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

Version 6.0
(30.05.2026)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@Uni-Koeln.de

Universität zu Köln

Vorwort

Aktuelle Version

Nach nunmehr 2 Jahren ist dieses Skript 2026 endlich wieder überarbeitet worden. Manches hatte sich inzwischen überholt und wurde entfernt. Neue Methoden, insbesondere solche, die in R verfügbar gemacht worden waren, zum Teil auch durch eigene R-Funktionen, wurden neu aufgenommen, insbesondere im Bereich heterogene Varianzen, sowohl Tests zur Überprüfung, aber auch neue robuste Varianzanalysen. Aber, und das ist das Wichtigste: Ich hatte mir die Zeit genommen, den ganzen Text einmal gründlich durchzulesen, wobei mir leider sehr viele Fehler aufgefallen waren, und zu korrigieren.

Entstehung

In den letzten Jahren meiner Tätigkeit an der Universität Köln hatte ich mehrfach Kurse zum Thema „nichtparametrische Methoden mit SPSS“ bzw. Kurse zur Programmiersprache S und dem System R sowohl am RRZK als auch an anderen Einrichtungen gehalten. Dort hatte sich gezeigt, dass ein großes Interesse an nichtparametrischen statistischen Verfahren besteht, insbesondere im Bereich Varianzanalyse. 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 nichtparamtrische 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 Teil der Ergebnisse wurde inzwischen veröffentlicht, u.a. in der Zeitschrift *Communications in Statistics - Simulation and Computation*. Weitere Artikel mit neuen Ergebnissen sind derzeit in Arbeit und sollen in Kürze ebenfalls veröffentlicht werden.

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 α-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 Bedienungskomfort 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` 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` 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 6.0 (30.3.2026): Komplette Überarbeitung des Textes, Korrektur von Fehlern, Ergänzung um robuste Methoden.

Version 5.0 (15.2.2024): Komplette Überarbeitung des Textes, Korrektur von Fehlern, Ergänzung um semi-parametrische Methoden, u.a. im Bereich GLM (general linear models).

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 Spehrizität, 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 Kriteriumsveriablen, 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

Datensätze

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 3
1. 1. 6 Pairing 3
1. 1. 7 Abhängige und unabhängige Stichproben 3
1. 2 Methoden für metrische und ordinale Merkmale 3
1. 3 Methoden für dichotome Merkmale 4
1. 4 Methoden für nominale Merkmale 4

2. Nichtparametrische Varianzanalysen - Übersicht der Methoden 5

2. 1 Kruskal-Wallis und Friedman (KWF) 6
2. 2 Rank transform Tests (RT) 7
2. 3 Inverse normal transform (INT) 8
2. 4 Aligned rank transform (ART und ART+INT) 8
2. 5 Puri & Sen-Tests (PS), L statistic 10
2. 6 van der Waerden (vdWS) 11
2. 7 Akritas, Arnold & Brunner ATS Tests 12
2. 8 Verteilungs-robuste Verfahren 13
2. 8. 1 Mair & Wilcox 14
2. 8. 2 Ananda und Weerahandi 14
2. 9 Varianzanalysen für heterogene Varianzen 14
2. 9. 1 Welch und Fligner-Policello 14
2. 9. 2 James 2nd order und Alexander & Govern 15
2. 9. 3 Welch & James (WJ) 15
2. 9. 4 Brown & Forsythe (BF) 15
2. 9. 5 Brunner, Dette und Munk (BDM) 15
2. 9. 6 Ananda und Weerahandi 16
2. 9. 7 Box-Korrektur 16
2. 10 Weitere Varianzanalysen für unabhängige Stichproben 16
2. 11 Weitere Varianzanalysen für abhängige Stichproben 16
2. 11. 1 Quade 16
2. 11. 2 Skillings & Mack 17
2. 11. 3 Multivariate Tests: Hotelling-Lawley, Pillai und Wilks 17
2. 11. 4 Agresti & Pendergast (AP) und Akritas & Arnold 17
2. 11. 5 Spatial Signs und Spatial Ranks Methoden 18
2. 12 Weitere Varianzanalysen für gemischte Versuchspläne 18
2. 12. 1 Koch 18
2. 12. 2 Beasley & Zumbo 18
2. 13 Varianzanalysen für heterogene Varianzen bei abhängigen Stichproben und gemischten Versuchsplänen 19

2. 13. 1 Welch & James (WJ) 19
2. 13. 2 Brown & Forsythe (mBF) 19
2. 13. 3 Adjustierungen der Freiheitsgrade: ε, GA und IGA 19
2. 14 Logistische Regression 20
2. 15 GEE und GLMM 20
2. 16 Alternative Rangberechnungen 22
2. 16. 1 Pseudo-Ränge 22
2. 16. 2 Spatial Ranks - multivariate Ränge 23
2. 17 Voraussetzungen 23
2. 17. 1 Versuchspläne ohne Messwiederholungen 24
2. 17. 2 Versuchspläne mit Messwiederholungen 24
2. 18 Vergleiche 25
2. 18. 1 Versuchspläne ohne Messwiederholungen 25
2. 18. 2 Versuchspläne mit Messwiederholungen 27
2. 19 Entscheidungshilfen zur Auswahl 29
2. 19. 1 Warnungen 29
2. 19. 2 Versuchspläne ohne Messwiederholungen 30
2. 19. 3 Versuchspläne mit Messwiederholungen 30
2. 20 Methoden zur Prüfung der Voraussetzungen 31
2. 20. 1 Normalverteilung 31
2. 20. 2 Varianzhomogenität bei unabhängigen Stichproben 33
2. 20. 3 Varianzhomogenität bei abhängigen Stichproben 35
2. 20. 4 Spherizität 35
2. 20. 5 Homogenität der Kovarianzmatrizen 36
2. 20. 6 Homogenität der Korrelationsmatrizen 37
2. 20. 7 Eine Warnung 38
2. 20. 8 Entscheidungen - die Qual der Wahl 38

3. Funktionen zur Varianzanalyse in R und SPSS 39

3. 1 Funktionen in R 39
3. 2 Funktionen in SPSS 41
3. 3 Fehler bei der Rangberechnung 41
3. 4 Fehlende Werte 42
3. 5 Beschränkungen 43

4. Unabhängige Stichproben 45

4. 1 Voraussetzungen der parametrischen Varianzanalyse 46
4. 2 Die 1-faktorielle Varianzanalyse 51
4. 2. 1 Kruskal-Wallis-Test 51
4. 2. 2 Varianzanalysen für inhomogene Varianzen 52
4. 2. 3 Verfahren für nichtnormalverteilte Variablen 54
4. 2. 4 Weitere Verfahren 54
4. 3 Die 2-faktorielle Varianzanalyse 54
4. 3. 1 Anmerkungen zur 2-faktoriellen Varianzanalyse 55
4. 3. 1. 1 Balancierte und nichtbalancierte Versuchspläne 55
4. 3. 1. 2 Die Interaktion 55
4. 3. 1. 3 Reduzierung des statistischen Fehlers 57
4. 3. 1. 4 Interpretation der Ergebnisse 57
4. 3. 2 Das parametrische Verfahren und Prüfung der Voraussetzungen 58

4. 3. 3 Varianzanalysen für inhomogene Varianzen 65
4. 3. 3. 1 Verfahren von Box, Brown & Forsythe sowie Welch & James 66
4. 3. 3. 2 BDM-Tests (Brunner, Dette und Munk) 67
4. 3. 3. 3 Verfahren von Ananda und Weerahandi 68
4. 3. 3. 4 Variablentransformationen 68
4. 3. 4 Verteilungs-robuste Varianzanalysen 69
4. 3. 4. 1 Robuste Mittelwerte 69
4. 3. 4. 2 Robuste Residuen 70
4. 3. 5 Rank transform-Tests (RT) 71
4. 3. 6 Puri & Sen-Test (Verallgemeinerte Kruskal-Wallis-Analysen) 73
4. 3. 7 Aligned rank transform (ART und ART+INT) 76
4. 3. 8 normal scores- (INT-) und van der Waerden-Tests 81
4. 3. 9 ATS-Tests von Akritas, Arnold & Brunner 83
4. 3. 10 Parametrisches Verfahren mit anderen Residuen-Verteilungen 84
4. 4 Nichtparametrische Verfahren zur mehrfaktoriellen Varianzanalyse 86

5. Abhängige Stichproben - Messwiederholungen 87

5. 1 Datenstruktur 89
5. 1. 1 Besonderheiten bei R und SPSS 89
5. 1. 2 Umstrukturierungen in R 91
5. 2 Voraussetzungen der parametrischen Varianzanalyse 95
5. 3 Die 1-faktorielle Varianzanalyse 99
5. 3. 1 Parametrischer Test und Prüfung der Voraussetzung 99
5. 3. 2 Friedman-Test 105
5. 3. 3 Rank transform (RT) und normal scores (INT) 106
5. 3. 4 Puri & Sen-Test 108
5. 3. 5 van der Waerden-Test (vdWS) 110
5. 3. 6 ATS-Tests von Akritas, Arnold & Brunner 113
5. 3. 7 Quade-Test 114
5. 3. 8 Skillings-Mack-Test 114
5. 3. 9 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Agresti-Pendergast 114
5. 3. 10 Multivariate Tests: Spatial Signs und Spatial Ranks Methoden 117
5. 3. 11 Verwendung robuster Mittelwerte 119
5. 4 Die 2-faktorielle Varianzanalyse 120
5. 4. 1 Das parametrische Verfahren und Prüfung der Voraussetzungen 120
5. 4. 2 Rank transform-Tests (RT) und normal scores-Tests (INT) 124
5. 4. 3 Puri & Sen-Test 127
5. 4. 4 Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF) und van der Waerden-Test 131
5. 4. 5 Aligned rank transform (ART und ART+INT) 134
5. 4. 6 ATS-Tests von Akritas, Arnold & Brunner 139
5. 4. 7 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und Akritas & Arnold 141

6. Gemischte Versuchspläne 143

6. 1 Voraussetzungen der parametrischen Varianzanalyse 143
6. 2 Parametrische Varianzanalyse und Prüfung der Voraussetzungen 149
6. 3 Rank transform-Tests (RT) und normal scores-Tests (INT) 161
6. 4 Puri & Sen-Test 165
6. 4. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor 166

6. 4. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren 167
6. 5 Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF) 170
6. 5. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor 170
6. 5. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren 173
6. 5. 3 Zwei Gruppierungs- und ein Messwiederholungsfaktoren 176
6. 6 van der Waerden-Tests (vdWS) 176
6. 6. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor 177
6. 6. 2 Zwei Gruppierungs- und ein Messwiederholungsfaktor 180
6. 6. 3 Ein Gruppierungs- und zwei Messwiederholungsfaktoren 183
6. 7 Aligned rank transform (ART und ART+INT) 185
6. 7. 1 Ein Gruppierungs- und ein Messwiederholungsfaktor 186
6. 7. 2 Ein Gruppierungs- und zwei Messwiederholungsfaktoren 189
6. 7. 3 Zwei Gruppierungs- und ein Messwiederholungsfaktor 193
6. 8 ATS-Tests von Akritas, Arnold & Brunner 198
6. 9 Robuste Mittelwerte 200
6. 10 Verfahren ohne Spherizitäts-Voraussetzungen 201
6. 10. 1 Multivariate Tests: Hotelling-Lawley, Wilks, Pillai und nichtparametrisch 201
6. 10. 2 Multivariate Analysen: Spatial Signs und Spatial Ranks Methoden 204
6. 10. 3 Welch & James 206
6. 10. 4 Koch 208
6. 10. 5 GEE 208
6. 10. 6 GLMM 213
6. 10. 7 GA- und IGA-Approximationen von Huynh 216
6. 10. 8 modifizierter Brown-Forsythe-Test (mBF) 217
6. 10. 9 Versuchspläne mit 2 Messwiederholungen 218
6. 11 Alternative parametrische Modelle 220
6. 11. 1 Andere Residuen-Modelle 220
6. 11. 2 Andere Kovarianz-Modelle 221

7. Analysen für dichotome Merkmale 224

7. 1 Anwendung der Verfahren für metrische Merkmale 225
7. 1. 1 Unabhängige Stichproben 226
7. 1. 2 Gemischte Versuchspläne 227
7. 2 Anwendung der Verfahren für ordinale Merkmale 229

8. Logistische Regression 230

8. 1 dichotome abhängige Variablen 230
8. 2 ordinale abhängige Variablen 233
8. 3 dichotome abhängige Variablen und Messwiederholungen 238
8. 4 ordinale abhängige Variablen und Messwiederholungen 242

9. Mittelwertvergleiche, Kontraste und Kodierungen 244

9. 1 Grundlagen 244
9. 2 Standard-Kontraste 246
9. 3 Auswahl der Kontraste 248
9. 4 nichtparametrische Kontraste für die RT-, ART- und Puri & Sen-Verfahren 249
9. 5 universelles Verfahren für Kontraste 253
9. 6 Kontraste bei logistischen Regressionen 254
9. 7 Kontraste für Messwiederholungen und Interaktionen 254

9. 8 Zusammenfassen von Kontrasten 258

10. Simple effects - einfache Effekte 260

10. 1 Unabhängige Stichproben 260
10. 2 Gemischte Versuchspläne 263

11. Beipiele mit problematischen Datensätzen 267

11. 1 Extrem heterogene Varianzen 267
11. 2 Lognormal verteilte abhängige Variable 269
11. 3 Negatives Pairing 271
11. 4 Gemischter Versuchsplan mit Varianzheterogenitäten 274
11. 5 Gemischter Versuchsplan: Prüfung der Voraussetzungen 278

Anhang 285

1. Umstrukturieren von Messwiederholungen in SPSS 285

1. 1 Umstrukturieren von Messwiederholungen in Fälle 285
1. 1. 1 ein Faktor und eine Analyse-Variable 285
1. 1. 2 mehrere Faktoren und eine Analyse-Variablen 288
1. 1. 3 ein Faktor und mehrere Analyse-Variablen 291
1. 2 Umstrukturieren von Fälle in Messwiederholungen 295

2. Spezielle robuste F-Tests und andere Statistiken 298

2. 1 Box-Korrektur für heterogene Varianzen 298
2. 2 Brown-Forsythe F-Test für inhomogene Varianzen 298
2. 3 Box-Andersen F-Test für nichtnormalverteilte Variablen 299
2. 4 Box-Cox-Transformationen 299
2. 5 Fishers combined probability test 299
2. 6 Levene-Test auf Gleichheit von Kovarianzmatrizen 300
2. 7 Wilcox-Test auf Gleichheit von Varianzen bei Messwiederholungen 300

3. anova.lib: R-Funktionen 301

3. 1 box.f: Box-F-Test für inhomogene Varianzen 301
3. 2 bf.f: Brown & Forsythe-F-Test für inhomogene Varianzen 301
3. 3 mbf.f: modified Brown & Forsythe-F-test für inhomogene Varianzen in gemischten Versuchsplänen (Split-Plot Designs) 302
3. 4 box.andersen.f: F-Test für nichtnormalverteilte Variablen 302
3. 5 check.covar: Test auf Homogenität von Kovarianzmatrizen 303
3. 6 check.corr: Test auf Homogenität von Korrelationsmatrizen 304
3. 7 check.var: Test auf Homogenität von Varianzen bei abhängigen Stichproben 305
3. 8 check.sphere: Test auf Spherizität 306
3. 9 ats.2 und ats.3: 2- bzw. 3-faktorielle Varianzanalyse 307
3. 10 np.anova: nichtparametrische Varianzanalyse mittels des KWF-Verfahrens und der von Puri & Sen und van der Waerden 307
3. 11 art1.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Gruppierungsfaktoren) 308
3. 12 art2.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (nur Messwiederholungsfaktoren) 308

3. 13 art3.anova: nichtparametrische Varianzanalyse nach dem ART-Verfahren (für gemischte Versuchspläne) 309
3. 14 wj.anova: Welch-James-Varianzanalyse für heterogene Varianzen (nur für Gruppierungsfaktoren) 309
3. 15 wj.spanova: Welch-James-Varianzanalyse für heterogene Varianzen (für gemischte Versuchspläne) 309
3. 16 koch.anova: nichtparametrische Varianzanalyse für gemischte Versuchspläne nach dem Verfahren von G.Koch 310
3. 17 ap.anova: nichtparametrische Varianzanalyse für Messwiederholungen und Split-Plot Versuchspläne von Agresti & Pendergast 310
3. 18 iga und iga.anova: general approximation test (GA) und improved general approximation test (IGA) von H.Huynh 311
3. 19 simple.effects: parametrische Analyse von simple effects 312
3. 20 gee.anova: Anova-like tests for GEE and GLMM models 312
3. 21 rob.anova: 1- / 2-faktorielle robuste Varianzanalysen 313

Literaturhinweise 314

Datensätze

Beispieldaten 1 (mydata1): 45
Beispieldaten 2 (mydata2): 45
Beispieldaten 3 (mydata3): 45
Beispieldaten 4 (winer518): 87
Beispieldaten 5 (mydata5): 87
Beispieldaten 6 (winer568): 88
Beispieldaten 7 (irish): 224
Beispieldaten 8 (koch): 225
Beispieldaten (industrial waste): 267
Beispieldaten (lognormal): 269
Beispieldaten 11: 271
Beispieldaten 12: 274
Beispieldaten 13: 278
Beispieldaten 14 (mydata14): 88

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 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 μ und Standardabweichung σ 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. Darüberhinaus gibt es noch *semiparametrische* Verfahren, Mischungen aus parametrischen und nichtparametrischen Verfahren, wenn etwa eine bestimmte Verteilung zugrunde liegt, die untersuchte Größe aber transformiert wird, z.B. um Heterogenitäten zu kompensieren.

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 (für das parametrische Verfahren) 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 u.a. Sawilowsky (1990) in seiner Zusammenstellung auch diesen allgemeinen Punkt betrachtet. Demnach schneiden die meisten (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.

Es gibt aber auch inzwischen (30 Jahre später) wieder andere Sichtweisen, die darauf aufmerksam machen, dass man vielzu wenig über die exakten Voraussetzungen der nichtparametrischen Verfahren weiß und mit deren Anwendung andere Risiken in Kauf nehmen muss. Siehe z.B. die Veröffentlichung mit dem viel sagenden Titel „Violating the normality assumption may be the lesser of two evils“ von Kief & Forstmeier (2021.

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 α vor. Dieses ist der *Fehler 1. Art*. Er bedeutet, dass z.B. bei einer Vorgabe α=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 wird mit β bezeichnet, und 1-β heißt die *Teststärke* oder *Power*. β ist zunächst unbekannt, kann aber für zahlreiche Tests bei Vorgabe einiger Daten, wie z.B. n oder der Effektgröße, errechnet werden. Allerdings gilt: $\beta \to 0$ für $n \to \infty$, d.h. je größer das n, desto größer die Teststärke (Gesetz der großen Zahlen).

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 Grenzverhältnis 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 β, bei einem beliebigen (aber festen) α und unter der Annahme, dass die Voraussetzungen des parametrischen Tests erfüllt sind. (Dieser Grenzwert ist allerdings abhängig von α.) In der Praxis ist die Effizienz eines Tests bekannt, woraus sich der Mehrbedarf n_B-n_A ermitteln lässt. So bedeutet z.B. die asymptotische Effizienz eines nichtparametrischen Tests A von 95% oder 67 % gegenüber einem parametrischen Test B, 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-β.

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 α ist. D.h. wenn z.B. bei einem α=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 α überschreiten kann, d.h. wenn z.B. bei einem α=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, denn 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 α und n die Wahrcheinlichkeit β 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 α eingehalten, d.h. bei z.B. α=0.05 sind auch nur 5 von 100 Ergebnissen zufällig signifikant. Zum anderen verändert sich die Wahrscheinlichkeit für einen Fehler 2. Art β 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 s_i^2, aber auch Korrelationen. Haben die größeren Stichproben auch die größeren Parameterwerte, z.B. Varianzen, spricht man von *positivem* oder *direktem Pairing*, haben dagegen die größeren Stichproben die kleineren Parameterwerte, z.B. Varianzen, so pricht man von *negativem* oder *inversem Pairing*.

1. 1. 7 Abhängige und unabhängige Stichproben

Beim Vergleich von Stichproben wird zwischen *abhängigen* und *unabhängigen* unterschieden. Wenn die Werte einer Stichprobe vollkommen unabhängig von denen der anderen Stichproben sind, handelt es sich um unabhängige Stichproben (entspricht einem *Gruppierungsfaktor*). Andernfalls sind es abhängige Stichproben (entspricht einem *Messwiederholungsfaktor*). Üblicherweise werden Stichproben als unabhängig betrachtet, wenn in diesen Merkmale von unterschiedlichen Personen erhoben werden, z.B. einmal Männer und einmal Frauen. Das muss allerdings nicht immer stimmen. Wenn z.B. Eltern, Mutter und Vater unabhängig voneinander, zum Verhalten ihres gemeinsamen Kindes befragt werden, so enthalten zwar beide Gruppen unterschiedliche Personen, die erhobenen Werte sind jedoch nicht unabhängig, da jeweils 2 Personen Aussagen zu demselben Kind machen. Klarer ist die Situation, wenn dieselben Personen mehrmals befragt werden, z.B. vor und nach einer Behandlung. Dann handelt es sich um abhängige Stichproben. Enthält ein Versuchsplan sowohl unabhängige wie auch abhängige Stichproben, spricht man von *gemischten Versuchplänen* oder *split-plot designs*.

1. 2 Methoden für metrische und ordinale Merkmale

Bei den älteren nichtparametrischen Methoden werden die n Werte der Variablen x_i sortiert und in Ränge 1,...,n umgerechnet. Auf diese werden dann die klassischen parametrischen Verfahren angewandt. So ist z.B. der Spearman-Rangkorrelations-koeffizient nichts anderes als der Pearson-Produkt-Moment-Korrelationskoeffizient der Ränge. Lediglich die Signifikanztests sind dann nicht mehr korrekt. Die *korrekten Signifikanzen* errechnen sich für kleine n (etwa <20) mit Mitteln der Kombinatorik, während für große n *asymptotische Signifikanztests* angeboten. Es konnte allerdings gezeigt werden, dass die Anwendung der klassischen parametrischen Verfahren auf die rangtransformierten Daten (ohne Anpassung der Signifikanztests, meistens F-Tests) zu i.a. gültigen Ergebnissen führt (vgl. dazu Conover & Iman, 1981), z.B. das RT-Verfahren. Parallel dazu wurden auch Tests entwickelt, die nicht auf der Normalverteilung basieren. Diese resultieren dann meistens in χ^2-Tests, z.B. die Puri & Sen-Methode.

Die oben erwähnten, primär für stetige Merkmale konzipierten Verfahren setzen voraus, dass eine Variable keine gleichen Werte hat. Indem bei identischen Werten, sog. *Bindungen* (*ties*),

deren Ränge gemittelt werden und dieser Tatsache bei der Berechnung der Statistiken durch die sog. *Bindungskorrekturen* Rechnung getragen wird, werden diese Verfahren auch für ordinale und diskrete Variablen anwendbar, bei denen typischerweise Werte mehrfach vorkommen. In den letzten Jahren sind auch zunehmend Modelle für ordinale Merkmale entwickelt worden, z.B. basierend auf dem Modell der *relativen Effekte* von Akritas, Arnold und Brunner (2013). Fast alle diese Verfahren werden zunehmend kritisch gesehen, weil es zahlreiche Situationen gibt, in denen das α-Risiko nicht konsequent eingehalten wird (vgl. Kapitel 2.19.1 sowie Erläuterungen zu den einzelnen Verfahren).

1. 3 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 auch 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 dichtomome 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. 4 Methoden für nominale Merkmale

Hier sind die polychotomen Merkmale angesprochen, also solche mit drei oder mehr Ausprägungen, ohne ordinales Skalenniveau. 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.

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 α-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(var)/min(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 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). Beide Texte sind allerdings nicht mehr auf dem aktuellen Stand. Denn 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. 2.7) oder “*spatial signs and ranks*“, räumlichen Vorzeichen und Rängen (vgl.

2.16). Hiervon können nur die „wichtigsten“ hier erwähnt werden. Dabei stehen zwar solche im Vordergrund, die sich leicht mit Standardsoftware wie SAS und SPSS durchführen lassen, doch zunehmend sind neuere Verfahren nur in R (und natürlich S-Plus), manchmal auch in SAS verfügbar.

Entscheidend für die Beurteilung eines Verfahrens ist das Verhalten hinsichtlich der Fehler 1. Art (*Irrtumswahrscheinlichkeit* α) und 2. Art (β, aber meistens über die *Power* 1-β). Dabei geht es um die Frage, in wieweit das vorgegebene α eingehalten wird, bzw. in wieweit 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 α-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. Der Anwender muss entscheiden, welches Risiko für ihn das wesentlichere ist.

Sofern nicht anders erläutert seien im Folgenden I die Anzahl der Gruppen eines Gruppierungsfaktors A, J die Anzahl der Gruppen eines weiteren Faktors B, z.B. die Anzahl der Messwiederholungen, n_i bzw. n_{ij} die Anzahl der Merkmalsträger (Versuchspersonen) pro Gruppe bzw. Zelle, sowie x_{ijm} die beobachteten Werte mit m=1,...,n_i , und i=1,...,I sowie j=1,...,J und $N=\sum n_{ij}$ insgesamt. Dann liegt in etwa folgendes parametrisches Modell zugrunde (mehr dazu in den Kapiteln 4.1, 5.2 und 6.1):

$$x_{ijm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijm} \quad (i=1,...,I \;, j=1,...,J \text{ und } m=1,...,n_{ij}) \qquad (2 - 0)$$

wobei α_i bzw. β_j die (festen) Effekte der beiden Faktoren A und B sind, $\alpha\beta_{ij}$ die Interaktionseffekte (Wechselwirkung) von A und B (mehr dazu in Kapitel 4.3.1.2) sowie e_{ijm} die Residuen (Abweichungen vom Modell), an die in der Regel die Voraussetzungen geknüpft werden. Die zu testenden Hypothesen lauten dann

H_A: $\alpha_i = 0$ H_B: $\beta_j = 0$ und H_{AB}: $\alpha\beta_{ij} = 0$ für i=1,..,I und j=1,..,J

was im 1-faktoriellen Fall gleichbedeutend ist mit H_A: μ_1=...=μ_I .

Es sei darauf hingewiesen, dass die Modelle für die nichtparametrischen Tests meistens anders formuliert werden. Sie basieren auf den Verteilungsfunktionen F von x, im 1-faktoriellen Fall: $F_i(x) = G(x-\mu_i)$ (für i=1,..,I), was bedeutet, dass alle Gruppen die gleicheVerteilungsfunktion haben, bis auf eine Verschiebung um den Mittelwert μ_i . Mehr hierzu findet man z.B. bei Beasley & Zumbo (2009), Serlin & Harwell (2001), sowie Koch, (1969).

Zunächst werden die bekanntesten Verfahren vorgestellt (2.1. bis 2.8), die durchweg alle für mehrfaktorielle Analysen sowohl ohne und als auch mit Messwiederholungen anwendbar sind.

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 beobachteten Werte x_{im} über alle Gruppen hinweg in Ränge R_{im} (m=1,..,n), sog. *Wilcoxon-Ränge*, transformiert und daraus eine χ^2-verteilte Testgröße errechnet, über die die Gleichheit der Mittelwerte geprüft wird. Beim Friedman-Test werden für jeden Merkmalsträger m die x_{im} in Ränge R_{jm} (j=1,..,J), sog. *Friedman-Ränge,* transformiert und daraus eine χ^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**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 α-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 (1997, vgl. Kapitel 2.8) entwickelt.

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

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

mit *J*-1 Zähler-FG und (*J*-1)(*n*-1) Nenner-FG, wobei *n* die Anzahl der Merkmalsträger ist. Der F-Wert hat deutlich bessere Eigenschaften und verleiht dem Friedman-Test eine etwas höhere Teststärke (siehe Iman und Davenport, 1976). Allerdings zeigen Harwell & Serlin (1994) das Gegenteil. Die Anwendung dieser Korrektur erübrigt sich selbstverständlich, wenn der χ^2-Wert bereits als signfikant 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 χ^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 χ^2-Werts in einen F-Wert näher untersucht. Sie zitieren Studien, wonach die F-Werte für kleine und mittlere *n* vorteilhafter sind.

Lüpsen (2023b) 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, insbesondere auch gemischte Versuchspläne, anwendbar. Der Algorithmus wird in Kapitel 6.5 erläutert. In R wird für die KWF-Methode die Funktion `np.anova` bereitgestellt (vgl. Anhang 3).

2. 2 Rank transform Tests (RT)

Dies sind die klassischen Anova-F-Tests angewandt auf Rangdaten. D.h. alle Werte der abhängigen Variablen, über Gruppen und Messwiederholungen hinweg, werden in Ränge R_i 1,2,...,*N,* bzw. 1,2,...,*N***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 α-Risikos für den Test der Interaktion berichtet, wenn zugleich signifikante Haupteffekte bestehen (vgl. u.a. Toothaker & 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 & Zu-

mbo, 2009). Auf der anderen Seite konnten Hora und Conover (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 klasssischen 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. Für den Fall von Messwiederholungen wird empfohlen, die Adjustierung der Freiheitsgrade von Huynh-Feldt (HF) wie im Fall des parametrischen F-Tests anzuwenden (vgl. Kapitel 5.2). Die Anwendung des Verfahrens in SPSS wird allerdings dadurch erschwert, dass im Fall von Messwiederholungen der Datensatz zunächst transformiert werden muss, um die Ränge ausrechnen zu können, und anschließend in die ursprüngliche Form zurück transformiert werden muss. (Siehe dazu die Beispiele in den Kapiteln 5.4.2 sowie 6.3.)

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:

$$\Phi^{-1}(R_i/(N+1)) \qquad (2 - 2)$$

wobei Φ die Standardnormalverteilung und N die Anzahl aller Werte insgesamt ist. (Die Division durch N+1 ist erforderlich, um den Wertebereich 1...N in das Intervall 0...1 zu transformieren.) 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:

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

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 α-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 α-Risiko leicht verletzt wird.

Das INT-Verfahren geht u.a. auf van der Waerden zurück (vgl. Kapitel 2.6). 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 eine hohe Effizienz hat. Wie beim RT empfiehlt sich im Fall von Messwiederholungen die HF-Adjustierung (vgl. Kapitel 5.2).

2. 4 Aligned rank transform (ART und ART+INT)

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 „stö-

renden“ Effekte, z.B. der Haupteffekte im Fall der Analyse einer Interaktion, bereinigt, das sog. *alignment*. Hierzu gibt es zwei Methoden, 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:
  $$x'_{ijm} = x_{ijm} - \bar{a}_i - \bar{b}_j + 2\bar{x} \qquad (2 - 4)$$
  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 x (mit allen Effekten) durchgeführt. Zu den daraus resultierenden Residuen e_{ijm} 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
  $$x'_{ijm} = e_m + (\overline{ab}_{ij} - \bar{a}_i - \bar{b}_j + 2\bar{x}) \qquad (2 - 5)$$
  untersucht, wobei e_{ijm} die Residuen des kompletten varianzanalytischen Modells, $\bar{a}_i$, $\bar{b}_j$, $\overline{ab}_{ij}$, $\bar{x}$ 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 mittels robuster Mittelwerte wie Median oder getrimmer Mittelwerte anstatt des arihmetischen 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.

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 er-

haltenen 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 & Claustriaux (2005) sowie Lüpsen (2018). Die Anwendung der INT-Transformation führt übrigens auch zu einer deutlichen Vergrößerung der Power. Allerdings zeigte Lüpsen (2024), dass die INT-Transformation im Fall von gemischten Versuchsplänen weniger hilfreich ist als in solchen nur mit Gruppierungsfaktoren.

Eine große Warnung:
Das ART-Verfahren kann aber nicht empfohlen, sondern eher von der Verwendung abgeraten werden, da es eine Reihe von Situationen gibt, in denen es das α-Risoko krass verletzt, so u.a. in den Fällen:

- diskreter abhängiger Variablen, insbesondere bei größeren *n* (vgl. Lüpsen, 2017)
- stark schiefer Verteilungen wie der Exponential-Verteilung (vgl. Lüpsen, 2016),
- heterogener Varianzen (s. z.B. Leys & Schumann, 2010, und Carletti & Claustriaux, 2005)
- von Tests der Haupteffekte bei größeren Zellenbesetzungszahlen $n>20$ (vgl. Lüpsen, 2017).

Gerade der erste 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.

Aktuell ist eine ausführliche und exzellent dargestellte Untersuchung des ART-Verfahrens von Tsandilas & Casiez erschienen, in der die o.a. Punkte detailliert analysiert und bestätigt werden. Die Autoren raten eindringlich von der Benutzung von ART ab und empfehlen statt dessen die o.a. INT-Methode.

2. 5 Puri & Sen-Tests (PS), L statistic

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 χ^2-Test durchgeführt, auch *L statistic* genannt. Bei Versuchsplänen ohne Messwiederholungen sind dies Verallgemeinerungen des Kruskal & Wallis-Tests auf mehrfaktorielle Versuchspläne. Die Testgröße errechnet sich im Fall von Versuchsplänen ohne Messwiederholungen als

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

bzw. für Gruppierungsfaktoren im Fall von Versuchsplänen mit Messwiederholungen als

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

bzw. im Fall von Messwiederlungsfaktoren als

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

wobei

- SS_{Effekt} die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,
- MS_{total} die Gesamtvarianz (Mean Square),
- $MS_{zwischen}$ 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,
- $MS_{innerhalb}$ die Varianz (Mean Square) innerhalb der Versuchspersonen und
- SS_{Fehler} die Streuungsquadratsumme des zum getesteten Effekt gehörenden Fehlers ist.

Alle *SS* und *MS* können aus den üblichen Anova-Tabellen abgelesen werden. Falls nur ein Messwiederholungsfaktor vorliegt, entspricht der Nenner in 2-7 der Varianz $MS_{innerhalb}$. Als Freiheitsgrade für den χ^2-Test nimmt man die Zählerfreiheitsgrade aus der Varianzanalyse. Für die Haupteffekte ergibt dies die Tests 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_{total} bzw. $MS_{innerhalb}$, 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 *a*-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 Spherizität der Kovarianzmatrizen bzw. deren Homogenität gefordert wird, da hier χ^2-Tests anstatt F-Tests verwendet werden. Dennoch wird auch hier empfohlen, wie beim K-W- oder Friedman-Test auf gleiche Verteilungsformen, insbesondere gleiche Streuungen zu achten. (Adäquate Verfahren werden in Abschnitt 2.20 vorgestellt.)

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).

Die Umsetzung in R, insbesondere in 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 muss hierzu im Fall von Messwiederholungen bei SPSS wie in Kapitel 2.2 angedeutet die Datenmatrix zweimal transformiert werden. Hinzu kommt die Durchführung der χ^2-Tests über o.a. Quotienten, die insbesondere in SPSS mit dem Taschenrechner erfolgen muss, während für R die Funktion `np.anova` zur Verfügung steht (vgl. Anhang 3).

2. 6 van der Waerden (vdWS)

Die Methode von van der Waerden vereinigt gewissermaßen die INT-Methode sowie die Verfahren von Kruskal & Wallis und Friedman bzw. das o.a. KWF-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 anstatt des F-Tests der χ^2-Test wie beim Kruskal-Wallis-Test angewandt. 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 (2023b) hat das Verfahren für gemischte Versuchspläne erweitert und vdWS (van der Waerden Scores) ge-

nannt. Allgemein werden die Effekttests mittels χ^2-Tests wie beim Puri & Sen-Verfahren errechnet, lediglich mit anders transformierten *x*-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, 1980, 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 hier 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 Stichproben $n>10$. Positiv ist noch zu bemerken, dass im Fall von Messwiederholungen nicht die sonst kritische Sphärizität der Kovarianzmatrizen bzw. deren Homogenität gefordert wird, da hier χ^2-Tests anstatt F-Tests verwendet werden. Dadurch ist der v.d.Waerden-Test das empfehlendswerteste nichtparametrische Verfahren, zumindest bei unabhängigen Stichproben, während bei gemischten Versuchsplänen die Power geringer, eher durchschnittlich, ausfällt, wie Lüpsen (2023b) gezeigt hat.

Der Aufwand ist exakt derselbe wie bei den im vorigen Kapitel skizzierten Puri & Sen-Tests. In R wird für die KWF-Methode die Funktion `np.anova` bereitgestellt (vgl. Anhang 3).

2. 7 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 Brunner & Munzel (2013) 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 χ^2-verteilte Statistik WTS (*Wald type statistic*) an, die aber hier nicht berücksichtigt wird, da die ATS wegen der Berücksichtigung der n_i bessere Eigenschaften aufweist. Letztlich werden dabei die relativen (Behandlungs-) Effekte p_i, anstatt Mittelwerte, verglichen, mit

$$p_i = (\overline{R}_i - 0,5)/n \qquad (\text{mit } \overline{R}_i = \text{mittlerer Rang und } N = \sum n_{ij})$$

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 dichtotome abhängige Variablen anwendbar (vgl. Noguchi et al., 2012). Es sei darauf aufmerksam gemacht, dass es zwei Varianten des ATS gibt: eine semiparametrische und eine nichtparame-

trische (vgl. Formeln 5 bzw. 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 α-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, 2018). 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. Im Fall von gemischten Versuchsplänen fand Lüpsen (2023b) eine erhebliche Verletzung des α-Risikos für kleine Stichproben ($n_i \leq 15$). 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. Vom Autor werden die Funktionen `ats.2` und `ats.3` angeboten (vgl. Anhang 3). In SPSS sind diese Tests wegen der umfangreichen Matrizenrechnungen nicht durchführbar.

2. 8 Verteilungs-robuste Verfahren

Hierunter sind solche Verfahren zu verstehen, die insbesondere robust gegen nichtnormale Verteilungen sind, nicht jedoch robust gegen inhomogene Varianzen. Auf der anderen Seite können allerdings auch ungleiche Varianzen durch Extremwerte hervorgerufen werden. Üblicherweise werden die klassischen Varianzanalysen als Basis genommen und bei der Berechnung die arithmetischen Mittel und Varianzen durch robustere Statistiken ersetzt. In der Regel werden dabei die Mittelwerte als getrimmte arithmetische Mittel ersetzt und die Varianzen durch winsorisierte Varianzen. Bei getrimmten Mittelwerten wird ein festgelegter Prozentsatz der größten und kleinsten Werte bei der Berechnung weggelassen, z.B. jeweils 10%. Bei winsorisierten Varianzen wird ähnlich verfahren, allerdings werden die größten und kleinsten Werte nicht einfach weggelassen, sondern durch den nächstgrößeren bzw. nächstkleineren Wert ersetzt. Dies ist anschaulich beschrieben z.B. von Dennis Twesmann (2022). Lix & Keselman (1998) haben gezeigt, dass diese getrimmten Varianzanalysen nicht nur nichtnormale Verteilungen kompensieren, sondern auch Varianzheterogenitäten abschwächen können. Ähnlich positive Ergebnisse berichtet Wilcox (1993) über seine Simulationen hinischtlich Varianzheterogenitäten in Split-Plot Designs.

Allerdings gibt es in der Literatur zahlreiche robuste Methoden für Regressionsanalysen. Dies ist insofern interessant, da die Varianzanalyse ein Spezialfall der Regressionsanalyse ist und damit Regressionsmethoden- mit ein wenig Aufwand - verwendet werden können. Als Methoden sind zu nennen: zum einen die Abwandlungen der *least squares* (LS) Methode: die *least median of squares* (LMS, vgl. Rousseeuw, 1984), bei der anstatt der Quadratsumme der Median der Residuenquadrate minimiert wird, sowie eine Variante hiervon, die weniger bekannte *least quantile of squared residuals* (LQS, vgl. Bertsimas & Mazumder, 2014), bei der, ähnlich der LMS, ein Quantil gewählt wird. Zum anderen Methoden, bei denen Ausreißer der Rohwerte bzw. der Residuen identifiert und gewichtet werden: die Verfahren basierend auf *Hubers M-Schätzer* und deren Erweiterungen, den *MM-Schätzern* von *Yohai* (vgl. Yohai, 1987), sowie die *least trimmed squared residuals* (LTS, vgl. Atkinson & Cheng,1999), bei denen, ähnlich dem *trimmed mean*, Extremwerte nicht verwendet werden. Robuste Varianzanalysen basierend auf

den o.a. Verfahren werden unter Verwendung der entsprechenden Regressionsmethoden vom Autor mit der Funktion `rob.anova` zur Verfügung gestellt (vgl. Anhang 3).

2. 8. 1 Mair & Wilcox

Wilcox (Mair & Wilcox, 2019) hat ein R-Paket (`WRS2`) von robusten Funktionen entwickelt, sowohl für unabhängige, als auch für abhängige Stichproben und gemischte Versuchspläne. Wilcox bietet zudem nicht nur die o.a. robusten Statistiken (getrimmte Mittelwerte und winsorisierte Varianzen an), sondern auch zwei Versionen des *M-estimators* mit *Huber's Psi*, den Median und *Bootstrapping*-basierte Schätzungen, die auch häufig bei (leicht) heterogenen Varianzen eingesetzt werden (vgl. Luh, 1999). Mehr dazu in Abschnitt 4.3.4.

2. 8. 2 Ananda und Weerahandi

Ananda & Weerahandi (siehe Dag et al., 2024) bieten ein R-Paket `twowaytests` an, in dem einige Funktionen für verteilungsrobuste Varianzanalysen enthalten sind. Dazu zählen Analysen auf Basis von Medianen und getrimmten arithmetischen Mitteln, aber auch unter Verwendung robuster Schätzer, z.B. Hubers Psi, Tukeys biweight (bisquare) estimator sowie Hampels estimator. Ein Hinweis: Es können nur 2-faktorielle Analysen durchgeführt werden.

2. 9 Varianzanalysen für heterogene Varianzen

Varianzhomogenität ist die wichtigste Voraussetzung nicht nur der parametrischen Varianzanalyse, sondern auch der meisten nichtparametrischen Analysen. Daher wurden (inzwischen zahlreiche) Methoden entwickelt, die die Ungleichheit der Varianzen berücksichtigen. Im Fall von unabhängigen Stichproben sind dies die Varianzen der einzelnen Gruppen/Stichproben. Im Fall von abhängigen Stichproben sind dies im einfachsten Fall die Varianzen der Stichproben, im Normalfall die Varianzen der Messwiederholungsvariablen. Mehr dazu in den Kapiteln 2.13 und 6. An dieser Stelle werden zunächst nur die "wichtigsten" Methoden für unabhängige Stichproben aufgeführt. In R gibt es dazu insbesondere die Pakete `onewaytests`, `twowaytests`, `vartest` und `doex`, die noch eine Reihe Funktionen für weitere Methoden enthalten. An dieser Stelle sei darauf hingewiesen, dass es zwar viele gute Methoden für 1-faktorielle Analysen, aber nur wenige robuste für 2- oder mehrfaktorielle Varianzanalysen gibt.

2. 9. 1 Welch und Fligner-Policello

Das wohl bekannteste Verfahren stammt von *Welch*. Er entwickelte einen Zweistichproben- t-Test für ungleiche Varianzen. Diesen gibt es auch in einer Version für *I* Gruppen (unabhängige Stichproben), der sowohl in R (Funktion `oneway.test`) als auch in SPSS (Prozedur `Oneway`) verfügbar ist. Für 2-faktorielle Analysen gibt es den u.a. Test von *Welch und James,* der eine Erweiterung des 1-faktoriellen Welch-Tests 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. 9. 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). Beide Methoden sind allerdings nur robust gegen heterogene Varianzen, wenn die Daten symmetrisch verteilt sind (siehe z.B. Myers, 1998). 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. 9. 3 Welch & James (WJ)

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, auch für 2-faktorielle Versuchspläne. Er ist beschrieben von Algina & Olejnik (1984) und basiert auf einem χ^2-Test. Darüber hinaus gibt es auch eine neuere Approximation mittels einem F-Test. Derzeit ist dieses Verfahren in den Standardprogrammen nicht verfügbar. Für R wird es jedoch als Funktion `wj.anova` vom Autor angeboten (vgl. Anhang 3).

2. 9. 4 Brown & Forsythe (BF)

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 `oneway-tests`, in der verbesserten Version als Funktion `MBF` im Paket `doex`, sowie in SPSS (Prozedur `Oneway`) verfügbar. Für R wird die Funktion `bf.f` für 1- und 2-faktorielle Varianzanalysen vom Autor angeboten (vgl. Anhang 3).

2. 9. 5 Brunner, Dette und Munk (BDM)

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. Das Verfahren ähnelt dem o.a. von Akritas, Arnold und Brunner (ATS), 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 (2013) beschrieben worden. R bietet dafür folgende Pakete: `GFD` für die parametrische mehrfaktorielle (Friedrich et al.,2017) sowie `asbio` für die nichtparametrische 1- und 2-faktorielle Varianz-analyse.

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

2. 9. 6 Ananda und Weerahandi

Ananda & Weerahandi (Dag et al. 2024, Ananda et al. 2023, Ananda & Weerahandi 1997) bieten ein (oben bereits erwähntes) R-Paket `twowaytests` an, in dem eine Funktion `gpTwoWay` für zwei Varianzanalysen bei heterogenen Varianzen enthalten ist, die auf der Methode von *generalized p-values* sowie *Bootstrapping* basieren: gPB (Parametric Bootstrap) und GPQ (Generalized Pivotal Quantity). Einzelheiten sind den o.a. Artikeln zu entnehmen sowie bei Tsui und Weerahandi (1989). Allerdings haben eigene Simulationen gezeigt, dass es manchmal irritierende Ergebnisse mit dieser Funktion gibt: unter anderem im Fall einer Interaktion werden häufig auch die Haupteffekte als signifikant ausgewiesen, im Gegensatz zu den Resultaten aus anderen robusten Verfahren. Diess machen diese Funktion leider nicht empfehlenswert. Im Paket sind auch die Tests von Levene, Bartlett und Fligner auf Varianzhomogenität enthalten. Ein Hinweis: Es können nur 2-faktorielle Analysen durchgeführt werden.

2. 9. 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 werden üblicherweise Varianzinhomogenitäten Rechnung getragen. Diese Box-Korrektur ist allerdings als vergleichsweise konservativ einzustufen. Eine entsprechende R-Funktion `box.f` ist im Anhang 3 zu finden.

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, 2018), werden sie hier nicht näher vorgestellt. Und von einer Benutzung wird abgeraten.

Hettmansperger & 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 (2018) zeigte sich allerdings, dass das α-Risiko selbst bei einem Modell ohne Effekte krass überschritten wird.

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 Vetfügung.

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 ($m=1,...,n$) die Werte $j=1,..., J$ 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 Wertesprektrum 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 Variaben 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 auch 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` ebenfalls fehlende Werte zulässt.

2. 11. 3 Multivariate Tests: Hotelling-Lawley, Pillai und Wilks

Neben der „klassischen“ parametrischen (univariaten) Varianzanalyse, die Spherizität (vgl. Kapitel 5.2) 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 aufwändiger zu überprüfen. Hierbei werden zunächst für die *J* Messwiederholungen $x_1,...,x_J$ 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, der zweite eher liberal reagieren. Wilks Test ist insbesondere bei nichtnormalen Daten und ungleichen Kovarianzmatrizen zu empfehlen. Der Test von *Roy* ist dagegen nicht zu empfehlen (vgl Olson, 1976).

2. 11. 4 Agresti & Pendergast (AP) und Akritas & Arnold

Agresti & Pendergast (1986) haben ein nichtparametrisches Äquivalent zum o.a. multivariaten Test entwickelt. Letztlich werden dabei nur die *x*-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 χ^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). Hierzu gibt es die Funktion `ap.anova` (vgl. Anhang 3).

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 χ^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. 11. 5 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 auf den mehrdimensionalen Raum, die im 1-dimensionalen Raum zum üblichen Werkzeug gehören. 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 die oben aufgeführten parametrischen, mit dem Vorteil ohne Spherizität auszukommen, und zum anderen zum Test auf Spherizität selbst. 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 Koch

Das Verfahren von *G.G. Koch* (1969) basiert auf dem oben erwähnten Ansatz einer multivariaten Varianzanalyse (vgl. Kapitel 5.2). Dieses wird auf Rangdaten übertragen. Eine R-Funktion `koch.anova` 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 (2023b). Es hat aber auch Schwächen: zum einen ist es recht liberal für $n_i \leq 10$, zum anderen hat es eine relativ geringe Power.

2. 12. 2 Beasley & Zumbo

Beasley (2000) sowie Beasley & Zumbo (2009) haben eine Reihe von Tests für die Interaktion bei gemischten Versuchsplänen zusammengestellt. Neben einigen weniger bekannten Verfahren 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. 13 Varianzanalysen für heterogene Varianzen bei abhängigen Stichproben und gemischten Versuchsplänen

Oben wurden bereits Methoden für unabhängige Stichproben vorgestellt. Im Fall von abhängigen Stichproben, also Messwiederholungen, wird die Varianzhomogenität zu dem Begriff Sphärizität verallgemeinert. Bei gemischten Versuchsplänen, also solchen mit Gruppierungs- als auch Messwiederholungsfaktoren, sind zwei Varianzhomogenitäten relevant: die des oder der Gruppierungsfaktoren und die des oder der Messwiederholungsfaktoren. Erstere ist dann allerdings nicht mehr eine einfache Varianzhomogenität, sondern eine Homogenität der Kovarianzmatrizen, da die Varianzen auf alle Messwiederholungen ausgedehnt werden. Dies wird etwas ausführlicher in Kapitel 6.1 erläutert.

Es gibt praktisch keine speziellen Verfahren für reine Messwiederholungs-Versuchspläne. Statt dessen sind solche für gemischte Versuchspläne anzuwenden und dabei keine Gruppierungsfaktoren anzugeben.

2. 13. 1 Welch & James (WJ)

Eine Variante des bereits oben erwähnten Tests von Welch & James für gemischte Versuchspläne wurde von Keselman, Carriere & Lix (1993) vorgestellt. Diverse Simulationen (vgl. Algina, 1994, oder Lüpsen, 2024) haben jedoch gezeigt, dass dieses Verfahren nur für den Test des reinen Messwiederholungsfaktors zuverlässig ist. Dagegen hat es für den Test der Interaktion, je nach Anzahl J der Messwiederholungen, nur für sehr große Stichproben (etwa n_i>40) die Fehlerrate unter Kontrolle, genauer: $min(n_i) \geq 3(J-1)$ für normalverteilte Daten und $min(n_i) \geq 5(J-1)$ für nicht-normalverteilte, insbesondere rechtsschiefe Daten. Derzeit sind diese Verfahren in den Standardprogrammen nicht verfügbar. Für R werden jedoch beide Varianten des Verfahrens als Funktion `wj.spanova` vom Autor angeboten (vgl. Anhang 3). Für Split-Plot Designs gibt es daüber hinaus die Funktion `welchADF.test` im Paket `welchADF`.

2. 13. 2 Brown & Forsythe (mBF)

Es gibt eine Erweiterung dieses bereits oben erwähnten Verfahrens von *Coombs & Algina* (1992) für gemischte Versuchspläne, die sowohl heterogene Varianzen des Gruppierungsfaktors wie auch des Messwiederholungsfaktors berücksichtigt (vgl. z.B. Vallejo et al., 2004, und Algina, 1994), auch *modifizierter Brown & Forsythe-Test* (mBF) genannt. Hierzu wird die Funktion `mbf.f` vom Autor zur Verfügung gestellt. Simulationen (z.B. Lüpsen, 2024) haben gezeigt, dass dieses Verfahren, im Gegensatz zu der Version für unabhängige Stichproben, extrem konservativ ist, insbesondere bei nichtnormalen Verteilungen.

2. 13. 3 Adjustierungen der Freiheitsgrade: ε, GA und IGA

Für den Fall ungleicher Varianzen des Messwiederholungsfaktors, insbesondere fehlender Sphärizität, haben *Huynh & Feldt* sowie *Greenhouse & Geisser* eine Adjustierung des F-Tests vorgeschlagen, derart dass Zähler- und Nennerfreiheitsgrade des F-Tests mit ε‘ multipliziert werden, wobei ε‘ ≤ 1 auf Box ε, dem Grad der Heterogenität, basiert, Je nach Größe von ε‘ 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 Sphärizität in den einzelnen Gruppen unter-

schiedlich ausfällt. *Huynh* (1978) hat dazu eine Alternative GA (*general approximation*) für die Adjustierung entwickelt, die nicht auf Box ε basiert, sondern lediglich auf den Varianzen und Kovarianzen. Er hat diese weiter verbessert zur IGA (*improved general approximation*), die auch Heterogenitäten der Kovarianzmatrizen berücksichtigt. Hierbei werden nicht nur die Freiheitsgrade, sondern auch der F-Wert gemäß der Heterogenität verkleinert (vgl. auch Algina, 1994). Wie bei o.a. ε‘ führt dieses zu einem konservativerem F-Test. Entsprechende R-Funktionen (`iga` und `iga.anova`) werden vom Autor angeboten (vgl. Anhang 3).

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 dichtomoe 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 Regressions-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 anwenbar sind, ist die ordinale logistische Regression eine Methode, die speziell auf ordinale Merkmale zugeschnitten ist. Die Anwendung ist allerdings nicht 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, in R über die Parameter `family` und `link`, in SPSS über `distribution` und `link`. (Einigermaßen) verständliche Einführungen in diese Verfahren bieten u.a. Baltes (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 Varianzhomo-

genitäten. Dagegen sind die Voraussetzungen für GLMM vergleichsweise komplex und zudem schwer verifizierbar (siehe Thiele & Markussen, 2012). Allerdings haben die wenigen Studien eine große Robustheit gezeigt (siehe Schielzeth et al., 2020). Bei beiden Methoden ist die Struktur der Korrelationsmatrix der Messwiederholungen wesentlicher Bestandteil des Modells (vgl. auch Abschnitt 5.2.). Gängige Strukturen für die Korrelationen r_{ij} (für Messwiederholungen *i, j* 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 erfordert keine entsprechende Spezifikation.

Beide Methoden basieren auf asymptotischer Statistik, d.h. benötigen 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_i als zuverlässiger gelten, erlauben die GLMM auch Versuchspläne mit fehlenden Werten auf den Messwiederholungen, ohne dass entsprechende Fälle eliminiert werden müssen.

Beide Verfahren sind primär Regressionsanalysen, die den Einfluss mehrerer Prädiktoren auf eine abhängige Variable untersuchen. Somit besteht zunächst einmal das Ergebnis aus der Schätzung der Regressions-Modell-Parameter und der dazugehörigen Tests auf Verschiedenheit von 0. Für eine Varianzanalyse müssen daher wie bei logistischen Regression (s.o.) die Faktoren in dichotome Prädiktoren umgewandelt werden. Hat ein Faktor mehr als 2 Ausprägungen, so müssen die entsprechenden 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 (1986) das Standard-Verfahren ist. Eine Übersicht geben Wang et al. (2016). Nicht viel besser sieht es bei GLMM aus, wozu Li & Redden (2015) eine Reihe von Methoden zusammengestellt haben. Vielfach wird die Methode von Kenward & Roger empfohlen, die allerdings manchmal zu unbrauchbaren Resultaten führt (vgl. Lüpsen, 2024, Guerin &Stroup, 2000, sowie Arnau et al. 2009). Eine Alternative kann der Wald Type II sein, der leider für kleine n_i recht liberal ist.

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 für Varianzanalysen 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) und Harville (1977) 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 (siehe z.B. Nooraee et al. 2014).

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). McNeish & Stapleton (2016), Stiger et al. (1998) wie auch Lüpsen (2023a) verglichen GEE und GLMM für typische varianzanalytische Datensaätze und geben zwar GLMM den Vorzug, halten aber beide Methoden für varianzanalytische Probleme weniger geeignet.

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,..,*N* zugeordnet werden. Aber es gibt auch andere Rangberechnungen, 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:

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

Hierbei ist *c*(..) 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:

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

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) 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 im Abschnitt 2.11 angedeutet, können multivariate Verfahren für Analysen mit Messwiederholungen eingesetzt werden.

Spatial signs und s*patial ranks* sind zunächst einmal Verallgemeinerungen von Vorzeichen und Rängen auf den mehrdimensionalen Raum, die im 1-dimensionalen Raum zum üblichen Werkzeug gehören. 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 $y=(y_1,..,y_J)$ 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 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,..,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 oder nachfolgend 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 KWF- und 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:

$F_1(x) = F_2(x) = ... = F_I(x)$

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 α=0.10) für die alternative Methode entscheiden. Hierfür plädieren u.a. Wilcox et al. (1986) und Zimmermann (2004), d.h. in diesem Fall: Wahl einer Anova-Methode, die robust gegen inhomogene Varianzen bzw. nichtnormale Verteilung ist.

2. 17. 1 Versuchspläne ohne Messwiederholungen

Im Fall von unabhängigen Stichproben wurde bislang in den meisten Textbüchern zum Test der Varianzhomogenität von x oder R(x) der *Levene-Test* empfohlen, da dieser robust gegen Abweichungen von der Normalverteilung und auch für ordinale Variablen anwendbar ist. In den letzten Jahren sind allerdings viele Vergleiche der inzwischen zahlreichen Verfahren publiziert worden, die diese Empfehlung in Frage stellen. Allerdings bei Verwendung der Differenzen zum Median anstatt zum arithmetischen Mittel, erweist er sich immer noch als einer der besseren (vgl. Sharma & Kibria, 2012). Alternative Tests sind in Abschnitt 2.20.2 aufgeführt. 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, jedoch in SPSS nicht standardmäßig verfügbar ist. Für R wird die 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 robusten 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 Abschnitt 20.2.3 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 Homo-

genitätsvoraussetzungen 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 χ^2- anstatt F-Tests. Harwell & Serlin (1994) weisen in Berufung auf E.L. Lehmann (1975) 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, 2023b), 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 Spherizität, insbesondere also der Varianzhomogenität. Dennoch ist es empfehlenswert, auch bei den Methoden, die auf χ^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 α-Risikos reagieren. Allerdings verlangen die drei o.a. Verfahren Gleichheit der Kovarianzmatrizen, wenn diese auch relativ robust gegen Verletzungen dieser Art sind, ganz besonders das KWF-Verfahren (vgl. Lüpsen, 2023b).

2. 18 Vergleiche

2. 18. 1 Versuchspläne ohne Messwiederholungen

An dieser Stelle ist ein Vergleich von Lüpsen (2018) zu erwähnen, in dem viele der 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 α-Risiko nicht unter Kontrolle, gelegentlich selbst bei gleichen n_{ij}. Dabei steigen die Fehlerraten deutlich über 0,10 (bei α=0,05) für n_{ij}->50. Auf die Tücken des ART wurde bereits in Abschnitt 2.4 hingewiesen. Einzig bei negativem Pairing (siehe Abschnitt 1.1.6) wird die ATS-Methode empfohlen. In der o.a. 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 bieten Hahn, Konietschke und Salmaso (2013). Sie bestätigen, dass die ATS-Methode zwar generell das α-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 beaurteilt, u.a. von Lei et al. (2004) und Mansouri et al. (2004), die aber meistens kritische Fälle wie ungleiche Varianzen oder diskrete abhängige Variablen 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 Vorausstzungen 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 Zellhäufigkeiten, 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.

Lüpsen (2026) hat ein einer kleinen (noch nicht veröffentlichten) Studie die für 2-faktorielle Designs zur Verfügung stehenden Verfahren verglichen. Dabei hat sich gezeigt - wie auch bei den wenigen anderen Vergleichen (z.B. von Hsiung & Olejnik, 1996), dass der schwierigste und selten berücksichtigte Fall stark heterogene Varianzen gepaart mit schiefen Verteilungen ist. Dazu kommt das Problem, dass die sehr robusten Methoden meistens eine vergleichsweise geringe Teststärke besitzen. Nicht nur bei Lüpsen, sondern allgemein schneiden die oben erwähnten Tests von Brown & Forsythe sowie von Welch & James „overall“ am besten ab, gefolgt von der Box-Approximation und den robusten Verfahren von Wilcox. Die verschiedentlich vorgeschlagene Variante, vor der Anova die Werte in Ränge zu transformieren (vgl. Cribbie, 2007), hat hier nicht den gewünschten Erfolg gezeigt, da dadurch zu häufig zu liberale p-Werte erzeugt wurden. Nachfolgend einige Resultate:

- Brown & Forysythe: nur in pairing-Situationen bei großen n_i wurde das α-Risiko verletzt. Durchweg (vergleichsweise) gute Power.
- Welch & James: nur bei schiefen Verteilungen in pairing-Situationen wurde das α-Risiko der Interaktion verletzt. Durchweg (vergleichsweise) gute Power.
- Box: nur bei der Normalverteilung wurde in pairing-Situationen das α-Risiko verletzt. (Wider Erwarten) passable Power.
- Wilcox Methoden: Beide Methoden schneiden bei normal-verteilten Daten sehr gut ab, die trimmed means-Methode besser für große n_i, die M-Schätzer-Methode besser bei kleinen n_i.

Bei schief-verteilten Daten gepaart mit stark heterogenen Varianzen reagieren beide liberal. Gute Power.

- BDM-Tests: Die parametrische Variante hält das α-Risiko komplett unter Kontrolle, auf Kosten einer (vergleichsweise) schlechten Power. Die nichtparametrische Variante reagiert dagegen sehr liberal bei schiefen Verteilungen mit stark heterogenen Varianzen oder in pairing-Situationen, dafür mit einer deutlich besseren Power.
- ATS: Schneider bei normal-verteilten Daten gut ab, dagegen sehr liberal bei schiefen Verteilungen mit stark heterogenen Varianzen oder in pairing-Situationen, dazu eine relativ schwache Power.
- van der Waerden: Insgesamt gut, außer in pairing-Situationen oder bei stark heterogenen Varianzen. Gute Power.
- Weerahandi: Beide Methoden sind nicht zu empfehlen, da sie nur bei normal-verteilten Daten das α-Risiko einhalten, und dann auch nur, wenn keine Effekte vorhanden sind.

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 Spherizität, Homogenität der Kovarianzmatrizen, gewählte Verteilungen, Anzahl der Messwiederholungen oder Größe der Stichproben abhängig und daher schwer wiederzugeben sind. Ein paar sollten dennoch erwähnt werden:

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 (1995) 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 Spherizität 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 Kriterumsvariablen auf einer 4er-Skala untersucht hatten.

Lüpsen (2023b) 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 und 2024). Er machte auch darauf aufmerksam, dass das KWF- sowie das van der Waerden-Verfahren als einzige das α-Risiko auch bei heterogenen Kovarianzmatrizen unter Kontrolle halten und dadurch eine gute Wahl sind, zumal sie zwar keine Spherizität 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 enttäusschend 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 α=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 einer neueren, noch nicht veröffentlichten Studie vergleicht Lüpsen (2024) neben den o.a. Methoden auch den robusten IGA von Huynh, den modifizierten Brown-Forsythe-Test (mBF), den Test von Welch & James, die parametrischen multivariate Verfahren (z.B. von Wilks), GLMM, die nichtparametrischen von Agresti & Pendergast sowie 4 auf spatial ranks basierende Methoden. Insbesondere werden diese unter heterogenen Bedingungen verglichen: heterogene Kovarianzmatrizen (ungleiche Gruppenvarianzen), fehlende Spherizität (ungleiche Messwiederholungsvarianzen oder ungleiche Korrelationen zwischen den Messwiederholungen), Pairing-Situationen bei ungleichen n_i sowie ungleiche Korrelationen der Kovarianzmatrizen. Die Resultate:

- der Test der Interaktion ist durchweg kritischer als der der Haupteffekte,
- bei inhomogenen Varianzen bzw. Kovarianzmatrizen sind die Fälle des Pairing kritisch - für die Prüfung siehe Beispiele in Kapitel 11 - weil nur wenige Methoden diesen Fall handhaben können (IGA, mBF, WJ, KWF, vdWS und ATS),
- alle rangbasierten Methoden (u.a. KWF, vdWS, ART, ATS, Koch) sollten nicht für den Test des Messwiederholungsfaktors (B) verwendet werden, wenn die Varianzen des Messwiederholungsfaktors ungleich sind und eine rechtsschiefe Verteilung (z.B. exponential oder lognormal) zugrunde liegt,
- bei parametrischen Tests kann es bei heterogenen Kovarianzmatrizen auch im Fall gleicher n_i zu erhöhten Fehlerraten kommen,
- der HF (Huynh-Feldt)-adjustierte F-Test schneidet relativ gut ab, solange keine Pairing-Situation vorliegt, reagiert allerdings liberal bei großer Anzahl Messwiederholungen und heterogenen Bedingungen,
- die multivariaten Tests, insbesondere der von Wilks, sind gut, solange keine Pairing-Situation vorliegt, haben aber eine geringere Power als der adjustierte F-Test,
- der ATS-Test zeigt in Pairing-Situationen gute Resultate, jedoch in vielen Situationen zu hohe Fehlerraten (kleine n_i<15, rechtsschiefe Verteilungen, Test des Gruppierungsfaktors),
- der Welch-James-Test benötigt für den Test der Interaktion im Falle ungleicher n_i mindestens n_i >35, hat aber eine geringe Power,
- die GLMM-Methode ist nur bei homogenen Kovarianzmatrizen zu empfehlen, bei kleinen Stichproben vorzugsweise mit dem Kenward-Rogers Test,
- die robusten IGA und mBF halten zwar das α-Risiko komplett unter Kontrolle, sind aber extrem konservativ mit geringer Power, insbesondere im Fall von schiefen Verteilungen,
- das INT-Verfahren eignet sich nur bei homogenen Kovarianzmatrizen, hat aber eine vergleichsweise hohe Power, und sollte ansttat RT und ART eingesetzt werden,

- die KWF- und vdWS-Verfahren schneiden durchweg gut ab, insbesondere mit einer hohen Power, zeigen erhöhte Fehlerraten lediglich bei inhomogenen Kovarianzmatrizen, wenn Varianzen und n_i unabhängig sind, jedoch die niedrigsten Fehlerraten bei negativem Pairing,
- Kochs Methode verhält sich ähnlich wie die KWF-Methode, jedoch mit deutlich geringerer Power,
- die Spatial Signs und Spatial Ranks Methoden haben zwar keine krassen Schwächen, allerdings auch keine nennenswerten Stärken, um sie anderen Methoden vorzuziehen.
- die Methoden von Puri & Sen sowie Agresti & Pendergast sind wegen sehr wechselhafter Ergebnisse insgesamt nicht zu empfehlen,

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 ordinales 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 sinifikanten Ergebnissen führen (vgl. Lüpsen, 2016 und 2024). Und im Fall einer Exponentialverteilung hält der F-Test das α-Risiko komplett unter Kontrolle und hat zugleich die größte Power. Vgl. dazu Zimmerman (2004) sowie Carletti & Claustriaux (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, sogar von ChatGPT. Deren Anwendung sollte jedoch vermieden werden, wie ausführliche Untersuchungen gezeigt haben (vgl. dazu Lüpsen, 2016 sowie Tsandilas & Casiez).

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 stark abgeschwächt auch für die van der Waerden- und Koch-Methode gelten. Diese zielen 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 gestestet

wurden. Leider haben die anderen Methoden wie ART und ATS, oder GEE und GLMM ebenfalls Schwächen, etwa 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 in 2.9. bzw. 2.18 aufgeführten Methoden die einzigen, die den Fehler 1. Art unter Kontrolle halten, u.a. die Methoden von Welch & James (WJ), Brown & Forsythe (BF), Akritas & Brunner (ATS) sowie Brunner, Dette & Munk (BDM). Der WJ ist vielleicht trotz Schwächen bei der Interaktion der empfehlenswerteste. Von allen anderen sind der KWF- und der van der Waerden-Test diejenigen, die noch am besten abschneiden, wenn auch sie das α-Risiko verletzen, allerdings in Maßen. (Das Puri & Sen-Verfahren ist hier mit KWF 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 2.9. 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 α-Risiko unter Kontrolle. Wem der Rechenaufwand zu groß ist, kann ersatzweise auch die INT-Methode wählen.

2. 19. 3 Versuchspläne mit Messwiederholungen

Auch im Fall von Messwiederholungen ist der F-Test durchweg zu empfehlen, aber er sollte generell mit der HF (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 und der modifierte 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.

2. 20 Methoden zur Prüfung der Voraussetzungen

2. 20. 1 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.

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

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* die Test-Statistiken die erforderlichen Verteilungsvoraussetzungen erfüllen. Zum anderen sprechen statistische Tests bei kleinem *n* nicht an, d.h. die Nullhypothese wird 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 $N=\sum n_{ij}$ 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 Mermalsausprägungen, 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.

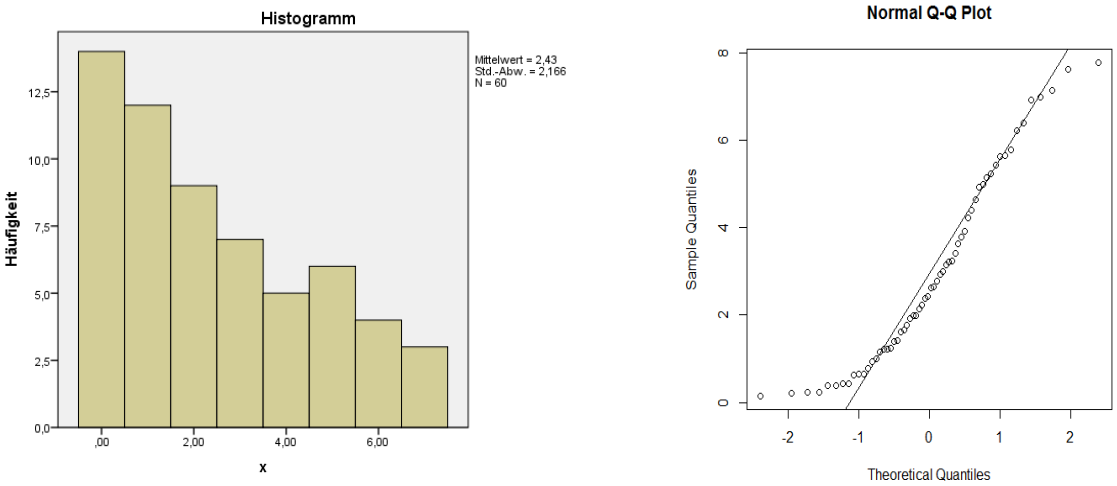
Bei den *normal probabilty Plots*, oder allgemein bei den *Quantile-Quantile-Plots*, kurz *Q-Q-Plots* genannt (vgl. http://en.wikipedia.org/wiki/Normal_probability_plot), wird die empirische (kumulative) Verteilung mit der theoretischen, hier der Normalverteilung, verglichen. Üblicherweise ist die empirische 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 Punke auf einer Geraden. Im Gegensatz zu den Histogrammen sind diese Grafiken unabhängig von Intervalleinteilungen, die möglicherweise ein Bild „verzerren“ können.

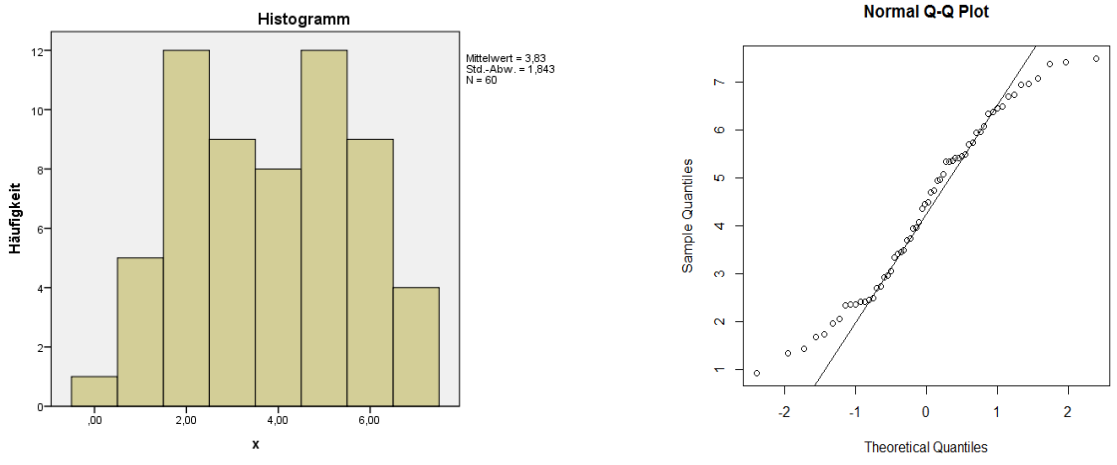
Aber sowohl die Interpretation von Histogrammen auch der Q-Q-Plots bedarf ein wenig Erfahrung. Die wichtigsten Kennzeichen einer Normalverteilung sind Symmetrie und Eingipfligkeit. Nachfolgend werden einige typische Verteilungsformen aufgezeigt, die zum Teil nicht mehr als normal eingestuft werden können.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

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.

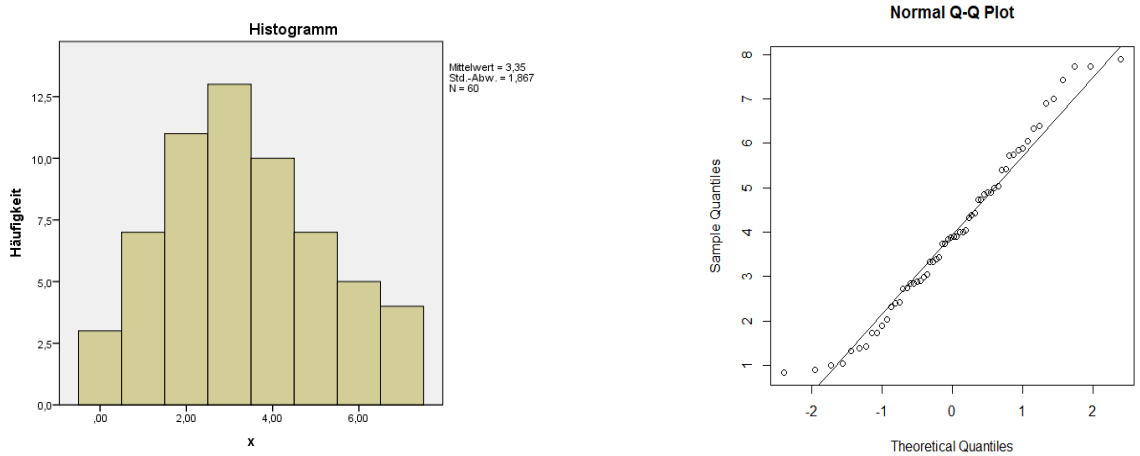
Die parametrischen multivariaten Tests, die auch für univariate Analysen mit Messwiederholungen eingesetzt werden, verlangen eine *multivariate* Normalverteilung der Residuen. In R werden dazu mehrere Pakete angeboten. So bietet z.B. das Paket `mvnormalTest` die Tests von Fattorini, Mardia, Henze-Zirkler, 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.



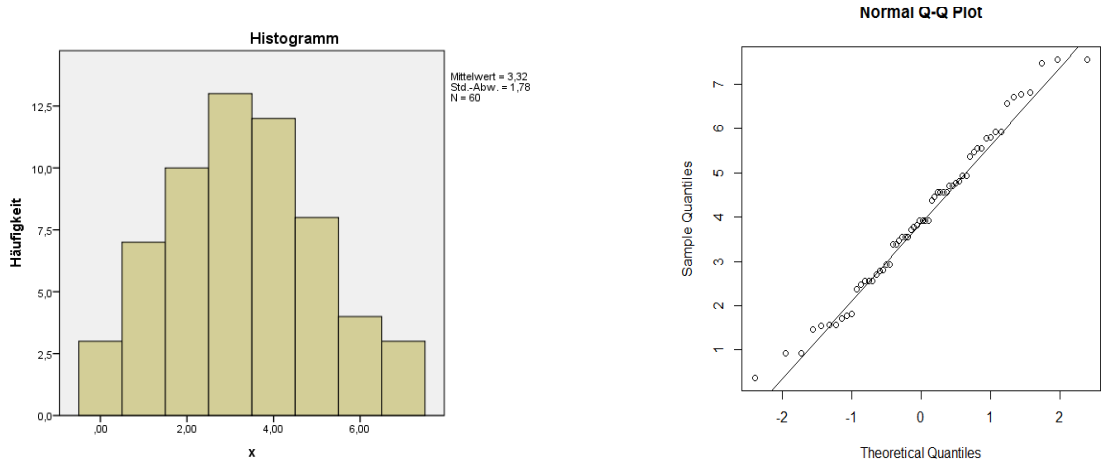
stark rechtsschiefe Verteilung (W=0.894 - p=0.001)



zweigipflige Verteilung (W=0.944 - p=0.008)



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



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

2. 20. 2 Varianzhomogenität bei unabhängigen Stichproben

Im Fall von unabhängigen Stichproben wurde bislang in den meisten Textbüchern zum Test der Varianzhomogenität (Gleichheit der Gruppenvarianzen $\sigma_1^2 = ... = \sigma_I^2$) der *Levene-Test* empfohlen, da dieser robust gegen Abweichungen von der Normalverteilung und auch für ordinale Variablen anwendbar ist. In den letzten Jahren sind allerdings viele Vergleiche der inzwischen zahlreichen Verfahren publiziert worden, die diese Empfehlung in Frage stellen. Dazu folgender Hintergrund zum Levene-Test. Es werden zunächst die absoluten Abweichungen der beobachteten Werte x_{im} vom jeweiligen Gruppenmittelwert $\bar{x}_i$ in verschiedenen Varianten errechnet:

$$(a) \quad \left|x_{im}-\bar{x}_i\right| \qquad (b) \quad \left|x_{im}-\tilde{x}_i\right|$$

$$(c) \quad \left|x_{im}-\bar{x}_i\right|^2 \qquad (d) \quad \left|R(x_{im})-\overline{R(x_i)}\right|$$

$$(e) \quad (n_i/(n_i-1))\left|x_{im}-\bar{x}_i\right|^2$$

in der Variante (a) vom arithmentischen Mittel, in (b) vom Median, in (c) als quadrierte Abweichung vom arithmentischen Mittel und in (d) nach einer Rangtransformation. Diese Differenzen werden anschließend mit der „normalen“ Varianzanalyse für die *k* Gruppen verglichen. (Ana-

zumerken ist, dass die Anwendung einer „robusten“ Varianzanalyse (siehe 2.9) in diesem Fall sich eher kontraproduktiv auswirkt. (a) entspricht dem „klassischen“ weit verbreiteten Levene-Test. (d) entspricht (a), wenn vorher die x_{im} (über alle Gruppen hinweg) in Ränge transformiert werden. Allerdings gehört die Variante (b), die auch als *Brown-Forsythe*-Test bekannt ist und nicht mit der gleichnamigen o.a. robusten Varianzanalyse verwechselt werden sollte, noch immer zu den besten Methoden (vgl. Sharma & Kibria, 2012). Insbesondere (a) sollte nicht für ungleiche Stichprobenumfänge n_i benutzt werden. Die Variante (e) kann dieses Defizit lösen und wird in einem Test von *O'Brien* angeboten. Ungleiche n_i, zu kleine n_i<15 sowie schiefe Verteilungen beeinträchtigen allerdings auch die Varianten (b) und (c). In R sind die Varianten (a), (b) und (c) in der Funktion `levene.test` (Paket `vartest`) sowie (a) und (b) in der Funktion `leveneTest` (Paket `car`) enthalten. SPSS bietet sie in der Prozedur `oneway` an.

Aber nun zu den anderen Methoden. Die klassischen wie Fishers F-Test, Hartleys F_{max}-Test, Cochrans C-Test oder der Bartlett-Test sind überholt, weil sie alle eine Normalverteilung von x voraussetzen und nicht robust gegen Abweichungen sind. Alternative Tests werden in diversen Methoden-Vergleichen vorgestellt, insbesondere bei Farrar et al. (2025), Katsileros et al. (2024), Gaonkar & Beasley (2023), Wang et al. (2017) und Sharma & Kibria (2012), von denen die hilfreichsten hier aufgeführt werden:

- Test von *O'Brien*, der auf o.a. Differenzen (e) aufbaut, wahlweise vom arithmetischen Mittel oder vorzugsweise vom Median. Er liefert auch bei kleinen und ungleichen n_i zuverlässige Ergebnisse und wird daher von den meisten favorisiert. In R verfügbar als `obrien.test` (Paket `vartest`), leider nur 1-faktoriell.
- Test von *Keyes & Levy*, der fast identisch mit dem von *O'Brien* ist, mit gleichen Eigenschaften. In R verfügbar als `levene.test` (Paket `lawstst`) mit der Option `correction.method=correction.factor`.
- *Ramsey's Bamtest*, eine Verbesserung des Bartlett-Tests. In R verfügbar als `bamset` (Paket `skedastic`).
- Test von *Fligner & Killeen* in der Variante von *Conover and Johnson*, der auf den Differenzen (d), allerdings zum Median, basiert und damit ein nichtparametrischer Test ist. Er gilt allerdings dennoch nicht als robust, z.B. bei kleinen n_i und großem k sowie stark schiefen Verteilungen. Er ist in R verfügbar als `fligner.test` (Paket `vartest`).
- *NPL* (nichtparametrischer Levene) Test, der lediglich die rangtransformierte Variable R(x) auf den Levene-Test anwendet, vorzugsweise mit Abweichungen vom Median. Er ist robust gegen schiefe Verteilungen und Ausreißer.

An dieser Stelle eine Warnung vor den rangbasierten Tests wie *Fligner & Killeen* und *NPL*. Bei diesen können ungleiche $\overline{x_i}$ (die ja eigentlich geprüft werden sollen) sich unangenehm auswirken. Dies haben Shear et al. (2018) an eindrucksvollen Beispielen gezeigt. Man kann zwar die x_{im} auf gleiche $\overline{x_i}$ vor Prüfung der Varianzhomogenität transformieren, aber auch das wird als nicht unproblematisch angesehen.

Weitere Tests sind in R in den Paketen `vartest` sowie `skedastic` zu finden. Letzteres basiert auf dem o.a. Artikel von Farrar et al. und behandelt den Test auf Varianzhomogenität insbesondere bei linearen Modellen. Da die Varianzanalyse allerdings hiervon ein Spezialfall ist, können die dort aufgeführten Methoden auch hier angewandt werden. Die Autoren favorisieren die Methoden *Evans-King*, *Verbyla, Cook-Weisberg*, *White* und *Breusch-Pagan.* SPSS bietet hiervon den *Breusch-Pagan*- und den damit verwandten *White*-Test in der Prozedur `GLM` an, die allerdings in den Beispielen nicht berücksichtigt werden.

2. 20. 3 Varianzhomogenität bei abhängigen Stichproben

Im Fall von abhängigen Stichproben (Messwiederholungen) gibt es vergleichsweise kaum Tests auf Gleichheit der Varianzen. (Es gibt einen „klassischen“ Test zum Vergleich von nur zwei Varianzen: den *Pitman-Morgan-Test*, in R als Funktion `pitman` im Paket `ANSM5`.) Das liegt daran, dass bei parametrischen Varianzanalysen mit Messwiederholungen wesentlich mehr als die Varianzhomogenität gefordert wird, nämlich Spherizität (siehe nächsten Abschnitt). Allerdings wird bei nichtparametrischen Tests i.d. Regel „nur“ gleiche Verteilungsform für alle zu vergleichen Gruppen gefordert, was in etwa gleiche Varianzen voraussetzt. Das heißt auch: Wenn ein Test auf Spherizität negativ ausfällt, können die Varianzen als gleich angesehen werden, aber umgekehrt kann bei einem positiven Ergebnis nicht auf heterogene Varianzen geschlossen werden.

Zwei brauchbare nichtparametrische Tests sind bei R.R. Wilcox (1989) beschrieben. Zum einen ein Levene-like Test wie oben beschrieben, d.h. es werden für jede Variable die absoluten Differenzen vom Median berechnet, die dann mit einer Varianzanalyse verglichen werden, vorzugsweise mit der Friedman-Anova (vgl. Abschnitt 2.1). Der andere („Q-Methode“), etwas zuverlässigere basiert auf einem Verfahren von Quade (vgl.Abschnitt 2.11.1). Beide können in der vom Autor angebotenen Funktion `check.var` angewandt werden (siehe Anhang 3). Beispiele dazu kommen in den Abschnitten 5.3.4 und 5.4.3 (mit einem bzw. zwei Messwiederholngsfaktoren). Paarweise Varianzvergleiche sind im R-Paket `PairedData` enthalten. Damit kann man entweder alle *J* Streuungen paarweise oder die kleinste gegen die größte vergleichen.

2. 20. 4 Spherizität

Die Spherizität bei Analysen mit Messwiederholungen ist eine spezielle Form einer Kovarianzmatrix (mehr dazu in Kapitel 5.2). Sie impliziert u.a. die Gleichheit der Varianzen der Messwiederholungsvariablen sowie gleiche Korrelationen zwischen den Variablen. Zur Prüfung der Spherizität wird allgemein der *Mauchly-Test* verwendet (siehe z.B. Winer, 1981, S. 255), so auch standardmäßig in R (u.a. mit der Funktion `mauchly.test` sowie in den Paketen `ez` und `superb`) 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 Spherizität 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):

- *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 ebenfalls auf der etwas allgemeineren elliptischen Verteilung basiert (s.o.),
- ein Test auf *Compound Symmetry*, beschrieben im Buch von Winer (S. 517), auch als Funktion `WinerCompoundSymmetryTest` im Paket `superb`.

Darüber hinaus gibt es relativ neue nichtparametrische Verfahren, die auf räumlichen Vorzeichen und Rängen basieren (vgl. Abschnitt 2.16.2). Die Methode ist nicht leicht zu verstehen (kurze Einführungen z.B. bei Sirkiä, 2007, sowie Zou et al., 2014). Immerhin gibt es R-Funktionen zum Test auf Spherizität 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, und somit auch keine Empfehlung.

2. 20. 5 Homogenität der Kovarianzmatrizen

Die Homogenität der Kovarianzmatrizen ist eine wichtige Voraussetzung bei gemischten Versuchsplänen. 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) sowie in der Funktion `HOMOGENEITY` im Paket `DFA.CANCOR`. 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):

- *LR-Bartlett*-Test, basierend auf der multivariaten Normalverteilung,
- *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,
- *Fligner & Killeen*, eine Erweiterung des unter Abschnitt 2.20.2 genannten parameterfreien Tests auf Gleichheit der Varianzen (verfügbar in der o.a. R-Funktion `HOMOGENEITY`),
- multivariater *Levene-like* Test, der lediglich ordinales Skalenniveau voraussetzt, sowie
- zwei multivariate *Dispersionstests*, ebenfalls lediglich auf ordinalem Skalenniveau basierend, die ausschließlich auf unterschiedliche Varianzen ansprechen.

Bis auf die letzten drei sind alle Tests sowie einige andere in Hallin & Paindaveine (2009) beschrieben. Der Levene-like und die Dispersionstests sind bei O‘Brien (1992) skizziert. Noch einige Anmerkungen zu den Tests:

- zum Box M-Test: Zum einen gibt es eine Test-Variante, die über die χ^2-Verteilung geprüft wird, und eine Variante, die über die F-Verteilung geprüft wird. Letztere ist vorzuziehen für kleine $n_i<20$, große $I>6$ oder große $J>6$. Zum Anderen variieren die Formeln für den Test, da Box (1949) in seiner Arbeit diverse Fälle unterscheiden musste, die zu (geringfügig) unterschiedlichen Statistiken führten.
- zu Schotts Tests: Bei elliptischen Verteilungen handelt es sich um einen größeren Bereich von Verteilungen, die die Normalverteilung als Spezialfall enthalten. Die entsprechenden Verfahren haben also einen größeren Anwendungsbereich.

- Zum Levene-like Test: Ähnlich wie beim univariaten Levene-Test, bei dem die Abweichungen vom Mittelwert verglichen werden, werden hier die für jede Erhebungseinheit ermittelten Kreuzprodukte $(y_{ij_1} - m_{j_1})(y_{ij_2} - m_{j_2})$ verglichen, die die Basis für die Varianzen und Kovarianzen sind.
- Zu den Dispersionstests: Der erste vergleicht das Streuungsniveau für die Gruppen. Dieses kann für die Gruppen gleich sein, obwohl die Kovarianzmatrizen nicht gleich sind, nämlich wenn z.B. in der ersten Gruppe die Variable 1 die größere Streuung hat und die Variable 2 die kleinere, aber in der zweiten Gruppe die Variable 1 die kleinere Streuung hat und die Variable 2 die größere. Genau solche Unterschiede soll der zweite Dispersionstest erkennen.
- Zum Fligner & Killeen Test: Dieser prüft wie die o.a. Dispersionstests die Gleichheit der Streuungen für die Gruppen, während er weniger auf Unterschiede der Kovarianzen reagiert.

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: z.B. der Levene-Test. Er kann allerdings nicht ungleiche Korrelationsmatrizen erkennen. Doch im Fall ungleicher (durchschnittlicher) 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 der Grund nicht klar, ob wegen Gleichheit der Kovarianzmatrizen oder wegen Ungleichheit der Korrelationen.

2. 20. 6 Homogenität der Korrelationsmatrizen

Die Homogenität der Korrelationsmatrizen ist bei gemischten Versuchsplänen eine fast so wichtige Voraussetzung wie die o.a. Homogenität der Kovarianzmatrizen (vgl. Luepsen, 2024). Hierbei ist zu beachten, 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. Die o.a. Tests auf Homogenität der Kovarianzmatrizen sprechen im Wesentlichen auf heterogene Varianzen an. D.h. zum einen sagt ein signifikantes Ergebnis nichts darüber aus, ob sich auch die (durchschnittlicher) Korrelationskoeffizienten r_i für die Gruppen unterscheiden, und zum anderen schließt ein nicht-signifikantes Ergebnis nicht aus, dass sich die r_i für die Gruppen unterscheiden. Da aber heterogene Korrelationsmatrizen in ähnlichem Maße wie heterogene Kovarianzmatrizen zu einem starken Anstieg des Fehlers 1. Art führen können, wird dringend geraten, bei einem nicht-signifikanten Ergebnis für die Kovarianzmatrizen noch zusätzlich einen Test auf Homogenität der Korrelationsmatrizen durchzuführen.

Die Anzahl der Tests auf Homogenität der Korrelationsmatrizen ist recht begrenzt. Zu erwähnen sind:

- *Jennrich*-Test, der bekannteste, der allerdings große Stichproben (n_i>50) verlangt,
- *Larntz & Perlman*-Test, der speziell für kleinere Stichproben (n_i<40) angelegt ist,
- ein etwas abgeänderter, auf Korrelationen beschränkter *Levene-like*-Test von *O'Brien*,
- ein etwas abgeänderter, auf Korrelationen beschränkter *Box-M*-Test.

Hierfür bietet der Autor eine R-Funktion `check.corr` mit 4 verschiedenen Tests an (vgl. Anhang 3), wobei der von Larntz & Perlman (1991) der im Normalfall empfehlenswerteste und

der von O'Brien der robusteste bei schiefen Verteilungen ist. Details und Literaturhinweise zu den Tests sind auch bei Lüpsen (2020a) zu finden.

2. 20. 7 Eine Warnung

Eine Warnung soll dieses Kapitel beenden. Am Thema „Prüfung von Voraussetzungen" scheiden 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. 20. 8 Entscheidungen - die Qual der Wahl

Normalerweise wird die folgende Vorgehensweise nahegelegt: Nach Festlegung der Nullhypothese zunächst Prüfung der Voraussetzung(en) und je nach deren Ergebnis Wahl des einen oder anderen Tests zur Verifizierung der Hypothese. Doch schon bei der Wahl des Verfahrens zur Prüfung einer Voraussetzung fängt die Qual der Wahl an. Zum Test auf Gleichheit der Varianzen (im univariaten) Fall mag man sich schnell auf den Levene-Test festlegen. Aber im multivariaten Fall (bei Messwiederholungen) wird häufig vom Mauchly-Test (zum Test auf Spherizität) abgeraten. Aber welche Alternative? Man ist versucht, mehrere zu probieren. Schaut man sich deren Ergebnisse dann an, so sind diese häufig sehr konträr: von nicht signifikant (p-Wert nahe bei 1) bis stark signifikant (p-Wert nahe bei 0). (Für die Prüfung der Voraussetzungen sollte man übrigens bei kleinem $N \leq 20$ eine größere Irrtumswahrscheinlichkeit wählen, etwas α=0.10.) Und wenn die Varianzen nun als heterogen anzusehen sind, welches der möglichen Verfahren sollte man wählen? Auch dann wird es passieren, dass widersprüchliche Resultate herauskommen, wenn unterschiedliche Tests „ausprobiert" werden. Dieses Szenario ist auch ein Grund dafür, dass einige Autoren von der Prüfung der Voraussetzung(en) nach Möglichkeit abraten, wie dies an verschiedenen Stellen dieses Textes zitiert wird.

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 4.5) und SPSS (Version 25).

Im Folgenden wird vielfach von „Streuungsquadraten vom Typ I“ bzw. „vom Typ III“ gesprochen. Hierbei handelt es sich um verschiedene Berechnungs- aber letztendlich Gewichtungsmethoden für die (parametrische) Varianzanalyse. Diese werden in Abschnitt 4.3.1.1 kurz erläutert. Weitere Hinweise hierzu bieten Barron (2015) und Meyer (2008).

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 dafür die in der Programmiersprache S vorgesehene Funktion `aov`, welche die Berechnung der Streuungsquadrate vom Typ I vornimmt, eine Methode, die problematisch ist und von kaum einem anderen Programm benutzt wird (siehe o.a. Hinweis).
- 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. Zum einen solche, die auf den Standardfunktionen aufbauen und dem Benutzer das Arbeiten erleichtern sollen, z.B. die Pakete `afex` und `ez`. Zum anderen solche für speziellere Methoden, z.B. die vom Autor angebotene Bibliothek `anova.lib` (siehe Anhang 3). Von denen können hier nur wenige erwähnt werden, vornehmlich zusammen mit den Beispielen in den folgenden Kapiteln.

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.

Nachfolgend die wichtigsten Funktionen zur Varianzanalyse.
Vorsicht: Es gibt 2 verschiedene Funktionen `Anova` und `anova`.

- `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 einer der folgenden Schritt erforderlich:
  `drop1 (`*`model`*`, ~. , test="F")` oder
  `Anova (`*`model`*`, type="III")` mit `Anova` (im Paket `car`)
  wobei *`model`* das Ergebnis von `aov` enthält.

- `lm` zusammen mit `Anova` (im Paket `car`)
  `Anova (lm (abh.Variable ~ Faktor1*Faktor2*..., Dataframe), type="III")`
  für unabhängige Stichproben
  Vorteil gegenüber `aov`: Direkte Wahl der Quadratsummen-Berechnung (`type`), und die Ergebnisse, wie z.B. die Quadratsummen lassen sich als Dataframe weiterverarbeiten, was vielfach erforderlich ist.

- `lm` zusammen 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.

- `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.

- `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, Welch sowie zahlreiche andere Tests.

- `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.

- in der `anova.lib` u.a.:
  `np.anova, koch.anova, ap.anova, ats.2, ats.3, art1.anova, art2.anova`
  mehrfaktorielle nichtparametrische Varianzanalysen
  `bf.f, mbf.f, iga, wj.anova, wj.spanova`
  mehrfaktorielle robuste Varianzanalysen

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“ oder „metrisch“ gesetzt wird, insbesondere bei der Anwendung nichtparametrischer Verfahren. Zur Verfügung stehen:

- `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` *`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` *`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 Kovarianzamatrizen angefordert werden.
  Unter „Modell“ kann die Methode zur Berechnung der Streuungsquadrate gewählt werden (Typ I, II oder III).
  Mauchlys Test auf Spherizität 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 `0.000001` , 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 (ART) 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. Einzig die GLMM-Methode erlaubt (generell) fehlende Werte, sowie die ATS-Methode in R bei einigen der Funktionen.

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

R: 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 Kriterum) 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. Dazu gehören insbesondere die in den Kapiteln 2.1 bis 2.7 vorgestellten Verfahren. 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 meistens 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 der 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`. Sind z.B. `v11, v12, v13,..` und `v21, v22, v23,..` die Messwiederholungen der 1. bzw. 2. Stufe von Faktor C, sowie `v11, v21, v31,..` und `v12, v22, v32,..` die Messwiederholungen der 1. bzw. 2. Stufe von Faktor D, dann sind für die Analyse von Faktor C erforderlich

```
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. Entsprechend für die Analyse von Faktor D

```
compute t1=sum(v11, v21, v31,...).
compute t2=sum(v12, v22, v32,...).
....
```

In R muss unterschieden werden, ob die Funktion die Daten im *breiten* Format oder im *langen* Format benötigt. Genaueres 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 multivariaterTest, 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 (i=1,...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_i=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 `drugs` 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: drugs 1, 2, 3 oder 4) erhalten. Die Zellenbestzungszahlen sind in diesem Datensatz ungleich. Die abhängige Variable ist eine Beurteilung auf einer Skala von 1 bis 9, also ordinal.

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

In R wie auch in SPSS werden hierfür die Variablennamen `group`, `drugs` und `x` verwendet. In R müssen `group` und `drugs` 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 verschieden hohen Dosierungen (Faktor B: dosis 1, 2, 3 oder 4) erhalten. Die Zellenbestzungszahlen sind in diesem Datensatz ungleich. Die abhängige Variable ist eine Beurteilung der Reaktion auf einer Skala von 1 bis 20. Durch Abbruch der Therapie kommt es hier zu unterschiedlichen n_i. Das Skalenniveau ist dasselbe wie im ersten Beispiel, also quasi metrisch, wenn auch streng genommen als Beurteilung ordinal.

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

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

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. Doch insbesondere die Normalverteilungsvoraussetzung kann auch anders formuliert werden: Die Residuen e_{im} müssen normalverteilt sein, wobei sich die Residuen aus dem varianzanalytischen Modell (vgl. o.a. Formel 2-0) ergeben, hier für den 1-faktoriellen Fall eines Faktors A mit *I* Stufen/Gruppen:

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

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

$$\mu_1 = \mu_2 = \ldots = \mu_I$$

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

$$\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$$

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:

$$F_1(x) = F_2(x) = ... = F_I(x)$$

woraus zu folgern ist, dass bei Ablehnung von H_0 sich nicht notwendigerweise die Mittelwerte μ_i unterscheiden, sondern möglicherweise andere Verteilungsparameter wie Varianz oder Schiefe. Deswegen wird als Voraussetzung gefordert, dass alle F_i dieselbe Form haben.

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 β_j:

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

(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 Normalverteilug sind in Kapitel 2.20.1 zu finden.

Die Varianzhomogenität ($\sigma_1^2 = ... = \sigma_I^2$) wird üblicherweise mit dem *Levene-Test* überprüft (vgl. Kapitel 2.20.2 mit Hinweisen auf weitere Tests), wenn auch inzwischen der Test von *O'Brien* vorgezogen wird. Hierbei wird empfohlen, die Abweichungen vom Median anstatt vom arithmetischen Mittel zu wählen. Es ist zu beachten, dass bei mehrfaktoriellen Analysen für jeden Effekttest 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 β (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 nichtnormalverteilte 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. Akzeptabelsind 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 μ als gewichtetes Mittel der Gruppenmittelwerte μ_i angenommen:

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

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:

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

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 De-

signs 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 notwenigerweise 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 β. 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 (Wölbung), machen den F-Test konservativer. Breitgipflige Verteilungen machen dagegen den Test liberaler, können aber auch das α-Risiko vergrößern, wenn auch 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. Sie 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.

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

  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 in der Regel 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 größeren Stichproben die kleineren Varianzen (*negative pairing)*, reagiert der F-Test liberal (vgl. u.a. Luepsen, 2018, 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 β. 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 $\max(n_i)/\min(n_i) < 4$.
- 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 Varianzheterogenitäten betroffen als die Haupteffekte.
- Bei der Anwendung rangbasierter nichtparametrischer Verfahren bleibt in der Regel eine Varianzheterogenität 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 manchen der in Kapitel 2.9 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 perfektionieren diese Idee (vgl. https://onlinestatbook.com/2/transformations/box-cox.html). Auf der anderen Seite warnen Feng et al. (2014) insbesondere vor der log-Transformation.
- In Kapitel 2.9 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.
- Box (1954) hat 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) eingehend 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. In Kapitel 1.1.7 war dazu ein Beispiel aufgeführt.

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 Würfelexperiments. Die Augenzahl (1,..,6) ist gleichverteilt. Würfelt man n-mal und berechnet den Mittelwert der jeweils erzielten Augenzahl, so kann man beobachten, wie dieser gegen 3.5 konvergiert. Und ein Histogramm der Mittelwerte zeigt, wie sich die Verteilung der Mittelwerte 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 χ^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). Allerdings gelten der Levene-Test sowie der Test von O'Brien auf Basis der Mediane als relativ robust und zuverlässig.
- 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.
- 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, selbst nicht signifikant verschiedenen Varianzen zu verfälschten Ergebnissen führen. Daher sollte die Varianzhomogenität und ein Pairing geprüft werden, und es ist in solchen Fällen eine Varianzanalyse für heterogene Varianzen sofern möglich in Erwägung zu ziehen. Im Fall von Pairing $|r|>0.5$, insbesondere bei $|r|>0.8$, ist es generell ratsam, Verfahren für heterogene Varianzen (vgl. Kapitel 2.9) anzuwenden.
- Liegt kein Zusammenhang zwischen den n_i und s_i (Pairing) vor, kann entweder der van der Waerden-Test (insbesondere bei $n_i > 10$) oder das INT-Verfahren ($n_i \leq 10$) verwendet werden (vgl. Lüpsen, 2018 und Dijkstra, 1987).
- Alternativ sind im Fall einer ordinalen abhängigen Variablen der van der Waerden-Test oder im Fall einer extremen Verteilung, z.B. mit Ausreißern, robuste Verfahren (Kapitel 4.3.4) vorzuziehen.

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 bietet derzeit nur R Verfahren, 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 halten zwar den Fehler 1. Art besser unter Kontrolle, gelten 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 INT noch die am besten geeigneten.

Beispiele zur Prüfung der Voraussetzungen für die parametrische Varianzanalyse 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: alle Werte werden in Ränge 1,...,*N* transformiert, so dass letztlich anstatt der Mittelwerte die mittleren Rangsummen verglichen werden. Für den Test wird ein Wert *H* errechnet, der χ^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 signikant unterscheiden.

mit SPSS:

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

mit folgender Ausgabe:

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

Übersicht über Hypothesentest

	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.

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. Wie in Kapitel 2.9 aufgeführt gibt es eine Reihe Methoden für den Fall ungleicher Varianzen. Trivialerweise dürfen diese Tests auch angewandt werden, wenn die Varianzen homogen sind. Allerdings ist in der Regel damit ein mehr oder weniger großer Verlust an Power verbunden. Die bekanntesten sind die Tests von Welch bzw. von Brown & Forsythe, wovon letzterer der neuere und bessere ist. Beide Tests sind in R (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 einfaktoriell der Faktor `dosis` untersucht.

mit R:

Zunächst die Prüfung der Varianzhomogenität mittels des Levene-Tests aus dem Paket `car`:

```
library(car)
leveneTest(x~dosis,center=median,data=mydata3)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3  4.1561 0.01448 *
      29
```

Infolge der signifikanten Inhomogenität ist anstatt des normalen F-Tests ein dafür geeigneter robuster F-Test zu wählen. Der Welch-Test ist durchführbar über die Funktion `welch.test`, alternativ ein für kleine Stichproben adjustierter Welch-Test über `aw.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:

```
library(onewaytests)
welch.test(x~dosis, mydata3)
bf.test(x~dosis, mydata3)
```

```
        Welch's Heteroscedastic F Test

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

  Brown-Forsythe Test (alpha = 0.05)

  statistic  : 3.217746
  num df     : 3
  denom df   : 18.61813
  p.value    : 0.04654974
```

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

```
  Alexander-Govern Test (alpha = 0.05)

  statistic  : 8.782242
  parameter  : 3
  p.value    : 0.03233067
  Result     : Difference is statistically significant.

  James Second Order Test (alpha = 0.05)

  statistic     : 12.80273
  criticalValue : 11.38424
  Result        : Difference is statistically significant.
```

Beim James-Test wird die Teststatistik (`12.80`) zusammen mit dem kritischen Wert (`11.38`) 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 zunächst der Levene-Test auf Homogenität der Varianzen in 4 Varianten, wobei die zweite, auf dem Median basierende, vorzuziehen ist, darunter das Ergebnis für homogene Varianzen:

Levene-Statistik		df1	df2	Sig.
Basiert auf dem Mittelwert	4.965	3	29	.007
Basiert auf dem Median	4.156	3	29	.014
Basierend auf dem Median und mit angepaßten df	4.156	3	17.479	.022
Basiert auf dem getrimmten Mittel	5.076	3	29	.006

	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[a]	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. In Abschnitt 4.3.4 werden einige Methoden vorgestellt, allerdings nur für R. Auf zwei allgemein anwendbare Lösungen 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 wiederspiegelt, wie dies üblicherweise auch bei den entsprechenden modfizierten 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 & Mirosevich (2008) erinnern an die Methoden der *Winsorisierung* und des *Trimmens*, die relativ selten angewandt werden, weil sie den Verdacht der Datenmanipulation aufkommen lassen, die aber statistisch durchaus sinnvoll sind. Hierbei werden ein fester Prozentsatz der größten und kleinsten Werte einer Variablen durch die nächstkleinere bzw. durch die nächstgrößere ersetzt (*Winsorisierung)* oder komplett eliminiert (*Trimmen*). 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`. Unter anderem hat Wilcox auf dieser Basis Varianz-analysen entwickelt, die in R verfügbar sind (siehe Abschnitte 4.3.4 und 5.3.11).

4. 2. 4 Weitere Verfahren

Die nachfolgend für die 2-faktorielle Varianzanalyse beschriebenen Verfahren sind fast alle auch als 1-faktorielle Analyse einsetzbar. Lediglich macht das ART-Verfahren nur im mehrfaktoriellen Design Sinn. Die ATS von Akritas & 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 Zellenbestzungszahlen zeilenweise oder spaltenweise proportional zueinander, z.B. bei einem Versuchsplan mit den Faktoren A (4 Stufen) und B (3 Stufen)

	B_1	B_2	B_3
A_1	10	12	16
A_2	15	18	24
A_3	20	24	32
A_4	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 nicht-balancierten 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 Faktors, 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 in Erscheinung tritt, dass dieser in Zusammenhang mit einer anderen Einflussgröße analysiert wird.

Im mathematischen Modell für die 2-faktorielle Varianzanalyse

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

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

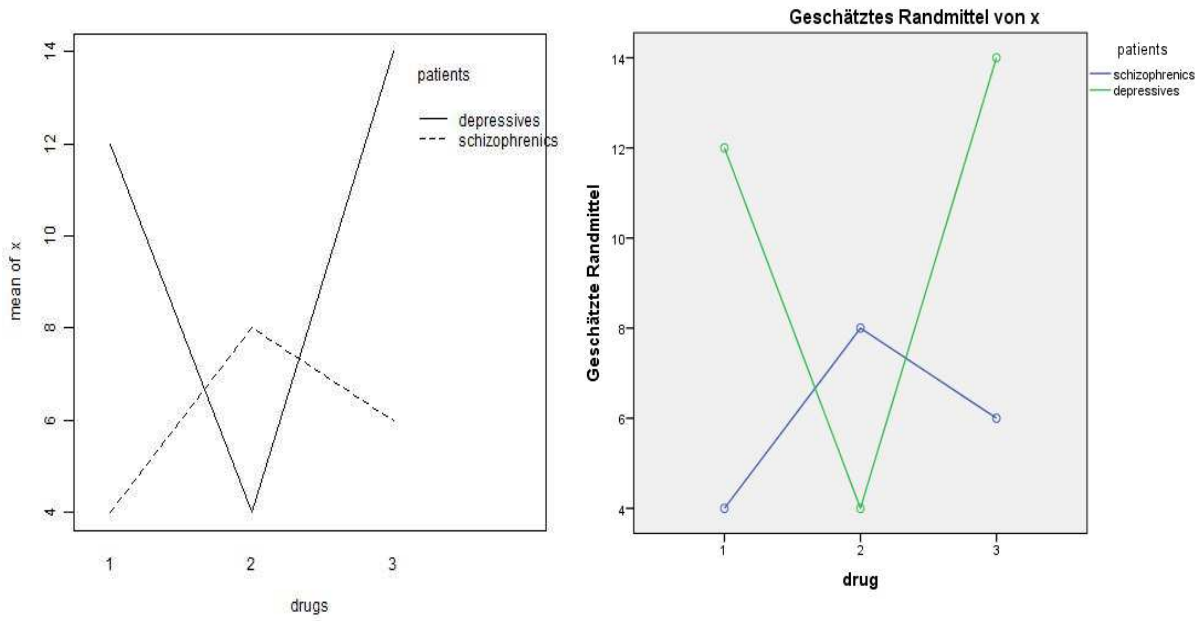
H_A: $\alpha_i = 0$ für i=1,..,*I* (entspricht $\mu_{A1} = \mu_{A2} = \ldots = \mu_{AI}$)

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

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

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

d.h. sowohl die durch A erklärten Abweichungen α_i sind für alle Stufen von B gleich groß als auch die durch B erklärten Abweichungen β_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 Varianzanalyse 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 Aufsplittung 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 Teststärke bedeutet.

Ausnahme: Falls ein hinzugenommener Faktor 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 einnmal 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 α_i ist derselbe für alle Stufen *j* von B. Gleichermaßen ist der Effekt von B β_j 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 gleichgroß. 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 einam anderen Skript *„Multiple Mittelwertvergleiche - parametrisch und nichtparametrisch“* (Lüpsen, 2020b) 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:

```
options (contrasts=c("contr.sum","contr.poly"))
mydata1 <- within(mydata1,{drugs<-factor(drugs);
                patients<-factor(patients)})
aov1    <- aov(x~patients*drugs,mydata1)
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` , alternativ `residuals(aov1)`, gewonnen werden. Der Shapiro-Wilk-Test und der Levene-Test zur Prüfung der Homogenität der Varianzen (wobei inzwischen standardmäßig die Differenzen zum Median berechnet werden, andernfalls über den Zusatz `center=median`) können über folgende Anweisungen erfolgen:

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

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 = median)
      Df F value Pr(>F)
group  5  0.4377  0.814
      12

Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  3.4299 0.08257 .
      16

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.1163 0.3532
      15
```

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

Nachfolgend nun die Varianzanalyse für die Beispieldaten 2. Da es sich dabei um einen nicht-balancierten Versuchsplan handelt, weichen die erforderlichen Kommandos von den oben aufgeführten etwas ab. Dazu bietet R u.a. die folgenden Lösungswege:

1) Funktion `aov` kombiniert mit `Anova` aus dem Paket `car`

2) Funktion `lm` kombiniert mit `Anova` aus dem Paket `car`

3) Funktion `aov` kombiniert mit `drop1`

4) Funktionen `aov_ez` oder `aov_car` aus dem Paket `afex`

5) Funktionen `ezANOVA` aus dem Paket `ez`

Diese werden nun in dieser Reihenfolge vorgestellt. Zunächst die allgemein erforderlichen Anweisungen:

```
options (contrasts=c("contr.sum","contr.poly"))
mydata2 <- within(mydata2,{drugs<-factor(drugs); group<-factor(group)})
```

Lösung 1

```
library(car)
Anova( aov( x~group*drugs,mydata2), type="III")       # Lösung 1
```

```
Anova Table (Type III tests)
             Sum Sq Df F value    Pr(>F)
(Intercept) 100.000  1 61.0998 3.577e-08 ***
group         9.333  1  5.7026   0.02481 *
drugs        46.560  3  9.4827 0.0002319 ***
group:drugs  17.932  3  3.6521   0.02604 *
Residuals    40.917 25
```

Tabelle 4-2a

Lösung 2

```
library(car)
Anova( lm( x~group*drugs,mydata2), type="III")        # Lösung 2
```

Die Ausgabe ist wegen der in beiden Fällen benutzten Funktion `Anova` identisch mit der von Lösung 1.

Lösung 3

```
drop1( aov( x~group*drugs,mydata2), ~. , test="F")    # Lösung 3
```

```
x ~ group * drugs
            Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                   40.917 23.096
group        1     9.333 50.250 27.877  5.7026 0.0248129 *
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 4-2b

Lösung 4
Ein Vorteil: unbalancierte Versuchspläne müssen nicht gesondert behandelt werden. Ein Nachteil: auch bei Analysen ohne Messwiederholungen muss eine Fallkennung vorhanden sein. Eine entsprechende Variable, hier `Vpn`, kann leicht erzeugt werden.

```
library(afex)                                          # Lösung 4
mydata2 <- cbind(mydata2, Vpn=factor(1:33))
aov_ez(data=mydata2, id="Vpn", dv="x", between=c("group","drugs"))
```

```
Anova Table (Type 3 tests)
       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
```

Tabelle 4-2c

alternativ mit der Funktion `aov_car` aus dem Paket `afex` mit derselben Ausgabe:

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

Lösung 5
Zwei Vorteile: unbalancierte Versuchspläne müssen nicht gesondert behandelt werden, und der Levene-Terst auf Gleichheit der Varianzen wird automatisch ausgegeben. Ein Nachteil: auch bei Analysen ohne Messwiederholungen muss eine Fallkennung vorhanden sein. Eine entsprechende Variable, hier `Vpn`, kann leicht erzeugt werden.

```
library(ez)                                          # Lösung 5
mydata2 <- cbind(mydata2, Vpn=factor(1:33))
ezANOVA(mydata2, x, Vpn, between=.(group,drugs), type=3 )
```

```
Warnung: Data is unbalanced (unequal N per group). Make sure you
specified a well-considered value for the type argument to ezANOVA().
$ANOVA
       Effect DFn DFd        F            p p<.05       ges
2       group   1  25 5.702648 0.0248129418     * 0.1857380
3       drugs   3  25 9.482730 0.0002318871     * 0.5322573
4 group:drugs   3  25 3.652053 0.0260398963     * 0.3047089
```

```
$`Levene's Test for Homogeneity of Variance`
  DFn DFd      SSn      SSd        F         p p<.05
1   7  25 4.125758 17.11667 0.860848 0.5498155
```

Anmerkungen:

- Bei den Lösungen mit `aov_ez` und `ezANOVA` enthält die Spalte `ges` ein Eta^2 als Schätzung für die Effektgröße. Unter der Anova-Tabelle wird noch Levenes Test auf Gleichheit der Varianzen ausgegeben.
- Die Lösungen 1 und 2 erlauben bequem die Weiterverarbeitung der Ergebnisse, was später erforderlich sein wird.
- Die Residuen, z.B. zum Test auf Normalverteilung, erhält man über `residuals(..)` angewandt auf das Ergebnisobjekt der Funktion `aov`, z.B.
  `residuals( aov( x~gruppe*dosis,mydata2))`
- `aov_ez` und `aov_car` liefern vereinzelt (allerdings sehr selten) falsche Ergebnisse

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
```

Zum Vergleich noch die Überprüfung der Varianzhomogenität mit dem empfohlenen Test von O‘ Brien, der allerdings nur die Haupteffekte möglich ist. Die Ergebnisse sind hier allerdings qualitativ gleich:

```
obrien.test(x~group,mydata2)
obrien.test(x~drugs,mydata2)
```

```
data:  x and group
F = 7.3812, num df = 1, denom df = 31, p-value = 0.01068

data:  x and drugs
F = 0.15489, num df = 3, denom df = 29, p-value = 0.9257
```

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

Es sei noch darauf aufmerksam gemacht, dass das Paket `performance` eine Reihe von Voraussetzungen prüft, u.a. Normalität und Homogeneität. Die Basis ist meistens ein zuvor mit `aov_ez` (aus dem `afex`-Paket) erstellten Anova-Ergebnis, verschiedentlich allerdings ein mit `aov` erstelltes Anova-Ergebnis. Welche `class` das Ergebnis haben muss, geht nicht

immer eindeutig aus der Dokumentation hervor. Aufbauend auf der o.a. Lösung 4 für den Datensatz `mydata2`, mit einer recht dürftigen Ausgabe:

```
library(afex)
erg <- aov_ez(data=mydata2,id="Vpn",dv="x",between=c("group","drugs"))
check_homogeneity(erg, method = "levene")
aov2<-aov(x~group*drugs,data=mydata2)
check_normality(aov2,effects="fixed")
```

```
OK: There is not clear evidence for different variances across groups
(Levene's Test, p = 0.314).

OK: residuals appear as normally distributed (p = 0.535).
```

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 = homogenity
  /design = patients drugs patients*drugs.
Oneway x by patients
  /statistics homogenity.
Oneway x by drugs
  /statistics homogenity.
```

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	264,000[a]	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[a]			
F	df1	df2	Sig.
,438	5	12	,814

sowie der Prüfung der Varianzhomogenität bzgl. der beiden Haupteffekte mittels `Oneway`:

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

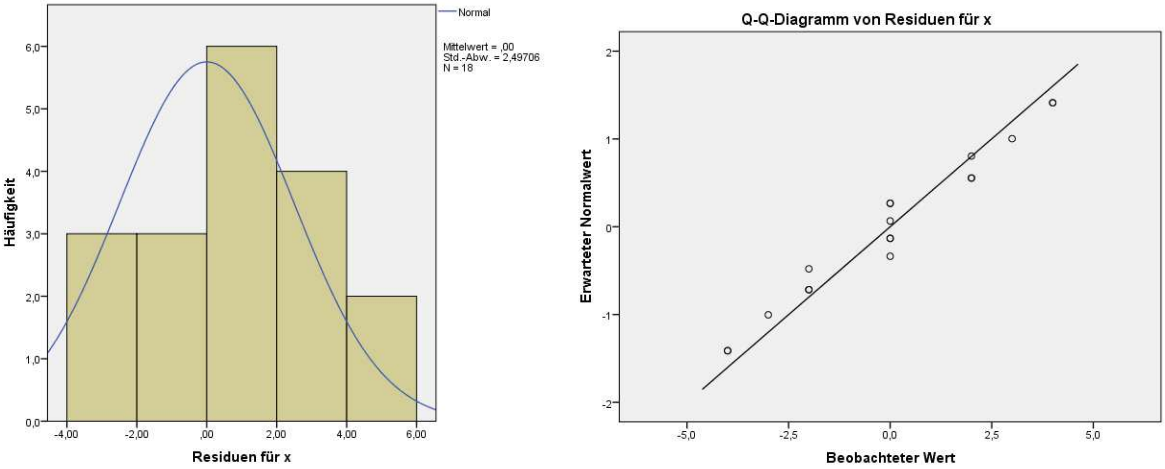
Test der Homogenität der Varianzen			
Levene-Statistik	df1	df2	Signifikanz
1,132	2	15	,348

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-Wilk-Tests. Beides zusammen kann man über das Menü „Deskriptive Statistiken -> Explorative Datenanalyse“ erzeugen. Die SPSS-Syntax dazu:

```
Examine  variables=ZRE_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 2.20.1). 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[a]			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[a]	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[a]			
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 signfikant 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).

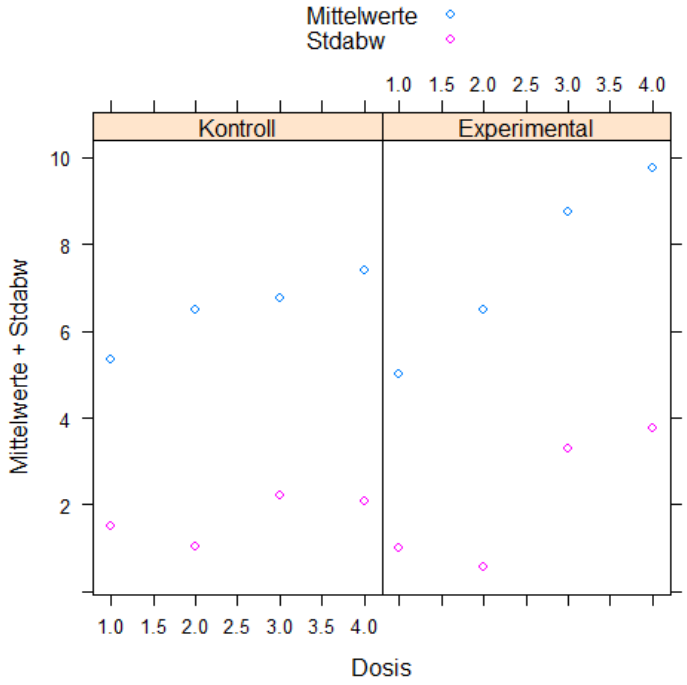
An dieser Stelle soll noch einmal auf die Ausführungen des Kapitels 4.3.1.3 zurückgekommen werden. Dort war darauf hingewiesen worden, dass durch die Hinzunahme eines Faktors häufig der statistische Fehler reduziert werden kann und Effekte erst bei mehrfaktoriellen Analysen als signifkant nachgewiesen werden können. Aus Tabelle 4-3 (Beispieldaten 1) konnten signifkante 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 2.9 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 ob ige Grafik, dass bei diesem tatsächlich die Streuungen 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. Vorab zu Vergleichszwecken die Ergebnisse auf Basis der “normalen“ Varianzanalyse:

```
              Sum Sq Df  F value    Pr(>F)
gruppe          7.93  1   1.6446   0.21146
dosis          48.57  3   3.3559   0.03476 *
gruppe:dosis   11.07  3   0.7650   0.52440
Residuals     120.62 25
```

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), sowie (zum Vergleich) des Tests von O‘Brien:

Effekt	Levene-Test ohne Transf.	Levene-Test mit Rangtransf.	Levene-Test mit normal score-Transf.	O‘Brien Test ohne Transf.
gruppe	0.1112	0.39388	0.33077	0.03883
dosis	0.01448	0.054783	0.17853	0.05244
gruppe*dosis	0.02306	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(mydata3,tapply(x,list(gruppe,dosis),sd))
ni <- with(mydata3,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 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:

```
attach("path/anova.lib")
box.f(x~gruppe*dosis, mydata3)
bf.f(x~gruppe*dosis, mydata3)
wj.anova(mydata3, "x", "gruppe", "dosis", Ftest=T)
```

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 21.326   9.116  9.1162  1.4458 0.24239
dosis        2.135 18.618  45.922 15.3074  3.2354 0.05953 .
gruppe:dosis 3.000 12.422  11.072  3.6908  0.7499 0.54246
Residuals   25.000        120.617  4.8247
```

Im Gegensatz zu den anderen Verfahren gibt es den Test von Welch & James in zwei Varianten: zum einen in einer „klassischen“ Variante basierend auf der χ^2-Verteilung, zum anderen eine neuere Approximation mittels eines F-Tests. Die Auswahl erfolgt über den Parameter `Ftest=...`.

```
                        F   df1       df2  P(F>value)
gruppe          1.6532692     1 12.258021  0.22227358
dosis           3.3849945     3  8.548393  0.07059026
gruppe : dosis 0.6497985     3  8.548393  0.60362853
```

Wie zu sehen ist, differieren die Resultate kaum. Für die Ergebnisse der Varianzanalyse mit der transformierten Variable x sei auf den Abschnitt „SPSS“ verwiesen.

4. 3. 3. 2 BDM-Tests (Brunner, Dette und Munk)

mit R:

Der BDM-Test in der nichtparametrischen Version ist im Paket `asbio` u.a. als Funktion `BDM.2way` für eine 2-faktorielle Varianzanalyse bzw. `BDM.test` für eine 1-faktorielle 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 heterogene Varianzen.

Der BDM-Test in der parametrischen Version ist im Paket `GFD` als Funktion `GFD` für eine robuste Varianzanalyse enthalten. Dieser hält das α-Risiko besser als die o.a. nichtparametrische Version, besonders im Fall von schiefen Verteilungen sowie in Pairing-Situationen. Nachfolgend ein Beispiel mit demselben oben benutzten Datensatz, das qualitativ dieselben Ergebnisse liefert:

```
GFD(x~gruppe*dosis,mydata3)$ATS
```

```
               Test statistic      df1      df2    p-value
  gruppe            1.6532692 1.000000 12.25802 0.22227358
  dosis             3.6367026 2.211855 12.25802 0.05408544
  gruppe:dosis      0.7651647 2.211855 12.25802 0.49851443
```

4. 3. 3. 3 Verfahren von Ananda und Weerahandi

mit R:

Die Verfahren von Ananda und Weerahandi haben sich als sehr robust gegen heterogene Varianzen u.a. in Pairing-Situationen erwiesen, verhalten sich allerdings sehr liberal im Fall von schiefen Verteilungen. Sie sind in der Funktion `gpTwoWay` (Paket `twowaytests`) verfügbar. Die beiden Methoden gPB und gPQ unterscheiden sich in der Praxis kaum. Auch hier ein Beispiel mit dem Datensatz `mydata3`. (Der Zusatz `verbose=T` ist zur Ausgabe des Ergebnisses erforderlich.)

```
gpTwoWay(x~gruppe*dosis,mydata3,verbose=T).
```

```
  Generalized p-value by PB method (alpha = 0.05)
-----------------------------------------------
       Factor P.value     Result
       gruppe  0.7508 Not reject
        dosis  0.2927 Not reject
 gruppe:dosis  0.6096 Not reject
```

4. 3. 3. 4 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 `lnx=ln(x)` ergibt:

Abhängige Variable: lnx					
Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
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			

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.

Insgesamt hat sich gezeigt, dass der bei der „normalen“ Varianzanalyse signifikante Dosis-Effekt bei allen o.a. Methoden für heterogene Varianzen nicht als signfikant nachgewiesen werden konnte. Lediglich die log-Transformation *log(x)* konnte den Einfluss von Dosis bestätigen.

4. 3. 4 Verteilungs-robuste Varianzanalysen

Während im vorigen Abschnitt solche Verfahren vorgestellt wurden, die speziell für den Fall heterogener Varianzen konzipiert sind, geht es hier um solche, die vorzugsweise den Fall nichtnormalverteilter Residuen berücksichtigen, z.B. durch Bewertung der Residuen (vgl. Kapitel 2.8). Da Ausreißer sich auf die Streuungen auswirken, können die hier vorgestellten Methoden auch im Fall ungleicher Varianzen eine Lösung bieten. Die abhängige Variable sollte metrisch sein. Vergleiche der verschiedenen Verfahren sind nicht bekannt. Das Prüfen der Varianzhomogenität sowie der Normalverteilung der Residuen erübrigt sich hier. Es sei darauf aufmerksam gemacht, dass die im Folgenden benutzte Funktion `rob.anova` auf Methoden basiert, die iterativ eine Lösung suchen. Dadurch kann es passieren, dass mitunter Warnmeldungen ausgegeben werden oder das Programm sogar abbricht. Da `rob.anova` auf Funktionen anderer Pakete zugreift, müssen diese vorher installiert werden (siehe dazu u.a. Listen).

4. 3. 4. 1 Robuste Mittelwerte

Anstatt arithmetischer Mittel und Varianzen werden robuste Mittelwerte und winsorisierte Varianzen verwendet. Die Methodik wurde in Abschnitt 2.8 kurz skizziert. Es gibt verschiedene Ansätze für die Berechnung (Verteilungs-) robuster Mittelwerte:

- getrimmte Mittelwerte, d.h. ein fester Prozentsatz der kleinsten und größten Werte von *x* wird weggelassen,
- robuste Mittelwerte auf Basis von M-Schätzern (maximum likelihood type), u.a. *Hubers psi*, *one-step* und *modified one-step*,
- der Median, oder allgemein Quantile.

Mehr Informationen hierzu bieten Mair & Wilcox (2019) sowie detailierter Wilcox (2017).

mit R:

Wilcox (siehe Mair & Wilcox, 2019) bietet dazu das Paket `WRS2` einerseits mit den Funktionen `t1way`, `t2way` und `t3way` für 1-, 2- bzw. 3-faktorielle Analysen unter Verwendung getrimmter Mittelwerte an. Der Anteil der getrimmten (gestutzten) Werte (standardmäßig 0.2, jeweils 0.2 Prozent am unteren und oberen Ende) kann über den Parameter `tr=...` bestimmt werden. Es wird wie im vorigen Abschnitt der Datensatz `mydata3` verwendet.

```
library(WRS2)
t2way(x~gruppe*dosis, mydata3)
```

```
              value p.value
gruppe        1.6073   0.227
dosis        11.0988   0.064
gruppe:dosis  2.1815   0.613
```

Wilcox bietet zum anderen eine weitere Funktion `pbad2way` an, bei der Mittelwerte auf Basis von M-Schätzern als Gruppenmittel verwendet werden. Das Schätzverfahren wird über den Parameter est=... ("`onestep`", "`mom`" oder "`median`") angegeben.

```
library(WRS2)
pbad2way(x~gruppe*dosis, mydata3, est="mom")
```

```
             p.value
gruppe        0.3122
dosis         0.0618
gruppe:dosis  0.7262
```

Die Ausgabe ist bei allen Varianten recht dürftig, und Erläuterungen zu `value` gibt es nicht.

Zwei der o.a. Methoden werden im Paket `twowaystats` angeboten: getrimmte Mittelwerte in der Funktion `tmeanTwoWay`, bei der über den Paramer `tr=...` der Anteil der zu stutzenden Werte (jeweils oben und unten) vorgegeben werden kann, und Mediane anstatt arithmetischer Mittel in der Funktion `medTwoWay`. Als Beispiel wird wieder Datensatz `mydata3` verwendet. Nachfolgend Ein- und Ausgabe für beide Funktionen

```
library(twowaystats)
tmeanTwoWay(x~gruppe*dosis,mydata3,tr=0.1,verbose=T)
medTwoWay(x~gruppe*dosis,mydata3,verbose=T)
```

```
  Two-way ANOVA for Trimmed Mean (alpha = 0.05)
------------------------------------------------------------------
       Factor Statistic P.value     Result
       gruppe  1.699902   0.214 Not reject
        dosis 13.610346   0.030     Reject
 gruppe:dosis  2.321031   0.578 Not reject

  Two-way ANOVA for Median (alpha = 0.05)
------------------------------------------------------------------
       Factor       df Statistic    P.value     Result
       gruppe F(1,Inf)  3.268422 0.07062553 Not reject
        dosis F(3,Inf)  6.267329 0.00030042     Reject
 gruppe:dosis Chisq(3)  3.360148 0.33935850 Not reject
```

4. 3. 4. 2 Robuste Residuen

Bei diesen Verfahren werden in einem Iterationsverfahren die Residuen der zugrunde liegenden Regression gewichtet.

mit R:

Hierzu gibt es die Funktion `rob.anova` innerhalb der `anova.lib` (vgl. Anhang 3). Diese greift auf Funktionen aus anderen Paketen zu, die vorher installiert sein müssen. Hierzu gibt es u.a. die folgenden Optionen:

- *Robust Fitting using Yohai's MM estimates*, basierend auf Yohai (1987) (`method="MM"`) Paket `robustbase` zu installieren,
- *Robust Fitting using Bisquare Psi Function*, basierend auf Yohai (1987) (`method="bi"`) Paket `robustreg` zu installieren, und
- *Robust Fitting using Huber Psi Function,* basierend auf Yohai (1987) (`method="Huber"`) Paket `robustreg` zu installieren,
- *Robust Fitting using Huber Psi Function,* basierend auf Huber (1981) (`method="M"`) Paket `MASS` zu installieren,

von denen die Benutzung letzterer kurz gezeigt wird:

```
attach("path/anova.lib")
rob.anova(x~gruppe*dosis, mydata3, method="M")
```

```
Robust Regression with Huber Function
             Df  Sum Sq Mean Sq F value  Pr(>F)
gruppe        1   3.566   3.566   0.734 0.39973
dosis         3  37.628  12.543   2.583 0.07583 .
gruppe:dosis  3  21.692   7.231   1.489 0.24170
Residuals    25 121.388   4.856
```

Eigene Simulationen haben gezeigt, dass das Verfahren von Wilcox das zuverlässigste ist, insbesondere das α-Risko weitgehend halten kann. Die Methoden `Huber` und `M` basieren beide auf Hubers M-Funktion, erzielen aber wegen unterschiedlicher Ansätze verschiedene Ergebniss: im Fall von annähernd normalverteilten Daten verhält sich die Methode `Huber` relativ liberal und `M` behält die Kontrolle, während bei starken Ausreißern es genau umgekehrt ist. Das `bi`-Verfahren ist beim Test der Interaktion sehr liberal. Die Verfahren `bi` und `Huber` reagieren liberal auf heterogene Varianzen. Daher sollte ein Verfahren für heterogene Varianzen den robusten vorgezogen werden, wenn die größere Streuung in manchen Gruppen erklärbar ist.

4. 3. 5 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. den v.d.Waerden-Test, der in den folgenden Kapiteln vorgestellt wird. Für die beiden nachfolgend benutzten Datensätze erübrigt sich dies allerdings, da in Kapitel 4.3.2 für diese keine Varianzinhomogenitäten 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:

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

```
options (contrasts=c("contr.sum","contr.poly"))
mydata2 <- within(mydata2,{drugs<-factor(drugs);
                group<-factor(group); rx<-rank(x)})
Anova( aov(rx~group*drugs,mydata2),type="III")
```

mit der Ausgabe:

```
Anova Table (Type III tests)

Response: Rx
             Sum Sq Df  F value    Pr(>F)
(Intercept) 8537.4   1 196.9344 2.339e-13 ***
group        364.2   1   8.4003 0.0076982 **
drugs       1157.72  3   8.9018 0.0003464 ***
group:drugs  464.61  3   3.5724 0.0281287 *
Residuals   1083.79 25
```

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 `Rx` resultiert. Die Varianzanalyse für Rx:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	346,167[a]	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		
Korrigierte Gesamtvariation	475,500	17			

Tabelle 4-7

Für das o.a. zweite Beispiel:

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	1820,713[a]	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		
Korrigierte Gesamtvariation	2904,500	32			

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. 6 Puri & Sen-Test (Verallgemeinerte Kruskal-Wallis-Analysen)

Dieses Verfahren geht 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 χ^2-Tests konstruiert. Dieses ist als Verallgemeinerung des Kruskal-Wallis-H-Test anzusehen, da es im 1-faktoriellen Fall mit letzterem identisch ist, sowie bei mehrfaktoriellen Designs (ohne Messwiederholungen) mit dem KWF-Verfahren.

Die χ^2-Werte haben den Aufbau (vgl. Formel 2-6):

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

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. χ^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 α=0.05 3,84 beträgt, bestätigt der errechnete χ^2-Wert, dass die Patientengruppen keinen signifikanten Einfluss haben.

Mit SPSS ist man auch auf diese Vorgehensweise 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 Datansatz.

mit R:

Zunächst wird die Durchführung der Tests wie oben erläutert auf der Basis der Ergebnisse der parametrischen Analyse gezeigt. Danach wird der einfachere Weg mittels der Funktion `np.anova` 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` (Paket `car`) zu wandeln, damit die Werte in einer Matrix einzeln ansprechbar sind.
- Zunächst muss MS_{total} als Summe der Sum Sq-Spalte (1. Spalte) dividiert durch die Summe der df-Spalte (2. 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:

```
options (contrasts=c("contr.sum","contr.poly"))
aov1x   <- Anova(aov1r, type="III")
mstotal <- sum(aov1x[,1])/sum(aov1x[,2])
chisq   <- aov1x[,1]/mstotal
df      <- aov1x[,2]
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 (nach der Lösung 1) mit `Anova` (Paket `car`) die Streuungsquadrate vom Typ III ermittelt (vgl. Tabelle 4-6). Die Berechnung von MS_{total} erfolgt wie oben aus der Varianazanalyse `aov2r` durch Summation der Streu-

ungs-Quadratsummen `aov2r[2:5,1]` und Residuen `aov2r[2:5,2]`, wobei die erste Zeile (Intercept) ausgelassen werden muss.

```
library(car)
options (contrasts=c("contr.sum","contr.poly"))
mydata2 <- within(mydata2,{drugs<-factor(drugs);
                group<-factor(group); rx<-rank(x)})
aov2r   <- Anova(aov(rx~group*drugs,mydata2),type="III")
mstotal <- sum(aov2r[2:5,1])/sum(aov2r[2:5,2])
chisq   <- aov2r[2:5,1]/mstotal
df      <- aov2r[2:5,2]
pvalues <- 1-pchisq(chisq,df)
aov2y   <- data.frame(chisq,df,pvalues)
row.names(aov2y) <- row.names(aov2r)[2:5]
aov2y[2:4,]
```

mit der Ausgabe:

```
                chisq df     pvalues
group        2.089032  1 0.148360014
drugs       12.796395  3 0.005098255
group:drugs  5.135381  3 0.162148059
```

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` aus der Bibliothek `anova.lib` (vgl. Anhang 3) durchgeführt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Nachfolgend die Ein- und Ausgabe:

```
attach("path/anova.lib")
np.anova(x~group*drugs,mydata2, method=2)
```

```
Puri & Sen 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

```
options (contrasts=c("contr.sum","contr.poly"))
attach("path/anova.lib")
np.anova(x~group*drugs,mydata2, method=2, pseudo=T)
```

```
Puri & Sen 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 χ^2-Werte errechnet:

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

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

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

Die 5%-Schranken für die χ^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. 7 Aligned rank transform (ART und ART+INT)

Verschiedene Studien, u.a. von Sawilowsky et al. (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 signifkant sind, als es das vorgegebene α 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:

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

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

  bzw. wenn beide Haupteffekte separat getestet werden sollen:

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

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

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

  Interaktionseffekt: $e_m + (\overline{abc}_{ijl} - \overline{ab}_{ij} - \overline{ac}_{il} - \overline{bc}_{jl} + \bar{a}_i + \bar{b}_j + \bar{c}_l)$
- 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 anschließend die Ränge in normal scores (vgl. Kapitel 2.4) 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 garaniert ist.

mit R:

Zunächst die Durchführung des Verfahrens „per Hand“, d.h. das Alignment, also die Umrechnung der Werte, wird elementar wie oben beschrieben vorgenommen. Danach wird der (etwas) einfachere Weg mittels der Funktionen `art` sowie `art1.anova` gezeigt.

Als erstes wird 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 (`rab`) sowie einmal zur Ermittlung der Haupteffekte die entsprechende Effekt addiert (`ra`). Anschließend werden die bereinigten Residuen in Ränge transformiert (`rabr` bzw. `rar`). 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.

```
options (contrasts=c("contr.sum","contr.poly"))
mydata2 <- within(mydata2,{drugs<-factor(drugs); group<-factor(group)})
aov3    <- aov(x~group*drugs,mydata2)
ra <- rab <- aov3$residuals
                                                  # 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:

```
rabr ~ 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 *
```

sowie für die Haupteffekte:

```
rar ~ group * drugs
            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 etwas bequemer 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, z.B. in `a.d`.

```
options (contrasts=c("contr.sum","contr.poly"))
library(ARTool)
library(car)
mydata2a <- art(x~group*drugs,mydata2)$aligned.ranks
mydata2x <- cbind(mydata2[,1:3],aligned=mydata2a)
names(mydata2x)[4:6] <- c("a.g","a.d","a.gd")
Anova(aov(a.gd~group*drugs,mydata2x),type="III")
```

```
Response: a.gd
             Sum Sq Df F value   Pr(>F)
(Intercept) 1972.01  1 23.5333 5.48e-05 ***
group        126.30  1  1.5072  0.23100
drugs          2.48  3  0.0099  0.99862
group:drugs  876.49  3  3.4866  0.03058 *
Residuals   2094.92 25
```

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$a.gd/(nc+1))
Anova(aov(n.gd~group*drugs,mydata2x),type="III")
n.g    <- qnorm(mydata2x$a.g/(nc+1))
Anova(aov(n.g~group*drugs,mydata2x),type="III")
n.d    <- qnorm(mydata2x$a.d/(nc+1))
Anova(aov(n.d~group*drugs,mydata2x),type="III")
```

Hier lediglich die Ausgabe für den Test der Interaktion:

```
Response: n.gd
             Sum Sq Df F value  Pr(>F)
(Intercept)  0.0010  1  0.0013 0.97163
group        0.8859  1  1.1645 0.29084
drugs        0.0384  3  0.0168 0.99695
group:drugs  7.8930  3  3.4582 0.03144 *
Residuals   19.0198 25
```

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:

```
options (contrasts=c("contr.sum","contr.poly"))
attach("path/anova.lib")
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 (`rab`) sowie einmal zur Ermittluing der Haupteffekte deren Effekte addiert (`ra` und `rb`).
- Anschließend werden die bereinigten Residuen in Ränge transformiert (`rabr` bzw. `rar`).
- 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. 8 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 ϕ^{-1} die Umkehrfunktion der Normalverteilung. Die statistischen parametrischen F-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 χ^2-Tests des Kruskal-Wallis-Tests bzw. wie bei der o.a. Puri & Sen-Methode gerechnet. Die χ^2-Werte haben den Aufbau (vgl. Formel 2-6a):

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

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. (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]                # Ermittlung von N
Rx     <- rank(mydata2$x)                # Umrechnung in Ränge
nsx    <- qnorm(Rx/(nc+1))               # Umrechnung in normal scaores
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 χ^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
```

Alternativ kann das van der Waerden-Verfahren auch mit der Funktion `np.anova` aus der `anova.lib` (vgl. Anhang 3) durchgeführt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Über den Zusatz method=1 werden anstatt Puri & Sen-Tests van der Waerden-Tests durchgeführt. Nachfolgend die Ein- und Ausgabe:

```
options (contrasts=c("contr.sum","contr.poly"))
attach("path/anova.lib")
np.anova(x~group*drugs, mydata2, method=1)
```

```
generalized van der Waerden tests
            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
```

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.

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[a]	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) χ^2-Werte errechnet werden. 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):

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

Anschließend werden für jeden Effekt die χ^2-Werte errechnet:

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

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

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

Die 5%-Schranken für die χ^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 normal scores-Tests mit den van der Waerden-Tests reproduziert werden können.

4. 3. 9 ATS-Tests von Akritas, Arnold & Brunner

Im Gegensatz zum RT-, dem INT- oder dem ART-Verfahren 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 Brunner und Munzel (2013). Entsprechende R-Funktionen wurden in Abschnitt 3.1 aufgeführt.

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 (vgl. Abschnitt 4.3.3). Dafür müssen allerdings konservativere Ergebnisse in Kauf genommen werden. Nachfolgend die Anweisungen für beide Funktionen sowie deren Ausgabe:

```
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
```

```
            Test statistic     df1      df2     p-value
group             7.779701 1.00000 14.07807 0.014423402
drugs            10.154690 2.41235 14.07807 0.001270577
group:drugs       3.640870 2.41235 14.07807 0.046539673
```

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. 10 Parametrisches Verfahren mit anderen Residuen-Verteilungen

Wie in Abschnitt 4.1 erläutert, basiert das „normale“ parametrische Verfahren auf einer Normalverteilung der Residuen. Diesem liegt ein allgemeineres Verfahren zugrunde: die allgemeinen linearen Modelle (GLM, *general linear models*). Diese erlauben auch andere Verteilungsmodelle für die Residuen, z.B. die Poisson-Verteilung, typischerweise für Variablen, die auf Häufigkeiten basieren, ordinale oder rechtsschief verteilte Variablen. Daher kann diese Methode eine gute Alternative zu nichtparametrischen Verfahren sein, z.B. wenn die Normalverteilungs-Voraussetzung wegen starker Schiefe nicht erfüllt ist. Andere Verteilungsmodelle sind z.B. die Binomialverteilung für dichotome (binäre) abhängige Variablen, aus der dann die Logistische Regression gebildet wird. Die Variable y im Datensatz `lognormal` (vgl. auch Kapitel 11.2) ist leicht rechtsschief. Daher wird dieser Datensatz im Folgenden benutzt.

mit R:

Die entsprechende Funktion in R ist `glm`, und die Anova-Tabelle wird durch `Anova` (Paket `car`) ausgegeben. Wählt man als Verteilungsmodell `gaussian` , so würde man für den Datensatz `mydata2` exakt das Resultat wie mit `aov` erhalten (vgl. Tabelle 4-2) :

```
library(car)
ano1 <- glm(x~group*drugs, family=gaussian, mydata2)
Anova(ano1,type=3,test="F")
```

Und nun zumDatensatz `lognormal`. Auf Basis normalverteilter Residuen erhält man

```
anog <- glm(y~A*B, family=gaussian, lognormal)
Anova(anog,type=3,test="F")
```

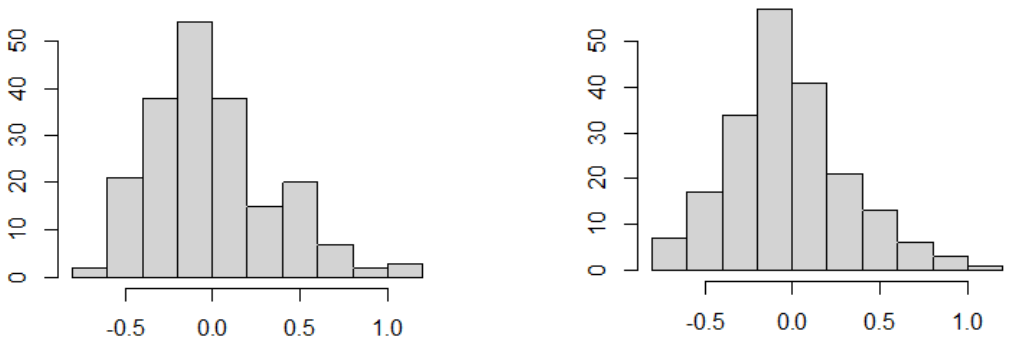
```
           Sum Sq  Df F values Pr(>F)
A            0.95   3     2.30  0.079 .
B            0.81   4     1.47  0.213
A:B          0.76  12     0.46  0.934
Residuals   24.79 180
```

und auf Basis des Poisson-Modells, wobei automatisch die Link-Funktion „log“ ist:

```
anop <- glm(y~A*B, family=poisson, lognormal)
Anova(anop,type=3,test="F")
```

```
           Sum Sq  Df F values Pr(>F)
A            0.86   3     2.29   0.08 .
B            0.73   4     1.46   0.22
A:B          0.68  12     0.45   0.94
Residuals   22.53 180
```

Die Ergebnisse sind zwar fast identisch, die Residuen auf Basis des Poisson-Modells sind allerdings deutlich besser normalverteilt. D.h. der Einfluss von Faktor A ist in jedem Fall nicht gesichert.



links die Residuen auf Basis der Normalverteilung, rechts auf Basis der Poisson-Verteilung

mit SPSS:

SPSS bietet zwar auch Funktionen für die Analyse mittels GLM (general linear models) an, aber die Prozedur `GLM` erlaubt nicht die Vorgabe von Verteilungsmodellen. Dafür jedoch die Prozedur `GENLIN` (Menü: Verallgemeinerte lineare Modelle -> Verallgemeinerte lineare Modelle) für die Analyse mittels GEE-Modellen (vgl. Kapitel 2.15 und 6.10), die allerdings nur Wald χ^2- aber keine F-Tests in der Anova-Tabelle durchführt. Theoretisch sollten die folgenden Anweisungen eine Analyse auf Basis Poisson-verteilter Residuen durchführen, doch leider führen sie zu einer Fehlermeldung:

```
GENLIN y by A B
 /model A B A*B
  distribution=Poisson   Link=Log.
```

Es sei angemerkt, dass dies im Fall von gemischten Versuchsplänen problemlos möglich ist (vgl. Kapitel 6.10.5).

4. 4 Nichtparametrische Verfahren zur mehrfaktoriellen Varianzanalyse

Die meisetn der in 4.3. vorgestellten Verfahren lassen sich 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. Wie man dennoch z.B. 3-faktorielle Versuchspläne mit solchen Funktionen analysieren kann, ist in Abschnitt 3.5 skizziert.

5. Abhängige Stichproben - Messwiederholungen

Es wird im Folgenden davon ausgegangen, dass ein Merkmal *x J*-mal (unter verschiedenen Bedingungen) erhoben wurde, so dass Variablen $x_1,...,x_J$ 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 Männer und Frauen dreimal im Abstand von mehreren Monaten auf einer ordinalen Skala von 1 - 9 (negativ - positiv) erfasst:

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

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

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.

		Kontrolle K			Medikament A			Medikament B		
		Aufgabe			Aufgabe			Aufgabe		
Geschlecht	Versuchsperson	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`.

Beispieldaten 14 (mydata14):

Der folgende Datensatz ist einem Artikel von Huynh (1978) entnommen. Von 17 Personen wurden auf Basis von Aufgaben Scores zu 3 Zeitpunkten erhoben, zunächst als Vortest (Variable `prae`), dann nach einer Störung und schließlich nach Verabreichung eines Wirkstoffs (Variablen `score1` und `score2`). Die Personen sind in 3 Gruppen (Faktor `Gruppe`) aufgeteilt, die sich hinsichtlich des Wirkstoffs unterscheiden.

Gruppe 1			Gruppe 2			Gruppe 3		
prae	score1	score2	prae	score1	score2	prae	score1	score2
10	24	17	25	26	20	15	24	17
10	26	20	20	25	20	15	26	19
15	27	20	35	28	21	20	27	19
35	29	21	40	29	21	40	28	20
30	29	22				35	30	20
5	24	17				55	32	23
						10	24	16

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
...	...	...	...	...	...	...

Vpn	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 die Analyse von 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. Die Beispiele hier beziehen sich alle auf Analysen von nur einer abhängigen Variablen. In der Praxis werden allerdings häufiger mehrere abhängige Variablen oder zusätzlich Kovariate analysiert. Dazu am Ende ein entsprechendes Beispiel.

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
.... . .    . .  ..
```

Beispieldaten 14 (mydata14):

Da `mydata14` nur einen Messwiederholungsfaktor beinhaltet, erfolgt die Umstrukturierung ähnlich wie oben gezeigt für `winer518`. Der Datensatz hat bereits eine Fallkennung (Vpn).

```
mydata14t <- reshape(mydata14, direction="long", timevar="Phase",
      v.names="score", varying=c("prae","score1","score2"), idvar="Vpn")
mydata14t <- within(mydata14t, {Gruppe<-factor(Gruppe);
      Phase<-factor(Phase); Vpn<-factor(Vpn) }
```

Der erzeugte Dataframe `mydata14t` hat dann folgende Gestalt:

```
     Gruppe  Vpn Phase score

1.1       1    1     1    10
2.1       1    2     1    10
3.1       1    3     1    15
.......
16.1      3   16     1    55
17.1      3   17     1    10
1.2       1    1     2    24
2.2       1    2     2    26
3.2       1    3     2    27
.......
1.3       1    1     3    17
2.3       1    2     3    20
3.3       1    3     3    20
.......
16.3      3   16     3    23
17.3      3   17     3    16
```

Wandlung in langes Format und Rückwandlung in breites Format

Soll ein Datensatz zuerst in das lange Format, danach wieder in das breite Format zurücktransformiert werden, z.B. um alle Werte einer Variablen, 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 Zurückwandlung 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)
```

Wandlung von mehr als einer abhängigen Variablen

Verschiedentlich sollen mehrere Variablen gleichzeitig transformiert werden, z.B. mehrere abhängige Variablen für dasselbe Design oder zusätzlich Kovariate. Das wird am o.a. Datensatz `mydata5` gezeigt, wozu die Bedeutung der 9 Messwiederholungsvariablen abgeändert wird: die 3 Stufen des Faktors Aufgabe haben nun folgende Bedeutung: 1 ist eine Kovariate (`covar`), 2 und 3 sind zwei abhängige Variablen (`fehler.a` und `fehler.b`), so dass nun ein 2-faktorieller gemischter Versuchsplan (Split-Plot Design) vorliegt.

```
Vpn      <- factor(1:8)
mydata5  <- cbind(Vpn,mydata5)
names(mydata5)[2] <- "Geschlecht"
mydata5  <- within(mydata5, Geschlecht<-factor(Geschlecht))
reshape(mydata5, direction="long", timevar="Medikament", times=1:3,
   v.names=c("covar","fehler.a","fehler.b"),
   varying=c("v1","v2","v3","v4","v5","v6","v7","v8","v9"),
   idvar="Vpn")->mydata5t
mydata5t <- within(mydata5t, Medikament<-factor(Medikament))
```

Einige Angaben sind selbsterklärend, zusätzlich sei angemerkt: `timevar` gibt den Namen des Messwiederholungsfaktors an und `times` die entsprechenden Stufennummern (levels), `varying` sind die Namen (oder Nummern) aller Messwiederholungsvariablen, zuerst Kovariate und beide abhängige Variablen für die erste Messwiederholungsstufe, dann Kovariate und beide abhängige Variablen für die zweite Messwiederholungsstufe, usw. Der Datensatz `mydata5t` hat em Ende folgende Struktur:

```
     Geschlecht Vpn Medikament covar fehler.a fehler.b
1.1     Männer   1          1     3        3        1
2.1     Männer   2          1     2        0        0
3.1     Männer   3          1     5        4        3
4.1     Männer   4          1     3        5        2
5.1     Frauen   5          1     2        2        1
6.1     Frauen   6          1     4        1        0
7.1     Frauen   7          1     3        2        1
8.1     Frauen   8          1     1        3        0
1.2     Männer   1          2     4        4        2
2.2     Männer   2          2     3        2        2
...     ......  ..         ..    ..       ..       ..
5.2     Frauen   5          2     2        2        2
...     ......  ..         ..    ..       ..       ..
1.3     Männer   1          3     5        4        3
...     ......  ..         ..    ..       ..       ..
8.3     Frauen   8          3     6        3        3
```

Eine abschließende Warnung:
Aus (dem Autor) nicht ersichtlichen Gründen kann es vorkommen, dass in der Ergebnisdatei die Variablennamen und Datenspalten falsch zugeordnet sind. Das kann man nur so lösen, dass die Variablennamen in einer anderen Reihenfolge vorgegeben werden.

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 für einen Faktor C mit *J* Messwiederholungen/Stufen - nachfolgend wird gelegentlich auch wieder die Anzahl mit *K* bezeichnet:

$$x_{jm} = \mu + \gamma_j + \pi_m + e_{jm} \quad (j=1,...,J \;\; m=1,...,N) \qquad (5\text{-}1)$$

wobei *N* die Anzahl der Merkmalsträger/Versuchspersonen ist und γ_j die Effekte von Faktor C. Gegenüber dem entsprechenden Modell ohne Messwiederholungen (vgl. Kapitel 4.1) gibt es einen personenspezifischen Effekt: π_m. Die Aufgabe der Varianzanalyse ist die Prüfung gleicher Gruppenmittelwerte, d.h. H_0 lautet wie in Kapitel 4.1:

$$\mu_1 = \mu_2 = \ldots = \mu_J$$

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

$$\gamma_1 = \gamma_2 = \ldots = \gamma_J = 0$$

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

$$F_1(x) = F_2(x) = ... = F_J(x)$$

woraus zu folgern ist, dass bei Ablehnung von H_0 nicht notwendigerweise sich die Mittelwerte μ_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

Die parametrischen Tests auf Basis des o.a. Modells (5-1) verlangen eine multivariaten Normalverteilung der Residuen e_{jm} der *J* Variablen. In der Praxis dürfte es genügen, die Residuen univariat zu überprüfen, allerdings wie in Kapitel 2.20.1 empfohlen, nicht zellweise oder für jede der *J* Variablen separat, sondern alle *J* Residuenvariablen e_j zusammen. Zusätzlich müssen die personenspezifischen Effekte π_m auf Normalverteilung überprüft werden. Letzteres wird nur vereinzelt erwähnt (z.B. Schaefer, 1994, und Stiger et al., 1998). Während die multivariate Normalverteilung der Residuen e_{jm} eher als essentiell angesehen wird, ist über die Auswirkungen fehlender Normalverteilung der π_m praktisch nichts bekannt. Die Ermittlung der e_{jm} und π_m erfolgt in R und SPSS unterschiedlich und wird in den Kapiteln 5.3.1 und 6.2 gezeigt. Zur univariaten Prüfung kann wieder zum einen der Shapiro-Wilk-Test, zum anderen grafische Verfahren herangezogen werden, bzw. zur multivariaten Prüfung z.B. der Test von Mardia. 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 in den Kapiteln 5.3.1 und 6.2.

Varianzhomogenität - Spherizität

Dazu kommt wieder die Voraussetzung der Varianzhomogenität, in diesem Fall der J zu vergleichenden Variablen: $\sigma_1^2 = ... = \sigma_J^2$. Allerdings nur für den Fall $J > 2$. Denn im Fall $J = 2$ kann zum Vergleich der beiden Variablen einfach deren Differenz verwendet werden (vgl.Abschnitt 6.10.9). 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 Σ ($\sigma_{ij,}$ i,j=1,...,J) der Variablen x_j muss besondere Eigenschaften haben.

Damit die F-Tests der Varianzanalyse korrekt getestet werden können, muss Σ *zirkulär* (*circular*) sein, das entspricht in etwa der Bedingung

$$\sigma^2_{x_1 - x_2} = \sigma^2_{x_1 - x_3} = \sigma^2_{x_2 - x_3} = \ldots$$

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 Σ auf Zirkularität (*circularity*) erfolgt über eine andere Matrix-Eigenschaft: *Spherizität*, 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 Σ 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 Spherizität 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 Spherizität 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 Spherizität wird allgemein der *Mauchly-Test* verwendet (siehe z.B. Winer, S. 255), so auch standardmäßig in R und SPSS, der allerdings im Vergleich zu einigen anderen Tests Nachteile aufweist. Mehr dazu und Hinweis auf weitere Tests in Kapitel 2.20.4. 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):

- 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 darauf basierende R-Funktionen zum Test auf Sphärizität 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 ε vom Idealwert 1 (für Sphärizität) messen, z.B. über die Korrelation von ε 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 ε 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 Abschnitt 5.3.1 beschrieben.

Der Test auf Sphärizität dient in erster Linie dem Test auf Gleichheit der Varianzen der Messwiederholungsvariablen. Huynh & Feldt (1970) haben gezeigt, dass Sphärizität nicht unbedingt eine Gleichheit der Korrelationen der Messwiederholungsvariablen $\rho_{x_1x_2}=\rho_{x_2x_3}=\rho_{x_3x_4}=...$ 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. Sphärizitäts-Tests 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 Sphärizität, sind dagegen gravierender, können aber statistisch aufgefangen werden. Ein Maß für die Abweichung ist das von Box errechnete ε (mit $1/(J-1) \leq \varepsilon \leq 1$), und $\varepsilon=1$ im Fall perfekter Sphärizität). Sowohl *Geisser & Greenhouse* als auch *Huynh & Feldt* haben auf diesem ε basierende modifizierte F-Tests entwickelt, die auch bei Abweichungen von der Sphärizität angewandt werden können. (Greenhouse-Geisser verwenden das Box ε, daher häufig mit GG bezeichnet, während das ε von Huynh & Feldt etwas modifiziert wurde, größer als das Box ε ist, und meist mit $\tilde{\varepsilon}$ (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 ε verkleinert. Der F-Wert selbst bleibt davon unberührt. Als Konsequenz daraus reagiert der F-Test konservativer, je stärker die Abweichung ε 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` (siehe auch Kapitel 2.20.4).

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 Sphärizität

erfüllen, sowohl der Fehler 1. Art α 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 Spherizität interessiert in dem Zusammenhang nicht, da dessen Voraussetzungen ohnehin kaum erfüllt sein werden. Das Verhalten der Kovarianzmatrizen, um die es ja bei der Spherizität geht, ist von Bryan (2009) ausführlich im Zusammenhang mit Rangtransformationen untersucht worden, ist aber zu umfangreich, um hier kurz wiedergegeben zu werden.

Eine „automatisierte“ Anwendung der HF-Adjustierung mag zwar praktisch erscheinen, man sollte allerdings auch bedenken, dass in der Regel damit eine Reduktion der Power verbunden ist. Und zwar beim parametrischen F-Test ca. 5 % im Fall von vorhandener Spherizität und 8-10% andernfalls. Bei der ART-Methode ist der Verlust deutlich höher (10-15%), dagegen bei RT und INT praktisch zu vernachlässigen (siehe Lüpsen, 2024).

Häufiger wird auch vorgeschlagen, zum Test eines Messwiederholungsfaktors anstatt der klassichen univariaten Tests einen multivariaten Test (vgl. Abschnitt 2.12) zu verwenden. Der Vorteil: Dieser verlangt nicht die Voraussetzung der Varianzhomogenität (Spherizität). Der Nachteil: Es wird eine multivariate Normalverteilung gefordert. Allerdings ist ein multivariater Test generell weniger robust gegenüber Verletzungen der Normalverteilung als der univariate F-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 χ^2-Tests konstruiert.

Die oben angeführten Rank transform Tests sind im Wesentlichen für metrische abhängige Variablen und als Alternative zum F-Test gedacht. Einige der in diesem Kapitel vorgestellten Verfahren verwenden χ^2-Tests und sind gut geeignet für ordinale Variablen, so z.B. der Puri & Sen-Test, das verallgeinerte Kruskal-Wallis-Friedman-Verfahren (KWF), das verallgeinerte van der Waerden- und das Koch-Verfahren. Wie schon früher erwähnt, verlangen diese zwar keine Spherizität, jedoch etwa gleiche Verteilungsformen, d.h. u.a. gleiche Varianzen der Messwiederholungsvariablen. Dies mittels eines der o.a. Spherizitäts-Tests zu überprüfen, wäre zu viel. Besser ist ein nichtparametrischer Test auf Gleichheit abhängiger Varianzen (vgl. Kapitel 2.20.3). 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. Spherizität voraussetzen, so z.B. autoregressive und 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 + Vpn bzw. C*D + Vpn

wobei Vpn 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_i 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 π_m subtrahiert werden. Die erforderlichen Schritte sind dann:

- Speichern der Residuen: e'_{jm} ,
- Ermitteln des Personeneffekts π_m aus $p_m = \left(\sum_j^J x_{jm}\right)/I$ 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 π_m kann entfallen, da sie für die Beurteilung der Residuen e_{jm} ohne Bedeutung ist.)

Für größere N (N>20) können diese e_{jm} auch einzeln für j=1,...,J auf Normalverteilung überprüft werden. Die J Testergebnisse 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*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 - eine Alternative mit dem Paket `ez` wird weiter unten gezeigt:

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

mit dem Ergebnis:

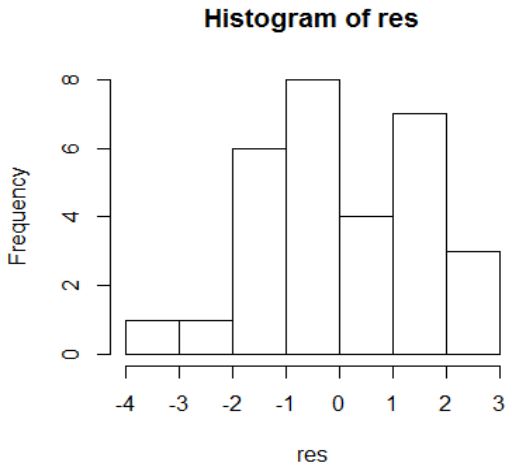
```
Error: Vpn
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9  59.87   6.652

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 π_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 <- rowMeans(winer518[,3:5])
hist(pm)
shapiro.test(pm)
```

Es sei noch darauf aufmerksam gemacht, dass R zwar in dem o.a. `aov`-Ergebnisobjekt `aov1` Residuen zur Verfügung stellt, die aber für die Prüfung der Normalverteilungsvoraussetzung wenig hilfreich sind. Das Objekt `aov1` selbst ist vom Typ `aovlist` und hat einen relativ komplizierten Aufbau. Über `residuals(aov1$"Vpn")` und

`residuals(aov1$"Vpn:Zeit")` erhält man die Residuen, deren Quadratsummen in der o.a. Anova-Tabelle unter den entsprechenden Überschriften ausgegeben werden und für die Berechnung der F-Tests benötigt wird. (Die Funktion `aov_ez` gibt zusammen mit der Funktion `residuals.afex_aov` (aus dem Paket `afex`) ebenfalls Residuen aus, allerdings ohne Erläuterungen, wie sich diese errechnen und welche Bedeutung diese haben.)

Soll die Normalverteilung der Residuen nicht nur univariat, sondern wie eigentlich gefordert multivariat geprüft werden, müssen diese vorher in eine Matrix gewandelt werden. Zunächst werden diese als ein Vektor `aov2$residuals`, allerdings variablenweise angeordnet, ausgegeben werden, Als Test wird hier der relativ bekannte und robuste von Mardia aus dem Paket `mvnormalTest` benutzt:

```
res.mat <- matrix(aov2$residuals,10,3)  # 10 Fälle und 3 Variablen
library(mvnormalTest)
mardia(res.mat)
```

```
$mv.test
          Test Statistic p-value Result
1     Skewness    3.3438  0.9721    YES
2     Kurtosis   -1.7717  0.0764    YES
3 MV Normality      <NA>    <NA>    YES

$uv.shapiro
   W      p-value UV.Normality
V1 0.9235 0.3868  Yes
V2 0.9335 0.4831  Yes
V3 0.9551 0.7284  Yes
```

der zuerst die Ergebnisse des multivariaten Tests zusammen mit einer Prüfung von Schiefe und Exzess, danach die univariaten ausgibt.

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 mlm- oder ein SSD-Objekt, doch der Aufruf ist einfach, solange man sich keine Gedanken über deren Bedeutung macht. Basis ist 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~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:

```
attach("path/anova.lib")
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 schon 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 Sphärizität enthalten ist. Auch dieses Verfahren reagiert empfindlich auf Abweichungen von der Sphärizität. 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 Mauchlys 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). Anzumerken ist, dass darin „`ges`“ die *generalized effect size* (Effektgröße Eta^2) ist. Darunter das Ergebnis des Tests von Mauchly (p~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 $\tilde{\varepsilon}$ bzw. dem Zusatz „p“ für die Irrtumswahrscheinlichkeit. Der unter GGe angegebene Korrekturfaktor `0.7375` ist das eigentliche Box ε.

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 Sphärizität anbetrifft, so ist der Benutzer hier in Anbetracht der recht unterschiedlchen 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 „Sphärizität angenommen“ abzulesen. Die beiden Zeilen „Greenhouse-Geisser“ und „Huynh-Feldt“ bieten alternative Tests für den Fall, dass die Voraussetzung der Sphärizität, 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~0.172), dass die Varianzhomogenität erfüllt ist. Die rechten Spalten „Epsilon“ enthalten den Korrekturfaktor der Freiheitsgrade $\tilde{\varepsilon}$ 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.

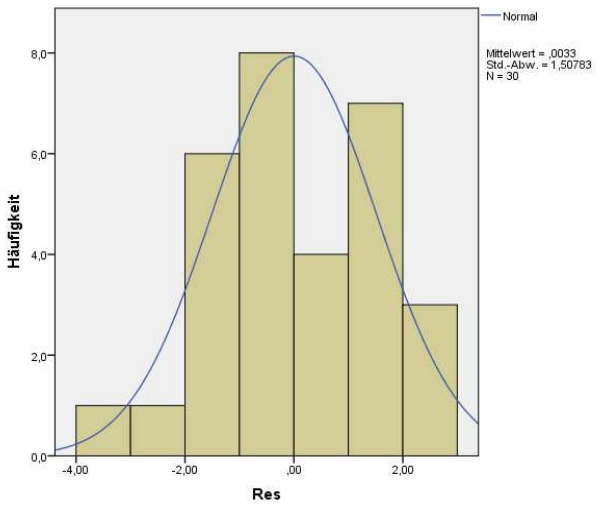
Über o.a. Aufruf wird für jede Messwiederholungsvariable (`t1,t2,t3`) eine Residuenvariable erzeugt: `Res_1, Res_2, Res_3`. Von diesen muss nun zunächst der Personeneffekt π_m abgezogen werden, der allerdings vorher noch berechnet werden muss. Nachfolgend die Kommandos hierfür, wobei im zweiten Schritt der Mittelwert von π 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 π 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[a]			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 bösartig angesehen werden muss. Diese resultiert zum Teil auch aus der zu großen Intervallzahl. Dahingegen weist der Shapiro-Wilk-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 J 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 χ^2-Test durchgeführt wird. 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),

```
friedman.test (as.matrix(winer518[,3:5]))          # Variante 1
friedman.test (score~Zeit | Vpn, data=winer518t)   # Variante 2
```

Die Ausgabe ist bei beiden natürlich identisch:

```
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
Teststatistik	9,556
Freiheitsgrade	2
Asymptotische Sig. (zweiseitiger Test)	,008

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Verteilungen von , and sind gleich.	Friedmans Zweifach-Rangvarianzanalyse verbundener Stichproben	,008	Nullhypothese ablehnen.

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 $R(x_m)$ transformiert. Beim einfachen *inverse normal transformation* (INT) werden anschließend zusätzlich die Ränge $R(x_m)$ über die inverse Normalverteilung in normal scores umgerechnet:

$$nscore_m = \Phi^{-1}(R(x_m)/(M+1))$$

wobei M die Anzahl aller Werte ist, also $M=N*J$ (mit N Anzahl der Merkmalsträger und J Anzahl der Messwiederholungen), m=1,..,M, und ϕ^{-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 `nscore` gelten. So ist es sinnvoll, auch $R(x)$ bzw. `nscore` 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. Lediglich 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 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~`0.0996`). 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[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, am besten Huynh-Feldt, entnommen werden.

Mauchly-Test auf Sphärizität[a]						
Innersubjekt-effekt	Mauchly-W	Approximiertes Chi-Quadrat	df	Sig.	Epsilon[b]	
					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 signfikant.

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,...,*N*J* transformiert.

Folgende Schritte sind durchzuführen:

- Alle *N*J* Werte werden in Ränge (1,...,*N*J*) transformiert.
- Mit den Rängen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle wird folgender χ^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 (A), SS_{Fehler} die Sum of Squares des Fehlers ist sowie *df* die entsprechenden Freiheitsgrade.

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

Die Überprüfung der Spherizität entfällt, da kein F-Test, sondern ein χ^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 $x_1,..,x_J$ wird der Median m_j ermittelt.
- Berechnen der absoluten Differenz $d_{jm} = abs(x_{jm} - m_j)$ für jede Fall *m*=1,...,*N*,
- 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):

$$\chi^2 = \frac{6\dot{9}8{,}6}{(6\dot{9}8{,}6 + 751{,}2)/(2 + 18)} = 9{,}64$$

die bei 2 FG auf dem 1%-Niveau signifikant ist.

Der Puri & Sen-Test kann auch bequemer mit der Funktion `np.anova` (vgl. Anhang 3) 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 Variablen $d_1,..,d_J$ als gleich und damit die Varianzen von $x_1,..,x_J$ als gleich angesehen werden. Inzwischen gibt es auch eine entsprechende Funktion `check.var` (vgl. Anhang 3). Die Daten können sowohl im breiten als auch im langen Format eingegeben werden. Neben dem Levene-like-Test (in 3 Varianten) wird noch ein weiterer Q-Test sowie die Varianzen der zu vergleichenden Variablen ausgegeben (siehe Abschnitt 2.20.3):

```
check.var(winer518[,3:5])                              # breites Format
with(winer518t,check.var(score,trial=Zeit,id=Vpn))  # langes Format
```

```
              Q method  Levene/Chi  Levene/F  Levene/Quade
Chisq/F-value    2.2961      1.1875    0.4500       0.5487
df               3.9477      2.0000   18.0000      18.0000
p-value          0.6737      0.5523    0.6446       0.5870
```

Hiervon soll laut Wilcox (1989) der Q-Test der zuverlässigste sein. Die anderen 3 basieren alle auf der o.a. Berechnung: die Chi-Variante verwendet den Friedman-Test, die F-Variante transformiert den Friedman- χ^2-Wert in einen F-Wert (vgl.Abschnitt 2.1), die Quade-Variante verwendet anstatt des Friedman- den Quade-Test (vgl. Abschnitt 2.11).

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:

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):

$$\chi^2 = \frac{6\dot{9}8{,}6}{(6\dot{9}8{,}6 + 751{,}2)/(2 + 18)} = 9{,}64$$

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 $d_1,..,d_J$ als gleich und damit die Varianzen von $x_1,..,x_J$ 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-Test (vdWS)

Bei dem Verfahren von *van der Waerden* werden ähnlich dem o.a. Puri & Sen-Test anstatt der „klassischen“ F-Tests die χ^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) m=1,..,N werden die Werte x_{m1},..,x_{mj} ,..,x_{mJ} in Ränge R(x_{mj}) (j=1,...,J) transformiert.
- Die Ränge werden in normal scores umgerechnet (vgl. Formel 2-2):
  $nscore_m = \Phi^{-1}(R(x_{mj})/(J+1))$.
- Mit diesen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle wird folgender χ^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 χ^2-Wert ist dann anhand der Tafeln für den χ^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 Sphärizität kann entfallen, da hier χ^2- anstatt F-Tests durchgeführt werden. Statt dessen sollte ein Test auf Gleichheit der Varianzen, wie im vorigen Abschnitt gezeigt, durchführt 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 χ^2-Wert des van der Waerden-Tests, der 2 FG hat, errechnet sich nun leicht per Hand:

χ^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) durchgeführt werden:

```
attach("path/anova.lib")
```

```
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 Quadrasummen interessieren, nicht aber die verschiedenen Testergebnisse in Abhängigkeit von der Sperizität, wird hier nur jeweils die 1. Zeile wiedergegeben:

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 χ^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 näherungsweisen 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` im Normalfall 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 (vgl. dazu Kapitel 2.7). 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. Über `time.name` wird optional ein Name für den Messwiederholungsfaktor angegeben:

```
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 Statistc (WTS):
     Statistic df      p-value
Zeit  43.42399  2 3.720494e-10

ANOVA-Type Statistc (ATS):
     Statistic       df     p-value
Zeit  8.369437 1.433543 0.001127567
```

5. 3. 7 Quade-Test

Das Verfahren von *Quade* war in Kapitel 2.11 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 2.11 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(Skillings.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 2.11 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 werden hierfür u.a. die Pakete `mvnormalTest` 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 wird aber hier nicht vorgestellt.

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,...,*N*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 χ^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 `V1,..,V4` 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 `V1,..,V4` erforderlich: `V4-V3, V3-V2` und `V2-V1`. 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 Wilks, wonach der Faktor Zeit einen signifikanten Einfluss hat:

```
with(winer568, anova(lm(cbind(V4-V3,V3-V2,V2-V1)~1), test="Wilks"))
```

```
Analysis of Variance Table

            Df    Wilks approx F num Df den Df    Pr(>F)
(Intercept)  1 0.027739   105.15      3      9 2.522e-07 ***
Residuals   11
```

Für die Durchführung des Tests von Akritas & Arnold müssen die Werte der Variablen `V1,..,V4` 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 `V1,..,V4` haben im Ergebnis-Dataframe die Namen `Rx.1,..,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 `V4-V3,V3-V2,...` 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:

```
attach("path/anova.lib")
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 `V1,..,V4` selbst, sondern die Residuen des Anova-Modells überprüft werden sollen, müssen diese zunächst ermittelt werden. Die Vorgehensweise war in Abschnitt 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, 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 Varianazanalysen 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 Wurzel nach Roy	35,051	105,152	3,000	9,000	,000

Für den Durchführung des Tests von Akritas & Arnold müssen die Werte der Variablen `V1,..,V4` 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.
Zeit	Pillai-Spur	,972	104,562	3,000	9,000	,000
	Wilks-Lambda	,028	104,562	3,000	9,000	,000
	Hotelling-Spur	34,854	104,562	3,000	9,000	,000
	Größte charakteristische Wurzel nach Roy	34,854	104,562	3,000	9,000	,000

Das Ergebnis des nichtparametrischen Tests weicht nur unbedeutend vom dem des parametrischen ab.

5. 3. 10 Multivariate Tests: Spatial Signs und Spatial Ranks Methoden

In den Kapiteln 2.11 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 multivariate Methoden für univariate Analysen mit Messwiederholungen eingesetzt werden

können. In R gibt es dazu mehrere Funktionen bzw. Pakets:

- Funktion `sr.loc.test` im SpatialNP (Sirkiä et al., 2007),
  eine Übertragung von Hotellings T^2 auf räumliche Ränge,
- Funktion `mv.1sample.test` im Paket MNM (Nordhausen & Oja, 2011),
  eine multivariate Erweiterung des Kruskal-Wallis-Test auf Basis räumlicher Ränge,
- Funktion `HP.loc.test` im Paket ICSNP (Hallin & Paindaveine, 2002),
  basierend auf den Rängen der pseudo-Mahalanobis-Abstände,
- Funktion `rank.ctest` im Paket ICSNP (Nordhausen et al., 2006 und Puri & Sen, 1971),
  Transformation der Variablen in ein affin-invariantes Koordinatensystem mit anschließenden score-Tests basierend auf Rängen oder normal scores.

die alle in gleicher Weise benutzt werden können. Der Unterschied besteht im Wesentlichen in den sehr komplexen 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 (Lüpsen, 2024) 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 negativem Pairing (vgl. 1.1.6) das α-Risiko fast so stark wie die parametrischen Methoden. Die Power liegt ca. 10-15% unter der des F-Tests, was als gut zu bewerten ist. Siehe auch Abschnitt 6.10.2.

mit R:

Diese Verfahren werden anhand des Datensatzes `winer568` mit den Variablen `V1,..,V4` vorgestellt. Die Struktur der Datenmatrix muss hier, wie oben beim parametrischen Verfahren, die „normale“, also die 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 standardization

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)
```

Beide Ergebnisse p=0.0002 bzw. p=0.0104 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: bei `HP.loc.test` Parameter `score` , jedoch bei `rank.ctest` Parameter `scores`):

```
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. 3. 11 Verwendung robuster Mittelwerte

Anstatt arithmetischer Mittel und Varianzen werden hier getrimmte Mittelwerte und winsorisierte Varianzen verwendet. Die Methodik wurde in Abschnitt 2.8 kurz erläutert. Das Verfahren kompensiert heterogene Varianzen der Messwiederholungsvariablen und die damit verbundene fehlende Spherizität.

mit R:

Wilcox (vgl. Mair & Wilcox, 2019) bietet dazu das Paket `WRS2` mit der Funktion `rmanova` für eine 1-faktorielle Analyse an, leider keine entsprechende für 2-faktorielle oder Split-Plot-Designs. Der Anteil der getrimmten (eliminierten) Werte, standardmäßig 0.2, kann über den Parameter `tr=...` bestimmt werden (jeweils `tr` Prozent am unteren und `tr` Prozent am oberen Ende). Es wird wie oben der Datensatz `winer518` verwendet, allerdings in der transformierten Form (siehe Abschnitt 5.1.2).

```
library(WRS2)
with(winer518t, rmanova(score,Zeit,Vpn,tr=0.2))
```

```
Test statistic: F = 6.1621
Degrees of freedom 1: 1.67
Degrees of freedom 2: 8.34
p-value: 0.02607
```

Basierend auf demselben Modell können auch paarweise Vergleiche durchgeführt werden:

```
with(winer518t, rmmcp(score,Zeit,Vpn,tr=0.2))
```

```
           psihat ci.lower ci.upper p.value p.crit   sig
T1 vs. T2 1.33333 -3.61816  6.28483 0.38497 0.0500 FALSE
T1 vs. T3 3.00000  1.24389  4.75611 0.00180 0.0169  TRUE
T2 vs. T3 2.33333 -1.26563  5.93230 0.07054 0.0250 FALSE
```

Die Spalte `p.crit` enthält den p-Wert, und die Spalte `p.value` den mit dem Hochbergh-Verfahren α-adjustierten p-Wert.

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 Spherizität. Für jeden der drei Tests C, D und C*D (im Beispiel: `Medikament, Aufgabe` und Wechselwirkung) wird die dafür relevante Spherizität 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.

```
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

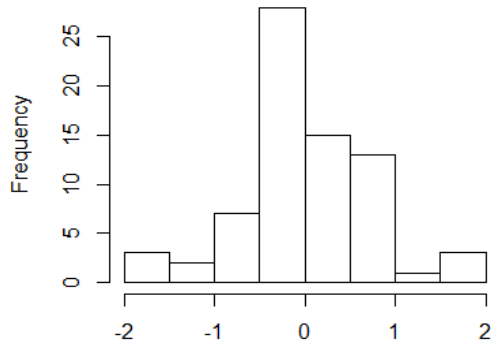
Das Ergebnis: Sowohl zwischen den beiden Medikamenten bzw. der Kontrollmessung als auch zwischen den drei Aufgaben bestehen hinsichtlich der Bearbeitung der Aufgaben (Fehlerzahl) signifikante Unterschiede. Hinzu kommt eine signifikante Wechselwirkung beider Faktoren. Auf Details der Interpretation soll hier nicht eingegangen werden.

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 vorher 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 π_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. Spherizität 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 Spherizität 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 Spherizität mittels der Funktion `check.sphere` ist dagegen etwas aufwändiger. Überprüft man "einfach" die 9 Variablen `v1,..,v9` , so entspricht das dem Test der Interaktion auf Spherizität:

```
attach("path/anova.lib")
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 `m1,m2,m3` und haben im Dataframe die Positionen 11-13:

```
attach("path/anova.lib")
mydata5 <- within(mydata5,{m1<-v1+v2+v3; m2<-v4+v5+v6; m3<-v7+v8+v9})
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
```

Wie bereits in Kapitel 2.20.8 angedeutet, war damit zu rechnen, dass bei der Prüfung einer Voraussetzung mit mehreren verschiedenenTests nicht nur unterschiedliche, sondern sogar recht widersprüchliche Ergenisse auftreten können. Eine Entscheidung ist häufig recht schwierig. In diesem Fall gelten die Tests von John (beta) und John-Nagao als (einigermaßen) zuverlässig. (Und erfreulicherweise mit qualitativ ähnlichen Ergebnissen.)

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: der Anova-Tabelle und des Mauchly-Tests :

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

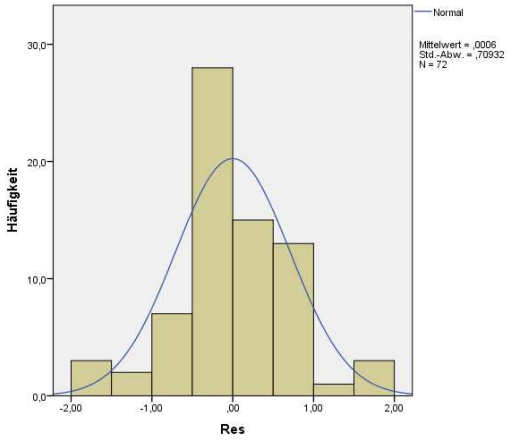
Mauchly-Test auf Sphärizität							
Innersubjekteffekt	Mauchly-W	Approximiertes Chi-Quadrat	df	Sig.	Epsilon		
					Greenhouse -Geisser	Huynh -Feldt	Unter grenze
Medikament	,350	6,297	2	,043	,606	,665	,500
Aufgabe	,869	,845	2	,655	,884	1,000	,500
Medikament * Aufgabe	,020	21,075	9	,016	,426	,549	,250

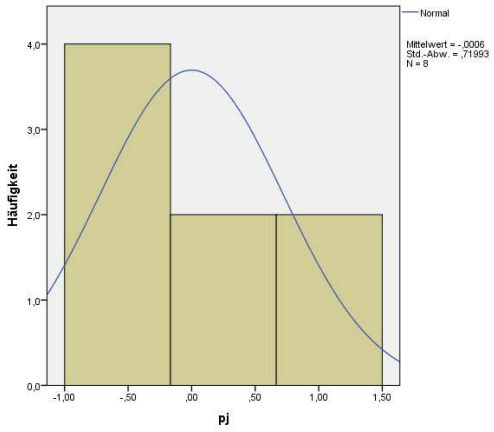
Tabelle 5-7

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			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 π_m abzuziehen. Der Shapiro-Wilk-Test ergibt dann ein p=0,173. Unten links das dazugehörende Histogramm, unten rechts das Histogramm für die π_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 Varianzinhomogenitäten zu begegnen.

Das INT-Verfahren unterscheidet sich von dem RT-Verfahren nur marginal: Die Ränge R(x_m) werden noch in normal scores (INT) umgerechnet:

$$nscore_m = \Phi^{-1}(R(x_m)/(M+1))$$

wobei *M* die Anzahl aller Werte ist, also *N*I*J* (mit *N* Anzahl der Merkmalsträger und *I* und *J* Anzahl der Messwiederholungen der Faktoren C und D), *m*=1,..,*M*, sowie ϕ^{-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 Spherizität 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 u.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, 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
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		

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
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		
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
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		

Tabelle 5-9

Nun das Ergebnis für das normal score (INT)-Verfahren, zunächst der Mauchly-Test:

Mauchly-Test auf Sphärizität[a]						
Innersubjekteffekt	Mauchly-W	Approx.. Chi-Quadrat	df	Sig.	Epsilon	
					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 χ^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 χ^2-Werte sind dann in den Tafeln für den χ^2-Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade (df_{Effekt}) des entsprechenden F-Tests sind.

- Die χ^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 Spherizität. 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 Messwiederholugen 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:

```
library(ez)
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_{Effekt} bzw. den dazugehörigen Fehlerterm SS_{Fehler}, die Spalten `DFn` und `DFd` (2. und 3. Spalte) die entsprechenden Freiheitsgrade. Mit folgenden Anweisungen lassen sich die χ^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(pfvalue,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) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`mydata5t`). Eingabe und Ausgabe:

```
attach("path/anova.lib")
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 die „elementare“ Lösung. Dazu muss ein 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.

```
attach("path/anova.lib")
mydata5t <- within(mydata5t,            # Bilden von Faktor Int
           Int<-(as.integer(Medikament)-1)*3 + as.integer(Aufgabe))
mydata5t <- within(mydata5t,Int<-factor(Int))
mydata5t <- within(mydata5t,            # Differenzen fuer Medikament
```

```
              diff.M<-abs(Fehler-ave(Fehler,Medikament,FUN=median)))
mydata5t <- within(mydata5t,            # Differenzen fuer Aufgabe
              diff.A<-abs(Fehler-ave(Fehler,Aufgabe,FUN=median)))
mydata5t <- within(mydata5t,            #  Differenzen fuer Interaktn
              diff.I<-abs(Fehler-ave(Fehler,Int,FUN=median)))
                                        # Friedman-Tests der Differenzen
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 Int       56 379.37
```

Nun die Lösung mittels der Funktion `check.var`. Direkt ist allerdings nur der Homogenitäts-Test für die Interaktion möglich, und zwar über den Datensatz `mydata5` (im breiten Format):

```
attach("path/anova.lib")
check.var(mydata5[,2:10])
```

```
    variance
y.1    1.554            y.6    0.696
y.2    2.571            y.7    0.696
y.3    1.143            y.8    0.571
y.4    1.125            y.9    0.571
y.5    0.839
               Q method  Levene/Chi  Levene/F  Levene/Quade
Chisq/F-value   20.5083      8.7536    0.1094       1.2636
df              19.8749      8.0000   56.0000      56.0000
p-value          0.4187      0.3635    0.9987       0.2810
```

Die Ausgabe enthält die Varianzen für die 9 Situationen sowie die 4 Testergebnisse. Die Homogenitäts-Test für die beiden Haupteffekte (Medikament und Aufgabe) erfordern leider auch erst die Erstellung zweier Datensätze mittels der Funktion `aggregate`, bei denen die Werte einmal über die Aufgaben und einmal über die Medikamente gemittelt werden:

```
med <-with(mydata5t,aggregate(Fehler,by=list(Vpn,Aufgabe),FUN=mean))
with(med,check.var(x,trial=Group.2,id=Group.1))
aufg<-with(mydata5t,aggregate(Fehler,by=list(Vpn,Medikament),FUN=mean))
with(aufg,check.var(x,trial=Group.2,id=Group.1))
```

Bei `aggregate` werden die Faktoren angegeben, über die *nicht* gemittelt werden soll. Die Variablen der Ergebnisdatei sind `Group.1` für die erste Variable `Vpn` und `Group.2` für die zweite Variable Medikament bzw Aufgabe. Die Ergebnisse (oben für Medikament, darunter für Aufgabe) unter Verwendsung des Medians zeigen, dass die Streuungen für beide Haupteffekte als homogen angesehen werden können:

```
               Q method  Levene/Chi  Levene/F  Levene/Quade
Chisq/F-value    3.2999      0.0645    0.4375        0.5255
df               4.9687      2.0000   14.0000       14.0000
p-value          0.6497      0.9683    0.6542        0.6025

               Q method  Levene/Chi  Levene/F  Levene/Quade
Chisq/F-value    4.5556      0.2500    0.4375        0.8715
df               4.9687      2.0000   14.0000       14.0000
p-value          0.4682      0.8825    0.6542        0.4398
```

mit SPSS:

Die Puri & Sen-Tests bauen auf der RT-Analyse (siehe vorigen Abschnitt, Tabelle 5-9) auf. Da hier χ^2-Tests anstatt F-Tests verwendet werden, spielt die Spherizität keine Rolle, so dass in der o.a. Tabelle nur die Zeilen „Sphärizität angen." relevant sind.

Die χ^2-Werte müssen nun „mit der Hand" aus den Werten der o.a. Tabelle (Spalten „Quadratsumme" und „df") berechnet werden:

$$\chi^2_{Medikament} = \frac{5419,1}{(5419,1 + 1733,8)/(2 + 14)} = 12,12 \qquad df_{Medikament} = 2$$

$$\chi^2_{Aufgabe} = \frac{8057,8}{(8057,8 + 3036,6)/(2 + 14)} = 11,61 \qquad df_{Aufgabe} = 2$$

$$\chi^2_{Interaktion} = \frac{1099,7}{(1099,7 + 2774,5)/(4 + 28)} = 9,08 \qquad df_{Interaktion} = 4$$

Die 5%- bzw. 1%-Schranken der χ^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 *m* (Vpn) Ränge für die einzelnen Messwiederholungen x_{mk} vergeben (*Friedman-Ränge*), hier die Ränge 1,...,*I*J*, und damit eine parametrische Varianzanalyse durchgeführt. Es werden nicht die F-Tests verwendet, sondern aus den Streuungsquadratsummen (SS, Sum of Sq) werden χ^2-Tests konstruiert:

Für die Effekte (Haupteffekte und Interaktionen) mit Messwiederholungsfaktoren z.B. C, D, C*D,... (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 zum Effekt gehörenden Fehlers ist sowie
- df_{Effekt} und df_{Fehler} die entsprechenden Freiheitsgrade.

Die χ^2-Werte sind dann anhand der Tafeln für den χ^2-Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade (df_{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 (Vpn) *m* ermittelten Ränge $R(x_{mk})$ (k=1,...,I*J) noch zusätzlich in normal scores umgerechnet werden:

$$nscore_{mk} = \Phi^{-1}(R(x_{mk})/(IJ + 1))$$

Die Überprüfung der Sphärizität kann entfallen, da hier χ^2- anstatt F-Tests durchgeführt werden. Statt dessen sollte ein Test auf Gleichheit der Varianzen, wie im vorigen Abschnitt ausführlich gezeigt, durchführt werden, bei dem allerdings vorab die x_{mk} wie oben beschrieben in Ränge zu transformieren sind.

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. Ausgangabasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `mydata5t`. Der Aufruf für das KFW-Verfahren:

```
attach("path/anova.lib")
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 Datensatzess 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:

$$\chi^2_{Medikament} = \frac{111,063}{(111,063 + 28,27)/(2 + 14)} = 12,75 \qquad df_{Medikament} = 2$$

$$\chi^2_{Aufgabe} = \frac{150,58}{(150,58 + 64,75)/(2 + 14)} = 11,19 \qquad df_{Aufgabe} = 2$$

$$\chi^2_{Interaktion} = \frac{28,41}{(28,41 + 46,92)/(4 + 28)} = 11,52 \qquad df_{Interaktion} = 4$$

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 χ^2-Tests wie folgt ermitteln:

$$\chi^2_{Medikament} = \frac{10,145}{(10,145 + 2,395)/(2 + 14)} = 12,94 \qquad df_{Medikament} = 2$$

$$\chi^2_{Aufgabe} = \frac{13,704}{(13,704 + 5,3)/(2 + 14)} = 11,54 \qquad df_{Aufgabe} = 2$$

$$\chi^2_{Interaktion} = \frac{2,7}{(2,7 + 3,722)/(4 + 28)} = 13,45 \qquad df_{Interaktion} = 4$$

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

  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. Sphärizität zu überprüfen, um dann gegebenenfalls anstatt des normalen F-Tests die Variante von Huynh & Feldt anzuwenden, oder alternativ ohne Beachtung des Mauchly-Tests. Nach Beasley (2002) spielt bei dieser ART-Methode die Sphärizität keine Rolle, so dass ein Blick auf den Mauchly-Test sogar entfallen kann und in der Anova-Tabelle ausschließlich der „normale“ F-Test von Bedeutung ist.

Es wird empfohlen (siehe Mansouri & Chang, 1995, sowie Carletti & Claustriaux, 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 garaniert 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.
- Ansschließend 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). Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz `mydata5t`. Eingabe und Ausgabe:

```
attach("path/anova.lib")
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:

```
attach("path/anova.lib")
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 (`ai`), Aufgaben (`bj`), Zellen (`mij`) und gesamt (`mm`) berechnet, um die Effekte von den Residuen abzuziehen und das Ergebnis in Ränge umzurechnen:

- `rab` bzw. die Ränge `rabr` zum Test der Interaktion
- `ra` bzw. `rar` 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 Spherizität keine Rolle, so dass ein Blick auf den Mauchly-Test entfallen kann und in der Anova-Tabelle ausschließlich die Zeile „Spherizität 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		Quadratsumme 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 „Spherizität angenommen“:

Tests der Innersubjekteffekte						
Quelle		Quadratsumme 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. Über `time1.name` und `time1.name` werden optional Namen für die Messwiederholungsfaktoren angegeben.

```
with(mydata5t, nparLD(Fehler~Medikament*Aufgabe,mydata5t,mydata5t$Vpn))
ano <- with(mydata5t, ld.f2(score,Medikament,Aufgabe,Vpn,
                      time1.name="Medikament",time2.name="Aufgabe"))
round(ano$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`:

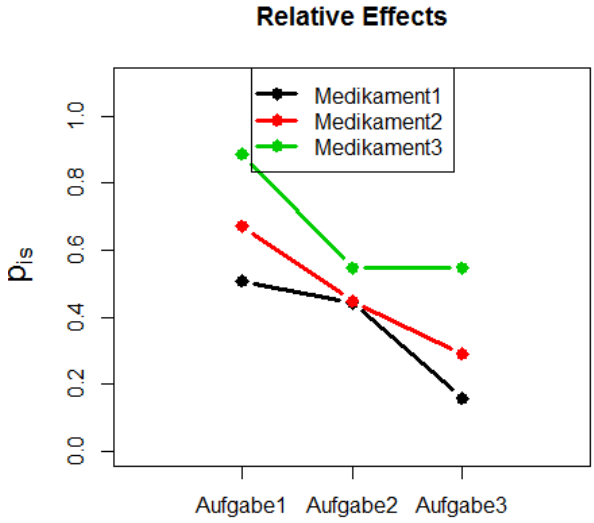
```
Call:
Fehler ~ Medikament * Aufgabe

Wald-Type Statistc (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 Statistc (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 denVorteil, 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,...,*N*I*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      # Medikament 1
rowSums(mydata5[,c("v4","v5","v6")])->m2      # Medikament 2
rowSums(mydata5[,c("v7","v8","v9")])->m3      # Medikament 3
rowSums(mydata5[,c("v1","v4","v7")])->a1      # Aufgabe 1
rowSums(mydata5[,c("v2","v5","v8")])->a2      # Aufgabe 2
rowSums(mydata5[,c("v3","v6","v9")])->a3      # Aufgabe 3
cbind(mydata5, m1,m2,m3, a1,a2,a3)->mydata5

anova(lm(cbind(a2-a1,a3-a2)~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
```

```
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 0.11026   24.209      2      6 0.00134 **
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 Vorabeit 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. (Die einzelnen Schritte sollen hier nicht weiter erläutert werden. Das Beispiel soll lediglich als ein „Muster" dienen.)

```
Medikament <- factor(rep(c("Kontrolle","Med A","Med B"),each=3),
              levels=c("Kontrolle","Med A","Med B"))
Aufgabe    <- factor(rep(c("A1","A2","A3"),3),
              levels=c("A1","A2","A3"))
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 Sphärizität 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 lediglich nicht wiedergegeben worden).

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 * Aufgabe	Pillai-Spur	,769	3,336[b]	4,000	4,000	,135
	Wilks-Lambda	,231	3,336[b]	4,000	4,000	,135
	Hotelling-Spur	3,336	3,336[b]	4,000	4,000	,135

6. Gemischte Versuchspläne

Unter *gemischten Versuchsplänen,* auch *Split Plot Designs* genannt*,* versteht man in der Regel solche, die sowohl Messwiederholungsfaktoren als auch Gruppierungsfaktoren enthalten. So wird im Folgenden davon ausgegangen, dass ein Merkmal x *J*-mal (unter verschiedenen Bedingungen) erhoben wurde, so dass Variablen $x_1,...,x_J$ 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, ein- oder mehrfaktoriell 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 Gruppierungfaktor A (mit *I* Gruppen) und einem Messwiederholungsfaktor C (mit *J* Wiederholungen) sowie $N = \sum n_i$ beobachteten Merkmalsträgern lautet das Modell dann:

$$x_{ijm} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \pi_m + \varepsilon_{ijm} \quad (i=1,...,I \;, j=1,...,J \text{ und } \; m=1,...,n_i) \qquad (6\text{-}1)$$

Auch hier gibt es einen personenspezifischen Effekt: π_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 ε_{ijm} sowie der Personeneffekte π_m: Hier sind im Wesentlichen dieselben Schritte erforderlich wie in Kapitel 5.2 beschrieben. Jedoch mit einem kleinen Unterschied die π_m betreffend: Diese berechnen sich prinzipiell auch hier pro Erhebungseinheit *m* als Summe p_m über alle Messwiederholungen, allerdings sind sie hier in den Gruppierungsfaktor A eingebettet. Somit wird am Ende nicht deren Mittelwert $\pi_m = (p_m - \bar{p})$ subtrahiert, sondern je nach Zugehörigkeit von *m* zur Gruppe *i* der entsprechende Mittelwert x_i also: $\pi_m = (p_m - x_i)$. Wie in 5.2 erwähnt, ist die, eigentlich multivariate, Normalverteilung der Residuen essentiell, allerdings die der Personeneffekte weniger folgenreich.

An dieser Stelle sei darauf aufmerksam gemacht, dass, streng genommen, auch die π_m auf Gleichheit der Varianzen bzgl. des oder der Gruppierungsfaktoren geprüft werden müssten, z.B. über einen einfachen Levene-Test, was aber in der Literatur weitgehend ignoriert wird. Dies erübrigt sich allerdings ohnehin, falls die u.a. Homogenität der Kovarianzmatrizen gegeben ist, da diese die Gleichheit der Varianzen der π_m impliziert.

Varianzhomogenität - Sphärizität

Zur Varianzhomogenität hinsichtlich der Messwiederholungen: Auch hier ist wie in 5.2 beschrieben ein Test auf Sphärizität (für alle Messwiederholungsfaktoren und deren Interaktionen miteinander) durchzuführen. Doch hier gibt es noch eine weitere Hürde: Die Kovarianzmatrizen der einzelnen Gruppen (für *i=1,...,I)* müssen (statistisch) gleich sein. Näheres dazu weiter unten. Die Verfahren zur Überprüfung der Sphärizität wurden in den Kapiteln 2.20.4 und 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 Sphärizität

Für den Fall von Verletzungen der Sphärizität 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 $\tilde{\varepsilon}$ von Huynh & Feldt nicht korrekt, insbesondere etwas zu groß. Lecoutre (1991) hat ein korrektes $\tilde{\varepsilon}$ angegeben. (Im Zähler des $\tilde{\varepsilon}$ von Huynh & Feldt ist N durch $(N\text{-}I\text{+}1)$ 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 I bemerkbar. Es sei noch einmal an den möglichen Verlust der Power der Tests für den Messwiederholungs- und Interaktionseffekt bei Anwendung der Adjustierungen der Freiheitsgrade erinnert (siehe Abschnitt 5.2).

Andere Verfahren, die keine Sphärizität voraussetzen, sind:

- den in Kapitel 2.12 erwähnten und in Kapitel 5.2 kurz vorgestellten multivariaten Test (z.B. von *Hotelling-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 Sphärizität umgeht,
- die Kapitel 2.15 vorgestellten *Generalized Linear Models* (GEE und GLMM),
- die in Kapitel 2.13 vorgestellte Methode für Adjustierungen der Freiheitsgrade IGA, quasi eine Verbesserung der in Kapitel 5.2 beschriebenen ε-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 Sphärizität überprüfen zu können. Und zwar verlangen sowohl die “normalen“ univariaten F-Tests, auch mit den oben vorgestellten Korrekturen, als auch die in 2.12 vorgestellten und in 6.10 besprochenen multivariaten Tests die Gleichheit der Kovarianzmatrizen (vgl. z.B. Keselman et al., 1993). Simulationsstudien (u.a. Algina, 1994 und Keselman et al., 1993) haben gezeigt, dass zwar, ähnlich den Varianzanalysen ohne Messwiederholungen, die Varianzheterogenität bei gleichen Zellenbestzungszahlen 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 Sphärizität voraussetzen.

Lüpsen (2024 und 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 Varianzen (heteroskedastische Kovarianzmatrizen) als auch durch heterogene Korrelationsmatrizen verursacht werden können. Dabei haben sich die o.a. Ergebnisse bestätigt, denen allerdings allen Kovarianzmatrizen mit ungleichen Varianzen, nicht jedoch mit ungleichen Korrelationen zugrunde lagen. Aber es zeigte sich darüber hinaus, dass zum einen der Fehler 1. Art und Teststärke auch bei gleichen n_i im Fall von großem J oder großen n_i betroffen sind, insbesondere für die Interaktion, und zwar beim multivariaten Test von Hotelling-Lawley (weniger jedoch Wilks und Pillai), beim GLMM sowie bei den nicht-HF-adjustierten Tests (parametrisch F, RT und ART), und zum anderen dass sich das Verhältnis von n_i zu den Korrelationen r_i der Messwiederholungen innerhalb der Gruppen erheblich auswirkt. Nachfolgend im Detail die Auswirkungen auf die Interaktion, zunächst beim parametrischen F-Test, der unbedingt HF-adjustiert sein sollte:

- 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 α=0.05).
- Bei ungleichen n_i und hinsichtlich der Streuung ungleichen Kovarianzmatrizen ist das Pairing (Zusammenhang zwischen n_i und Gruppenvarianzen s_{ij}^2, j=1,..,J) wesentlich:
  Liegt kein Pairing vor, ist auch der F-Test nicht stark beeinträchtigt. Bei negativem Pairing steigt die Fehlerrate (bei nominellem α=0.05) schnell auf 0.10 (corr(n_i,s_{ij}^2)=-0.3) bis 0.24 (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 α=0.05) schnell auf 0.20 (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,..,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 α=0.05) für corr(n_i,s_{ij}^2)=-0.3 bzw. bis 0.25 für corr(n_i,s_{ij}^2) = -0.7. Bei positivem Pairing sinkt wie beim F-Test die Power der multivariatenTests.
- Hinsichtlich des Einflusses ungleicher Korrelationen r_i innerhalb der Gruppen ist kein Unterschied zum Verhalten des F-Tests erkennbar.

Oben war darauf aufmerksam gemacht worden, dass der Fall des negativen Pairings (im Fall von heterogenen Kovarianzmatrizen), und analog im Fall des positiven Pairings (im Fall von Kovarianzmatrizen mit ungleichen Korrelationen), zu exzessiven Verletzungen des α-Risikos führt, zumindest bei den klassischen Methoden. Alternative Verfahren werden weiter unten aufgeführt.

Überprüfung der Homogenität von Kovarianzmatrizen

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 standardmäßig den Box-Test aus, und für R gibt es auch entsprechende Funktionen (siehe 2.20.5). 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:

- *LR-Test* (*Bartlett*-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,
- *Fligner & Killeen*, ein parameterfreier Tests,
- multivariater *Levene*-Test, der lediglich ordinales Skalenniveau voraussetzt, sowie
- zwei multivariate Dispersionstests, die ausschließlich auf unterschiedliche Varianzen ansprechen.

Weitere Erläuterungen zu den Verfahren sind in Kapitel 2.20.5 nachzulesen. Einige davon werden vom Autor in der R-Funktion `check.covar` angeboten wird (vgl. Anhang 3). 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-Tests sinnvoll bei symmetrischen 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: z.B. der Levene-Test. Er kann 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 nicht erkannter Ungleichheit der Korrelationen. Daher wird dringend geraten, bei einem nicht-signifikanten Ergebnis noch zusätzlich einen Test auf Homogenität der Korrelationsmatrizen durchzuführen (siehe auch Kapitel 2.20.5 und 2.20.6). Hierfür bietet der Autor eine R-Funktion `check.corr` mit 4 verschiedenen Tests an (vgl. Anhang 3), wobei der von Larntz & Perlman (1991) der empfehlenswerteste ist.

Insbesondere bei nichtparametrischen Tests, die einen χ^2- anstatt eines F-Tests benutzen, ist nicht unbedingt die Spherizität 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 die Beispiele in Kapiteln 5.3.4, 5.4.3 und 11.5.

Es gibt aber auch noch einen "Notbehelf" zur Überprüfung der Homogenität der Kovarianzmatrizen: statt dessen wird die Homogenität der Residuen-Varianzen geprüft. In Kapitel 5.2 war gezeigt worden, wie die Residuenvariablen bzw. die Residuen insgesamt ermittelt werden. Diese können nun auf Homogenität der Varianzen geprüft werden, 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 2.13.1 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: insbesondere für den Test der Interaktion muss der Stichprobenumfang eine Mindestgröße haben, um zuverlässige Ergebnisse zu erhalten. Algina (1994) hat auf Basis eigener Untersuchungen diese Bedingung sogar noch weiter verschärft.
- Weiterhin ist eine multivariate Version des bekannten Tests von Brown & Forsythe (mBF) anzuführen (siehe Kapitel 2.13.2), wofür in R eine Funktion zur Verfügung steht (siehe Anhang 3). Der mBF verfügt allerdings nur über eine geringe Power, die bis zu 50% unter dem Durchschnitt anderer Tests liegt.

- Algina (1994) favorisiert modifizierte F-Tests zur Kompensierung solcher Kovarianzheterogenitäten, wie etwa in Kapitel 4.2.2 oder 4.3.3 vorgestellt. Huynh (1978) hat für diesen Fall u.a. eine *improved general approximate procedure (*IGA*)* entwickelt*,* die zusätzlich Abweichungen von der Spherizität kompensieren kann (siehe Kapitel 2.13.3). 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). Die IGA ist allerdings wie der mBF sehr konservativ.
- Lüpsen (2020a und 2024) hat gezeigt, dass der verallgemeinerte Kruskal-Wallis-Friedman-Test (KWF) das α-Risiko sowohl bei heterogenen Kovarianzmatrizen als auch bei heterogenen Korrelationsmatrizen unter Kontrolle hält. Das Verfahren verlangt keine Spherizität, 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. Ähnliche Ergebnisse erzielen die Methoden ATS und van der Waerden (vdWS). Das KWF- und das vdWS-Verfahren haben jedoch eine akzeptable Power, insbesondere im Fall von negativem Pairing.
- Auch die Spatial Ranks Methoden (siehe Abschnitte 5.3.10 and 6.10.2) zeigen eine relativ gute overall-Performance, verletzen zwar in den Pairing-Situationen leicht das α-Risioko, aber weniger krass als die o.a. parametrischen Tests. Allerdings leidet auch hier die Teststärke darunter, die bis zu 30% unter dem Durchschnitt anderer Tests liegt.

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-Spherizität, jedoch nicht zur Abmilderung der Heterogenität der Kovarianzmatrizen verwendet werden können. Dies zeigte sich deutlich in Simulationsstudien (z.B. Lüpsen, 2020a und 2024).

Wie schon erwähnt, befreien nichtparametrische Verfahren in der Regel nicht von der Überprüfung der Homogenitätsvoraussetzung, da die Rangtransformationen in der Regel Heterogenitäten bestenfalls abschwächen. Andererseits basieren fast alle Prüfverfahren auf der Normalverteilung. Daher sind besonders bei ordinalen Kriterumsvariablen nur der Test auf Spherizität von Sirkiä bzw. der Levene-Test auf Gleichheit der Kovarianzmatrizen zu empfehlen.

Empfehlungen für die Methodenwahl

Bei den vielen in Kapitel 2 aufgeführten Methoden, aber zugleich all den o.a. Einschränkungen scheint eine kleine Hilfestellung zur Auswahl angebracht, die allerdings nicht auf Einzelheiten eingehen kann (vgl. auch Kapitel 2.19).

- Für den Test des Messwiederholungsfaktors (B) gelten generell folgende Einschränkungen: Bei rangbasierten Verfahren (u.a. RT, INT, KWF, vdWS, PS, ART) führt der Test im Fall von rechtsschiefen Verteilungen und zugleich heterogenen Varianzen der Messwiederholungen zu stark erhöhten p-Werten. Dasselbe gilt bei allen anderen Verfahren (u.a. alle parametrischen) im Fall von modalen (leicht mehrgipfligen) Verteilungen.
- Im Fall gleicher n_i ist der F-Test mit HF-Adjustierung der Freiheitsgrade eine gute Wahl, wobei im Fall von ordinalen Merkmalen die INT-Methode mit HF-Adjustierung bzw im Fall von extremen Werten die Analyse mittels getrimmter Mittelwerte vorzuziehen ist.
- Im Fall ungleicher n_i und gleicher, oder ungleicher Varianzen (Kovarianzmatrizen) ohne Pairing (also $|\text{corr}(n_i,s_i)| < 0.2$), ist der F-Test mit HF-Adjustierung der Freiheitsgrade ebenfalls eine gute Wahl, wobei im Fall von ordinalen Merkmalen die INT-Methode mit HF-Adjustierung vorzuziehen ist.

- Im Fall ungleicher n_i und ungleicher Varianzen (Kovarianzmatrizen), aber mit Pairing (also $|\text{corr}(n_i,s_i)| > 0.2$) sind andere Methoden vorzuziehen. Wenn das α-Risiko komplett unter Kontrolle gehalten werden soll, ist der Welch-James-Test eine gute Wahl, solange die n_i hinreichend groß sind, sonst die IGA-Methode (siehe Kapitel 2.13). Beide haben eine extrem geringe Power, die mBF-Methode sogar eine noch geringere. Andernfalls ist die KWF-Methode vorzuziehen, die eine sehr gute Power besitzt, aber in wenigen Situationen (u.a. bei schiefen Verteulungen) das α-Risiko leicht verletzen kann.
- Im Fall ungleicher n_i und ungleicher Korrelationen (Kovarianzmatrizen), aber mit Pairing (also $|\text{corr}(n_i,r_i)| > 0.2$) ist die KWF-Methode die beste Wahl.
- Folgende Verfahren sind nur in wenigen Situationen den anderen überlegen:
  - Die multivariaten Verfahren schneiden nicht besser ab als F-Test mit HF-Adjustierung.
  - GLMM und GEE haben in zu vielen Situationen ein sehr liberales Verhalten.
  - Die RT-Methode ist nur selten besser als die INT-Methode, doch meistens schlechter.
  - Das vdWS-Verfahren verhält sich fast identisch wie das KWF-Verfahren, jedoch manchmal etwas schlechter.
  - Die übrigen Verfahren (ART, ATS, PS, Koch, Agresti & Pendergast, Akritas & Arnold sowie die Spatial Signs und Spatial Ranks Methoden) sind ebenfalls zu vernachlässigen.

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 + Vpn (6-2)

wobei Vpn 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 gleichermaßen 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: $e'_{ijm} = x_{\text{ijm}} - \alpha\gamma_{ij} - \alpha_i - \gamma_j$. 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 π_m subtrahiert werden. Im Gegensatz zum Design ohne Gruppierungsfaktoren müssen hier allerdings von den Zeilensummen p_m die Gruppenmittelwerte x_i anstatt des Mittelwertes $\bar{p}$ subtrahiert werden. Die erforderlichen Schritte sind dann:

- Speichern der Residuen: e'_{ijm} ,
- Ermitteln des Personeneffekts $\pi_m = (p_m - \bar{x}_i)$ aus $p_m = \left(\sum_j^J x_{jm}\right) / J$ und
  $\bar{x}_i$ = Mittelwert der p_m für Gruppe *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 unterschiedlichen 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 zweckmäßig ist, weil die Funktion `ezANOVA` zugleich den Mauchly-Test durchführt (siehe unten). Hier 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 die Prüfung der Residuen ε_{ijm} auf univariate 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. Eine Überprüfung auf multivariate Normalverteilung der Residuen war in Kapitel 5.3.1 gezeigt. worden. Die Überprüfung der Normalverteilung der versuchspersonenspezifischen Abweichungen π_m erfolgt ähnlich wie in Kapitel 5.3.1:

```
pm <- rowMeans(winer518[,3:5])                    # Zeilenmittelwerte
xm <- with(winer518,ave(pm,Geschlecht,FUN=mean))  # Gruppenmittelwerte
pm<-pm-xm
shapiro.test(pm)
leveneTest(pm)
```

mit dem Ergebnis `W = 0.96756, p-value = 0.8673` für den Shapiro-Test sowie einem `F value` von 0 und einer `Pr(>F)` von 1, wonach alle Voraussetzungen für π_m erfüllt sind.

Die Überprüfung der Spherizität kann z.B. mittels des Mauchly-Tests und der Funktion `ezANOVA` des Pakets `ez` vorgenommen werden:

```
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 für beide Tests von `Zeit` 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 Spherizität, aber auch einen multisample Spherizitäts-Test nach Mauchly, der nicht die Homogenität der Kovarianzmatrizen verlangt. Hier werden wieder beide Aufrufvarianten gezeigt:

```
attach("path/anova.lib")
check.sphere(winer518[,3:5],winer518[,2)              # wide Format
with(winer518t,check.sphere(score,                    # long Format
                groups=Geschlecht,trial=Zeit,id=Vpn))
```

```
                     Chisquare df p value
Mauchly                 0.5752  2  0.7500
multisample Mauchly    11.5774 11  0.3962
John's V               10.8084  5  0.0553
Nagao                  10.8084  5  0.0341
LR                     11.0201  5  0.0005
Muirhead & Waternaud    4.5477  5  0.4735
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:

```
attach("path/anova.lib")
check.covar(winer518[,3:5],winer518$Geschlecht)      # wide Format
with(winer518t,check.covar(score,                    # long Format
                groups=Geschlecht,trial=Zeit,id=Vpn))
```

```
             statistic  type       df p value
LR/Bartlett     7.5870 chisq   6.0000  0.2699
Box M (C)       4.5048 chisq   6.0000  0.6087
Box M (F)       0.7344     F 463.6981  0.6221
Schott T1       3.8944 chisq   6.0000  0.6910
Schott T2       4.3395 chisq   6.0000  0.6308
Schott T3       3.9365 chisq   6.0000  0.6853
Levene          0.0351     F   8.0000  0.9657
Dispersion 1    0.2630     F   8.0000  0.6219
Dispersion 2    0.0944     F   8.0000  0.9111
```

Der empfehlenswerteste Test ist der Levene-Test. Der p-Wert `0.9657` indiziert die Homogenität der Kovarianzmatrizen, die, wie vorher erwähnt, auch Voraussetzung der Varianzanalyse ist.

Noch eine Erläuterung zu den beiden Dispersionstests: der erste vergleicht das Streuungsniveau für die beiden Gruppen. Dieses kann für beide Gruppen gleich sein, obwohl die Kovarianzmatrizen nicht gleich sind, nämlich wenn z.B. in der ersten Gruppe die Variable 1 die größere Streuung hat und die Variable 2 die kleinere, aber in der zweiten Gruppe die Variable 1 die kleinere Streuung hat und die Variable 2 die größere. Genau solche Unterschiede soll der zweite Dispersionstest erkennen.

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):

```
attach("path/anova.lib")
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, winer518$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 vorgstellte 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 2, 16 1.32 22.05 *** .428   <.001
3 Geschlecht:Zeit 2, 16 1.32 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 Sphärizität (in der neuesten Version in das Paket `performance` ausgelagert):

```
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
```

In der neuesten Version kommen anstatt o.a. Übersicht lediglich die folgenden Zeilen:

```
OK: There is not clear evidence for different variances across groups
(Levene's Test, p = 0.962).

OK: Data seems to be spherical (p > 0.778).
```

Eine Warnung: `residuals(..)` , angewandt auf ein Ergebnisobejekt, gibt bei gemischten Designs nicht die korrekten Residuen aus.

Zusätzlich wird auch die parametrische Analyse für den Datensatz `mydata5` gezeigt, der zwei Messwiederholungsfaktoren beinhaltet. Zunächst die Analyse mit `ezANOVA`, wobei die Syntax für die Angabe mehrerer Faktoren zu beachten ist:

```
ezANOVA(mydata5t, Fehler, Vpn, within=.(Medikament,Aufgabe),
    between=Geschlecht, type=3)
```

```
$ANOVA
                      Effect DFn DFd         F           p p<.05
                   Geschlecht   1   6  4.1482014 0.0878432774
                   Medikament   2  12 19.5000000 0.0001696940
                      Aufgabe   2  12 20.2018349 0.0001441841
         Geschlecht:Medikament   2  12  0.5526316 0.5894038922
            Geschlecht:Aufgabe   2  12  0.8807339 0.4396417315
            Medikament:Aufgabe   4  24  3.0560748 0.0361232196
 Geschlecht:Medikament:Aufgabe   4  24  0.3644860 0.8314342497

$`Mauchly's Test for Sphericity`
                        Effect           W           p p<.05
                   Medikament 0.10595568 0.003654354     *
         Geschlecht:Medikament 0.10595568 0.003654354     *
                      Aufgabe 0.90396431 0.776923421
            Geschlecht:Aufgabe 0.90396431 0.776923421
            Medikament:Aufgabe 0.01419791 0.037430986     *
 Geschlecht:Medikament:Aufgabe 0.01419791 0.037430986     *

$`Sphericity Corrections`
                  Effect       GGe       p[GG]       HFe        p[HF]
             Medikament 0.5279707 0.003726609 0.5452596 0.0033216294
  Geschlecht:Medikament 0.5279707 0.493274120 0.5452596 0.4979998431
                Aufgabe 0.9123791 0.000257679 1.2901416 0.0001441841
     Geschlecht:Aufgabe 0.9123791 0.432634492 1.2901416 0.4396417315
     Medikament:Aufgabe 0.3968182 0.101394908 0.5161724 0.0822208807
 Geschlt:Medikt:Aufgabe 0.3968182 0.657041002 0.5161724 0.7081566468
```

Die Prüfung der Voraussetzungen soll hier nur kurz skizziert werden, da dies in Kapitel 5.4.1 bereits ausführlich behandelt worden war. Zunächst zur Prüfung der Residuen ε_{ijm} auf univariate Normalverteilung. Diese lassen sich, wie oben erläutert, bequem als Residuen eines Anova-Modells ohne Messwiederholungen ermitteln (hier ohne Ausgabe der Ergebnisse):

```
aov3<-aov(Fehler~Geschlecht*Medikament*Aufgabe+Vpn, mydata5t)
res<-aov3$residuals
hist(res)
shapiro.test(res)
```

Die Prüfung der Normalverteilung der Personenmittel π_{im} erfolgt über

```
pm <- rowMeans(mydata5[,3:11])                    # Zeilenmittelwerte
xm <- with(mydata5, ave(pm,Geschlecht,FUN=mean)) # Gruppenmittelwerte
pm<-pm-xm
shapiro.test(pm)
leveneTest(pm)
```

Die Prüfung auf Sphärizität ist bereits in der o.a. Varianzanalyse unter „`Mauchly's Test for Sphericity`" durchgeführt worden, und zwar für jeden der Tests, bei denen Messwiederholungsfaktoren involviert waren.

Schließlich zur Prüfung der Homogenität der Kovarianzmatrizen, auch hier für jeden der Tests, bei denen Messwiederholungsfaktoren involviert waren. Da für den Test des Haupteffekts Medikament die Summen über den Faktor Aufgabe und für den Test des Haupteffekts Aufgabe die Summen über den Faktor Medikament relevant sind, müssen auch für den Test der Kovarianzhomogenität die entsprechenden Summen gebildet werden. Dies war in Kapitel 5.4.7 bereits einmal gezeigt worden. Am einfachsten ist dies über den nichttransformierten Datensatz möglich:

```
rowSums(mydata5[,c("v1","v2","v3")])->m1      # Medikament 1
rowSums(mydata5[,c("v4","v5","v6")])->m2      # Medikament 2
rowSums(mydata5[,c("v7","v8","v9")])->m3      # Medikament 3
rowSums(mydata5[,c("v1","v4","v7")])->a1      # Aufgabe 1
rowSums(mydata5[,c("v2","v5","v8")])->a2      # Aufgabe 2
rowSums(mydata5[,c("v3","v6","v9")])->a3      # Aufgabe 3
cbind(mydata5, m1,m2,m3, a1,a2,a3)->mydata5
```

Zunächst der Test der Kovarianzhomogenität für den Haupteffekt Medikament:

```
check.covar(mydata5[,c("m1","m2","m3")],mydata5x[,"Geschlecht"])
```

```
             statistic  type       df p value
LR/Bartlett    16.9132 chisq   6.0000  0.0096
Box M (C)       7.7519 chisq   6.0000  0.2569
Box M (F)       1.2271     F 260.8302  0.2926
Schott T1       5.8641 chisq   6.0000  0.4386
Schott T2       6.3972 chisq   6.0000  0.3802
Schott T3       6.3394 chisq   6.0000  0.3863
Levene          3.3343     F   6.0000  0.1202
Dispersion 1    2.0268     F   6.0000  0.2044
Dispersion 2    1.5511     F   6.0000  0.2992
```

analog der Test der Kovarianzhomogenität für den Haupteffekt Aufgabe:

```
check.covar(mydata5[,c("a1","a2","a3")],mydata5x[,"Geschlecht"])
```

```
             statistic  type       df p value
LR/Bartlett     9.7098 chisq   6.0000  0.1374
Box M (C)       4.4503 chisq   6.0000  0.6160
Box M (F)       0.7045     F 260.8302  0.6462
Schott T1       4.7206 chisq   6.0000  0.5801
Schott T2       5.1272 chisq   6.0000  0.5276
Schott T3       4.7651 chisq   6.0000  0.5743
Levene          0.3488     F   6.0000  0.7214
Dispersion 1    3.3781     F   6.0000  0.1157
Dispersion 2    0.5806     F   6.0000  0.5933
```

Schließlich der Test der Kovarianzhomogenität für die Interaktion, dem die 9 erhobenen Variablen `V1`,...,`V9` zugrunde liegen. Da für die parametrischen Homogenitäts-Tests die Stichprobenumfänge mindestens $n_i > J+1$ gelten muss (wobei J die Anzahl zu testenden Messwiederholungsvariablen ist), können hier bestenfalls die Levene-like nichtparametrischen Tests durchgeführt werden. Allerdings gelten für diese ebenfalls Einschränkungen, für die NAs ausgegeben werden, wenn diese nicht erfüllt sind:

```
check.covar(mydata5[,2:10],mydata5[,"Geschlecht"],par=F)
```

```
             statistic type df p value
Levene              NA <NA> NA      NA
Dispersion 1    0.8311    F  6  0.3971
Dispersion 2        NA <NA> NA      NA
```

Insgesamt können die Kovarianzmatrizen für alle Tests als homogen angenommen werden.

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 „Spherizität angenommen“ verwendet werden können:

Mauchly-Test auf Sphärizität							
Innersubjekt-effekt	Mauchly-W	Approx. Chi-Quadrat	df	Sig.	Epsilon		
					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		Quadratsumme 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 Kovarianzamatrizen für die beiden Gruppen sowie für alle 3 Variablen ein Levene-Test auf Gleichheit der Zellenvarianzen ausgegeben:

Box-Test auf Gleichheit der Kovarianzenmatrizen	
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				
	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 `m1,..,m3` für die Variablen `t1,..,t3` getrennt für die Gruppenvariable `Geschlecht` mittels `aggregate`,
- Berechnung der Kreuzprodukte $(y_i\text{-median}_i)(y_j\text{-median}_j)$ für i=1,..,3 und j=i,..,3, wodurch 3*4/2=6 neue Variablen `c1,..,c6` erzeugt werden,

- Transformation der c-Variablen: c=sign(c)*sqrt(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 charakter. Wurzel nach Roy	6,999	3,500	6,000	3,000	,166
Geschlecht	Pillai-Spur	,486	,472	6,000	3,000	,801
	Wilks-Lambda	,514	,472	6,000	3,000	,801
	Hotelling-Spur	,944	,472	6,000	3,000	,801
	Größte charakter. Wurzel nach Roy	,944	,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:

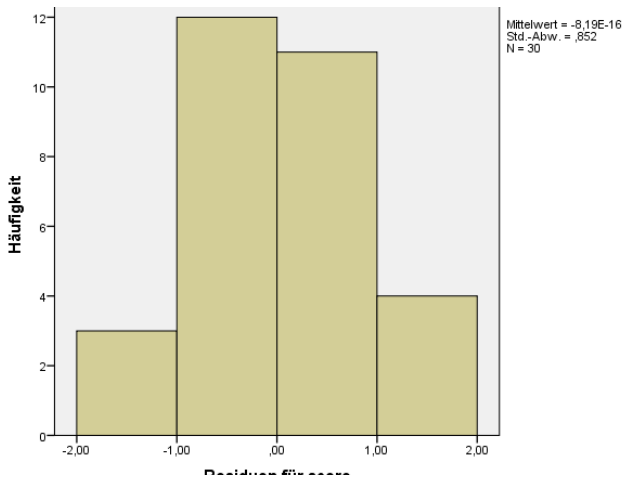
```
Varstocases
  /id=Vpn         /make score from t1 t2 t3
  /index=Zeit(3) /keep=Geschlecht  /null=keep.
```

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 n=30 keinen Sinn macht. Möglich wären hier 4, 5 oder 6 Intervalle (vgl. Kapitel 2.20.1), so dass eine Nachbereitung mit dem Grafikeditor erforderlich ist und o.a. Abbildung erzeugt, wonach die Residuen als normalverteilt angenommen werden können.

Tests auf Normalverteilung						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
RES_1	,126	30	,200*	,959	30	,288

Zusätzlich wird auch die parametrische Analyse für den Datensatz `mydata5` gezeigt, der zwei Messwiederholungsfaktoren beinhaltet (hier ohne Ausgabe der Ergebnisse, die weitgehend schon in Kapitel 5.4.1 zu sehen war):

```
GLM v1 v2 v3 v4 v5 v6 v7 v8 v9 by Geschlecht
  /wsfactor=Medikament 3 polynomial Aufgabe 3 polynomial
  /wsdesign=Medikament Aufgabe Medikament*Aufgabe
  /PRINT HOMOGENEITY.
```

Die Prüfung der Voraussetzungen soll hier nur kurz skizziert werden, da dies in den Kapiteln 5.3.1 und 5.4.1 bereits ausführlich behandelt worden war, u.a. die Prüfung der Residuen ε_{ijm} auf univariate Normalverteilung. Bezüglich der Personeneffekte π_{im} erfordert der zusätzliche Gruppierungsfaktor Geschlecht eine Abwandlung der Berechnung, weil von diesen die Gruppenmittel der π_{im} subtrahiert werden müssen:

```
compute pm = mean(v1 to v9).
aggregate
    /outfile=* /mode = addvariables
    /break = Geschlecht  /pm_mean = mean(pm).
compute pm = pm-pm_mean.
execute.
```

Die Personeneffekte `pm` können nun wie gewohnt auf Normalverteilung überprüft werden.

Schließlich zur Prüfung der Homogenität der Kovarianzmatrizen, auch hier für jeden der Tests, bei denen Messwiederholungsfaktoren involviert waren. Da für den Test des Haupteffekts Medikament die Summen über den Faktor Aufgabe, und für den Test des Haupteffekts Aufgabe die Summen über den Faktor Medikament relevant sind, müssen auch für den Test der Kovarianzhomogenität die entsprechenden Summen gebildet werden.

```
compute m1=v1+v2+v3.
compute m2=v4+v5+v6.
compute m3=v7+v8+v9.
compute a1=v1+v4+v7.
compute a2=v2+v5+v8.
compute a3=v3+v6+v9.
execute.
```

Anschließend wird mit diesen in zwei Schritten wieder eine Varianzanalyse mit `GLM` durchgeführt, um die Box-Tests zu erhalten.

```
GLM m1 m2 m3 by Geschlecht
  /wsfactor=Medikament 3 polynomial
  /wsdesign=Medikament
  /PRINT HOMOGENEITY.
```

```
GLM a1 a2 a3 by Geschlecht
  /wsfactor=Aufgabe 3 polynomial
  /wsdesign=Aufgabe
  /PRINT HOMOGENEITY.
```

Box-Test auf Gleichheit der Kovarianzenmatrizen

Box-M-Test	16,913
F	1,227
df1	6
df2	260,830
Sig.	,293

oben der Test für den Medikament-Haupteffekt, unten für den Aufgaben-Haupteffekt

Box-Test auf Gleichheit der Kovarianzenmatrizen

Box-M-Test	9,710
F	,704
df1	6
df2	260,830
Sig.	,646

Der für die Interaktion Medikament*Aufgaben relevante Test Box-Test wurde eigentlich im o.a. `GLM` -Kommando angefordert, führt aber zu einer Fehlermeldung, da für die parametrischen Homogenitäts-Tests die Stichprobenumfänge mindestens $n_i > J+1$ gelten muss (wobei J die Anzahl zu testenden Messwiederholungsvariablen ist), was hier mit $n_i=4$ und $J=9$ nicht gegeben ist. Abhilfe könnte bestenfalls der o.a. , allerdings recht aufwändige Levene-Test auf Gleichheit der Kovarianzmatrizen bringen.

6. 3 Rank transform-Tests (RT) und normal scores-Tests (INT)

Bei dem *Rank transform Test* werden die Werte der abhängigen Variablen x über alle Messwiederholungen und Gruppen hinweg, also z.B. $N*J$ Werte, in Ränge R(x) gewandelt, um mit diesen dann eine „normale“ parametrische Varianzanalyse zu rechnen. Bei dem *normal score*- bzw. *inverse normal transform-Verfahren* (INT) werden die Ränge R(x) darüber hinaus noch in normal scores umgerechnet:

$$nscore_m = \Phi^{-1}(R(x_m)/(M+1))$$

wobei M die Anzahl aller Werte ist, also z.B. $M=N*J$. Mit diesen scores wird dann eine „normale“ parametrische Varianzanalyse gerechnet. In beiden Fällen sollte man einen Test auf Spherizität, z.B. den Mauchly-Test, durchführen, oder direkt die korrigierten F-Tests von Huynh & Feldt benutzen, auch wenn die Spherizität gegeben ist. Beide Verfahren sollen wieder am Beispieldatensatz 4 demonstriert werden. Die Ergebnisse zeigen, dass das INT-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 Funktion `ezANOVA` durchgeführt, um HF- oder GG-adjustierten Ergebnisse zu erhalten, wenn auch die Darstellung der Werte unter Verwendung von `e-`.. das Lesen erschwert:

```
library(ez)
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 6-4a

Auch wenn der Mauchly-Test keine Signifikanzen zeigt, wird vielfach empfohlen, die korrigierten F-Tests von Huynh-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 5.3.3. Die Varianzanalyse enthält allerdings hier zusätzlich Tests des Faktor `Geschlecht`. Die Analyse wird wieder mit `ezANOVA` durchgeführt. Die Sphärizität war schon in Kapitel 5.3.3 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 Spherizität (Varianzhomogenität), danach die Ergebnisse der Varianzanalyse für den Effekt des Gruppierungsfaktors und zuzletzt die Effekte des Messwiederholungsfaktors (Innersubjekteffekte). Bei diesen wird vielfach empfohlen, die Resultate aus der Zeile „Huynh-Feldt“ abzulesen, auch wenn der entsprechende Mauchly-Test keine Signifikanzen aufweist.

Mauchly-Test auf Sphärizität						
Innersubjekteffekt	Mauchly W	Approxi. Chi-Quadrat	df	Sig.	Epsilon	
					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		Quadrat-summe	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 6-4b

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 `nc` 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 aus den `t1,t2,t3` die abhängige Variable `score` gebildet.
- Über `Aggregate` wird die Anzahl der Werte `nc` 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 `nscore` die Variablen `nscore.1`, `nscore.2`,...
- Schließlich kann die parametrische Varianzanalyse auf die Variablen `nscore.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 x wie beim o.a. RT-Verfahren über alle *N* Merkmalsträger und alle *J* Messwiederholungen hinweg in Ränge R(*x*) (Wilcoxon-Ränge) transformiert. Da die F-Tests hier nicht interessieren, ist auch eine Überprüfung der Spherizität bei den Puri & Sen-Tests nicht erforderlich, allerdings eine Überprüfung der Gleichheit der Varianzen der Messwiederholungsvariablen ratsam, wie sie in den Abschnitten 5.3.4 und 5.4.3 vorgenommen worden war.

Da zum einen die Konstruktion der χ^2-Tests beim Puri & Sen-Test nicht ganz trivial ist, und zum anderen diese χ^2-Tests in gleicher Weise 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 Messwiederholungsfaktoren 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 R(*x*) wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle werden folgende χ^2-Tests aufgestellt:
  Für die Effekte ohne Messwiederholungsfaktoren, z.B. A, B, A*B (vgl. Formel 2-6b):

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

  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):

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

  wobei
  - SS_{Effekt} die Streuungsquadratsumme (Sum of Squares) des zu testenden Effektes,
  - $MS_{zwischen}$ die Varianz der gesamten Zwischensubjektstreuung (MS, Mean Square), die die Streuung aller Gruppierungsfaktoren und deren Interaktionen sowie der damit verbundenen Fehlerstreuung beinhaltet,
  - SS_{Fehler} die Streuungsquadratsumme des zum Effekt gehörenden Fehlers ist sowie
  - SS_X die Streuungsquadratsummen aller Effekte, die SS_{Fehler} als Fehlerterm haben, also insbesondere der zu testende Effekt SS_{Effekt} sowie Interaktionen mit allen Gruppierungsfaktoren,

- df_X die entsprechenden Freiheitsgrade.

(Der Nenner der χ^2-Tests für die Messwiederholungsfeffekte entspricht genau $MS_{innerhalb}$, also der Varianz innerhalb der Versuchspersonen.)

- Die χ^2-Werte sind dann in den Tafeln für den χ^2-Test auf Signifikanz zu überprüfen, wobei die Freiheitsgrade die Zählerfreiheitsgrade (df_{Effekt}) des entsprechenden F-Tests sind.
- Die χ^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:

$$F = \frac{(N-1)\chi^2}{df_X + df_{Fehler} - \chi^2}$$

wobei df_X und df_{Fehler} die o.a. Freiheitsgrade sind.

Dazu ein paar Beispiele für die χ^2-Tests, wobei letztlich nur die Nenner interessieren, da der Zähler einheitlich die Streuung des zu testenden Effekts SS_{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 χ^2-Werte erforderlichen Streuungsquadratsummen aus der Tabelle 6-4a (Kapitel 6.3) zu entnehmen: Die Spalten `SSn` enthalten die SS_{Effekt} , die Spalten `SSd` die SS_{Fehler} und natürlich `DFn` und `DFd` die dazugehörigen Freiheitsgrade. In diesem Fall ist es am einfachsten, die χ^2-Werte daraus „mit der Hand“ auszurechnen:

$$MS_{zwischen} = \frac{53,3 + 701,8}{1+8} = 83,9$$

$$\chi^2_{Geschlecht} = \frac{53,3}{83,9} = 0,635$$

$$\chi^2_{Zeit} = \frac{698,6}{(698,6 + 501,3 + 249,9)/(2+2+16)} = 9,64$$

$$\chi^2_{Interaktion} = \frac{501,3}{(698,6 + 501,3 + 249,9)/(2+2+16)} = 6,915$$

Ergebnisse 6-1

Die für die Tests erforderlichen Freiheitsgrade enstsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse, d.h. der Spalte `DFn`. Die 5%- bzw. 1%-Schranken der χ^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) dazu benutzt werden:

```
attach("path/anova.lib")
np.anova(score~Geschlecht*Zeit+Error(Vpn/Zeit),winer518t,method=2)
```

```
Puri & Sen tests
                          Df Sum Sq  Chisq Pr(>Chi)
Geschlecht                 1  53.33 0.6356 0.425301
Residuals Btw.Subj         8 701.83
Zeit                       2 698.60 9.6370 0.008079 **
Geschlecht:Zeit            2 501.27 6.9148 0.031511 *
Residuals Zeit            16 249.97
```

mit SPSS:

Zunächst die relativ einfache Durchführung der klassischen Puri & Sen-Tests: Dazu genügt es, die für die Berechnung der χ^2-Werte erforderlichen Streuungsquadratsummen aus der Tabelle 6-4b (Kapitel 6.3) zu entnehmen: Die Spalten `Quadratsummen` enthalten die SS_{Effekt} bzw. SS_{Fehler} und natürlich `df` die dazugehörigen Freiheitsgrade. Die χ^2-Werte sind daraus „mit der Hand“ auszurechnen. Die Berechnung der χ^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 and damit eine parametrische Varianzanalyse mittels `ezANOVA` durchgeführt.

```
library(ez)
mydata5t<-within(mydata5t,RFehler<-rank(Fehler))
ezANOVA(mydata5t, RFehler, Vpn, within=.(Medikament,Aufgabe),
        between=Geschlecht, detailed=T)
```

```
                      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 χ^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 χ^2-Werte werden daraus „per Hand“ ausgerechnet (zu deren Aufbau vgl. den Anfang von Kapitel 6.4). Das Ergebnis ist der Übersicht der Ergebnisse 6-2 im nachfolgenden Abschnitt zu SPSS wiedergegeben.

Die für die Tests erforderlichen Freiheitsgrade enstsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse. Die 5%- bzw. 1%-Schranken der χ^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) dazu benutzt werden:

```
attach("path/anova.lib")
np.anova(Fehler~Geschlecht*Medikament*Aufgabe+
                Error(Vpn/(Medikament*Aufgabe)), mydata5t, method=2)
```

mit der folgenden Ausgabe:

```
Puri & Sen tests
                                        Df Sum Sq   Chisq Pr(>Chi)
Geschlecht                               1 3741.1  3.4505 0.063234 .
Residuals Btw.Subj                       6 3848.5
Medikament                               2 5419.1 12.1218 0.002332 **
Geschlecht:Medikament                    2  159.3  0.3562 0.836849
Residuals Medikament                    12 1574.5
Aufgabe                                  2 8037.8 11.6128 0.003008 **
Geschlecht:Aufgabe                       2  523.6  0.7565 0.685072
Residuals Geschlecht:Aufgabe            12 2513.0
Medikament:Aufgabe                       4 1099.7  9.0831 0.059056 .
Geschlecht:Medikament:Aufgabe            4  189.2  1.5625 0.815518
Residuals Geschlecht:Medikament:Aufgabe 24 2585.3
```

mit SPSS:

Nach Transformation der abhängigen Variablen `Fehler` in Ränge `RFehler`, wozu zunächst wieder eine Umstrukturierung des Datensatzes erforderlich ist, 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 χ^2-Test benutzt wird, spielt die Spheriztät 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 χ^2-Werte „per Hand“ berechnet (zu deren Aufbau vgl. den Anfang von Kapitel 6.4).

$$\chi^2_{Geschlecht} = \frac{3741}{1084,2} = 3,45$$

$$MS_{innerhalb(Medikamente))} = \frac{5419,06 + 159,2 + 1574,5}{2+2+12} = 447,0$$

$$\chi^2_{Medikamente} = \frac{5419,06}{447,0} = 12,12$$

$$\chi^2_{Medikamente \times Geschlecht} = \frac{159,2}{447,0} = 0,36$$

$$MS_{innerhalb(Aufgabe)} = \frac{8037,75 + 523,6 + 2513}{2+2+12} = 692,1$$

$$\chi^2_{Aufgabe} = \frac{8037,75}{692,1} = 11,62$$

$$\chi^2_{Aufgabe \times Geschlecht} = \frac{523,6}{692,1} = 0,755$$

$$MS_{innerhalb(Interaktion)} = \frac{1099,7 + 189,17 + 2585,3}{4+4+24} = 121,06$$

$$\chi^2_{Interaktion} = \frac{1099,7}{121,06} = 9,08$$

$$\chi^2_{Interaktion \times Geschlecht} = \frac{189,17}{121,06} = 1,56$$

Ergebnisse 6-2

Die für die Tests erforderlichen Freiheitsgrade enstsprechen den Zählerfreiheitsgraden der o.a. parametrischen Varianzanalyse. Die 5%- bzw. 1%-Schranken der χ^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.

6. 5 Verallgemeinerte Kruskal-Wallis-Friedman-Tests (KWF)

Die Rangtransformation werden beim KWF-Verfahren anders als beim Puri & Sen- und dem RT-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) χ^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 Spherizität 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, 2023b).

Folgende Schritte sind für eine Analysevariable `x` durchzuführen:

- Für die Analyse-Variable *x* (Variablen $x_1,...,x_J$) für jede Erhebungseinheit (Versuchsperson) *m* die Summe aller *J* Messwiederholungen $\texttt{Sum}_m$ errechnen.
- Diese Summen $\texttt{Sum}_m$ in Ränge $R(\texttt{Sum}_m)$ 1,...,*N* umrechnen.
- Für jede Erhebungseinheit (Versuchsperson) *m* werden die Werte $x_{m1},...,x_{mJ}$ in Friedman-Ränge (1,...,*J*) transformiert und ergeben $R(x_{m1}),...,R(x_{mJ})$.
- Für jede Erhebungseinheit *m* und Messwiederholung j=1,...,*J* berechnen von $(R(\texttt{Sum}_m) - 1) * J + R(x_{mj})$
- Mit diesen Rängen wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Schließlich werden dieselben χ^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, weiter unten dann etwas bequemer mit der Funktion `np.anova` (vgl. Anhang 3). Die ersten Schritte sind weitgehend dieselben wie in Kapitel 5.1.2. Zusätzlich ist am Anfang in `winer518` erforderlich:

- Die Summe der Variablen `t1,.,t3` errechnen und diese in Ränge (`Rsum`) wandeln.

Nach der Umstrukturierung in `winer518t` noch folgende Schritte:

- Die Messwiederholungsvariablen pro Vpn in Friedman-Ränge `Rscore` umrechnen.
- Aus `Rsum` und `Rscore` die zu analysierende Variable `Ry` bilden.
- Schließlich wird die Anova mit `aov` oder `ezANOVA` durchgeführt. Falls die χ^2-Werte „mit der Hand“ ausgerechnet werden, empfiehlt sich die Verwendung von `aov`.

```
Rsum     <- rank(rowSums(winer518[,3:5]))
winer518 <- cbind(Vpn=1:10, 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)
```

```
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 χ^2-Werte aus den Spalten „Sum Sq“ und „Df“ „mit der Hand“ auszurechnen:

$$MS_{zwischen} = \frac{24,3 + 2176,2}{1 + 8} = 244,5$$

$$\chi^2_{Geschlecht} = \frac{24,3}{244,5} = 0,1$$

$$\chi^2_{Zeit} = \frac{8,6}{(8,6 + 7,4 + 2,0)/(2 + 2 + 16)} = 9,56$$

$$\chi^2_{Interaktion} = \frac{7,4}{(8,6 + 7,4 + 2,0)/(2 + 2 + 16)} = 8,22$$

Ergebnisse 6-3

Die für die Tests erforderlichen Freiheitsgrade enstsprechen den Zählerfreiheitsgraden der parametrischen Varianzanalyse. Die 5%- bzw. 1%-Schranken der χ^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.

Soll dagegen die Berechnung in R programmiert werden, ist `ezANOVA` vorzuziehen, 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:

```
library(ez)
winer518t <- cbind(winer518t,Rscore,Ry)
ezANOVA(winer518t,Ry,Vpn,within=Zeit,between=Geschlecht,detailed=T)
```

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 `SSn` die Sum of Squares des jeweiligen Effekts und `SSd` die Streuung des dazugehörenden Fehler- (Residuen) Terms. Bei diesem vergleichsweise einfachen Design sind die χ^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 χ^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.

Etwas bequemer ist die Analyse mittels der Funktion `np.anova` (vgl. Anhang 3). Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`winer518t`). Eingabe und Ausgabe:

```
attach("path/anova.lib")
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 (`RSum`).
- Umstrukturieren des Datensatzes, so dass aus den 3 Messwiederholungen jeweils 3 Fälle erzeugt werden. Das ist im Anhang 1.1.1 ausführlich beschrieben.
- Pro Vpn aus den Werten von `score` die Ränge `Rscore` errechnen.
- Aus `RSum` und `Rscore` die zu analysierende Variable `Ry` errechnen.
- Zurücktransformieren des Datensatzes, wobei aus `Ry` für die 3 Zeitpunkte die Variablen `Ry.1, Ry.2, Ry.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 (Zwischensubjekteffekte). 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 χ^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 6-3).

Die Umrechnung der χ^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. Dies wird am Beispieldatensatz 5 (`mydata5`) gezeigt.

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 χ^2-Werte aus der Lösung mit SPSS.

```
attach("path/anova.lib")
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 (Zwischensubjekteffekte). Da eine Prüfung der Spherizität 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 χ^2-Werte (vgl. dazu deren Aufbau am Anfang von Kapitel 6.4):

$$MS_{zwischen} = \frac{5832 + 24421,5}{1+6} = 4321,93$$

$$\chi^2_{Geschlecht} = \frac{24421,5}{4321,93} = 1,35$$

$$MS_{innerhalb(Medikamente))} = \frac{111,06 + 0,9 + 27,38}{2+2+12} = 8,709$$

$$\chi^2_{Medikamente} = \frac{111,06}{8,709} = 12,75$$

$$\chi^2_{Medikamente \times Geschlecht} = \frac{0,9}{8,709} = 0,10$$

$$MS_{innerhalb(Aufgabe)} = \frac{150,58 + 2,58 + 62,17}{2+2+12} = 13,458$$

$$\chi^2_{Aufgabe} = \frac{150,58}{13,458} = 11,19$$

$$\chi^2_{Aufgabe \times Geschlecht} = \frac{2,58}{13,458} = 0,19$$

$$MS_{innerhalb(Interaktion)} = \frac{26,42 + 1,71 + 45,21}{4+4+24} = 2,292$$

$$\chi^2_{Interaktion} = \frac{26,42}{2,292} = 11,53$$

$$\chi^2_{Interaktion \times Geschlecht} = \frac{1,71}{2,292} = 0,75$$

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 χ^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`:

```
attach("path/anova.lib")
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 KWF weitgehend identisch ist. Bei den Rechenvorgängen ist lediglich zu beachten, dass die bei van der Waerden übliche Transformation in normal scores hier entfällt und die kombinierten Ränge sich über

```
compute Ry=(Rsum-1)*4 + Rscore.
```

errechnen. Die Bildung der χ^2-Werte erfolgt bei beiden Verfahren nach demselben Prinzip.

6. 6 van der Waerden-Tests (vdWS)

Das Verfahren von van der Waerden verläuft zunächst ähnlich dem KWF-Verfahren (vgl. Abschnitt 6.5). D.h. zum einen erhalten die einzelnen Erhebungseinheiten (Vpn) *m* Ränge R($\mathtt{Sum}_m$) entsprechend der Summe der Messwiederholungen $\mathtt{Sum}_m$, und zum anderen werden die Werte der Messwiederholungen pro Fall analog dem Friedman-Test in Friedman-Ränge R($\mathtt{x}_{mj}$) transformiert. Die Ränge werden jeweils in normal scores umgerechnet. Beide scores werden addiert. Schließlich werden die χ^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 Spherizität 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, 2023b).

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 $x_1,...,x_J$) pro Erhebungseinheit *m* die Summe Sum_m aller *J* Messwiederholungen errechnen.
- Diese Summen Sum_m in Ränge $R(Sum_m)$ 1,...,*N* umrechnen.
- Umrechnung der Rangsummen $R(Sum_m)$ in normal scores:

  $$n(sum)_m = \Phi^{-1}(R(sum_m)/(N+1)) \ ,$$

  wobei *N* die Anzahl der Erhebungseinheiten ist.
- Für jede Erhebungseinheit (Versuchsperson) *m* werden die Werte $x_{m1},...,x_{mJ}$ in Friedman-Ränge (1,...,*J*) transformiert und ergeben $R(x_{m1}),..., R(x_{mJ})$.
- 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,...,*J*

  $nsx_{mj} = nsum_m + nscore_{mj}$

  berechnen.
- Mit diesen normal scores nsx_{mj} wird eine parametrische Varianzanalyse mit Messwiederholungen durchgeführt.
- Auf Basis der Anova-Tabelle werden χ^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 Spherizität kann entfallen, da hier χ^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 χ^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 χ^2-Werte:
Aus dem oberen Teil der Anova-Tabelle ist zu entnehmen:

$$MS_{zwischen} = \frac{0,368 + 17,833}{1 + 8} = 2,022$$

$$\chi^2_{Geschlecht} = \frac{0,368}{2,022} = 0,182$$

Aus dem unteren Teil der Anova-Tabelle ist zu entnehmen:

$$\chi^2_{Zeit} = \frac{3,847}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,847}{0,4044} = 9,513$$

$$\chi^2_{Interaktion} = \frac{3,331}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,331}{0,4044} = 8,24$$

Ergebnisse 6-5

Die 5%- bzw. 1%-Schranken der χ^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) 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:

```
attach("path/anova.lib")
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 χ^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:

$$MS_{zwischen} = \frac{0,368 + 17,833}{1 + 8} = 2,022$$

$$\chi^2_{Geschlecht} = \frac{0,368}{2,022} = 0,182$$

Aus Tabelle 6-9b ist zu entnehmen:

$$\chi^2_{Zeit} = \frac{3,847}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,847}{0,4044} = 9,513$$

$$\chi^2_{Interaktion} = \frac{3,331}{(3,847 + 3,331 + 0,910)/(2 + 2 + 16)} = \frac{3,331}{0,4044} = 8,24$$

Die 5%- bzw. 1%-Schranken der χ^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 (`winer568`) demonstriert werden. Die Überprüfung der Spherizität kann wieder entfallen, da hier χ^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 Ergbnis. Dagegen wurden diese Effekte in der ART- wie auch in der ART+INT-Analyse (Kapitel 6.5.3) als hochsignifikant ausgewiesen. Die gleichen signifikanten Ergbenisse erhielte 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 χ^2-Werte aus der Lösung mit SPSS. Ausgangsbasis ist wieder der in Kapitel 5.1.2 erstellte Dataframe `winer568t`.

```
attach("path/anova.lib")
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 Gruppierungsfaktoren A und B (Zwischensubjekteffekte), danach die Tabelle für alle Tests, bei denen die Messwiederholung Zeit involviert ist (Innersubjekteffekte). Da die Spherizität 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		Quadrat-summe	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 χ^2-Werte aus den o.a. Quadratsummen:

$$MS_{zwischen} = \frac{4,761 + 16,071 + 0,12 + 10,147}{1 + 1 + 1 + 8} = 2,827$$

$$\chi^2_A = \frac{4,761}{2,827} = 1,68$$

$$\chi^2_B = \frac{16,071}{2,827} = 5,68$$

$$\chi^2_{A \times B} = \frac{0,12}{2,827} = 0,04$$

$$MS_{innerhalb} = \frac{15,9 + 0,83 + 0,23 + 0,08 + 0,49}{3 + 3 + 3 + 3 + 24} = 0,487$$

$$\chi^2_{Zeit} = \frac{15,9}{0,487} = 32,65$$

$$\chi^2_{A \times Zeit} = \frac{0,83}{0,487} = 1,70$$

$$\chi^2_{B \times Zeit} = \frac{0,23}{0,487} = 0,47$$

$$\chi^2_{A \times B \times Zeit} = \frac{0,08}{0,487} = 0,16$$

Ergebnisse 6-6

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 χ^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 Sphärizität kann wieder entfallen, da hier χ^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, bzw. aus der u.a. Ergebnistabelle 6-7:

```
attach("path/anova.lib")
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 Abschnitt 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.
compute nSum=Idf.normal(RSum/9,0,1).            #  9 ~ N+1
```

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).     # 10 ~ J+1
compute nscore = nFehler + nSum.
execute.

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 (Zwischensubjekteffekte). Da eine Prüfung der Sphärizität hier entfällt, interessieren in der Anova-Tabelle nur die Zeilen mit den unkorrigierten F-Tests.

Tests der Zwischensubjekteffekte

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	,001	1	,001	172,548	,000
Geschlecht	6,380	1	6,380	1,113	,332
Fehler	34,406	6	5,734		

Tests der Innersubjekteffekte

Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angenommen	10,145	2	5,073	26,492	,000
Medikament * Geschlecht	Sphärizität angenommen	,097	2	,049	,254	,780
Fehler(Medikament)	Sphärizität angenommen	2,298	12	,191		
Aufgabe	Sphärizität angenommen	13,704	2	6,852	16,239	,000
Aufgabe * Geschlecht	Sphärizität angenommen	,237	2	,119	,281	,760
Fehler(Aufgabe)	Sphärizität angenommen	5,063	12	,422		
Med * Aufg	Sphärizität angenommen	2,700	4	,675	4,514	,007
Med * Aufg * Geschl	Sphärizität angenommen	,133	4	,033	,222	,923
Fehler(Med*Aufg)	Sphärizität angenommen	3,589	24	,150		

Nun zur Berechnung der χ^2-Werte aus den o.a. Quadratsummen wie in Abschnitt 6.4 erläutert und ähnlicherweise in 6.4.2 bereits durchgeführt:

$$MS_{zwischen} = \frac{0,411 + 6,453}{1 + 6} = 0,98$$

$$\chi^2_{Geschlecht} = \frac{0,411}{0,98} = \ddot{0},419$$

$$MS_{innerhalb(Medikamente))} = \frac{10,145 + 0,097 + 2,298}{2 + 2 + 12} = 0,784$$

$$\chi^2_{Medikamente} = \frac{10,145}{0,784} = 12,94$$

$$\chi^2_{Medikamente \times Geschlecht} = \frac{0,097}{0,784} = 0,124$$

$$MS_{innerhalb(Aufgabe)} = \frac{13,704 + 0,237 + 5,063}{2 + 2 + 12} = 1,188$$

$$\chi^2_{Aufgabe} = \frac{13,704}{1,188} = 11,19$$

$$\chi^2_{Aufgabe \times Geschlecht} = \frac{0,237}{1,188} = 0,199$$

$$MS_{innerhalb(Interaktion)} = \frac{2,7 + 0,133 + 3,589}{4 + 4 + 24} = 0,20$$

$$\chi^2_{Interaktion} = \frac{2,7}{0,20} = 13,5$$

$$\chi^2_{Interaktion \times Geschlecht} = \frac{0,133}{0,20} = 0,665$$

Ergebnisse 6-7

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 α_i lässt sich nicht vom Personeneffekt π_m trennen. Daher muss hier ein anderer Weg einge-

schlagen 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 Kovarianzamatrizen und nichtnormalen Daten hingewiesen, dennoch haben Kowalchuk et al. (2003) gezeigt, dass dies nicht mehr gilt, wenn die Kovarianzamatrizen 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 2.4 sowie an die Bemerkungen in Kapitel 5.4.4 erinnert, wonach empfohlen wird, nach der Berechnung der Ränge diese noch in normal scores (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 Hauptffekts γ_j der Messwiederholungen sowie des Personeneffekts π_m aus der Kriteriumsvariablen x:

  $$e_{jm} = x_{jm} - (\overline{p_m} + \overline{c_l} - \bar{x})$$

  alternativ per *standard approach* (vgl. Formel 2-5): Berechnung der Residuen e_{jm} wie in Kapitel 6.2 beschrieben, anschließend Addition des „reinen“ Interaktionseffekts:

  $$e_{jm} = e_{jm} + \overline{ac}_{ij} - (\overline{p_m} + \bar{c}_j - \bar{x})$$

  wobei $\overline{c_l}$, $\overline{ac}_{il}$ die Mittelwerte von C bzw. AC sowie $\bar{p}_m$ und $\bar{x}$ 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 Interaktionsffekt 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). Zunächst die „elementare“ Berechnung, anschließend bequemer mit der R-Funktion `art3.anova`.

Für die Berechnung der Residuen e_{jm} wird hier der o.a. „naive approach“ gewählt. Dazu müssen zunächst die Effekte γ_j (mb) und π_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) angewandt werden. Der Aufruf ist praktisch identisch mit dem der Standardfunktion `aov`. Basis ist auch hierfür der umstrukturierte Datensatz (`winer518t`). Eingabe und Ausgabe:

```
attach("path/anova.lib")
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 `ek` in normal scores `nsek` 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 „Spherizität angenommen“ 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* (`nc`) 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 5.4.5 errechnet werden.
- Die Interaktion C*D, ein reiner Messwiederholungseffekt, wird mit der ART wie in Kapitel 5.4.5 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 γ_j (mb) und π_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	4	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 weitegehend 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 „Spherizität angenommen“ relevant sind. Demnach liegt für die Interaktion keine Signifikanz vor.

Tests der Innersubjekteffekte						
Quelle		Quadrat summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	7,750	2	3,875	,046	,955
Medikament * Geschlecht	Sphärizität angen.	110,583	2	55,292	,656	,537
Fehler(Medikament)	Sphärizität 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* (`nc`) 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		Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Medikament	Sphärizität angen.	,088	2	,044	,031	,970
	Huynh-Feldt	,088	1,357	,065	,031	,922
Medikament * Geschlecht	Sphärizität angen.	1,511	2	,756	,532	,600
	Huynh-Feldt	1,511	1,357	1,114	,532	,540
Fehler(Medikament)	Sphärizität 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 α_i , β_j bzw. γ_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 ist zu den Residuen zunächst der Interaktionseffekt zu addieren, und danach der Messwiederholungseffekt γ_l sowie. der Personeneffekt π_m zu subtrahieren:

  $$e_m\ (a) = e_m + \overline{ac}_{il} - (\overline{p_m} + \bar{c}_l - \bar{x})$$

- Für die Interaktion B*C wird analog A*C vorgegangen.

  $$e_m\ (b) = e_m + \overline{bc}_{jl} - (\overline{p_m} + \bar{c}_l - \bar{x})$$

  wobei $\bar{c}_l$, $\overline{ac}_{il}$, $\overline{bc}_{jl}$ 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 Interaktionsffekt 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 R(x), 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 wird wieder der umstrukturierte Dataframe `winer568t` aus Kapitel 5.1.2 genommen. Zunächst die „elementare“ Berechnung, anschließend die bequemere Möglichkeit mit der R-Funktion `art3.anova`. Damit werden für die Analyse der Interaktionen A*C und B*C die Residuen (`ek`) des Modells A*B*C+Vpn 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 `V1,.,V4` 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) bequem durchführbar. Basis ist auch hier der umstrukturierte Datensatz `winer568t`. Nachfolgend Eingabe und Ausgabe:

```
attach("path/anova.lib")
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 und bequemer kann die Analyse auf Basis des ART+INT-Verfahrens für alle Effekte auch bequem mittels der Funktion `art3.anova` durchgeführt werden:

```
attach("path/anova.lib")
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,.,rema.4` bzw. `remb.1,.,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		Quadrat-summe	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 `V1,.,V4` 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` die Konventionen für die Angabe der Variablennamen recht komplex sind. Daher empfiehlt es sich, beide Funktionen in der Form `with(winer518t,...)` aufzurufen:

```
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 über `group.name` bzw. `time.name` in "..." als frei wählbarer Name des Faktors für die Ausgabe. Diese Funktion gibt noch zusätzlich einen Interaktionsplot aus, allerdings der relativen Effekte (vgl. Kapitel 2.4) anstatt der Mittelwerte, da sich ja die Hypothesen auf erstere beziehen:

Die Ergebnisse von nparLD:

```
Wald-Type Statistc (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 Statistc (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 denVorteil, dass man über die Funktion `round` die Zahlendarstellung der Art xxxe-nn ändern kann, wie oben gezeigt.

```
                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. Mit `no.subf` wird hier die Anzahl der Messwiederholungsfaktoren (hier Zeit) angegeben.

```
library(MANOVA.RM)
RM(score~Geschlecht*Zeit,winer518t,"Vpn",iter=100,no.subf=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"

p-values resampling:
                Perm (WTS)
Geschlecht      "0.42"
Zeit            "0.02"
Geschlecht:Zeit "<0.001"
```

6. 9 Robuste Mittelwerte

Anstatt arithmetischer Mittel und Varianzen werden hier getrimmte Mittelwerte und winsorisierte Varianzen verwendet. Die Methodik wurde in Abschnitt 2.9 kurz erläutert.

mit R:

Wilcox (Mair & Wilcox, 2019) bietet dazu im Paket `WRS2` mit der Funktion `bwtrim` ein Verfahren für ein 2-faktorielles Split-Plot-Design an. Es kompensiert allerdings nicht heterogene Varianzen der Messwiederholungsvariablen und die damit verbundene fehlende Spherizität wie etwa das in Abschnitt 5.3.11 beschriebene, sondern lediglich nicht-normalverteilte Residuen. Der Anteil der getrimmten (ausgeschlossenen Werte, standardmäßig 0.2, jeweils `tr` Prozent am unteren und `tr` Prozent am oberen Ende) kann über den Parameter `tr=...` bestimmt werden. Es wird der Datensatz `mydata11` (vgl. Abschnitt 11.4) verwendet, der nicht-normalverteilte Residuen aufweist, aber auch ungleiche Varianzen der Messwiederholungen. Diese sollten jedoch durch das Trim-Verfahren abgemildert werden

(vgl. Lix & Keselman, 1998). Der Datensatz wird hier in der transformierten Form verwendet (siehe Abschnitt 5.1.2). Auch hier ist eine Fallkennung (`Id`) erforderlich.

```
library(WRS2)
bwtrim(x~A*Zeit, Id, data=mydata11t, trim=0.2)
```

```
        value df1     df2 p.value
A      0.2186   3 10.2492  0.8813
Zeit   6.8099   3 10.7790  0.0076
A:Zeit 1.4876   9 11.4427  0.2600
```

6. 10 Verfahren ohne Spherizitäts-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, komplett verteilungsfreie, auf sphärischen Rängen basierende Verfahren sowie das Verfahren für nichthomogene Varianzen von Welch & James. Die multivariaten 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 für die Analyse gebildeten Differenzen, gefordert, nicht jedoch die Spherizität. 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 (vgl. Kapitel 3.5). 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 der Datensätze `winer568` oder `mydata5` 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 Beipiel vorgestellt worden waren. Bei der Besprechung der Voraussetzungen in Kapitel 5.2 sowie in 5.3.9 wurde bereits darauf hingewiesen, dass die parametrischen multivariaten Tests 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. 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. Zunächst wird an das Beispiel aus Abschnitt 5.3.9 angeknüpft, bei dem die Berechnung der Differenzen der 4 Messwiederholungsvariablen `V1,..,V4` erforderlich ist: `V4-V3, V3-V2` und `V2-V1`. 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 Wilks:

```
with(winer568,anova(lm(cbind(V4-V3,V3-V2,V2-V1)~A), test="Wilks"))
```

```
Analysis of Variance Table

            Df    Wilks approx F num Df den Df    Pr(>F)
(Intercept)  1 0.024088  108.039      3      8 8.205e-07 ***
A            1 0.032135   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"))
Aufgabe    <- factor(rep(c("1","2","3"),3))
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
```

Der Vollständigkeit wegen sei noch die Lösung mittels der Funktion `manova` für die o.a. 2-faktorielle Fragestellungvorgestellt.

```
library(car)
Anova(manova(cbind(V4-V3,V3-V2,V2-V1)~A,data=winer568),type=3,
      test="Wilks"
```

```
           Df    Wilks approx F num Df den Df    Pr(>F)
(Intercept)  1 0.024088  108.039      3      8 8.205e-07 ***
A            1 0.032135   80.316      3      8 2.590e-06 ***
Residuals   10
```

Abschließend noch einige Anmerkungen zu den nichtparametrischen Varianten. Für das Verfahren von Agresti & 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) `Vsum` von `V1,..,V4`, danach z.B. ein Kruskal-Wallis-Test:

```
attach("path/anova.lib")
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 5.3.9 beschrieben, die abhängige Variable, bestehend aus `V1`,..,`V4`, in Ränge umgerechnet werden:

```
mat568 <- matrix(rank(as.vector(as.matrix(winer568[,3:6]))),12,4)
```

Wie dort beschrieben können die Differenzen `V4-V3`, `V3-V2` und `V2-V1` einfach durch die Differenz aus der Matrix der Variablen `V2,V3,V4` und der Matrix der Variablen `V1,V2,V3` 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[b]	3,000	8,000	,000
	Wilks-Lambda	,024	108,039[b]	3,000	8,000	,000
	Hotelling-Spur	40,514	108,039[b]	3,000	8,000	,000
	Größte charakteristische Wurzel nach Roy	40,514	108,039[b]	3,000	8,000	,000
Zeit * A	Pillai-Spur	,968	80,316[b]	3,000	8,000	,000
	Wilks-Lambda	,032	80,316[b]	3,000	8,000	,000
	Hotelling-Spur	30,118	80,316[b]	3,000	8,000	,000
	Größte charakteristische Wurzel nach Roy	30,118	80,316[b]	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 Beipiel 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 Methoden

In Kapitel 2.11.5 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 einige der 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. Ein Vergleich mittels Simulationen (siehe Lüpsen, 2024) der in Abschnitt 5.3.10 aufgeführten verschiedenen Modelle bzw. Funk-

tionen zeigte, dass von diesen lediglich das von Hallin & Paindaveine unbefriedigende Ergebnisse erzielte. Von den beiden Methoden, die auf Puri & Sen zurückgehen, ist die auf Rängen basierende Variante der vorzuziehen, die normal scores verwendet. Das Verfahren von Sirkiä ist relativ zuverlässig, bis auf den Test des Haupteffekts des Messwiederholungsfaktors bei kleinen n_i. Alle schneiden besser ab bei großen n_i und kleinem J als umgekehrt. Auch hier müssen für eine univariate Analyse mit Messwiederholungen die Differenzen zweier aufeinander folgenden Variablen wie im vorigen Abschnitt gebildet werden.

mit R:

Bei den hier vorgestellten Funktionen werden zwar der Haupteffekt des Messwiederholungsfaktors und die Interaktion mit derselben Funktion getestet, 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. Das Prozedere wird wieder anhand des Datensatzes `winer568` mit den Variablen `V1,..,V4` 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) :

```
library(SpatialNP)
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`,

```
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.1sample.test` genommen werden, hier ohne den Parameter `g`:

```
library(MNM)
winer568 <- within(winer568, A <- factor(A))
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)
```

Im Paket `ICSNP` bietet die Funktion `rank.ctest` einen Test der Interaktion, für den 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 Sphärizität der Kovarianzmatrix noch deren Homogenität über die einzelnen Gruppen voraus. Damit ist es unproblematischer anzuwenden als die parametrischen Varianzanalysen unter Verwendung der ε-Korrekturen. Keselman, Carriere & Lix (1993) haben sich intensiv mit dem Verfahren von Welch & James auseinandergesetzt. 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 gravierenden 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 $n_i > 4*(J\text{-}1)$, 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 Sphärizität nicht erfüllt ist ($p < 0.01$). Dazu dient wieder die umstrukturierte Version `winer568t`.

```
library(ez)
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) erfolgen

(hier über `round` besser lesbar, indem über `$results` die Ausgabe auf die Testergebnisse beschränkt wird):

```
attach("path/anova.lib")
round(check.sphere(winer568[,3:6])$results,digits=3)
```

```
                     statistic    df p value
Mauchly test            21.204 5.000   0.001
Likelihood Ratio        23.991 9.000   0.004
John's test chisq       32.546 5.000   0.000
John's test beta         0.317 8.769   0.004
John-Nagao              17.121 9.000   0.027
Muirhead & Waternaud    45.603 9.000   0.000
compound symmetry       33.529 8.000   0.000
```

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.

```
attach("path/anova.lib")
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, 1969). Eines davon entspricht einer Übertragung des multivariaten Ansatzes des Messwiederholungsmodells (vgl. Kapitel 5.2) auf rangtransformierte Daten. Damit entfallen auch hier die entsprechenden Prüfungen von Voraussetzungen. Es hat aber auch Schwächen: zum einen ist es recht liberal für $n_i \leq 10$, zum anderen hat es eine relativ geringe Power.

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):

```
attach("path/anova.lib")
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
```

In der Ausgabe werden die Faktoren einfach mit „A“ und „B“ bezeichnet, d.h. in diesem Beispiel entspricht „A“ wirklich dem Faktornamen, und „B“ dem Faktor Zeit.

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 z.B. einen weiteren Versuch ohne Interaktionen starten, weil dadurch die zu schätzende Parameteranzahl deutlich reduziert wird, oder man versucht es in R mit einer anderen Funktion. 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 sowie die Varianz-Kovarianzmatrix der Parameterschätzungen, aber keine Anova-Tabelle. Diese kann vielfach über die Funktionen `gee.anova` und `gee.robanova` erzeugt werden (siehe Anhang 3). Gegebenenfalls muss

man aus diese, wie weiter unten in Kapitel 9.8 beschrieben, für einen Faktor einen Gesamttest mit der Hand ausrechnen. Wie in Kapitel 2.15 erwähnt, ist zwar die Struktur der Korrelationsmatrix ein wichtiger Bestandteil des Modells, allerdings ist der Einfluss relativ gering. Üblicherweise wird die Struktur über `corstr=...` vorgegeben. Die realistischsten sind `exchangeable` und `ar1`.

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. Als Modell für die Verteilung der abhängigen Variablen wird zunächst die Normalverteilung angenommen (`family=gaussian`), anschließend das Modell für ordinale Variablen (`family=poisson`). Ein- und Teil der Ausgabe, zunächst für den Standard-Aufruf:

```
library(gee)
erg <- gee(score~Geschlecht*Zeit,id=Vpn,family=gaussian,
           corstr="exchangeable",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 dazu ein Test auf linearen (`Zeit.L`) bzw. quadratischen Trend (`Zeit.Q`) ausgegeben, jeweils eine „naive“ und eine „robuste“ Ermittlung des Schätzfehlers, sowie ein daraus resultierender, annähernd normalverteilter z-Wert. Zweckmäßigerweise werden anschließend anova-like Tests durchgeführt, z.B. der Wald-Test mittels der Funktion `gee.anova` (vgl. Anhang 3) durchgeführt werden, die als Eingabe erwartet:

- 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 (für die F-Tests)

```
attach("path/anova.lib")
gee.anova(erg$coefficients,erg$"robust.variance",c(1,2,2),n=10)
```

```
     df    Chi P.Chi      F   P.F nerror err.invert
1  1  1  1.289 0.256  1.289 0.286      0          0
2  2  2 20.944 0.000 10.472 0.006      0          06
3  2  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 χ^2-Test für größere Stichproben konzipiert ist, sollte der F-Test (vorzugsweise) bei kleineren *N* angewandt werden. (Die beiden letzten Spalten protkollieren die Anzahl der Probleme bei der Berechnung des klassischen Wald-Tests.) Die alternative Funktion `gee.robanova` (vgl. Anhang 3) ist zwar im allgemeinen vorzuziehen, bietet allerdings keinen zusätzlichen
F-Test für kleinere *N*.

```
gee.robanova(erg$coefficients,erg$"robust.variance",c(1,2,2))
```

```
           df    Chi P.Chi
1 1.000000  1.289 0.256
2 1.983347 22.262 0.000
3 1.983347 17.201 0.000
```

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. Die Ergebnisse für die Koeffizienten (`Estimates`) weichen zwar deutlich ab, die anova-like-Tests zeigen aber vergleichbare Resultate an:

```
library(gee)
erg <- gee(score~Geschlecht*Zeit,id=Vpn,family=poisson,data=winer518t)
round((summary(erg)$coefficients),digits=4)
gee.anova(erg$coefficients,erg$"robust.variance",c(1,2,2),n=10)
```

```
                   Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)          1.3482     0.0954 14.1336      0.0936  14.4111
Geschlecht1          0.0989     0.0954  1.0366      0.0936   1.0570
Zeit.L              -0.5994     0.1653 -3.6266      0.1682  -3.5635
Zeit.Q              -0.0217     0.1652 -0.1312      0.1556  -0.1393
Geschlecht1:Zeit.L   0.0804     0.1653  0.4865      0.1682   0.4781
Geschlecht1:Zeit.Q  -0.4795     0.1652 -2.9034      0.1556  -3.0814
```

```
  df    Chi P.Chi     F   P.F nerror err.invert
1  1  1.117 0.291 1.117 0.318      0          0
2  2 15.641 0.000 7.820 0.013      0          0
3  2 14.159 0.001 7.080 0.017      0          0
```

Da es GEE-Funktionen gibt, deren Ausgabe nicht mit der Funktion `gee.anova` kompatibel ist, wird hier die (näherungsweise) Berechnung der anova-Tests „per Hand“ gezeigt. Aus den z-Werten der Kontraste ergibt sich (bei einem α von 0.05):

- Geschlecht: kein signifikanter Haupteffekt (z=`1.13519`)
  (kritischer Wert der Normalverteilung: 1.96).
- Zeit: χ^2 = `(-4.57625)`^{2} + `(-0.33241)`^{2} = `21.05` ist signifikant bei FG=2
  (kritischer Wert der χ^2-Verteilung: 6.0).
- Interaktion: χ^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.

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.
```

SPSS bietet für GENLIN eine Vielzahl an Parametern zur Steuerung der Analyse und der Ausgebe. Hier sollen nur wenige erwähnt werden. 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:

Parameter	Regressionskoeffizient 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
[Geschlecht=1]	-,200	1,1027	-2,361	1,961	,033	1	,856
[Geschlecht=2]	0[a]	.	.	.	.	.	.
[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[a]	.	.	.	.	.	.
[Geschlecht=1] * [Zeit=1]	-1,600	1,0198	-3,599	,399	2,462	1	,117
[Geschlecht=1] * [Zeit=2]	4,200	,9121	2,412	5,988	21,202	1	,000
[Geschlecht=1] * [Zeit=3]	0[a]	.	.	.	.	.	.
[Geschlecht=2] * [Zeit=1]	0[a]	.	.	.	.	.	.
[Geschlecht=2] * [Zeit=2]	0[a]	.	.	.	.	.	.
[Geschlecht=2] * [Zeit=3]	0[a]	.	.	.	.	.	.
(Skala)	3,233						

Den wichtigsten Teil der Ausgabe, die Anova-like Tests, enthält die Tabelle mit den Tests der Modelleffekte:

Tests der Modelleffekte			
Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	96.604	1	.000
Geschlecht	.590	1	.443
Zeit	44.526	2	.000
Geschlecht * Zeit	57.801	2	.000

Über die /EMMEANS Anweisungen werden die Mittelwertvergleiche über Kontraste ausgegeben, zusammen mit einem Test des jeweiligen Haupteffekts:
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

und für Zeit:

Individuelle Testergebnisse

Zeit Wiederholter Kontrast	Kontrastschätzer	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 signfikanten Effekt (p<0.001) anzeigt.

Fügt man z.B. noch

```
/EMMEANS  TABLES = Geschlecht*Zeit
  compare = Geschlecht
  contrast=pairwise.
```

an, so erhält man u.a. Tests des Faktors Geschlecht für die 3 Zeitpunkte, also quasi simple effect-Tests (vgl. Kapitel 10): für jeden Zeitpunkt 1, 2, 3 einen Test auf Unterschiede für den Faktor Geschlecht:

Gesamttestergebnisse

Zeit	Wald-Chi-Quadrat	df	Sig.
1	3,266	1	,071
2	17,857	1	,000
3	,033	1	,856

Auch hier kann ein anderes Verteilungsmodell gewählt werden, z.B. für eine diskrete, schief verteilte Variable *x*, indem oben in den Anweisungen /MODEL ersetzt wird durch:

```
/MODEL Geschlecht Zeit Geschlecht*Zeit
       DISTRIBUTION=POISSON   LINK=LOG
```

Von der Ausgabe wird hier nur die Anova-Tabelle, der Test der Modelleffekte, wiedergegeben, wobei sich die Ergebnisse nicht qualitativ von den o.a. auf der Normalverteilung basierenden unterscheiden:

Tests der Modelleffekte Typ III			
Quelle	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	190.075	1	.000
Geschlecht	1.023	1	.312
Zeit	13.004	2	.000
Geschlecht * Zeit	9.349	2	.000

Weitere Anwendungen für die Prozedur GENLIN, u.a. Zur Analyse von Modellen mit alternativen Kovarianz-/Korrelationsmatrizen, sind in Abschnitt 6.11.2 zu finden.

6. 10. 6 GLMM

In Kapitel 2.15 war darauf hingewiesen worden, dass die GLMM-Methode schwächere Voraussetzungen hat als die parametrische Varianzanalyse, die Anwendung allerdings problematisch ist, insbesondere wenn die Fallzahl nicht hinreichend groß ist. Zwei Vorteile von GLMM: Zum einen können Datensätze mit fehlenden Werten verarbeitet werden, zum anderen erlaubt es mehrere Verteilungsmodelle, so u.a. metrisch normalverteilt, diskret Polisson-verteilt und dichotom. 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` für `family=gaussian` und `glmer` für `family=poisson, binary` (Paket `lme4`)
- `glmmML` nur für `family=poisson, binary` (Paket `glmmML`)
- `glmmPQL` für `family=gaussian, poisson, binary` (Paket `MASS`)
- `mixed` für `family=gaussian, poisson, binary` (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`) aus dem Ergebnisobjekt für die Effekte varianzanalytische Tests auszugeben, u.a. den in Kapitel 9.8 erwähnten Typ II Wald-Test. Alternativ werden die varianzanalytischen Tests von Kenward-Rogers mittels der Funktion `anova` angeboten, die insbesondere für kleinere Stichproben zu bevorzugen sind. Bei anderen GLMM-Funktionen kann vielfach eine Anova-Tabelle über die Funktionen `gee.anova` und `gee.robanova` erzeugt werden (siehe Anhang 3). Gegebenenfalls muss man aus diesen, wie weiter unten in Kapitel 9.8 beschrieben, für einen Faktor einen Gesamttest mit der Hand ausrechnen.

Bei `lmer` und `glmer` hilft die Funktion `nlminb` des Optimierungspaket `optimx` die in 2.15 beschriebenen 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, zunächst für den Fall einer metrischen abhänigen Variablen, einmal mit Wald Typ II-Test und einmal mit dem Kenward-Roger-Test als Ergebnis:

```
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 ***
```

```
anova(erg, type=3, ddf="Kenward-Roger"))
```

```
Type III Analysis of Variance Table with Kenward-Roger's method
               Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
Geschlecht      0.621  0.6211     1     8  0.4717 0.5116204
Zeit           58.067 29.0333     2    16 22.0506 2.523e-05 ***
Geschlecht:Zeit 44.867 22.4333     2    16 17.0380 0.0001086 ***
```

Dieselbe Analyse mit den varianzanalytischen Tests von Kenward-Rogers bietet `mixed` aus dem `afex`-Paket, die auch andere anova-like-Tests anbietet:

```
library(afex)
mixed(score~Geschlecht*Zeit+(1|Vpn), data=winer518t, method="KR")
```

```
Model: score ~ Geschlecht * Zeit + (1 | Vpn)
Data: winer518t
           Effect    df         F p.value
1      Geschlecht  1, 8      0.47    .512
2            Zeit 2, 16 22.05 ***   <.001
3 Geschlecht:Zeit 2, 16 17.04 ***   <.001
```

Darüber hinaus können über `check_contrasts=T` Kontraste angefordert und mittels `summary` angezeigt werden (mehr dazu in Kapitel 9):

```
summary(mixed(score~Geschlecht*Zeit+(1|Vpn), data=winer518t,
        method="KR", check_contrasts=T)
```

```
                   Estimate Std. Error      df t value Pr(>|t|)
(Intercept)          4.2667     0.4853  8.0000   8.791 2.20e-05 ***
Geschlecht1          0.3333     0.4853  8.0000   0.687    0.512
Zeit.L              -2.4042     0.3629 16.0000  -6.626 5.83e-06 ***
Zeit.Q              -0.1633     0.3629 16.0000  -0.450    0.659
Geschlecht1:Zeit.L   0.5657     0.3629 16.0000   1.559    0.139
Geschlecht1:Zeit.Q  -2.0412     0.3629 16.0000  -5.625 3.80e-05 ***
```

Auch für GLMM gibt es die Möglichkeit, ein Modell für Häufigkeiten oder ordinale abhängige Variablen über die Option `family=poisson` zu analysieren. Dazu ist lediglich die Funktion `glmer` anstatt `lmer` aufzurufen, die diverse Verteilungsfamilien erlaubt, sowie die Optimierungsoptionen wie folgt 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.glm)
```

```
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 **
```

Auch `mixed` erlaubt Poisson- oder binär-verteilte *x*. Dann kann allerdings nur der likelihood-ratio-Test (LRT) für die Anova-like Tests verwendset werden:

```
library(afex)
mixed(score~Geschlecht*Zeit+(1|Vpn), data=winer518t, family=poisson,
      method="LRT")
```

```
Mixed Model Anova Table (Type 3 tests, LRT-method)

Model: score ~ Geschlecht * Zeit + (1 | Vpn)
Data: winer518t
Df full model: 7
           Effect df  Chisq p.value
1      Geschlecht  1   0.48    .489
2            Zeit  2   5.05    .080
3 Geschlecht:Zeit  2   4.03    .133
```

Da, wie oben angedeutet, eine Lösung nicht immer gesichert ist, andererseits den verschiedenen Funktionen auch verschiedene Algorithmen zugrunde liegen, ist man gelegentlich gezwungen, eine andere Funktion für die Analyse zu probieren. Deswegen wird hier zusätzlich die Anwendung von `glmmML` gezeigt, deren Ergbnis leider nicht so einfach mit den Funktionen `Anova` oder `anova` dargestellt werden kann. Statt dessen ist eine spezielle Funktion zur Ausgabe einer Anova-Tabelle anzuwenden: `gee.anova` für den in Kapitel 9.8 erwähnten Typ II Wald-Test oder `gee.robanova` für einen robusteren Wald-Test nach Fan & Zhang (2014). Letzterer ist i.a. der zuverlässigere. Beide werden vom Autor zur Verfügung gestellt (siehe Anhang 3). Diese benötigen als Parameter die Regressionskoeffizienten (`$coefficients`), die Kovarianzmatrix der Regressionskoeffizienten (`$variance`) sowie die Freiheitsgrade für die 3 Tests (i.a. Anzahl der Gruppen/Wiederholungen -1 für die beiden Haupteffekte sowie deren Produkt für die Interaktion), `gee.anova` allerdings noch das *N* für die Anwendung eines F-Tests. Beide werden hier wiedergegeben.

```
library(glmmML)
attach("path/anova.lib")
erg <- glmmML(score~Geschlecht*Zeit,cluster=Vpn,data=winer518t,
           family=poisson)

gee.robanova(erg$coefficients,erg$variance,c(1,2,2))
gee.anova(erg$coefficients,erg$variance,c(1,2,2),10)
```

```
        df    Ch  P.Chi
1 1.000000  0.806 0.369
2 1.900727 11.918 0.002
3 1.900727  7.832 0.018
```

```
  df    Chi P.Chi     F   P.F nerror err.invert
1  1  0.806 0.369 0.806 0.393      0          0
2  2 13.004 0.002 6.502 0.021      0          0
3  2  9.349 0.009 4.674 0.045      0          0
```

oben die χ^2-Tests des robusten Tests, unten die χ^2- zusammen mit den F-Tests (vorzugsweise für kleinere Stichproben) sowie Anzahl der Probleme bei der Berechnung des klassischen Wald-Tests.

Weitere Anwendungen für GLMM-Funktionen, u.a. Zur Analyse von Modellen mit alternativen Kovarianz-/Korrelationsmatrizen, sind in Abschnitt 6.11.2 zu finden.

mit SPSS:

In SPSS gibt es zwar die Prozedur GENLINMIXED für die GLMM-Methode, doch sie bricht für die Analyse von Versuchsplänen mit Messwiederholungen generell mit Fehlermeldungen ab, manchmal mit der Meldung, dass bei der Schätzung eine Matrix nicht „positiv definit“ ist, vielfach auch mit nicht näher spezifizierten Meldungen. Dieser Fehler ist schon seit Jahren SPSS gemeldet, allerdings noch nicht behoben worden

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.
```

6. 10. 7 GA- und IGA-Approximationen von Huynh

Huynh (1978) hat die bekannte $\tilde{\varepsilon}$-Adjustierung für die Freiheitsgrade des F-Tests verbessert: die GA-Approximation zur Kompensierung fehlender Spherizität 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 redziert. Funktionen hierfür sind nur in R verfügbar.

mit R:

Vom Autor werden dafür die Funktionen `iga` und `iga.anova` angeboten (Anhang 3). `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.

```
attach("path/anova.lib")
iga.anova(winer568[,3:6],winer568[,1])      # Eingabe im breiten Format
```

```
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)) # im langen Format
```

6. 10. 8 modifizierter Brown-Forsythe-Test (mBF)

Coombs & Algina (1992) sowie Vallejo et al. (2004) haben die bekannte Varianzanalyse-Methode von Brown-Forsythe für heterogene Varianzen auf Split-Plot-Designs verallgemeinert. Diese berücksichtigt sowohl fehlende Spherizität als auch heterogene Kovarianzmatrizen. Eine entsprechende Funktionen ist nur in R verfügbar.

mit R:

Vom Autor wird dafür die Funktion `mbf.f` angeboten (Anhang 3), die eine komplette Varianzanalyse durchführt. Dabei wird für den Test des Gruppierungsfaktors der „normale" Brown-Forsythe-Test (vgl. Kapitel 2.9.4) eingesetzt. 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.

```
attach("path/anova.lib")
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))
```

6. 10. 9 Versuchspläne mit 2 Messwiederholungen

Bei Versuchsplänen mit nur 2 Messwiederholungen kann die Differenz der beiden Messwiederholungsvariablen gebildet und damit die Varianzanalyse durchgeführt werden. Dadurch entfällt die Bedingung der Sphärizität bzw. gleicher Varianzen der Messwiederholungsvariablen und die Notwendigkeit deren Überprüfung. Unten wird gezeigt, wie die einzelnen Teile der Analyse durchzuführen sind, und dass diese zu demselben Ergebnis wie die Analyse des Split-Plot-Designs führen. D.h. in der Praxis: die Varianzanalyse mit einem Messwiederholungsfaktor kann "normal" durchgeführt werden, und der Test auf Sphärizität kann entfallen. Da im Gegensatz zu SPSS in R bei Versuchsplänen mit Messwiederholungen in der Regel eine Datenumstrukturierung erforderlich, kann der Weg über die Differenzvariable eine Erleichterung sein. Daher wird dieser unten nur für R gezeigt. Die einzelnen Schritte:

- Zunächst wird die Differenz `d` der beiden Messwiederholungsvariablen errechnet,
- eine 1-faktorielle Analyse von `d` mit dem Gruppierungsfaktor ergibt den Test der Interaktion,
- ein 1-Stichproben-t-Test von `d` prüft den Messwiederholungseffekt,
- es wird die Summe `s` der beiden Messwiederholungsvariablen errechnet und damit eine 1-faktorielle Analyse mit dem Gruppierungsfaktor durchgeführt, der den Effekt der Gruppierung prüft.

Als Beispiel wird der Datensatz `mydata14` verwendet, bei dem Scores von insgesamt 17 Personen in 3 Gruppen zu 3 Zeitpunkten vorliegen: einem Vortest sowie 2 Phasen des Haupttests. Diese beiden Phasen sollen verglichen werden. zunächst zum Vergleich über die "normale" 2-faktorielle Varianzanalyse, danach wie erläutert in 3 Einzelschritten.

mit R:

Die Transformation des Datensatzes `mydata14` in `mydata14t` wurde in Kapitel 5.1.2 gezeigt. Für die "normale" 2-faktorielle Varianzanalyse müssen die beiden Phasen zuvor ausgewählt werden:

```
subset(mydata14t,Phase=="2"|Phase=="3") -> mydata14.sub
summary(aov(score~Gruppe*Phase+Error(Vpn/Phase), mydata14.sub))
```

```
Error: Vpn
          Df Sum Sq Mean Sq F value Pr(>F)
gruppe     2   2.76   1.380   0.143  0.868
Residuals 14 135.36   9.668

Error: Vpn:Phase
             Df Sum Sq Mean Sq F value   Pr(>F)
Phase         1  459.6   459.6 928.593 3.36e-14 ***
gruppe:Phase  2    4.0     2.0   4.054   0.0408 *
Residuals    14    6.9     0.5
```

Nun die Analyse mittels der Differenzvariablen:

```
within(mydata14, d <- score2-score1 ) -> mydata14
summary(aov(d~Gruppe, mydata14))
t.test(mydata14$d)
within(mydata14, s <- score2+score1 ) -> mydata14
summary(aov(s~Gruppe, mydata14))
```

mit folgenden Ergebnissen, die mit denen der o.a. 2-faktoriellen Analyse übereinstimmen:

```
Response: d
          Df  Sum Sq Mean Sq F value  Pr(>F)
Gruppe     2  8.0252  4.0126   4.054 0.04084 *
Residuals 14 13.8571  0.9898
```

```
        One Sample t-test

data:  mydata14$d
t = -25.924, df = 16, p-value = 1.697e-14
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7.954224 -6.751658
sample estimates:
mean of x
-7.352941
```

```
Response: s
          Df  Sum Sq Mean Sq F value Pr(>F)
Gruppe     2   5.521  2.7605  0.1428 0.8682
Residuals 14 270.714 19.3367
```

mit SPSS:

Zunächst die "normale" 2-faktorielle Varianzanalyse:

```
GLM score1 score2 BY Gruppe
  /WSFACTOR=Phase 2 Polynomial
  /WSDESIGN=Phase
  /DESIGN=Gruppe.
```

Tests der Zwischensubjekteffekte

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Konstanter Term	17497,026	1	17497,026	1809,719	,000
Gruppe	2,761	2	1,380	,143	,868
Fehler	135,357	14	9,668		

Tests der Innersubjekteffekte

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Phase	418,582	1	418,582	845,795	,000
Phase * Gruppe	4,013	2	2,006	4,054	,041
Fehler(Phase)	6,929	14	,495		

Nun die Analyse mittels der Differenzvariablen:

```
compute  d=score2-score1.
compute  s=score2+score1.
execute.

UNIANOVA d BY Gruppe.
T-TEST   /VARIABLES=d /TESTVAL=0.
UNIANOVA s BY Gruppe.
```

Tests der Zwischensubjekteffekte

Abhängige Variable: d

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Gruppe	8,025	2	4,013	4,054	,041
Fehler	13,857	14	,990		
Gesamt	941,000	17			

Test bei einer Sichprobe

	Testwert = 0					
	T	df	Sig. (2-seitig)	Mittlere Differenz	95% Konfidenzintervall der Differenz	
					Untere	Obere
d	-25,924	16	,000	-7,35294	-7,9542	-6,7517

Tests der Zwischensubjekteffekte

Abhängige Variable: s

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Gruppe	5,521	2	2,761	,143	,868
Fehler	270,714	14	19,337		
Gesamt	37081,000	17			

6. 11 Alternative parametrische Modelle

6. 11. 1 Andere Residuen-Modelle

In Kapitel 4.3.10 war kurz erläutert und ein Beispiel gezeigt worden, wie mittels allgemeinen linearen Modellen (GLM, *general linear models*) parametrische Analysen durchgeführt werden können, denen auch andere Verteilungsmodelle für die Residuen als die Normalverteilung zugrunde liegen. Z.B. die Poisson-Verteilung, typischerweise für Variablen, die auf Häufigkeiten basieren, ordinal oder rechtsschief verteilt sind.

mit R:

In R gibt es für die Analyse von allgemeinen linearen Modellen (GLM) u.a. die schon in Abschnitt 6.10.6 vorgestellten Funktionen `lmer` für `family=gaussian` und `glmer` für `family=poisson, binary` (Paket `lme4`) dazu `Anova` (Paket `car`) für die Anova-Tests, alternativ die Funktion `mixed` (Paket `afex`), die allerdings lediglich ein Skript für die manchmal einfachere Benutzung von `lmer` und `glmer` ist. Dort war auch ein Beispiel für ein Modell Poisson-verteilter Residuen mit dem Datensatz `winer518` vorgestellt wordern.

mit SPSS:

In SPSS gibt es für die Analyse von allgemeinen linearen Modellen (GLM) u.a. die schon in Abschnitt 6.10.5 vorgestellte Prozedur `GENLIN`, die auch andere Verteilungsmodelle für die Residuen zulässt u.a. die Poisson-Verteilung. In SPSS muss neben der Verteilung auch die Link-Funktion angegeben werden, im Fall der Poisson-Verteilung ist dies `Link=Log`. Ein Beispiel dazu mit dem Datensatz `winer518` war dort vorgestellt wordern.

6. 11. 2 Andere Kovarianz-Modelle

In Kapitel 5.2 war erläutert worden, dass die parametrische Varianzanalyse die Bedingung der *Spherizität* oder der *Compound Symmetry* für die Kovarianzmatrix der Messwiederholungsvariablen verlangt. Es gibt aber auch parametrische Lösungen für alternative Formen der Kovarianzmatrix (vgl. Kapitel 2.15.5), die nicht die Bedingung der Spherizität erfüllen. Dabei werden die verschiedenen Formen hinsichtlich der Korrelationen unterschieden. In der Praxis spielen die folgenden eine Rolle (in Klammern die Bezeichnungen in R):

- *exchangeable* (*Compound Symmetry*): alle r_{ij} $(i \neq j)$ sind gleich (`corCompSymm`), entspricht der Voraussetzung für die parametrische Varianzanalyse,
- *unstructured*: alle r_{ij} $(i \neq j)$ sind beliebig (`corSymm`), entspricht einer Korrelationsmatrix ohne besondere Strukturen,
- *autogressive*: die r_{ij} $(i \neq j)$ errechnen sich als r^{i-j} $(i > j)$ (`corAR1`), sinnvoll für zeitabhängige Messwiederholungsfaktoren.

Allerdings dürfte es in der Praxis schwierig sein, aus einer vorhandenen Kovarianzmatrix eine der o.a. Formen zu identifizieren. Es gibt jedoch Techniken basierend auf dem Log Likehood-Test, verschiedene Lösungen miteinander zu vergleichen und somit die passendste auszuwählen.

mit R:

R bietet für Varianzanalysen auf Basis diverser Kovarianzmatrix-Formen die Funktion `gls` im Paket `nlme`, das die Analyse über 8 verschiedene Formen gestattet. Die Anova-Tabellen erhält man über `anova`. Der zu analysierende Datensatz muss - wie fast immer - im langen Format vorliegen. Im Fall des „klassischen“ Compound Symmetry-Modells liefert `gls` dieselben Ergebnisse wie oben die Funktion `lmer` für Analysen des GLMM-Modells. Zunächst einmal zum Vergleich das Compound Symmetry-Modell (`corCompSymm`) mit dem Datensatz `winer518`, das (wie zu erwarten) mit dem früher errechneten (vgl. Tabelle 6-1) identisch ist:

```
library(nlme)
gls1 <- gls(score~Geschlecht*Zeit, data=winer518t,
            corr = corCompSymm(, form = ~ 1 | Vpn))
anova(gls1)
```

```
                numDF  F-value p-value
(Intercept)         1 77.28302  <.0001
Geschlecht          1  0.47170  0.4988
Zeit                2 22.05063  <.0001
Geschlecht:Zeit     2 17.03797  <.0001
```

Nun ein anderer Datensatz (`mydata13`, vgl. Kapitel 11.5) mit zwei Messwiederholungsfaktoren (Phase und Wdh), der, wie dort gezeigt, die Spherizitätsvoraussetzung verletzt. Für diesen werden nun Modelle mit den 3 o.a. Typen von Korrelationsmatrizen durchgerechnet:

```
glcompS <- gls(score~Gruppe*Phase*Wdh, data=mydata13,
               corr = corCompSymm(, form = ~ 1 | Id))

glSymm  <- gls(score~Gruppe*Phase*Wdh, data=mydata13,
               corr = corSymm(, form = ~ 1 | Id))

glAR    <- gls(score~Gruppe*Phase*Wdh, data=mydata13,
               corr = corAR1(, form = ~ 1 | Id))
```

Die 3 Lösungen werden nun paarweise über `anova` verglichen:

```
anova(glcompS,glSymm)
```

```
    Model     df      AIC      BIC    logLik   Test L.Ratio p-value
glcompS     1 26 2271.709 2378.894 -1109.855
glSymm      2 40 2271.754 2436.653 -1095.877 1 vs 2 27.9555  0.0144
```

```
anova(glcompS,glAR)
```

```
    Model     df      AIC      BIC    logLik
glcompS     1 26 2271.709 2378.894 -1109.855
glAR        2 26 2314.898 2422.083 -1131.449
```

```
anova(glSymm,glAR)
```

```
    Model    df      AIC      BIC    logLik   Test  L.Ratio p-value
glSymm     1 40 2271.754 2436.653 -1095.877
glAR       2 26 2314.898 2422.083 -1131.449 1 vs 2 71.14427  <.0001
```

Die Maßzahlen `AIC` und `logLik` messen die Güte der Modellanpassung: je kleiner deren Werte, desto besser die Anpassung. Der Likelihood Ratio vergleicht jeweils die beiden `logLik` -Werte. Somit zeigen die Vergleiche nun, dass

- `glSymm` (wegen des kleineren `logLik`-Wertes) deutlich besser abschneidet als `glcompS`,
- `glcompS` und `glAR` fast gleich abschneiden,
- `glSymm` (wegen des kleineren `logLik`-Wertes) deutlich besser abschneidet als `glAR`,

d.h. das Standardmodell Compound Symmetry ist für die Varianzanalyse nicht geeignet, was wegen der fehlenden Sphärizität zu erwarten war.

Ein interessanter und hilfreicher Zusatzparameter ist

`weight = varIdent(form = ~ 1 | Faktorname)`

der ungleiche Varianzen der Messwiederholungsvariablen kompensieren kann, die meistens die Ursache für die Verletzung der Sphärizität sind. Zur Demonstration wird der Datensatz `mydata14` verwendet, der stark heterogene Streuungen auf den 3 Messwiederholungsvariablen zeigt (Standardabweichungen: 13.9 , 2.4 , 1.9). Eine Rechnung ohne den Zusatzparameter sowie eine mit:

```
gls1 <- gls(score~Gruppe*Phase, data=mydata14t,
            corr = corCompSymm(, form = ~ 1 | Vpn))
anova(gls1)

gls2 <- gls(score~Gruppe*Phase, data=mydata14t,
            corr = corCompSymm(, form = ~ 1 | Vpn))
            weight = varIdent(form = ~ 1 | Phase)
anova(gls2)
```

```
             numDF   F-value p-value
(Intercept)      1 268.66605  <.0001
Gruppe           2   0.86786  0.4272
Phase            2   5.16302  0.0099
Gruppe:Phase     4   1.55766  0.2033
```

```
             numDF  F-value p-value
(Intercept)      1 9765.642  <.0001
Gruppe           2    5.278  0.0090
Phase            2  496.184  <.0001
Gruppe:Phase     4    2.874  0.0343
```

Der Unterschied könnte kaum krasser sein. Einen statistischen Vergleich bietet hier ebenfalls die Funktion `anova`:

```
anova(gls1,gls2)
```

```
    Model df      AIC      BIC     logLik   Test L.Ratio p-value
gls1     1 11 329.0413 348.1557 -153.52065
gls2     2 13 224.7333 247.3230  -99.36663 1 vs 2 108.308  <.0001
```

Der Likelihood Ratio vergleicht die beiden `logLik` -Werte und zeigt einen stark signifikanten Unterschied an.

mit SPSS:

SPSS bietet dafür die bereits in Abschnitt 6.10.5 vorgestellte Prozedur `GENLIN` an. Doch leider gibt SPSS weder Maßzahlen wie z.B. `AIC` und `logLikelihood` aus, noch Tests, die mehrere Ergebnisse vergleichen, lediglich einen wenig zuverlässigen Quasi-likelihood (der bei allen Modellen immer denselben Wert hat.) Dennoch soll hier kurz am Datensatz `mydata13` (vgl. Kapitel 11.5) gezeigt werden, wie eine Korrelations-Modell vorgegeben werden kann:

```
GENLIN score BY Gruppe Phase
 /MODEL Gruppe Phase Gruppe*Phase
        DISTRIBUTION=normal LINK=IDENTITY
 /REPEATED SUBJECT=Vpn WITHINSUBJECT=Phase CORRTYPE = EXCHANGEABLE
          ADJUSTCORR=YES COVB=MODEL .
```

mit folgender Anova-Tabelle:

Tests der Modelleffekte

	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	258.400	1	.000
Gruppe	1.736	2	.420
Phase	9.677	2	.008
Gruppe * Phase	6.231	4	.183

und bei Wahl von `CORRTYPE = UNSTRUCTURED`

Tests der Modelleffekte

	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	202.469	1	.000
Gruppe	1.360	2	.507
Phase	8.376	2	.015
Gruppe * Phase	91.511	4	.000

Andere sinnvolle Optionen sind: `INDEPENDENT` und `AR(1)`.

7. Analysen für dichotome Merkmale

Für dichotome abhängige Variablen gibt es grundsätzlich zwei Möglichkeiten zur Durchführung einer Varianzanalyse: die bisher 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 eigentlich 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 α-Level wie auch β 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 *p*, 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, 2023a) 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~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 fehlender Sphärizität, also der Varianzinhomogenität, 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 Varianzinhomogenitäten 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 Wilks oder 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 bzgl. 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    0.0312 262.61 -1582.7  0.1309    0.7176
income      2    7.4363 270.02 -1553.9 15.5901 2.106e-07 ***
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 zwei 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 Varaible (`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 Spherizität) 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 Erläuterungen zum o.a. Output siehe Kapitel 5.3.1 und 5.4. 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 Messiwederholungsfaktors 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[b]	2,000	335,000	,000
Zeit * severity	,016	2,660[b]	2,000	335,000	,071
Zeit * treat	,073	13,184[b]	2,000	335,000	,000
Zeit * severity * treat	,003	,429[b]	2,000	335,000	,652

7. 2 Anwendung der Verfahren für ordinale Merkmale

Zur 1-faktoriellen Varianzanalyse eines dichotomen Merkamls verwendet man üblicherweise den χ^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) ausdrücklich darauf hin, dass die 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 $x_1,x_2,..,x_v$ 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]:

$$P(y = 1) = \frac{e^{b_0 + b_1x_1 + \ldots + b_vx_v}}{1 + e^{b_0 + b_1x_1 + \ldots + b_vx_v}}$$

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 Effekttest 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*(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 χ^2-Test des log likehood-Wertes (LR-Test). Sind Effekte der Faktoren vorhanden, so sollte dieser Test signifikant sein.
- Ein „klassischer“ χ^2-Anpassungstest des Modells, der also prüft, in wieweit die Daten mit dem Modell vereinbar sind. Dieser sollte nicht signifkant sein.
- Die Signifikanzüberprüfung eines Regressionskoeffizienten (auf Verschiedenheit von 0) oder eines Effekts über die Wald-Statistik mittels des χ^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 χ^2-Werte beider Tests gemittelt werden und dann dieser Mittelwert, der bei 1 FG χ^2-verteilt ist, per Hand auf Signifikanz überprüft wird. Darüberhinaus verletzen beide Tests das α-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, 2023a).

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, da 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(dvocabul~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(irish.glm)` 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
```

Nachfolgend 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.

	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

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 $P(y \le j)$ als die abhängige Variable verwendet wird, mit $j=1,...,m$, wenn *m* die Anzahl der Merkmalsausprägungen von *y* ist:

$$P(y \le j) = \frac{e^{b_{0j} + b_{1j}x_1 + \ldots + b_{vj}x_v}}{1 + e^{b_{0j} + b_{1j}x_1 + \ldots + b_{vj}x_v}}$$

(*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* ($i=1,...,v$) die Koeffizienten der jeweiligen Merkmalsausprägungen als gleich angenommen werden: $b_{i1}=b_{i2}=...=b_{i(m-1)}$. 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*)**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 ein 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 Regregression 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`
  Diese Funktion führt allerdings in der derzeitigen Version 1.0-4 zu Programmabbrüchen.

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 ein Gesamttest 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, {vocabula<-ordered(vocabula);
                        sex<-factor(sex); type<-factor(type)} )
options(contrasts=c("contr.sum","contr.poly"))
lr.clm <- clm(vocabula~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       -0.10271    0.07199  -1.427  0.15368
type1       0.28935    0.08189   3.534  0.00041 ***
type2       0.69361    0.10133   6.845 7.64e-12 ***
sex1:type1  0.36922    0.08198   4.504 6.67e-06 ***
sex1:type2 -0.24184    0.10019  -2.414  0.01579 *
```

```
         Df   Chisq Pr(>Chisq)
sex       1 257.077  < 2.2e-16 ***
type      2 341.272  < 2.2e-16 ***
sex:type  2  17.192  0.0001848 ***
```

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 $p_{sex} = 0.1537$ ablesbar, da der Faktor nur zwei Stufen hat.
- Der Effekt von `type` wird aus den z-Werten der Kontraste `type1` und `type2` ermittelt:
  $\chi^2_{type} = 3.534^2 + 6.845^2 = 59.3$
  und die Signifikanzprüfung ergibt $p_{type} < 0.001$.
- Der Effekt von `sex*type` wird aus den z-Werten der entsprechenden Kontraste ermittelt:
  $\chi^2_{sex*type} = 4.504^2 + 2.414^2 = 26.1$
  und die Signifikanzprüfung ergibt $p_{sex*type} < 0.001$.

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 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 `clm` (oder einer der anderen Funktionen) vornehmen, so erfordert das ein wenig Aufwand, da die Ergebnisobjekte selbst sind äußerst komplex und nicht einheitlich aufgebaut sind. Im Fall von `clm` erhält man (wie oben zu sehen) die z-Werte als 3. Spalte der Koeffizienten-Matrix und damit deren Quadrate:

```
c <- (summary(lr.clm))$coefficients
z2.values <- c[,3]^2
z2.values
```

```
...      sex1        type1         type2     sex1:type1     sex1:type2
...  2.035358    12.486226     46.856660     20.284316       5.826058
```

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. Für `sex` braucht kein Test angefordert werden, da er nur 2 Stufen hat und das Ergebnis aus der Tabelle der Parameterschätzer direkt ablesbar ist.

Das wesentliche Ergebnis steckt in der Tabelle der Regressionsparameter, oben unter „Schwelle“ die Parameter b_{oj} 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:

Parameterschätzer								
		Schätzer	Standard fehler	Wald	Fg	Sig.	Konfidenz intervall 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[a]	.	.	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[a]	.	.	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[a]	.	.	0	.	.	.
	[sex=2] * [type=1]	0[a]	.	.	0	.	.	.
	[sex=2] * [type=2]	0[a]	.	.	0	.	.	.
	[sex=2] * [type=3]	0[a]	.	.	0	.	.	.

Anschließend folgen die gewünschten Gesamttests, und zwar in der Reihenfolge, wie diese spezifiert worden waren, auch ersichtlich aus der jeweils davor angezeigten Tabelle der Kontrastkoefiizienten. Also unten zunächst der Test für `type`, danach für `sex*type`

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[a]				
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 (Steigungkoeffizienten) 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 Kriterumsvariable 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 diesem 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 cri-
terion 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 Interaktionen 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 erhoben worden sind, z.B. unter verschiedenen Versuchsbedingungen, 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.

Eine andere Möglichkeit besteht im Fall von nur zwei Messwiederholungen (J=2), die in Kapitel 6.10.9 vorgestellt worden war: die Bildung der Differenz beider Variablen. Diese hat dann ordinales Skalenniveau: bei der üblichen Skalierung 0/1 also die Werte -1/0/+1. Formal ist dies praktisch generell möglich, allerdings inhaltlich nicht immer. Z.B. im Fall von 0~schlecht und 1~gut könnte man von einer Verschlechterung, ohne Veränderung bzw. einer Verbesserung sprechen. Diese Differenz kann dann über eine geeignete Varianzanalyse ohne Messwiederholung analysiert werden.

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, 2023a) 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
winer518t     <- 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)
     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 Männer 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 6.10.5 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 Gruppierunsgfaktor 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.

Parameterschätzer

Parameter	Regressionskoeffizient B	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[a]	.	.	.	.	.	.
[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[a]	.	.	.	.	.	.
(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							
		Mittlere Differenz (I-J)	Standard Fehler	df	Sig.	95% Wald-Konfidenzintervall für die Differenz	
(I)	(J)					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, „logistischen Regression“ derzeit allerdings nur in R. Normalerweise wird das in Kapitel 8.2 kurz beschriebene *proportion odds model* angewandt. Es gibt aber auch andere Lösungen für den Fall ordinaler abhängige Variablen: zum einen mittels der in Kapitel 6.10.4 vorgestellten GEE-Modelle, zum anderen im Fall von nur zwei Messwiederholungen die Möglichkeit, die bereits im vorigen Abschnitt erläutert und in Kapitel 6.10.9 vorgestellt worden war: die Bildung der Differenz beider Variablen. Diese hat dann wieder ordinales Skalenniveau und kann mit einem geeigneten Verfahren ohne Messwiederholungen analysiert werden.

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` unterstützt auch eine Anova-Tabelle 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            -2.3427   0.4801    -4.8796    0.0000
cuts2|3            -0.8230   0.4036    -2.0391    0.0414
cuts3|4             1.2852   0.3000     4.2840    0.0000
cuts4|5             3.4874   0.6743     5.1719    0.0000
Geschlecht1        -0.2482   0.2008    -1.2361    0.2164
Zeit.L              2.5607   0.5196     4.9282    0.0000
Zeit.Q             -0.0863   0.2696    -0.3201    0.7489
Geschlecht1:Zeit.L -0.7117   0.2978    -2.3899    0.0169
Geschlecht1:Zeit.Q  2.1244   0.4336     4.8994    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 χ^2-Test zu ermitteln. Für den Faktor Zeit (Zeilen `Zeit.L` und `Zeit.Q`) wäre das z.B.:

$\chi^2 = 4.9282^2 + 0.3201^2 = 24.39$

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.

```
PO Score Test: 8.121 (d.f. = 15 and p.value = 0.9188)
```

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 α-Adjustierungen*.

Alpha-Adjustierungen und multiplen Vergleichen ist ein separates Skript gewidmet (vgl. Lüpsen, 2020b). 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 $1*\mu_K - 0.5*(\mu_A + \mu_B)$. Theoretisch können sogar bei der Zusammenfassung von Gruppen gewichtete Mittel gebildet werden, etwa $(0.333*\mu_A + 0.667*\mu_B)$, wenn etwa die B-Gruppe doppelt so stark berücksichtigt werden soll wie die A-Gruppe.

Hat ein Faktor *I* Gruppen (Schichten), so ist ein Kontrast C über *I* Koeffizienten c_i definiert:

$$C = c_1\mu_1 + c_2\mu_2 + \ldots + c_I\mu_I$$

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

$$SS_C = \frac{(c_1\bar{x}_1 + c_2\bar{x}_2 + \ldots + c_I\bar{x}_I)^2}{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \ldots + \frac{c_I^2}{n_I}}$$

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

$$F = \frac{SS_C}{MS_{Fehler}}$$

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

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

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

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

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

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

- Eine Menge von Kontrasten heißt *orthogonal*, wenn alle Paare orthogonal sind.
- Werden (I-1) orthogonale Kontraste $C_1, C_2,..,C_{I-1}$ mit Streuungen SS_{C1}, SS_{C2},.., $SS_{C(I-1)}$ getestet, dann gilt $SS_{C1} + SS_{C2} +..+ SS_{C(I-1)} = SS_{Effekt}$, 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 α-Korrektur (siehe dazu Lüpsen, 2020b) 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:

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

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

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

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

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

Die Kontraste oder Kodierungen haben auch eine andere Funktion: Bei der Regression müssen Prädiktoren mit nominalem Skalenniveau dichotomisiert werden. Die „naive" Art, ein nominales Merkmal f mit m Ausprägungen in mehrere dichotome $d_1,.., d_m$ 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. $d_m = 1\text{-}d_1\text{-}d_2\text{-}...\text{-}d_{m-1}$, 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 $c_1+c_2+..+c_I = 0$ 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:

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

	Kontraste R				Kontraste SPSS			
Gruppe	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 vergleichen: 1-2, (1,2)-3, (1,2,3)-4 usw. wobei mit (.,.) der Mittelwert der entsprechenden Gruppen bezeichnet wird.

	Kontraste R und SPSS			
Gruppe	1	2	...	(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 vergleichen, die zweite gegen alle folgenden usw. (Diese Kontraste sind in R nicht verfügbar.)

	Kontraste SPSS			
Gruppe	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 vergleichen: 1-2, 2-3, 3-4 usw. Diese werden sinnvollerweise bei Messwiederholungsfaktoren eingesetzt. (Diese Kontraste sind in R nicht verfügbar.)

	Kontraste SPSS			
Gruppe	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

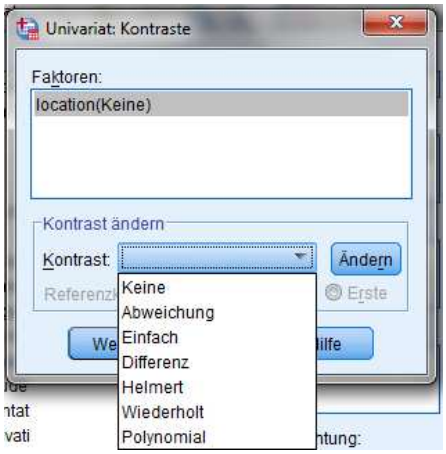
`options(contrasts=c("contr.`*name1*`","contr.`*name2*`"))`

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:

```
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)
```

einzufügen, wobei das „(1)“ die Nummer der Vergleichsgruppe angibt, also hier die erste:

```
Unianova  x by patients drugs
  /Contrast(drugs)=Simple(1)
  /save = zresid
  /print = homogenity
  /design = patients drugs patients*drugs.
```

Die Ausgabe dazu sollte selbsterklärend sein:

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

Niveau 4 vs. Niveau 1	Kontrastschätzer		16,713
	Hypothesenwert		0
	Differenz (Schätzung - Hypothesen)		16,713
	Standardfehler		3,347
	Sig.		,000
	95% Konfidenzintervall für die Differenz	Untergrenze	9,820
		Obergrenze	23,605
a. Referenzkategorie = 1			

Tabelle 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 χ^2-Werte für jeden Vergleich „mit der Hand“ ausgerechnet werden, was ein wenig mühselig ist, zumal SPSS nicht die Testgröße ausgibt:

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

wobei

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

Die χ^2-Werte haben jeweils 1 Fg und müssen anhand der Tabellen der χ^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, ohne den Aufruf von `Anova` oder `drop1`, errechnet werden und ergibt `msfehler` mit dem Wert 43,35. Zur Berechnung der χ^2-Werte müssen die t-Werte aus der Tabelle 9-1 quadriert, mit MS_{Fehler} multipliziert sowie durch MS_{total} dividiert werden. Das kann in R programmiert werden. (Die Berechnung „per Hand“ kann dem Abschnitt „SPSS“ entnommen werden.) Wenn `aovc` das oben ermittelte Ergebnisobjekt von `gls` ist, dann lässt sich mit folgenden Anweisungen daraus zunächst die Kontrasttabelle `ctabelle`, die t-Werte `twerte` und schließlich die χ^2-Werte `chisq`:

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

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

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

Tabelle 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 χ^2-Wert bei 1 Fg beträgt 3,84, so dass kein Unterschied zwischen drug1 und drug2 nachgewiesen werden kann.

Das vorige Beispiel wird dahingehend modifiziert, dass drugs1 und drugs2 als etablierte Präparate angenommen werden, während drugs3 und drugs4 als neu angesehen werden. Daher sollen zum einen die beiden alten Präparate (1-2) sowie die beiden neuen Prärparate (3-4) verglichen werden, zum anderen die alten zusammen gegen die neuen zusammen ((1,2)-(3,4)). Daraus resultiert folgende Kontrastmatrix:

	Kontraste		
Gruppe	1	2	3
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*,...,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: $C*(C`*C)^{-1}$ . 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)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           16.266      1.159  14.033 2.34e-13 ***
group1                 3.359      1.159   2.898 0.007698 **
drugsA1-A2            -5.975      3.347  -1.785 0.086328 .
drugsA3-A4            -7.337      3.209  -2.287 0.030950 *
drugsA12-A34         -20.112      4.636  -4.338 0.000207 ***
group1:drugsA1-A2     -1.725      3.347  -0.515 0.610759
group1:drugsA3-A4      6.537      3.209   2.037 0.052326 .
group1:drugsA12-A34   11.013      4.636   2.375 0.025518 *
```

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.

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

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

d <- mydata2$drugs                  # Vergleich 3-4
```

```
d[d==1|d==2] <- NA
d<-factor(d,levels=c(3,4))
with(mydata2,kruskal.test(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))
with(mydata2,kruskal.test(x,d))
```

Der globale χ^2-Wert beträgt 11,2 . Die χ^2-Werte der drei Kontraste: 1,97 (1-2), 2,61 (3-4) und 7,32 ((1,2)-(3,4)) mit der Summe von 11,9, die ungefähr dem globalen Wert entspricht, da die Kontraste orthogonal sind. Fazit: alle 3 Tests haben jeweils 1 Fg (kritischer Wert c=3,84) , so dass lediglich der Vergleich (1,2)-(3,4) signifikant ist.

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 χ^2-Wert beträgt 11,2 . Die χ^2-Werte der drei Kontraste: 1,97 (1-2), 2,61 (3-4) und 7,32 ((1,2)-(3,4)) mit der Summe von 11,9, die ungefähr dem globalen Wert entspricht, da die Kontraste orthogonal sind. Fazit: alle 3 Tests haben jeweils 1 Fg (kritischer Wert c=3,84) , so dass lediglich der Vergleich (1,2)-(3,4) signifikant ist.

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.

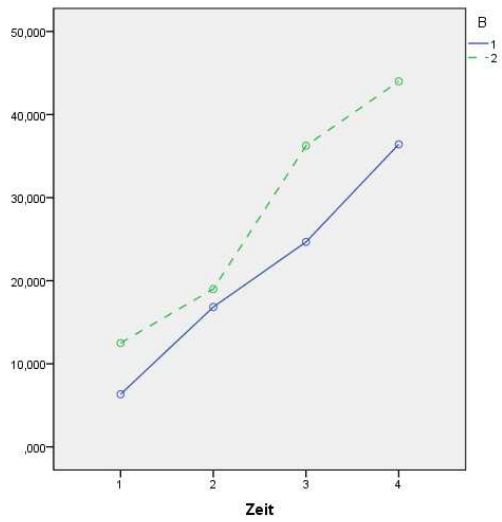
9. 7 Kontraste für Messwiederholungen und Interaktionen

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

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

```
/wsdesign=Zeit
/design=A B A*B.
```

Die Ergebnisse der Varianzanalyse sind in Tabelle 6-7 (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 die Kontraste für die Faktoren 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 χ^2-Wert, was der Prüfstatistik des *Wald-Tests* entspricht - verschiedentlich wird auch direkt dieser Test ausgegeben.
- Aufsummieren der χ^2-Werte aller Kontraste, die zum selben Effekt gehören, was wiederum einen χ^2-Wert ergibt.

- Testen dieser Summe auf Signifkanz anhand der χ^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_β die dazugehörige Kovarianzmatrix sind. Diese Statistik ist χ^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 $\hat{\beta}$ unabhängig sind, also V_β 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 Ergenbnis-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) vom Autor angeboten (vgl. Anhang 3), insbesondere für die Anwendung auf GEE- und GLMM-Ergebnisse. Beide Funktionen erwarten als Argumente die Koeffizienten, die Kovarianzmatraix, 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 unabhängig voneinander 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 Interprätationsmöglichkeiten 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 `drugs` (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 zwei 1-faktorielle Analysen von `drugs` 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 *
drugs          2     48   24.00   2.717 0.10634
patients:drugs 2    144   72.00   8.151 0.00581 **
Residuals     12    106    8.83
```

Die beiden 1-faktoriellen Analysen (mit Kommando):

```
> summary(aov(x~drugs,subset(mydata1,patients=="1")))
            Df Sum Sq Mean Sq F value Pr(>F)
drugs        2     24      12     1.5  0.296
Residuals    6     48       8

> summary(aov(x~drugs,subset(mydata1,patients=="2")))
            Df Sum Sq Mean Sq F value Pr(>F)
drugs        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 α-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 `drugs` analysiert, danach der Faktor `drugs` für die beiden Stufen von `patients`, wobei standardmäßig eine α-Adjustierung vorgenommen wird, und zwar die Methode von Holm. Soll dies vermieden werden, ist explitit "none" anzugeben:

```
library(phia)
ano1 <- aov(x~patients*drugs,mydata1)
testInteractions(ano1,fixed="drugs",across="patients",
                 adjustment="none")
testInteractions(ano1,fixed="patients",across="drugs",
                 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
         drugs1 drugs2 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 `drugs` 1 und 3 besteht und zum anderen der Faktor `drugs` 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 drugs
  /EMMEANS=TABLES(patients*drugs) COMPARE (patients) ADJ(LSD)
  /EMMEANS=TABLES(patients*drugs) COMPARE (drugs) ADJ(LSD)
  /DESIGN = patients drugs patients*drug.
```

Nach der 2-faktoriellen Varianzanalyse wird zunächst der Faktor `patients` für die 3 Stufen von Faktor `drugs` 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 `drugs` für die beiden Stufen von `patients`:

Tests auf Univariate

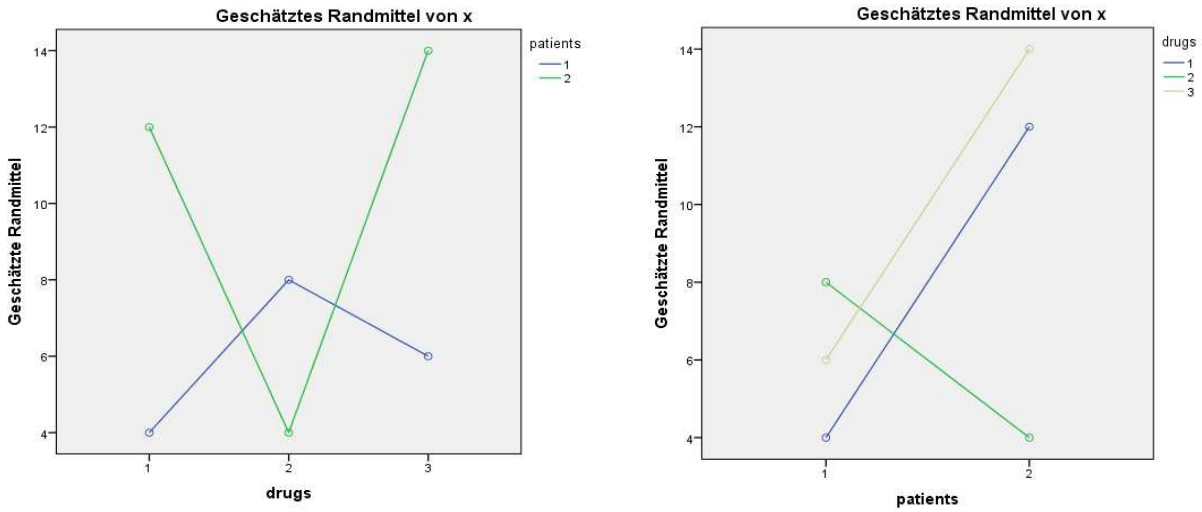
		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Kontrast	24,000	2	12,000	1,358	,294
	Fehler	106,000	12	8,833		
2	Kontrast	168,000	2	84,000	9,509	,003
	Fehler	106,000	12	8,833		

hieran schließen sich, ausgelöst durch den Befehl `POSTHOC=drugs(TUKEY)`, paarweise Mittelwertvergleiche nach dem Verfahren von Tukey für den Faktor `drugs` an, da dieser mehr als 2 Stufen hat:

Paarweise Vergleiche

	(I)	(J)	Mittlere Differenz (I-J)	Standardfehler	Sig.[b]	95% Konfidenzintervall für die Differenz	
						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, nicht auf der parametrischen Analyse basierenden Methoden wie Puri & Sen, KWF, 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), die sowohl bei Versuchsplänen mit mehreren Gruppierungsfaktoren alas auch mit maximal einem Messwiederholungsfaktor die Analyse von simple effects durchführt. Die erforderlichen Anweisungen (vgl. auch Tabelle 6-1):

```
attach("path/anova.lib")
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 α-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 Männer 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.	95% Konfidenzintervall für die Differenz	
						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:

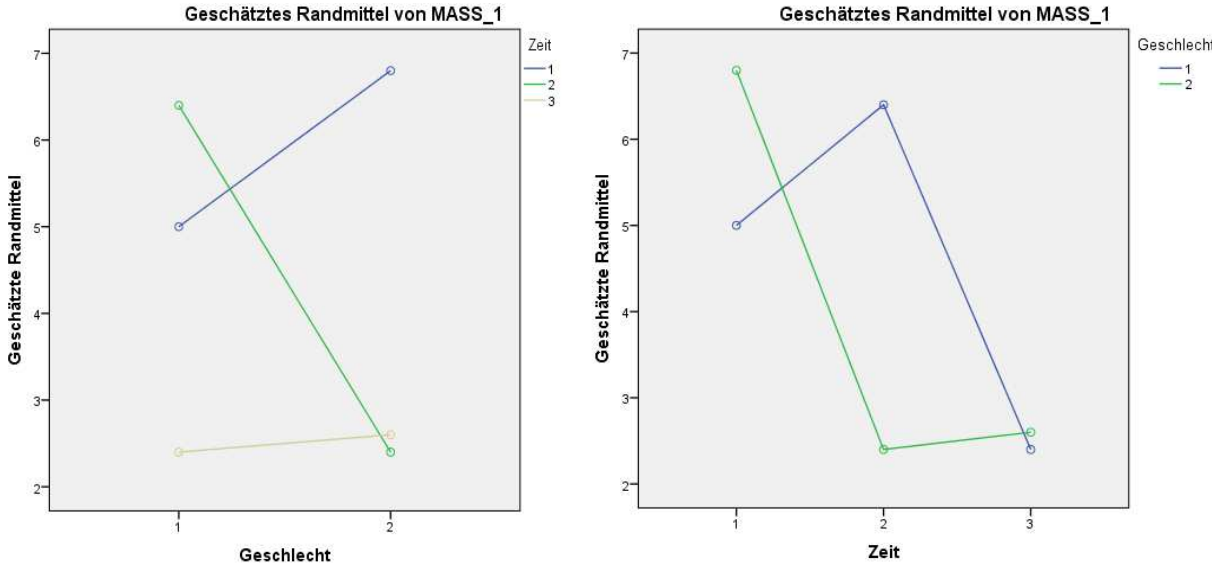
Multivariate Tests

		Wert	F	Hypothese df	Fehler df	Sig.
1	Pillai-Spur	,795	13,584[a]	2,000	7,000	,004
	Wilks-Lambda	,205	13,584[a]	2,000	7,000	,004
	Hotelling-Spur	3,881	13,584[a]	2,000	7,000	,004
	Wurzel nach Roy	3,881	13,584[a]	2,000	7,000	,004
2	Pillai-Spur	,864	22,230[a]	2,000	7,000	,001
	Wilks-Lambda	,136	22,230[a]	2,000	7,000	,001
	Hotelling-Spur	6,351	22,230[a]	2,000	7,000	,001
	Wurzel nach Roy	6,351	22,230[a]	2,000	7,000	,001

Paarweise Vergleiche

	(I)Zeit	(J)Zeit	Mittlere Differenz (I-J)	Standardfehler	Sig	95% Konfidenzintervall für die Differenz	
						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 Männer 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 1von den Zeitpunkten 2 und 3.



Interaktionsplots zur Visualisierung der simple effects in beiden Ansichten

Ein weiteres Beispiel für simple effects, im Anschluss eine Analyse mittel general linear models (GLM), war in Kapitel 6.10.5 gezeigt worden.

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. Beipiele 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
Standard-abweichungen	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 oder van der Waerden, die Varianzinhomogenität nicht hinreichend 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)
      Df    F value    Pr(>F)
group 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. 2.8 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 (`lognormal`), 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	0.94	0.95	1.04	1.18	1.12
	A 2	1.17	1.07	1.31	1.17	1.24
	A 3	1.05	0.97	1.02	1.21	1.24
	A 4	1.19	1.17	1.34	1.25	1.14
Standard-abweichungen	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

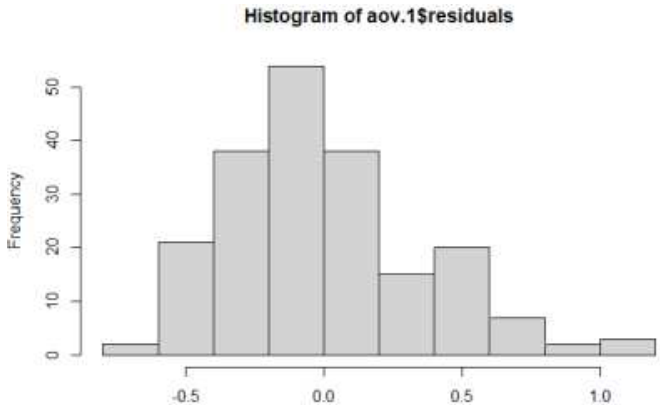
Die Tabelle lässt vermuten, dass die Standardabweichungen für Faktor B ungleich sind:

```
leveneTest(y~B,lognormal)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   4  5.5881 0.0002801 ***
      195
```

Als nächstes werden die Residuen untersucht:

```
aov.1 <- aov(y~A*B,lognormal)
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 0.726, der Test auf Schiefe>0 (S/(6/*n*)) ergibt:

0.726/(6/200) = 24.21

Dieser Wert ist χ^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 3) verfügbar:

```
attach("path/anova.lib")
wj.anova(lognormal,'y','A','B', Ftest=T)
```

```
              F   df1      df2  P(F>value)
A     2.2775442     3 72.19293  0.08682882
B     1.1882087     4 76.69976  0.32276977
A : B 0.6010449    12 66.84131  0.83371794
```

```
bf.f(y~A*B,lognormal)
```

```
                Df Df.err  Sum Sq Mean Sq F value  Pr(>F)
A           2.9838 194.42  0.9488 0.31625  2.3506 0.07402 .
B           3.3816 156.77  0.8109 0.20272  1.4913 0.21470
A:B        12.0000 128.19  0.7644 0.06370  0.4625 0.93304
Residuals 180.0000        24.7946 0.13775
```

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:

```
lognormal <- within(lognormal,ly<-log(y))
leveneTest(ly~A*B,lognormal)
aov.2 <- aov(ly~A*B,lognormal)
shapiro.test(aov.2$residuals)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group  19  2.2652 0.002932 **

Test Name:                       Shapiro-Wilk normality test
Test Statistic:                  W = 0.991115
P-value:                         0.2514
```

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(lognormal,'ly','A','B')
bf.f(ly~A*B,lognormal)
```

mit den folgenden Ergebnissen:

```
          Chi Sq   df  P(Chi>value)
A       8.302701    3    0.05150485
B      11.404768    4    0.03150685
A : B   9.282005   12    0.74850000

Response: ly
                 Df Df.err  Sum Sq Mean Sq F value  Pr(>F)
A           2.9095 187.26  0.9320 0.31068  2.6456 0.05221 .
B           3.0725 139.04  1.6294 0.40736  3.5590 0.01527 *
A:B        12.0000 114.43  0.6892 0.05744  0.4995 0.91128
Residuals 180.0000        20.6984 0.11499
```

Auch hier unterscheiden sich beide Ergebnisse nicht. Aber: diese Resultate widersprechen den zuerst gefundenen mit den nicht-transformierten y-Werten, insbesondere für Faktor B. 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 signfikant werden. Dieses wird durch die Berechnungen des lognormal-Datensatzes mittels `glm(y~A*B,family=poisson)` (vgl. Kapitel 4.3.10) bestätigt, wo das hier passendere Poisson-Verteilungsmodell gewählt worden war.

11. 3 Negatives 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

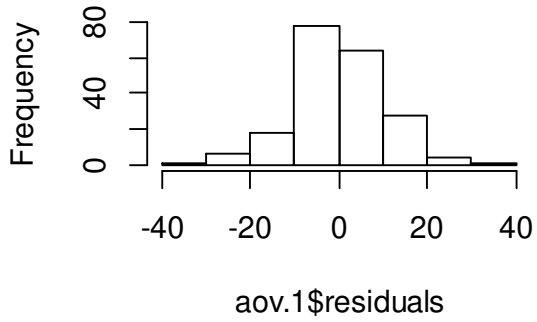
Zunächst einmal wird die parametrische Varianzanalyse durchgeführt, um auch die Residuen auf Normalverteilung überprüfen zu können. Hier muss `drop1` verwendet werden, da der Datensatz ungleiche Zellenbesetzungszahlen aufweist.

```
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 als 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

& Forsythe, 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, die alle in der eigenen Bibliothek (siehe Anhang 3) verfügbar sind:

```
attach("path/anova.lib")
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:

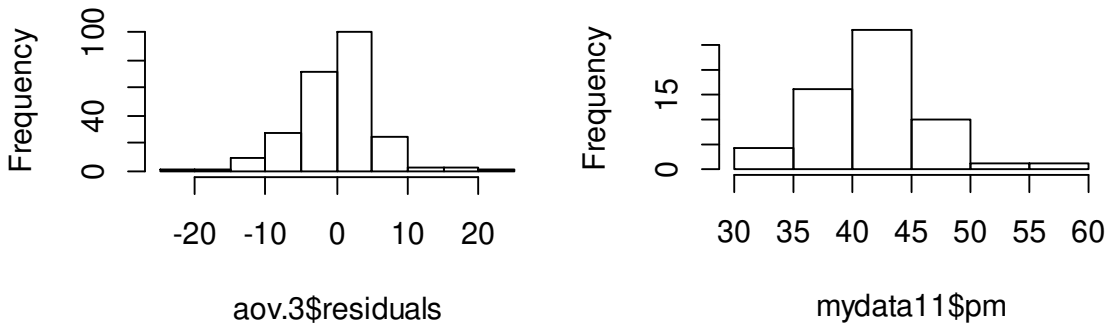
			Zeitpunkt			
		n	1	2	3	4
Mittelwerte	A 1	10	38.50	41.50	46.50	44.5
	A 2	10	36.50	46.00	43.00	36.5
	A 3	20	39.50	40.25	40.25	42.5
	A 4	20	40.25	39.00	41.00	43.0
Standard-abweichungen	A 1	10	4.74	14.62	9.56	14.54
	A 2	10	7.09	8.82	6.58	15.10
	A 3	20	4.84	7.48	5.45	6.05
	A 4	20	3.43	3.77	3.48	6.07

Zunächst einmal muss der Datensatz in das für Messwiederholungsanalysen erforderliche „long“-Format transformiert werden:

```
mydata11t <- reshape(mydata11,direction="long",timevar="Zeit",
             v.names="x",varying=1:4,idvar="Id")
mydata11t <- within(mydata11t,Zeit <- factor(Zeit))
```

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)
```




```
        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. Eine genauere Untersuchung über Schiefe und Ex-

zess, zu berechnen mittels `skewness` bzw. `kurtosis` aus dem Paket `EnvStats`, ergibt S=-.136 bzw. E=1.89 und deren Tests auf Abweichung von 0:

$S^2/(6/N)) = 0.0031/(6/240) = 0.74$ bzw.
$E^2/(24/N)) = 3.57/(24/240) = 35.72$

Beide Wert sind χ^2-verteilt mit 1 FG und deuten auf einen nennenswerten Exzess hin. Wegen des relativ großen *N* von 240 könnte allerdings die Abweichung von der Normalverteilung vernachlässigt werden.

Als nächstes müssen noch die Personeneffekte π_i auf Normalverteilung überprüft werden:

```
mydata11 <- within(mydata11,pm<-(V1+V2+V3+V4)/4)
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 Spherizität 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 π_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 Spherizität 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 Varianzinhomogenitäten ist. In R gibt es dazu die Verfahren von Huynh (IGA), von Brown & Forsythe (mBF) sowie von Welch & James. Wegen des starken negatives Pairings haben diese Methoden allerdings nur eine geringe Teststärke. Daher sollte das KWF-Verfahren bevorzugt werden, das in diesem Fall den anderen überlegen ist. Für alle hier genannten Methoden sind entsprechende Funktionen in der eigenen Bibliothek (siehe Anhang 3) verfügbar.

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. Das Ergebnis für die Interaktion:

```
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 siginifikant eingestuft werden konnte. Der Interaktionseffekt wird hier als nicht signifikant ausgewiesen, wobei an die Bedingung $min(n_i) \geq 3(J-1)$ erinnert sei, die hier mit 20>3*3 erfüllt war.

Für das KWF-Verfahren gibt es die Funktion `np.anova`. Während meistens für die nichtparametrischen Methoden die Gleichheit der Varianzen der Messwiederholungen überprüft werden sollte, kann dies bei der KWF-Methode entfallen (vgl. Kapitel 2.17.2). Aufruf und Ergebnis:

```
np.anova(x~A*Zeit+Error(Id/Zeit), mydata11t)
```

```
generalized Kruskal-Wallis/Friedman tests including Iman & Davenport
F-tests
                     Df  Sum Sq   Chisq Pr(>Chi) F value    Pr(>F)
A                     3    8908  0.4652  0.92646  0.1484 0.9303196
Residuals Btw.Subj   56 1120820
Zeit                  3      32 19.1400  0.00026  7.0201 0.0001781 *
A:Zeit                9      23 13.6200  0.13650  1.7267 0.0864036 .
Residuals Zeit      168     245
```

Damit weichen diese Ergebnisse doch erheblich von dem oben mit `ezANOVA` erstellten ab und ist auf die nicht erfüllten Voraussetzungen zurückzuführen. Fasst man die Ergebnisse der 4 robusteren Verfahren zusammen, so kommt man zu dem Schluss, dass ledigleich der `Zeit`-Effekt einen deutlichen Einfluss hat.

11. 5 Gemischter Versuchsplan: Prüfung der Voraussetzungen

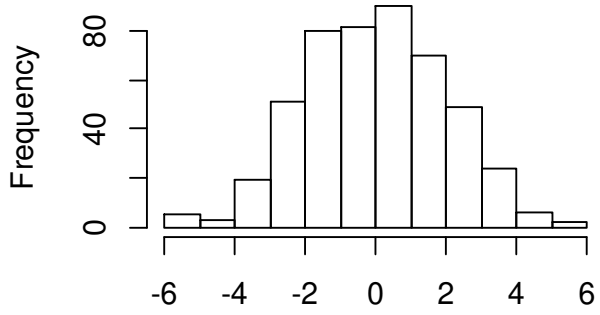
Der Datensatz `mydata13` besteht aus der abhängigen Variablen `score` zu 6 Zeitpunkten: jeweils 3 Werte in einer 1. Phase sowie in einer 2. Phase (Faktor `Wdh`). `score` hat einen ganzzahligen Wertebereich zwischen 0 und 12. Die Werte wurden in 4 unterschiedlichen Versuchsgruppen (Faktor `Gruppe`) erhoben mit einer Fallidentifikation `Id` und umfasst 80 Versuchspersonen. Im Gegensatz zum vorherigen Datensatz `mydata11` (siehe Abschnitt 11.4) liegen hier zwei Messwiederholungsfaktoren vor, was insbesondere die Prüfung der Voraussetzungen erschwert:

		Phase 1			Phase 2		
Gruppe	n	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 multvariate Normalverteilung überprüft werden, was sich insbesondere dann empfiehlt, wenn eine multvariate 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` zu einem mit den Stufen 1,..,6 (2*3) zusammengefasst (Variable `PW`), mit deren Hilfe die Transformation erfolgen kann (neuer Dataframe: `mydata13t`). Hierzu müssen die Faktoren für die arithmetische Berechnung in “integer“ gewandelt werden, wobei die “3“ die Anzahl der Stufen des Faktors “Wdh“ ist:

```
res       <- aov.1$residuals
mydata13  <- cbind(mydata13,res)
mydata13  <- within(mydata13,
             PW<-(as.integer(Phase)-1)*3 + as.integer(Wdh))
mydata13t <- reshape(mydata13, direction="wide", timevar="PW",
             v.names=c("score","res"),times=1:6,idvar="Id")
```

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 `mvnormalTest` :

```
msw(mydata13t[,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 sowie eine von Villasenor-Alva und Gonzalez-Estrada (`VA-GE`). Beide erkennen keinen Widerspruch zur multvariaten 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:

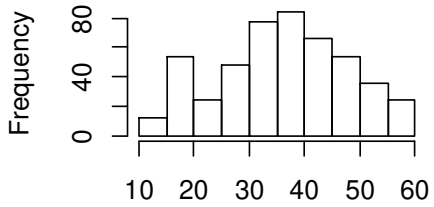
```
mardia(mydata13t[,c(6,8,10,12,14,16)])
```

```
$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 π_i auf Normalverteilung überprüft werden:

```
pm <- with(mydata13,ave(score,Id,FUN=sum))
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, weil diese letztlich die Basis für die Varianzanalyse sind. Mittels der Funktion `ave` können diese in einer solchen Struktur errechnet werden, dass sie als neue Variablen an `mydata13` angehängt werden können:

```
Wdh.s    <- with(mydata13,ave(score,list(Id,Wdh),FUN=mean))
Phase.s  <- with(mydata13,ave(score,list(Id,Phase),FUN=mean))
mydata13 <- cbind(mydata13,Wdh.s,Phase.s)
```

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(mydata13, 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 Sphärizität 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                19.3992 14  0.1502
John's V              202.6776 20  0.0000
Nagao                 202.6776 20  0.0000
LR                    139.2011 20  0.0000
Muirhead & Waternaud  221.4102 20  0.0000
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 Spherizität 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=mydata13))
```

```
        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 Spherizitäts-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 Namen `Wdh.s.1, Wdh.s.2, Wdh.s.3 bzw. die` Positionen 9 bis 11. Beide Varianten zur Spezifikation der Variablen angegeben werden:

```
myW <- reshape(mydata13, direction="wide", timevar="Wdh",
               v.names=c("Wdh.s"), times=1:3, idvar="Id")
check.sphere(myW[,7:9])                          # ueber Variablennummern
check.sphere(myW[,c("Wdh.s.1", "Wdh.s.2", "Wdh.s.3")]) # ueber VarNamen
```

```
$results
                     Chisquare df p value
Mauchly                 0.3866  2  0.8242
John's V               88.1595  5  0.0000
Nagao                  88.1595  5  0.0000
LR                     86.2647  5  0.0000
Muirhead & Waternaud  108.2067  5  0.0000
Compound Symmetry       9.9130  4  0.0419

$Box.epsilon
[1] 0.9950802
```

Für den Test auf Spherizität für den Phasen-Effekt (Variablen `Phase.s.1, Phase.s.2`) sind folgende Anweisungen erforderlich (hier ohne Ausgabe):

```
myP <- reshape(mydata13, direction="wide", timevar="Phase",
               v.names=c("Phase.s"), times=1:2, idvar="Id")
check.sphere(myP[,9:10])
```

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 Sphärizität, mit beiden Datenformaten möglich ist:

```
with(mydata13, check.covar(score,Gruppe,trial=PW,id=Id))
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
```

Die Ergebnisse der Tests von Larntz & Perlman sowie von Box zeigen 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 (oder Mittelwerte) über den nicht berücksichtigten Messwiederholungsfaktor gebildet werden.
- Es muss eine Teildatenmatrix für eine beliebige Ausprägung des nicht berücksichtigten Messwiederholungsfaktors gewählt werden.

Hier soll nur die Varianzanalysen mit `Gruppe` und `Phase` gezeigt werden. Zunächst die Mittelwertbildung und Teildatenmatrix, hier für `Wdh`=1:

```
mydata13.P  <- within(mydata13, Sy<-ave(score,Id,Phase,FUN=sum))
mydata13.P1 <- subset(mydata13.P,Wdh==1)
```

Die Varianzanalyse nach Brown-Forsythe nun:

```
with(mydata13.P1,mbf.f(Sy,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
```

Anzumerken ist, dass für `Wdh`=2 und `Wdh`=3 dieselben Ergebnisse herauskommen müssen. Der Vollständigkeit wegen soll wenigstens noch das Ergebnis der Analyse für `Gruppe` und `Wdh` angeführt werden:

```
modified Brown-Forsythe method for mixed repeated measures designs
                   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, d.h. nur der Faktor `Phase` hat einen Einfluss.

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 `c1,..,c4` errechnet, und daraus die durchsschnittliche Korrelation `sc1,..,sc4` 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 `sc1,..,sc4` werden die

Summen der transformierten Korrelationskoeffizienten über die untere Dreiecksmatrix berechnet und anschließend durch v*(v-1)/2 dividiert, wobei hier die Variablenzahl v=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(c2[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.

Eine ganz andere Möglichkeit zur Analyse bieten die GLM-Methoden (generalized linear models) mit der Option, andere Kovarianz-/Korrelationsmatrizen als Modellgrundlage zu nehmen als die Compound Symmetry. Diese war in Kapitel 6.11.2 an diesem Datensatz `mydata13` vorgestellt worden. Und es zeigte sich, dass eine Kovarianzmatrix ohne spezielle Eigenschaften am besten zu den Daten passt:

```
library(nlme)
glSymm  <- gls(score~Gruppe*Phase*Wdh, data=mydata13,
           corr = corSymm(, form = ~ 1 | Id))
anova(glSymm)
```

Die Anova-Tabelle deckt sich mit den o.a. Ergebnissen, die mittels `np.anova` auf Basis des KWF-Modelles errechnet wurden:

```
                numDF  F-value p-value
(Intercept)         1 807.4647  <.0001
Gruppe              3   0.5240  0.6660
Phase               1  24.9179  <.0001
Wdh                 2   0.3055  0.7369
Gruppe:Phase        3   0.5813  0.6275
Gruppe:Wdh          6   0.7082  0.6431
Phase:Wdh           2   1.4112  0.2449
Gruppe:Phase:Wdh    6   1.2952  0.2578
```

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“

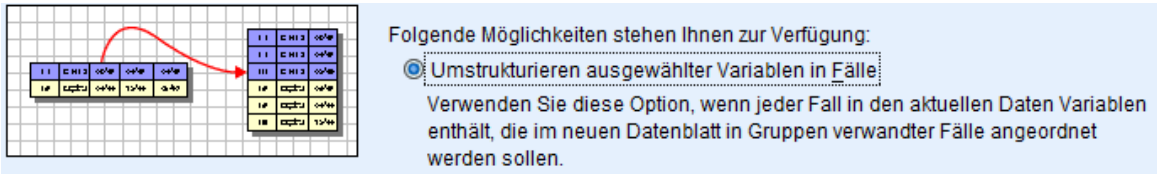
1. 1. 1 ein Faktor und eine Analyse-Variable

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**

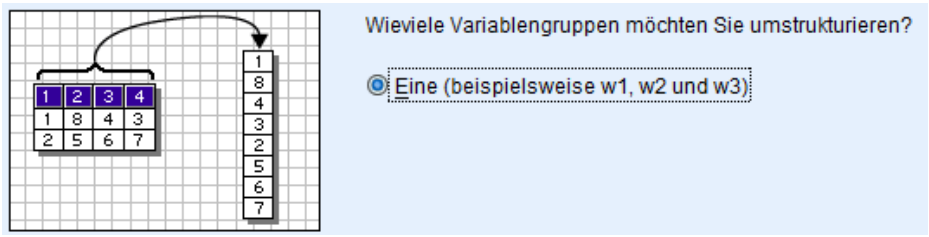
  1. Option:
  Umstrukturieren ausgewählter Variablen in Fälle



  -> Weiter

- **Anzahl der Variablengruppen**

  1. Option:
  Eine (Variablengruppe)



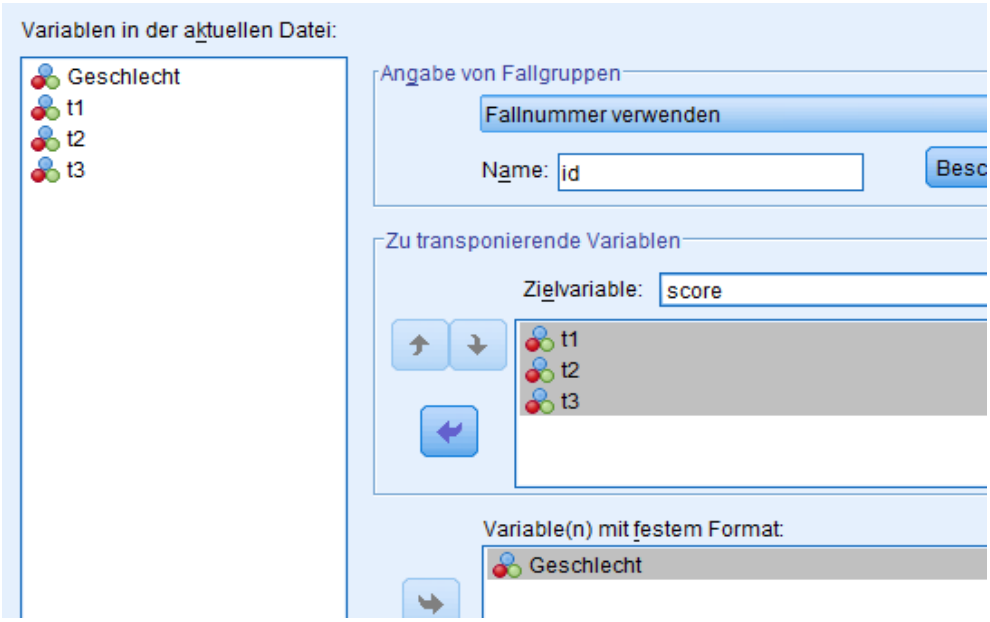
  -> Weiter

- **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: „score“

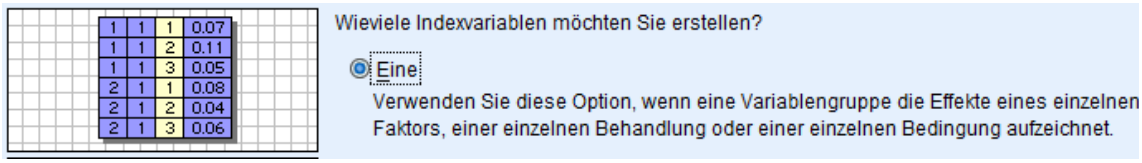
  c. Variablen mit festem Format:
  hier die "konstanten" Variablen (ohne Messwiederholung) eintragen
  (z.B. Alter, Geschlecht etc)



-> Weiter

- **Wieviel Indexvariablen möchten Sie erstellen?**

  1. Option:
  Eine (Indexvariablen)

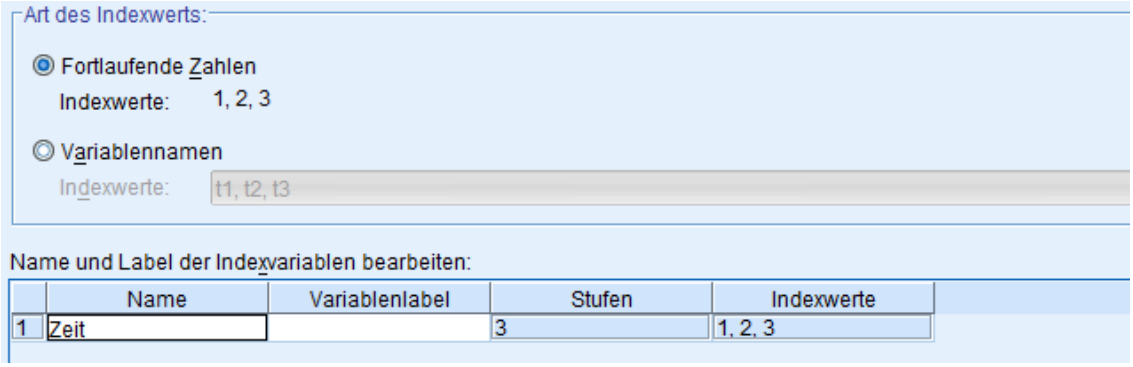


-> Weiter

- **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: „Zeit“

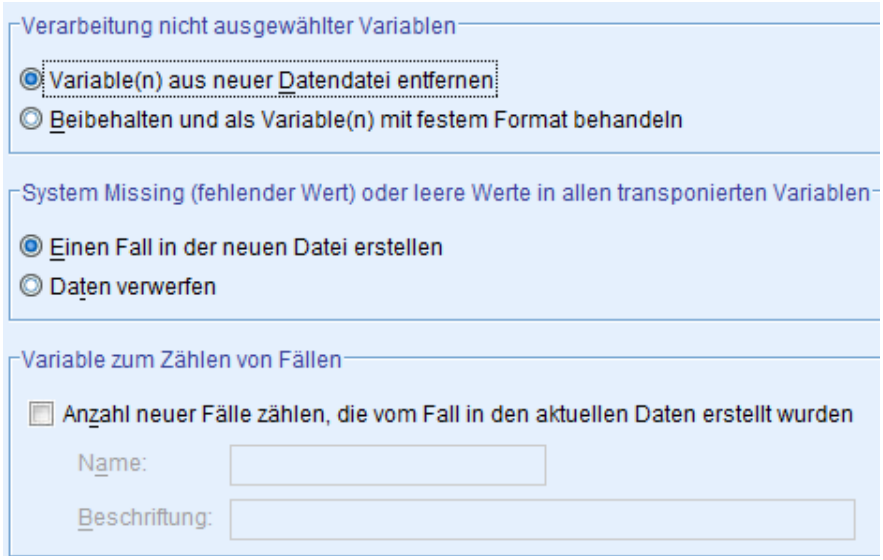


-> Weiter (es folgen dann noch Optionen) oder Fertigstellen

- **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



-> 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

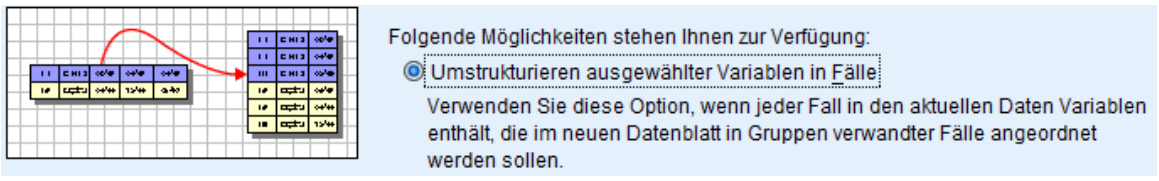
1. 1. 2 mehrere Faktoren und eine Analyse-Variablen

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**

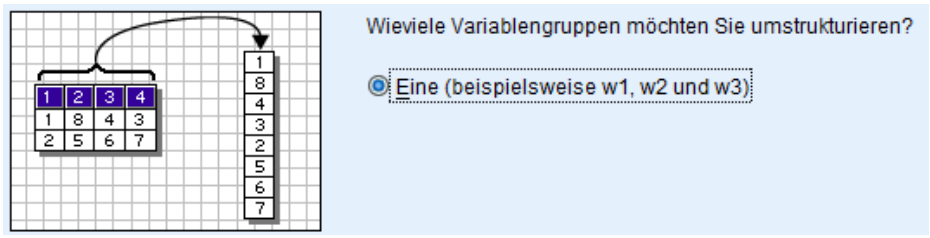
  1. Option:
  Umstrukturieren ausgewählter Variablen in Fälle



  -> 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

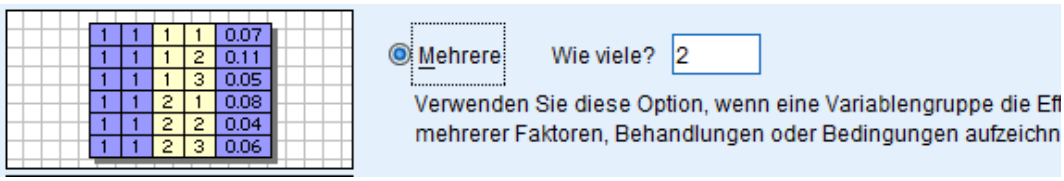
  b. zu transponierende Variablen:
  hier die Messwiederholungsvariablen eintragen
  und einen gemeinsamen Namen geben, hier: „Fehler“

  c. Variablen mit festem Format:
  hier die "konstanten" Variablen (ohne Messwiederholung) eintragen
  (z.B. Alter, Geschlecht etc)

  -> Weiter

- **Wieviel Indexvariablen möchten Sie erstellen?**

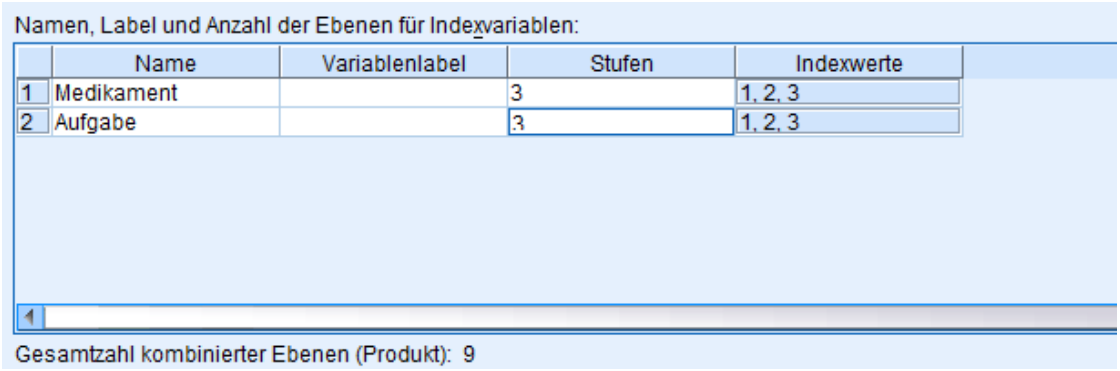
  2. Option:
  Mehrere (Indexvariablen) und Anzahl der Messwiederholungsfaktoren festlegen



-> 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 festleget 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:

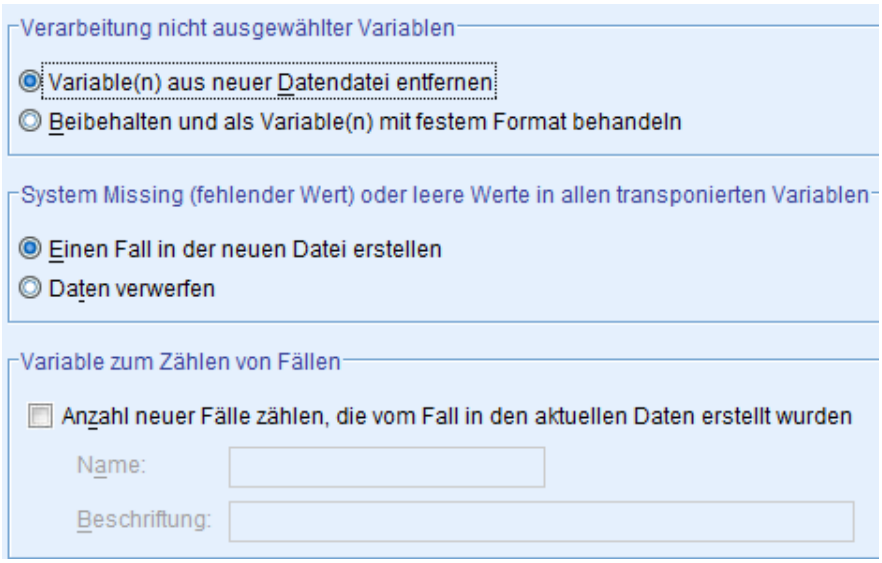


-> Weiter

- 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



-> Weiter

-> Fertigstellen

Wenn keine Namen festegelgt 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

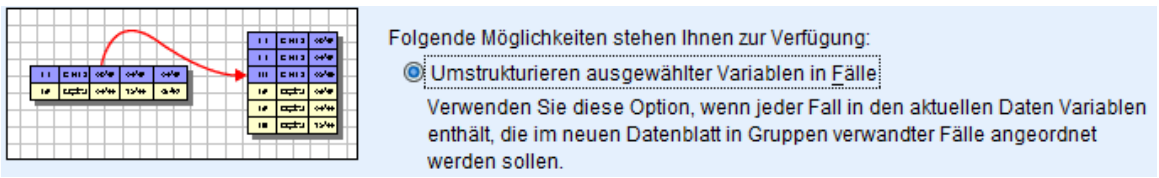
1. 1. 3 ein Faktor und mehrere Analyse-Variablen

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**

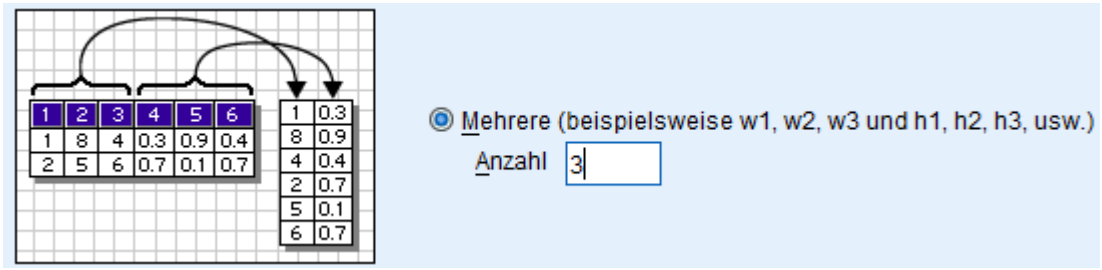
  1. Option: Umstrukturieren ausgewählter Variablen in Fälle

  -> Weiter



- **Anzahl der Variablengruppen**

  2. Option:
  Mehrere (Variablengruppen) sowie Anzahl der Analyse-Variablen festlegen (hier 3)



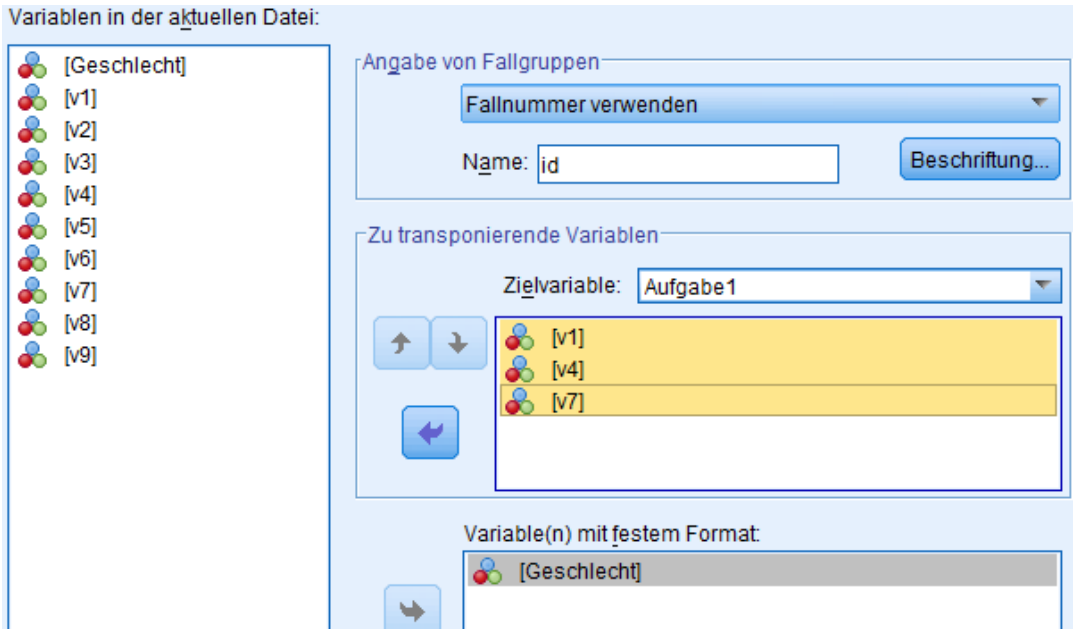
-> Weiter

- **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 weren, deren Voreinstellung `trans1, trans2,` ... ist.

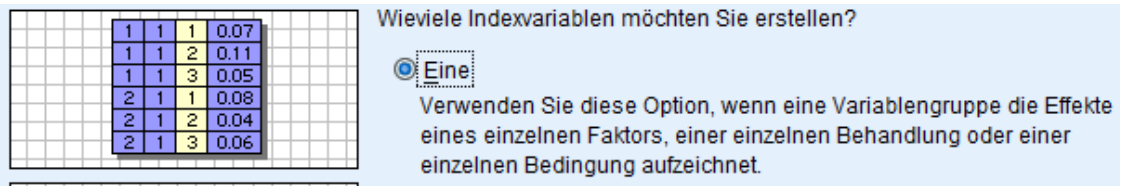
  c. Variablen mit festem Format:
  hier die "konstanten" Variablen (ohne Messwiederholung) eintragen
  (z.B. Alter, Geschlecht etc)



-> Weiter

- **Erstellen von Indexvariablen**)

  1. Option:
  Eine (Indexvariablen)



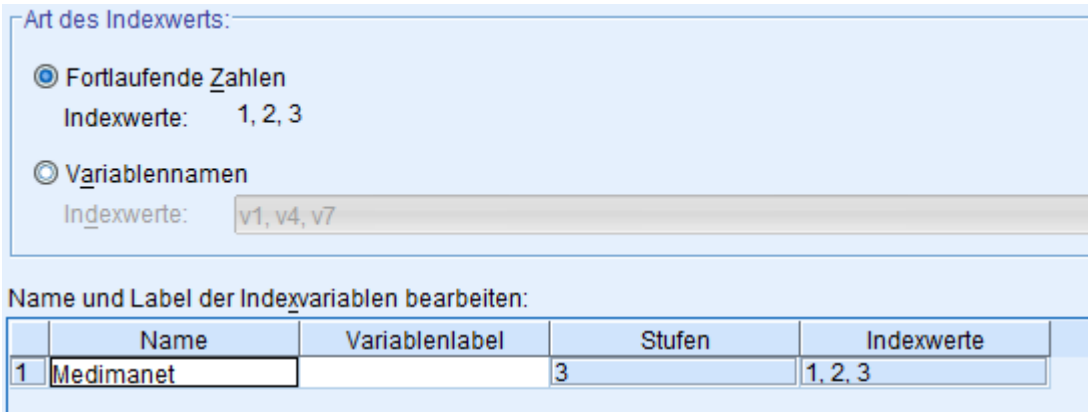
Liegt ein mehrfaktorielles Design wir, wie etwa im vorigen Abschnitt, so können bei der 2. Option die Anzahl der Messwiederholungsfaktoren festgelegt werden.

-> 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.

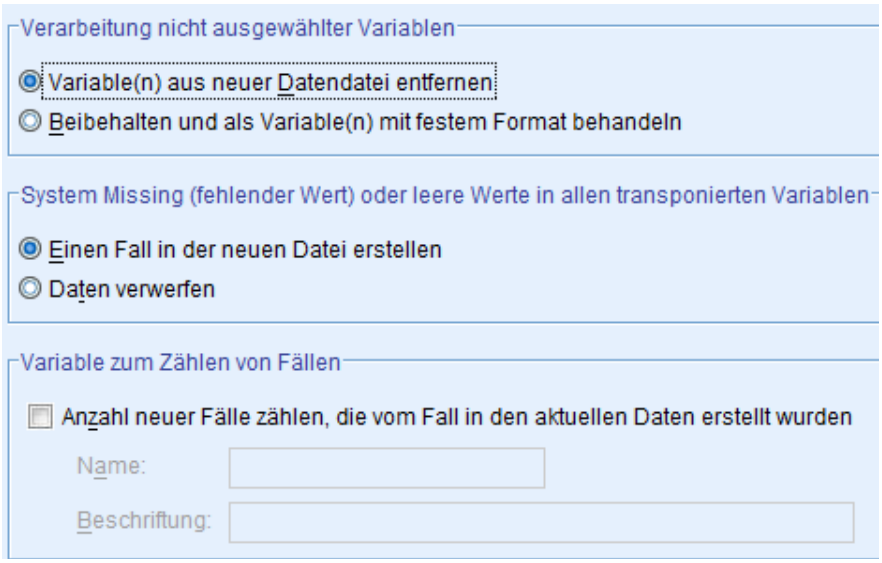


-> Weiter

- **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



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:

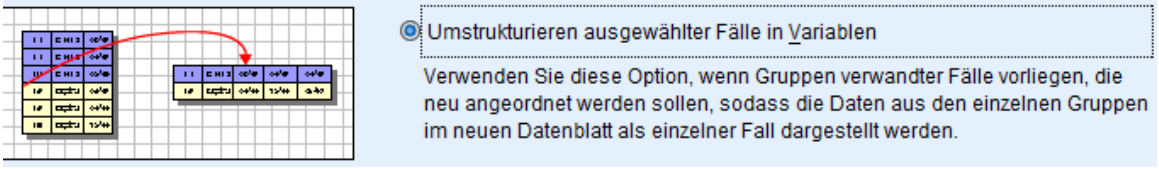
	id	Geschlecht	Medikament	Aufgabe1	Aufgabe2	Aufgabe3
1	1	1	1	3	3	1
2	1	1	2	4	4	2
3	1	1	3	5	4	3
4	2	1	1	2	0	0
5	2	1	2	3	2	2
6	2	1	3	4	3	3
7	3	1	1	5	4	3
8	3	1	2	5	3	3
9	3	1	3	6	3	4
10	4	1	1	3	5	2
11	4	1	2	4	4	3
12	4	1	3	4	4	4

1. 2 Umstrukturieren von Fälle in Messwiederholungen

Vorzunehmen im Menü: „Daten -> Umstrukturieren“

- **Datenumstrukturierung**

  2. Option: Umstrukturieren ausgewählter Variablen in Fälle

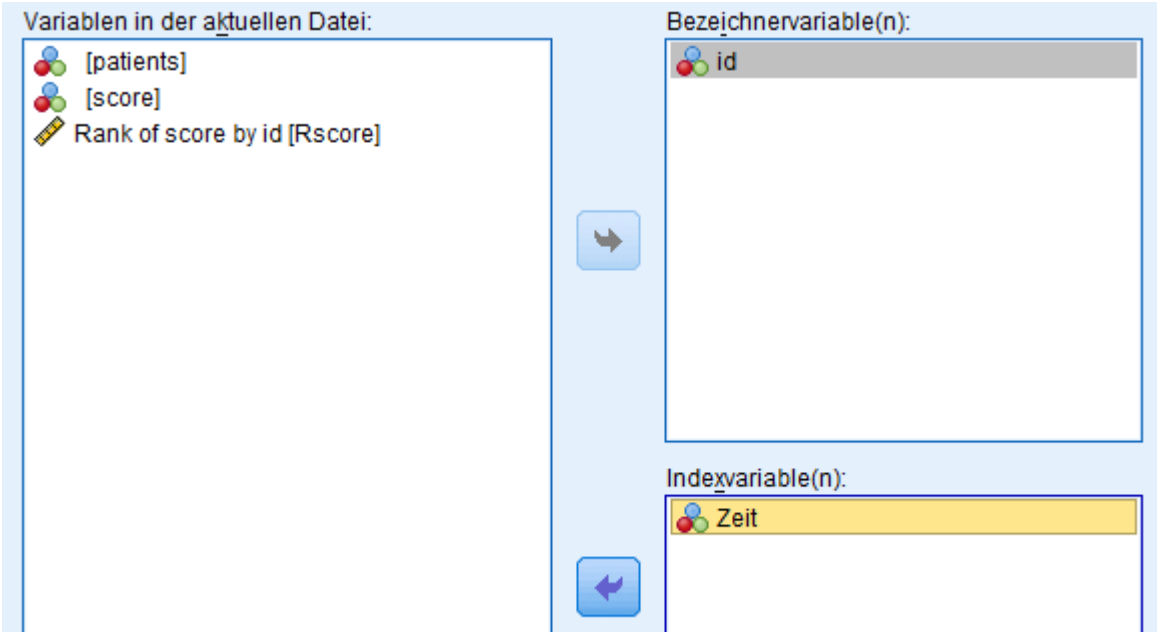


  -> Weiter

- **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,...)



  Alle übrigen Variablen werden automatisch „sinnvoll“ als konstante oder Messwiederholungsvariable zugeordnet.

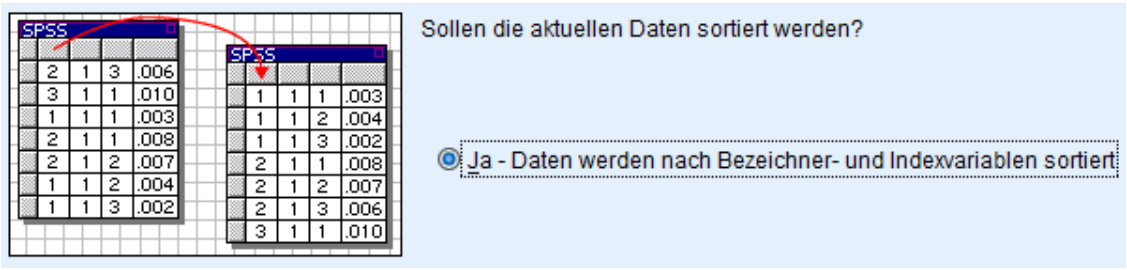
  -> Weiter

- **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)



-> Weiter

- **Optionen**

Die Optionen sind i.a. nicht erforderlich.

-> Weiter

-> Fertigstellen

Die neuen Namen der Messwiederlungen 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 mit Varianzen s_i^2 vor. Der F-Test (Haupteffekt oder Interaktion)

$$F = \frac{MS_{Effekt}}{MS_{Fehler}}$$

kann bzgl. der Heterogenität der Varianzen korrigiert werden, indem die Zähler- und Nennerfreiheitsgrade adjustiert (genauer: reduziert) werden. Die Zählerfreiheitsgrade *df1* werden dabei mit ε_1 multipliziert, die Nennerfreiheitsgarde *df2* mit ε_2. Diese Korrekturfaktoren errechnen sich wie folgt:

$$\bar{s}^2 = (\sum s_i^2)/k$$

$$c^2 = \left(\sum (s_i^2 - \bar{s}^2)^2\right)/(k \cdot \bar{s}^4)$$

$$\varepsilon_1 = \left(1 + \frac{k-2}{k-1}c^2\right)^{-1} \qquad \varepsilon_2 = (1 + c^2)^{-1}$$

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 ε_1 und ε_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:

$$F = \frac{SS_{Effect}}{SS_{Error}}$$

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

$$df = \left(\sum \frac{m_i^2}{n_i - 1}\right)^{-1} \qquad m_i = \left(1 - \frac{n_i}{n}\right) s_i^2 / (SS_{Error})$$

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 Nennerfreitsgrade 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.

$$S_2 = \sum_i^k \sum_j^n (x_{ij} - \bar{x})^2 \qquad S_4 = \sum_i^k \sum_j^n (x_{ij} - \bar{x})^4$$

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) -> log(x) für a -> 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 $P_1,...,P_k$ erzielt, dann wird mit der folgenden Testgröße *X* die Hypothese geprüft, dass für alle *k* Tests H_0 richtig ist:

$$X = -2[ln(P_1) + ln(P_2) + \ldots + ln(P_k)]$$

X ist χ^2-verteilt mit 2*k* Freiheitsgraden.

Mehr dazu unter https://en.wikipedia.org/wiki/Fishers_method

2. 6 Levene-Test auf Gleichheit von Kovarianzmatrizen

Der Levene-Test auf Gleichheit von mehreren Varianzen aus unbahängigen 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,...,I Gruppen, k=1,...,n_i , N ($N = \sum n_i$) Erhebungseinheiten und j=1,..,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:

$$s_{k(i)j_1j_2} = (y_{k(i)j_1} - m_{ij_1})(y_{k(i)j_2} - m_{ij_2}) \quad (j_1, j_2 = 1,...,J)$$

die anschließend transformiert werden in

$$\hat{s}_{k(i)j_1j_2} = \mathrm{sgn}(s_{k(i)j_1j_2}) \cdot \sqrt{|s_{k(i)j_1j_2}|}$$

wobei sgn die Vorzeichen-Funktion bezeichnet. Im nächsten Schritt wird für jede Erhebungseinheit $k(i)$ die untere Dreiecksmatrix von $\hat{s}_{k(i)j_1j_2}$ in einen Vektor umgewandelt, woraus eine Datenmaztrix 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,...,I testet.

2. 7 Wilcox-Test auf Gleichheit von Varianzen bei Messwiederholungen

Dieser Test prüft die Gleichheit der Varianzen von J abhängigen Variablen $x_1,..,x_J$. Er basiert auf dem Prinzip des bekannten Levene-Tests (vgl. Wilcox, 1989).

Die dazu erforderlichen Schritte:

- Für jede Variable $x_1,..,x_J$ wird der Median m_j ermittelt.
- Für jede Beobachtungseinheit i=1,..,N berechnen der absoluten Differenz $d_{ij} = \mathrm{abs}(x_{ij} - m_j)$.
- 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. anova.lib: 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. Der bequemste Weg ist Folgender:

- Herunterladen der Datei `http://www.uni-koeln.de/~luepsen/R/anova.lib` in ein lokales Verzeichnis *`path`*,
- danach in R verfügbar machen über `attach("`*`path`*`/anova.lib")`

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 aov) Beispiel: x ~ A*B
Dataframe	Datensatz, Objekt vom Type Dataframe

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	varianzanalystisches Modell (vgl. Funktion aov) Beispiel: x ~ A*B
Dataframe	Datensatz, Objekt vom Type Dataframe

3. 3 mbf.f: modified Brown & Forsythe-F-test für inhomogene Varianzen in gemischten Versuchsplänen (Split-Plot Designs)

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 beötigt 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:

df	Dataframe, Object vom Typ data.frame, bei weitem Format: nur die Messwiederholungsvariablen
group	Gruppierungsfaktor
trial	Messwiederholungsfaktor
id	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 ~ 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:

1. Likelihood Ratio Test,
2. Box M-test, basierend auf der multivariaten Normalverteilung, sowohl mit χ^2-Test als auch mit F-Test für kleine Stichproben,
3. Schott‘s T1, eine Verbesserung des Box M-Tests,
4. 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,
5. Schott‘s T3, wie T2, jedoch unter der Annahme, dass nicht für alle Gruppen die Verteilungen gleich sind,
6. multivariater Levene Test, der nur ordinales Skalenniveau voraussetzt.
7. Zwei Dispersionstests, die ebenfalls nur ordinales Skalenniveau voraussetzen, basierend auf Verfahren von O‘Brien.

Die Tests 1-5 benötigen Stichprobenumfänge n_i>Anzahl Variablen+1.

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

Allgemeine Aufrufparameter:

transf=T	Transformation der Covarianz Matrix CSC' mit einer Contrast Matrix
rank=T	Transformation der Rohwerte in Ränge
par=F	keine parametrischen Homogenitäts-Tests
npar=F	keine nichtparametrischen Homogenitäts-Tests

Beispiele:
```
check.covar (winer[ ,c("V3","V4","V5")], winer$V2)
with(winer,check.covar(v,sex,time,id))
```

Ausgabe:

dataframe mit χ^2-values, df und p Wert für alle 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 Tests benötigen Stichprobenumfänge n_i>Anzahl Variablen.

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.var: Test auf Homogenität von Varianzen bei abhängigen Stichproben

Durchführung von Tests zur Prüfung der Gleichheit von Varianzen bei Messwiederholungen:

- ein nichtparametrischer Test, von Wilcox „Q“ benannt, basierend auf dem Quade-Test,
- nichtparametrische Levene-like-Tests, unter Verwendung des Friedman-Tests sowohl mit χ^2-Test als auch Iman-Davenports F-Test sowie des Quade-Tests.

Die Funktion kann sowohl auf Datensätze im “wide format“ als auch im “long format“ angewandt werden.

Aufruf (wide format): `check.var (abh. Variablen)`
Aufruf (long format): `check.var (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

Zusätzliche Parameter :

method	1: Abweichungen vom Median (Standard) 2: Abweichungen vom 15% getrimmten arithmetischen Mittel
sel	Variablennumerrn der zu vergleichenden Variablen innerhalb der vorgegebenen Variablen (Standard: alle Variablen)

Beispiele:
```
check.var (winer[ ,c("V3","V4","V5")])
check.var (winer[ ,3:5)], sel=c(1,3), method=2)
with(winer,check.var(score,trial=time,id=Vpn))
```

3. 8 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-Wertes (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 Tests benötigen einen Gesamt-Stichprobenumfang *N*>Anzahl Variablen+2.

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 χ^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 ε

`$error`: error code:
2: Datenmatrix und Gruppierungsvektor haben unterschiedliche Länge.

3. 9 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 aov) Beispiel: x ~ A*B
Dataframe	Datensatz, Objekt vom Type Dataframe

3. 10 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-Transformation, verallgemeinerte Kruskal-Wallis- und Friedman-Ananlysen (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 aov) Beispiele: x ~ A*B oder score ~ gruppe*Zeit+Error(Vpn/Zeit)
Dataframe	Datensatz, Objekt vom Type Dataframe
method	0 : KWF-Methode 1 : van der Waerden-Methode 2 : Puri & Sen (L statistic) 3 : Puri & Sen mit INT-Transformation
compact	im Falle von Messwiederholungen: T: alle Tests in einer Dataframe-Tabelle (default) F: für jeden Fehlerterm eine getrennte Tabelle (wie bei `summary(aov)`)
pseudo	F (klassische Ränge) oder T (Pseudo-Ränge), vgl. Kapitel 2.16

3. 11 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 aov) Beispiel: x ~ A*B)
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
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. 12 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 ~ Medi*Aufgabe+Error(Vpn/(Medi*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. 13 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: score ~ gruppe*Zeit+Error(Vpn/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. 14 wj.anova: Welch-James-Varianzanalyse für heterogene Varianzen (nur für 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. 15 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. 16 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, 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. 17 ap.anova: nichtparametrische Varianzanalyse für Messwiederholungen und Split-Plot Versuchspläne von Agresti & Pendergast

1- und 2-faktortielle Varianzanalyse mit einem Messwiederholungsfaktor nach A. Agresti & J. Pendergast, basierend auf dem multivariaten Test von Hotelling & Lawley, der keine Sphetrizität 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:

dataframe	data, object of class data.frame
dependent variable	abhängige Variable
case id	Fallidentifikation (class factor)
trial factor	Messwiederholungsfaktor (class factor)
grouping factor	Gruppierunghsfaktor (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 iga und iga.anova: general approximation test (GA) und improved general approximation test (IGA) von H.Huynh

Adjustierungen für den parametrischern 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:

df	Dataframe, Object vom Typ data.frame, bei weitem Format: nur die Messwiederholungsvariablen
group	Gruppierungsfaktor
trial	Messwiederholungsfaktor
id	Fallidentifikation
ga	F: IGA adjustment, T: GA adjustment

Ergebnis von `iga`: Liste mit 4 Vektoren

`GA.B`	GA-adjustments for the repeated measures main effect
`GA.AB`	GA-adjustments for the interaction effect
`IGA.B`	IGA-adjustments for the repeated measures main effect
`IGA.AB`	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. 19 simple.effects: parametrische Analyse von simple effects

Analyse der simple effcts 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 `aov`
Interaktion	Spezifikation der Interaktion, z.B. `"Geschlecht*Zeit"` , mehrere zu analysierende Interaktionen können mittels `c(..)` zusammengefasst werden.
Dataframe	Datensatz vom Type Dataframe, der auch für `aov` verwendet wurde
adjust	optional: α-Adjustierung, vgl. R-Funktion `p.adjust` (default: „none“)

3. 20 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 Wals-Test nach Fan & Zhang. Ersterer ist sehr liberal insbesondere bei GEE- und GLMM-Modellen, bei denen die Kovarianzamatrizen der Parameterschätzungen generell zu klein geschätzt werden und dadurch zu große χ^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:

coefficients	regression coefficients (details see below)
covariance matrix	
degrees of freedom	Array with 3 df for 2 factors and the interaction
n	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
χ^2-value
p value

`gee.robanova` degrees of freedom
χ^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

3. 21 rob.anova: 1- / 2-faktorielle robuste Varianzanalysen

1- / 2-faktorielle robuste Varianzanalysen

Aufruf: `rob.anova (Modell, Dataframe, method=...)`

Aufrufparameter:

Modell	varianzanalytisches Modell (vgl. Funktion aov)	
Dataframe		Datensatz
method =	bi:	Bisquare Psi Function using iteratively reweighted least squares (IRLS), the function calculates the optimal weights to perform m-estimator or bounded influence regression.
	Huber	Huber Psi Function using iteratively reweighted least squares (IRLS), the function calculates the optimal weights to perform m-estimator or bounded influence regression.
	MM	MM-type estimators for linear models as described in Yohai (1987), using a bi-square redescending score function (
	M	robust regression using an standard M estimator
	lts	Robust Fitting using least trimmed squared residuals minimize the sum of the quantile smallest squared residuals.
	lqs	Robust Fitting using least quantile of squared residuals minimize the quantile squared residual
	lms	Robust Fitting using least median squared residuals

Literatur:

Yohai, V.J. (1987) High breakdown-point and high efficiency estimates for regression. The Annals of Statistics 15, 642–65

P. J. Rousseeuw and A. M. Leroy (1987) Robust Regression and Outlier Detection. Wiley.

Literaturhinweise

Adebayo, O. P., Ogunjimi, O. A., & Ahmed, I. (2024). On identifying a robust alternative method for ANOVA with unequal variance. group, 1, 2.

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, M. G., & Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *Journal of the American Statistical Association*, 89(425), 336-343.

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).

Ananda, M. M., & Weerahandi, S. (1997). Two-way ANOVA with unequal cell frequencies and unequal variances. *Statistica Sinica*, 631-646.

Ananda, M.M., Dag, O., Weerahandi, S. (2023). Heteroscedastic two-way ANOVA under constraints. *Communications in Statistics-Theory and Methods*, 52:22, 8207-8222.

Arnau J, Bono R, Vallejo G. 2009. Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics - Simulation and Computation*. 38(5): 1083-1103.

Atkinson, A. C., & Cheng, T. C. (1999). Computing least trimmed squares regression with the forward search. *Statistics and Computing*, 9 (4), 251-263.

Baltes-Götz, Bernhard (2016): Generalisierte lineare Modelle und GEE-Modelle in SPSS Statistics, ZIMK, Universität Trier

Barron, Sheila et al. (2015): Sums of Squares: The Basics and a Surprise. https://support.sas.com/resources/papers/proceedings15/1521-2015.pdf

Beasley, T. M. (2000). Nonparametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs. *Journal of Educational and Behavioral Statistics*, 25(1), 20-59.

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? *Behavourial 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, J*ournal of Modern Applies Statistical Methods*, Vol 8, N0. 1 , pp 16-50

Bennett, B. M. (1968). Rank-order tests of linear hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(3), 483-489.

Bertsimas, D., & Mazumder, R. (2014). Least quantile regression via modern optimization. *The Annaly of Statistics*, Vol 42, No 6, 2494-2525.

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

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4), 317-346.

Box, G.E.P. (1953): Non-normality and tests on variances, *Biometrika* 40, pp. 318-335

Box, G.E.P. (1954): Some theorems on quadrative 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 Socierty*, Series B, Vol XVII, No 1

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 StatisticalAssociation*, 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, 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

Carletti, I. , Claustriaux, J.J. (2005). Anova or Aligned Rank Transform Methods: Which one use when Assumptions are not fulfilled ? *Buletinul USAMV-CN*, nr. 62, 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

Coombs, W.T. & Algina, J. (1992): *Four new solutions to the multivatriate G-sample Behrens-Fisher problem*. Paper presented at the meeting of the Psychometric Society, Ohio

Conover, W.J. (1980): *Practical nonparametric Ststitsics*, (Vol 350), John Wiley

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

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4), 351-361.

Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods,* 6(1), 12.

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

Dag, O., Kasikci, M., Yilmaz, M.A., Weerahandi, S., Ananda, M.M.A. (2024). twowaytests: An R Package for Two-Way Tests in Independent Groups Designs. *SoftwareX*, 27, 1-8.

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)

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.

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 l. (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 Heterogenity 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.

Farrar, T., Blignaut, R., Luus, R., & Steel, S. (2025). A review and comparison of methods of testing for heteroskedasticity in the linear regression model. *Journal of Applied Statistics*, 52(16), 3121-3150.

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.

Friedrich, S., Konietschke, F., Pauly, M. (2017). GFD - An R-package for the Analysis of General Factorial Designs. Journal of Statistical Software, Code Snippets 79(1), 1–18

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.

Gaonkar, M. P., & Beasley, T. M. (2023). Comparison of tests for heteroscedasticity in between-subjects ANOVA models. *General Linear Model Journal,* 47(1), 15-32.

García, P. F., Vallejo, G., Livacic-Rojas, P., Herrero, J., & Cuesta, M. (2010). Compative 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

Guerin L, Stroup WW. 2000. A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Annual Conference on Applied Statistics in Agriculture, Kansas State University*.

Hahn, S., Konietschke, 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, SanFrancisco, CA

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, 72(358), 320-338.

Hettmansperger, Thomas P. & McKean, Joseph W. (2011): *Robust Nonparametric Statistical Methods*, CRC Press

Higgins, James J. & Tashtoush, Souleiman (1994). An aligned rank transform test for interaction. *Nonlinear World* 1, pp 201-211.

Hodges Jr, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, Vol. 33, No. 2 .

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

Hsiung, T. H., & Olejnik, S. (1996). Type I error rates and statistical power for the James second-order test and the univariate F test in two-way fixed-effects ANOVA models under heteroscedasticity and/or nonnormality. *The Journal of Experimental Education,* 65(1), 57-71.

Huang, M.L. (2007): A Quantile-Score Test for Experimental Design, *Applied Mathematical Sciences*, Vol. 1, No 11, pp 507-516.

Huber, P. J. (1981): *Robust Statistics*, Wiley, N. Y.

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*, Communications in Statistics - Theory and Methods*, 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.

Katsileros, A., Antonetsis, N., Mouzaidis, P., Tani, E., Bebeli, P. J., & Karagrigoriou, A. (2024). A comparison of tests for homoscedasticity using simulation and empirical data. *Communications for Statistical Applications and Methods,* 31(1), 1-35.

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.

Keselman, H. J., Wilcox, R. R., Taylor, J., & Kowalchuk, R. K. (2000). Tests for mean equality that do not require homogeneity of variances: Do they really work?. *Communications in Statistics-Simulation and Computation*, 29(3), 875-895.

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

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior research methods*, 53(6), 2576-2590.

Koch, Gary (1969): Some aspects of the statistical analysis of split plot experiments in completely randomized layouts. *Journal of the American Statistical Association*, 64, No. 326.

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 ε Approximate Test in Repeated Measures Designs With Two or More Independent Groups. *Journal of Educational Statistics*, Vol. 16, No. 4, pp. 371-372

Lehmann, E.L. (1975): *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

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

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*, 15(1), 1-12.

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

Liang, K.Y. & Zeger S.L. (1986): A Comparison of Two Bias-Corrected Covariance Estimators for Generalized Estimating Equations. *Biometrika* 73, pp 13–22.

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.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, 58(3), pp. 409-429.

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

Luepsen, Haiko (2016): *The lognormal distribution and nonparametric anovas - a dangerous alliance*, Universität zu Köln,
URL: http://www.uni-koeln.de/~luepsen/statistik/texte/lognormal-anova.pdf

Luepsen, H. (2017). The aligned rank transform and discrete variables: A warning. *Communications in Statistics-Simulation and Computation*, 46(9), 6923-6936.

Luepsen, H. (2018). Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. *Communications in Statistics-Simulation and Computation*, 47(9), 2547-2576.

Luepsen, Haiko (2020a): *Checking the Homogeneity of Covariance Matrices: some practical aspects.*
URL:http://www.uni-koeln.de/~luepsen/statistik/texte/Checking.Homogeneity.pdf

Luepsen, Haiko (2020b): *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

Luepsen, Haiko (2020c): *Anmerkungen zum Testen der Spherizität.*
URL: http://www.uni-koeln.de/~luepsen/statistik/texte/Spherizitaet.pdf

Luepsen, Haiko (2023a). ANOVA with binary variables: the F-test and some alternatives. *Communications in Statistics-Simulation and Computation*, 52(3), 745-769.

Luepsen, Haiko (2023b): Generalizations of the Tests by Kruskal-Wallis, Friedman and van der Waerden for Split-plot Designs. *Austrian Journal of Statistics*, 52(5), pp 101-130.

Luepsen, Haiko (2024): Rating of Anova Methods for Split-plot Designs in heterogeneous Conditions, *to be published*.

Luh, W. M. (1999). Developing trimmed mean test statistics for two-way fixed-effects ANOVA models under variance heterogeneity and nonnormality. *The Journal of Experimental Education*, 67(3), 243-264.

Mair, P., & Wilcox, R. (2019). Robust statistical methods using WRS2. *The WRS2 package*.
URL: https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.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.

McNeish, D. & Stapleton, L.M. (2016): Modeling Clustered Data with Very Few Clusters, *Multivariate Behavioral Research*, 51 (4), pp 495-518.

Mehrotra, Devan V. (1997): Improving the brown-forsythe solution to the generalized behrens-fisher problem, *Communications in Statistics - Simulation and Computation*, 26:3, pp 1139-1145.

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

Myers, L. (1998). Comparability of the James' second-order approximation test and the Alexander and Govern A statistic for non-normal heteroscedastic data. J*ournal of Statistical Computation and Simulation,* 60(3), 207-222.

Nagao, Hisao (1973): On Some Test Criteria for Covariance Matrix. T*he Annals of Statistics,* Vol. 1, No. 4, pp 700-709.

Nguyen, D., Kim, E., Wang, Y., Pham, T. V., Chen, Y. H., & Kromrey, J. D. (2020). Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. J*ournal of Modern Applied Statistical Methods*, 18(2), 2.

Noguchi, K., Gel, Y.R., Brunner, E. , Konietschke, 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.

Nooraee, N., Molenberghs, G., & van den Heuvel, E. R. (2014). GEE for longitudinal ordinal data: comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics & Data Analysis*, 77, 70-83.

Nordhausen, K., Oja, H. and Tyler, D.E. (2006): *On the Efficiency of Invariant Multivariate Sign and Rank Tests*, in Festschrift of Tarmo Pukkila on his 60th Birthday, University of Tampere, pp 217–231.

Nordhausen, K. and Oja, H. (2011): Multivariate L1 Methods: The Package MNM, *Journal of Statistical Software*,43, pp 1-28.

Nordstokke, D. W., & Zumbo, B. D. (2010). A new nonparametric Levene test for equal variances. *Psicológica*, 31(2), 401-430.

O'Brien, Peter C. (1992): Robust Procedures for Testing Equality of Covariance Matrices. *Biometrics*, Vol. 48, No. 3 (Sep., 1992), pp. 819-827

Olson, Chester L. (1976): On Choosing a Test Statistic in Multivariate Analysis of Variance. *Psychological Bulletin*, Vol 83, No 4, pp 579-586

Osborne, Jason W. (2008): *Best Practices in Quantitative Methods*, Sage Publications

Peterson, Kathleen (2002): Six Modifications Of The Aligned Rank TransformTest For Interaction, *Journal Of Modem Applied Statistical Methods*, 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.

Reed, J. F., & Stark, D. B. (1995). Robust analysis of variance: a simulation study: Robust analysis of variance. *Journal of Applied Statistics*, 22(1), 87-104.

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.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79 (388), 871-880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley

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.

Schaefer, R. L. (1994). Using default tests in repeated measures: how bad can it get?. *Communications in Statistics-Simulation and Computation*, 23(1), 109-127.

Scheirer, J., Ray, W.S. , Hare, N. (1976): The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs. *Biometrics*. 32(2). pp 429–434

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C. & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, 11(9), 1141-1152.

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.

Serlin, R. C., & Harwell, M. R. (2001). *A review of nonparametric test for complex experimental designs in educational research*. In annual meeting of the American Educational Research Association, Seattle, WA.

Sharma, D., & Kibria, B. G. (2013). On some test statistics for testing homogeneity of variances: a comparative study. *Journal of Statistical Computation and Simulation*, 83(10), pp 1944-1963.

Shear, B. R., Nordstokke, D. W., & Zumbo, B. D. (2018). A Note on Using the Nonparametric Levene Test When Population Means Are Unequal. *Practical Assessment, Research & Evaluation,* 23(13), n13.

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.

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.

Thiele, J., & Markussen, B. (2012). Potential of GLMM in modelling invasive spread. *CABI Reviews,* (2012), 1-10.

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.

Thompson, G. L., & Ammann, L. P. (1990). Efficiencies of interblock rank statistics for repeated measures designs. *Journal of the American Statistical Association*, 85(410), 519-528.

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

Tsandilas T, Casiez G. [forthcoming]. The illusory promise of the aligned rank transform. *Journal of Visualization and Interaction*. https://statransform.github.io/jov .

Tsui, K. W., & Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, 84(406), 602-607.

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.

Twesman, Dennis (2022): *Robust ANOVA*. https://rpubs.com/DeTwes/robANOVA

Ujian, K., Abdullah, N. F., & Munda, N. (2022). An Overview of Homogeneity of Variance Tests on various Conditions based on Type 1 Error Rate and Power of a Test. *Journal of Quality Measurement and Analysis* JQMA, 18(3), 111-130.

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*. 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.

Wang, Y., Rodríguez de Gil, P., Chen, Y. H., Kromrey, J. D., Kim, E. S., Pham, T., ... & Romano, J. L. (2017). Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and psychological measurement,* 77(2), 305-329.

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

Wilcox, R. R., Char1in, V. L., & Thompson, K. L. (1986). New monte carlo results on the robustness of the anova f, w and f statistics. *Communications in Statistics-Simulation and Computation*, 15(4), 933-943.

Wilcox, Rand R. (1989): Comparing the Variances of dependent Groups. *Psychometrika*, vol 54, No. 2, pp 305--315

Wilcox, Rand R. (1993). Analysing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology*, 46(1), pp. 63-76.

Wilcox, Rand R. (2003): *Applying Contemporary Statistical Techniques*, Elsevier

Wilcox, Rand R. (2005): *Introduction to robust estimation and hypothesis testing*, Burlington MA; Elsevier

Wilcox, Rand R. (2017): *Introduction to Robust Estimation and Hypothesis Testing*, Elsevier, Amsterdam

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

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

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173-181.

Yigit, E., & Gokpinar, F. (2010). A simulation study on tests for one-way ANOVA under the unequal variance assumption. *Commun Fac Sci Univ Ankara,* Ser A, 1, 15-34.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, Vol 15, No 20, 642-656.

Zhang, S. (1998). Fourteen Homogeneity of Variance Tests: When and How To Use Them. Paper presented at the - ERIC - Department of Education

Zou, C., Peng, L., Feng, L., Wang, Z. (2014): Multivariate sign-based high-dimensional tests