soll wieder am Beispieldatensatz 6 (`winer568`) demonstriert werden, der zwei Gruppierungsfaktoren A und B enthält. Die Daten können sowohl im breiten als auch im langen Format eingegeben werden. Zunächst die Version mit breitem Format.

```
mbf.f(winer568[,3:6],winer568[,1])
```

```
modified Brown-Forsythe method for mixed repeated measures designs
              Df Df.err  F value    Pr(>F)
grouping factor  1  9.9573   2.0455   0.1833
trial factor     3  7.2208 423.0254 1.810e-08 ***
interaction      3  7.2208  78.6192 7.141e-06 ***
```

Die Eingabe der Daten im langen Format sieht folgendermaßen aus:

```
with(winer568t, mbf.f(x,A,trial=Zeit,id=Vpn))
```

## 7. Analysen für dichotome Merkmale

Für dichotome abhängige Variablen gibt es grundsätzlich zwei Möglichkeiten zur Durchführung einer Varianzanalyse: die oben beschriebenen Verfahren oder die weiter unten angeführte logistische Regression (siehe Kapitel 8.1).

### Beispieldatensatz 7 (irish):

Hier wurden 1107 irische Schulkinder zu ihrer Einstellung und Gebrauch der irischen Sprache befragt. Erhoben wurden u.a.:

Variablenname	Bedeutung	Ausprägungen
(school) type	Schultyp	1=secondary (Gymnasium) 2=community (Mischung aus Gymn. und Berufsschule) 3=vocational (Berufsschule)
(school) location	Lage	1=urban (städtisch) 2=rural (ländlich)
sex	Geschlecht	1=male 2=female
income	Einkommen	1=high 2=medium 3=low
vocabula	Vokabular	1=bad 2=poor 3=good 4=excellent
usage	Nutzung	1=never 2=little 3=regular
attitude	Einstellung	1=negative 2=neutral 3=positive

Diesen Daten liegt kein Versuchsplan zugrunde, wie sonst vielfach bei Varianzanalysen. D.h. die Daten wurden erhoben, ohne dass darauf geachtet wurde, dass die Gruppierungsvariablen (Schultyp, Schullage und Geschlecht) orthogonal zueinander oder zumindest unabhängig voneinander sind. Dies erschwert Varianzanalysen insofern, als dass zum einen die Effekte nicht unabhängig voneinander sind und zum anderen die Hinzunahme z.B. von Interaktionseffekten die Tests der anderen Effekte deutlich beeinflusst und somit keine klare Interpretation der Effekte möglich ist. Konkret: Geschlecht und Schultyp sowie Schultyp und Einkommen sind voneinander abhängig. Da nicht orthogonale Faktoren aber bei Untersuchungen häufig der Fall sind, wurde dieser Datensatz bewusst als Gegenstück zu den bislang vorgestellten ausgewählt, die allesamt Versuchspläne beinhalten.

Die Daten wurden früher als Beispieldatensatz mit SPSS ausgeliefert. Die primäre Herkunft der Daten lässt sich nicht mehr klären.

### Beispieldaten 8 (koch):

Bei diesem Datensatz handelt es sich um klinische Daten von 340 Patienten, die in ein Krankenhaus eingeliefert worden waren. Zu Beginn wurden die Leiden der Patienten in leicht (0) und schwer (1) klassifiziert (Faktor *severity*). Ein Teil der Patienten wurde daraufhin behandelt (Faktor *treat*). Anschließend wurden alle im Abstand von mehreren Tagen dreimal untersucht (Faktor *time*). Dabei wurde eine Person entweder als krank (0) oder normal (1) eingestuft (Variable *outcome*). Der Datensatz stammt von Koch et al. (1977) und umfasst eigentlich noch weitere Informationen, wie z.B. Behandlungen zwischen den Untersuchungsterminen. Deren Analyse würde jedoch eine Kovarianzanalyse erfordern. Daher werden diese hier nicht berücksichtigt. Im „Original“ liegt der Datensatz „umstrukturiert“ vor, d.h. die Werte der 3 Zeitpunkte als jeweils 3 Fälle. Nachfolgend ein Auszug:



	case_id	severity	treat	outcome	time013	t013trea	time012	t012trea
1	1	0	0	1	0	0	0	0
2	1	0	0	1	1	0	1	0
3	1	0	0	1	3	0	2	0
4	2	0	0	1	0	0	0	0
5	2	0	0	1	1	0	1	0
6	2	0	0	1	3	0	2	0
7	3	0	0	1	0	0	0	0
8	3	0	0	1	1	0	1	0
9	3	0	0	1	3	0	2	0
10	4	0	0	1	0	0	0	0

In der Standardform für Messwiederholungen sehen die ersten Fälle folgendermaßen aus:

	case_id	severity	treat	outcome.0	outcome.1	outcome.2
1	1	0	0	1	1	1
2	2	0	0	1	1	1
3	3	0	0	1	1	1
4	4	0	0	1	1	1
5	5	0	0	1	1	1

## 7. 1 Anwendung der Verfahren für metrische Merkmale

Dichotome Merkmale verhalten sich vielfach wie metrische Merkmale. Simulationen haben gezeigt, dass man dichotome Variablen bei größeren Fallzahlen vielfach genauso handhaben kann wie metrische Variablen. So auch bei der Varianzanalyse (vgl. dazu Cochran, W.G., 1950 und Lunney, G.H., 1970.) Danach werden sowohl  $\alpha$ -Level wie auch  $\beta$  eingehalten. Für das erforderliche  $n$  gilt: Liegen die relativen Häufigkeiten der beiden Ereignisse über 0,2, so genügen 20 Freiheitsgrade für den Fehlerterm, andernfalls sind mindestens 40 Freiheitsgrade erforderlich. Die Untersuchungen betrafen allerdings nur Versuchspläne mit gleichen Zellenbesetzungszahlen und Tests des Null-Modells, also ohne Effekte anderer Faktoren. D'Agostino (1971) sowie Cleary & Angel (1984) haben die Untersuchungen von Lunney zwar bestätigt, allerdings etwas abgeschwächt mit der Bedingung, dass die relativen Häufigkeiten  $p$  zwischen 0,25 und 0,75 liegen sollten, da andernfalls die Varianzen zu unterschiedlich werden können. Hierbei sei daran erinnert, dass ungleiche Varianzen durch ungleiche relative Häufigkeiten der abhängigen Variablen in den einzelnen Gruppen zustande kommen, da bei einem dichotomen Merkmal Mittelwert, also relative Häufigkeit, und Varianz über  $s^2 = p(1-p)$  zusammenhängen. Dieses wirkt sich allerdings erst bei  $p < 0.25$  bzw.  $p > 0.75$  aus. Bogard (2011) hat die wichtigste Literatur zu diesem Thema mit Zitaten zusammengestellt. Erstaunlicherweise gibt es hierzu kaum neuere Ergebnisse bzw. Veröffentlichungen. Im Gegensatz zur u.a. Logistischen Regression kann diese Vorgehensweise auch bei Messwiederholungen angewandt werden.

Eigene Simulationen (Lüpsen, 2021) haben gezeigt, dass es doch eine Reihe von Situationen gibt, bei denen der Fehler 1. Art nicht mehr eingehalten wird. Zunächst das Positive: Solange die relativen Häufigkeiten  $p$  der abhängigen Variablen zwischen 0,25 und 0,75 liegen oder die Zellenbesetzungszahlen gleich sind, ist wenig zu befürchten. Lediglich bei gemischten Ver-

suchsplänen kann es vereinzelt zu leicht erhöhten Fehlerraten kommen, aber nur bei  $p \sim 0.1$  wenn die Korrelationen der Messwiederholungsvariablen deutlich unterschiedlich sind. Liegt  $p$  außerhalb des Intervalls  $[0.25, 0.75]$  und sind die Zellenbesetzungszahlen ungleich, wird es schwieriger. In Versuchsplänen ohne Messwiederholungen ist die L-Statistik von Puri & Sen die bessere Wahl, zumal die Power annähernd mit der des F-Tests identisch ist. Diese Wahl gilt auch generell für große Designs, mit etwa 15-20 Zellen oder mehr. In gemischten Versuchsplänen ist die Wahl des Verfahrens vom zu testenden Effekt abhängig: Für den Gruppierungsfaktor kann die „normale“ Varianzanalyse angewandt werden, da wie schon früher erwähnt, dieser von der Sphärizität, also der Varianzhomogenität, nicht betroffen ist, für alle Effekte, die einen Messwiederholungsfaktor beinhalten, also z.B. die Interaktion, ist der ATS die erste Wahl. Dieser hat zwar eine deutlich geringere Power (bis zu 50% Verlust), aber es ist das einzige Verfahren, das bei ungleichen  $n_i$  und Vorliegen von Varianzhomogenitäten die Fehlerrate unter Kontrolle hält. Falls dieser nicht verfügbar ist, kann ersatzweise der in Kapitel 6.10.1 vorgestellte multivariate Test, z.B. der von Pillai, alternativ auch die parametrische Analyse mit der Huynh-Feldt-Korrektur benutzt werden, die zwar beide relativ liberal sind, insbesondere beim Test der Interaktion, dafür aber eine relativ große Power besitzen. Das gute Abschneiden der beiden zuletzt genannten Verfahren bei dichotomen abhängigen Variablen in gemischten Designs erklärt sich daraus, dass diese keine Sphärizität voraussetzen (vgl. Kapitel 6.10).

Tests auf Homogenität der Varianzen bzw. auf Sphärizität im Fall von Messwiederholungen entfallen hier, da die Varianzen  $p(1-p)$  sich aus den Mittelwerten  $p$  errechnen lassen.

### 7. 1. 1 Unabhängige Stichproben

An dieser Stelle soll ein Beispiel gerechnet werden, und zwar für den Datensatz 7. Als Kriteriumsvariable wird `vocabula` gewählt, allerdings dichotomisiert: 0=(1/bad, 2/poor) und 1=(3/good, 4/excellent). Als Faktoren: Geschlecht, Schultyp und Einkommen. Wegen der Problematik der Abhängigkeit der Faktoren, auf die bei der Beschreibung des Datensatzes kurz aufmerksam gemacht wurde, wird zum einen eine 2-faktorielle Varianzanalyse mit den Faktoren `sex` und `income` durchgeführt, da diese voneinander unabhängig sind. Der Einfluss von `type` wird wegen der Abhängigkeit von `sex` und `income` separat untersucht, wenn auch der Effekt des Schultyps vom Geschlecht und Einkommen ein wenig mitbeeinflusst wird. Die Interaktionen `sex*type` und `income*type` machen wegen der Abhängigkeit keinen Sinn. Die relativen Häufigkeiten des Kriteriums liegen mit 0,21 bzw. 0,68 im geforderten Bereich.

#### mit R:

Zunächst muss die 4-stufige abhängige Variable `vocabula` dichotomisiert werden (Variable `dvocabul`), bevor „wie gewohnt“ mit `aov` und `drop1` die parametrische Varianzanalyse darauf angewandt wird:

```
irish <- within(irish, dvocabul<-as.integer(vocabula)>2)
options (contrasts=c("contr.sum", "contr.poly"))
drop1(aov(dvocabul~sex*income, irish), ~. , test="F")
drop1(aov(dvocabul~type, irish), ~. , test="F")
```

mit folgendem Ergebnis für die Analyse der Effekte von `sex` und `income`:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			262.58	-1580.8		
sex	1	0.6298	263.21	-1580.2	2.6408	0.1044
income	2	12.5531	275.13	-1533.1	26.3175	6.843e-12 ***
sex:income	2	0.4187	263.00	-1583.0	0.8777	0.4160

sowie für die Analyse des Effekts von `type` :

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			261.52	-1591.3		
type	2	15.009	276.53	-1533.5	31.68	4.186e-14 ***

### mit SPSS

Zunächst muss die 4-stufige abhängige Variable `vocabula` dichotomisiert werden (Variable `dvocabul`), bevor „wie gewohnt“ mit `Unianova` die parametrische Varianzanalyse darauf angewandt wird.

```
compute dvocabula=vocabula gt 2.
Unianova dvocabula by Sex Income
  /Design = Sex Income Sex*Income.
Unianova dvocabula by Type
  /Design = Type.
```

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
sex	,630	1	,630	2,641	,104
income	12,553	2	6,277	26,317	,000
sex * income	,419	2	,209	,878	,416
Fehler	262,581	1101	,238		
type	15,009	2	7,505	31,680	,000
Fehler	261,524	1104	,237		

### 7.1.2 Gemischte Versuchspläne

Gemäß den eingangs gemachten Empfehlungen werden zur Varianzanalyse 2 verschiedene Methoden angewandt: die „normale“ zum Test der Haupteffekte und entweder das ATS-Verfahren zum Test der Messwiederholungseffekte, oder ersatzweise die multivariate Varianzanalyse. Als Beispiel wird hier der Datensatz 8 von Koch verwendet, der zum einen eine dichotome abhängige Variable (`outcome`) und zum anderen 2 Gruppierungsfaktoren (`severity` und `treat`) sowie einen Messwiederholungsfaktor (`time`) beinhaltet. `outcome` hat mit 48 bzw. 52 Prozent ideale relative Häufigkeiten. Der Mauchly-Test auf Varianzhomogenität (genauer Sphärität) entfällt hier wie oben bereits erläutert. Damit erübrigen sich auch die in Kapitel 5.1 erwähnten robusten Tests von Huynh & Feldt bzw. Greenhouse & Geisser.

#### mit R:

Der Datensatz muss zwar nicht umstrukturiert werden, jedoch die Variablen `severity`, `treat`, `time012` sowie `case_id` als Faktoren deklariert werden. Darüber hinaus muss gegebenenfalls `outcome` über `as.numeric` numerische Werte erhalten. Zunächst erfolgt die „normale“ Varianzanalyse zum Test der Effekte der Gruppierungsfaktoren `severity` und `treat`, hier einmal über `ezANOVA`, wobei zu beachten ist, dass wegen ungleicher Zellenbesetzungszahlen über `type=3` die Quadratsummen vom Typ III angefordert werden müssen:

```
ezANOVA (koch, outcome, case_id, between=.(severity,treat),
  within=time012, type=3)
```

	Effect	DFn	DFd	F	p
2	severity	1	336	90.89621790	3.166354e-19
3	treat	1	336	40.81026220	5.591147e-10
5	time012	2	672	60.68707191	5.908176e-25
4	severity:treat	1	336	0.09022516	7.640769e-01
6	severity:time012	2	672	2.68142786	6.919789e-02
7	treat:time012	2	672	12.79599590	3.515413e-06
8	severity:treat:time012	2	672	0.41843893	6.582447e-01

`Sphericity Corrections`					
	Effect	GGe	p[GG]	HFe	p[HF]
	time012	0.9981284	6.503010e-25	1.004088	5.908176e-25
	severity:time012	0.9981284	6.930510e-02	1.004088	6.919789e-02
	treat:time012	0.9981284	3.577777e-06	1.004088	3.515413e-06
	severity:treat:time012	0.9981284	6.578639e-01	1.004088	6.582447e-01

Für den Test der Effekte der Messwiederholungsfaktoren wird noch das ATS-Verfahren (vgl. auch 6.8) eingesetzt:

```
nparLD(outcome~severity*treat*time012,koch,koch$case_id)$ANOVA.test
```

	Statistic	df	p-value
severity	90.52414737	1.000000	1.827331e-21
treat	40.64321129	1.000000	1.827209e-10
time012	62.50376884	1.999345	7.297548e-28
severity:treat	0.08985584	1.000000	7.643605e-01
severity:time012	2.76169770	1.999345	6.320280e-02
treat:time012	13.17905024	1.999345	1.895950e-06
severity:treat:time012	0.43096510	1.999345	6.498157e-01

In diesem Fall decken sich allerdings die Ergebnisse für die 4 Effekte, bei denen `time012` involviert ist, zum einen bei den Huynh-Feldt-Tests ( $p[HF]$ ) der parametrischen Analyse, zum anderen bei dem ATS-Verfahren.

### mit SPSS:

Gemäß den eingangs gemachten Empfehlungen wird für die Effekte der Gruppierungsfaktoren eine „normale“ Varianzanalyse und für die Effekte der Messwiederholungsvariablen ersatzweise ein multivariater Test angewandt. Für eine Varianzanalyse mit Messwiederholungen muss der Datensatz in die entsprechende Form umstrukturiert werden (vgl. Anhang 1.2), wobei die Messwiederholungsvariablen `outcome.0`, `outcome.1`, `outcome.2` entstehen. Die Syntax für die Anova lautet dann:

```
GLM outcome.0 outcome.1 outcome.2 BY severity treat
  /WSfactor=Zeit 3 Polynomial
  /WSdesign=Zeit
  /Design=severity treat severity*treat.
```

Nachfolgend zunächst die Tabelle für die Effekte der Gruppierungsfaktoren `severity` und `treat`, danach die Tabelle der Effekte mit dem Faktor `Zeit`, wobei die Zeile mit den Huynh-Feldt-adjustierten Werten von Interesse ist, sowie die multivariaten Tests, die den Huynh-Feldt-Tests vorzuziehen ist:

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	296,013	1	296,013	1510,983	,000
severity	17,807	1	17,807	90,896	,000
treat	7,995	1	7,995	40,810	,000
severity * treat	,018	1	,018	,090	,764
Fehler	65,825	336	,196		

Tests der Innersubjekteffekte						
Quelle		Quadrat summe	df	Mittel der Quadrate	F	Sig.
Zeit	Huynh-Feldt	23,844	2	11,922	60,687	,000
Zeit * severity	Huynh-Feldt	1,054	2	,527	2,681	,069
Zeit * treat	Huynh-Feldt	5,028	2	2,514	12,796	,000
Zeit * severity * treat	Huynh-Feldt	,164	2	,082	,418	,658
Fehler(Zeit)	Huynh-Feldt	132,017	672	,196		

Schließlich die Tabelle der multivariaten Tests, nachfolgend der von Hotelling-Lawley, zur Beurteilung der Effekte des Messwiederholungsfaktors Zeit, die sich weitgehend mit denen der o.a. Huynh-Feldt-Tests decken:

Effekt	Wert	F	Hypothese df	Fehler df	Sig.
Zeit	,269	61,789 <sup>b</sup>	2,000	335,000	,000
Zeit * severity	,016	2,660 <sup>b</sup>	2,000	335,000	,071
Zeit * treat	,073	13,184 <sup>b</sup>	2,000	335,000	,000
Zeit * severity * treat	,003	,429 <sup>b</sup>	2,000	335,000	,652

## 7.2 Anwendung der Verfahren für ordinale Merkmale

Zur 1-faktoriellen Varianzanalyse eines dichotomen Merkmalms verwendet man üblicherweise den  $\chi^2$ -Test im Fall eines Gruppierungsfaktors bzw. Cochrans Q-Test im Fall eines Messwiederholungsfaktors. Diese Tests sind aber nichts anderes als der Kruskal-Wallis H-Test bzw. die Friedman-Varianzanalyse, wenn man bei diesen die ordinale Variable nur zwei Werte annehmen lässt und die Bindungskorrekturen verwendet. Somit lassen sich trivialerweise die in den Kapiteln 4.3.5, 5.3.4 und 6.4 beschriebenen Puri & Sen-Tests auf dichotome Merkmale anwenden.

Ferner weisen Akritas, Arnold und Brunner (1997) und Akritas, Arnold & Brunner (1997) ausdrücklich darauf hin, dass ihre ATS (Anova type statistic) nicht nur für ordinale, sondern auch für dichotome Merkmale anwendbar sind. Im Gegensatz zur u.a. Logistischen Regression können diese Methoden auch bei Messwiederholungen angewandt werden.

Auf Beispiele soll hier verzichtet werden, da die Anwendung dieser Verfahren in den vorangegangenen Kapiteln ausführlich beschrieben wurde.

## 8. Logistische Regression

### 8. 1 dichotome abhängige Variablen

Die bekannteste logistische Regression ist die *binär-logistische Regression*, bei der ein Modell mit einer dichotomen (d.h. binären) abhängigen Variablen  $y$  (mit Werten 0 und 1) und  $v$  Prädiktoren  $x_1, x_2, \dots, x_v$  aufgestellt wird. Typischerweise ist dabei die abhängige Variable nicht  $y$  selbst, sondern  $P(y=1)$ , d.h. die Wahrscheinlichkeit, dass  $y$  den Wert 1 annimmt. Dadurch ist der Wertebereich der Funktion das komplette Intervall  $[0,1]$ :

$$P(y = 1) = \frac{e^{b_0 + b_1 x_1 + \dots + b_v x_v}}{1 + e^{b_0 + b_1 x_1 + \dots + b_v x_v}}$$

Für die unabhängigen Variablen (Prädiktoren) gelten die üblichen Bedingungen, d.h. für nominale Prädiktoren müssen Kontrastvariablen gebildet werden.

Zum weiteren Verständnis im Kontext der Varianzanalyse ist es an dieser Stelle nicht erforderlich, auf dieses Modell näher einzugehen. Die logistische Regression ist inzwischen soweit etabliert, dass sie in vielen einführenden Statistik-Lehrbüchern beschrieben wird. Eine Einführung bieten z.B. Diaz-Bone & Künemund (2003) oder auch Wikipedia.

Allerdings ist an dieser Stelle noch nicht die Beziehung zur Varianzanalyse direkt erkennbar. Dazu sei angemerkt, dass die (parametrische) Varianzanalyse nichts anderes als eine lineare Regression mit nominalen Prädiktoren ist, nämlich den Faktoren, die wie oben angedeutet in Kontrastvariable transformiert werden. Und wenn genau diese Transformation bei der binären oder ordinalen logistischen Regression angewandt wird, erhält man ein Modell für eine dichotome oder ordinale Varianzanalyse. Hierbei gibt es jedoch einen Stolperstein: Für die Transformation der nominalen Faktoren in Kontraste gibt es zahlreiche Lösungen (vgl. Kapitel 9.1.2), die allerdings hinsichtlich der Tests der einzelnen Kontraste nicht immer zu demselben Ergebnis führen. Hinzu kommt, dass zunächst einmal, wie bei der Regression üblich, der Effekt jeder einzelnen Kontrastvariablen separat getestet wird. Einige Programme, insbesondere der binär-logistischen Regression ohne Messwiederholungen, fassen allerdings die Tests für die Kontrastvariablen eines Faktors zu einem Gesamtergebnis zusammen, z.B. mit dem Wald-Test (vgl. Kapitel 9.8), woraus der Effekt dieses Faktors zu entnehmen ist. Wünschenswert wäre, dass dieser globale Effektttest von dem gewählten Kontrasttyp unabhängig ist. Doch das ist nur beim 1-faktoriellen Modell sowie bei einer 2-faktoriellen Analyse für die Interaktion der Fall. Die Wahl der Kontraste bietet zwar eine Reihe von Möglichkeiten, auf die allerdings in diesem Kontext nicht eingegangen werden soll. Für die hier im Fokus stehenden varianzanalytischen Fragestellungen wird empfohlen, sofern nicht anders vermerkt, für alle Faktoren die Kontraste zu wählen, die man in R mittels `contr.sum` bzw. in SPSS über `deviation` (vgl. Kapitel 9.2 sowie 3.1) erhält. Andernfalls läuft man Gefahr, Ergebnisse falsch zu interpretieren.

Ein Nachteil gegenüber den o.a. varianzanalytischen Verfahren liegt in der nicht immer befriedigenden Möglichkeit zur Behandlung von Messwiederholungen. Auf der anderen Seite gibt es die Möglichkeit zur Verarbeitung von Versuchsplänen mit leeren Zellen. Wie auch insgesamt die Logistische Regression relativ liberal hinsichtlich der Voraussetzungen ist. Schaut man in die Literatur, so sucht man vergebens nach „handfesten“ Voraussetzungen, obwohl die zur Lösung eingesetzte Maximum-Likelihood-Methode sehr sensibel ist. (So kann es durchaus vorkommen, dass keine Lösung gefunden werden kann, weil die mathematische Schätzmethode nicht konvergiert. Das liegt an der mathematischen „Kondition“. Denn im Gegensatz zur Varianzanalyse wird die Lösung der Logistischen Regression nicht „direkt“ errechnet, sondern über

ein Iterationsverfahren näherungsweise gefunden. Oder aber auch nicht.) Um Probleme zu vermeiden, sind nur zwei Dinge zu beachten:

- ein hinreichend großer Stichprobenumfang  $n$ , mindestens 10 pro Prädiktor bzw. geschätztem Parameter (wobei die Empfehlungen, sofern erwähnt, zum Teil stark divergieren). Da bei der Varianzanalyse ein Faktor als nominal skalierte Variable mit  $I$  Merkmalsausprägungen in  $(I-1)$  Kontrastvariable transformiert und für die Interaktionen auch deren Produkte als Prädiktoren verwendet werden, bedeutet das für das  $n$ : ca.  $10 \cdot (\text{Anzahl der Zellen})$ .
- ein „vernünftiges“ Modell, d.h. u.a. ohne überflüssige (nicht erklärende) und ohne kollineare Variablen. Diese Forderung erübrigt sich allerdings beim Einsatz als Varianzanalyse.

Mit der logistischen Regression sind i.a. drei Signifikanztests verbunden:

- Ein Test des gesamten Modells, d.h. aller Effekte zusammen, über einen  $\chi^2$ -Test des log likelihood-Wertes. Sind Effekte der Faktoren vorhanden, so sollte dieser Test signifikant sein.
- Ein „klassischer“  $\chi^2$ -Anpassungstest des Modells, der also prüft, in wieweit die Daten mit dem Modell vereinbar sind. Dieser sollte nicht signifikant sein.
- Die Signifikanzüberprüfung eines Regressionskoeffizienten (auf Verschiedenheit von 0) oder eines Effekts über die Wald-Statistik mittels des  $\chi^2$ -Tests.

Bei der binär-logistischen Regression wird zunächst für jeden Regressionskoeffizienten bzw. Kontrast ein Wald-Test automatisch ausgegeben, womit man noch kein Ergebnis für einen varianzanalytischen Effekt hat. Hierzu dienen die in Kapitel 9.8 besprochenen Wald- und LR-Tests. Bei der ordinalen Regression müssen die Wald-Tests recht aufwändig angefordert werden. Da kann es nützlich sein, über die Modell-Tests vorab zu erfahren, ob dieser Aufwand überhaupt erforderlich ist.

Hierbei wird darauf hingewiesen, dass der LR-Test bei kleinem  $n_i \leq 10$  sowie beim Test der Interaktion sehr liberal reagiert (mit Fehlerraten bis zu 20%), während der Wald-Test sich in solchen Fällen sehr konservativ verhält. Dem kann man begegnen, indem die  $\chi^2$ -Werte beider Tests gemittelt werden und dann dieser Mittelwert, der bei 1 FG  $\chi^2$ -verteilt ist, per Hand auf Signifikanz überprüft wird. Darüberhinaus verletzen beide Tests das  $\alpha$ -Risiko für den Test eines Haupteffekts, wenn ein Interaktionseffekt vorhanden ist. Hier steigt die reale Fehlerrate sogar bis auf 30-40% bei einem  $n=50$ . Dies macht die logistische Regression zur Durchführung von Varianzanalysen unattraktiv (vgl. Lüpsen, 2021).

Als Beispiel wird hier wie in Kapitel 7.1.1 der Datensatz 7 mit `dvocubul`, der dichotomisierten Variable `vocubula` (Wortschatz), als abhängige Variable verwendet. Mit Hilfe der Logistischen Regression können allerdings alle drei Einflussfaktoren simultan untersucht werden, was die Interpretation der Effekte nicht gerade vereinfacht. Allerdings werden die Interaktionen `sex*type` und `income*type` auch hier weggelassen, die die beteiligten Faktoren nicht unabhängig voneinander sind. Für die oben angesprochene Transformation der Faktoren in Kontrastvariablen wird hier, wie in der Varianzanalyse üblich, die Effekt-Kodierung („Deviation“) vorgenommen. Mit dem Test eines Kontrasts wird dann die Abweichung der entsprechenden Ausprägung vom Mittelwert getestet. Alternativ könnten auch die einfache Kodierung gewählt werden, bei der Unterschiede einer Ausprägung zur letzten Ausprägung getestet werden. Die Anzahl von Zellen beträgt 36, so dass ein  $n$  von ca. 360 wünschenswert ist, was mit 1107 mehr als erfüllt ist.

**mit R:**

Zur Logistischen Regression bietet R u.a. die Funktion `glm` an. Hierbei ist die Angabe der Verteilungsfamilie `binomial` als Fehlerverteilung erforderlich, um das logistische Regressionsmodell zugrunde zu legen. Die oben angesprochene Effekt-Kodierung der Faktoren wird hier über den Parameter `contr.sum` der `options`-Anweisung vorgenommen. Die `Anova`-Funktion (Paket `car`) erlaubt hier die Ausgabe einer Anova-Tabelle:

```
options(contrasts=c("contr.sum", "contr.poly"))
irish.glm <- glm(dvocabulary~sex+income+type+sex:income,
               family=binomial, irish)
Anova(irish.glm, test="Wald", type="III")
```

	Df	Chisq	Pr(>Chisq)	
(Intercept)	1	9.9765	0.001586	**
sex	1	0.3529	0.552462	
income	2	19.7510	5.142e-05	***
type	2	38.3248	4.763e-09	***
sex:income	2	1.8746	0.391690	

Fordert man über `summary(...)` eine Zusammenfassung der Ergebnisse, erhält man eine Tabelle der Einzelvergleiche, bei denen jeweils eine Stufe eines Faktors gegen den Mittelwert verglichen wird:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-0.74771	0.23672	-3.159	0.001586	**
sex1	0.19568	0.32939	0.594	0.552462	
income1	0.93233	0.27527	3.387	0.000707	***
income2	0.14968	0.26591	0.563	0.573507	
type1	0.27648	0.10134	2.728	0.006365	**
type2	0.65142	0.11727	5.555	2.78e-08	***
sex1:income1	-0.05088	0.39064	-0.130	0.896375	
sex1:income2	0.30780	0.37303	0.825	0.409304	

**mit SPSS:**

Die Logistische Regression ist in SPSS über das Menü „Regression -> binär logistisch...“ erreichbar. Nominale Prädiktoren, also Faktoren, müssen in der Menü-Oberfläche als „kategorial“ vereinbart werden. Hierbei bietet SPSS die Möglichkeit, für jeden Faktor die Kontraste individuell zu wählen. Meistens empfiehlt es sich, den Typ „Deviation“ zu wählen, da dann die Tests, die für alle Stufen (bis auf die letzte) ausgegeben werden, die Abweichungen der jeweiligen Kategorie vom Mittelwert überprüfen. Interaktionen müssen explizit angefordert werden. Die Syntax für die Analyse lautet:

```
compute dvocabula=vocabula gt 2.
Logistic regression variables=dvocabula
  /method = enter Sex Income Type Sex*Income
  /contrast(Sex) =Deviation
  /contrast(Income) =Deviation
  /contrast(Type) =Deviation
```



	Regr.koeff B	Standardfehler	Wald	df	Sig.	Exp(B)
sex(1)	,141	,073	3,764	1	,052	1,151
income			27,460	2	,000	
income(1)	,503	,096	27,308	1	,000	1,654
income(2)	-,100	,088	1,284	1	,257	,905
type			38,325	2	,000	
type(1)	,276	,101	7,444	1	,006	1,318
type(2)	,651	,117	30,854	1	,000	1,918
income * sex			1,875	2	,392	
income(1) by sex(1)	-,068	,094	,524	1	,469	,934
income(2) by sex(1)	,111	,088	1,589	1	,208	1,117
Konstante	-,246	,092	7,187	1	,007	,782

Oben die Ergebnistabelle für alle Effekte, in der sowohl die globalen Tests als auch die der einzelnen Kontraste (Variablennamen mit (..)), den Einzelvergleichen der ( $K-1$ ) ersten Stufen eines Faktors gegen den Mittelwert (aller Stufen), enthalten sind.

## 8. 2 ordinale abhängige Variablen

Das Modell der binär-logistischen Regression lässt sich in ein Modell für eine ordinale abhängige Variable  $y$  verallgemeinern, indem nicht mehr  $P(y=1)$ , sondern  $P(y \leq j)$  als die abhängige Variable verwendet wird, mit  $j=1, \dots, m$ , wenn  $m$  die Anzahl der Merkmalsausprägungen von  $y$  ist:

$$P(y \leq j) = \frac{e^{b_{0j} + b_{1j}x_1 + \dots + b_{vj}x_v}}{1 + e^{b_{0j} + b_{1j}x_1 + \dots + b_{vj}x_v}}$$

( $v$  ist wieder die Anzahl der Prädiktoren.) Während bei der binär-logistischen Regression nur eine Modellgleichung aufgestellt wird, sind es bei der ordinalen  $m-1$  Modellgleichungen. D.h. es müssten  $(m-1)*v$  Parameter geschätzt werden. Dieses Modell wird üblicherweise vereinfacht, indem für jeden Prädiktor  $i$  ( $i=1, \dots, v$ ) die Koeffizienten der jeweiligen Merkmalsausprägungen als gleich angenommen werden:  $b_{i1}=b_{i2}=\dots=b_{i(m-1)}$ . Dies Modell heißt dann *proportional odds model*.

Zu den Voraussetzungen der dichotomen logistischen Regression kommt im Falle ordinaler Kriteriumsvariablen allerdings erschwerend die Anzahl der Ausprägungen von  $y$  hinzu, weil sich dadurch die Anzahl der Zellen vervielfacht. Daher ist dieses Verfahren i.a. nur für abhängige Variablen  $y$  mit 3 bis 5 Ausprägungen empfehlenswert.

Wie kann man sich die Bedingung gleicher Regressionskoeffizienten vorstellen? Dazu ein Beispiel: Eine Aufgabe wird mit Schulnoten 1 bis 6 beurteilt, und es soll der Einfluss von Geschlecht und Alter untersucht werden. Hinsichtlich des Geschlechts besagt die Bedingung: Wenn sich Mädchen und Jungen bei guten Noten (1 und 2) unterscheiden, dann unterscheiden sie sich auch bei guten bis mittleren Noten (1 bis 3) sowie bei guten bis schwachen (1 bis 4). Oder umgekehrt: wenn sie sich in einer Gruppe nicht unterscheiden, dann auch in keiner anderen. Die Gleichheit der Koeffizienten geht sogar noch soweit, dass die Mädchen-Jungen-Unterschiede in allen Notengruppen gleich groß sind. Ähnlich verhält es sich mit dem Alter. Wenn mit zunehmendem Alter die Wahrscheinlichkeit für eine gute Note steigt, dann gilt das ebenso für die Wahrscheinlichkeit einer guten bis mittleren Note oder einer nicht schlechten Note (1 bis 4).

Für die Anwendung des *proportion odds model* muss allerdings die Gleichheit der Koeffizienten mit den Daten vereinbar sein. Das wird mit dem „Parallelitätstest für Linien“ (*parallel lines test*) überprüft. Bei diesem werden die Abweichungen (ähnlich den Residuen) beider Modelle (einmal mit gleichen und einmal mit individuellen Koeffizienten) verglichen. Fällt dieser signifikant aus, bedeutet dies zunächst, dass die individuellen Koeffizienten eine signifikante Verbesserung der Anpassung erbringen. Das heißt aber, dass das vereinfachte Regressionsmodell nicht angewandt werden kann. Um diesen Test durchzuführen, müssen allerdings alle  $(m-1) \cdot v$  Parameter geschätzt werden, was ein hinreichend großes  $n$  erfordert. R bietet allerdings mit der Funktion `vglm` im Paket `VGAM` auch eine Lösung des o.a. Modells, bei dem die Gleichheit der Koeffizienten nicht gefordert wird.

Wenn für den Test ohnehin schon das Modell mit den individuellen Koeffizienten geschätzt werden muss, dann könnte man ja einfach damit anstatt mit dem vereinfachten Modell arbeiten. Nur: man hat dann eine riesige Anzahl von Koeffizienten, die einzeln kaum interpretierbar sind. Für einen Faktor mit  $I$  Gruppen (Stufen) resultieren alleine  $(I-1)(m-1)$  Koeffizienten. Daher ist man bestrebt, das Modell mit gleichen Koeffizienten zu wählen.

Aber damit sind noch nicht alle Probleme aus dem Weg geräumt. Sollte man „zufällig“ ein Modell zum einen mit R und zum anderen mit SPSS rechnen, so wird man direkt irritiert sein, dass die Ergebnisse überhaupt nicht in Einklang zu bringen sind. Die Ursache: Das Modell ist ja zunächst einmal ein Regressionsmodell. Bei diesem werden in beiden Fällen automatisch Faktoren, d.h. nominale Prädiktoren, in Kontraste transformiert (vgl. Kapitel 9.1). Doch die Wahl des Kontrastes fällt bei beiden Programmen verschieden aus: R nimmt standardmäßig „einfache“ Kontraste mit der ersten Gruppe als Referenzgruppe, SPSS zwar auch „einfache“ Kontraste, aber mit der letzten Gruppe als Referenzgruppe. Dadurch fallen die Tests der Kontraste verschieden aus. Erschwerend kommt hinzu, dass beide Programme apriori neben den Einzeltests der Kontraste keinen globalen, zusammenfassenden Test ausgeben, aus dem der Effekt eines Faktors abzulesen wäre. Sowohl bei SPSS als auch bei R kann allerdings solcher ein Test angefordert werden.

Als Beispiel wird hier der Datensatz 7 (`irish`) benutzt, und zwar soll der Einfluss von Geschlecht (`sex`) und Schultyp (`type`) auf den Wortschatz (`vocabula`) untersucht werden.

### mit R:

In R stehen eine Reihe von Funktionen zur ordinalen logistischen Regression zur Verfügung, u.a.:

- `polr` (*Modell*, `data=Dataframe`) aus dem Paket `MASS`
- `clm` (*Modell*, `data=Dataframe`) aus dem Paket `ordinal`
- `vglm` (*Modell*, `family=cumulative(parallel=T/F)`) aus dem Paket `VGAM`, die sowohl das vereinfachte Modell (`parallel=T`) als auch das Modell mit individuellen Regressionskoeffizienten (`parallel=F`) handhaben kann.
- `npmlt` (*Modell*, `link="clogit"`) aus dem Paket `mixcat`

R bietet zum einen die Funktion `Anova` (Paket `car`) für globale Tests der Effekte. Alternativ wird hier gezeigt, wie er sich näherungsweise aus den Tests für die einzelnen Kontraste des Faktors ermittelt lässt, wie in Kapitel 9.8 näher beschrieben.

Nachfolgend die Anweisungen für die ordinale Regression, hier mit `clm`, wobei zu beachten ist, dass nicht nur die Faktoren (hier `sex` und `type`) vom Typ „factor“ sein müssen, sondern auch die abhängige Variable vom Typ „ordered factor“. Die `options`-Anweisung

bewirkt, dass bei der Transformation der Faktoren das Effekt-Kodieren (`contr.sum`) angewandt wird.

```
irish <- within(irish, {vocabulary<-ordered(vocabulary);
                      sex<-factor(sex); type<-factor(type)} )
options(contrasts=c("contr.sum", "contr.poly"))
lr.clm <- glm(vocabulary~sex*type, data=irish)
summary(lr.clm)
Anova(lr.clm, test="Chisq")
```

mit folgender Ausgabe für die Koeffizienten sowie die Anova-Tabelle:

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z )
sex1	-0.20542	0.14399	-1.427	0.1537
type1	-0.07987	0.12311	-0.649	0.5165
type2	0.93545	0.15661	5.973	2.33e-09 ***
sex1:type1	0.73844	0.16396	4.504	6.67e-06 ***
sex1:type2	-0.48368	0.20039	-2.414	0.0158 *

	Df	Chisq	Pr(>Chisq)
sex	1	193.1891	<2e-16 ***
type	2	367.3082	<2e-16 ***
sex:type	2	2.0626	0.3565

Darüber hinaus werden noch die Koeffizienten  $b_{oj}$  ausgegeben, die aber für die Interpretation des Einflusses von `sex` und `type` ohne Bedeutung sind:

Threshold coefficients:			
	Estimate	Std. Error	z value
bad poor	-1.04596	0.08457	-12.367
poor good	0.13791	0.07852	1.756
good excellent	1.18223	0.08500	13.909

Zu den Tests der Effekte:

- Der Effekt von `sex` ist direkt aus der Tabelle mit  $p_{\text{sex}} = 0.1537$  ablesbar, da der Faktor nur zwei Stufen hat.
- Der Effekt von `type` wird aus den z-Werten der Kontraste `type1` und `type2` ermittelt:  
 $\chi^2_{\text{type}} = 0.649^2 + 5.973^2 = 36.1$   
 und die Signifikanzprüfung ergibt  $p_{\text{type}} < 0.001$ .
- Der Effekt von `sex*type` wird aus den z-Werten der entsprechenden Kontraste ermittelt:  
 $\chi^2_{\text{sex*type}} = 4.504^2 + 2.414^2 = 26.1$   
 und die Signifikanzprüfung ergibt  $p_{\text{sex*type}} < 0.001$ .

Bleibt noch zu prüfen, ob das vereinfachte *proportional odds model* überhaupt angewandt werden darf. Dazu wird mit der Funktion `vglm` einmal das einfache Modell (`fit1`) gerechnet und einmal das Modell mit den individuellen Koeffizienten (`fit2`). Der Unterschied der Anpassungsgüte wird mittels der Maßzahl „deviance“ auf Signifikanz überprüft:

```
fit1 <- vglm(vocabula~sex*type, data=irish, family=cumulative(parallel=T))
fit2 <- vglm(vocabula~sex*type, data=irish, family=cumulative(parallel=F))
pchisq(deviance(fit1)-deviance(fit2),
       df=df.residual(fit1)-df.residual(fit2), lower.tail=F)
```

Der p-Wert von 0.196 indiziert die Verträglichkeit des vereinfachten Modells mit den Daten.

Möchte man oben die Quadrierung der z-Werte direkt aus dem Ergebnisobjekt der Funktion `c1m` (oder einer der anderen Funktionen) vornehmen, so steht man vor einem kleinen Problem. Die Ausgabe von `summary` erlaubt leider nicht die Adressierung der einzelnen Spalten. Und die Ergebnisobjekte selbst sind äußerst komplex aufgebaut. Einzig `npmlt` bietet die Koeffizienten und Standardfehler als Ergebnisobjekte an:

```
erg <- npmlt(irish$vocabula~irish$sex*irish$type)
zvalues <- erg$coefficients/erg$SE.coefficients
chisq <- zvalues^2
```

Abschließend sei noch angemerkt, dass nicht nur die Eingabe der o.a. 4 Funktionen für die Analyse der ordinalen Regression quasi identisch ist, sondern gleichermaßen die Ausgabe.

### mit SPSS:

In SPSS steht für die ordinale logistische Regression der Modul `PLUM` (*polytomous universal model*) zur Verfügung, im Menü über Regression -> Ordinal. Faktoren, d.h. nominal skalierte Prädiktoren mit  $K$  Ausprägungen, werden automatisch in  $K-1$  Kontraste transformiert (vgl. Kapitel 9.1), derart dass für diese  $b_i=0$  getestet wird. Die oben erwähnten globalen Tests der Effekte sind nur über die Syntax anforderbar. Für den Faktor `sex` erübrigt sich solch ein Test, da für eine 2-stufige Variable dieser mit dem Test des Koeffizienten identisch ist.

```
PLUM vocabula BY sex type
  /link = logit
  /location = sex type sex*type
  /print = fit parameter summary tparallel
  /test (0,0) = type 1 0 0;
                type 0 1 0
  /test (0,0) = sex*type 1 0 0 0 0 0 ;
                sex*type 0 1 0 0 0 0 .
```

Erläuterungen hierzu: Über `location` werden die zu testenden Effekte angegeben. Über `test` wird jeweils ein globaler Effekttest angefordert, wobei auf der rechten Seite so viele Kontraste aufgeführt werden müssen, wie Parameter geschätzt werden, also  $(I_A-1)$  (mit  $I_A$  als Anzahl Stufen/Gruppen von Faktor A). Für jeden Kontrast wird hinter `test` ein Hypothesenwert in `(..)` angegeben, also i.a. 0. Bei Interaktionen beträgt die Anzahl der Kontraste  $(I_A-1)(I_B-1)$  mit jeweils  $I_A \cdot I_B$  Kontrastkoeffizienten. Als Koeffizienten werden zweckmäßigerweise nur 0 und 1 gewählt, wodurch die Hypothese lautet: alle Koeffizienten sind gleich 0.

Das wesentliche Ergebnis steckt in der Tabelle der Regressionsparameter, oben unter „Schwelle“ die Parameter  $b_{0j}$  sowie unter „Lage“ die Parameter  $b_i$ , die nach Annahme nicht von der Merkmalsausprägung  $j$  abhängen. Durch die nominalen Prädiktoren und deren Transformation in  $(I-1)$  Kontraste und damit  $(I-1)$  Parameter sind davon einige redundant, die dann mit 0 ausgegeben werden.

Bei den „globalen“ Effekttests werden zunächst die Kontraste noch einmal einzeln getestet, deren Ergebnis mit den o.a. identisch ist. Anschließend folgen die gewünschten Gesamttests. Auf welchen Faktor sich diese beziehen, ist nur über die davor angezeigten Kontrastkoeffizienten erkennbar. Also unten zunächst der Test für *type*, danach für *sex\*type*:

		Parameterschätzer						
		Schätzer	Standardfehler	Wald	Fg	Sig.	Konfidenzintervall 95%	
							Unterg.	Oberg.
Schwelle	[vocabula = 1]	-,293	,275	1,139	1	,286	-,831	,245
	[vocabula = 2]	,891	,276	10,414	1	,001	,350	1,432
	[vocabula = 3]	1,935	,280	47,694	1	,000	1,386	2,484
Lage	[sex=1]	-,460	,341	1,823	1	,177	-1,128	,208
	[sex=2]	0 <sup>a</sup>	.	.	0	.	.	.
	[type=1]	,776	,288	7,246	1	,007	,211	1,341
	[type=2]	1,791	,333	28,937	1	,000	1,138	2,444
	[type=3]	0 <sup>a</sup>	.	.	0	.	.	.
	[sex=1] * [type=1]	,993	,367	7,337	1	,007	,275	1,712
	[sex=1] * [type=2]	-,229	,415	,304	1	,581	-1,043	,585
	[sex=1] * [type=3]	0 <sup>a</sup>	.	.	0	.	.	.
	[sex=2] * [type=1]	0 <sup>a</sup>	.	.	0	.	.	.
	[sex=2] * [type=2]	0 <sup>a</sup>	.	.	0	.	.	.
	[sex=2] * [type=3]	0 <sup>a</sup>	.	.	0	.	.	.

Testergebnisse		
Wald	Freiheitsgrade	Sig.
35,100	2	,000

Testergebnisse		
Wald	Freiheitsgrade	Sig.
23,614	2	,000

Von besonderem Interesse ist noch der Parallelitätstest. Da dieser nicht signifikant ist, darf das vereinfachte *proportional odds model* angewandt werden.

Parallelitätstest für Linien <sup>a</sup>				
Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Sig.
Nullhypothese	99,933			
Allgemein	86,421	13,511	10	,196

Die Nullhypothese gibt an, daß die Lageparameter (Steigungskoeffizienten) über die Antwortkategorien übereinstimmen.

Was passiert, wenn das *n* bezogen auf die Anzahl der Zellen nicht ausreichend ist? Wollte man z.B. eine ordinale Regression mit den Daten des Beispiels 2 (*mydata2*) rechnen, dann stößt man auf dieses Problem: Die Kriteriumsvariable hat 8 Ausprägungen und das Design hat 8 Zellen, also gibt es insgesamt 64 Zellen. Aber auf der anderen Seite nur 33 Beobachtungen. Man könnte zunächst das Problem abmildern, indem Merkmalsausprägungen der abhängigen Variablen zusammengefasst werden, z.B. von 8 auf 4 reduzieren. Das kann gelegentlich gut gehen, in die-

sem Fall aber nicht. Es kann nämlich keine „gesicherte“ Lösung gefunden werden. Sowohl R als auch SPSS geben in solchen Fällen Warnungen aus, etwa in R:

```
Warning message:
(1) Hessian is numerically singular: parameters are not uniquely
determined
In addition: Absolute convergence criterion was met, but relative
criterion was not met
```

oder in SPSS:

Warnungen
Es gibt 15 (46.9%) Zellen (also Niveaus der abhängigen Variablen über Kombinationen von Werten der Einflußvariablen) mit Null-Häufigkeiten.
Es wurden unerwartete Singularitäten in der Fisher-Informationsmatrix gefunden. Möglicherweise liegt eine quasi-vollständige Trennung der Daten vor. Einige Parameter werden sich Unendlich nähern.
Die PLUM-Prozedur wird trotz der obigen Warnung(en) fortgesetzt. Die anschließend angezeigten Ergebnisse basieren auf der letzten Iteration. Die Zulässigkeit der Anpassungsgüte des Modells ist unsicher.

Zwar kann sowohl in R als auch in SPSS die Anzahl der Iterationen zur Berechnung der Lösung vergrößert werden, was aber selten hilft. In solchen Fällen kann nur davon abgeraten werden, die Ergebnisse zu verwenden.

### 8.3 dichotome abhängige Variablen und Messwiederholungen

Es gibt Methoden für die logistische Regression mit dichotomen Kriteriumsvariablen, wenn diese für die Versuchspersonen mehrfach, z.B. unter verschiedenen Versuchsbedingungen, erhoben worden sind, also bei Messwiederholungen. Zu nennen sind hier die in 2.15 vorgestellten *Generalized Linear Mixed-Effects Models* (GLMM) und *Generalized Estimating Equation* (GEE). Doch diese Verfahren führen sehr häufig zum Abbruch, insbesondere bei mehrfaktoriellen Versuchsplänen. Die Ursache ist meistens eine nicht ausreichend große Fallzahl. So ist es z.B. nicht immer möglich, Interaktionen mit dem Messwiederholungsfaktor zu testen.

Als Beispiel wird hier der Datensatz 4 (`winer518`) verwendet, allerdings wird die abhängige Variable dichotomisiert: 1-5->0 bzw. 6-9->1.

#### mit R:

In R gibt es u.a. die folgenden Funktionen für eine dichotome logistische Regression mit Messwiederholungen:

- `glmer` (Paket `lme4`) (GLMM-Methode)
- `glmmML` (Paket `glmmML`) (GLMM-Methode)
- `geeglm` (Paket `geepack`) (GEE-Methode)
- `gee` (Paket `gee`) (GEE-Methode)
- `geem` (Paket `geem`) (GEE-Methode)

Simulationen (vgl. Lüpsen, 2021) haben gezeigt, dass die GEE-Methode gefährlich ist, da Interaktionseffekte sich auf die Haupteffekte auswirken, d.h. die Tests sind nicht unabhängig, wie man es sonst von der Varianzanalyse gewohnt ist. Das gleiche gilt zwar

auch für GLMM, allerdings kann die Verwendung des Wald-Tests vom Typ II mittels der Funktion `Anova` (Paket `car`) den Fehler weitgehend unter Kontrolle halten, falls ein Interaktionseffekt vorhanden ist. Allerdings lässt sich diese Funktion nur auf Ergebnisse von `glmer` anwenden. Weiterhin hat sich gezeigt, dass die Teststärke (Power) von GEE und GLMM äußerst gering ist (Ausnahme: `glmer` unter Verwendung des o.a. Wald-Tests). Daher sind die in Kapitel 7 vorgeschlagenen Methoden vorzuziehen.

Die Anweisungen sind für alle Funktionen ähnlich, allerdings sind die Ergebnisse wegen der unterschiedlichen Schätzmethoden recht unterschiedlich. Es sind auch mehrere Messwiederholungs- und Gruppierungsfaktoren möglich.

Basis ist immer der umstrukturierte Datensatz, hier also `winer518t`. Es ist zu beachten, dass viele Funktionen die Kodierung 0/1 für die abhängige Variable erwarten. Zunächst wird hier `glmer` vorgestellt, allerdings nur mit der Möglichkeit zur Ermittlung der beiden Haupteffekte Geschlecht und Zeit, da bei Anforderung eines Interaktionseffektes keine Lösung gefunden werden kann. Für das Ergebnis wird mittels der Funktion `Anova` eine Anova-Tabelle erstellt. Die Eingabe:

```
winer518t[,3]<-winer518t[,3]%/%5 # Dichotomisierung
within(winer518t, {Geschlecht<-factor(Geschlecht); Zeit<-factor(Zeit);
              Vpn<-factor(Vpn) })
g <- glmer(score~Geschlecht+Zeit+(1|Vpn), data=winer518t, family=binomial)
Anova(g, test="Chisq")
summary(g)
```

	Chisq	Df	Pr(>Chisq)
Geschlecht	1.4276	1	0.23216
Zeit	5.0600	2	0.07966

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: score ~ Geschlecht + Zeit + (1 | Vpn)
Data: winer518.5
```

	AIC	BIC	logLik	deviance	df.resid
	44.2	51.2	-17.1	34.2	25

```
Scaled residuals:
   Min      1Q  Median      3Q      Max
-2.0209 -0.6201  0.0608  0.6252  2.6793
```

```
Random effects:
 Groups Name          Variance Std.Dev.
 Vpn     (Intercept)  2.752e-20 1.659e-10
Number of obs: 30, groups: Vpn, 10
```

```
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.4636     0.5990   0.774   0.4390
Geschlecht2   -1.0153     0.8498  -1.195   0.2322
Zeit.L        -1.6708     0.7740  -2.159   0.0309 *
Zeit.Q        -0.5828     0.6999  -0.833   0.4051
```

Die obige Ausgabe mittels `summary` enthält für die Gruppierungsfaktoren lediglich Tests für die einzelnen Kontraste. Bei Verwendung anderer Funktionen als `glmer` muss man gegebenenfalls aus diesen wie in Kapitel 9.8 beschrieben und in Kapitel 6.10 bereits demonstriert für einen Faktor einen Gesamttest mit der Hand ausrechnen. Für den Messwiederholungsfaktor (hier `Zeit`) wird ein Test auf linearen (`Zeit.L`) bzw. quadratischen Trend (`Zeit.Q`) ausgegeben. Hiernach besteht ein Unterschied zwischen den Zeitpunkten, aber nicht zwischen Männern und Frauen.

Die Funktion `geeglm` kann im Gegensatz zu `glmer` auch bei kleinerem  $n$  Interaktionen mit dem Messwiederholungsfaktor testen. Allerdings ist das Ergebnis nicht mit der Funktion `Anova` kompatibel, sondern nur mit `anova`, bei der die Reihenfolge der Faktoren eine Rolle spielt. Zunächst die Eingabe, wobei vorher noch die Dichotomisierung und Wandlung in den Typ `factor` wie im vorigen Beispiel vorzunehmen ist:

```
g <- geeglm(score~Geschlecht*Zeit,id=Vpn,data=winer518t,family=binomial)
summary(g)
anova(g)
```

zunächst mit der Ausgabe der Ergebnisse für die Kontraste, danach die Anova-Tabelle:

```

Coefficients:
              Estimate Std. err Wald Pr(>|W|)
(Intercept)   6.71e+00  1.51e+06  0.00   1.000
Geschlecht1   7.17e+00  1.34e+06  0.00   1.000
Zeit.L        -1.61e+00  7.57e-01  4.55   0.033 *
Zeit.Q        -1.70e+01  3.55e+06  0.00   1.000
Geschlecht1:Zeit.L  3.47e-01  7.57e-01  0.21   0.647
Geschlecht1:Zeit.Q -1.82e+01  3.69e+06  0.00   1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Estimated Scale Parameters:
              Estimate Std. err
(Intercept)   0.833   0.349

```

```

Analysis of 'Wald statistic' Table
Model: binomial, link: logit
Response: dscore
Terms added sequentially (first to last)

              Df   X2 P(>|Chi|)
Geschlecht    1  1.18   0.28
Zeit          2  4.36   0.11
Geschlecht:Zeit  2  0.21   0.90

```

Eine andere Funktion, `gee`, wurde bereits im Abschnitt 6.10.4 vorgestellt. Für die Analyse einer dichotomen Variablen ist dort lediglich bei dem Parameter `family` die Spezifikation `gaussian` durch `binomial` zu ersetzen. Allerdings wird darauf aufmerksam gemacht, dass das  $n$  für eine Schätzung nicht mehr ausreicht.

### mit SPSS:

SPSS bietet für die dichotome logistische Regression mit Messwiederholungen die Prozedur `Genlin` (GEE-Methode) an. Auch hier sind Interaktionen mit dem Messwiederholungsfaktor nicht immer möglich. SPSS erwartet hier wie in 6.10.4 beschrieben ausnahmsweise



die Daten nicht in der „normalen“ Struktur (alle Werte pro Fall in einer Zeile), sondern in der Form, in der die Werte jeder Messwiederholung in einer separaten Zeile angeordnet sein müssen. Die Umstrukturierung ist im Anhang 1.1 beschrieben. Nachfolgend zunächst die Eingabe:

```
COMPUTE dscore=score>5.

GENLIN dscore (REFERENCE=LAST)
  BY Geschlecht Zeit (order = DESCENDING)
/MODEL Geschlecht Zeit
  DISTRIBUTION=BINOMIAL
  LINK=LOGIT
/REPEATED SUBJECT=Vpn CORRTYPE = EXCHANGEABLE
/EMMEANS TABLES = Zeit
  compare = Zeit
  contrast=repeated
/EMMEANS TABLES = Geschlecht
  compare = Geschlecht
  contrast=pairwise.
```

Mittels der beiden EMMEANS-Befehle werden Einzelvergleiche durchgeführt und ein Gesamttest für den Faktor ausgegeben. Für den Messwiederholungsfaktor empfiehlt sich häufig die Option von „repeated“-Kontrasten (siehe Kapitel 9), für den Gruppierungsfaktor wäre in diesem Fall der Befehl entbehrlich, da er nur 2 Gruppen hat. Nachfolgend zunächst der wesentliche Teil der Standardausgabe, danach die jeweilige Ausgabe der beiden EMMEANS-Befehle, Mittelwertvergleiche und Gesamttest.

Parameter	Parameterschätzer						
	RegressionskoeffizientB	Standard Fehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	-,757	1,0461	-2,808	1,293	,524	1	,469
[ Geschlecht=2]	,584	,9711	-1,320	2,487	,361	1	,548
[ Geschlecht=1]	0 <sup>a</sup>	.	.	.	.	.	.
[Zeit=3]	2,631	1,0191	,634	4,628	6,665	1	,010
[Zeit=2]	1,282	,9135	-,509	3,072	1,969	1	,161
[Zeit=1]	0 <sup>a</sup>	.	.	.	.	.	.
(Skala)	1						

/EMMEANS TABLES = Zeit: Hier ist zu beachten, dass der globale Test für Zeit signifikant ist, während dies aus den beiden folgenden Einzelvergleichen nicht ersichtlich ist.

Individuelle Testergebnisse					
Zeit Wiederholter Kontrast	Kontrastschätzer	Standard Fehler	Wald-Chi-Quadrat	df	Sig.
Niveau 3 vs. Niveau 2	,20	,121	2,816	1	,093
Niveau 2 vs. Niveau 1	,31	,208	2,193	1	,139

Gesamttestergebnisse		
Wald-Chi-Quadrat	df	Sig.
11,526	2	,003

/EMMEANS TABLES = Geschlecht: Hier ist zu anzumerken, dass beide Vergleiche (natürlich) identische Ergebnisse liefern und das Gesamtergebnis mit dem aus der ersten Tabelle übereinstimmt.

		Paarweise Vergleiche				95% Wald-Konfidenzintervall für die Differenz	
(I)	(J)	Mittlere Differenz (I-J)	Standard Fehler	df	Sig.	Unterer Wert	Oberer Wert
2	1	,12	,221	1	,579	-,31	,56
1	2	-,12	,221	1	,579	-,56	,31

Gesamttestergebnisse		
Wald-Chi-Quadrat	df	Sig.
,307	1	,579

## 8. 4 ordinale abhängige Variablen und Messwiederholungen

Auch für den Fall ordinaler Kriteriumsvariablem gibt es Methoden der logistischen Regression mit Messwiederholungen, derzeit allerdings nur in R. Normalerweise wird das in Kapitel 8.2 kurz beschriebene *proportion odds model* angewandt.

Als Beispiel wird hier wieder der Datensatz 4 (`winer518`) verwendet, allerdings die abhängige Variable transformiert: (1,2)->1, (3,4)->2,..., 9->5.

### mit R:

R bietet hierzu u.a. die folgenden zwei Funktionen an:

- `repolr` (Paket `repolr`)
- `nomLORgee` (Paket `multgee`)

Bei diversen Tests hat sich `repolr` als die robustere und zuverlässigere Funktion erwiesen. Basis ist auch hier der umstrukturierte Datensatz, hier also `winer518t`. Die Funktion bietet zum einen die Möglichkeit an, die Struktur für die Korrelationen der Messwiederholungen festzulegen (vgl. Abschnitt 2.15): gleiche Korrelationen (`uniform`), Unabhängigkeit der Messwiederholungen (`independence`) oder autoregressive (`ar1`), falls ein Trend vermutet wird, wobei der default (`uniform`) der Normalfall sein wird. Zum anderen bietet die Funktion einen Test (`po.test`) zur Überprüfung der Gültigkeit des *proportion odds model*. Die Zeitpunkte (`times`) können angegeben werden, falls diese nicht äquidistant sind. Die Anzahl der Ausprägungen von `y` muss dagegen mit `categories` spezifiziert werden. `repolr` erlaubt auch eine Anova-Table mittels der Funktion `Anova` (Paket `car`).

Die Werte müssen 1,2,... sein, also größer 0. Nachfolgend Ein- und Ausgabe:

```
winer518t[,3]<-winer518t[,3]%/%2+1 # Transformation von y in 1,...,5
fit.r <- repolr(score~Geschlecht*Zeit, subjects="Vpn",
               data=winer518t, times=c(1,2,3), categories=5, po.test=T)
summary(fit.r)
Anova(fit.r)
```

Coefficients:					
		coeff	se.robust	z.robust	p.value
cuts1 2	2	-2.5910	0.6551	-3.9551	0.0001
cuts2 3	3	-1.0712	0.5263	-2.0353	0.0418
cuts3 4	4	1.0370	0.0543	19.0976	0.0000
cuts4 5	5	3.2392	0.6981	4.6400	0.0000
Geschlecht2		0.4965	0.4016	1.2363	0.2163
Zeit.L		1.8490	0.4503	4.1062	0.0000
Zeit.Q		2.0381	0.5275	3.8637	0.0001
Geschlecht2:Zeit.L		1.4234	0.5955	2.3903	0.0168
Geschlecht2:Zeit.Q		-4.2489	0.8672	-4.8996	0.0000

Correlation Structure: independence  
Fixed Correlation: 0  
PO Score Test: 8.121 (d.f. = 15 and p.value = 0.9188)

Analysis of Deviance Table (Type II tests)				
Response: score				
	Df	Chisq	Pr(>Chisq)	
cuts	4	37.585	1.365e-07	***
Geschlecht	1	341.387	< 2.2e-16	***
Zeit	2	2182.997	< 2.2e-16	***
Geschlecht:Zeit	2	32.801	7.538e-08	***

Die Koeffizienten `cuts1|2,...` sind die absoluten Glieder des Modells und spielen bei der varianzanalytischen Interpretation der Ergebnisse keine Rolle. Darunter folgen die Tests für die Kontraste der Gruppenvariablen, hier `Geschlecht`, sowie die linearen und quadratischen Kontraste des Messwiederholungsfaktors (`zeit.L` und `zeit.Q`). Darunter dann die Tests für die daraus resultierenden Interaktionen. Hieraus ist abzulesen (vgl. auch Abschnitt 6.10.4), dass die `Zeit` einen Einfluss hat, der für Männer und Frauen verschieden ausfällt. Häufig kann allerdings aus den Tests der Kontraste nicht unmittelbar ein Gesamttest für den Faktor abgelesen werden. Dann ist es erforderlich, wie in Kapitel 9.8 beschrieben aus den z-Werten der Kontraste, die zu einem Faktor bzw. zu einer Interaktion gehören, einen  $\chi^2$ -Test zu ermitteln. Für den Faktor `Zeit` (Zeilen `zeit.L` und `zeit.Q`) wäre das z.B.:

$$\chi^2 = 4.1062^2 + 3.8637^2 = 31.79$$

ein Wert, der bei 2 Freiheitsgraden auf dem 1%-Niveau signifikant ist.

Zuletzt wird der Test zur Überprüfung des *proportion odds model* ausgegeben, der mit  $p=0.92$  nicht signifikant ausfällt und somit die Anwendung der Methode legitimiert.

## 9. Mittelwertvergleiche, Kontraste und Kodierungen

In der Regel ist es erforderlich, im Anschluss an eine Varianzanalyse Mittelwertvergleiche durchzuführen. Denn signifikante Effekte besagen nur, dass zwischen irgendwelchen Gruppen Mittelwertunterschiede bestehen, geben aber keinen weiteren Aufschluss darüber, welche Gruppen oder Stufen dies nun sind. 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 schwächste Verfahren in dieser Kategorie sind die *paarweisen Vergleiche mit  $\alpha$ -Adjustierungen*.

Alpha-Adjustierungen und multiplen Vergleichen ist ein separates Skript gewidmet (vgl. Lüpssen, 2014). Dieses Skript beschränkt sich auf allgemeine Grundlagen zu Kontrasten, da diese zum Verständnis in den Kapiteln 7 und 8 erforderlich sind. Ausführliche Darstellungen sind auch im Internet zu finden, so z.B. bei Gonzalez (2009).

### 9. 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  $1 \cdot \mu_K - 0.5 \cdot (\mu_A + \mu_B)$ . Theoretisch können sogar bei der Zusammenfassung von Gruppen gewichtete Mittel gebildet werden, etwa  $(0.333 \cdot \mu_A + 0.667 \cdot \mu_B)$ , wenn etwa die B-Gruppe doppelt so stark berücksichtigt werden soll wie die A-Gruppe.

Hat ein Faktor  $I$  Gruppen (Schichten), so ist ein Kontrast  $C$  über  $I$  Koeffizienten  $c_i$  definiert:

$$C = c_1 \mu_1 + c_2 \mu_2 + \dots + c_I \mu_I$$

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

$$SS_C = \frac{(c_1 \bar{x}_1 + c_2 \bar{x}_2 + \dots + c_I \bar{x}_I)^2}{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \dots + \frac{c_I^2}{n_I}}$$

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

$$F = \frac{SS_C}{MS_{\text{Fehler}}}$$

wobei dieser F-Wert 1 Zähler-Fg hat und Nenner-Fg dem Test von  $SS_{\text{Effekt}}$  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 wie 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  $c_{11}+c_{12}+..+c_{1I}$  und  $C_2$  mit Koeffizienten  $c_{21}+c_{22}+..+c_{2I}$  heißen *orthogonal*, d.h. sind unabhängig voneinander, wenn die folgende Bedingung erfüllt ist:

$$\frac{c_{11}c_{21}}{n_1} + \frac{c_{12}c_{22}}{n_2} + \dots + \frac{c_{1I}c_{2I}}{n_I} = 0$$

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

Sind die zu untersuchenden Kontraste nicht orthogonal oder sollen mehr als  $(I-1)$  Kontraste geprüft werden, so sind die einzelnen Testergebnisse nicht mehr unabhängig voneinander. In solchen Fällen ist eine  $\alpha$ -Korrektur (siehe dazu Lüpsen, 2014) vorzunehmen. Speziell hierfür ist u.a. das Verfahren von *Dunn & Bonferroni* 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  $c_j$ :

Gruppe	Kontraste	
	C <sub>1</sub>	C <sub>2</sub>
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  $c_j$ :

Gruppe	Kontraste	
	C <sub>1</sub>	C <sub>2</sub>
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  $m$  Ausprägungen in mehrere dichotome  $d_1, \dots, d_m$  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  $m$  Variablen zwangsläufig eine redundant ist - jede beliebige von diesen lässt sich aus den übrigen errechnen, z.B.  $d_m = 1 - d_1 - d_2 - \dots - d_{m-1}$ , 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.

## 9.2 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  $c_1 + c_2 + \dots + c_I = 0$  eingehalten werden, bei Kodierungen nicht.

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

Statistisch werden alle Gruppen gegen eine vorgegebene, üblicherweise die erste oder letzte, paarweise 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:

Gruppe	Kontraste R				Kontraste SPSS			
	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  $(I-1)$  Vergleiche erlaubt sind, muss der Test für eine Gruppe entfallen. Dies ist üblicherweise (in R und SPSS) die letzte Gruppe. Die Koeffizienten:

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

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

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

Gruppe	Kontraste R und SPSS			
	1	2	...	(I-1)
1	- 1	$- 1/2$		$- 1/(I-1)$
2	1	$- 1/2$		$- 1/(I-1)$
...	0	1		
I-1	0	0		$- 1/(I-1)$
I	0	0		1

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

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

Gruppe	Kontraste SPSS			
	1	2	...	(I-1)
1	1	0		0
2	$- 1/(I-1)$	1		0
...	$- 1/(I-1)$	$- 1/(I-2)$		
I-1	$- 1/(I-1)$	$- 1/(I-2)$		1
I	$- 1/(I-1)$	$- 1/(I-2)$		- 1

### Wiederholt bzw. Repeated (SPSS)

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

Gruppe	Kontraste SPSS			
	1	2	...	(I-1)
1	1	0		0
2	- 1	1		0
...	0	- 1		
I-1	0	0		1
I	0	0		- 1

### Polynomial

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

## 9.3 Auswahl der Kontraste

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

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

wobei  $I$  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(I, base=I) für „normale“ Faktoren
contr.poly(I) 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"))
```

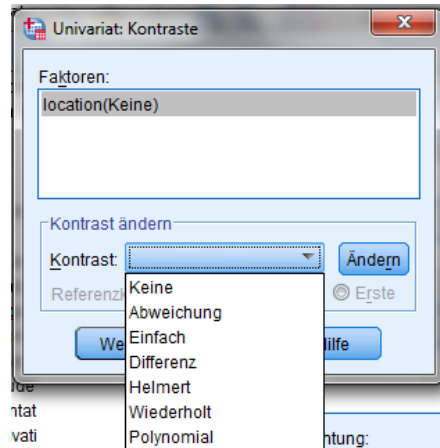
geändert werden und über `getOption("contrasts")` abgefragt werden. Hierbei wird `contr.name1` für „normale“ Faktoren und `contr.name2` für für „ordered factors“ übernommen. (Vgl. auch Anmerkungen zur Funktion `aov` in Kapitel 3.1.)

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 Auswahl führt:



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

## 9. 4 nichtparametrische Kontraste für die RT-, ART- und Puri & Sen-Verfahren

Einige der im Kapitel 2 vorgestellten nichtparametrischen Varianzanalysen lassen sich ja auf die parametrischen Standardverfahren zurückführen, so insbesondere die RT-, die ART-, die INT- sowie die Puri & Sen-Tests. 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 *drugs* demonstriert werden. Zunächst einmal wird angenommen, dass die erste Gruppe eine Vergleichsgruppe ist, gegen die die anderen drei Gruppen getestet werden sollen.

### mit R:

Die Tabelle 4.6 in Kapitel 4.3.4 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 *lm*, alternativ *gls* aus dem Paket *nlme*, als Varianzanalysefunktion zu verwenden, die zwar keine Anova-Tabelle ausgeben, dafür aber die Kontraste:

```
library(nlme)
contrasts(mydata2$drugs) <- contr.treatment(4, base=1)
aovc <- lm(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 *drugs1*:

	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 9-1

**mit SPSS:**

Die Tabelle 4.8 in Kapitel 4.3.4 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 „simple“-Kontrasten geprüft werden. Dazu ist bei den Anweisungen für die oben erwähnte Analyse die Zeile

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

einzufragen, wobei das „(1)“ die Nummer der Vergleichsgruppe angibt, also hier die erste:

```
Unianova x by patients drugs
/Contrast(drugs)=Simple(1)
/save = zresid
/print = homogeneity
/design = patients drugs patients*drugs.
```

Die Ausgabe dazu sollte selbsterklärend sein:

Kontrastergebnisse (K-Matrix)			
Einfacher Kontrast <sup>a</sup>	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 9-2

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

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

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

wobei

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

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

### mit R:

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

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

mit der nachfolgenden Ausgabe, worin die Zeilen `drugs2,...,drugs4` die gewünschten 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 9-3

### mit SPSS:

Die Berechnung soll nur für den ersten Vergleich (`drugs1 - drugs2`) gezeigt werden:

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

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

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

Gruppe	Kontraste		
	1	2	3
drugs1	1	0	1
drugs2	- 1	0	1
drugs3	0	1	- 1
drugs4	0	- 1	- 1

Tabelle 9-4

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

### mit R:

Auch hier dient natürlich wieder die Funktion `lm` zur Analyse der Kontraste. Lediglich die Spezifikation der Koeffizienten differiert erheblich. 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. Variante 1: Diese müssen zusätzlich als erste Spalte die Werte  $(1/I, \dots, 1/I)$  enthalten. Anschließend wird die Inverse der transponierten Matrix gebildet. Schließlich werden daraus die Spalten 2,...,k als Kontrastmatrix genommen. Variante 2: Aus der Matrix  $C$  der eigenen Kontraste wird die Kontrastmatrix errechnet:  $C \cdot (C \cdot C)^{-1}$ . Die zweite Variante wird nachfolgend verwendet, wobei `%*%` die Matrix-Multiplikation, `t(...)` die Transponierte und `solve(...)` die Inverse einer Matrix ist:

```

cmatrix <- cbind("A1-A2"=c(1, -1, 0, 0), "A3-A4"=c(0, 0, 1, -1),
  "A12-A34"=c(1, 1, -1, -1))
cont <- cmatrix%*(solve(t(cmatrix)%*cmatrix))
contrasts(mydata2$drugs) <- cont
aovc <- lm(rx~group*drugs, mydata2)
summary(aovc)

```

**mit SPSS:**

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.

**9. 5 universelles Verfahren für Kontraste**

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

Es sollen die Kontraste aus Tabelle 9-4 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`.

**mit R:**

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

```

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

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

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

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

```

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

### mit SPSS:

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

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

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

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

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

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

Aus diesem Beispiel geht das generelle Prozedere hervor. So lassen sich auch die im vorigen Abschnitt vorgenommenen Vergleiche der `drugs2`, ..., `drugs4` gegen `drugs1` durchführen.

## 9.6 Kontraste bei logistischen Regressionen

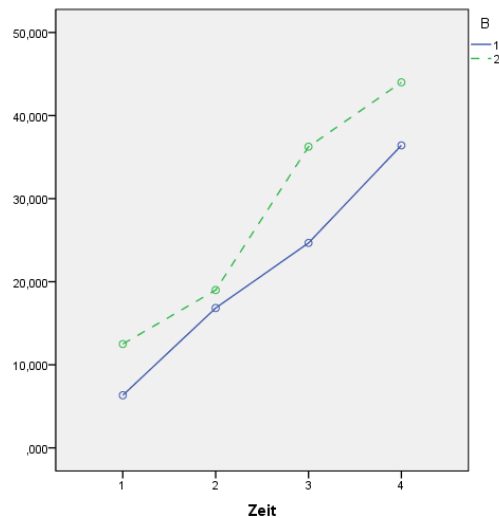
Bei der logistischen Regression gibt es für nominale Prädiktoren Standard-Kontraste. Wenn in R ein Prädiktor als „factor“ deklariert ist, wird für diesen automatisch die Kodierung gewählt, die in der `options(contrasts...)`-Anweisung festgelegt wurde (vgl. Kapitel 9.3). In SPSS kann bei der binär-logistischen Regression wie oben in 9.3 dargestellt die Kodierung gewählt werden. Speziellere Kontraste müssen wie oben in 9.5 skizziert über Umkodierungen analysiert werden. Beispiele sind in Kapitel 8 zu finden.

## 9.7 Kontraste für Messwiederholungen und Interaktionen

Aus dem eingangs (Kapitel 9.1) 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-, INT- und Puri & Sen-Verfahren Kontrastanalysen gleichermaßen für Versuchspläne mit Messwiederholungen durchführbar.

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

Dies soll am Datensatz 6 (`winer568`) demonstriert werden. Dieser umfasst die Gruppierungsfaktoren A und B sowie den Messwiederholungsfaktor `zeit`. Tabelle 6-7 in Kapitel 6.5.3 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

#### mit SPSS:

Hierzu werden zunächst analog den Berechnungen in Kapitel 6.3 die Daten umstrukturiert, so dass aus den Variablen  $v_1, \dots, v_4$  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  $RV.1, \dots, RV.4$  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 6-7 (Kapitel 6.5.3) zusammengefasst (dort allerdings in der Ausgabe von R). Nachfolgend nun die Ausgabe der Kontraste für den Faktor `Zeit`.

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.

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

**mit R:**

Ausgangsbasis ist der in Kapitel 6.5.3 erstellte Datensatz `winer568t`.

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

Die Kommandos lauten dann:

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

Zunächst vorab die oben erzeugte Kontrastmatrix `cont4`:

```
> cont4
      [,1] [,2] [,3]
[1,]    1    0    0
[2,]   -1    1    0
[3,]    0   -1    1
[4,]    0    0   -1
```

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



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

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

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

Denom. DF: 32			
	numDF	F-value	p-value
(Intercept)	1	415.9648513	<.0001
A	1	3.3160463	0.0780
B	1	8.1886042	0.0074
Zeit	3	235.4226927	<.0001
A:B	1	0.1732465	0.6800
A:Zeit	3	25.8348225	<.0001
B:Zeit	3	4.8246777	0.0070
A:B:Zeit	3	0.9709950	0.4185

Abschließend noch zur Illustration die Kontraste für den Interaktionseffekt, die sich als Kronecker-Produkt, in R über den Operator `%x%`, errechnen lassen:

```
> contrasts(win568t$A)
[,1]
1  1
2 -1
> contrasts(win568t$Zeit)
[,1] [,2] [,3]
1  1  0  0
2 -1  1  0
3  0 -1  1
4  0  0 -1
```

```

> contrasts(win568t$A) %x% contrasts(win568t$Zeit)
      [,1] [,2] [,3]
[1,]    1    0    0
[2,]   -1    1    0
[3,]    0   -1    1
[4,]    0    0   -1
[5,]   -1    0    0
[6,]    1   -1    0
[7,]    0    1   -1
[8,]    0    0    1

```

## 9. 8 Zusammenfassen von Kontrasten

In den vorangegangenen Abschnitten dienten die Kontraste primär dazu, den Effekt eines (signifikanten) Faktors zu erklären. Kontraste können aber auch die umgekehrte Funktion haben: aus mehreren Kontrasten eines Faktors einen Test für diesen zu ermitteln. Anzumerken ist vielleicht, dass dies auch die implizite Vorgehensweise bei linearen Modellen, und damit auch bei der Varianzanalyse, ist, wovon der normale Anwender allerdings nichts merkt. Denn zum einen muss er keine Kontraste vorgeben und zum anderen werden daraus automatisch für alle Effekte Tests ausgegeben werden. Wann aber ist es erforderlich, aus Kontrasten den Test für einen Faktor abzuleiten? Zahlreiche Funktionen für die Methoden zur Durchführung einer logistischen Regression mit und ohne Messwiederholungen, das sind insbesondere die in 2.15 erwähnten GEE (*Generalized Estimating Equations*) sowie die GLMM (*Generalized Linear Mixed Models*), geben lediglich Tests für die Kontraste bzw. für die Modell-Parameter aus, nicht jedoch einen „Gesamttest“ (*anova-like test*) für einen Faktor oder eine Interaktion. Nachfolgend wird kurz skizziert, wie aus den Tests der Kontraste für einen Faktor näherungsweise ein Gesamttest ermittelt werden kann.

Eine Voraussetzung dafür: die Kontraste müssen orthogonal sein. Dies sind z.B. die, die man in R mittels `contr.sum` bzw. in SPSS über `deviation` (vgl. Kapitel 9.2) erhält. Die Funktionen geben für jeden Kontrast immer eine Testgröße aus, nämlich den Quotienten aus Parameterschätzung und Schätzfehler. Dieser ist normalerweise ein z-Wert, der für größere  $n$  immer normalverteilt ist, gelegentlich auch einen t-Wert, der allerdings wie ein z-Wert behandelt werden kann. Die folgende Vorgehensweise setzt Unabhängigkeit der Parameterschätzungen voraus und ist eher ein Notbehelf:

- Durch Quadrieren jedes z-Wertes erhält man jeweils einen  $\chi^2$ -Wert, was der Prüfstatistik des *Wald-Tests* entspricht - verschiedentlich wird auch direkt dieser Test ausgegeben.
- Aufsummieren der  $\chi^2$ -Werte aller Kontraste, die zu einem Effekt gehören, was wiederum einen  $\chi^2$ -Wert ergibt.
- Testen dieser Summe auf Signifikanz anhand der  $\chi^2$ -Verteilung, wobei die Anzahl der Freiheitsgrade der Anzahl Summanden entspricht.

Beispiele dazu sind in den Kapiteln 8.2 und 8.4 zu finden.

Es gibt aber auch „klassische“ Verfahren hierfür, wovon der Wald-Test der bekannteste sein dürfte, in der einfachsten Form:

$$\hat{\beta}' V_{\beta}^{-1} \hat{\beta}$$

wobei  $\hat{\beta}$  die Parameterschätzungen und  $V_{\beta}$  die dazugehörige Kovarianzmatrix sind. Diese Statistik ist  $\chi^2$ -verteilt und hat so viele Freiheitsgrade wie die entsprechende F-Statistik der Varianzanalyse Zählerfreiheitsgrade hat, also z.B.  $k-1$  für den Test eines Haupteffekts. Wenn die  $\beta$  unabhängig sind, also  $V_{\beta}$  die Einheitsmatrix, ist diese Statistik mit der oben beschriebenen identisch. Diese wird auch mit Wald-Test vom Typ III bezeichnet, zur Unterscheidung von dem Wald-Test vom Typ II, der die Haupteffekte stärker bewertet und den Interaktionseffekt schwächer. Letzterer wird z.B. für GLMM-Verfahren empfohlen vom Fox & Weisberg (2011, Kapitel 4.4.4). Der Wald-Test kann auch in einen F-Test transformiert werden, was insbesondere für kleinere Stichproben vorteilhaft ist. Alle 3 sind in der Funktion `Anova` im R-Paket `car` verfügbar, die allerdings nur auf wenige Ergebnis-Objekte anwendbar ist. Alternativ werden die Funktionen `gee . anova` und `gee . robanova` für den Wald-Test bzw. eine robuste Variante des Wald-Tests nach Fan & Zhang (2014) angeboten (vgl. Anhang 3), insbesondere für die Anwendung auf GEE- und GLMM-Ergebnisse. Beide Funktionen erwarten als Argumente die Koeffizienten, die Kovarianzmatrix, die Freiheitsgrade sowie die Fallzahl  $n$  (nur `gee . anova`)

An dieser Stelle ist auch die Funktion `anova` zu erwähnen, die in R häufig in Zusammenhang mit der logistischen Regression angeführt wird. Deren Gebrauch ist allerdings problematisch, da das Ergebnis von der Reihenfolge der Faktoren abhängig ist. Ein weiteres Verfahren, um aus mehreren Kontrasten einen varianzanalytischen Test zu erhalten, ist der Likelihood-Ratio-Test (LR), der ebenfalls in der o.a. Funktion `Anova` enthalten ist.

## 10. Simple effects - einfache Effekte

In Kapitel 4.3.1.4 war darauf hingewiesen worden, dass bei mehrfaktoriellen Varianzanalysen (globale) Haupteffekte von Faktoren nicht interpretiert werden dürfen, wenn diese in signifikanten Interaktionen enthalten sind. Ist z.B. bei Faktoren A, B und C die Interaktion AC signifikant, so können die Haupteffekte der Faktoren A und C nicht interpretiert werden, da sowohl Faktor A sich für die einzelnen Stufen von C unterschiedlich verhält, wie auch Faktor C für die einzelnen Stufen von A. Statt dessen ist die Analyse dieser Faktoren über sog. *simple effects* (*einfache Effekte*) erforderlich. Dies sind 1-faktorielle Varianzanalysen eines Faktors, z.B. A, für jede Stufe des anderen Faktors, z.B. C. Im parametrischen Fall jedoch mit einem kleinen Unterschied: die Fehlerterme und Freiheitsgrade für die 1-faktoriellen F-Tests werden aus der globalen 2- oder 3-faktoriellen Analyse übernommen. Diese Analysen zeigen nun detailliert auf, in welchen Fällen der Faktor A oder C überhaupt einen Einfluss hat, oder aber für welche Stufen von C der Einfluss von A geringer ist bzw. für welche der Einfluss größer ist. Dazu kann sowohl Faktor A für jede der Stufen von Faktor C als auch Faktor C für jede der Stufen von Faktor A untersucht werden. Man kann sich dann aussuchen, welche Variante bessere Interpretationsmöglichkeiten bietet. Eine visuelle Hilfe bieten dabei auch die Interaktionsplots (vgl. Kapitel 4.3.1.2). In R und SPSS sind zum Teil Routinen zur Analyse der simple effects vorhanden.

### 10.1 Unabhängige Stichproben

Zunächst soll die exakte Analyse der simple effects erklärt werden, wie sie z.B. bei Winer (1991, pp 419-432) beschrieben ist, und zwar am Beispieldatensatz `mydata1`, bei dem ein signifikanter Interaktionseffekt von `patients` (A) und `drug` (B) besteht (vgl. Tabellen 4-1 für R bzw. 4-3 für SPSS). Soll z.B. der Faktor B für die 2 Stufen von Faktor A untersucht werden, dann werden 2 1-faktorielle Analysen von `drug` für die Gruppen `patients=1` und `patients=2` durchgeführt. Dabei erhält man 2 Streuungsquadratsummen:  $SS_{B(a=1)}$  und  $SS_{B(a=2)}$ , entsprechende Freiheitsgrade (jeweils 2) und Varianzen. Diese werden aber mittels F-Test nicht zu dem Fehlerterm der 1-faktoriellen Analyse in Bezug gesetzt, sondern zu dem der globalen 2-faktoriellen Analyse: 106.0 mit 12 FG. Dazu kurz die Analysen mit R:

Zunächst noch einmal die globale Analyse:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
patients	1	72	72.00	8.151	0.01449	*
drug	2	48	24.00	2.717	0.10634	
patients:drug	2	144	72.00	8.151	0.00581	**
Residuals	12	106	8.83			

Die beiden 1-faktoriellen Analysen (mit Kommando):

```
> summary(aov(x~drug, subset(mydata1, patients=="1")))
              Df Sum Sq Mean Sq F value Pr(>F)
drug           2     24      12      1.5  0.296
Residuals     6     48       8

> summary(aov(x~drug, subset(mydata1, patients=="2")))
              Df Sum Sq Mean Sq F value Pr(>F)
drug           2    168     84.00   8.69 0.0169 *
Residuals     6     58     9.67
```

Hieraus resultieren  $SS_{B(a=1)}=24$  und  $SS_{B(a=2)}=168$  sowie die Varianzen  $MS_{B(a=1)}=12$  und  $MS_{B(a=2)}=84$ . Da die globale Fehlervarianz  $106.0/12=8.33$  ist, erhält man für die simple effects die F- bzw. p-Werte  $F_{B(a=1)}=1.44$  ( $p=0.275$ ) und  $F_{B(a=2)}=10.08$  ( $p=0.003$ ).

In der Regel werden die F-Werte der simple effects eher im signifikanten Bereich liegen als die der normalen 1-faktoriellen Analyse, da wie in Kapitel 4.3.1.3 dargelegt durch die Einbeziehung weiterer Faktoren die Fehlervarianz reduziert wird. Verschiedentlich wird der Einwand geäußert, dass wegen der mehrfachen Tests eine  $\alpha$ -Adjustierung vorgenommen werden müsste. Winer (1991) erwähnt zwar diese Option, hält sie aber nicht für erforderlich. Nachfolgend werden die Verfahren in R und SPSS vorgestellt.

### mit R:

Zunächst wird die 2-faktorielle Varianzanalyse durchgeführt, anschließend über `testInteractions` aus dem Paket `phia` zunächst der Faktor `patients` für die 3 Stufen von Faktor `drug` analysiert, danach der Faktor `drug` für die beiden Stufen von `patients`, wobei standardmäßig eine  $\alpha$ -Adjustierung vorgenommen wird, und zwar die Methode von Holm. Soll dies vermieden werden, ist explizit "none" anzugeben:

```
library(phia)
ano1<-aov(x~patients*drug,mydata1)
testInteractions(ano1,fixed="drug",across="patients",adjustment="none")
testInteractions(ano1,fixed="patients",across="drug",adjustment="none")
```

Nachfolgend nur die Ausgabe von `testInteractions`:

P-value adjustment method: none						
	Value	Df	Sum of Sq	F	Pr(>F)	
1	-8	1	96	10.868	0.00638	**
2	4	1	24	2.717	0.12520	
3	-8	1	96	10.868	0.00638	**
Residuals		12	106			

P-value adjustment method: none						
	drug1	drug2	Df	Sum of Sq	F	Pr(>F)
1	-2	2	2	24	1.3585	0.293883
2	-2	-10	2	168	9.5094	0.003352 **
Residuals			12	106		

Hieraus ist ersichtlich, dass zum einen ein Geschlechtsunterschied nur bei drug 1 und 3 besteht und zum anderen der Faktor drug nur bei Frauen einen Einfluss hat.

### mit SPSS:

Die erforderlichen Kommandos sind im Wesentlichen die in Kapitel 4.3.2 angegebenen, jedoch ergänzt um die `EMMEANS`-Kommandos, bei denen zunächst der zu analysierende Interaktionseffekt anzugeben ist und bei `COMPARE` der zu analysierende Faktor. Nachfolgend zunächst A für die B-Stufen sowie B für die A-Stufen:

```
UNIANOVA x by patients drug
  /EMMEANS=TABLES(patients*drug) COMPARE (patients) ADJ(LSD)
  /EMMEANS=TABLES(patients*drug) COMPARE (drug) ADJ(LSD)
  /DESIGN = patients drug patients*drug.
```

Nach der 2-faktoriellen Varianzanalyse wird zunächst der Faktor `patients` für die 3 Stufen von Faktor `drug` analysiert,

Tests auf Univariate

		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Kontrast	96,000	1	96,000	10,868	,006
	Fehler	106,000	12	8,833		
2	Kontrast	24,000	1	24,000	2,717	,125
	Fehler	106,000	12	8,833		
3	Kontrast	96,000	1	96,000	10,868	,006
	Fehler	106,000	12	8,833		

anschließend der Faktor `drug` für die beiden Stufen von `patients`:

Tests auf Univariate

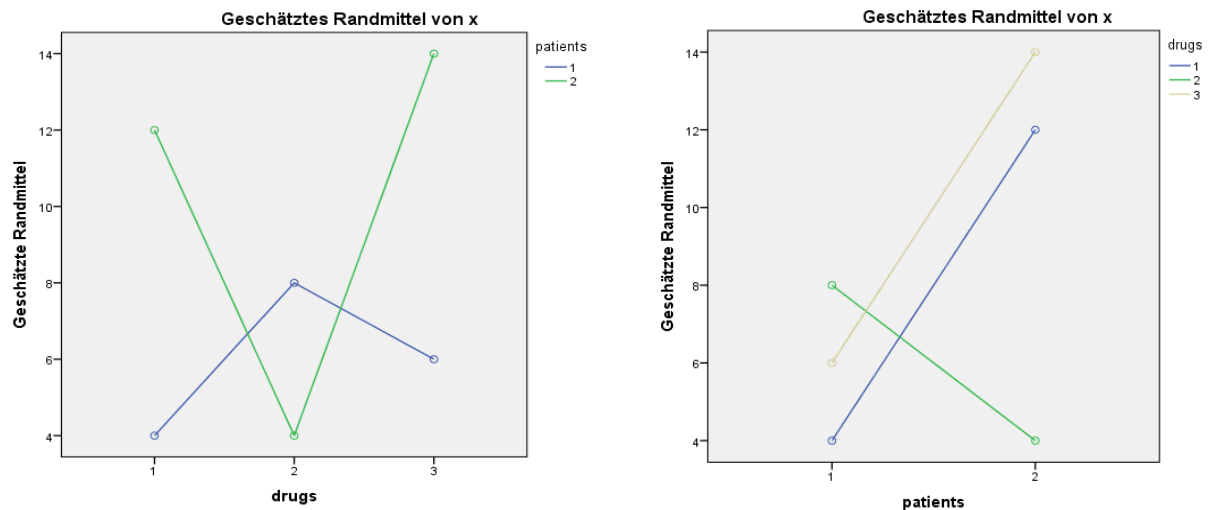
		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Kontrast	24,000	2	12,000	1,358	,294
	Fehler	106,000	12	8,833		
2	Kontrast	168,000	2	84,000	9,509	,003
	Fehler	106,000	12	8,833		

hieran schließen sich, ausgelöst durch den Befehl `POSTHOC=drug(TUKEY)`, paarweise Mittelwertvergleiche nach dem Verfahren von Tukey für den Faktor `drug` an, da dieser mehr als 2 Stufen hat:

Paarweise Vergleiche

	(I)	(J)	Mittlere Differenz (I-J)	Standardfehler	Sig. <sup>b</sup>	95% Konfidenzintervall für die Differenz <sup>b</sup>	
						Untergrenze	Obergrenze
1	1	2	-4,000	2,427	,125	-9,287	1,287
		3	-2,000	2,427	,426	-7,287	3,287
	2	1	4,000	2,427	,125	-1,287	9,287
		3	2,000	2,427	,426	-3,287	7,287
	3	1	2,000	2,427	,426	-3,287	7,287
		2	-2,000	2,427	,426	-7,287	3,287
2	1	2	8,000*	2,427	,006	2,713	13,287
		3	-2,000	2,427	,426	-7,287	3,287
	2	1	-8,000*	2,427	,006	-13,287	-2,713
		3	-10,000*	2,427	,001	-15,287	-4,713
	3	1	2,000	2,427	,426	-3,287	7,287
		2	10,000*	2,427	,001	4,713	15,287

Als Ergebnis zeigt sich, dass die beiden Patientengruppen sich nur bei den Präparaten 1 und 3 unterscheiden, nicht jedoch bei Präparat 2. Umgekehrt unterscheiden sich die 3 Präparate nur in der 2. Patientengruppe. Wie oben erwähnt helfen hier die Interaktionsplot bei der Interpretation der Ergebnisse:



*Interaktionsplots zur Visualisierung der simple effects in beiden Ansichten*

Diese simple effects-Analyse lässt sich natürlich problemlos auf die nichtparametrischen Methoden RT, INT und ART bzw. ART+INT übertragen. Für die anderen Methoden wie Puri & Sen, van der Waerden oder ATS gibt es nur die Möglichkeit, die „normalen“ 1-faktoriellen Analysen durchzuführen.

## 10.2 Gemischte Versuchspläne

Bei gemischten Versuchsplänen, also solchen mit mindestens einem Gruppierungsfaktor und einem Messwiederholungsfaktor ist das Prinzip dasselbe wie oben erläutert: Zur Analyse eines Faktors werden 1-faktorielle Analysen für jede Stufe eines anderen Faktors gerechnet, und die resultierenden Varianzen, z.B.  $MS_{B(A=1)}$ ,  $MS_{B(A=2)}$ ,... werden für den F-Test zu der Fehlervarianz der globalen Varianzanalyse in Bezug gesetzt, die auch für den entsprechenden Test des globalen Haupteffekts verwendet wird. Wird z.B. ein Gruppierungsfaktor für die einzelnen Messwiederholungen analysiert, so ist dies die Streuung zwischen den Versuchspersonen, bei R (vgl. Tabelle 6-1) die Zeile `Residuals` im ersten Block (`Error: vpn`) bzw. bei SPSS (vgl. Tabelle 6-3) die Zeile `Fehler` im Block `Zwischensubjekteffekte`. Wird ein Messwiederholungsfaktor für die Gruppen eines Gruppierungsfaktors analysiert, so ist dies die Streuung innerhalb der Versuchspersonen, bei R (vgl. Tabelle 6-1) die Zeile `Residuals` im zweiten Block (`Error: vpn:...`) bzw. bei SPSS (vgl. Tabelle 6-3) die Zeile `Fehler(...)` im Block `Innersubjekteffekte`. Das Verfahren ist bei Winer (1991, pp 526-531) beschrieben.

Ein Beispiel soll mit dem Datensatz `winer518` gerechnet werden, der ebenfalls eine signifikante Interaktion aufzeigt (vgl. Tabellen 6-1 und 6-3).

### mit R:

Die o.a. Funktion `testInteractions` für R kann leider keine gemischten Versuchspläne verarbeiten. Allerdings wird eine Funktion `simple.effects` vom Autor angeboten (vgl. Anhang 3.15), die sowohl bei Versuchsplänen mit mehreren Gruppierungsfaktoren wie auch mit maximal einem Messwiederholungsfaktor die Analyse von simple effects durchführt. Die erforderlichen Anweisungen (vgl. auch Tabelle 6-1):

```
aov1 <- aov(score~Geschlecht*Zeit+Error(Vpn/Zeit),winer518t)
simple.effects(aov1,"Geschlecht*Zeit",winer518t)
```

Hierbei sind das Ergebnis der Varianzanalyse (`aov1`), der verwendete Dataframe (`winer518t`) sowie die zu analysierende Interaktion anzugeben. Sollen in einem mehrfaktoriellen Versuchsplan mehrere Interaktionen aufgeschlüsselt werden, so sind diese über `c(...)` zusammenzufassen. Optional kann eine  $\alpha$ -Adustierung über `adjust=..` (vgl. R-Funktion `p.adjust`) angefordert werden. Die Ausgabe:

```

Response: score
              Df Sum Sq Mean Sq F value Pr(>F)
Geschlecht (Zeit=T1)  1  8.100   8.100   1.1462 0.3156
Geschlecht (Zeit=T2)  1 40.000  40.000   5.6604 0.0446 *
Geschlecht (Zeit=T3)  1  0.100   0.100   0.0142 0.9082
Error (Geschlecht)    8 56.533   7.067
Zeit (Geschlecht=1)   2 41.200  20.600  15.6456 0.0002 ***
Zeit (Geschlecht=2)   2 61.733  30.867  23.4430 <2e-16 ***
Error (Zeit)         16 21.067   1.317

```

Das Ergebnis zeigt, dass zum einen ein Unterschied zwischen Männern und Frauen nur zum Zeitpunkt 2 besteht und zum anderen die Ergebnisse sich zu den 3 Zeitpunkten unterscheiden, sowohl für Männer als auch für Frauen (vgl. auch Grafiken weiter unten).

### mit SPSS:

Die erforderlichen Kommandos sind im Wesentlichen die in Kapitel 6.2 angegebenen, jedoch ergänzt um die `EMMEANS`-Kommandos, bei denen zunächst der zu analysierende Interaktionseffekt anzugeben ist und bei `COMPARE` der zu analysierende Faktor. Nachfolgend zunächst `Geschlecht` für die `Zeit`-Stufen sowie `Zeit` für die `Geschlecht`-Stufen:

```

GLM t1 t2 t3 by Geschlecht
/wsfactor=Zeit 3 polynomial
/wsdesign=Zeit
/design=Geschlecht
/EMMEANS=TABLES(Geschlecht*Zeit) COMPARE (Geschlecht) ADJ(LSD)
/EMMEANS=TABLES(Geschlecht*Zeit) COMPARE (Zeit) ADJ(LSD).

```

Die Ausgabe umfasst nach der globalen Varianzanalyse zunächst die Tests des Faktors `Geschlecht` für die 3 Zeitstufen, sowie eine Tabelle der Mittelwertvergleiche,

**Tests auf Univariate**

Zeit		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Kontrast	8,100	1	8,100	2,613	,145
	Fehler	24,800	8	3,100		
2	Kontrast	40,000	1	40,000	14,286	,005
	Fehler	22,400	8	2,800		
3	Kontrast	,100	1	,100	,026	,875
	Fehler	30,400	8	3,800		



**Paarweise Vergleiche**

Zeit	(I)	(J)	Mittlere Differenz (I-J)	Standardfehler	Sig. <sup>b</sup>	95% Konfidenzintervall für die Differenz <sup>b</sup>	
						Untergrenze	Obergrenze
1	1	2	-1,800	1,114	,145	-4,368	,768
	2	1	1,800	1,114	,145	-,768	4,368
2	1	2	4,000*	1,058	,005	1,560	6,440
	2	1	-4,000*	1,058	,005	-6,440	-1,560
3	1	2	-,200	1,233	,875	-3,043	2,643
	2	1	,200	1,233	,875	-2,643	3,043

danach die Tests des Faktors *Zeit* für die beiden Gruppen, ebenfalls gefolgt von einer Tabelle der Mittelwertvergleiche. Allerdings werden für den Messwiederholungsfaktor die multivariaten Tests (vgl. Kapitel 5.3.9) anstatt der „normalen“ F-Tests ausgegeben:

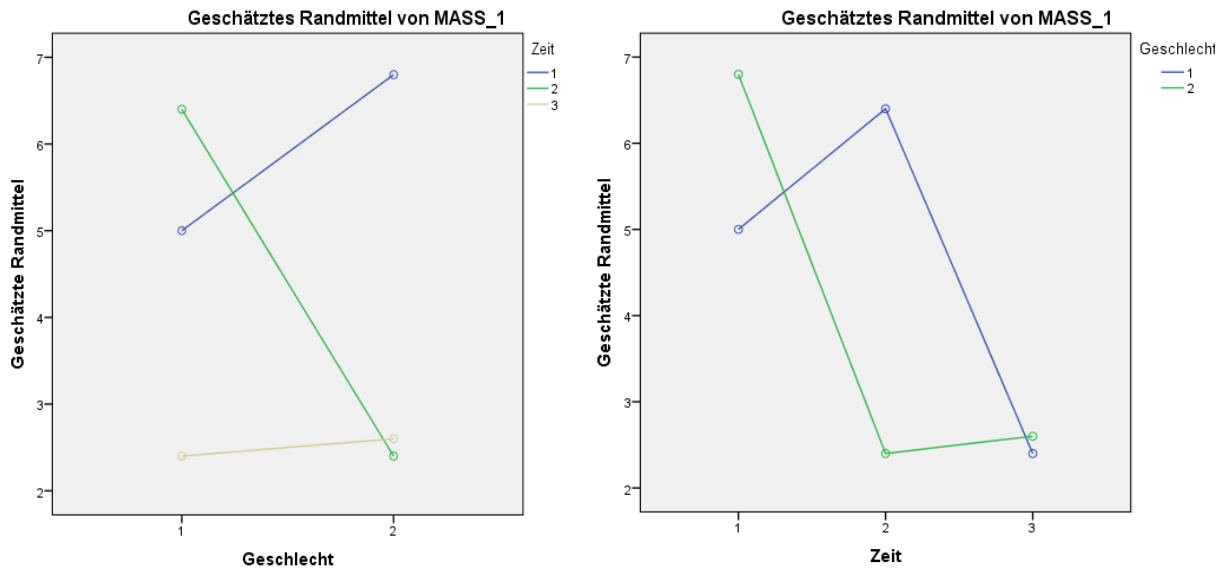
**Multivariate Tests**

		Wert	F	Hypothese df	Fehler df	Sig.
1	Pillai-Spur	,795	13,584 <sup>a</sup>	2,000	7,000	,004
	Wilks-Lambda	,205	13,584 <sup>a</sup>	2,000	7,000	,004
	Hotelling-Spur	3,881	13,584 <sup>a</sup>	2,000	7,000	,004
	Größte charakteristische Wurzel nach Roy	3,881	13,584 <sup>a</sup>	2,000	7,000	,004
2	Pillai-Spur	,864	22,230 <sup>a</sup>	2,000	7,000	,001
	Wilks-Lambda	,136	22,230 <sup>a</sup>	2,000	7,000	,001
	Hotelling-Spur	6,351	22,230 <sup>a</sup>	2,000	7,000	,001
	Größte charakteristische Wurzel nach Roy	6,351	22,230 <sup>a</sup>	2,000	7,000	,001

**Paarweise Vergleiche**

	(I)Zeit	(J)Zeit	Mittlere Differenz (I-J)	Standardfehler	Sig. <sup>b</sup>	95% Konfidenzintervall für die Differenz <sup>b</sup>	
						Untergrenze	Obergrenze
1	1	2	-1,400	,640	,060	-2,877	,077
		3	2,600*	,806	,012	,741	4,459
	2	1	1,400	,640	,060	-,077	2,877
		3	4,000*	,721	,001	2,337	5,663
	3	1	-2,600*	,806	,012	-4,459	-,741
		2	-4,000*	,721	,001	-5,663	-2,337
2	1	2	4,400*	,640	,000	2,923	5,877
		3	4,200*	,806	,001	2,341	6,059
	2	1	-4,400*	,640	,000	-5,877	-2,923
		3	-,200	,721	,789	-1,863	1,463
	3	1	-4,200*	,806	,001	-6,059	-2,341
		2	,200	,721	,789	-1,463	1,863

Das Ergebnis zeigt, dass zum einen ein Unterschied zwischen Männern und Frauen nur zum Zeitpunkt 2 besteht (links grüne Linie) und zum anderen die Ergebnisse sich zu den 3 Zeitpunkten unterscheiden, sowohl für Männer als auch für Frauen, und zwar wie die Mittelwertvergleiche zeigen, bei den Männern (rechts blaue Linien) Zeitpunkte 1 und 2 von Zeitpunkt 3 und bei den Frauen (rechts grüne Linien) Zeitpunkt 1 von den Zeitpunkten 2 und 3.



*Interaktionsplots zur Visualisierung der simple effects in beiden Ansichten*

Diese simple effects-Analyse lässt sich natürlich problemlos auf die nichtparametrischen Methoden RT, INT und ART bzw. ART+INT übertragen. Für die anderen Methoden wie Puri & Sen, van der Waerden oder ATS gibt es nur die Möglichkeit, die „normalen“ 1-faktoriellen Analysen durchzuführen.

## 11. Beispiele mit problematischen Datensätzen

Während in den vorangegangenen Kapiteln lediglich kleinere, überschaubare Datensätze behandelt wurden, bei denen in der Regel die passende Methode quasi auf der Hand lag, geht es in diesem Kapitel um die Bearbeitung von größeren Datensätzen, die manchmal einigen Aufwand erfordert. Da hier Varianzheterogenitäten im Vordergrund stehen, auf der anderen Seite SPSS dazu keine Lösungen bietet, beschränken sich die Beispiele auf R.

### 11. 1 Extrem heterogene Varianzen

Der Datensatz `ind.waste` (industrial waste) umfasst 2 Faktoren: Temperatur (low, medium, high) und die Komplexität der Produktionsbedingung (environment) mit Stufen 1,...,5. Die abhängige Variable ist die Abfallmenge (Waste). Für jede der insgesamt 15 Bedingungen liegen nur 2 Messungen vor. Dies kann zu beträchtlichen Schwankungen der Streuung zwischen den jeweils 2 Messungen führen.

		Environment					
		Temperatur	1	2	3	4	5
Mittelwerte	low		6.495	8.545	9.540	6.58	8.165
	medium		6.415	6.685	7.095	8.58	9.345
	high		7.755	9.585	9.085	11.56	11.415
Standardabweichungen	low		0.841	0.856	0.438	1.626	1.789
	medium		0.841	0.714	1.082	0.127	0.389
	high		0.035	1.138	0.262	0.863	2.341

Wie problematisch der Datensatz ist, zeigt das Verhältnis der Standardabweichungen  $\max(s_i)/\min(s_i)$ :  $2.341/0.035 = 66.89$ , d.h. ein Verhältnis von etwa 4474 für die Varianzen.

Prüft man die Varianzhomogenität mit dem Levene-Test, erhält man mit:

```
leveneTest(Waste~Temperatur*Environment, ind.waste)
leveneTest(Waste~Temperatur, ind.waste)
leveneTest(Waste~Environment, ind.waste)
```

```
Levene's Test for Homogeneity of Variance (center = median)
  Df    F value    Pr(>F)
group 14 4.0688e+29 < 2.2e-16 ***

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 2  0.2571 0.7752

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 4  0.701 0.5986
```

d.h. die Tests der beiden Haupteffekte sind von der Varianzheterogenität nicht betroffen, wohl aber der Test der Interaktion. Daher wäre die „normale“ Varianzanalyse für den Test von Temperatur und Environment einsetzbar:

```
summary(aov(Waste~Temperatur*Environment, ind.waste))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Temperatur	2	30.69	15.346	13.063	0.000519	***
Environment	4	24.68	6.171	5.253	0.007546	**
Temperatur:Environment	8	22.91	2.864	2.438	0.065134	.
Residuals	15	17.62	1.175			

die beide als signifikant ausgewiesen werden. Es bleibt der Test der Interaktion, dann mit einem Verfahren für inhomogene Varianzen, da die nichtparametrischen Methoden, wie etwa INT, die Varianzhomogenität nicht mildern können:

```
ns.waste <- qnorm(rank(ind.waste$Waste)/(dim(ind.waste)[1]+1))
leveneTest(ns.waste~Temperatur*Environment, ind.waste)
```

Levene's Test for Homogeneity of Variance (center = median)			
group	Df	F value	Pr(>F)
14		4.9823e+30	< 2.2e-16 ***

Für 2-faktorielle Versuchspläne stehen dazu in R zur Verfügung: ATS (vgl. Kapitel 2.9 und 4.3.8), Welch-James, Brown & Forsythe sowie BDM (vgl. 2.13 und 4.3.3). Wegen der extremen Bedingungen werden „vorsichtshalber“ alle Verfahren durchgerechnet:

```
ats.2(Waste~Temperatur*Environment, ind.waste)
```

	Df	F value	Pr(>F)
Temperatur	1.856124	12.447298	0.01159451
Environment	2.987679	6.498499	0.03479792
Temperatur*Environment	4.102722	3.006421	0.12764763

```
wj.anova(ind.waste, "Waste", "Temperatur", "Environment")
```

	Chi Sq	df	P(Chi>value)
Temperatur	25.65739	2	0.00450955
Environment	29.93999	4	0.01850815
Temperatur : Environment	26.44512	8	0.23550000

```
bf.f(Waste~Temperatur*Environment, ind.waste)
```

Response: Waste							
	Df	Df.err	Sum Sq	Mean Sq	F value	Pr(>F)	
Temperatur	2	25.5059	30.693	15.3464	6.3533	0.005768	**
Environment	4	18.0464	24.685	6.1713	2.1661	0.114138	
Temperatur:Environment	8	5.8907	22.912	2.8639	2.4378	0.149116	
Residuals	15		17.622	1.1748			

```
library(GFD)
```

```
GFD(Waste~Temperatur*Environment, ind.waste, nperm=1)
```

ANOVA-Type Statistic (ATS):				
	Test statistic	df1	df2	p-value
Temperatur	13.063178	1.852204	5.890664	0.00723086
Environment	5.253177	2.742812	5.890664	0.04337361
Temperatur:Environment	2.437850	4.373908	5.890664	0.15834953

Erfreulicherweise sagen die Ergebnisse weitgehend dasselbe aus: Der oben schon erkannte signifikante Einfluss der beiden Haupteffekte kann als gesichert angesehen werden, während die Interaktion, wie bei der parametrischen Varianzanalyse, nicht gesichert ist.

## 11.2 lognormal verteilte abhängige Variable

Hierbei handelt es sich um einen „synthetischen“ Datensatz, d.h. die Daten wurden mittels Zufallszahlen erzeugt. Er enthält eine abhängige Variable  $y$  mit Werten im Bereich  $[-2, 2]$  sowie zwei Gruppierungsfaktoren A (4 Stufen) und B (5 Stufen) mit einem  $n_i=10$  und  $N=200$ .

		B 1	B 2	B 3	B 4	B 5
Mittelwerte	A 1	0.89	0.90	0.99	1.13	1.07
	A 2	1.07	0.97	1.21	1.07	1.14
	A 3	0.91	0.82	0.87	1.06	1.09
	A 4	1.00	0.97	1.14	1.05	0.94
Standardabweichungen	A 1	0.44	0.53	0.24	0.36	0.19
	A 2	0.46	0.59	0.33	0.26	0.27
	A 3	0.42	0.49	0.28	0.22	0.29
	A 4	0.36	0.38	0.50	0.22	0.25

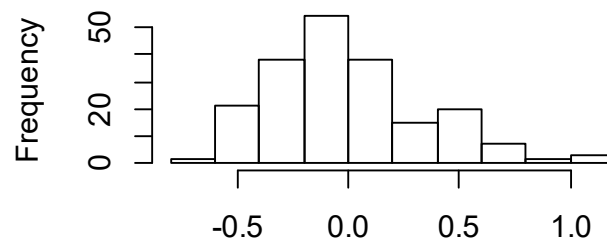
Die Tabelle lässt vermuten, dass die Standardabweichungen für Faktor B ungleich sind:

```
leveneTest(y~B, df)
```

Levene's Test for Homogeneity of Variance (center = median)			
	Df	F value	Pr(>F)
group	4	5.1568	0.0005705 ***

Als nächstes werden die Residuen untersucht:

```
aov.1 <- aov(y~A*B, df)
hist(aov.1$residuals)
```



Das Histogramm widerspricht nicht notwendigerweise einer Normalverteilung, zeigt aber dennoch eine leichte Rechtsschiefe, die wegen der hohen Fallzahl nicht außer Acht gelassen werden sollte. Das bestätigt auch der Shapiro-Test

```
shapiro.test(aov.1$residuals)
```

Test Statistic:	W = 0.9658659
P-value:	9.013545e-05

Die Schiefe, zu berechnen mittels `skewness` aus dem Paket `EnvStats`) beträgt 0.763, der Test auf Schiefe > 0 ( $S/(6/n)$ ) ergibt:

$$0.763/(6/200) = 24.21$$

Dieser Wert ist  $\chi^2$ -verteilt mit 1 FG und zeigt somit eine deutliche Rechtsschiefe an, was auf eine Lognormalverteilung hindeutet. Die erste Möglichkeit besteht darin, eine Methode für heterogene Varianzen anzuwenden. Bei Vorliegen einer Lognormalverteilung sollten allerdings keine rangbasierten Tests angewandt werden, womit u.a. ATS und BDM ausscheiden. Es bleiben noch die Verfahren von Welch & James sowie von Brown & Forsythe, beide in der eigenen Bibliothek (siehe Anhang) verfügbar:

```
wj.anova(df, 'y', 'A', 'B', Ftest=T)
```

	F	df1	df2	P(F>value)
A	1.1796886	3	72.19293	0.3235286
B	1.1882087	4	76.69976	0.3227698
A : B	0.6010449	12	66.84131	0.8337179

```
bf.f(y~A*B, df)
```

	Df	Df.err	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	194.42	0.5190	0.173002	1.2859	0.2805
B	4	157.97	0.8109	0.202723	1.5159	0.2001
A:B	12	128.19	0.7644	0.063704	0.4625	0.9330
Residuals	180		24.7946	0.137748		

Es gibt aber noch eine zweite Möglichkeit: die Daten durch eine log-Transformation in eine Normalverteilung zu wandeln, um dann „wie gewohnt“ weiter zu verfahren:

```
within(df, ly<-log(y)) -> df
leveneTest(ly~A*B, df)
aov.2 <- aov(ly~A*B, df)
shapiro.test(aov.2$residuals)
```

Levene's Test for Homogeneity of Variance (center = median)			
	Df	F value	Pr(>F)
group	19	2.3283	0.002153 **
Test Name: Shapiro-Wilk normality test			
Test Statistic:		W = 0.9931843	
P-value:		0.4847057	

Durch die log-Transformation ist zwar die Varianzheterogenität erhalten geblieben, aber die Daten können als normalverteilt angenommen werden. Jetzt werden noch einmal die Tests von Welch & James sowie von Brown & Forsythe angewandt, jedoch auf die Variable  $\log(y)$ :

```
wj.anova(df, 'ly', 'A', 'B')
bf.f(ly~A*B, df)
```

mit den folgenden Ergebnissen:

	Chi Sq	df	P(Chi>value)
A	3.268109	3	0.36850000
B	12.187069	4	0.02350765
A : B	8.801682	12	0.78250000

Response: ly

	Df	Df.err	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	192.60	0.5048	0.16827	1.1013	0.349845
B	4	139.06	2.2536	0.56341	3.8959	0.004966 **
A:B	12	116.88	0.8525	0.07104	0.4764	0.925047
Residuals	180		26.8426	0.14913		

Auch hier unterscheiden sich beide Ergebnisse nicht. Aber: diese Resultate widersprechen den zuerst gefundenen mit den nicht-transformierten y-Werten. Eine Entscheidung, welchem Ergebnis man nun vertrauen darf, geben Feng et al. (2014). Sie warnen davor, auf rechtsschiefe Daten die log-Transformation anzuwenden und geben ein Beispiel, allerdings mit einer Regressionsanalyse, in dem Effekte durch die Transformation kleinere Standardfehler bekommen und somit signifikant werden.

### 11.3 negative pairing

Der Datensatz `mydata12` besteht aus der abhängigen Variablen  $x$  mit ganzzahligen Werten zwischen 0 und 80 sowie zwei Faktoren A (4 Gruppen) und B (5 Gruppen) und umfasst 200 Fälle.

		B 1	B 2	B 3	B 4	B 5
Zellenbesetzungen	A 1	8	6	14	10	16
	A 2	16	14	4	14	10
	A 3	6	14	10	4	10
	A 4	12	10	4	6	12
Mittelwerte	A 1	32.00	36.17	36.57	34.40	41.06
	A 2	30.81	30.43	36.25	39.36	39.00
	A 3	30.67	35.50	38.60	27.25	36.70
	A 4	29.92	34.40	30.00	47.67	38.08
Standard-abweichungen	A 1	11.69	20.31	11.97	8.83	5.71
	A 2	3.71	9.29	17.73	10.29	12.55
	A 3	14.39	8.12	7.60	15.17	9.12
	A 4	8.72	10.52	16.87	23.04	7.89

Zunächst einmal wird die parametrische Varianzanalyse durchgeführt, um auch die Residuen auf Normalverteilung überprüfen zu können. Hier muss `drop1` verwendet werden, da der Datensatz ungleiche Zellenbesetzungszahlen aufweist.

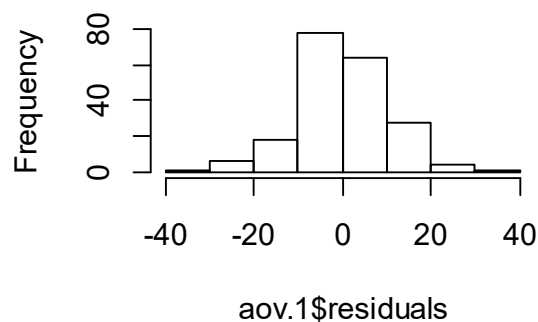
```
aov.1 <- aov(x~A*B,mydata12)
drop1(aov.1, ~. , test="F")
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			21017	970.95		
A	3	133.47	21151	966.22	0.3810	0.76679
B	4	1412.07	22429	975.96	3.0234	0.01915 *
A:B	12	1738.56	22756	962.85	1.2408	0.25835

Prüfen der Voraussetzungen:

```
hist(aov.1$residuals)
shapiro.test(aov.1$residuals)

leveneTest(x~A*B, mydata12)
leveneTest(x~A, mydata12)
leveneTest(x~B, mydata12)
```



```
Shapiro-Wilk normality test

data: aov.1$residuals
W = 0.99257, p-value = 0.406

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 19  2.5401 0.0007507 ***

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group  3  0.9544 0.4154

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group  4  3.1109 0.01645 *
```

Sowohl das Histogramm wie auch der Shapiro-Test erlauben die Annahme, dass  $y$  normalverteilt ist. Die Levene-Tests zeigen, dass eine sehr starke Varianzheterogenität vorhanden ist, insbesondere für den Test der Interaktion. Dies legt nahe, ein mögliches pairing zu überprüfen. Dazu wird die Korrelation zwischen  $n_i$  und  $s_i^2$  errechnet:

```
ni <- as.vector(with(mydata12, table(A, B)))
si <- as.vector(with(mydata12, tapply(x, list(A, B), sd)))
cor(ni, si^2)
```

Der Wert  $r = -0.75$  bestätigt ein negatives pairing, was die Durchführung der normalen Varianzanalyse nicht ratsam macht. Statt dessen sollte eine Methode benutzt werden, die robust gegen heterogene Varianzen ist. Dazu zählen die Tests von Welch & James sowie von Brown & For-



sythe, wobei letzterer als leicht liberal gilt. Das ATS-Verfahren ist zwar auch bei negativem pairing anwendbar, gilt allerdings als extrem konservativ. Dazu alle drei Verfahren im Vergleich, sie alle in der eigenen Bibliothek (siehe Anhang) verfügbar sind:

```
wj.anova(mydata12, 'x', 'A', 'B')
```

	Chi Sq	df	P(Chi>value)
A	0.7351891	3	0.87550000
B	11.7029784	4	0.05150485
A : B	10.2629290	12	0.74150000

```
bf.f(x~A*B,mydata12)
```

	Df	Df.err	Sum Sq	Mean Sq	F value	Pr(>F)
A	2.8923	175.018	147.0	49.00	0.3850	0.75662
B	3.6008	148.315	1781.3	445.34	3.5931	0.01024 *
A:B	12.0000	35.649	1738.6	144.88	0.9478	0.51323
Residuals	180.0000		21017.1	116.76		

```
ats.2(x~A*B,mydata12)
```

	Df	F value	Pr(>F)
A	2.914354	0.1708253	0.9111070
B	3.559425	2.1158563	0.1032501
A*B	7.981024	0.8242638	0.5858894
Residuals	40.747127		

Die 3 Ergebnisse geben exakt den Trend dieser Verfahren wider: die leicht liberale Methode BF, das konservative ATS und das ausgewogene WJ. Wem soll man nun trauen? In diesem Fall ist es der Test des Haupteffekts B, für den die Resultate nicht einheitlich sind. Somit können hier die „besten“ Methoden von James sowie von Alexander & Govern die Entscheidung treffen, die leider beide nur als 1-faktorielle Varianzanalyse bekannt sind:

```
library(oneWayTests)
james.test(x~B,mydata12)
ag.test(x~B,mydata12)
```

James Second Order Test	
-----	
statistic	: 22.72145
criticalValue	: 10.10662
Result	: Difference is statistically significant.
Alexander-Govern Test	
-----	
statistic	: 20.34211
parameter	: 4
p.value	: 0.0004274263
Result	: Difference is statistically significant.

Somit ist Faktor B als signifikant nachgewiesen, während Faktor A und die Interaktion keinen Einfluss haben.

## 11.4 Gemischter Versuchsplan mit Varianzheterogenitäten

Der Datensatz `mydata11` besteht (in dieser Reihenfolge) aus der abhängigen Variablen  $x$  zu 4 Zeitpunkten (Variablen  $V1, V2, V3, V4$ ) mit einem ganzzahligen Wertebereich zwischen 10 und 80, einer Fallidentifikation  $Id$  sowie einem Gruppierungsfaktor  $A$  (4 Gruppen) und umfasst 60 Fälle:

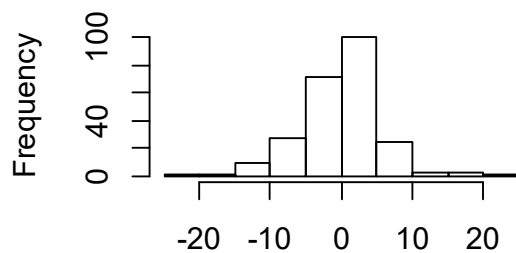
		n	Zeitpunkt			
			1	2	3	4
Mittelwerte	A 1	10	38.50	41.50	46.50	44.5
	A 2	10	36.50	46.00	43.00	36.5
	A 3	20	39.50	40.25	40.25	42.5
	A 4	20	40.25	39.00	41.00	43.0
Standard- abweichungen	A 1	10	4.74	14.62	9.56	14.54
	A 2	10	7.09	8.82	6.58	15.10
	A 3	20	4.84	7.48	5.45	6.05
	A 4	20	3.43	3.77	3.48	6.07

Zunächst einmal muss der Datensatz in das für Messwiederholungsanalysen erforderliche „long“-Format transformiert werden:

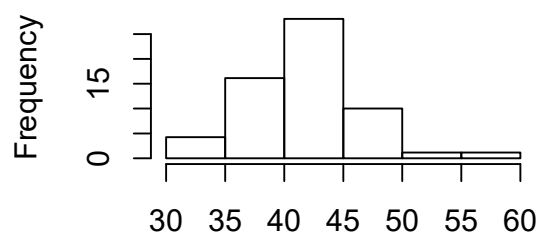
```
reshape(mydata11, direction="long", timevar="Zeit", v.names="x",
        varying=1:4, idvar="Id") -> mydata11t
within(mydata11t, Zeit <- factor(Zeit)) -> mydata11t
```

Zunächst wird eine parametrische Varianzanalyse ohne Messwiederholungen durchgeführt, um die Residuen zu ermitteln und auf Normalverteilung überprüfen zu können.

```
aov.3 <- aov(x ~ A * Zeit + Id, mydata11t)
hist(aov.3$residuals)
shapiro.test(aov.3$residuals)
```



`aov.3$residuals`



`mydata11$pm`

Shapiro-Wilk normality test

W = 0.97197, p-value = 0.0001113

Das Histogramm (oben links) zeigt eine leichte Linksschiefe, und der Shapiro-Test indiziert eine Abweichung von der Normalverteilung. Die Schiefe, zu berechnen mittels `skewness` aus

dem Paket `EnvStats`) beträgt  $-0.136$ , der Test auf Schiefe  $> 0$  ( $S/(6/n)$ ) ergibt:

$$0.136/(6/200) = -1.36$$

Dieser Wert ist  $\chi^2$ -verteilt mit 1 FG und deutet auf keine nennenswerte Schiefe hin. Wegen des relativ großen  $n$  von 60 (hier sogar 240 durch die Messwiederholungen) kann allerdings die Abweichung von der Normalverteilung vernachlässigt werden.

Als nächstes müssen noch die Personeneffekte  $\pi_i$  auf Normalverteilung überprüft werden:

```
within(mydata11, pm <- (V1+V2+V3+V4)/4) -> mydata11
shapiro.test(mydata11$pm)
hist(mydata11$pm)
```

Test Name:	Shapiro-Wilk normality test
Test Statistic:	W = 0.9288996
P-value:	0.001784088

Auch hier zeigen Histogramm (oben rechts) wie auch der Shapiro-Test eine Abweichung von der Normalverteilung. Als nächstes die Überprüfung der Varianzhomogenitäten: zuerst die Überprüfung der Sphärität mittels des Mauchly-Test in der Funktion `ezANOVA` (Paket `ez`):

```
ezANOVA(mydata11t, x, Id, between=. (A), within=. (Zeit))
```

Effect	DFn	DFd	F	p	p < .05	ges
2	A	3	56	0.567528	0.63870443	0.01145754
3	Zeit	3	168	2.437313	0.06642195	0.02622517
4	A:Zeit	9	168	2.344058	0.01625495	* 0.07210060

```
$`Mauchly's Test for Sphericity`
Effect      W          p p < .05
3  Zeit 0.6153137 6.927358e-05 *
4 A:Zeit 0.6153137 6.927358e-05 *
```

```
$`Sphericity Corrections`
Effect      GGe      p[GG] p[GG] < .05      HFe      p[HF] p[HF] < .05
3  Zeit 0.7493929 0.08490472      0.7822755 0.08221934
4 A:Zeit 0.7493929 0.02946172      * 0.7822755 0.02722404      *
```

Der Mauchly-Test (in der Mitte) zeigt eine deutliche Abweichung von der Annahme gleicher Varianzen der Messwiederholungsvariablen und gleicher Korrelationen an. Als nächstes die Überprüfung der Gleichheit der Varianzen der 4 Messwiederholungsvariablen für die 4 Gruppen von Faktor A sowie des Personeneffekts  $\pi_i$ . Hierzu ist der nichttransformierte Datensatz zu verwenden:

```
leveneTest(V1~A, mydata11)
leveneTest(V2~A, mydata11)
leveneTest(V3~A, mydata11)
leveneTest(V4~A, mydata11)
leveneTest(pm~A, mydata11)
```

```

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  1.1629 0.3321

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  1.8395 0.1505

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  2.8034 0.04803 *

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  5.1022 0.003424 **

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 3  3.7228 0.01639 *

```

woraus zu entnehmen ist, dass zumindest für die Variablen V3 und V4 sowie für den Personeneffekt keine Varianzhomogenität gegeben ist. Wegen der ungleichen Zellenbesetzungszahlen muss daher auch ein pairing überprüft werden:

```

si <- with(mydata11, tapply(pm, A, sd))      # sd von pm für jede A-Gruppe
ni <- with(mydata11, table(A))             # Häufigkeitstabelle für A
cor(ni, si^2)

```

Der Wert  $r = -0.86$  bestätigt ein negatives pairing, was die Durchführung der normalen Varianzanalyse weder zum Test des Faktors A noch zum Test der Interaktion ratsam macht. Lediglich der Messwiederholungsfaktor Zeit kann über o.a. Varianzanalyse geprüft werden, und zwar wegen der Verletzung der Sphärität im unteren Abschnitt (``Sphericity Corrections``), in dem hinteren Teil der Zeile „zeit“: Dort findet man den p-Wert 0.0822. Sinnvoller erscheint es, von vorneherein eine Methode zu verwenden, die robust gegen Varianzhomogenitäten ist. In R gibt es dazu die Verfahren von Huynh (IGA), von Brown & Forsythe (MBF) sowie von Welch & James mit entsprechenden Funktionen, die in der eigenen Bibliothek (siehe Anhang) verfügbar sind.

Für die IGA-Adjustierung ist es die Funktion `iga.anova`, die allerdings nicht den Gruppeneffekt A testet. Die Daten werden hier im breiten Format eingegeben. Die abhängigen Variablen sind die ersten 4, der Gruppierungsfaktor die Variable 6, der als `factor` deklariert werden muss. Eingabe und Ausgabe, in der mit `groups` der Gruppierungsfaktor und mit `w` der Messwiederholungsfaktor bezeichnet werden:

```

iga.anova(mydata11[, 1:4], factor(mydata11[, 6]))

```

```

Hyunh IGA (improved general approximation)
      Df Sum Sq Mean Sq F value Pr(>F)
W      2.4614  340.3  113.437  2.3941 0.08512 .
groups:W  5.1379  981.9  109.097  1.7302 0.13343
Residuals 93.3934 7819.1  46.542

```

Für das MBF-Verfahren gibt es die Funktion `mbf.f`, die im Gegensatz zu o.a. Funktion den Gruppeneffekt A testet. Auch hier werden die Daten im breiten Format eingegeben. Die

Eingabe ist identisch mit o.a.:

```
mbf.f(mydata11[,1:4], factor(mydata11[,6]))
```

```
modified Brown-Forsythe method for mixed repeated measures designs
              Df Df.err F value Pr(>F)
grouping factor 2.1840 21.827  0.4051 0.6893
trial factor    3.0000 21.090  4.7442 0.0111 *
interaction     7.3587 54.401  1.6096 0.1491
```

Für das Verfahren von Welch & James steht die Funktion `wj.spanova` zur Verfügung. Allerdings werden darin nur die Messwiederholungseffekte getestet. Der Gruppeneffekt A muss separat mit der korrespondierenden Funktion `wj.anova` getestet werden:

```
wj.spanova(mydata11t, 'x', 'A', 'Zeit', 'Id')
wj.anova(mydata11t, 'pm', 'A')
```

```
              F value df num df denom    p value
Zeit         4.857511      3 19.24192 0.01113892
A:Zeit       1.351924      9 23.58293 0.26440049

              Chi Sq  df  P(Chi>value)
A 3.017588      3      0.4035
```

Das Ergebnis: Lediglich der Faktor Zeit ist signifikant, wobei beim IGA-Verfahren Zeit höchstens als schwach signifikant eingestuft werden kann. Damit weicht dieses Ergebnis doch erheblich von dem oben mit `ezANOVA` erstellten ab und ist auf die nicht erfüllten Voraussetzungen zurückzuführen. Das Verfahren von Koch (Funktion `koch.anova`) verlangt weder Normalverteilungsforderung noch Sphärität, aber die Robustheit bei negativem pairing ist nicht gewährleistet.

## 11.5 Gemischter Versuchsplan: Prüfung der Voraussetzungen

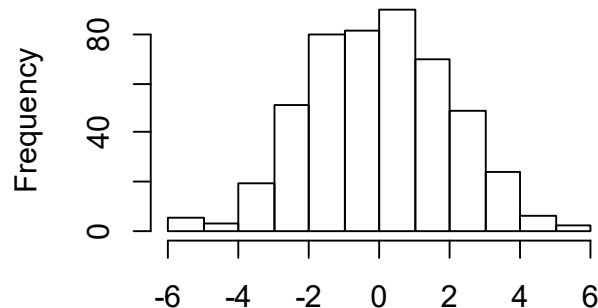
Der Datensatz `mydata13` besteht aus der abhängigen Variablen `score` zu 6 Zeitpunkten: jeweils 3 Werte (Faktor "Wdh") in einer 1. Phase sowie in einer 2. Phase. `score` hat einen ganzzahligen Wertebereich zwischen 0 und 12. Die Werte wurden in 4 unterschiedlichen Versuchsgruppen (Faktor "Gruppe") erhoben mit einer Fallidentifikation "Id" und umfasst 80 Versuchspersonen. Im Gegensatz zum vorherigen Datensatz `mydata11` (siehe 11.4) liegen hier zwei Messwiederholungsfaktoren vor, was insbesondere die Prüfung der Voraussetzungen erschwert:

Gruppe	n	Phase 1			Phase 2		
		Wdh 1	Wdh 2	Wdh 3	Wdh 1	Wdh 2	Wdh 3
Kontroll	24	5.5	5.3	5.1	6.8	6.8	5.8
Behandlung 1	16	5.1	6.1	4.9	5.5	5.8	6.8
Behandlung 2	8	5.6	7.5	6.5	7.1	6.1	7.0
Behandlung 3	32	5.3	5.9	5.7	7.1	6.3	6.8

Der Datensatz liegt bereits im long-Format vor. Für diesen sollen ausführlich die Voraussetzungen geprüft werden.

Zunächst zur Normalverteilung: Es wird eine parametrische Varianzanalyse ohne Messwiederholungen durchgeführt, um die Residuen zu ermitteln und auf Normalverteilung überprüfen zu können.

```
aov.1<-aov(score~Gruppe*Phase*Wdh+Id,mydata13)
hist(aov.1$residuals)
```



Das Histogramm zeigt keine erkennbaren Abweichungen von der Normalverteilung. Alternativ können auch die Residuen auf eine multivariate Normalverteilung überprüft werden, was sich insbesondere dann empfiehlt, wenn eine multivariate Varianzanalyse durchgeführt werden soll. Dazu müssen die Residuen (Variable `res`) mit dem Dataframe `mydata13` verbunden werden. Anschließend muss dieser in das breite Format transformiert werden. Dazu werden die Messwiederholungsfaktoren `wdh` und `Phase` in einer zusammengefasst (Variable `PH`), mit deren Hilfe die Transformation erfolgen kann (neuer Dataframe: `mydata13t`):

```
res      <- aov.1$residuals
mydata13 <- cbind(mydata13,res)
mydata13 <- within(mydata13x, PH<-(Phase-1)*3 + Wdh)
reshape(mydata13x,direction="wide", timevar="PH",
v.names=c("score","res"),times=1:6,idvar="Id")->mydata13t
```

Zur Kontrolle sollte man sich mit `names` die Namen des neuen Dataframes ausgeben lassen:

```
[1] "Id" "Gruppe" "Phase" "Wdh" "score.1" "res.1" "score.2" "res.2" "score.3" "res.3"
[11] "score.4" "res.4" "score.5" "res.5" "score.6" "res.6"
```

Die Überprüfung der Normalverteilung (für die Variablen 6, 8,...,14, 16) erfolgt dann über die Funktion `msw` im Paket `mvnnormalTest`:

```
msw(mydata13t1[,c(6,8,10,12,14,16)])
```

```
$mv.test
      Test      Statistic p-value Result
[1,] Royston 9.5833      0.1383 YES
[2,] VA-GE   0.9813      0.1985 YES

$uv.shapiro
      W      p-value UV.Normality
res.1 0.9928 0.9372 Yes
res.2 0.9838 0.41 Yes
res.3 0.9825 0.3451 Yes
res.4 0.9876 0.6388 Yes
res.5 0.9837 0.4023 Yes
res.6 0.957 0.0089 No
```

msw gibt 2 Varianten einer multivariaten Version des klassischen Tests von Shapiro & Wilk aus: eine von Royston und eine von Villaseñor-Alva und Gonzalez-Estrada (VA-GE). Beide erkennen keinen Widerspruch zur multivariaten Normalverteilung. Unter diesen beiden multivariaten Tests werden noch die Ergebnisse für die univariaten Tests von Shapiro & Wilk ausgegeben.

Interessanterweise erhält man bei Verwendung des Tests von Mardia ein anderes Ergebnis:

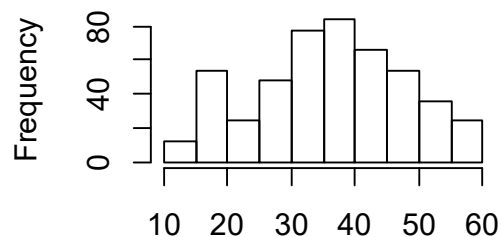
```
mardia(mydata13$t1[,c(6,8,10,12,14,16)])
```

\$mv.test	Test	Statistic	p-value	Result
1	Skewness	101.9943	2e-04	NO
2	Kurtosis	10.969	0	NO
3	MV Normality	<NA>	<NA>	NO

Ein solches “Phänomen“ ist zwar nicht selten, mag aber dennoch verwirren. Bei Verwendung des klassischen F-Tests ist dieses Ergebnis ohnehin von untergeordneter Bedeutung.

Als nächstes müssen noch die Personeneffekte  $\pi_i$  auf Normalverteilung überprüft werden:

```
with(mydata13, ave(score, Id, FUN=sum)) -> pm
hist(pm)
```



Hier ist zwar eine Abweichungen von der Normalverteilung offensichtlich, aber nicht gravierend.

Die Überprüfung der Sphärizität mit dem Mauchly-Test war im vorigen Abschnitt 11.4 gezeigt worden. Sollen andere Tests mittels der Funktion `check.sphere` durchgeführt werden, so ist das zwar ohne Weiteres als Voraussetzung für die Interaktion `Phase*Wdh` möglich, jedoch müssen für die Haupteffekte `Phase` und `Wdh` pro `Id` die Summen (oder Mittelwerte) über den jeweils anderen Faktor berechnet werden. Mittels der Funktion `ave` können diese in einer solchen Struktur errechnet werden, dass sie als neue Variablen an `mydata13` angehängt werden können. Zweckmäßigerweise wird noch aus beiden Messwiederholungsfaktoren eine Kombination “PW“ mit den Stufen 1,...,6 gebildet, wobei die Faktoren für die arithmetische Berechnung in “integer“ gewandelt werden müssen und die “3“ die Anzahl der Stufen des Faktors “Wdh“ ist:

```
with(mydata13, ave(score, list(Id, Wdh), FUN=mean)) -> Wdh.s
with(mydata13, ave(score, list(Id, Phase), FUN=mean)) -> Phase.s
mydata13 <- cbind(mydata13, Wdh.s, Phase.s)
mydata13 <- within(mydata13,
  PW<-factor((as.integer(Phase)-1)*3 + as.integer(Wdh)))
```

Der Test auf Sphärizität für die Interaktion kann zwar auch auf Basis des Datensatzes im langen Format durchgeführt werden, nämlich über

```
with(mydata13, check.sphere(score, trial=PW, id=Id))
```

doch für die Tests in Zusammenhang mit den beiden Haupteffekten muss der Dataframe in das "breite" Format gewandelt werden:

```
myPW <- reshape(mydata13x, direction="wide", timevar="PW",
                v.names=c("score"), times=1:6, idvar="Id")
```

Die Variablen von `myPW` sind nun:

Id, Gruppe, Phase, Wdh, Wdh.s, Phase.s, score.1, score.2, score.3, score.4, score.5, score.6

Zunächst der Test auf Sphärität mittels der Funktion `check.sphere` für die Interaktion, alternativ der nichtparametrische Test mittels "spatial signs" und der Funktion `sr.sphere.test` aus dem Paket `SpatialNP`. Anzumerken ist, dass die zu analysierenden Variablen die Positionen 7 bis 12 haben, wie aus o.a. Liste zu entnehmen ist:

```
check.sphere(myPW[, 7:12])
```

```
$results
              Chisquare df p value
Mauchly test      19.3992 14  0.1502
Likelihood Ratio  19.8773 20  0.4656
John's test       4305.4637 14  0.0000
John-Nagao        4305.4637 20  0.0000
Muirhead & Waternaud 32.4219 20  0.0390
compound symmetry 31.9650 19  0.0315

$Box.epsilon
[1] 0.9010927
```

```
library(SpatialNP)
sr.sphere.test(myPW[, 7:12])
```

```
Test of sphericity using spatial signs
Q.2 = 1158.7, df = 20, p-value < 2.2e-16
alternative hypothesis: true shape is not equal to diag(6)
```

Fast alle Ergebnisse deuten auf eine deutliche Abweichung von der Sphärität hin. Alternativ kann auch ein Test auf Gleichheit der Varianzen der 6 Messwiederholungen mittels des Levene-like-Tests durchgeführt werden (vgl. Kapitel 5.3.4), hier wieder auf Basis von `mydata13` (im "langen" Format):

```
mydata13<-within(mydata13, diff<-abs(score-ave(score,PW,FUN=median)))
with(mydata13, friedman.test(diff~PW | Id, data=mydata13y))
```

```
Friedman rank sum test

data:  diff and PW and Id
Friedman chi-squared = 4.7339, df = 5, p-value = 0.4492
```

Dieser zeigt keine Ungleichheit der Varianzen an. Der Unterschied der Ergebnisse kann dadurch erklärt werden, dass die Kovarianzmatrix zwar gleiche Varianzen aber unterschiedliche Korrelation hat. Ähnlich die Durchführung der Sphäritäts-Tests für die Haupteffekte, hier für den Faktor `wdh`. Wie oben muss ein entsprechender Datensatz im breiten Format erstellt werden, bevor `check.sphere` oder `sr.sphere.test` aufgerufen werden können. Die zu analysierenden Variablen haben hier die Positionen 7 bis 9:



```
myW <- reshape(mydata13x, direction="wide", timevar="Wdh",
               v.names=c("Wdh.s"), times=1:3, idvar="Id")
check.sphere(myW[, 7:9])
```

	Chisquare	df	p value
Mauchly test	0.3866	2	0.8242
Likelihood Ratio	0.3916	5	0.9956
John's test	238.4257	2	0.0000
John-Nagao	238.4257	5	0.0000
Muirhead & Waternaud	0.5037	5	0.9920
compound symmetry	9.9130	4	0.0419
\$Box.epsilon			
[1]	0.9950802		

Für den Test auf Spherizität für den Phasen-Effekt sind folgende Anweisungen erforderlich (hier ohne Ausgabe):

```
myP <- reshape(mydata13x, direction="wide", timevar="Phase",
               v.names=c("Phase.s"), times=1:2, idvar="Id")
check.sphere(myP[, 7:8])
```

Nun zur Überprüfung der Homogenität der Kovarianzmatrizen. Auch hier muss für alle 3 Messwiederholungseffekte (Phase, Wdh und Phase\*Wdh) ein Test durchgeführt werden. Zunächst für die Interaktion, was wie schon oben beim Test der Spherizität, mit beiden Datenformaten möglich ist:

```
with(mydata13x, check.covar(score, Gruppe, trial=PW, id=Id))
```

oder

```
check.covar(myPW[, 7:12], myW[, 2])
```

	Chisquare	df	p value
LR	81.8687	63.000	0.0553
Box M F	1.0190	2662.194	0.4357
Schott T1	52.3902	63.000	0.8274
Schott T2	76.2589	63.000	0.1219
Schott T3	96.3606	63.000	0.0043
Levene	0.9921	21.000	0.4862

Der Levene-Test ist der zuverlässigste, zumindest hinsichtlich der Varianz-Heterogenität. Da das Ergebnis nicht signifikant ist, ist es ratsam noch einen Test auf Homogenität der Korrelationsmatrizen zu machen, wobei auch hier in gleicher Weise beide Datenformate möglich sind:

```
with(mydata13, check.corr(score, Gruppe, trial=PW, id=Id))
```

	statistic	df	p value
Jennrich test (chisquare)	40.9123	45	0.6457
Levene correlation (F)	0.7413	45	0.8817
Larntz & Perlman (chisq)	11.5944	15	0.1256
Box M (chisq)	60.2983	45	0.0633

Das Ergebnis der Tests von Larntz & Perlman sowie von Box zeigt eine, zumindest leichte, Heterogenität der Korrelationsmatrizen. Die entsprechenden Anweisungen, z.B. für die Tests beim Faktor Wdh:

```
check.covar(myW[, 7:9], myW[, 2])
check.corr(myW[, 7:9], myW[, 2])
```

Da allerdings bereits für die Interaktion die Kovarianzmatrizen als ungleich angenommen werden, kann die Überprüfung für die beiden Haupteffekte entfallen. Als Anova-Methode ist das KWF-Verfahren die einfachste, aber eine gute Wahl, weil dies relativ robust gegen Heterogenitäten ist. Dazu die Funktion `np.anova` (vgl. Anhang 3), Aufruf und Ergebnis (allerdings aus Platzgründen ohne die Ergebnisse des parametrischen F-Tests):

```
np.anova(score~Gruppe*Phase*Wdh+Error(Id/(Phase*Wdh)), mydata13)
```

```
generalized Kruskal-Wallis/Friedman tests including Iman & Davenport
F-tests
```

	Df	Sum Sq	Chisq	Pr(>Chi)
Gruppe	3	125893	1.0808	0.78171
Residuals Btw.Vpn	76	9075923		
Phase	1	57	15.3901	0.00009
Gruppe:Phase	3	5	1.4415	0.69584
Residuals Phase	76	233		
Wdh	2	1	0.3587	0.83581
Gruppe:Wdh	6	14	4.6315	0.59186
Residuals Gruppe:Wdh	152	455		
Phase:Wdh	2	11	3.2757	0.19440
Gruppe:Phase:Wdh	6	25	7.5920	0.26955
Residuals Gruppe:Phase:Wdh	152	500		

Vom statistischen Standpunkt aus wäre als Anova-Methode das Verfahren von Brown & Forsythe das bessere gewesen. Doch leider berücksichtigt die Funktion `mbf.f`, wie auch `iga.anova` für das IGA-Verfahren, nur einen Messwiederholungsfaktor. Damit könnte nicht die Interaktion der beiden Messwiederholungsfaktoren getestet werden. Liegt keine Hypothese für diese Interaktion vor, kann diese Methode durchaus angewandt werden. Das Prozedere ist etwas aufwändig und wurde in Abschnitt 3.5 kurz skizziert:

- Es sind 2 Varianzanalysen durchzuführen, einmal mit `Gruppe` und `Phase`, sowie einmal mit `Gruppe` und `wdh`.
- Es müssen die Summen über den nicht berücksichtigten Messwiederholungsfaktor gebildet werden.
- Es muss eine Teildatenmatrix gewählt werden.

Hier soll nur die Varianzanalysen mit `Gruppe` und `Phase` gezeigt werden. Zunächst die Mittelwertbildung und Teildatenmatrix, hier für `wdh=1`:

```
within(mydata13, Sŷ<-ave(score, Id, Phase, FUN=sum)) ->mydata13.P
subset(mydata13.P, Wdh==1) ->mydata13.P1
```

Die Varianzanalyse nach Brown-Forsythe nun:

```
with(mydata13.P1, mbf.f(score, Gruppe, trial=Phase, id=Id))
```

```
modified Brown-Forsythe method for mixed repeated measures designs
```

	Df	Df.err	F value	Pr(>F)
grouping factor	2.7220	60.070	0.7455	0.517174
trial factor	1.0000	34.955	9.6221	0.003791 **
interaction	2.7786	55.220	0.7257	0.531088

Der Vollständigkeit wegen soll wenigstens noch das Ergebnis der Analyse für Gruppe und wdh angeführt werden:

	Df	Df.err	F value	Pr(>F)
grouping factor	2.8190	59.752	0.8531	0.4646
trial factor	2.0000	40.263	2.2946	0.1138
interaction	5.4673	126.846	0.6243	0.6957

so dass sich zeigt, dass zumindest in diesem Fall beide Methoden zum gleichen Ergebnis geführt haben.

Zum Abschluss noch eine mehr theoretische Betrachtung. Oben zeigte sich, dass sich die Kovarianzmatrizen nicht hinsichtlich der Varianzen sondern hinsichtlich der Korrelationen unterscheiden. In Kapitel 5.2 war beschrieben worden, wie sich ungleiche Korrelationen innerhalb der Gruppen bei ungleichen Stichproben  $n_i$  auswirken können. Welcher Fall, d.h. wie die Relation zwischen  $n_i$  und  $r_i$  aussieht, soll hier kurz ermittelt werden. Dazu werden für jede Gruppe mittels `subset` die Korrelationsmatrizen  $c_1, \dots, c_4$  errechnet, und daraus die durchschnittliche Korrelation  $sc_1, \dots, sc_4$  innerhalb einer Korrelationsmatrix mittels einer doppelten `for`-Schleife über die untere Dreiecksmatrix. Zum Mitteln der Korrelationskoeffizienten wird die Fisher-z-Transformation angewandt (Funktion `atanh`) und der Mittelwert zurücktransformiert (Funktion `tanh`). (Zur Berechnung der mittleren Korrelationen  $sc_1, \dots, sc_4$  werden die Summen der transformierten Korrelationskoeffizienten über die untere Dreiecksmatrix berechnet und anschließend durch  $nv*(nv-1)/2$  dividiert, wobei hier die Variablenzahl  $nv=6$  ist):

```

c1 <- cor(subset(myPW,Gruppe==1)[,7:12])
c2 <- cor(subset(myPW,Gruppe==2)[,7:12])
c3 <- cor(subset(myPW,Gruppe==3)[,7:12])
c4 <- cor(subset(myPW,Gruppe==4)[,7:12])
sc1<-0
for (i in 2:6) for (j in 1:(i-1)) sc1<-sc1+atanh(c1[i,j])
sc1<-sc1/(6*5/2)
sc2<-0
for (i in 2:6) for (j in 1:(i-1)) sc2<-sc2+atanh(c2[i,j])
sc2<-sc2/(6*5/2)
sc3<-0
for (i in 2:6) for (j in 1:(i-1)) sc3<-sc3+atanh(c3[i,j])
sc3<-sc3/(6*5/2)
sc4<-0
for (i in 2:6) for (j in 1:(i-1)) sc4<-sc4+atanh(c4[i,j])
sc4<-sc4/(6*5/2)

```

Die für die 4 Gruppen ermittelten durchschnittlichen Korrelationen:

```
[1] 0.5127412 0.6379254 0.4907381 0.2519690
```

Diese werden nun mit den  $n_i$  (24, 16, 8, 32) korreliert:

```
cor(c(24,16, 8,32),c(sc1,sc2,sc3,sc4))
```

mit dem Ergebnis  $r(n_i, r_i) = -0.674$ . Die relativ hohe negative Korrelation hätte zur Folge, dass der F-Test deutlich konservativ reagieren würde, dagegen der multivariate Test deutlich liberal.

## A. Anhang

### 1. Umstrukturieren von Messwiederholungen in SPSS

Dieses ist z.B. erforderlich zur Rangbildung von Messwiederholungen.

#### 1.1 Umstrukturieren von Messwiederholungen in Fälle

Vorzunehmen im Menü: „Daten -> Umstrukturieren“

##### 1.1.1 ein Faktor und eine Analyse-Variable

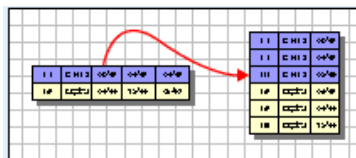
Als Beispiel dient der Datensatz 4 (vgl. Kapitel 5)

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

#### • Datenumstrukturierung

1. Option:

Umstrukturieren ausgewählter Variablen in Fälle



Folgende Möglichkeiten stehen Ihnen zur Verfügung:

- Umstrukturieren ausgewählter Variablen in Fälle

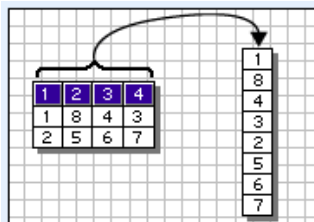
Verwenden Sie diese Option, wenn jeder Fall in den aktuellen Daten Variablen enthält, die im neuen Datenblatt in Gruppen verwandter Fälle angeordnet werden sollen.

-> Weiter

#### • Anzahl der Variablengruppen

1. Option:

Eine (Variablengruppe)



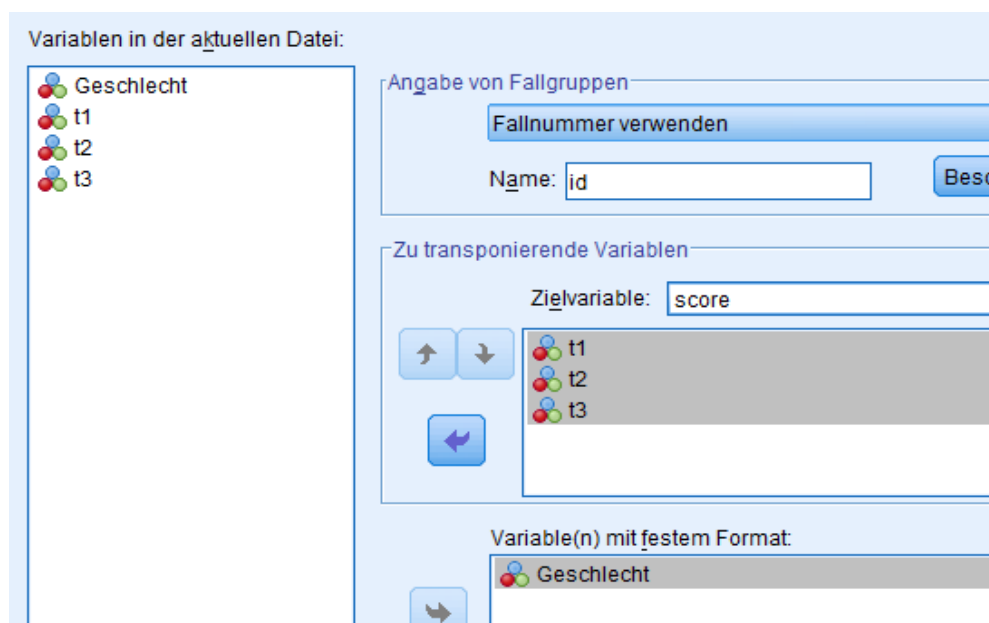
Wieviele Variablengruppen möchten Sie umstrukturieren?

- Eine (beispielsweise w1, w2 und w3)

-> Weiter

- **Auswählen von Variablen**

- Fallnummer verwenden,
  - kann eine vorhandene Fallkennung sein, z.B. Vpn
  - ist aber frei wählbar
  - erhält standardmäßig den Namen id
- zu transponierende Variablen:
  - hier die Messwiederholungsvariablen eintragen
  - und einen gemeinsamen Namen geben, hier: „score“
- Variablen mit festem Format:
  - hier die "konstanten" Variablen (ohne Messwiederholung) eintragen
  - (z.B. Alter, Geschlecht etc)



-> Weiter

- **Wieviele Indexvariablen möchten Sie erstellen?**

1. Option:

Eine (Indexvariablen)

1	1	1	0.07
1	1	2	0.11
1	1	3	0.05
2	1	1	0.08
2	1	2	0.04
2	1	3	0.06

Wieviele Indexvariablen möchten Sie erstellen?

Eine

Verwenden Sie diese Option, wenn eine Variablen­gruppe die Effekte eines einzelnen Faktors, einer einzelnen Behandlung oder einer einzelnen Bedingung aufzeichnet.

-> Weiter

- **Erstellen einer Indexvariablen**

(Diese kann numerisch oder alphanumerisch sein.)

- Art des Indexwertes:
  - fortlaufende Zahlen

## b. Name und Label der Indexvariablen:

kann frei gewählt werden (standardmäßig: Index1), hier: „Zeit“

Art des Indexwerts:

Fortlaufende Zahlen  
Indexwerte: 1, 2, 3

Variablennamen  
Indexwerte: t1, t2, t3

Name und Label der Indexvariablen bearbeiten:

	Name	Variablenlabel	Stufen	Indexwerte
1	Zeit		3	1, 2, 3

-> Weiter (es folgen dann noch Optionen) oder Fertigstellen

## • Optionen

a. Verarbeitung nicht ausgewählter Variablen (die oben weder als zu transponierende noch als "konstante" deklariert worden waren):

(normalerweise) beibehalten und als Variablen mit festem Format behandeln

b. System Missing: Einen Fall in der neuen Datei erstellen

Verarbeitung nicht ausgewählter Variablen

Variable(n) aus neuer Datendatei entfernen

Beibehalten und als Variable(n) mit festem Format behandeln

System Missing (fehlender Wert) oder leere Werte in allen transponierten Variablen

Einen Fall in der neuen Datei erstellen

Daten verwerfen

Variable zum Zählen von Fällen

Anzahl neuer Fälle zählen, die vom Fall in den aktuellen Daten erstellt wurden

Name:

Beschriftung:

-> Weiter

Die hier aufgeführten Schritte können auch über die SPSS-Syntax realisiert werden:

```
Varstocases
  /Id=id
  /Make score from t1 t2 t3
  /index=Zeit(3)
  /keep=Geschlecht
  /null=keep.
```

Das Ergebnis der Umstrukturierung:

	id	Geschlecht	Zeit	score
1	1	1	1	4
2	1	1	2	7
3	1	1	3	2
4	2	1	1	3
5	2	1	2	5
6	2	1	3	1
7	3	1	1	7
8	3	1	2	9
9	3	1	3	6
10	4	1	1	6
11	4	1	2	6
12	4	1	3	2

### 1.1.2 mehrere Faktoren und eine Analyse-Variablen

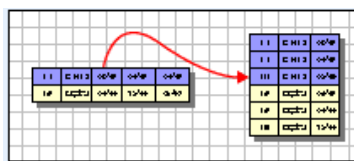
Als Beispiel dient der Datensatz 5 (vgl. Kapitel 5)

	Geschlecht	v1	v2	v3	v4	v5	v6	v7	v8	v9
1	1	3	3	1	4	4	2	5	4	3
2	1	2	0	0	3	2	2	4	3	3
3	1	5	4	3	5	3	3	6	3	4
4	1	3	5	2	4	4	3	4	4	4
5	2	2	2	1	2	2	2	5	2	3
6	2	4	1	0	3	2	1	5	2	2
7	2	3	2	1	3	2	1	4	3	2
8	2	1	3	0	5	2	1	6	3	3

#### • Datenumstrukturierung

1. Option:

Umstrukturieren ausgewählter Variablen in Fälle



Folgende Möglichkeiten stehen Ihnen zur Verfügung:

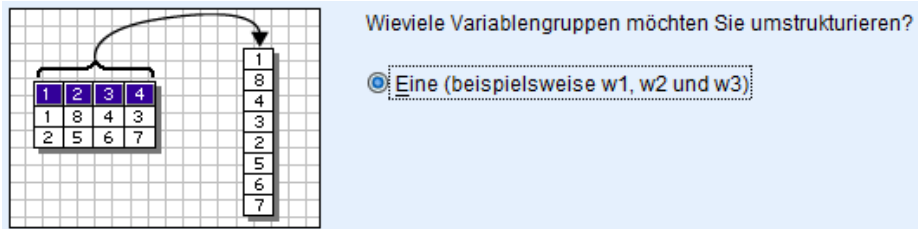
- Umstrukturieren ausgewählter Variablen in Fälle

Verwenden Sie diese Option, wenn jeder Fall in den aktuellen Daten Variablen enthält, die im neuen Datenblatt in Gruppen verwandter Fälle angeordnet werden sollen.

-> Weiter

- **Anzahl der Variablengruppen**

1. Option: Eine (Variablengruppe)



-> Weiter

- **Auswählen von Variablen**

a. Fallnummer verwenden,

- kann eine vorhandene Fallkennung sein, z.B. Vpn
- ist aber frei wählbar
- erhält standardmäßig den Namen id

b. zu transponierende Variablen:

hier die Messwiederholungsvariablen eintragen und einen gemeinsamen Namen geben, hier: „Fehler“

c. Variablen mit festem Format:

hier die "konstanten" Variablen (ohne Messwiederholung) eintragen (z.B. Alter, Geschlecht etc)



-> Weiter

- **Wieviel Indexvariablen möchten Sie erstellen?**

2. Option:

Mehrere (Indexvariablen) und Anzahl der Messwiederholungsfaktoren festlegen



1	1	1	1	0.07
1	1	1	2	0.11
1	1	1	3	0.05
1	1	2	1	0.08
1	1	2	2	0.04
1	1	2	3	0.06

 Mehrere
Wie viele? 

Verwenden Sie diese Option, wenn eine Variablengruppe die Effekte mehrerer Faktoren, Behandlungen oder Bedingungen aufzeichnet

-&gt; Weiter

- **Erstellen mehrerer Indexvariablen)**

In der folgenden Tabelle müssen für jeden Messwiederholungsfaktor Name und wahlweise Label frei gewählt (standardmäßig: Index1, Index2) sowie für jede die Anzahl der Stufen festgelegt werden, hier „Medikament“ und „Aufgabe“. Hierbei ist die Reihenfolge zu beachten: in der Variablenreihenfolge variiert der erste Faktor am langsamsten, der letzte am schnellsten. Und das Produkt der Stufen muss die Anzahl der Messwiederholungsvariablen ergeben:

Namen, Label und Anzahl der Ebenen für Indexvariablen:

	Name	Variablenlabel	Stufen	Indexwerte
1	Medikament		3	1, 2, 3
2	Aufgabe		3	1, 2, 3

Gesamtzahl kombinierter Ebenen (Produkt): 9

-&gt; Weiter

- Optionen

a. Verarbeitung nicht ausgewählter Variablen (die oben weder als zu transponierende noch als "konstante" deklariert worden waren): (normalerweise) beibehalten und als Variablen mit festem Format behandeln

b. System Missing: Einen Fall in der neuen Datei erstellen

Verarbeitung nicht ausgewählter Variablen

Variable(n) aus neuer Datendatei entfernen

Beibehalten und als Variable(n) mit festem Format behandeln

---

System Missing (fehlender Wert) oder leere Werte in allen transponierten Variablen

Einen Fall in der neuen Datei erstellen

Daten verwerfen

---

Variable zum Zählen von Fällen

Anzahl neuer Fälle zählen, die vom Fall in den aktuellen Daten erstellt wurden

Name:

Beschriftung:

-> Weiter

-> Fertigstellen

Wenn keine Namen festgelegt worden waren, hat die Analyse-Variablen anschließend die Namen `trans1` und `Index1`, `Index2`,... sind standardmäßig die Kennzeichnungen der Messwiederholung für die jeweiligen Faktoren.

Die hier aufgeführten Schritte können auch über die SPSS-Syntax realisiert werden:

```
Varstocases
  /Id=id
  /make Fehler from v1 v2 v3 v4 v5 v6 v7 v8 v9
  /index=Medikament(3) Aufgabe(3)
  /keep=Geschlecht
  /null=keep.
```

Das Ergebnis der Umstrukturierung:

	id	Geschlecht	Medikament	Aufgabe	Fehler
1	1	1	1	1	3
2	1	1	1	2	3
3	1	1	1	3	1
4	1	1	2	1	4
5	1	1	2	2	4
6	1	1	2	3	2
7	1	1	3	1	5
8	1	1	3	2	4
9	1	1	3	3	3
10	2	1	1	1	2
11	2	1	1	2	0
12	2	1	1	3	0

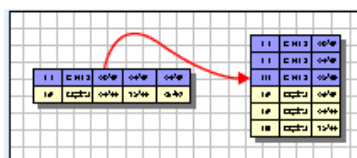
### 1.1.3 ein Faktor und mehrere Analyse-Variablen

Als Beispiel dient der Datensatz 4 (vgl. Kapitel 5), wobei die 3 Aufgaben nicht als Faktor, sondern als 3 Variablen interpretiert werden und lediglich ein Faktor Medikament vorhanden ist.

#### • Datenumstrukturierung

1. Option: Umstrukturieren ausgewählter Variablen in Fälle

-> Weiter



Folgende Möglichkeiten stehen Ihnen zur Verfügung:

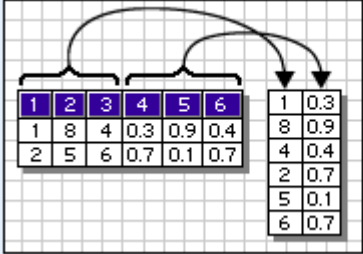
- Umstrukturieren ausgewählter Variablen in Fälle

Verwenden Sie diese Option, wenn jeder Fall in den aktuellen Daten Variablen enthält, die im neuen Datenblatt in Gruppen verwandter Fälle angeordnet werden sollen.

#### • Anzahl der Variablengruppen

2. Option:

Mehrere (Variablengruppen) sowie Anzahl der Analyse-Variablen festlegen (hier 3)



Mehrere (beispielsweise w1, w2, w3 und h1, h2, h3, usw.)  
 Anzahl

-> Weiter

- **Auswählen von Variablen**

a. Fallnummer verwenden,

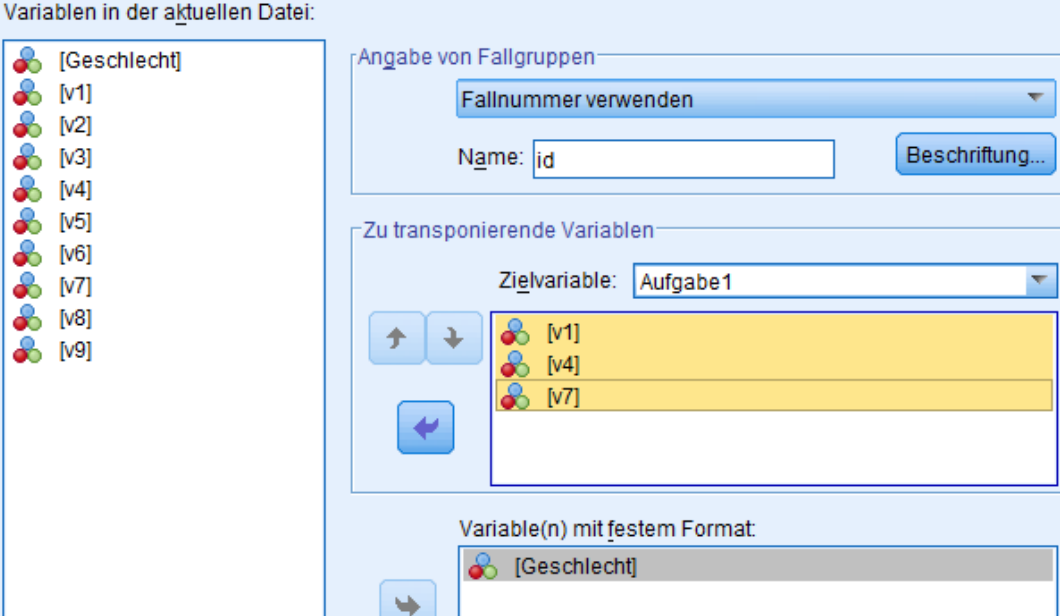
- kann eine vorhandene Fallkennung sein, z.B. Vpn
- ist aber frei wählbar
- erhält standardmäßig den Namen id

b. zu transponierende Variablen:

hier die Messwiederholungsvariablen für die 1. abhängige Variable eintragen und bei „Zielvariable“ einen gemeinsamen Namen geben, hier: „Aufgabe1“ diesen Schritt dann für die anderen abhängigen Variablen wiederholen, indem im Pulldown-Menü rechts neben der Zielvariablen nacheinander die nächsten Variablen ausgewählt werden, deren Voreinstellung `trans1`, `trans2`, ... ist.

c. Variablen mit festem Format:

hier die "konstanten" Variablen (ohne Messwiederholung) eintragen (z.B. Alter, Geschlecht etc)

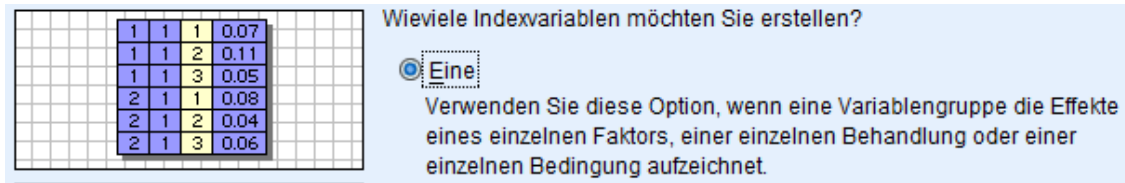


-> Weiter

- **Erstellen von Indexvariablen)**

1. Option:

Eine (Indexvariablen)



Wieviele Indexvariablen möchten Sie erstellen?

Eine

Verwenden Sie diese Option, wenn eine Variablengruppe die Effekte eines einzelnen Faktors, einer einzelnen Behandlung oder einer einzelnen Bedingung aufzeichnet.

Liegt ein mehrfaktorielles Design vor, wie etwa im vorigen Abschnitt, so können bei der 2. Option die Anzahl der Messwiederholungsfaktoren festgelegt werden.

-> Weiter

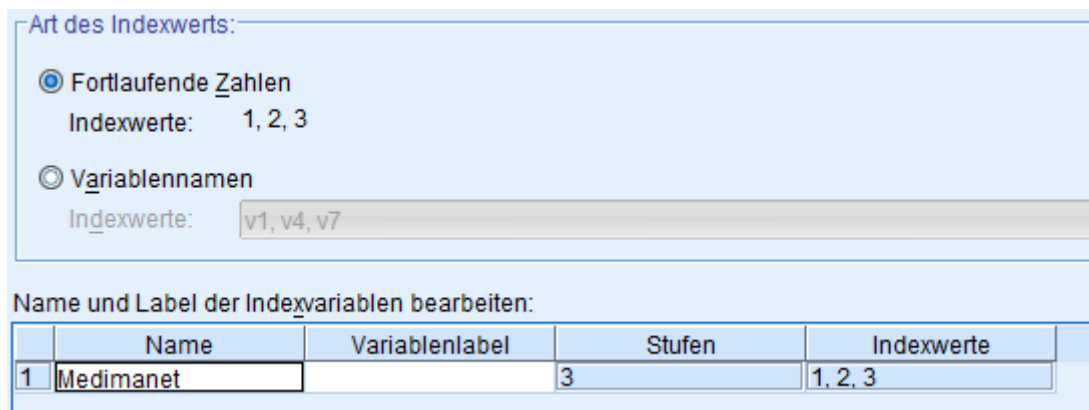
- **Erstellen einer Indexvariablen**

(Diese kann numerisch oder alphanumerisch sein.)

a. Art des Indexwertes:  
fortlaufende Zahlen

b. Name und Label der Indexvariablen:

kann frei gewählt werden (standardmäßig: Index1), hier: „Medikament“. Die Stufenzahl ergibt sich aus den anderen Angaben.



Art des Indexwerts:

Fortlaufende Zahlen  
Indexwerte: 1, 2, 3

Variablennamen  
Indexwerte: v1, v4, v7

Name und Label der Indexvariablen bearbeiten:

	Name	Variablenlabel	Stufen	Indexwerte
1	Medimanet		3	1, 2, 3

-> Weiter

- **Optionen**

a. Verarbeitung nicht ausgewählter Variablen (die oben weder als zu transponierende noch als "konstante" deklariert worden waren):

(normalerweise) beibehalten und als Variablen mit festem Format behandeln

b. System Missing: Einen Fall in der neuen Datei erstellen

Verarbeitung nicht ausgewählter Variablen

Variable(n) aus neuer Datendatei entfernen

Beibehalten und als Variable(n) mit festem Format behandeln

---

System Missing (fehlender Wert) oder leere Werte in allen transponierten Variablen

Einen Fall in der neuen Datei erstellen

Daten verwerfen

---

Variable zum Zählen von Fällen

Anzahl neuer Fälle zählen, die vom Fall in den aktuellen Daten erstellt wurden

Name:

Beschriftung:

Falls keine Namen vereinbart worden waren, haben die Analyse-Variablen anschließend die Namen `trans1`, `trans2`, ... und `Index1` ist standardmäßig der Kennzeichnung der Messwiederholung.

Die hier aufgeführten Schritte können auch über die SPSS-Syntax realisiert werden:

```
Varstocases
  /Id=id
  /make Aufgabe1 from v1 v4 v7
  /make Aufgabe2 from v2 v5 v8
  /make Aufgabe3 from v3 v6 v9
  /Index=Medikament(3)
  /Keep=Geschlecht
  /Null=keep.
```

Das Ergebnis der Umstrukturierung:

	id	Geschlecht	Medikament	Aufgabe1	Aufgabe2	Aufgabe3
1	1	1	1	3	3	1
2	1	1	2	4	4	2
3	1	1	3	5	4	3
4	2	1	1	2	0	0
5	2	1	2	3	2	2
6	2	1	3	4	3	3
7	3	1	1	5	4	3
8	3	1	2	5	3	3
9	3	1	3	6	3	4
10	4	1	1	3	5	2
11	4	1	2	4	4	3
12	4	1	3	4	4	4

## 1.2 Umstrukturieren von Fälle in Messwiederholungen

Vorzunehmen im Menü: „Daten -> Umstrukturieren“

- **Datenumstrukturierung**

2. Option: Umstrukturieren ausgewählter Variablen in Fälle

11	EM13	001e	001e	001e
11	EM14	001e	001e	001e
12	EM13	001e	001e	001e
12	EM14	001e	001e	001e
12	EM15	001e	001e	001e

11	EM13	001e	001e	001e
11	EM14	001e	001e	001e
12	EM13	001e	001e	001e
12	EM14	001e	001e	001e
12	EM15	001e	001e	001e

- **Umstrukturieren ausgewählter Fälle in Variablen**

Verwenden Sie diese Option, wenn Gruppen verwandter Fälle vorliegen, die neu angeordnet werden sollen, sodass die Daten aus den einzelnen Gruppen im neuen Datenblatt als einzelner Fall dargestellt werden.

-> Weiter

- **Auswählen von Variablen**

a. Bezeichnervariablen:

Fallkennzeichnung/nummer (z.B. id oder Vpn)

b. Indexvariable:

Kennzeichnungen der Messwiederholung, hier „Zeit“

(z.B. 1-faktoriell: Index1 bzw. mehrfaktoriell Index1, Index2,...)

Variablen in der aktuellen Datei:

- [patients]
- [score]
- Rank of score by id [Rscore]

Bezeichnervariable(n):

- id

Indexvariable(n):

- Zeit

Alle übrigen Variablen werden automatisch „sinnvoll“ als konstante oder Messwiederholungsvariable zugeordnet.

-> Weiter

- **Sortieren von Daten**

2. Option:

Nein - Daten wie gegenwärtig sortiert verwenden

(Bei 1. Option werden zuerst alle Wiederholungen einer Analyse-Variablen hintereinander ausgegeben, vor denen der nächsten Analysevariablen,

bei 2. Option werden zuerst die ersten Werte aller Analyse-Variablen hintereinander aus-

gegeben, vor allen Werten der zweiten Messwiederholung etc)

Sollen die aktuellen Daten sortiert werden?

Ja - Daten werden nach Bezeichner- und Indexvariablen sortiert.

-> Weiter

- **Optionen**

Anordnung der neuen Variablengruppen

Nach ursprünglicher Variable sortieren (z. B.: w1 w2 w3, h1 h2 h3)

Nach Index sortieren (z. B.: w1 h1, w2 h2, w3 h3)

Variable zum Zählen von Fällen

Anzahl der Fälle in den aktuellen Daten zählen, mit denen ein neuer Fall erstellt wird

Name:

Beschriftung:

Indikatorvariablen

Indikatorvariablen erstellen

Stammname:

Die Optionen sind i.a. nicht erforderlich.

-> Weiter

-> Fertigstellen

Die neuen Namen der Messwiederholungen der einzelnen Analyse-Variablen sind Name.1, Name2, .. (wenn ein Name vorgegeben wurde) andernfalls trans.1, trans.2, ... Bei mehrfaktoriellen Designs haben diese jeweils den Zusatz der Kennzeichnung der Messwiederholung z.B. .1.1, .1.2, ..., 2.1, 2.2, ...

Die hier aufgeführten Schritte können auch über die SPSS-Syntax realisiert werden:

```
Sort cases by id Zeit.
casestovars
  /Id=id
  /index=Zeit
  /groupby=variable.
```

Und das Ergebnis der Umstrukturierung:

	id	patients	score.1	score.2	score.3	Rscore.1	Rscore.2	Rscore.3
1	1	1	4	7	2	2,000	3,000	1,000
2	2	1	3	5	1	2,000	3,000	1,000
3	3	1	7	9	6	2,000	3,000	1,000
4	4	1	6	6	2	2,500	2,500	1,000
5	5	1	5	5	1	2,500	2,500	1,000
6	6	2	8	2	5	3,000	1,000	2,000
7	7	2	4	1	1	3,000	1,500	1,500
8	8	2	6	3	4	3,000	1,000	2,000
9	9	2	9	5	2	3,000	2,000	1,000
10	10	2	7	1	1	3,000	1,500	1,500



## 2. Spezielle robuste F-Tests und andere Statistiken

Im Folgenden werden drei robuste F-Tests vorgestellt, deren Formeln in der Literatur nicht weit verbreitet sind und daher hier zitiert werden.

### 2.1 Box-Korrektur für heterogene Varianzen

Es liegen  $k$  Gruppen (Haupteffekt oder Interaktion) mit Varianzen  $s_i^2$  vor. Der F-Test

$$F = \frac{MS_{Effekt}}{MS_{Fehler}}$$

kann bzgl. der Heterogenität der Varianzen korrigiert werden, indem die Zähler- und Nennerfreiheitsgrade adjustiert (genauer: reduziert) werden. Die Zählerfreiheitsgrade  $df1$  werden dabei mit  $\varepsilon_1$  multipliziert, die Nennerfreiheitsgrade  $df2$  mit  $\varepsilon_2$ . Diese Korrekturfaktoren errechnen sich wie folgt:

$$\begin{aligned}\bar{s}^2 &= (\sum s_i^2)/k \\ c^2 &= \left( \sum (s_i^2 - \bar{s}^2)^2 \right) / (k \cdot \bar{s}^4) \\ \varepsilon_1 &= \left( 1 + \frac{k-2}{k-1} c^2 \right)^{-1} \quad \varepsilon_2 = (1 + c^2)^{-1}\end{aligned}$$

Hierbei lassen sich  $\bar{s}^2$  als durchschnittliche Varianz und  $c^2$  als Streuung der Varianzen interpretieren. Es ist leicht zu erkennen, dass im Falle gleicher Varianzen  $c^2=0$  wird und damit  $\varepsilon_1$  und  $\varepsilon_2$  den Wert 1 bekommen.

### 2.2 Brown-Forsythe F-Test für inhomogene Varianzen

*1-faktorielle Analyse:*

Es liegen  $k$  Gruppen mit Varianzen  $s_i^2$ , Zellenbesetzungen  $n_i$  vor. Brown & Forsythe bilden den folgenden Quotienten, der annähernd F-verteilt ist:

$$F = \frac{SS_{Effekt}}{SS_{Fehler}}$$

Hierbei errechnet sich  $SS_{Error}$  (mit  $n = \sum n_i$ )

$$SS_{Error} = \sum \left( 1 - \frac{n_i}{n} \right) s_i^2$$

Die Nennerfreiheitsgrade des F-Tests berechnen sich

$$df = \left( \sum \frac{m_i^2}{n_i - 1} \right)^{-1} \quad m_i = \left( 1 - \frac{n_i}{n} \right) s_i^2 / (SS_{Error})$$

*2-faktorielle Analyse:*

Der Test der Interaktion erfolgt (relativ aufwändig) mittels Kontrasten. Einzelheiten hierzu sind der Veröffentlichung von Brown & Forsythe (1974) zu entnehmen

### 2. 3 Box-Andersen F-Test für nichtnormalverteilte Variablen

Bei diesem modifizierten F-Test werden dessen Zähler- und Nennerfreiheitsgrade mit dem Parameter  $d$  multipliziert. Dieser errechnet sich im Wesentlichen aus der Varianz und dem Exzess der Variablen  $x$ . Die folgende Berechnung des Korrekturparameters  $d$  ist gültig für annähernd gleiche  $n_i$ . Sei daher  $n$  die Anzahl der Beobachtungen pro Gruppe. Es sei erwähnt, dass es auch eine etwas kompliziertere Formel für stark differierende  $n_i$  gibt.

$$S_2 = \sum_i^k \sum_j^n (x_{ij} - \bar{x})^2 \quad S_4 = \sum_i^k \sum_j^n (x_{ij} - \bar{x})^4$$

Daraus werden zwei Zwischengrößen berechnet:

$$k_2 = S_2 / (n - 1)$$

$$k_4 = [n(n + 1)S_4 - 3(n - 1)S_2^2] / [(n - 1)(n - 2)(n - 3)]$$

Schließlich errechnet sich hieraus  $d$  als

$$d = 1 + \frac{1}{n} \frac{k_4}{k_2^2}$$

### 2. 4 Box-Cox-Transformationen

Hier geht es darum, einen passenden Parameter  $a$  zu finden, so dass die Funktion, angewandt auf die abhängige Variable, varianzstabilisierend wirkt.

$$f(x) = \frac{x^a - 1}{a}$$

Für den Parameter  $a$  gilt:

- $0 < a < 1$  rechtsschiefe Verteilungen symmetrisch machen
- $1 < a$  linksschiefe Verteilungen symmetrisch machen

Schließlich gilt, dass  $f(x) \rightarrow \log(x)$  für  $a \rightarrow 0$ .

Mehr dazu unter:

<http://de.wikipedia.org/wiki/Box-Cox-Transformation>

### 2. 5 Fishers combined probability test

Mit *Fishers combined probability test* können mehrere unabhängig voneinander gewonnene Testergebnisse zur gleichen Hypothese  $H_0$  über deren p-Werte zusammengefasst werden. Das Verfahren ist für beliebige Tests anwendbar, also z.B. auch für den W-Test von Shapiro und Wilk zur Überprüfung eines Merkmals auf Normalverteilung, etwa für  $k$  Variablen oder  $k$  Stichproben. Werden für  $k$  Tests die p-Werte  $P_1, \dots, P_k$  erzielt, dann wird mit der folgenden Testgröße  $X$  die Hypothese geprüft, dass für alle  $k$  Tests  $H_0$  richtig ist:

$$X = -2[\ln(P_1) + \ln(P_2) + \dots + \ln(P_k)]$$

$X$  ist  $\chi^2$ -verteilt mit  $2k$  Freiheitsgraden.

Mehr dazu unter [https://en.wikipedia.org/wiki/Fishers\\_method](https://en.wikipedia.org/wiki/Fishers_method)

## 2. 6 Levene-Test auf Gleichheit von Kovarianzmatrizen

Der Levene-Test auf Gleichheit von mehreren Varianzen aus unabhängigen Stichproben ist allgemein bekannt: Dabei werden die Absolutbeträge der Abweichungen der einzelnen Messungen vom Median ermittelt und diese über eine Varianzanalyse auf Gleichheit getestet. Ein ähnlicher Test auf Gleichheit mehrerer Kovarianzmatrizen (vgl. O'Brien, 1992) ist dagegen relativ unbekannt, obwohl er im Vergleich zu anderen Tests, etwa dem bekannten Box-Test, wesentlich robuster und effizienter ist. Hier die Berechnung:

Für  $i=1, \dots, I$  Gruppen,  $k=1, \dots, n_i$ ,  $N$  ( $N = \sum n_i$ ) Erhebungseinheiten und  $j=1, \dots, J$  Messwiederholungen wird mit  $y_{k(i)j}$  die  $j$ . Messung der Variablen  $y$  für Versuchsobjekt  $k(i)$  und  $m_{ij}$  der Median in Gruppe  $i$ . Nun werden für jede Erhebungseinheit  $k(i)$  die folgenden Kovarianzen  $s_{j_1 j_2}$  berechnet:

$$s_{k(i)j_1 j_2} = (y_{k(i)j_1} - m_{ij_1})(y_{k(i)j_2} - m_{ij_2}) \quad (j_1, j_2 = 1, \dots, J)$$

die anschließend transformiert werden in

$$\hat{s}_{k(i)j_1 j_2} = \text{sgn}(s_{k(i)j_1 j_2}) \cdot \sqrt{|s_{k(i)j_1 j_2}|}$$

wobei  $\text{sgn}$  die Vorzeichen-Funktion bezeichnet. Im nächsten Schritt wird für jede Erhebungseinheit  $k(i)$  die untere Dreiecksmatrix von  $\hat{s}_{k(i)j_1 j_2}$  in einen Vektor umgewandelt, woraus eine Datenmatrix  $Y$  mit  $N$  Reihen und  $(J+1)J/2$  Spalten resultiert. Schließlich wird darauf eine multivariate Varianzanalyse angewandt, z.B. Wilks Lambda Test, der die Homogenität der Kovarianzmatrizen für die Gruppen  $i=1, \dots, I$  testet.

## 2. 7 Wilcoxon-Test auf Gleichheit von Varianzen bei Messwiederholungen

Dieser Test prüft die Gleichheit der Varianzen von  $J$  abhängigen Variablen  $x_1, \dots, x_J$ . Er basiert auf dem Prinzip des bekannten Levene-Tests (vgl. Wilcoxon, 1989).

Die dazu erforderlichen Schritte:

- Für jede Variable  $x_1, \dots, x_J$  wird der Median  $m_j$  ermittelt.
- Für jede Beobachtungseinheit  $i=1, \dots, N$  berechnen der absoluten Differenz  $d_{ij} = \text{abs}(x_{ij} - m_j)$ .
- Durchführung eines Friedman-Tests über die  $d_j$ .  
Falls die Nullhypothese gleicher Mittelwerte für die  $d_j$  verworfen wird, kann daraus auf eine Varianzheterogenität der  $x_j$  geschlossen werden.

### 3. R-Funktionen

Die folgenden Funktionen zusammen mit einer Benutzungsanleitung sind alle im Verzeichnis

<http://www.uni-koeln.de/~luepsen/R/>

zu finden und können von dort heruntergeladen werden.

Hinweis: Fehlende Werte müssen mittels `na.omit(...)` eliminiert werden.

#### 3.1 **box.f: Box-F-Test für inhomogene Varianzen**

Durchführung einer 1- oder 2-faktoriellen Varianzanalyse (ohne Messwiederholungen) unter Verwendung der robusten F-Tests von Box (vgl. Anhang 2.1).

Aufruf: `box.f (Modell, Dataframe)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion <code>aov</code> ) Beispiel: $x \sim A*B$
Dataframe	Datensatz, Objekt vom Type <code>Dataframe</code>

#### 3.2 **bf.f: Brown & Forsythe-F-Test für inhomogene Varianzen**

Durchführung einer 1- oder 2-faktoriellen Varianzanalyse (ohne Messwiederholungen) unter Verwendung der robusten F-Tests von Brown & Forsythe (vgl. Anhang 2.2).

Aufruf: `bf.f (Modell, Dataframe)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion <code>aov</code> ) Beispiel: $x \sim A*B$
Dataframe	Datensatz, Objekt vom Type <code>Dataframe</code>

#### 3.3 **mbf.f: modified Brown & Forsythe-F-test für inhomogene Varianzen in gemischten Versuchsplänen (split-plot)**

2-faktorielle Varianzanalyse mit Messwiederholungen auf einem Faktor unter Verwendung der robusten F-Tests von Brown & Forsythe in der Bearbeitung von Vallejo et al. (2004), die sowohl heterogene als auch nicht-sphärische Kovarianzmatrizen erlaubt. Die Eingabe kann im breiten wie auch im langen Format erfolgen. Die Funktion benötigt o.a. Funktion `bf.f`.

Aufruf (wide format): `mbf.f (df, group)`

Aufruf (long format): `with(df,mbf.f (y, groups=g, trial=w, id=id))`

Parameter:

<code>df</code>	Dataframe, Object vom Typ <code>data.frame</code> , bei weitem Format: nur die Messwiederholungsvariablen
<code>group</code>	Gruppierungsfaktor
<code>trial</code>	Messwiederholungsfaktor
<code>id</code>	Fallidentifikation

Ergebnis:

anova table            object of type data.frame and anova

### 3.4      **box.andersen.f: F-Test für nichtnormalverteilte Variablen**

Durchführung einer 1- oder 2-faktoriellen Varianzanalyse (ohne Messwiederholungen) unter Verwendung der robusten F-Tests von Box & Andersen (vgl. Anhang 2.3) zur Kompensierung von Abweichungen von der Normalverteilung.

Aufruf: `box.andersen.f (Modell, Dataframe)`

Aufrufparameter:

Modell                    varianzanalytisches Modell (vgl.Funktion aov)  
                               Beispiel:  $x \sim A*B$   
 Dataframe                Datensatz, Objekt vom Type Dataframe

Ergebnisobjekte:

anova                    Anova-Tabelle  
 eps                      Korrekturfaktor d zur Korrektur der Freiheitsgrade

### 3.5      **check.covar: Test auf Homogenität von Kovarianzmatrizen**

Durchführung diverser Tests zur Prüfung der Homogenität von Kovarianzmatrizen:

- Likelihood Ratio Test,
- Box M-test, basierend auf der multivariaten Normalverteilung,
- Schott's T1, eine Verbesserung des Box M-Tests,
- Schott's T2, für elliptische Verteilungen, unter Verwendung des Exzesses zur Messung der Abweichung von der Normalverteilung, unter der Annahme, dass für alle Gruppen die Verteilungen gleich sind,
- Schott's T3, wie T2, jedoch unter der Annahme, dass nicht für alle Gruppen die Verteilungen gleich sind,
- multivariater Levene Test, der nur ordinales Skalenniveau voraussetzt.

Die Funktion kann sowohl auf Datensätze im "wide format" als auch im "long format" angewandt werden.

Aufruf (wide format): `check.covar (abh. Variablen, Gruppierungsfaktor)`

Aufruf (long format): `check.covar (abh. Variable, groups=., trial=., Id=.)`

Aufrufparameter (wide format):

Abh. Variablen        Variablen des Messwiederholungsfaktors als Dataframe oder Matrix  
 Gruppierungsfaktor    Vektor mit den Werten des Gruppierungsfaktors

Aufrufparameter (long format):

Abh. Variable        abhängige Variable als Dataframe  
 groups                Vektor mit den Werten des Gruppierungsfaktors  
 trial                  Vektor mit den Werten des Messwiederholungsfaktors  
 id                     Vektor mit den Kennzeichnungen der Erhebungseinheiten

Beispiele: `check.covar (winer[,c("V3","V4","V5")], winer$V2)`  
`with(winer,check.covar(v,sex,time,id))`

Ausgabe:

dataframe mit  $\chi^2$ -values, df und p Wert für die 6 Tests, NAs wenn ein Test nicht durchgeführt werden kann, z.B. wenn  $n$  zu klein in Relation zur Anzahl der Wiederholungen.

### 3.6 check.corr: Test auf Homogenität von Korrelationsmatrizen

Durchführung diverser Tests zur Prüfung der Homogenität von Korrelationsmatrizen:

- Jennrich-Test,
- Larntz & Perlman-Test,
- ein etwas abgeänderter, auf Korrelationen beschränkter Levene-like-Test von O'Brien,
- ein etwas abgeänderter, auf Korrelationen beschränkter Box-M-Test

Die Funktion kann sowohl auf Datensätze im "wide format" als auch im "long format" angewandt werden.

Aufruf (wide format): `check.corr (abh. Variablen, Gruppierungsfaktor)`

Aufruf (long format): `check.corr (abh. Variable, groups=., trial=., Id=.)`

Aufrufparameter (wide format):

Abh. Variablen	Variablen des Messwiederholungsfaktors als Dataframe oder Matrix
Gruppierungsfaktor	Vektor mit den Werten des Gruppierungsfaktors

Aufrufparameter (long format):

Abh. Variable	abhängige Variable als Dataframe
groups	Vektor mit den Werten des Gruppierungsfaktors
trial	Vektor mit den Werten des Messwiederholungsfaktors
id	Vektor mit den Kennzeichnungen der Erhebungseinheiten

Beispiele: `check.corr (winer[,c("V3","V4","V5")], winer$V2)`  
`with(winer,check.corr(v,sex,time,id))`

### 3.7 check.sphere: Test auf Spherizität

Durchführung diverser Tests Spherizität:

- John's V Test (see John, 1972),
- John's V Test, unter Verwendung der Exzesses zur Messung der Abweichung von der Normalverteilung (Li & Yao, 2016),
- John's V Test, mit einer exakteren Berechnung des p Werts (see Nagao, 1973),
- Mauchly's Test (sie z.B. Winer, 1991, p. 255),
- Likelihood Ratio Test, auf dem die o.a. Verfahren basieren,
- multisample Mauchly Test (Mendoza, 1980),
- multisample Mauchly Test (Harris, 1984),
- Likelihood Ratio Test unter Verwendung des Exzesses zur Messung der Abweichung von der

Normalverteilung (Muirhead & Waternaud, 1980)

- Test auf Zirkularität (compound symmetry), beschrieben in Winer (1991, p. 517).

Die Funktion kann sowohl auf Datensätze im “wide format“ als auch im “long format“ angewandt werden.

Aufruf (wide format): `check.sphere (abh. Variablen, Gruppierungsfaktor)`

Aufruf (long format): `check.sphere (abh. Variable, groups=., trial=., Id=.)`

Aufrufparameter (wide format):

Abh. Variablen	Variablen des Messwiederholungsfaktors als Dataframe oder Matrix
Gruppierungsfaktor	Vektor mit den Werten des Gruppierungsfaktors

Aufrufparameter (long format):

Abh. Variable	abhängige Variable als Dataframe
groups	Vektor mit den Werten des Gruppierungsfaktors
trial	Vektor mit den Werten des Messwiederholungsfaktors
id	Vektor mit den Kennzeichnungen der Erhebungseinheiten

Beispiele: `check.covar (winer[,c("V3", "V4", "V5")], winer$V2)`  
`with(winer, check.covar(v, sex, time, id))`

Output:

`$results`: dataframe mit  $\chi^2$ -values, df und p Wert für die 6 Tests.  
 Für John's Test mit Exzess-Berücksichtigung enthält die Spalte `chisquare` den Exzess .

`$Box.epsilon`: Wert für Box  $\epsilon$

`$error`: error code:  
 2: Datenmatrix und Gruppierungsvektor haben unterschiedliche Länge.

### 3. 8      **ats.2 und ats.3: 2- bzw. 3-faktorielle Varianzanalyse**

`ats.2` führt eine 2-faktorielle Varianzanalyse (ohne Messwiederholungen) nach dem Verfahren von Akritas, Arnold und Brunner (1997) durch sowie `ats.3` eine 3-faktorielle Analyse. Errechnet wird die F-verteilte ATS (anova type statistic). Leere Zellen sind nicht erlaubt.

Aufruf: `ats.2 (Modell, Dataframe)`

bzw. `ats.3 (Modell, Dataframe)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion <code>aov</code> ) Beispiel: <code>x ~ A*B</code>
Dataframe	Datensatz, Objekt vom Type Dataframe

### 3. 9      **np.anova: nichtparametrische Varianzanalyse mittels des KWF-Verfahrens und der von Puri & Sen und van der Waerden**

`np.anova` führt eine mehrfaktorielle Varianzanalyse (mit und ohne Messwiederholungen) wahlweise nach den Verfahren von Puri & Sen (L-Statistik), Puri & Sen mit INT-Transforma-

tion, verallgemeinerte Kruskal-Wallis- und Friedman-Analysen (KWF) oder van der Waerden durch. Im Fall von Messwiederholungen muss der Datensatz die gleiche Struktur haben, wie sie von `aov` oder `ezANOVA` gefordert wird.

Aufruf: `np.anova (Modell, Dataframe)` KWF-Methode  
 bzw. `np.anova (Modell, Dataframe, method=1)` van der Waerden  
 bzw. `np.anova (Modell, Dataframe, method=2)` Puri & Sen (L statistic)  
 bzw. `np.anova (Modell, Dataframe, method=3)` Puri & Sen mit INT-Transformation

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion <code>aov</code> ) Beispiele: $x \sim A*B$ oder $\text{score} \sim \text{gruppe}*\text{Zeit}+\text{Error}(\text{Vpn}/\text{Zeit})$
Dataframe	Datensatz, Objekt vom Type <code>Dataframe</code>
method	0 (Methode von Puri & Sen) oder 1 (Methode von van der Waerden)
compact	im Falle von Messwiederholungen: T: alle Tests in einer <code>Dataframe</code> -Tabelle (default) F: für jeden Fehlerterm eine getrennte Tabelle (wie bei <code>summary(aov)</code> )
pseudo	F (klassische Ränge) oder T (Pseudo-Ränge), vgl. Kapitel 2.16

### 3. 10 **art1.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Gruppierungsfaktoren)**

`art1.anova` führt eine mehrfaktorielle Varianzanalyse ohne Messwiederholungen nach dem ART-Verfahren (Aligned Rank Transform) durch. Eine Transformation der Ränge in normal scores ist möglich.

Aufruf: `art1.anova (Modell, Dataframe, method=..., main=..., adjust=..., INT=...)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion <code>aov</code> ) Beispiel: $x \sim A*B$
Dataframe	Datensatz, Objekt vom Type <code>Dataframe</code>
method	0: Berechnung der Residuen über eine Regression (default) 1: Berechnung der Residuen als Abweichungen vom Zellenmittelwert
main	F: für die Tests der Haupteffekte nur das RT-Verfahren (default) T: für die Tests der Haupteffekte ebenfalls das ART-Verfahren
adjust	0: Alignment (Adjustierung) mittels arithmetischem Mittel (default) 1: Alignment (Adjustierung) mittels Median
INT	F: ohne INT-Transformation nach der Rangbildung (default) T: mit INT-Transformation nach der Rangbildung

### 3. 11 **art2.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Messwiederholungsfaktoren)**

`art2.anova` führt eine mehrfaktorielle Varianzanalyse mit Messwiederholungen auf zwei Faktoren nach dem ART-Verfahren (Aligned Rank Transform) durch. Eine Transformation der Ränge in normal scores ist möglich.

Aufruf: `art2.anova (Modell, Dataframe, main=..., INT=...)`



Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion aov) Beispiel: $x \sim \text{Medi} * \text{Aufgabe} + \text{Error}(\text{Vpn}/(\text{Medi} * \text{Aufgabe}))$
Dataframe	Datensatz, Objekt vom Type Dataframe
main	F: für die Tests der Haupteffekte nur das RT-Verfahren (default) T: für die Tests der Haupteffekte ebenfalls das ART-Verfahren
INT	F: ohne INT-Transformation nach der Rangbildung (default) T: mit INT-Transformation nach der Rangbildung

### 3. 12 art3.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (für gemischte Versuchspläne)

art3.anova führt eine mehrfaktorielle Varianzanalyse für Versuchspläne mit mindestens einem Gruppierungsfaktor und ein oder zwei Messwiederholungsfaktoren nach dem ART-Verfahren (Aligned Rank Transform) durch. Im Fall von 3-faktoriellen Versuchsplänen wird keine Adjustierung für die 3er-Interaktion vorgenommen. Eine Transformation der Ränge in normal scores ist möglich.

Aufruf: `art3.anova (Modell, Dataframe, method=..., main=..., INT=...)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion aov) Beispiel: $\text{score} \sim \text{gruppe} * \text{Zeit} + \text{Error}(\text{Vpn}/\text{Zeit})$
Dataframe	Datensatz, Objekt vom Type Dataframe
method	0: Berechnung der Residuen über eine Regression (default) 1: Berechnung der Residuen als Abweichungen vom Zellenmittelwert
main	F: für die Tests der Haupteffekte nur das RT-Verfahren (default) T: für die Tests der Haupteffekte ebenfalls das ART-Verfahren
INT	F: ohne INT-Transformation nach der Rangbildung (default) T: mit INT-Transformation nach der Rangbildung

### 3. 13 wj.anova: Welch-James-Varianzanalyse für heterogene Varianzen (nur Gruppierungsfaktoren)

1- oder 2-faktorielle Varianzanalyse für unabhängige Faktoren nach dem Verfahren von Welch & James.

Aufruf: `wj.anova (Dataframe, abh. Variable, Faktor 1, Faktor 2)`

Aufrufparameter:

Dataframe	Datensatz, Objekt vom Type Dataframe
abh. Variable	Name in “...“
Faktor 1	Name in “...“
Faktor 2	Name in “...“ (optional)

### 3. 14 **wj.spanova: Welch-James-Varianzanalyse für heterogene Varianzen (für gemischte Versuchspläne)**

Aufruf: `wj.spanova (Dataframe, abh. Variable, F1, F2, Fallkennung)`

Aufrufparameter:

Dataframe	Datensatz, Objekt vom Type Dataframe
abh. Variable	Name in “...”
F1 (Gruppierungsfaktor)	Name in “...”
F2 (Messwiederholungsfaktor)	Name in “...”
Fallkennzeichnungsvariable	Name in “...”

### 3. 15 **koch.anova: nichtparametrische Varianzanalyse für gemischte Versuchspläne nach dem Verfahren von G.Koch**

Varianzanalyse für einen Gruppierungs- und einen Messwiederholungsfaktor. Entsprechend der Veröffentlichung (Gary Koch: *Some aspects of the statistical analysis of split plot experiments in completely randomized layouts*. Journal of the American Statistical Association, Vol. 64, No. 326 (Jun., 1969), pp. 485-505) sind mehrere Varianten des Verfahrens möglich. Die Eingabe verlangt den Datensatz im „wide format“, also alle Werte eines Falles in einer Zeile.

Aufruf: `koch.anova (Dataframe, Gruppierungsfaktor, A=..., B=...)`

Aufrufparameter:

Dataframe	Datensatz vom Type Dataframe, der ausschließlich die Messwiederholungen enthält
Gruppierungsfaktor	Vektor
A	0: univariater Kruskal-Wallis -Test für Fallmittelwerte 1: multivariate Kruskal-Wallis -Test
B	0: W Test, unter der Annahme beliebiger Verteilungsformen 1: $W_N^*$ Test, unter der Annahme gleicher Verteilungsformen 2: $W_{ni}^*$ Test, unter der Annahme gleicher Verteilungsformen

### 3. 16 **iga und iga.anova: general approximation test (GA) und improved general approximation test (IGA) von H.Huynh**

Adjustierungen für den parametrischen F-Test in 2-faktoriellen gemischten Versuchsplänen (split-plot designs) nach der GA- wie auch der IGA-Methode von Huynh (1978), zur Berücksichtigung von nicht-sphärischen und heterogenen Kovarianzmatrizen. Die Funktion `iga` berechnet die Korrekturfaktoren für die beiden Messwiederholungseffekte, `iga.anova` führt eine komplette Varianzanalyse unter Verwendung von `iga` durch. `iga.anova` erlaubt die Eingabe sowohl im breiten wie auch im langen Format.

Aufruf: `iga (df, group)`

Aufruf (wide format): `iga.anova (df, group)`

Aufruf (long format): `with(df, iga.anova (y, groups=g, trial=w, id=id, ga=T/F))`

Parameter:

<code>df</code>	Dataframe, Object vom Typ <code>data.frame</code> , bei weitem Format: nur die Messwiederholungsvariablen
<code>group</code>	Gruppierungsfaktor
<code>trial</code>	Messwiederholungsfaktor
<code>id</code>	Fallidentifikation
<code>ga</code>	F: IGA adjustment, T: GA adjustment

Ergebnis von `iga`: Liste mit 4 Vektoren

<code>GA.B</code>	GA-adjustments for the repeated measures main effect
<code>GA.AB</code>	GA-adjustments for the interaction effect
<code>IGA.B</code>	IGA-adjustments for the repeated measures main effect
<code>IGA.AB</code>	IGA-adjustments for the interaction effect

mit je 3 Elementen:

1. correction factor `c` for the F value
2. adjusted degrees of freedom for the numerator `df1`
3. adjusted degrees of freedom for the denominator `df2`

Wenn `F` der nichtadjustierte F-Wert der parametrischen Anova ist, dann ist  $F/c$  mit (`df1`, `df2`) Freiheitsgraden zu testen.

Ergebnis von `iga.anova`: Anova-Tabelle

### 3. 17 **ap.anova: nichtparametrische Varianzanalyse für Messwiederholungen und split-plot Versuchspläne von Agresti & Pendergast**

1- und 2-faktorielle Varianzanalyse mit einem Messwiederholungsfaktor nach A. Agresti & J. Pendergast, basierend auf dem multivariaten Test von Hotelling & Lawley, der keine Sphärität der Kovarianzmatrix und damit keine Varianzhomogenität voraussetzt (Tian & Wilcox, 2007). Es kann auch eine Interaktion mit einem Gruppierungsfaktor getestet werden (Beasley, 2002). Die Daten werden im langen Format eingegeben.

Aufruf: `ap.anova (dataframe, dependent var, case id, trial factor [,grouping factor])`

Parameter:

<code>dataframe</code>	data, object of class <code>data.frame</code>
<code>dependent variable</code>	abhängige Variable
<code>case id</code>	Fallidentifikation (class factor)
<code>trial factor</code>	Messwiederholungsfaktor (class factor)
<code>grouping factor</code>	Gruppierungsfaktor (class factor, optional)

Variablesnamen müssen in “...” eingeschlossen werden.

Ergebnis:

Anova-Tabelle      object of class `data.frame`

Example:

```
ap.anova (winer518t, "score", "Vpn", "Geschlecht", "Zeit")
```

### 3. 18 **simple.effects: parametrische Analyse von simple effects**

Analyse der simple effects für ein oder mehrere Gruppierungs- und maximal einen Messwiederholungsfaktor. (Literatur: B.J.Winer et al, 1991, 422 ff und 526 ff).

Die Eingabe verlangt ausnahmsweise den Datensatz im „wide format“, also alle Werte eines Falles in einer Zeile.

Aufruf: `simple.effects (Anova, Interaktion, Dataframe, adjust=...)`

Aufrufparameter:

Anova	Ergebnis-Objekt der Varianzanalyse der Funktion <code>aov</code>
Interaktion	Spezifikation der Interaktion, z.B. „Geschlecht*Zeit“ , mehrere zu analysierende Interaktionen können mittels <code>c(...)</code> zusammengefasst werden.
Dataframe	Datensatz vom Type Dataframe, der auch für <code>aov</code> verwendet wurde
adjust	optional: $\alpha$ -Adjustierung, vgl. R-Funktion <code>p.adjust</code> (default: „none“)

### 3. 19 **gee.anova: Anova-like tests for GEE and GLMM models**

2 Anova-like Wald-Tests für 2-faktorielle Designs: `gee.anova` für einen klassischen Wald-Test (vgl. Kapitel 9.8) sowie `gee.robANOVA` für einen robusten Wald-Test nach Fan & Zhang. Ersterer ist sehr liberal insbesondere bei GEE- und GLMM-Modellen, bei denen die Kovarianzmatrizen der Parameterschätzungen generell zu klein geschätzt werden und dadurch zu große  $\chi^2$ -Werte erzeugen. (Literatur: Li, Peng & Redden, David T., 2015, sowie Fan, C. & Zhang, D., 2014).

Aufruf: `gee.anova (coefficients, covariance matrix, degrees of freedom, n)`  
`gee.robANOVA (coefficients, covariance matrix, degrees of freedom)`

Parameter:

<i>coefficients</i>	regression coefficients (details see below)
<i>covariance matrix</i>	
<i>degrees of freedom</i>	Array with 3 df for 2 factors and the interaction
<i>n</i>	sample size (required for the F test)

Ergebnis:

The result is a dataframe with 3 rows, one for each of the 3 effects with columns:

```

gee.anova: degrees of freedom
            $\chi^2$ -value
           p value
gee.robANOVA degrees of freedom
            $\chi^2$ -value
           corresponding p value
           F-value
           corresponding p value
nerror: 0 for no errors
err.invert: 0 for no errors while computing the inverse

```

## Literaturhinweise

- Agresti, A. & Pendergast, J. (1986): Comparing mean ranks for repeated measures data. *Communications in Statistics - Theory and Methods*, 15, No 5, pp 1417-1433.
- Akritas, Michael G. , Arnold, Steven F. & Brunner, Edgar (1997): *Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs*, Journal of the American Statistical Association, Volume 92, Issue 437 , pages 258-265
- Akritas, Michael & Brunner, Edgar (2003): *Nonparametric Models for ANOVA and ANCOVA, a Review* . in "Recent Advances and Trends in Nonparametric Statistics" (Eds. M.G. Akritas and D.N. Politis), 79-91.
- Alexander, R.A., Govern, D.M. (1994). A New and Simpler Approximation for ANOVA Under Variance Heterogeneity. *Journal of Educational Statistics*, 19 (2), pp. 91-101.
- Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and psychological measurement*, 44(1), pp 39-48.
- Algina, J. (1994): Some Alternative Approximate Tests for a Split Plot Design. *Multivariate Behavioral Research*, 29 (4), pp.365-384).
- Beasley, T.Mark (2002): Multivariate Aligned Rank Test for Interactions in multiple Group repeated Measures Design, *Multivariate Behavioral Research*, 37 (2), 197-226
- Beasley & Zumbo (2003): Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics & Data Analysis*, 42, pp 569 – 593
- Beasley, T.M., Erickson, S., Allison, D.B. (2009): Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavioural Genetics*, 39 (5), pp 380-395
- Beasley, T.Mark & Zumbo, Bruno D. (2009): Aligned Rank Tests for Interactions in Split-Plot Designs: Distributional Assumptions and Stochastic Heterogeneity, *Journal of Modern Applies Statistical Methods*, Vol 8, NO. 1 , pp 16-50
- Bennett, B.M. (1968): Rank-order tests of linear hypotheses, *J. of Stat . Society*, B 30: 483-489.
- Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., Bendayan, R. (2017): *Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?* Behavior Research Methods, pp 1-26.
- Bogard, Matt (2011): *Linear Regression and Analysis of Variance with a Binary Dependent Variable*, <http://econometricsense.blogspot.de/2011/08/linear-regression-and-analysis-of.html>
- Boik, Robert J (1981): A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, Vol 46, No 3, pp 241-255.
- Boos, D. D., & Brownie, C. (1995): ANOVA and rank tests when the number of treatments is large. *Statistics & Probability Letters*, 23(2), pp 183-191.
- Bortz, Jürgen (1984): *Statistik*, Springer Lehrbuch, Berlin

- Bortz, J. , Lienert, G.A. , Boehnke, K. (2008): *Verteilungsfreie Methoden in der Biostatistik*, Springer, (gekürzte Neuauflage des Klassikers)
- Box, G.E.P. (1953): *Non-normality and tests on variances*, *Biometrika* 40, pp. 318-335
- Box, G.E.P. (1954): *Some theorems on quadrature forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification*. *Annals of Mathematical Statistics*, 25, pp 290-302
- Box, G.E.P. & Andersen, S.L. (1955): *Permutation Theory in the Derivation of robust criteria and the study of departures from assumption*, *Journal of the Royal Statistical Society, Series B*, Vol XVII, No 1
- Bredenkamp, J. (1974): *Nonparametrische Prüfung von Wechselwirkungen*, *Psychologische Beiträge* 16, 398-416
- Brown, M.B. & Forsythe, A.B. (1974): *The Anova and Multiple Comparisons for Data with Heterogeneous Variances*. *Biometrics*, Vol. 30, No. 4, pp. 719-724
- Brunner, E., Dette, H. and Munk, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association*, 92, pp 1494-1502.
- Brunner, E., Munzel, U. and Puri, M.L. (1999): *Rank-Score Tests in Factorial Designs with Repeated Measures*, *Journal of Multivariate Analysis* 70, 286-317
- Brunner, E. & Munzel, U. (2002): *Nichtparametrische Datenanalyse - unverbundene Stichproben*, Springer, ISBN 3-540-43375-9
- Brunner, Edgar & Munzel, Ullrich (2013): *Nichtparametrische Datenanalyse, Unverbundene Stichproben*, Springer, 126 ff.
- Brunner, E., Konietschke, F., Bathke, A.C. & Pauly, M. (2020): Ranks and Pseudo-ranks - Surprising Results of Certain Rank Tests in Unbalanced Designs. *International Statistical Review*, doi:10.1111/insr.12418.
- Bryan, Jennifer Joanne (2009): *Rank transforms and tests of interaction for repeated measures experiments with various covariance structures*, Oklahoma State University, Dissertation
- Cardinal, Rudolf N. (2004): *ANOVA in practice, and complex ANOVA designs*, [http://egret.psychol.cam.ac.uk/psychology/graduate/Guide\\_to\\_ANOVA.pdf](http://egret.psychol.cam.ac.uk/psychology/graduate/Guide_to_ANOVA.pdf)
- Carletti, I. , Clautriaux, J.J. (2005). Anova or Aligned Rank Transform Methods: Which one use when Assumptions are not fulfilled ? *Buletinul USAMV-CN*, nr. 62/2005 and below, ISSN, pp 1454-2382.
- Chatfield, Mark & Mander, Adrian (2009): *The Skillings–Mack test*, *Stata Journal*, 9(2): pp 299–305.
- Cleary, Paul D. & Angel, Ronald (1984): *The Analysis of Relationships Involving Dichotomous Dependent Variables*, *Journal of Health and Social Behavior*, 25, pp. 334-348.
- Clinch, Jennifer J. & Keselman, H. J. (1982): *Parametric Alternatives to the Analysis of Variance*, *Journal of Educational Statistics*, Vol. 7, No. 3, pp. 207-214

- Cochran, W.G. (1950): *The comparison of percentages in matched samples*. Biometrika 3
- Conover, W.J. (1980): *Practical nonparametric Statistics*, Wiley,  
(Standardverfahren mit einigen Zusatzinformationen)
- Conover, W. J. & Iman, R. L. (1981): *Rank transformations as a bridge between parametric and nonparametric statistics*. American Statistician 35 (3): 124–129.
- Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk, R. E. (1992). *Power comparisons of eight tests for sphericity in repeated measures designs*. Journal of Educational Statistics, 17, 233-249.
- D'Agostino, Ralph B. (1971): *A Second Look at Analysis of Variance on Dichotomous Data*, Journal of Educational Measurement, Vol. 8, No. 4, pp. 327-333
- Danbaba, Abubakar (2009): *A Study of Robustness of Validity and Efficiency of Rank Tests in AMMI and Two-Way ANOVA Tests*, Thesis, University of Ilorin (Nigeria)
- Dawson, Robert J. MacG. (1995): *The 'Unusual Episode' Data Revisited*, Journal of Statistics Education, 3
- Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F-test instead of the Classical F-test in One-Way ANOVA. International Review of Social Psychology, 32 (1), 13.  
DOI: <http://doi.org/10.5334/irsp.198>
- Diaz-Bone, Rainer & Künemund, Harald (2003): *Einführung in die binäre logistische Regression*, Freie Universität Berlin, Mitteilungen aus dem Schwerpunktbereich Methodenlehre, Heft Nr. 56  
<http://www.rainer-diaz-bone.de/Logreg.pdf>
- Dijkstra, J. B. (1987). Analysis of means in some non-standard situations. Technische Universiteit, Eindhoven DOI: 10.6100/IR272914.
- Erceg-Hurn, David M. & Mirosevich, Vikki M. (2008): *Modern robust statistical methods*, American Psychologist, Vol. 63, No. 7, 591–601
- Ernst, Michael D. & Kepner, James I. (1993) A monte carlo study of rank tests for repeated measures designs, *Communications in Statistics - Simulation and Computation*, 22:3, pp 671-678,
- Fan, Weihua (2006): *Robust means modelling: An Alternative to Hypothesis Testing of Mean Equality in Between-subject Designs under Variance Heterogeneity and Nonnormality*, Dissertation, University of Maryland  
<http://drum.lib.umd.edu/bitstream/1903/3786/1/umi-umd-3627.pdf>
- Fan, C. & Zhang, D. (2014): Robust small sample inference for generalised estimating equations: An application of the Anova-type test.  
*Australian & New Zealand Journal of Statistics*, 56(3), pp 237–255.
- Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M. (2014): *Log-transformation and its implications for data analysis*. Shanghai Archives of Psychiatry, Vol. 26, No. 2, pp 105-109.

- Field, Andy (2009): *Discovering Statistics using SPSS*, Sage Publications, London
- Fischer D., Mosler K., Möttönen J., Nordhausen K., Pokotylo O., Vogel D. (2020). Computing the Oja Median in R: The Package OjaNP. *Journal of Statistical Software*, 92(8), 36 p. <http://dx.doi.org/10.18637/jss.v092.i08>.
- Fox, J. & Weisberg, S. (2011): *An R Companion to Applied Regression*. SAGE Publications, Los Angeles.
- Friedrich, S., Brunner, E., & Pauly, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153, pp 255-265.
- Gao, X. and Alvo, M. (2005). A nonparametric test for interaction in two-way layouts. *Canadian Journal of Statistics*, Volume 33, Issue 4, pp 529–543.
- García, P. F., Vallejo, G., Livacic-Rojas, P., Herrero, J., & Cuesta, M. (2010). Comparative robustness of six tests in repeated measures designs with specified departures from sphericity. *Quality & Quantity*, 44(2), 289-301.
- Glass, G.V., Peckham, P.D. & Sanders, J.R. (1972): *Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance*, Review of Educational Research, 42(3), pp 237-288
- Gonzalez, Richard (2009): *Contrasts and Post Hoc tests (Lecture Notes)* , University of Michigan, Ann Arbor, <http://www-personal.umich.edu/~gonzo/coursenotes/file3.pdf>
- Götzte-Baltes, Bernhard (2016): *Generalisierte lineare Modelle und GEE -Modelle in SPSS Statistics*, Universität Trier, [https://www.uni-trier.de/fileadmin/urt/doku/gzlm\\_gee/gzlm\\_gee.pdf](https://www.uni-trier.de/fileadmin/urt/doku/gzlm_gee/gzlm_gee.pdf)
- Hahn, S., Konietzschke, F. and Salmaso, L. (2013): *A comparison of efficient permutation tests for unbalanced ANOVA in two by two designs - and their behavior under heteroscedasticity*, arXiv.org Cornell University, <http://arxiv.org/pdf/1309.7781.pdf>
- Hallin, M. and Paindaveine, D. (2002): Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Annals of Statistics*, 30, pp 1103–1133.
- Hallin, Marc & Paindaveine, Davy (2006): *Optimal Rank-Based Tests for Sphericity*, The Annals of Statistics, Vol. 34, No. 6, pp 2707–2756
- Hallin, Marc & Paindaveine, Davy (2009): *Optimal tests for homogeneity of covariance, scale, and shape*. *Journal of Multivariate Analysis* ,100, pp 422-444.
- Happ, M., Zimmermann, G., Brunner, E. & Brunner, E. & Bathke, A.C. (2020): Pseudo-Ranks: How to Calculate Them Efficiently in R. *Journal of Statistical Software*, October 2020, Volume 95, doi: 10.18637/jss.v095.c01.
- Harwell, M.R. & Serlin, R.C. (1994): A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, 17, pp 35-49.
- Harwell, M.R. & Serlin, R.C. (1995): Empirical Study of the Type I Error Rates of Five Multivariate Tests for the Single-Factor Repeated Measures Mode. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA



- Hettmansperger, Thomas P. & McKean, Joseph W. (2011): *Robust Nonparametric Statistical Methods*, CRC Press
- Hora, Stephen C. & Conover, W. J. (1984): *The F Statistic in the Two-Way Layout with Rank-Score Transformed Data*, Journal of the American Statistical Association, Vol. 79, No. 387, pp. 668-673
- Huang, M.L. (2007): *A Quantile-Score Test for Experimental Design*, Applied Mathematical Sciences, Vol. 1, No 11, pp 507-516.
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, 65(332), pp 1582-1589.
- Huynh, H. (1978): *Some approximate tests for repeated measurement designs*, Psychometrika 43, pp 161-175.
- Iman, R.L. & Davenport, J.M. (1976): *New approximations to the exact distribution of the Kruskal-Wallis test statistic*, Comm, Statist, A5, pp 1335-1348
- Ito, P.K. (1980): *Robustness of Anova and Manova Test Procedures* in Handbook of Statistics, Vol. 1, (P.R.Krishnaiah,ed.)
- James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38, pp 324-329.
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993): *Testing Repeated Measures Hypotheses When Covariance Matrices are Heterogeneous*. *Journal of Educational and Behavioral Statistics*, Vol. 18, no. 4, pp 305-319
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995): Robust and powerful nonorthogonal analyses. *Psychometrika*, 60, 395-418.
- Kloke, John D. & McKean, Joseph W. (2012): *Rfit : Rank-based estimation for linear models*, [http://journal.r-project.org/archive/2012-2/RJournal\\_2012-2\\_Kloke+McKean.pdf](http://journal.r-project.org/archive/2012-2/RJournal_2012-2_Kloke+McKean.pdf)
- Koch, Gary (1969): *Some aspects of the statistical analysis of split plot experiments in completely randomized layouts*. Journal of the American Statistical Association, Vol. 64, No. 326, pp. 485-505
- Koch, G. G. (1970): The use of non-parametric methods in the statistical analysis of a complex split plot experiment. *Biometrics*, pp 105-128.
- Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., and Lehnen, R.G. (1977): *A general methodology for the analysis of experiments with repeated measurement of categorical data*. *Biometrics*, 33, 133-158.
- Koch, G.G., Amara, I.A., Stokes, M.E. and Gillings, D.B. (1980): Some Views on Parametric and Non-Parametric Analysis for Repeated Measurements and Selected Bibliography. *International Statistical Review*, Vol. 48, No. 3, pp. 249-265
- Kowalchuk, Rhonda K. , Keselman, H. J. & Algina, James (2003): *Repeated Measures Interaction Test with Aligned Ranks*, Multivariate Behavioral Research, Volume 38, Issue 4

- Larntz, Kinley & Perlman, Michel D. (1985): A simple Test for the Equality of Correlation Matrices. University of Washington, Technical Report No. 63.
- Lecoutre, Bruno (1991): A Correction for the  $\epsilon$  Approximate Test in Repeated Measures Designs With Two or More Independent Groups. *Journal of Educational Statistics*, Vol. 16, No. 4, pp. 371-372
- Lei, X., Holt, J., Beasley, T.M. (2004): Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances. *Journal of Modern Applied Statistical Methods*, Vol 3, No 2, pp. 462-475
- Lemmer, H. H., & Stoker, D. J. (1967). *A distribution-free analysis of variance for the two-way classification*. *South African Statistical Journal*, 1, 67–74
- Leys, C., Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*.
- Li, Peng & Redden, David T. (2015): Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, <https://doi.org/10.1186/s12874-015-0026-x>
- Li, Zeng & Yao, Jianfeng (2016): Testing the sphericity of a covariance matrix when the dimension is much larger than the sample size. *Electronic Journal of Statistics*, Vol. 10, pp 2973–3010
- Lienert, G.A. (1987): *Verteilungsfreie Methoden in der Biostatistik* - Band 1 und 2, (der „Klassiker“)
- Lindman, H. R. (1974): *Analysis of variance in complex experimental designs*. San Francisco: W. H. Freeman & Co.
- Lix L.M., Keselman J.C. and Keselman, H.J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research*, Vol. 66, No. 4, pp. 579-619.
- Lunney, G.H. (1970): *Using Analysis of Variance with a dichotomous dependent variable: an empirical study*. *Journal of Educational Measurement* Volume 7, Issue 4
- Lüpsen, Haiko (2014): *Multiple Mittelwertvergleiche - parametrisch und nichtparametrisch - sowie alpha-Adjustierungen mit praktischen Anwendungen mit R und SPSS*, Universität zu Köln, URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/mult-comp.pdf>
- Lüpsen, Haiko (2016a): *The Aligned Rank Transform and discrete Variables - a Warning*, erschienen in: *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/03610918.2016.1217014 URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/ART-discrete.pdf>
- Lüpsen, Haiko (2016b): *The lognormal distribution and nonparametric anovas - a dangerous alliance*, Universität zu Köln, URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/lognormal-anova.pdf>

- Lüpsen, 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  
URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-1.pdf>
- Lüpsen, Haiko (2021): *Anova with binary variables - The F test and some Alternatives* to be published in *Communications in Statistics - Simulation and Computation*
- Lüpsen, Haiko (2020a): *Checking the Homogeneity of Covariance Matrices: some practical aspects*.  
URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/Checking.Homogeneity.pdf>
- Lüpsen, Haiko (2020b): *Some rank based ANOVA procedures for analyzing data from split-plot designs*. URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/algorithm.pdf>.
- Lüpsen, Haiko (2020c): *Anmerkungen zum Testen der Sphärizität*.  
URL: <http://www.uni-koeln.de/~luepsen/statistik/texte/Sphaerizitaet.pdf>
- Mansouri, H. & Chang, G. H. (1995): *A Comparative Study of Some Rank Tests for Interaction*, *Computational Statistics and Data Analysis*, 19, 85-96
- Mansouri, H., Paige, R. L. & Surles, J. G. (2004): *Aligned Rank Transform Techniques for Analysis of Variance and Multiple Comparisons*, *Communications in Statistics - Theory and Methods*, Volume 33, Issue 9
- Marascuilo, Leonard A. & McSweeney, Maryellen (1977): *Nonparametric and distribution-free methods for the social sciences*, Brooks/Cole Pub. Co.
- Mendes, Mehmet & Yiğit, Soner (2013): *Type I error and test power of different tests for testing interaction effects in factorial experiments*, *Statistica Neerlandica*, Vol 67 Issue 1, pp 1-26
- Mendoza, Jorge L. (1980): *A significance test for multisample sphericity*, *Psychometrika*, Vol 45, No 4, pp 495-498.
- Meyer, Bertolt (2008): *Obtaining the same ANOVA results in R as in SPSS - the difficulties with Type II and Type III sums of squares*,  
<http://myowelt.blogspot.de/2008/05/obtaining-same-anova-results-in-r-as-in.html>
- Muirhead, R.J. & Waternaud, C.M. (1980): *Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations*. *Biometrika*, 67, No 1, pp. 31-43.
- Munzel, Ullrich & Brunner, Edgar (2000): *Nonparametric methods in multivariate factorial designs*, *Journal of Statistical Planning and Inference*, Volume 88, Issue 1, Pages 117–132
- Nagao, Hisao (1973): *On Some Test Criteria for Covariance Matrix*. *The Annals of Statistics*, Vol. 1, No. 4, pp 700-709.
- Noguchi, K., Gel, Y.R., Brunner, E., Konietzke, F. (2012): *nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments*, *Journal of Statistical Software*, Volume 50, Issue 12.

- Nordhausen, K. and Oja, H. (2011): Multivariate L1 Methods: The Package MNM, *Journal of Statistical Software*, 43, pp 1-28.
- O'Brien, Peter C. (1992): Robust Procedures for Testing Equality of Covariance Matrices. *Biometrics*, Vol. 48, No. 3 (Sep., 1992), pp. 819-827
- Olejnik, Stephen F. & Algina, James (1985): *A Review of Nonparametric Alternatives To Analysis of Covariance*, Evaluation Review 9: p 51-83
- Olson, Chester L. (1976): On Choosing a Test Statistic in Multivariate Analysis of Variance. *Psychological Bulletin*, Vol 83, No 4, pp 579-586
- Online Statistics Education: <http://onlinestatbook.com/2/transformations/box-cox.html>
- Osborne, Jason W. (2008): *Best Practices in Quantitative Methods*, Sage Publications
- Peterson, Kathleen (2002): *Six Modifications Of The Aligned Rank Transform Test For Interaction*, Journal Of Modern Applied Statistical Methods Winter 2002, Vol. 1, No. 1, pp 100-109
- Puri, M.L. & Sen, P.K. (1971): *Nonparametric Methods in Multivariate Analysis*, Wiley, NY.
- Puri, M.L. & Sen, P.K. (1985): *Nonparametric Methods in General Linear Models*, Wiley, NY.
- Richter, S. J. and Payton, M. (2003). *An Improvement to the Aligned Rank Statistic for Two-Factor Analysis of Variance*. Joint Statistical Meeting of the American Statistical Association, Journal of Applied Statistical Science, 14(3/4), pp 225-236.
- Salazar-Alvarez, M.I. , Tercero-Gomez, V.G., Temblador-Pérez, M., Cordero-Franco, A.E., Conover, W.J. (2014): *Nonparametric analysis of interactions: a review and gap analysis*, Proceedings of the 2014 Industrial and Systems Engineering Research Conference, Y. Guan and H. Liao (eds.)
- Sawilowsky, S., Blair, R. C., & Higgins, J. J. (1989): *An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA*, Journal of Educational Statistics 14 (3): 255–267
- Sawilowsky, S. (1990): *Nonparametric tests of interaction in experimental design*. Review of Educational Research 60: 91–126.
- Scholer, Falk (2016): *ANOVA (and R)*, <http://goanna.cs.rmit.edu.au/~fscholer/anova.php>
- Scheirer, J., Ray, W.S. , Hare, N. (1976): *The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs*. *Biometrics*. 32(2). International Biometric Society, pp 429–434
- Schneider, P. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: Providing an alternative to ANOVA under variance heterogeneity. *The Journal of Experimental Education*, 65, pp 271-286.
- Sheskin, David J. (2004): *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall
- Shirley, E.A. (1981): *A distribution-free method for analysis of covariance based on ranked data*, Journal of Applied Statistics 30: pp 158-162.

- Sirkiä, S., Taskinen, S., Nevalainen, J., Oja, H. (2007): Multivariate nonparametrical methods based on spatial signs and ranks: The R package SpatialNP, *Journal of Statistical Software*
- Smith, C.E. & Cribbie, R. (2014): Factorial ANOVA with unbalanced data: A fresh look at the types of sums of squares. *Journal of Data Science* 12, pp 385-404.
- Statsoft: <https://www.statsoft.com/Textbook/ANOVA-MANOVA>
- Stiger, R.T., Kosinski, A.S., Barnhart, H.X. & Kleinbaum, D.G. (1998) Anova for repeated ordinal data with small sample size? A comparison of anova, manova, wls and gee methods by simulation. *Communications in Statistics - Simulation and Computation*, 27:2, pp 357-375.
- Tandon, P.K. & Moeschberger, M.L. (1989) Comparison of Nonparametric and Parametric Methods in Repeated Measures Designs - A Simulation Study, *Communications in Statistics - Simulation and Computation*, 18:2, pp 777-792.
- Thomas, J.R., Nelson, J.K. and Thomas, T.T. (1999). A Generalized Rank-Order Method for Nonparametric Analysis of Data from Exercise Science: A Tutorial. *Research Quarterly for Exercise and Sport, Physical Education, Recreation and Dance*, Vol. 70, No. 1, pp 11-23.
- Tian, Tian & Wilcox, Rand R. (2007): A Comparison of Two Rank Tests for Repeated Measure Designs. *Journal of Modern Applied Statistical Methods*, Vol. 6, No. 1, pp 331-335.
- Tomarken, A.J. and Serlin, R.C. (1986). Comparison of ANOVA Alternatives Under Variance Heterogeneity and Specific Noncentral Structures. *Psychological Bulletin*, Vol. 99, No 1, pp 90-99.
- Toothaker, Larry E. & De Newman (1994): *Nonparametric Competitors to the Two-Way ANOVA*, *Journal of Educational and Behavioral Statistics*, Vol. 19, No. 3, pp. 237-273
- Tuerlinckx, F., Rijmen, F., Verbeke, G. & De Boeck, P. (2006): Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59, pp 225–255.
- Vallejo, G, Fernandez, M.P. Herrero, F.J., Connejo, N.M. (2004): *Alternative Procedures for testing fixed effects in repeated measures designs when assumptions are violated*. *Psicothema* 2004. Vol. 16, no 3, pp. 498-508.
- Vallejo, G., Ato, M., Fernandez, M.P. (2010). A robust approach for analyzing unbalanced factorial designs with fixed levels. *Behavior Research Methods*, 42 (2), 607-617
- Vargha, András & Delaney, Harold D. (1998): *The Kruskal-Wallis Test and Stochastic Homogeneity*, *Journal of Education and Behavioral Statistics*, vol. 23 no. 2, pp 170-192
- Wang, M., Kong, L., Zheng, L. & Zhang, L. (2016): Covariance estimators for Generalized Estimating Equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine*, 35(10), pp 1706–1721.
- Wang, Q. & Yao, J. (2013): On the sphericity test with large-dimensional observations. *Electronic Journal of Statistics*, Vol. 7, pp 2164–2192.

- Weyer, Veronika (2008): *Modellwahl für die Analyse longitudinaler Daten einer Forschungsstudie des visuellen Systems*, Carl von Ossietzky Universität Oldenburg, [https://www.statistik.tu-dortmund.de/fileadmin/user\\_upload/Lehrstuehle/Ingenieur/Mueller/Diplomarbeiten/Weyer.pdf](https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/Ingenieur/Mueller/Diplomarbeiten/Weyer.pdf)
- Wikipedia: [http://en.wikipedia.org/wiki/ANOVA\\_on\\_ranks](http://en.wikipedia.org/wiki/ANOVA_on_ranks)
- Wikipedia: <https://de.wikipedia.org/wiki/Zweistichproben-t-Test>
- Wikipedia: [https://en.wikipedia.org/wiki/Fishers\\_method](https://en.wikipedia.org/wiki/Fishers_method)
- Wikipedia: [http://en.wikipedia.org/wiki/Logistic\\_regression](http://en.wikipedia.org/wiki/Logistic_regression)
- Wikipedia: [http://en.wikipedia.org/wiki/Van\\_der\\_Waerden\\_test](http://en.wikipedia.org/wiki/Van_der_Waerden_test)
- Wikipedia: [http://de.wikipedia.org/wiki/Friedman-Test\\_\(Statistik\)](http://de.wikipedia.org/wiki/Friedman-Test_(Statistik))
- Wilcox, Rand R. (1989): Comparing the Variances of dependent Groups. *Psychometrika*, vol 54, No. 2, pp 305--315
- Wilcox, Rand R. (2003): *Applying Contemporary Statistical Techniques*, Elsevier
- Wilcox, Rand R. (2012): *Introduction to Robust Estimation and Hypothesis Testing*, Elsevier
- Wilcox, Rand R. (2013): *New Statistical Procedures for the Social Sciences: Modern Solutions To Basic Problems*, Psychology Press, Lawrence Erlbaum Assoc
- Wilcox, Rand R. (2005): *Introduction to robust estimation and hypothesis testing*, Burlington MA; Elsevier
- Winer, B.J. et al. (1991): *Statistical Principles in Experimental Design*, pp 1028 ff bzw. pp 1024 ff)
- Wobbrock, J. O., Findlater, L., Gergle, D. & Higgins, J. (2011): *The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures*, Computer Human Interaction - CHI , pp. 143-146
- Wobbrock, J. O et al. (2011): ARTool: <http://depts.washington.edu/aimgroup/proj/art/>
- Zhang, Shuqiang (1998): *Fourteen Homogeneity of Variance Tests: When and how to use them*, Annual Meeting of the American Educational Research Association, San Diego
- Ziegler, A., Kastner, Ch., Blettner, M. (1998): The Generalised Estimating Equations: An Annotated Bibliography. *Biometrical Journal* 40 (2), pp 115-139.
- Zimmerman, D.W. (1998). Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education*, Vol. 67, No. 1 (Fall, 1998), pp. 55-68.
- Zimmerman, D.W. (2004). Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests. *Psicológica*, 25, pp 103-133.
- Zou, C., Peng, L., Feng, L., Wang, Z. (2014): Multivariate sign-based high-dimensional tests for sphericity. *Biometrika*, 101, No 1, pp. 229–236.