Die statistische Auswertung von ordinalen Daten bei zwei Zeitpunkten und zwei Stichproben

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Alexander Siemer
aus Göttingen

Göttingen 2002

D 7
Referent: Prof. Dr. Edgar Brunner
Koreferent: Prof. Dr. Manfred Denker
Tag der mündlichen Prüfung: 3. April 2002

Models of course, are never true

but fortunately it is only necessary

that they be useful.

George E. P. Box (1979)[1]

[1] Bickel und Doksum (2001)

Vorwort

Diese Arbeit entstand in den gut zwei Jahren nach meiner Diplomarbeit über die nichtparametrische Kovarianzanalyse. Nach dem Abschluß dieser Arbeit stellte sich die Frage, welche nichtparametrischen Verfahren bei Baseline-Werten anzuwenden sind. Zur Auswahl standen die auf der Diplomarbeit basierenden Kovarianzverfahren, welche die Baseline-Werte als Kovariablen verwenden, und schon etablierte Verfahren, die die Zeit als Faktor berücksichtigen. Um die Arbeit nicht ausufern zu lassen, wurde der Vergleich auf ein Design mit zwei Stichproben und zwei Zeitpunkten eingeschränkt. Nach einigen Überlegungen zu den beiden genannten Verfahren wurde der Vergleich noch erweitert. Es kamen andere Verfahren noch hinzu, weil sie entweder aktuell erschienen waren, wie das Schichtungsverfahren, oder weil sie weit verbreitet sind, wie die parametrischen Verfahren. Somit stellt die vorliegende Arbeit einen breiten Überblick über verschiedene Auswertungsmethoden für das genannte Design dar.

Wie schon die Diplomarbeit wurde auch diese Arbeit an der Abteilung Medizinische Statistik der Universität Göttingen erstellt. Dem Leiter dieser Abteilung, Prof. Dr. Edgar Brunner, gilt mein besonderer Dank für das interessante Thema, die wertvollen Ratschläge und Hinweise und die Bereitstellung der Arbeitsmittel. Außerdem danke ich ihm und Prof. Dr. Manfred Denker für die Erstellung der Gutachten.

Das sorgfältige Korrekturlesen einer Dissertation bereitet sehr viel Mühe. Daher bin ich Frau Iris Schacht, Frau Gudrun Schacht, Frau Hedda Uphoff und Herrn Dr. Sebastian Domhof sehr dankbar, dass sie diese Arbeit übernommen haben. Die Berücksichtigung der Tücken der neuen Rechtschreibung und die Verbesserung einiger Formulierungen ist ihnen zu verdanken. Herr Dr. Sebastian Domhof war zudem nicht nur während der Arbeit an der Dissertation sondern in meinem gesamten Studium ein Freund; und als kompetenter Gesprächspartner war er mir eine große Hilfe. Außer für das Korrekturlesen möchte ich mich bei Frau Iris Schacht für die liebe Unterstützung bedanken. Sie hat mich in anstrengenden Zeiten wieder aufgerichtet.

Außerhalb des Studiums habe ich in Göttingen viele sehr gute Freunde gefunden. Auch ihnen gilt hier mein Dank. Sie haben mich motiviert oder meine Gedanken von der

Mathematik abgelenkt, je nachdem was notwendig war. Die Zeit des Studiums erlebte ich somit wesentlich intensiver.

Nicht nur während des Studiums, sondern schon mein ganzes Leben lang unterstützt mich meine Familie. Ohne ihre materielle und vor allem immaterielle Hilfe hätte ich diese Arbeit nie schreiben können. Dafür gebührt meinem Bruder und meinen Eltern ein besonderer Dank.

Alexander Siemer, Göttingen im Februar 2002

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Motivation 1

1.2 Aufbau der Arbeit 2

2 Skalenniveau und Versuchsplan **5**

2.1 Ordinale Skalen 5

2.2 Versuchsplan 8

3 Vorstellung der Beispiele **11**

3.1 Neurologische Beeinträchtigung bei MS 11

3.2 Behandlung von Schlaflosigkeit 12

4 Modelle, Effekte und Hypothesen **15**

4.1 Nichtparametrische Kovarianzanalyse 15

4.2 Naive Verfahren 20

4.2.1 Pearson-Verfahren 20

4.2.2 Cochran-Armitage-Verfahren 23

4.3 Nichtparametrische faktorielle Analyse 26

4.4 Nichtparametrische Analyse durch Schichtung 35

4.5 Parametrische faktorielle Analyse 39

4.6 Parametrische Kovarianzanalyse 52

5 Test- und Schätzverfahren der nichtparametrischen Kovarianzanalyse **57**

5.1 Randomisierte Versuchsanlage 57

5.2 Nicht randomisierte Versuchsanlage 67

6 Anwendung der nichtparametrischen Kovarianzanalyse 71

6.1 Neurologische Beeinträchtigung bei MS 71

6.2 Behandlung von Schlaflosigkeit 73

7 Vergleich der Verfahren 75

7.1 Definition der Vergleichspunkte 75

7.2 Naive Verfahren 76

7.3 Nichtparametrische faktorielle Analyse 77

7.4 Nichtparametrische Analyse durch Schichtung 78

7.5 Parametrische faktorielle Analyse 80

7.6 Parametrische Kovarianzanalyse 82

8 Zusammenfassung und Ausblick 83

A Test- und Schätzverfahren 85

A.1 Nichtparametrische faktorielle Analyse 85

A.2 Nichtparametrische Analyse durch Schichtung 96

B Beweise 103

B.1 Beweis zu Satz 4.5.1 103

B.2 Beweis zu Satz 5.1.1 104

C Niveau- und Gütesimulationen 107

C.1 Naive Verfahren 107

C.2 Nichtparametrische faktorielle Analyse 112

C.3 Nichtparametrische Analyse durch Schichtung 115

C.4 Parametrische faktorielle Analyse 121

C.5 Parametrische Kovarianzanalyse 129

D Auswertungen **131**

D.1 Naive Verfahren 131

D.1.1 Neurologische Beeinträchtigung bei MS 131

D.1.2 Behandlung von Schlaflosigkeit 133

D.2 Nichtparametrische faktorielle Analyse 134

D.2.1 Neurologische Beeinträchtigung bei MS 134

D.2.2 Behandlung von Schlaflosigkeit 138

D.3 Nichtparametrische Analyse durch Schichtung 140

D.3.1 Neurologische Beeinträchtigung bei MS 140

D.3.2 Behandlung von Schlaflosigkeit 143

D.4 Parametrische faktorielle Analyse 145

D.4.1 Neurologische Beeinträchtigung bei MS 145

D.4.2 Behandlung von Schlaflosigkeit 150

D.5 Parametrische Kovarianzanalyse 154

D.5.1 Neurologische Beeinträchtigung bei MS 154

D.5.2 Behandlung von Schlaflosigkeit 155

Literaturverzeichnis **159**

Lebenslauf **165**

Abbildungsverzeichnis

3.1 Boxplots für die MS-Studie 11

3.2 Histogramme für die Insomniastudie 13

4.1 Erstes konstruiertes Beispiel zur Illustration der Probleme bei nichtrandomisierten Studien 18

4.2 Zweites konstruiertes Beispiel zur Illustration der Probleme bei nichtrandomisierten Studien 19

4.3 Drittes konstruiertes Beispiel zur Illustration der Probleme bei nichtrandomisierten Studien 20

4.4 Hypothetische Wahrscheinlichkeiten für zwei Behandlungen 23

4.5 Beispiel für den Vergleich der naiven Effekte und der Effekte der nichtparametrischen Kovarianzanalyse 25

4.6 Graphen der logistischen Verteilungsfunktion, der Standardnormalverteilung und der extremen Minimalwertverteilung 42

4.7 Graphen der logistischen Verteilungsfunktion, der Standardnormalverteilung und der extremen Minimalwertverteilung bei gleichem Erwartungswert und gleicher Varianz 43

4.8 Logarithmen von Odds im Proportional Odds Modell 46

7.1 Vergleich der Macht zwischen der nichtparametrischen faktoriellen Analyse und der nichtparametrischen Kovarianzanalyse 78

7.2 Vergleich der Macht zwischen der parametrischen faktoriellen Analyse und der nichtparametrischen Kovarianzanalyse 81

C.1 Powersimulation für die naiven Verfahren bei mittleren Stichprobenumfängen 110

C.2 Powersimulation für die naiven Verfahren bei großen Stichprobenumfängen 111

C.3 Powersimulation für die Schichtungsverfahren mit den Gewichten nach van Elteren 119

C.4 Powersimulation für die Schichtungsverfahren mit den Gewichten nach Bajorski und Petkau 120

C.5 Powersimulation für die parametrische faktorielle Analyse 128

D.1 Relative Häufigkeiten der MS-Studie bei den naiven Verfahren 131

D.2 Relative Häufigkeiten der Insomnia-Studie bei den naiven Verfahren 133

D.3 Relative Effekte der MS-Studie im faktoriellen Design 135

D.4 Relative Effekte der Insomnia-Studie im faktoriellen Design 139

D.5 Logarithmen der beboachteten Odds für die Insomnia-Studie 152

Tabellenverzeichnis

2.1 Versuchsplan des Designs 8

3.1 Daten der MS-Studie 12

3.2 Daten der Insomnia-Studie 13

4.1 Schwächen der naiven Effekte bei einem konstruierten Beispiel 22

5.1 Niveausimulation für die nichtparametrische Kovarianzanalyse im balancierten Fall 66

5.2 Niveausimulation für die nichtparametrische Kovarianzanalyse im unbalancierten Fall 67

5.3 Powersimulationen für die nichtparametrische Kovarianzanalyse im balancierten Fall 68

5.4 Niveausimulation für die nichtparametrische Kovarianzanalyse im balancierten und nicht randomisierten Fall 69

5.5 Wahrscheinlichkeiten für die Niveausimulation im nicht randomisierten Fall 70

6.1 Übersicht der nichtparametrischen Kovarianzanalyse für die MS-Studie . . 72

6.2 Übersicht der nichtparametrischen Kovarianzanalyse für die Insomnia-Studie 73

C.1 Erste Niveausimulation für die naiven Verfahren 107

C.2 Zweite Niveausimulation für die naiven Verfahren 108

C.3 Dritte Niveausimulation für die naiven Verfahren 109

C.4 Vierte Niveausimulation für die naiven Verfahren 109

C.5 Fünfte Niveausimulation für die naiven Verfahren 109

C.6 Niveausimulation der nichtparametrischen faktoriellen Analyse für den Gruppeneffekt 112

C.7 Niveausimulation der nichtparametrischen faktoriellen Analyse für den Zeiteffekt . . . 112
C.8 Niveausimulation der nichtparametrischen faktoriellen Analyse für die Wechselwirkung . . . 113
C.9 Powersimulation der nichtparametrischen faktoriellen Analyse für die Wechselwirkung . . . 114
C.10 Raten für keinen Abbruch der Durchführung des Schichtungsverfahrens bei unterschiedlich vielen Kategorien . . . 116
C.11 Niveausimulation für die Schichtungsverfahren bei drei Kategorien . . . 116
C.12 Niveausimulation für die Schichtungsverfahren bei fünf Kategorien . . . 117
C.13 Niveausimulation für die Schichtungsverfahren bei zehn Kategorien . . . 118
C.14 Powersimulation für die Schichtungsverfahren . . . 119
C.15 Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 3 Kategorien . . . 121
C.16 Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 3 Kategorien . . . 122
C.17 Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 3 Kategorien . . . 122
C.18 Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 5 Kategorien . . . 123
C.19 Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 5 Kategorien . . . 124
C.20 Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 5 Kategorien . . . 125
C.21 Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 10 Kategorien . . . 126
C.22 Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 10 Kategorien . . . 127
C.23 Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 10 Kategorien . . . 127
C.24 Niveausimulation für die parametrische Kovarianzanalyse . . . 129

D.1 Kontingenztafel für die MS-Studie bei den naiven Verfahren . . . 132

D.2 Kontingenztafel für die Insomnia-Studie bei den naiven Verfahren 134
D.3 Relative Effekte im faktoriellen Design für die MS-Studie 135
D.4 Übersicht der nichtparametrischen faktoriellen Testergebnisse für die MS-Studie 137
D.5 Relative Effekte im faktoriellen Design für die Insomnia-Studie 138
D.6 Übersicht der nichtparametrischen faktoriellen Testergebnisse für die Insomnia-Studie 140
D.7 Daten der MS-Studie in Schichten 141
D.8 Gewichte der Schichten bei der MS-Studie 142
D.9 Relative Effekte in den Schichten bei der MS-Studie 143
D.10 Testentscheidungen der Schichtungsanalyse der MS-Studie 144
D.11 Gewichte und Stichprobenumfänge der Schichten bei der Insomnia-Studie . 144
D.12 Relative Effekte in den Schichten bei der Insomnia-Studie 144
D.13 Testentscheidungen der Schichtungsanalyse der Insomnia-Studie 145
D.14 Parameterschätzer der parametrischen faktoriellen Analyse bei der MS-Studie 146
D.15 Beobachtete und erwartete Häufigkeiten der MS-Studie 147
D.16 Geschätzte Odds Ratios im parametrischen faktoriellen Modell bei der MS-Studie 149
D.17 Teststatistiken und p-Werte im parametrischen faktoriellen Modell bei der MS-Studie 149
D.18 Parameterschätzer der parametrischen faktoriellen Analyse bei der Insomnia-Studie 150
D.19 Beobachtete und erwartete Häufigkeiten der Insomnia-Studie 151
D.20 Geschätzte Odds Ratios im parametrischen faktoriellen Modell bei der Insomnia-Studie 153
D.21 Teststatistiken und p-Werte im parametrischen faktoriellen Modell bei der Insomnia-Studie 153
D.22 Schätzer der α_{js} in der MS-Studie 155
D.23 Schätzer der α_{js} in der Insomnia-Studie 156
D.24 Erwartete und beobachtete Häufigkeiten bei der Insomnia-Studie 157

Kapitel 1

Einleitung

1.1 Motivation

Eine der einfachsten biometrischen Fragestellungen ist, ob sich zwei verschiedene Behandlungen, zum Beispiel Therapien mit verschiedenen Medikamenten, unterscheiden. Zur Beantwortung einer solchen Frage werden Experimente an Patienten oder allgemein an Versuchseinheiten in der Weise durchgeführt, dass eine Gruppe von Versuchseinheiten die eine und eine andere Gruppe die andere Behandlung erfährt. Danach soll anhand einer Variablen entschieden werden, ob sich die Behandlungen unterscheiden. Diese Variable wird als Zielvariable oder primäre Variable bezeichnet. Sie steht aber häufig nicht nur zu der Behandlung, sondern auch zu anderen Größen, wie zum Beispiel Alter, Gewicht oder der Zielvariablen vor der Behandlung, in Beziehung. Es stellt sich die Frage, wie diese Größen, die als Kovariablen oder sekundäre Variablen bezeichnet werden, in die Analyse der Daten einbezogen werden können.

In der 1998 von einer *Efficacy Working Group* (EWG) fertig gestellten Richtlinie „Statistical Principles for Clinical Trials (ICH E9)[1]“ der *International Conference on Harmonisation* (ICH) wird dieses Thema angesprochen. In Kapitel 5.7 der Richtlinie heißt es: „The primary variable is ... often systematically related to other influences apart from treatment.“ Diese Richtlinie ist von Bedeutung für viele Statistiker, insbesondere für diejenigen, die an der Entwicklung neuer medizinischer Produkte beteiligt sind (Lewis, 1999). Die Richtlinie unterscheidet ferner zwischen zwei Zielen, die eine Berücksichtigung der zusätzlichen Einflussgrößen in der Analyse haben kann. So kann zum einen die Präzision der Auswertungsverfahren erhöht werden. Zum anderen kann eine Ungleichheit zwischen den Gruppen der Versuchseinheiten kompensiert werden.

[1] http://www.emea.eu.int/pdfs/human/ich/036396en.pdf

Eine besondere Art einer zusätzlichen Einflussgröße ist die Betrachtung der primären Zielvariablen vor der Behandlung. So besagt die ICH E9 Richtlinie: „Special attention should be paid ... to the role of baseline measurements of the primary variable.“ Dieser Hinweis ist von sehr allgemeiner Art, da die Richtlinie sich auf grundlegende statistische Prinzipien beschränkt. Die Auslegung dieser Prinzipien hat zu Meinungsverschiedenheiten und Diskussionen geführt. Es wurden verschiedene weitere Richtlinien gegeben, die einige der diskutierten Punkte näher beleuchten. Für die Berücksichtigung der Baseline-Werte wurde im Februar 2000 ein *concept paper* (CP) [2] mit dem Namen „Adjustement for Baseline Covariates“ entworfen, auf dessen Basis ein *Points to Consider Paper* (PtC) entstanden ist. Dieses ist seit Mitte Dezember 2001 öffentlich zugängig und stellt den aktuellen Stand der Diskussion dar. Die im CP gestellte Frage „What analyses should be planned to relate the primary outcome to the covariates ... ?“ wird im PtC konzeptionell diskutiert. In dieser Arbeit gehen wir noch einen Schritt weiter. Wir stellen mit der nichtparametrischen Kovarianzanalyse ein Verfahren vor, mit dem die Baseline-Werte berücksichtigt werden können. Durch einen Vergleich mit verschiedenen anderen Verfahren zur Berücksichtigung von Baseline-Werten aus der Literatur, werden die Vorteile der nichtparametrischen Kovarianzanalyse deutlich. Diese bestehen zunächst in den geringen Modellannahmen. Des Weiteren sind die Effekte und Hypothesen sehr gut zu interpretieren. Außerdem werden die Alternativen zu den Hypothesen gut nachgewiesen, da die Teststatistiken eine große Macht haben und zudem das Niveau auch bei kleinen Stichprobenumfängen einhalten.

Bei der Auswahl der statistischen Verfahren zur Analyse von Daten spielt die Skala, auf der die Beobachtungen liegen, eine entscheidende Rolle. Besondere Probleme treten auf, wenn die Skala ordinal ist. Eine wesentliche Eigenschaft aller in dieser Arbeit betrachteten Verfahren ist ihre Eignung für eine solche Skala. Dabei hat die nichtparametrische Kovarianzanalyse den Vorteil, dass sie kaum von der Anzahl der Kategorien der ordinalen Skala abhängt. Auch bei vielen Kategorien, wo andere Verfahren versagen, zeigt sie gute Eigenschaften.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich wie folgt. Nach der Einleitung werden im Kapitel 2 einige grundlegende Eigenschaften der Experimente, die mit den diskutierten Verfahren ausgewertet werden können, definiert und diskutiert. Dabei wird vor allem auf die Skala und das Versuchsdesign eingegangen. Im darauf folgenden Kapitel 3 werden zwei Beispiele vorgestellt, mit denen die Eigenschaften der Verfahren verdeutlicht werden sollen.

[2] http://www.emea.eu.int/pdfs/human/ewp/286399en.pdf

Danach werden im Kapitel 4 die Modelle, Effekte und Hypothesen dieser Verfahren vorgestellt. Als erstes wird auf die nichtparametrische Kovarianzanalyse eingegangen. Deren Voraussetzungen werden dann in den folgenden Abschnitten mit den Annahmen der anderen Verfahren verglichen. Dabei wird auch die unterschiedliche Interpretierbarkeit der Effekte und Hypothesen diskutiert. Um die Hypothesen zu testen, werden Teststatistiken benötigt. Diese werden für die nichtparametrische Kovarianzanalyse im Kapitel 5 hergeleitet, um dann im Kapitel 6 auf die schon vorgestellten Beispiele angewendet zu werden. Damit ist dann die nichtparametrische Kovarianzanalyse vorgestellt und angewendet worden. Wie sie sich, vor allem auch in Bezug auf die Anwendung, von den anderen Verfahren unterscheidet, wird im Kapitel 7 beschrieben. Die Ergebnisse dieser Vergleiche und der vorhergehenden Kapitel wird abschließend im Kapitel 8 zusammengefasst. Zudem wird dort ein Ausblick auf noch offene Fragestellungen gegeben, die über den Rahmen dieser Arbeit hinausgehen. Im Anhang befinden sich Überblicke über die Test- und Schätzverfahren der beiden anderen verwendeten nichtparametrischen Verfahren und einige Beweise. Außerdem sind dort die Auswertungen der Beispiele und die Simulationstabellen aller anderen Verfahren bereitgestellt.

Kapitel 2

Skalenniveau und Versuchsplan

2.1 Ordinale Skalen

Für die statistische Auswertung von Daten spielt die Skala, auf der diese gemessen werden, eine wichtige Rolle. Je besser das statistische Modell an diese Skala angepasst wird, desto aussagekräftiger und machtvoller sind die statistischen Verfahren, welche auf diesem Modell basieren. Da in dieser Arbeit auf Verfahren eingegangen wird, welche die ordinale Struktur der Daten berücksichtigen, werden wir im Folgenden die Ordinalskala in Beziehung zu drei anderen geläufigen Skalen, der Nominal-, Intervall- und Verhältnisskala, setzen. Der Charakter dieser Skalen wird darüber definiert, welche Transformationen man von ihnen durchführen kann, ohne dass wichtige Informationen verloren gehen. Es gibt verschiedene Arbeiten in der Literatur, die sich eingehend mit der Unterscheidung von Skalen befassen. Hier wären zum Beispiel Stevens (1951); Krantz et al. (1971); Suppes et al. (1989); Luce et al. (1990) oder Cliff (1993) zu nennen.

Zunächst wird auf die Verhältnisskala eingegangen, für die das klassische Beispiel die Temperatur ist. Diese wird in der Physik in Kelvin[1] K gemessen. Da die Temperatur die ungeordnete mikroskopische Bewegung der Teilchen, das heißt der Atome, Moleküle oder Ähnlichem, angibt, existiert ein ausgezeichneter Nullpunkt. Dies ist der Punkt, an dem die Teilchen stillstehen. Der dritte Hauptsatz der Wärmelehre besagt, dass dieser zwar nie zu erreichen, aber theoretisch berechenbar ist (Gerthsen und Vogel, 1993, Seite 238).

Die Einheiten der Skala entsprechen 1/100 des Abstandes zwischen dem Gefrier- und dem Siedepunkt des Wassers bei einem Druck von 1,013 bar. Dieser Druck entspricht dem Druck auf Meereshöhe. Damit liegt der Gefrierpunkt des Wassers bei 273,2 K, sein Siedepunkt bei 373,2 K. Nun ist die Wahl, die Differenz zwischen diesen beiden Punkten in

[1] nach W. T. Kelvin (1824-1907)

100 Teile einzuteilen beziehungsweise die Wahl gerade dieser beiden Punkte, willkürlich. Der Nullpunkt erscheint hingegen von der Natur gegeben zu sein. Da ein Maß für ungeordnete Bewegung nicht negativ sein kann, ist es sogar sinnlos, eine negative Temperatur in Kelvin anzugeben. Transformationen f, die diese Struktur der Skala berücksichtigen, sind von der Form $f(x) = ax$ mit einer positiven Konstante a, das heißt, sie sind streng monoton wachsend und linear. Will man zum Beispiel geringe Temperaturen messen, so ist es sinnvoller, die Differenz zwischen dem Gefrierpunkt und dem Siedepunkt des Wassers in 100000 Teile einzuteilen. Dann erhält man die Einheit Millikelvin mK. Der Faktor a der Transformation f wäre in diesem Fall $a = 1000$.

Intervallskalen entstehen häufig durch die Betrachtung von Differenzen auf Verhältnisskalen. So wird die Differenz zwischen zwei in Kelvin gemessenen Temperaturen auf einer Intervallskala gemessen. Häufig interessiert zum Beispiel die Differenz zwischen dem Gefrierpunkt des Wassers und der Lufttemperatur. Werden beide Größen in Kelvin gemessen, so wird die Differenz mit Celsius[2] C bezeichnet. Der Nullpunkt dieser Skala ist mithin der Gefrierpunkt des Wassers. Eine andere gebräuchliche Skala für Temperaturunterschiede ist die Fahrenheitskala[3] F. Ein Temperaturunterschied c in Celsius entspricht einem Temperaturunterschied $f = 9c/5 + 32$ in Fahrenheit. Dies ist ein Beispiel für die streng monoton wachsenden affinen Transformationen $f(x) = ax + b,\ a > 0$, die bei Intervallskalen zugelassen sind.

Nun betrachten wir die ordinalen Daten. Eine Weise, auf der diese Skalen, die auch häufig Skalen mit geordneten Kategorien genannt werden, entstehen, ist die Zusammenfassung von Intervallskalen in Klassen. Solche Skalen werden von Anderson (1984) als *grouped continuous* bezeichnet. Wo nach dem Augenschein gemessen wird, entstehen ebenfalls häufig ordinale Skalen. Diese heißen bei Anderson (1984) *assessed ordered*. So werden schulische Leistungen mit sehr gut, gut, befriedigend, ausreichend, mangelhaft oder ungenügend bewertet. Diese Kategorien werden auch häufig mit den natürlichen Zahlen von eins bis sechs bezeichnet. Eine solche Umbenennung ist bereits eine Transformation. Dass diese Bezeichnung, beziehungsweise Transformation, willkürlich ist, erkennt man daran, dass in anderen Ländern die ersten Buchstaben des Alphabetes verwendet werden. Der Vorteil der Transformation auf die Zahlen von eins bis sechs ist, dass man mit ihnen rechnen kann. Ist die Differenz $N_1 - N_2$ zwischen zwei Schulnoten N_1 und N_2 positiv beziehungsweise negativ, so ist die erste Schulnote N_1 schlechter beziehungsweise besser als die zweite N_2. Ist die Differenz hingegen gleich null, so sind beide Noten gleich gut. Der Nachteil der Zahlen ist, dass sie eine Genauigkeit vortäuschen, die nicht vorhanden ist. So ist die Differenz zwischen einer Vier und einer Drei die gleiche wie zwischen einer

[2] nach Anders Celsius (1701- 1744)

[3] nach Daniel Gabriel Fahrenheit (1686-1736)

Zwei und einer Eins, aber einmal wird der Abstand von ausreichend zu befriedigend und einmal von gut zu sehr gut beschrieben. Diese Abstände sind jedoch nicht miteinander vergleichbar. Wir wissen nur, dass beide Differenzen positiv sind und also die ersten Noten schlechter als die zweiten sind; um wie viel schlechter können wir jedoch nicht bestimmen.

Dieser Sachverhalt, dass die Differenzen zwischen zwei Kategorien der ordinalen Skala nur in ihrem Vorzeichen, nicht jedoch von ihrem Betrag her messbar sind, spiegelt sich in den Transformationen wider, die bei ordinalen Skalen zugelassen sind. Dies sind alle Transformationen f, die streng monoton wachsend sind. So folgt für zwei Beobachtungen x_1 und x_2 aus $x_1 > x_2$ nur $f(x_1) > f(x_2)$ und nicht $f(x_1 - x_2) = f(x_1) - f(x_2)$. Dies ist ein entscheidender Unterschied zu den Verhältnisskalen, bei denen dies gilt. Ein weiterer wichtiger Unterschied ist, dass bei einer ordinalen Skala nur endlich viele Werte auf der Skala, nämlich die Kategorien, vorhanden sind. Bei der Verhältnis- beziehungsweise Intervallskala gibt es hingegen grundsätzlich unendlich viele Werte. Zwischen zwei Werten liegen immer weitere Werte der Skala, weshalb die Verhältnis- und die Intervallskala auch als stetige Skalen bezeichnet werden. Es gibt zwar auch stetige ordinale Skalen; diese werden hier jedoch ausgeschlossen.

Zuletzt werden nun die nominalen Skalen charakterisiert. Diese werden verwendet, wenn kategoriale Beobachtungen vorliegen, die nicht in einer Reihenfolge angeordnet werden können. Dies ist zum Beispiel bei der Farbe der Augen der Fall. Teilt man die Augenfarben in die Kategorien blau, grün, grau und braun ein, so unterliegt diese Skala der Augenfarben keiner natürlichen Reihenfolge. Umso mehr macht es keinen Sinn, Abstände zwischen den Kategorien zu bilden. Die Transformationen, die bei einer nominalen Skala erlaubt sind, spiegeln diese Strukturen wider. Es sind alle bijektiven Transformationen erlaubt. Diese entsprechen den Umbenennungen der Kategorien.

Die vier genannten Typen von Skalen kann man auf natürliche Weise anordnen, indem man die Transformationen betrachtet, die für die Skalen zugelassen sind. Alle wachsenden linearen Transformationen sind auch wachsende affine Transformationen. Des Weiteren sind alle affinen Transformationen streng monoton wachsend. Und schließlich sind alle streng monoton wachsenden Transformationen auch bijektiv. Die Mengen der erlaubten Transformationen sind also echt ineinander verschachtelt, das heißt, die Menge der linearen Transformationen ist eine echte Teilmenge der affinen Transformationen und so weiter. Man kann sagen, dass die vier verschiedenen Typen von Skalen auf einer ordinalen Skala liegen, die durch die möglichen Transformationen vorgegeben ist.

2.2 Versuchsplan

Die vorliegende Arbeit befasst sich nicht nur mit einer bestimmten Skala, der ordinalen Skala, sondern auch mit einem bestimmten Versuchsplan. Bei diesem Versuchsplan werden an zwei Gruppen von Versuchseinheiten V_{ik} Beobachtungen angestellt. Die beiden Gruppen unterscheiden sich darin, dass die erste eine andere *Behandlung* erfährt als die zweite. Die Anzahlen der Versuchseinheiten in den beiden Gruppen werden mit n_1 und n_2 und die Gesamtzahl mit $n = n_1 + n_2$ bezeichnet. Der Index $i = 1, 2$ bezeichnet also die Gruppe und der Index $k = 1, \ldots, n_i$ beschreibt, um welche Versuchseinheit es sich in dieser Gruppe handelt.

Außerdem werden an jeder Versuchseinheit V_{ik} zwei Beobachtungen durchgeführt. Diese Beobachtungen sind Werte der Zufallsvariablen X_{ikj}, $j = 1, 2$. Des Weiteren nehmen wir an, dass die unterschiedliche Behandlung der Versuchseinheiten sich nur auf genau eine dieser beiden Beobachtungen auswirkt. Dieses Phänomen tritt häufig dann auf, wenn die beiden Beobachtungen an einer Versuchseinheit zeitlich getrennt liegen, und die Behandlung der Versuchseinheit zwischen diesen beiden Beobachtungszeitpunkten liegt. Sie kann sich demnach nur auf die zweite Beobachtung auswirken. Mit dieser praktischen Sichtweise wird im Folgenden gearbeitet, weshalb die Beobachtungen zum ersten Zeitpunkt auch abkürzend als Baseline-Werte und die Beobachtungen zum zweiten Zeitpunkt als Follow-Up-Werte bezeichnet werden.

Tabelle 2.1: Der Versuchsplan mit der Anordnung der Versuchseinheiten V_{ik} und der Zufallsvariablen X_{ikj}

		Zeitpunkte			
		$j = 1$ (vorher)		$j = 2$ (nachher)	
		Versuchseinheit	Zufallsvariable	Versuchseinheit	Zufallsvariable
Gruppe	$i = 1$	V_{11}	X_{111}	V_{11}	X_{112}
		$\vdots$	$\vdots$	$\vdots$	$\vdots$
		V_{1n_1}	X_{1n_11}	V_{1n_1}	X_{1n_12}
	$i = 2$	V_{21}	X_{211}	V_{21}	X_{212}
		$\vdots$	$\vdots$	$\vdots$	$\vdots$
		V_{2n_2}	X_{2n_21}	V_{2n_2}	X_{2n_22}

Es sei jedoch bemerkt, dass solche paarigen Beobachtungen auch auf anderem Wege entstehen können. Wenn Beobachtungen an unterschiedlichen Stellen der Versuchseinhei-

ten vorgenommen werden, und nur eine der beiden Stellen behandelt wurde, so ergibt sich der gleiche Versuchsplan. In diesem Fall entspricht die Beobachtung an der behandelten Stelle der Beobachtung zum zweiten Zeitpunkt, und die Beobachtung an der unbehandelten Stelle entspricht der ersten Beobachtung. Die Einschränkung der Sichtweise auf wiederholte Beobachtungen erweist sich daher als unwesentlich.

Für jede Versuchseinheit erhält man damit einen Zufallsvektor $\boldsymbol{X}_{ik} = (X_{ik1}, X_{ik2})^t$, welcher nach F_i verteilt sei. Diese Zufallsvektoren werden als unabhängig betrachtet. Des Weiteren gibt die Zufallsvariable $X_{ikj} \sim F_{ij}$ für $j = 1$ die Beobachtung zum ersten Zeitpunkt und für $j = 2$ die Beobachtung zum zweiten Zeitpunkt an. Es werden also $N = 2n$ Zufallsvariablen beobachtet. Die Anordnung dieser Zufallsvariablen und der Versuchseinheiten wird durch die Tabelle 2.1 verdeutlicht.

Kapitel 3

Vorstellung der Beispiele

3.1 Neurologische Beeinträchtigung bei MS

Zur Messung der neurologischen Beeinträchtigung von Patienten werden häufig ordinale Skalen herangezogen (Bajorski und Petkau, 1999). Von Kurtzke (1961) wurde eine solche Skala, die *Disability Status Scale (DSS)*, entwickelt. Auf ihr kann der Grad der Schwächung eines an Multipler Sklerose (MS) leidenden Patienten auf einer Skala mit 11 Kategorien von 0 (keine Schwächung) bis 10 (Tod) gemessen werden. 1983 fügte Kurtzke weitere Stufen hinzu. Diese aus 21 Kategorien bestehende Skala wird *Expanded Disability Status Scale (EDSS)* genannt. Jede höhere Kategorie der Skala entspricht einem schlechteren Zustand des Patienten. Je niedriger ein Patient eingestuft wird, desto geringer ist seine neurologische Beeinträchtigung. Die einzelnen Zahlen, welche die verschiedenen Kategorien kennzeichnen, haben jedoch keine weitere Bedeutung außer der Beschreibung der ansteigenden Ordnung der Kategorien. Nach Anderson (1984) handelt es sich hier also um eine assessed ordered Skala.

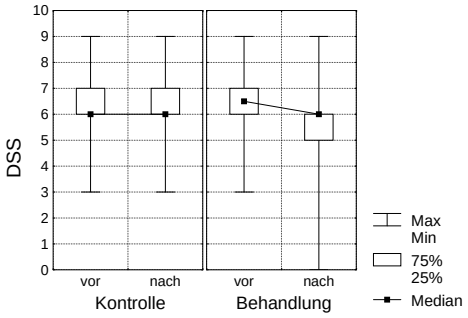

Abbildung 3.1: Boxplots der beiden Gruppen zu den beiden Zeitpunkten für die MS-Studie

Die Zielvariable bei klinischen Versuchen basiert häufig auf den Skalen DSS oder EDSS. So berichten Khatri et al. (1985) von einem doppelt blinden, kontrollierten klinischen

Behandlung						Kontrolle					
Nr.	vor	nach	Nr.	vor	nach	Nr.	vor	nach	Nr.	vor	nach
1	3	0	16	7	6	27	3	5	42	6	6
2	5	3	17	7	6	28	4	3	43	6	9
3	5	3	18	7	6	29	4	6	44	7	5
4	5	5	19	7	7	30	5	3	45	7	6
5	6	2	20	7	7	31	5	3	46	7	7
6	6	3	21	8	6	32	5	4	47	7	7
7	6	6	22	8	6	33	5	4	48	7	7
8	6	6	23	8	8	34	6	6	49	7	7
9	6	6	24	9	6	35	6	6	50	7	7
10	6	6	25	9	6	36	6	6	51	8	8
11	6	6	26	9	9	37	6	6	52	8	8
12	6	6				38	6	6	53	8	8
13	6	7				39	6	6	54	9	7
14	7	5				40	6	6	55	9	9
15	7	6				41	6	6			

Tabelle 3.1: Daten von Khatri et al. (1985) zur neurologischen Beeinträchtigung bei MS

Versuch von Plasmaaustausch (Plasmapheresis) bei Patienten mit chronischer progressiver MS. Die Patienten wurden randomisiert zwei Gruppen zugeteilt. Für 20 Wochen erhielt die eine Gruppe Plasmapheresis (Behandlung), während die andere Gruppe *sham* Plasmapheresis (Kontrolle) erhielt. Der Zustand der Patienten wurde vor Beginn der Behandlung und danach wöchentlich gemessen. Für die hier betrachtete Analyse werden nur die Daten vor und nach der 20 wöchigen Behandlung verwendet. Diese Werte der 55 Patienten sind in der Tabelle 3.1 zusammengefasst in der Abbildung 3.1 mit Hilfe von Boxplots illustriert.

3.2 Behandlung von Schlaflosigkeit

Das zweite Beispiel wurde Francom et al. (1989) beziehungsweise Agresti (1990) entnommen. Es ist ebenfalls ein kontrollierter klinischer Versuch, der doppelt verblindet durchgeführt wurde. Die untersuchten Patienten litten an Schlaflosigkeit und wurden zwei Versuchsgruppen randomisiert zugeteilt. Die erste Gruppe erhielt ein hypnotisches Medikament, die zweite ein Placebo.

Zur Messung der Einschlafzeit wurde die Frage „Wie schnell schlafen Sie ein, nachdem Sie zu Bett gegangen sind?“ gestellt. Diese Einschlafzeit wurde in die vier Kategorien weniger als 20 Minuten (1), zwischen 20 und 30 Minuten (2), zwischen 30 und 60 Minuten (3) und mehr als 60 Minuten (4) eingeteilt. Hier wurde eine stetige Skala in Kategorien eingeteilt, womit es sich nach Anderson (1984) um eine *grouped continuous* Skala handelt. Diese Kategorisierung geschieht, da eine genauere Messung nicht erzielt werden kann; die Patienten können zum Zeitpunkt des Einschlafens nicht die Zeit messen.

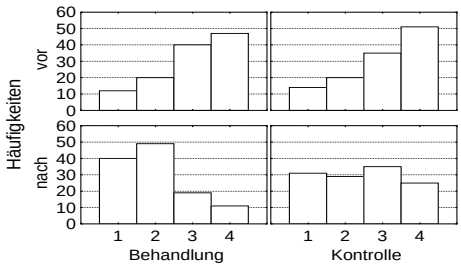

Abbildung 3.2: Histogramme der Insomnia-Studie

	vor	nach			
Behandlung		< 20	$20-30$	$30-60$	> 60
Verum	< 20	7	4	1	0
	$20-30$	11	5	2	2
	$30-60$	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	$20-30$	14	5	1	0
	$30-60$	6	9	18	2
	> 60	4	11	14	22

Tabelle 3.2: Daten von Francom et al. (1989) zur Insomnia-Studie

Vor und nach der zweiwöchigen Behandlung wurden die Patienten befragt. Die Ergebnisse der Studie sind in der Tabelle 3.2 dargestellt und die Histogramme 3.2 verdeutlichen die Unterschiede zwischen den Gruppen zu den beiden Zeitpunkten.

Kapitel 4

Modelle, Effekte und Hypothesen

4.1 Nichtparametrische Kovarianzanalyse

Jedes Verfahren zur Auswertung von Daten, wie sie gerade beispielhaft vorgestellt wurden, hat eine eigene Sichtweise. Das kovarianzanalytische Modell besitzt eine asymmetrische Betrachtungsweise der Komponenten der Zufallsvektoren $\boldsymbol{X}_{ik}$. Während die Follow-Up-Werte X_{ik2} als Zielvariablen betrachtet werden, sieht man die Baseline-Werte X_{ik1} als Kovariablen an. Der Schwerpunkt der Betrachtung liegt damit auf den Werten nach der Behandlung. Die Messungen vor der Behandlung dienen nur zur Erhöhung der Güte der verwendeten statistischen Verfahren und zur Adjustierung bezüglich eventueller Ungleichgewichte in den Gruppen.

An die Verteilungen F_{11} und F_{21} der beiden Gruppen zum ersten Zeitpunkt wird häufig eine wesentliche Voraussetzung gestellt:

$$F_{11} = F_{21}. \tag{4.1}$$

Diese Gleichheit der Verteilungen der Baseline-Werte kann angenommen werden, falls eine randomisierte Studie vorliegt. In diesem Fall stellen die beiden Verteilungen der Baseline-Werte die Verteilung der Auswahlpopulation dar, welche möglichst gut die Verteilung der Zielpopulation wiedergeben sollte. Im Folgenden werden wir von einer randomisierten Studie oder Versuchsanlage sprechen, falls die Voraussetzung (4.1) erfüllt ist. In der Kovarianzanalyse werden die Kovariablen auch als homogen bezeichnet, falls deren Verteilungen in allen Faktorstufenkombinationen gleich sind und damit (4.1) gilt.

In vielen Fällen ist eine Randomisierung jedoch nicht möglich. Dann kann die Annahme (4.1) nicht gestellt werden und wir sprechen von einer nicht randomisierten Studie oder Versuchsanlage. In der Kovarianzanalyse werden die Kovariablen in diesem Fall als

heterogen bezeichnet. Bei nicht randomisierten Studien muss ein Effekt diese Ungleichheit der Verteilungen der Baseline-Werte berücksichtigen, damit ein Unterschied zwischen den Follow-Up-Werten nicht irrtümlich als Behandlungseffekt interpretiert wird, obwohl er nur durch die Ungleichheit der Baseline-Werte verursacht wurde. In dieser Arbeit werden wir sowohl den bekannten Fall für homogene Kovariablen besprechen, als auch einen neuen Vorschlag für die nichtparametrische Analyse von heterogenen Kovariablen geben.

Um die Verteilungen der Teststatistiken der nichtparametrischen Kovarianzanalyse herleiten zu können, müssen Annahmen an die Verteilungsfunktionen F_{ij} der X_{ikj} gestellt werden. Die erste Annahme ist, dass keine Einpunktverteilungen vorliegen:

$$\mathrm{P}\left(X_{ikj} \neq X_{ik'j}\right) \quad > \quad 0 \quad \text{für} \quad i,j = 1,2 \quad \text{und} \quad k \neq k'. \tag{4.2}$$

Außerdem sollen sich die Träger der Verteilungen für die Ziel- und die Kovariablen jeweils überlappen:

$$\mathrm{P}\left(X_{i1j} \leq X_{i'1j}\right) \quad > \quad 0 \quad \text{für} \quad i,i',j = 1,2. \tag{4.3}$$

Diese beiden Annahmen sind sehr schwach, da sie die Menge der möglichen Verteilungsfunktionen kaum einschränken. Sie sind in der Regel erfüllt. Ebenfalls schwache Annahmen müssen an die Stichprobenumfänge n_i gestellt werden. Da bei allen hier betrachteten Verfahren asymptotische Aussagen bezüglich der Verteilungen getroffen werden, ist die folgende Voraussetzung natürlich:

$$\min_{i=1,2} n_i \quad \rightarrow \quad \infty. \tag{4.4}$$

Des Weiteren sollen beide Stichprobenumfänge in etwa gleicher Art gegen Unendlich streben:

$$n/n_i \quad < \quad n_0 \quad < \quad \infty, \quad i = 1,2. \tag{4.5}$$

Zur Beschreibung des Unterschiedes zwischen den beiden Gruppen kann der relative Effekt

$$p_j \quad = \quad \int F_{1j} dF_{2j}, \quad j = 1,2,$$

verwendet werden. Dieser Effekt ist bekannt vom Wilcoxon-Mann-Whitney Test (Wilcoxon, 1947; Mann und Whitney, 1947). Er wird in der Kovarianzanalyse jeweils innerhalb der Ziel- beziehungsweise der Kovariablen definiert. Dies ist natürlich, da die Ziel- und die

Kovariablen im Allgemeinen nicht vergleichbar sein müssen. Des Weiteren ist der Wertebereich des relativen Effektes aufgrund der Annahmen bezüglich der Marginalverteilungen das offene Einheitsintervall. Die folgende Schreibweise des relativen Effektes mit Hilfe von Wahrscheinlichkeiten für unabhängige Zufallsvariablen ist hilfreich für seine Interpretation:

$$p_j = \mathrm{P}\left(X_{11j} < X_{21j}\right) + \frac{1}{2}\mathrm{P}\left(X_{11j} = X_{21j}\right), \quad j = 1, 2. \tag{4.6}$$

Ist p_j größer als $1/2$, so tendieren die Ziel- beziehungsweise Kovariablen der zweiten Gruppe zu höheren Werten als die Ziel- beziehungsweise Kovariablen der ersten Gruppe. In diesem Fall sprechen wir davon, dass F_{2j} tendenziell größer als F_{1j} ist. Bei $p_j = 1/2$ heißen F_{1j} und F_{2j} tendenziell gleich. Die gleiche Sprechweise wird auch für die Zufallsvariablen X_{ikj} verwendet.

Die Verteilungen F_{11} und F_{21} der Baseline-Werte sind immer dann tendenziell gleich, wenn randomisierte Studien betrachtet werden. Denn aus der Annahme (4.1) folgt $p_1 = 1/2$. In diesem Fall kann der Einfluß der Behandlungen mit Hilfe des Effektes p_2 beschrieben werden. Die Hypothese, dass die Behandlung keinen Einluß hat, kann dann entweder mit Hilfe der Verteilungsfunktionen

$$H_0: \quad F_{12} = F_{22} \tag{4.7}$$

oder mit Hilfe des relativen Effektes

$$H_0: \quad p_2 = 1/2 \tag{4.8}$$

gestellt werden. Hierbei ist (4.7) stärker als (4.8), denn die Hypothese über die Verteilungsfunktionen impliziert die Hypothese über den relativen Effekt. Die Umkehrung gilt hingegen nicht.

Bei nicht randomisierten Studien muß p_2 bezüglich der heterogenen Baseline-Werte adjustiert werden, weswegen der Effekt

$$p^*(\gamma_{nkov}) = p_2 - \gamma_{nkov}(p_1 - 1/2) \tag{4.9}$$

verwendet wird (Siemer, 1999). Die Bestimmung von γ_{nkov} hängt von den Schätzern der relativen Effekte ab und wird daher erst später betrachtet (Domhof, 2001). Es sei hier nur erwähnt, dass γ_{nkov} je nach Korrelation das Vorzeichen ändert. So wird sicher gestellt, dass die Adjustierung immer korrekt erfolgt. Sind die Schätzer beispielsweise positiv korreliert, was in der Regel der Fall ist, und ist F_{21} tendenziell größer als F_{11}, so ist eine Adjustierung

von p_2 nach unten erforderlich. Diese erfolgt auch, da γ_{nkov} bei positiver Korrelation ebenfalls positiv ist. Damit ist der adjustierte relative Effekt $p^*(\gamma_{nkov})$ in diesem Fall kleiner als p_2.

Der Effekt $p^*(\gamma_{nkov})$ ist eine Erweiterung des relativen Effektes p_2 für nicht randomisierte Versuchsanlagen, denn im Fall von randomisierten Studien sind die beiden Effekte gleich. In diesem Fall gilt $p_1 = 1/2$ und es folgt $p^*(\gamma_{nkov}) = p_2$.

Die Nullhypothese wird bei nicht randomisierten Studien mit Hilfe des adjustierten relativen Effektes $p^*(\gamma_{nkov})$ gestellt:

$$H_0: \quad p^*(\gamma_{nkov}) \quad = \quad 1/2. \tag{4.10}$$

Die Betrachtung der Hypothese (4.7) ist bei nicht randomisierten Versuchsanlagen unsinnig, da die Verzerrung durch die Ungleichheit der Verteilungen der Baseline-Werte nicht berücksichtigt wird.

Die Hypothese (4.10) basiert auf der korrekten Adjustierung des Effektes durch die Gleichung (4.9). Diese Modellgleichung erkärt, wie man die Verzerrung durch die heterogenen Kovariablen beheben kann. Andere Verfahren für nichtrandomisierte Studien haben ähnliche Modellgleichungen, um die Effekte zu adjustieren. Es ergeben sich einige grundsätzliche Probleme bei der Interpretation von solchen adjustierten Effekten. Diese werden im Folgenden anhand von konstruierten Beispielen diskutiert. Dazu wird exemplarisch angenommen, dass alle Beobachtungen, auch *Responses* genannt, auf einer ordinalen Skala mit den vier Punkten $1, \ldots, 4$ liegen. Die Kategorie 1 bezeichnet dabei den besten und die Kategorie 4 den schlechtesten Wert.

Im ersten Beispiel, das in der Abbildung 4.1 illustriert ist, liegen die Baseline-Werte der ersten Gruppe in der Kategorie 2 und die der zweiten Gruppe in der Kategorie 3. Während sich die Werte der ersten Gruppe nicht ändern, verschlechtert sich die zweite Gruppe auf den Wert 4. Wird nun die Behandlung der ersten Gruppe gegenüber der Behandlung der zweiten Gruppe bevorzugt, weil sich in der ersten Gruppe die Werte nicht verschlechtert haben, so begeht man unter Umständen einen Fehler, da man nicht weiß, was mit den Versuchseinheiten der ersten Gruppe passiert wäre, wenn sie die gleichen Ausgangsvoraussetzungen gehabt hätten wie die zweite Gruppe. Es ist nicht klar, ob die

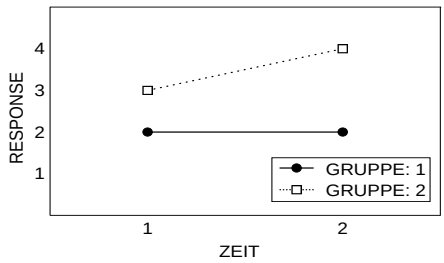

Abbildung 4.1: Die erste Gruppe bleibt konstant und die zweite Gruppe verschlechtert sich.

Behandlung der ersten Gruppe auch in diesem Fall erfolgreicher gewesen wäre als die Behandlung der zweiten Gruppe. Vielleicht sind Versuchseinheiten mit Kategorie 3 schon so stark geschädigt, dass sie in jedem Fall nach einiger Zeit in die schlechteste Kategorie fallen.

Für das zweite Beispiel, welches in Abbildung 4.2 dargestellt ist, wird angenommen, dass sich die erste Gruppe genauso verhält wie im vorhergehenden Beispiel. Die zweite Gruppe verbesserte sich jedoch von der Kategorie 4 auf die Kategorie 3. Hier könnte es sein, dass es einfacher ist, von der Kategorie 4 auf die 3 zu gelangen als von der 2 auf die 1, weshalb die Behandlung der zweiten Gruppe nicht unbedingt besser sein muss, als die Behandlung der ersten Gruppe.

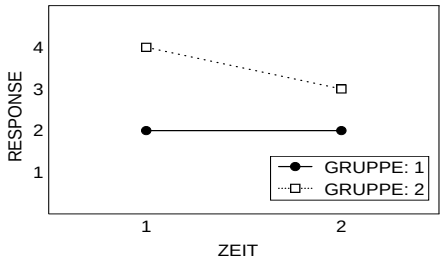

Abbildung 4.2: Die erste Gruppe bleibt konstant und die zweite Gruppe verbessert sich.

Die gleiche Argumentation kann auch dann geführt werden, wenn in beiden Gruppen zum zweiten Zeitpunkt die Kategorie 2 beobachtet wird und die Baseline-Werte die Gleichen wie im vorhergehenden Beispiel sind. Allerdings wird man in diesem Fall, wenn nicht andere Aspekte wie zum Beispiel Nebenwirkungen eine Rolle spielen, die Behandlung der zweiten Gruppe vorziehen. Man kann zwar nicht sagen, dass sie der ersten Gruppe überlegen ist, aber sie ist wohl mindestens genau so gut. Es ist nämlich nicht plausibel, dass sich die Werte der ersten Gruppe von einem schlechteren Ausgangswert, zum Beispiel der 3, auf einen Wert verbessern, der sogar noch besser ist als der ursprüngliche Wert, also die 2.

Ein weiterer Aspekt eröffnet sich, wenn die Kategorie 1 aus der Skala entfernt wird. Dann ist eine weitere Verbesserung nicht möglich. Wenn in diesem Fall die beiden Gruppen trotz unterschiedlicher Ausgangswerte zum zweiten Zeitpunkt die beste Kategorie erreichen, kann keine eindeutige Priorität für eine der beiden Gruppen festgestellt werden. Trotzdem wird man sich, der obigen Argumentation folgend, für die Behandlung derjenigen Gruppe entscheiden, welche die schlechteren Voraussetzungen hatte.

Im dritten Beispiel, das die Abbildung 4.3 verdeutlicht, kann eine Priorisierung durchgeführt werden. Dazu wird erneut angenommen, dass die Beobachtungen in der ersten Gruppe zu beiden Zeitpunkten den Wert 2 annehmen. Die zweite Gruppe verbessert sich hingegen von der Kategorie 3 auf die Kategorie 1. Da es in der Regel sehr unwahrscheinlich ist, dass schlechtere Ausgangswerte bessere Endwerte erzeugen, kann in diesem Fall die zweite Gruppe der ersten vorgezogen werden.

Im Allgemeinen kann man also nur dann eine eindeutige Entscheidung zu Gunsten einer Behandlung treffen, wenn sich der Effekt vom ersten zum zweiten Zeitpunkt umkehrt. Das bedeutet, dass die eine Gruppe zum ersten Zeitpunkt schlechtere und zum zweiten Zeitpunkt bessere Werte aufweist als die andere Gruppe. In der Kovarianzanalyse ist dies zum Beispiel der Fall bei $p_1 < 1/2$ und $p_2 > 1/2$.

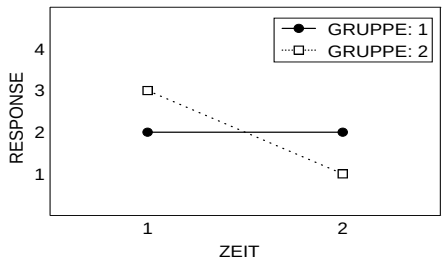

Abbildung 4.3: Die erste Gruppe ist zum ersten Zeitpunkt besser und zum zweiten Zeitpunkt schlechter als die zweite Gruppe.

Nachdem nun die nichtparametrische Kovarianzanalyse vorgestellt wurde, sollen verschiedene Verfahren aus der Literatur diskutiert und mit der nichtparametrischen Kovarianzanalyse verglichen werden. Dabei wird sich zeigen, dass die Annahmen des Modells sehr schwach sind. Ferner wird sich ergeben, dass die Effekte der nichtparametrischen Kovarianzanalyse anschaulich interpretiert werden können.

4.2 Naive Verfahren

4.2.1 Pearson-Verfahren

In diesem und dem nächsten Abschnitt werden zwei Ansätze, das Pearson- und das Cochran-Armitage-Verfahren, vorgestellt, die wir als naive Verfahren bezeichen, weil sie sehr einfach sind und nur einen bestimmten Teil der vorhandenen Information verwenden. Die naiven Verfahren betrachten bei jeder einzelnen Versuchseinheit nur, in welche Richtung sich die Werte an den jeweiligen Versuchseinheiten geändert haben. Man betrachtet also die Zufallsvariablen $Z_{ik} = 2(c(X_{ik1}, X_{ik2}) - 1/2)$ mit der Vergleichsfunktion

$$c(x,y) \quad = \quad \begin{cases} 0 & \text{falls} \quad x < y \\ 1/2 & \text{falls} \quad x = y \\ 1 & \text{falls} \quad x > y \end{cases} .$$

Mithin nehmen die Zufallsvariablen Z_{ik} die Werte -1, 0 oder 1 an, je nachdem in welche Richtung sich die Beobachtungen der Versuchseinheit k in der Gruppe i ändern. Damit interessieren also die Wahrscheinlichkeiten

$$\mathrm{P}(X_{ik1} < X_{ik2}) \quad = \quad q_{i1} \quad = \quad \mathrm{P}(Z_{ik} = -1),$$

$$\begin{aligned} \mathrm{P}(X_{ik1} = X_{ik2}) &= q_{i2} = \mathrm{P}(Z_{ik} = 0) \quad \text{und} \\ \mathrm{P}(X_{ik1} > X_{ik2}) &= q_{i3} = \mathrm{P}(Z_{ik} = 1). \end{aligned}$$

Im Gegensatz zur nichtparametrischen Kovarianzanalyse werden die beiden Zeitpunkte hier gewissermaßen symmetrisch betrachtet, da die Vorzeichen der Differenzen der Zufallsvariablen verwendet werden. An die Wahrscheinlichkeiten q_{is} wird die Voraussetzung

$$q_{is} > 0, \quad i = 1, 2, \; s = 1, 2, 3, \tag{4.11}$$

gestellt. Sie ist wichtig für die Herleitung der Verteilung der Teststatistik. Diese wird wie bei der nichtparametrischen Kovarianzanalyse für unendlich großen Stichprobenumfänge bestimmt. Es wird allerdings eine etwas strengere Forderung an das Verhältnis der Stichprobenumfänge gestellt. Für die naiven Verfahren soll

$$n \to \infty \quad \text{mit} \quad n_1/n \to \lambda \in (0; 1) \tag{4.12}$$

gelten.

Der Effekt des Pearson-Verfahrens betrachtet die Differenzen $q_{1s} - q_{2s}$. Verschwinden diese Differenzen alle, so sind die Verteilungen von Z_{1k} und Z_{2k} gleich. Damit sich bei der Summierung der Differenzen die Unterschiede nicht aufheben können, wird die Summe der Quadrate der Differenzen als Effekt verwendet:

$$p_{Pearson} = \sum_{s=1}^{3} (q_{1s} - q_{2s})^2 .$$

Der Effekt $p_{Pearson}$ wird als Pearson-Effekt bezeichnet, da auf ihm im Wesentlichen die Pearson-χ^2-Statistik für 2×3-Kontingenztafeln beruht. Die Hypothese des Pearson-Verfahrens ist

$$H_0: \quad p_{pearson} = 0. \tag{4.13}$$

Sie ist äquivalent dazu, dass die Verteilungen von Z_{1k} und Z_{2k} gleich sind.

Diese Hypothese hat einige Nachteile, welche auf die Eigenschaften des Pearson-Effektes zurückzuführen sind. Der Pearson-Effekt gibt nur an, ob sich die Verteilungen der Z_{ik} unterscheiden, aber nicht in welcher Weise. Da er nur nicht negative Werte annehmen kann, kann man aufgrund von $p_{Pearson}$ nicht feststellen, welche der beiden Gruppen die besseren Werte hat. Es ist somit auch nicht möglich, die Hypothese (4.13) in eine einseitige Hypothese umzuwandeln. Des Weiteren ist die verwendete Information von $p_{Pearson}$

Tabelle 4.1: Schwächen der naiven Effekte bei einem konstruierten Beispiel

Wahrscheinlichkeit	X_{1k1}	X_{1k2}	Z_{1k}	X_{1k2}	X_{2k2}	Z_{2k}
1/3	3	1	1	3	2	1
1/3	3	3	0	3	3	0
1/3	3	4	-1	3	4	-1

so grob, dass man aus der Gültigkeit von $H_0 : \ p_{Pearson} = 0$ keine Rückschlüsse auf die Verteilungen der ursprünglichen Zufallsvariablen X_{ik} machen kann. Weder über die gemeinsamen Verteilungen F_1 und F_2 noch über die Marginalverteilungen zum zweiten Zeitpunkt F_{12} und F_{22} kann etwas Wesentliches aufgrund von $p_{Pearson}$ ausgesagt werden.

Verdeutlichen lässt sich dies anhand eines konstruierten Beispiels, dessen Daten in der Tabelle 4.1 angegeben sind. Es wird eine ordinale Skala mit den Kategorien 1 bis 4 verwendet, wobei die Kategorie 1 die beste Kategorie sein soll. Es sind für jede Gruppe drei Realisierungen der Zufallsvektoren möglich. Diese sind gleich wahrscheinlich. Zum ersten Zeitpunkt haben alle Versuchseinheiten die Kategorie 3 als Wert. Zum zweiten Zeitpunkt ist die Wahrscheinlichkeit für die Verschlechterung auf die Kategorie 4 genauso wie die Wahrscheinlichkeit für keine Veränderung 1/3. Weiterhin ist die Wahrscheinlichkeit für eine Verbesserung in beiden Gruppen 1/3. Aber während die Patienten, welche sich verbessern, in der ersten Gruppe in die Kategorie 1 fallen, haben die Patienten mit einer Verbesserung in der zweiten Gruppe nur die Kategorie 2. Damit würde man ausgehend von den Wahrscheinlichkeiten für die Originalvektoren $(X_{ik1}, X_{ik2})^t$ die Behandlung der ersten Gruppe der Behandlung der zweiten Gruppe vorziehen. Der Pearson-Effekt $p_{Pearson}$ ist aber Null, da die Verteilungen der Z_{ik} identisch sind.

Ein weiterer Nachteil des Pearson-Effektes ist seine schlechte Interpretierbarkeit. Sie rührt von der Orientierung des Verfahrens an der Testentscheidung her. Es ist im Gegensatz zum relativen Effekt aus der nichtparametrischen Kovarianzanalyse nicht möglich, den Pearson-Effekt einfach mit Hilfe von Wahrscheinlichkeiten zu interpretieren.

Zuletzt sei noch ein struktureller Nachteil des naiven Ansatzes erwähnt. Er läßt sich schlecht auf mehrere Gruppen und Zeitpunkte erweitern. Selbst wenn man dies versuchen würde, so würde man den größten Vorteil des Ansatzes, nämlich die Einfachheit, verlieren.

4.2.2 Cochran-Armitage-Verfahren

Das Cochran-Armitage-Verfahren (Armitage, 1955) verwendet den gleichen Ansatz wie das Pearson-Verfahren. Es betrachtet ebenfalls die Zufallsvariablen Z_{ik}. Allerdings wird ein anderer Effekt verwendet:

$$p_{CA} \quad = \quad (q_{23} - q_{13}) - (q_{21} - q_{11}).$$

Der Cochran-Armitage-Effekt p_{CA} betrachtet nur die beiden Differenzen $q_{21} - q_{11}$ und $q_{23} - q_{13}$. Die dritte Differenz ergibt sich ohnehin aus diesen beiden. Der Effekt p_{CA} kann anschaulich als Differenz der Wahrscheinlichkeiten einer Verbesserung minus der Differenz der Wahrscheinlichkeiten einer Verschlechterung interpretiert werden. Die aus diesem Effekt folgende zweiseitige Hypothese ist

$$H_0: \quad p_{CA} \quad = \quad 0.$$

Sie kann im Gegensatz zur Hypothese (4.13) auch dann gelten, wenn die Verteilungen der Z_{ik} nicht identisch sind. Wird zum Beispiel durch eine neue Behandlungsmethode die Wahrscheinlichkeit für eine Verbesserung gegenüber einer alten Methode im gleichen Maß erhöht wie die Wahrscheinlichkeit für eine Verschlechterung, so gilt $H_0: \ p_{CA} = 0$. Dies wird in der Abbildung 4.4 deutlich, in welcher hypothetische Wahrscheinlichkeiten für zwei verschiedene Behandlungsmethoden aufgetragen sind. Für die Wahrscheinlichkeiten in dieser Abbildung verschwindet der Cochran-Armitage-Effekt im Gegensatz zum Pearson-Effekt, denn die Erhöhung der Wahrscheinlichkeit für eine Verbesserung wird durch eine gleich große Erhöhung der Wahrscheinlichkeit für eine Verschlechterung erkauft. Falls also der Cochran-Armitage-Effekt null ist, kann der Pearson-Effekt trotzdem positiv sein. Umgekehrt gilt hingegen

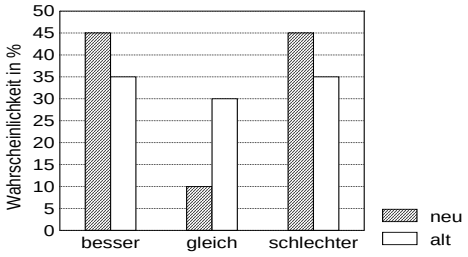

Abbildung 4.4: Hypothetische Wahrscheinlichkeiten bei einer neuen und einer alten Methode

$$p_{pearson} = 0 \quad \Rightarrow \quad p_{CA} = 0. \tag{4.14}$$

Der Pearson-Effekt betrachtet im Gegensatz zum Cochran-Armitage-Effekt, welcher nur die Erwartungswerte der Z_{ik} verwendet, auch die Varianzen der Z_{ik}.

Der Cochran-Armitage-Effekt hat wie der Pearson-Effekt strukturelle Schwächen. Zum einen ist die verwendete Information die gleiche, weshalb der Cochran-Armitage-Effekt die Daten ähnlich grob beschreibt wie der Pearson-Effekt. Zum anderen ist er genauso schlecht auf mehrere Gruppen oder Zeitpunkte erweiterbar.

Allerdings hat er zwei wesentliche Vorteile gegenüber dem Pearson-Effekt. Zunächst gibt er an, welche Gruppe die besseren Werte hat, womit im Gegensatz zum Pearson-Effekt auch einseitige Hypothesen möglich sind. So kann die Macht für die interessierenden Alternativen erhöht werden. Außerdem ist die Interpretation des Cochran-Armitage-Effektes anschaulicher als die des Pearson-Effektes.

Die Effekte der naiven Verfahren unterscheiden sich deutlich von den relativen Effekten der nichtparametrischen Kovarianzanalyse. Während die naiven Verfahren sich nur auf die individuellen Veränderungen der Versuchseinheiten beziehen, werden in der nichtparametrischen Kovarianzanalyse nur die Marginalverteilungen betrachtet. Die Prioritäten der Verfahren sind unterschiedlich. Die Effekte der naiven Verfahren betrachten die einzelne Versuchseinheit und bewerten, wie groß die Wahrscheinlichkeit ist, dass sich deren Werte ändern. Die relativen Effekte vergleichen hingegen die gesamten Gruppen.

Dass diese Betrachtungen sich wesentlich unterscheiden, lässt sich am folgenden extremen Beispiel veranschaulichen. Dazu wird zunächst nur eine Gruppe zu zwei Zeitpunkten betrachtet. Des Weiteren sei eine ordinale Skala mit den Kategorien von eins bis sechs gegeben, wobei die niedrigste die beste und die höchste die schlechteste sein soll. Nehmen wir weiter an, dass sich die gemeinsame Verteilung F_1 folgendermaßen darstellt:

$$\begin{aligned}
\mathrm{P}\left((X_{1k1}, X_{1k2})^t = (s, s+1)^t\right) &= \frac{1}{5} \quad \text{für} \quad s = 2, 3, 4, \\
\mathrm{P}\left((X_{1k1}, X_{1k2})^t = (s, s-4)^t\right) &= \frac{1}{5} \quad \text{für} \quad s = 5, 6 \quad \text{und} \\
\mathrm{P}\left((X_{1k1}, X_{1k2})^t = (s, s')^t\right) &= 0 \quad \text{sonst.}
\end{aligned}$$

Anschaulich ist dies links in der Abbildung 4.5 dargestellt. Betrachtet man nun die Marginalverteilungen, so ist eine Verbesserung festzustellen. Während zum ersten Zeitpunkt die Kategorien zwei bis sechs beobachtet wurden, ergeben sich zum zweiten Zeitpunkt die Kategorien eins bis fünf. Werden hingegen die einzelnen Individuen betrachtet, so werden im Mittel drei Verschlechterungen und nur zwei Verbesserungen erwartet. Mithin hat sich die Gruppe insgesamt verbessert, obwohl sich der größere Teil verschlechtert hat.

Dieser Unterschied zwischen den Effekten kann auch folgendermaßen veranschaulicht werden. Für die Marginalverteilungen betrachten wir die Wahrscheinlichkeit, dass eine

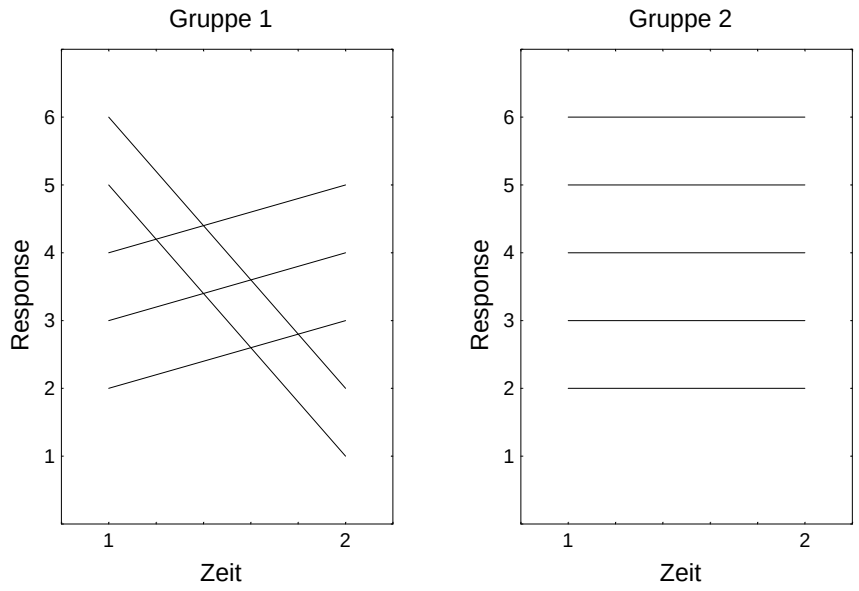

Abbildung 4.5: Beispiel von zwei gemeinsamen Verteilungen für den Vergleich der naiven Effekte und der Effekte der nichtparametrischen Kovarianzanalyse

zufällig zum ersten Zeitpunkt ausgewählte Versuchseinheit bessere Werte zeigt als eine zufällig zum zweiten Zeitpunkt gezogene Versuchseinheit. Dem gegenüber steht bei den naiven Verfahren die Wahrscheinlichkeit, dass sich eine zufällig ausgewählte Versuchseinheit verbessert.

Nun erweitern wir das Beispiel um die zweite Gruppe (Abbildung 4.5, rechte Seite). Die gemeinsame Verteilung F_2 soll durch die folgenden Wahrscheinlichkeiten gegeben sein:

$$\begin{aligned} \mathrm{P}\left((X_{1k1}, X_{1k2})^t = (s,s)^t\right) &= \frac{1}{5} \quad \text{für} \quad s = 2, \ldots, 6 \quad \text{und} \\ \mathrm{P}\left((X_{1k1}, X_{1k2})^t = (s,s')^t\right) &= 0 \quad \text{sonst.} \end{aligned}$$

Wenn sich der Wert jeder Versuchseinheit der zweiten Gruppe nicht ändert, so widersprechen sich die Effekte der beiden Verfahren. Während der Cochran-Armitage-Effekt die Behandlung der zweiten Gruppe der anderen Behandlung vorzieht, bewertet der relative Effekt p_2 die Behandlung der ersten Gruppe besser. Der relative Effekt p_1 ist hier $1/2$, da die Marginalverteilungen zum ersten Zeitpunkt identisch sind.

Es stellt sich nun die Frage, wodurch solche Probleme, die bei allen Verfahren auftreten, in der Praxis entstehen. An der fehlenden Randomisierung liegt es in diesem Fall

nicht, da die Probleme zum einen schon bei einer Gruppe auftreten und zum anderen die Marginalverteilungen zum ersten Zeitpunkt im vorgestellten Beispiel sogar gleich sind. Vielmehr kann die fehlende Homogenität der betrachteten Gruppe von Versuchseinheiten eine Rolle spielen. Wenn sich die Werte der meisten Versuchseinheiten kaum ändern und nur bei einer kleinen Untergruppe die Unterschiede sehr groß sind, so kann es daran liegen, dass sich diese Untergruppe in einem entscheidenden Merkmal von den anderen Versuchseinheiten unterscheidet. Eine Interpretation solcher Daten ohne die Berücksichtigung dieses Merkmals erscheint dann fragwürdig, weil der Einfluss dieses Merkmals auf den Versuchsausgang ein viel größerer ist als der Einfluss der zu untersuchenden Behandlung.

4.3 Nichtparametrische faktorielle Analyse

Der nichtparametrischen faktoriellen Analyse liegt, wie den naiven Verfahren, eine symmetrische Sichtweise der Komponenten der Zufallsvektoren $\boldsymbol{X}_{ik}$ zugrunde. Bei dieser Sichtweise, die wir als faktoriell bezeichnen, werden zwei feste Faktoren G (*Group*) und T (*Time*) und ein zufälliger Faktor Z, welcher die Blöcke beziehungsweise *Plots* darstellt, betrachtet. Während die beiden festen Faktoren vollständig gekreuzt sind, ist der zufällige Faktor unter dem Faktor G verschachtelt und mit dem Faktor T gekreuzt. Ein solches Versuchsdesign wird auch als *Split-Plot*-Plan (Kirk, 1982) bezeichnet.

Die Faktoren G und T haben jeweils zwei Stufen, welche mit g_1 und g_2 beziehungweise mit t_1 und t_2 bezeichnet werden. Der Faktor Z hat n Stufen. Bei den Stufen des Faktors G handelt es sich häufig um zwei verschiedene Behandlungen, die an den Versuchseinheiten, die die Stufen des Faktors Z darstellen, durchgeführt werden. Daher werden wir diese beiden Stufen Behandlungen nennen. Die Versuchseinheiten kann man folglich in zwei Gruppen mit den Anzahlen n_1 und n_2 an Versuchseinheiten einteilen, je nachdem ob die Versuchseinheiten der Stufe g_1 oder g_2 zugeordnet werden. Der Faktor G wird deshalb auch *Whole-Plot*-Faktor genannt, da seine Stufen auf die ganzen Blöcke wirken.

Während jeweils nur eine Behandlung an den Versuchseinheiten durchgeführt wird, das heißt die Versuchseinheiten den Gruppen zugeordnet sind, ist der Faktor Z mit dem Faktor T gekreuzt. Der Faktor T wird daher auch *Sub-Plot*-Faktor genannt. Jede Versuchseinheit wird unter beiden Stufen des Faktors T betrachtet. Dies geschieht häufig, wenn der Faktor T die Zeit ist und seine beiden Stufen zwei Zeitpunkte darstellen, zu denen die Beobachtungen an den Versuchseinheiten durchgeführt werden. Daher heißen die Stufen von T auch Zeitpunkte. Es kommen aber auch andere Faktoren als die Zeit vor.

Werden die Beobachtungen an zwei verschiedenen Stellen einer Versuchsperson, beispielsweise am linken und rechten Auge durchgeführt, so sind sie, wie die beiden Beobachtungen zu zwei Zeitpunkten, voneinander abhängig.

Die Faktoren G und T agieren häufig nicht unabhängig voneinander auf die Versuchseinheiten. So kann der Unterschied der Wirkungen der beiden Faktorstufen t_1 und t_2 davon abhängen, ob gleichzeitig die Faktorstufe g_1 oder g_2 betrachtet wird. Die zeitliche Entwicklung kann also von der Gruppe abhängen. Diese Abhängigkeit wird als Wechselwirkung GT oder W bezeichnet. Das primäre Ziel ist es, eine Wechselwirkung, falls sie vorhanden ist, nachzuweisen. Diese Fragestellung entspricht den beiden Fragestellungen aus den vorhergehenden Kapiteln.

Wie im Kapitel über die nichtparametrische Kovarianzanalyse müssen Annahmen an die Marginalverteilungen F_{ij} gestellt werden. Zum einen sollen, wie in (4.2) auf Seite 16 beschrieben, die Einpunktverteilungen ausgeschlossen werden. Zum anderen sollen sich die Träger der Marginalverteilungen überlappen. Für die nichtparametrische faktorielle Analyse reicht es aber nicht aus, dass dies nur für die Paare F_{1j} und F_{2j} vor und nach der Behandlung gilt, wie dies bei der nichtparametrischen Kovarianzanalyse der Fall ist. Es soll jeder Träger einer Marginalverteilung mit jedem anderen überlappen:

$$\mathrm{P}\left(X_{i1j} \leq X_{i'1j'}\right) \quad > \quad 0 \quad \text{für} \quad i, i', j, j' = 1, 2. \tag{4.15}$$

Damit unterscheiden sich die Annahmen der nichtparametrischen faktoriellen Analyse bezüglich der Marginalverteilungen etwas von den Annahmen der nichtparametrischen Kovarianzanalyse. Hingegen stimmen die Annahmen der nichtparametrischen faktoriellen Analyse, welche die Stichprobenumfänge betreffen, mit denen der nichtparametrischen Kovarianzanalyse überein.

Im Gegensatz zur nichtparametrischen Kovarianzanalyse werden im faktoriellen Ansatz nicht jeweils zwei sondern vier Marginalverteilungen miteinander verglichen. Es ergeben sich mithin drei Effekte: der Gruppeneffekt, der Zeiteffekt und die Wechselwirkung. Um diese Effekte zu beschreiben, schlagen Akritas und Arnold (1994) die folgende Zerlegung der Marginalverteilungen vor:

$$F_{ij}(x) \quad = \quad M(x) + G_i(x) + T_j(x) + W_{ij}(x), \quad i, j = 1, 2. \tag{4.16}$$

Um die Funktionen M, G_i, T_j und W_{ij} eindeutig zu bestimmen, müssen Reparametrisierungsbedingungen gestellt werden. Akritas und Arnold (1994) verwenden die Effektkodierung

$$\sum_{i=1}^{2} G_i^e \quad = \quad \sum_{j=1}^{2} T_j^e \quad = \quad \sum_{i=1}^{2} W_{ij}^e \quad = \quad \sum_{j=1}^{2} W_{ij}^e \quad = \quad 0.$$

Ebenso kann die Referenzkodierung gewählt werden. Dann sind die Reparametrisierungsbedingungen

$$G_1^r = T_1^r = W_{11}^r = W_{21}^r = W_{12}^r = 0.$$

Hier wurde die erste Gruppe zum ersten Zeitpunkt als Referenz gewählt. Spielt die Reparametrisierung eine entscheidende Rolle, so wird der entsprechende Superscript e oder r hinzugefügt. Ansonsten wird er vernachlässigt. Akritas et al. (1997) nennen die Funktionen G_i und T_j die nichtparametrischen Haupteffekte der Faktoren und W_{ij} deren nichtparametrische Wechselwirkung. Im Gegensatz zu Akritas und Arnold (1994) und Akritas et al. (1997) werden hier nur zwei Stufen pro Faktor betrachtet. Für die Zusammenhänge zwischen den Effekten der beiden Kodierungen erhält man in diesem Fall die folgenden Gleichungen für die beiden Haupteffekte

$$\begin{aligned}
\frac{1}{2}\sum_{j=1}^{2}(F_{2j} - F_{1j}) &= 2G_2^e = G_2^r + W_{22}^r, \\
F_{21} - F_{11} &= G_2^r = 2G_2^e - 2W_{22}^e, \\
\frac{1}{2}\sum_{i=1}^{2}(F_{i2} - F_{i1}) &= 2T_2^e = T_2^r + W_{22}^r, \\
F_{12} - F_{11} &= T_2^r = 2T_2^e - 2W_{22}^e
\end{aligned}$$

und für die Wechselwirkung

$$(F_{11} - F_{12}) - (F_{21} - F_{22}) = 4W_{22}^e = W_{22}^r.$$

Es ist wichtig zu erkennen, dass die beiden Parameter für die Wechselwirkung W_{22}^e und W_{22}^r sich nur durch eine Faktor 4 unterscheiden. Liegt damit eine Wechselwirkung für die eine Kodierung vor, so ist sie auch in der anderen Kodierung vorhanden. Man braucht daher bei einem Nachweis der Wechselwirkung die Kodierungen nicht zu unterscheiden. Unter der Hypothese

$$H_0: \quad F_{11} - F_{12} = F_{21} - F_{22} \tag{4.17}$$

verschwindet die Wechselwirkung für beide Kodierungen. Anders ist dies im Fall der beiden Haupteffekte. Liegt eine Wechselwirkung vor, so hängt die Interpretation der Effekte der Faktoren G und T von der Kodierung ab. Ansonsten unterscheiden sich die Parameter der beiden Kodierungen jeweils nur durch den Faktor 2. Diese Betrachtung gilt für alle additiven Zerlegungen dieser Art, die noch im weiteren Verlauf der Arbeit auftreten werden.

Diese auf den Marginalverteilungen beruhenden Effekte sind allerdings sehr testorientiert und wenig praxisnah. Um besser interpretierbare Effekte zu erhalten, werden Funktionale der Verteilungsfunktionen, die relativen Effekte, gebildet. Dafür wird zunächst die mittlere Verteilungsfunktion gebraucht. Diese lässt sich mit den Stichprobenumfängen gewichtet oder ungewichtet bilden. Die gewichtete mittlere Verteilungsfunktion sei

$$H_N = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{2} n_i F_{ij}.$$

Sie berücksichtigt die Anzahl aller Beobachtungen $N = 2n = 2(n_1+n_2)$. Die ungewichtete mittlere Verteilungsfunktion wird hingegen durch

$$H = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} F_{ij}$$

definiert. Der relative Behandlungseffekt einer Faktorstufenkombination setzt die entsprechende Verteilungsfunktion in Beziehung zur mittleren Verteilungsfunktion. Wie bei dieser gibt es auch für den Behandlungseffekt eine gewichtete und eine ungewichtete Version. Der gewichtete beziehungsweise der ungewichtete relative Behandlungseffekt wird durch

$$p_{N,ij} = \int H_N dF_{ij} \quad \text{beziehungsweise} \quad p_{ij} = \int H dF_{ij} \quad \text{für} \quad i,j = 1,2$$

definiert. Der gewichtete relative Behandlungseffekt ist vor allem im Fall von unverbundenen Stichproben sehr verbreitet, da sich dessen Schätzer einfach aus dem Mittelwert der Ränge berechnen lässt (Kruskal und Wallis, 1952). Im Gegensatz zum ungewichteten relativen Behandlungseffekt hängt er jedoch von den Stichprobenumfängen n_1 und n_2 ab. Diese Abhängigkeit ist nicht wünschenswert, weshalb der ungewichtete relative Effekt entwickelt wurde. Der Schätzer dieses Effektes ergibt sich aus dem Mittelwert der harmonischen Ränge (Siemer, 1999; Kulle, 1999), die jedoch nicht verbreitet sind. Der Name der harmonischen Ränge ergibt sich aus der Abhängigkeit vom harmonischen Mittel der Stichprobenumfänge. Für die harmonischen Ränge ist bei Kulle (1999) und Siemer (1999) auch der Begriff Pseudoränge üblich. Im balancierten Fall sind beide Behandlungseffekte gleich, da dann die gewichtete mittlere Verteilungsfunktion H_N mit der ungewichteten mittleren Verteilungsfunktion H übereinstimmt.

In der nichtparametrischen Kovarianzanalyse könnte ebenfalls der gewichtete relativen Effekt

$$p_{n,ij} = \frac{1}{n}\int n_1 F_{1j} + n_2 F_{2j}\, dF_{ij}, \quad i,j = 1,2,$$

verwendet werden. Hier wird der Index n anstatt N verwendet, da in der Ziel- beziehungsweise der Kovariablen nur n Beobachtungen auftreten. Aufgrund der Zusammenhangs

$$\begin{aligned}
p_{n,2j} - p_{n,1j} &= \frac{1}{n} \int n_1 F_{1j} + n_2 F_{2j} \, d\left(F_{2j} - F_{1j}\right) \\
&= \frac{1}{n} \left(\frac{n_2 - n_1}{2} + n_1 \int F_{1j} \, dF_{2j} - n_2 \int F_{2j} \, dF_{1j} \right) \\
&= \frac{1}{n} \left(\frac{n_2 - n_1}{2} + n_1 \int F_{1j} \, dF_{2j} - n_2 \left(1 - \int F_{1j} \, dF_{2j} \right) \right) \\
&= \int F_{1j} \, dF_{2j} - \frac{1}{2} \\
&= p_j - \frac{1}{2}
\end{aligned}$$

ist dies im Zwei-Stichproben-Fall jedoch nicht notwendig.

Für die Interpretation des relativen Effektes ist eine Darstellung mit Hilfe von Wahrscheinlichkeiten hilfreich. Seien die Zufallsvariablen $Y_N \sim H_N$, $Y \sim H$ und $X \sim F_{ij}$ unabhängig, dann gilt

$$p_{N,ij} = \mathrm{P}\left(Y_N < X\right) + \frac{1}{2} \mathrm{P}\left(Y_N = X\right) \quad \text{und} \tag{4.18}$$

$$p_{ij} = \mathrm{P}\left(Y < X\right) + \frac{1}{2} \mathrm{P}\left(Y = X\right). \tag{4.19}$$

Y_N beziehungsweise Y kann man sich vorstellen als Zufallsvariablen, die mit Wahrscheinlichkeit n_i/N beziehungsweise $1/4$ nach F_{ij} verteilt ist. Dies ist so zu verstehen, dass zur Bestimmung des Wertes von Y_N und Y zwei Zufallsexperimente nötig sind. Im ersten wird die Verteilungsfunktion F_{ij} mit der Wahrscheinlichkeit n_i/N beziehungsweise $1/4$ gezogen; im zweiten wird dann die Beobachtung gemäß der vorher bestimmten Verteilungsfunktion ermittelt. Bei dieser Interpretation wird besonders deutlich, wieso man die ungewichtete Version vorziehen sollte. Im ungewichteten Fall hängt die Wahrscheinlichkeit, mit der die Zufallsvariable nach F_{ij} verteilt ist, nicht vom Stichprobenumfang ab. Es gibt allerdings einen Fall, in dem die Gewichtung mit den Stichprobenumfängen gerechtfertigt scheint. Stellen die Verhältnisse der Stichprobenumfänge n_i/n das natürliche Verhältnis der Gruppen in einer zu beschreibenden Population dar, so spiegelt $p_{N,ij}$ das Ziehen aus dieser Population wider (Vargha und Delaney, 1998).

Die Interpretation der relativen Effekte mit Hilfe von Wahrscheinlichkeiten legt nahe, dass diese alle Werte im Einheitsintervall annehmen können. Das wirkliche Intervall für $p_{N,ij}$ und p_{ij} ist jedoch kleiner, da Y_N beziehungsweise Y auch immer mit der Wahrscheinlichkeit n_i/N beziehungsweise $1/4$ nach F_{ij} verteilt sind. Berechnet man aber die

Wahrscheinlichkeit $\mathrm{P}\,(Y < X) + 1/2\,\mathrm{P}\,(Y = X)$ von zwei identisch verteilten Zufallsvariablen Y und X, so ist diese immer gleich $1/2$. Somit ergibt sich für die Wertebereiche der relativen Effekte

$$p_{N,ij} \quad \in \quad \left[\frac{n_i}{2N};\, 1 - \frac{n_i}{2N}\right] \quad \text{und} \quad p_{ij} \quad \in \quad \left[\frac{1}{8};\, 1 - \frac{1}{8}\right].$$

Die Ränder der Intervalle werden angenommen, falls die Zufallsvariablen X_{ikj} fast sicher kleiner oder größer sind als alle anderen Zufallsvariablen X_{tkv}. In diesem Fall ist die Annahme (4.15) von Seite 27 allerdings verletzt. Da der Effekt $\int F_1 dF_2$, der im Wilcoxon-Mann-Whitney-Test (Wilcoxon , 1947; Mann und Whitney, 1947) verwendet wird, alle Werte im Einheitsintervall annehmen kann, sind die $p_{N,ij}$ und p_{ij} keine Verallgemeinerung von $\int F_1 dF_2$. Daher schlägt Domhof (2001) vor, einen modifizierten Effekt zu verwenden:

$$p^*_{ij} \quad = \quad \frac{1}{3} \sum_{(t,v)\neq(i,j)} \int F_{tv} dF_{ij}.$$

p^*_{ij} ist von den Stichprobenumfängen unabhängig und kann alle Werte im Einheitsintervall annehmen, da er nicht die Verteilungsfunktion F_{ij} mit sich selbst vergleicht. Domhof (2001) beschreibt die Eigenschaften von p^*_{ij} sehr ausführlich und schlägt verschiedene Ansätze vor, wie man Konfidenzintervalle berechnen kann. Viele dieser Techniken und Eigenschaften von p^*_{ij} können auf den ungewichteten relativen Effekt übertragen werden, da

$$p^*_{ij} \quad = \quad \frac{4}{3}\left(p_{ij} - \frac{1}{8}\right)$$

gilt. Auf ähnliche Weise kann man auch den gewichteten relativen Effekt auf das gesamte Einheitsintervall strecken.

Des Weiteren wird durch Domhof (2001) mit seiner Definition 3.3 eine Sprechweise für die Ordnung der relativen Effekte beziehungsweise der dazugehörigen Verteilungsfunktionen eingeführt. Dafür sei G eine beliebige Verteilungsfunktion und $Y \sim G$ sei unabhängig von den Zufallsvariablen $X_{i1j} \sim F_{ij}$ und $X_{t1v} \sim F_{tv}$. Dann heißt F_{ij} tendenziell kleiner (größer, gleich) als F_{tv} mit Bezug auf G, falls

$$\begin{aligned} &\int G dF_{ij} = \mathrm{P}(Y < X_{i1j}) + \frac{1}{2}\,\mathrm{P}(Y = X_{i1j}) \\ &< (>, =) \quad \int G dF_{tv} = \mathrm{P}(Y < X_{t1v}) + \frac{1}{2}\,\mathrm{P}(Y = X_{t1v}) \end{aligned}$$

gilt. Für die Zufallsvariablen wird die gleiche Sprechweise verwendet.

Der relative Behandlungseffekt lässt sich durch eine Änderung der Anzahl der Indices einfach auf mehrere Faktorstufen oder Faktoren erweitern. Außerdem kann die mittlere Verteilungsfunktion auch noch durch eine Scorefunktion J gewichtet werden. Für die dann von J abhängigen relativen Behandlungseffekte

$$p_{N,ij}(J) \quad = \quad \int J(H_N) dF_{ij}$$

ergeben sich aber einige technische Schwierigkeiten (Domhof, 1999). Außerdem ist eine Interpretation mit Hilfe von Wahrscheinlichkeiten dann nicht mehr möglich. In dieser Arbeit wird daher keine Scorefunktion verwendet.

Wie bei den Marginalverteilungen (siehe Gleichung (4.16) auf Seite 27) können auch die relativen Effekte in additive Komponenten zerlegt werden. Aufgrund der oben genannten Nachteile der gewichteten relativen Effekte werden nur die ungewichteten relativen Effekte betrachtet. Trotzdem kann man die folgenden Effekte auch im gewichteten Fall definieren. Für die Zerlegung

$$p_{ij} \quad = \quad p_M + p_{G,i} + p_{T,j} + p_{W,ij}, \quad i,j = 1,2,$$

gibt es wie bei den Verteilungsfunktionen zwei Kodierungen. Bei der Effektkodierung

$$\sum_{i=1}^{2} p^e_{G,i} \quad = \quad \sum_{j=1}^{2} p^e_{T,j} \quad = \quad \sum_{i=1}^{2} p^e_{W,ij} \quad = \quad \sum_{j=1}^{2} p^e_{W,ij} \quad = \quad 0.$$

werden die Effekte mit e und bei der Referenzkodierung

$$p^r_{G,1} \quad = \quad p^r_{T,1} \quad = \quad p^r_{W,11} \quad = \quad p^r_{W,21} \quad = \quad p^r_{W,12} \quad = \quad 0.$$

mit r gekennzeichnet. Zwischen den beiden Kodierungen ergeben sich die analogen Beziehungen und Probleme wie bei den Verteilungsfunktionen. Als relativen Effekt der Wechselwirkung kann man daher

$$p_W \quad = \quad \int \frac{1}{2}\left(F_{11} + F_{22}\right) d\left(\frac{1}{2}\left(F_{12} + F_{21}\right)\right)$$

bezeichnen. Auf die relativen Effekte der Haupteffekte G und T soll nicht näher eingegangen werden, da diese bei vorhandener Wechselwirkung von der Kodierung abhängen. Der relative Effekt der Wechselwirkung kann ähnlich wie die anderen relativen Effekte mit Hilfe von Wahrscheinlichkeiten interpretiert werden. Dazu definieren wir die unabhängigen Zufallsvariablen

$$Y_1 \sim (F_{11} + F_{22})/2 \quad \text{und} \quad Y_2 \sim (F_{12} + F_{21})/2.$$

Die Interpretation dieser Zufallsvariablen erfolgt wieder über eine zufällige Ziehung nach der einen oder anderen Verteilungsfunktion. So ist die Zufallsvariable Y_1 mit Wahrscheinlichkeit $1/2$ nach F_{11} und mit Wahrscheinlichkeit $1/2$ nach F_{22} verteilt. Es gilt der folgende Zusammenhang zwischen dem relativen Effekt der Wechselwirkung und den Wahrscheinlichkeiten:

$$p_W = \mathrm{P}(Y_2 < Y_1) + \frac{1}{2}\,\mathrm{P}(Y_2 = Y_1). \tag{4.20}$$

Die Hypothese über diesen Effekt

$$H_0: \quad p_W = 1/2 \tag{4.21}$$

ist schwächer als die Hypothese $H_0:\ F_{11} - F_{12} = F_{21} - F_{22}$ in (4.17) auf Seite 28, da

$$F_{11} - F_{12} = F_{21} - F_{22} \quad \Rightarrow \quad p_W = 1/2$$

gilt.

Unter der Annahme der Randomisierung (4.1) von Seite 15 vereinfachen sich die Effekte. Dazu wird nun die Referenzkodierung für den Haupteffekt G verwendet. Diese erscheint hier als sehr sinnvoll, da die Behandlungen auf die Baseline-Werte keinen Einfluss haben. Ein Gruppeneffekt existiert dann nicht, da dieser schon zum ersten Zeitpunkt, also vor der Behandlung der Versuchseinheiten, vorhanden sein müsste. Für den Zeiteffekt verwenden wir ebenfalls die Referenzkodierung mit dem ersten Zeitpunkt als Referenz. Da die Zufallsvariablen zufällig aus der Grundgesamtheit gezogen werden, ist hier die Effektkodierung nicht sinnvoll. Dann müsste man den Zeiteffekt $T(x)$ zum ersten Zeitpunkt von $M(x)$ abziehen. Die Zerlegung der Marginalverteilungen ist folglich

$$\begin{aligned}
F_{11}(x) &= M(x), \\
F_{21}(x) &= M(x), \\
F_{12}(x) &= M(x) + T(x) + W_1(x) \quad \text{und} \\
F_{22}(x) &= M(x) + T(x) + W_2(x)
\end{aligned}$$

mit der Referenzkodierung $W_1^r = 0$ oder der Effektkodierung $W_1^e = -W_2^e$. Der Verteilungshaupteffekt der Zeit ist mithin T. Er ist entweder die mittlere Veränderung über die Zeit oder die Veränderung in der Referenzgruppe über die Zeit, je nachdem welche Kodierung für die Wechselwirkung gewählt wird. Außerdem gilt die Beziehung

$$F_{22} - F_{12} = 2W_2^e = W_2^r,$$

wobei wir wie bisher mit e und r die Effekt- und Referenzkodierung kennzeichnen.

Wie bei den Verteilungsfunktionen kann auch bei den relativen Effekten die Annahme der Randomisierung (4.1) von Seite 15 berücksichtigt werden. Dazu werden neue relative Effekte vorgeschlagen. Für den relativen Haupteffekt der Zeit ergibt sich wie oben das Problem der Kodierung. Wird für die Wechselwirkung die erste Gruppe als Referenz gewählt, so ist der relative Effekt der Zeit

$$p_T^{r,M} = \int F_{12} dM.$$

Der relative Effekt $p_T^{r,M}$ ist mithin der Unterschied zwischen der Verteilung aller Baseline-Werte und der Verteilung der Follow-Up-Werte der Referenzgruppe. Im Fall der Effektkodierung wird hingegen der Unterschied zur mittleren Verteilung der Follow-Up-Werte betrachtet:

$$p_T^{e,M} = \int \frac{1}{2}(F_{12} + F_{22}) dM.$$

Wird diese Unterscheidung zwischen den Kodierungen fortgeführt, so erhält man für die Wechselwirkung im Fall der Referenzkodierung den Effekt

$$p_W^{r,M} = \int F_{12} dF_{22}$$

und für die Effektkodierung

$$p_W^{e,M} = \int \frac{1}{2} \left(F_{12} + F_{22}\right) dF_{22}.$$

Aufgrund der Beziehung

$$p_W^{e,M} = \frac{1}{2} p_W^{r,M} + \frac{1}{4}$$

ist die Unterscheidung der Kodierungen für die Wechselwirkung jedoch müßig. Außerdem unterscheidet sich der Effekt $p_W^{r,M}$ nicht von dem Effekt p_2 aus der Kovarianzanalyse, weshalb im Weiteren nur noch die Bezeichnung p_2 verwendet wird.

Die Hypothesen auf keine Wechselwirkung können im randomisierten Fall entweder über die Verteilungsfunktionen

$$H_0: \quad F_{12} = F_{22}$$

oder mit Hilfe des relativen Effektes

$$H_0: \quad p_2 = 1/2$$

gestellt werden. Diese Hypothesen entsprechen genau den Hypothesen aus der nichtparametrischen Kovarianzanalyse im randomisierten Fall. Die Hypothesen $H_0: F_{12} = F_{22}$ und $H_0: F_{12} - F_{11} = F_{22} - F_{21}$ sind im randomisierten Fall mit $F_{11} = F_{21}$ natürlich äquivalent. Für die Hypothesen $H_0: p_W = 1/2$ und $H_0: p_2 = 1/2$ gilt dies hingegen nicht, was man an der folgenden Darstellung von p_W erkennt:

$$\begin{aligned} p_W &= \int \frac{1}{2}(F_{11} + F_{22})\, d\left(\frac{1}{2}(F_{12} + F_{21})\right) \\ &= \frac{1}{4}\left(\int F_{11} dF_{21} + \int F_{22} dF_{12} + \int F_{11} dF_{12} + \int F_{22} dF_{21}\right). \end{aligned}$$

Das erste Integral $\int F_{11} dF_{21}$ ist nach der Voraussetzung der Randomisierung 1/2. Das zweite Integral $\int F_{22} dF_{21}$ ist gerade p_2 und nimmt daher unter der Hypothese (4.21) von Seite 33 den Wert 1/2 an. Die beiden anderen Integrale vergleichen die Marginalverteilungen innerhalb der Gruppen.Die Hypothesen über die relativen Effekte p_W und p_2 unter $F_{11} = F_{21}$ sind nicht äquivalent, da die Summe dieser beiden Integrale nicht eins sein muß.

4.4 Nichtparametrische Analyse durch Schichtung

In diesem Kapitel wird ein Ansatz von Bajorski und Petkau (1999) vorgestellt, welches auf der Schichtung der Zufallsvariablen zum zweiten Zeitpunkt nach den Werten der Zufallsvariablen zum ersten Zeitpunkt beruht. Es werden also nur die Zufallsvariablen zum zweiten Zeitpunkt miteinander verglichen, die denselben Ausgangswert, nämlich die Beobachtung zum ersten Zeitpunkt, haben. Damit werden für jeden möglichen Wert der Zufallsvariablen zum ersten Zeitpunkt Schichten der Zufallsvariablen zum zweiten Zeitpunkt gebildet. Ohne eine kategoriale Struktur der Beobachtungen zum ersten Zeitpunkt und eine Ordnungsstruktur der Beobachtungen zum zweiten Zeitpunkt ist dies nicht möglich. Diese Forderungen sind im Fall von zwei Gruppen, zwei Zeitpunkten und ordinalen Daten gegeben.

Bei dem Ansatz von Bajorski und Petkau (1999) ist es also im Gegensatz zu den vorhergehenden Verfahren zwingend notwendig, dass die Zufallsvariablen X_{ikj} auf einer geordnet kategoriellen Skala beobachtet werden. Diese Kategorien werden im Folgenden mit den natürlichen Zahlen von 1 bis r bezeichnet. Mithin ergeben sich r Schichten. In

jeder Schicht können ebenfalls r Werte angenommen werden. Die Anzahl der Versuchseinheiten aus der Gruppe i, welche in eine Schicht u fallen und dort den Wert v haben, bezeichnen wir mit M_{iuv}. Diese Anzahl ist zufällig und der Erwartungswert dieser Zufallsvariablen ist

$$\mathrm{E}\,(M_{iuv}) \;=\; n_i q_{iuv}, \quad i = 1, 2, \quad u, v = 1, \ldots, r,$$

wobei die Bezeichnung

$$\mathrm{P}\left((X_{ik1}, X_{ik2})^t = (u, v)^t\right) \;=\; q_{iuv}, \quad i = 1, 2, \quad u, v = 1, \ldots, r,$$

gelte. Folglich ist die Wahrscheinlichkeit, dass eine Beobachtung aus der Gruppe i in die Schicht u fällt, $q_{iu\cdot} = \sum_{s=1}^{r} q_{ius}$. Bajorski und Petkau (1999) stellen an diese Wahrscheinlichkeiten die Forderung

$$q_{1u\cdot} \;=\; q_{2u\cdot}, \quad u = 1, \ldots, r. \tag{4.22}$$

Diese Annahme ist äquivalent zur Annahme der Randomisierung (4.1) von Seite 15, da die Wahrscheinlichkeiten $q_{iu\cdot}$ die Marginalverteilungen zum ersten Zeitpunkt bestimmen. Sie ist erfüllt, wenn die Versuchseinheiten den beiden Versuchsgruppen zufällig zugeteilt werden. Daher bezeichnen Bajorski und Petkau (1999) diese Annahme als natürlich. Andererseits erscheint die Annahme (4.22) bei einer Schichtung nicht nötig zu sein, denn die Idee der Schichtung stammt aus den Beobachtungsstudien, bei denen man nicht von einer randomisierten Versuchsanlage ausgehen kann. Um trotzdem Vergleiche anstellen zu können, werden daher Schichten von Versuchseinheiten gebildet, welche die gleichen Ausgangswerte haben. Damit sind die Versuchseinheiten innerhalb der Schichten homogen. Trotzdem wird die Voraussetzung (4.22) von Bajorski und Petkau (1999) gestellt.

Eine schwächere Annahme als (4.22) an die Marginalverteilungen zum ersten Zeitpunkt ist hingegen wichtig für die Schichtungsanalyse:

$$\mathrm{P}(X_{1k1} = X_{2s1}) \;>\; 0.$$

Sie verhindert, dass fast sicher alle Beobachtungen der einen Gruppe in andere Schichten fallen als die Beobachtungen der anderen Gruppe. In diesem Fall könnte man keine Schicht auswerten, da die Vergleiche der Gruppenn nur innerhalb der Schichten durchgeführt werden. Fallen nun in eine Schicht nur die Beobachtungen aus einer Gruppe, so können diese nicht mit in die Analyse einfließen, da sie nicht mit Beobachtungen aus der anderen Gruppe verglichen werden können. Die Schichten u, in denen

$$q_{1u\cdot} q_{2u\cdot} \;=\; 0 \tag{4.23}$$

gilt, können folglich nicht in der Analyse berücksichtigt werden. Im Folgenden soll also das Produkt $q_{1u\cdot}q_{2u\cdot}$ für alle Schichten positiv sein.

Die Theorie zur Herleitung der Verteilungen der Teststatistiken basiert auf den Verteilungsfunktionen in den Schichten:

$$\begin{aligned} F_{i|u}(x) &= \mathrm{P}\left(X_{ik2} < x \mid X_{ik1} = u\right) + \frac{1}{2}\,\mathrm{P}\left(X_{ik2} = x \mid X_{ik1} = u\right), \\ & \quad i = 1,2, \quad u = 1,\ldots,r. \end{aligned}$$

Diese können auch mit den bedingten Wahrscheinlichkeiten

$$q_{iv|u} = \mathrm{P}\left(X_{ik2} = v \mid X_{ik1} = u\right), \quad i = 1,2, \quad u,v = 1,\ldots,r,$$

berechnet werden. Zusammen mit den Marginalwahrscheinlichkeiten $q_{iu\cdot}$ bestimmen sie die gemeinsamen Verteilungen der Gruppen, da

$$q_{iuv} = q_{iv|u}q_{iu\cdot}, \quad i = 1,2, \quad u,v = 1,\ldots,r,$$

gilt. Mit $m_{iu\cdot}$ bezeichnen wir die Anzahl der Versuchseinheiten der Gruppe i in der Schicht u. Diese Anzahlen werden bei der Betrachtung der bedingten Verteilungsfunktionen als fest angesehen.

An die bedingten Wahrscheinlichkeiten $q_{iv|u}$ beziehungsweise an die bedingten Verteilungen $F_{i|u}$ müssen für die asymptotischen Aussagen über die Teststatistiken ähnliche Voraussetzungen gestellt werden wie in der Kovarianzanalyse. Diese werden hier jedoch für jede einzelne Schicht gestellt. Sind sie in einer Schicht nicht erfüllt, so können die Beobachtungen in dieser Schicht nicht berücksichtigt werden. Die erste der Voraussetzungen schließt Einpunktverteilungen aus:

$$\mathrm{P}\left(X_{i12} \neq X_{i22} \mid X_{i11} = X_{i21} = u\right) > 0, \quad i = 1,2, \quad u = 1,\ldots,r.$$

Sie entspricht der Voraussetzung (4.2) auf Seite 16 der Kovarianzanalyse. Der Unterschied besteht darin, dass die Voraussetzung hier für jede Schicht gestellt wird. Genauso verhält es sich mit den Trägern der bedingten Verteilungen. Diese sollen sich für jede Schicht überlappen:

$$\mathrm{P}\left(X_{i12} \leq X_{i'12} \mid X_{i12} = X_{i'12} = u\right) > 0, \quad i,i' = 1,2, \quad u = 1,\ldots,r. \tag{4.24}$$

An die Stichprobenumfänge wird ebenfalls eine etwas stärkere Forderung gestellt als bei der Kovarianzanalyse. Wie bei den naiven Verfahren soll (4.12) von Seite 21 auch bei

den Schichtungsverfahren gelten. Zusammen mit der Annahme, dass $q_{1u\cdot}q_{2u\cdot} > 0$ in allen Schichten u gilt, folgt aus (4.12), dass der asymptotische Anteil der Beobachtungen der Gruppe 1 in der Schicht u dem Wert $\lambda q_{1u\cdot}$ und dem Wert $(1-\lambda)q_{2u\cdot}$ in der Gruppe 2 entspricht. Mit Hilfe dieser Annahmen können die Verteilungen der Teststatistiken nun hergeleitet werden.

Einige dieser Teststatistiken beruhen auf der Schätzung von Effekten. Diese werden neu vorgestellt. Da nur innerhalb der Schichten die Beobachtungen verglichen werden, ist es natürlich, diese Effekte innerhalb der Schichten zu definieren. Wie bei der nichtparametrischen Kovarianzanalyse werden also zwei Verteilungen verglichen, hier allerdings nicht für die Ziel- oder Kovariable, sondern für jede Schicht. Damit ist es natürlich, den relativen Effekt $p_{S,u}$ für jede Schicht u zu definieren. Der Index S soll dabei kennzeichnen, dass es sich um den relativen Effekt in der Schicht u handelt. Er wird durch

$$p_{S,u} = \int F_{1|u} dF_{2|u}$$

in jeder Schicht $u = 1, \ldots, r$ definiert. Dieses Vorgehen hat sowohl Vor- als auch Nachteile. Zum einen werden durch die r relativen Effekte die Verteilungsfunktionen der beiden Gruppen sehr detailliert beschrieben; zum anderen ist es schwierig, sich aufgrund von r verschiedenen Effekten für eine Gruppe zu entscheiden. Um trotzdem eine Gesamtaussage treffen zu können, müssen die r Effekte zusammengefasst werden. Motiviert durch die Teststatistiken von van Elteren (1960) und Bajorski und Petkau (1999) wird daher eine gewichtete Summation dieser Effekte vorgeschlagen.

Die Hypothesen können auf zwei Arten gestellt werden. Zum einen kann man sie über die Wahrscheinlichkeiten $q_{iv|u}$ oder äquivalent über die Verteilungsfunktionen $F_{i|u}$ stellen:

$$H_0: \quad F_{1|u} = F_{2|u} \quad \text{für alle} \quad u = 1, \ldots, r. \tag{4.25}$$

Zum anderen können dafür die relativen Effekte $p_{S,u}$ verwendet werden:

$$H_0: \quad p_{S,u} = \frac{1}{2} \quad \text{für alle} \quad u = 1, \ldots, r. \tag{4.26}$$

Die Hypothese (4.25) wird von Bajoski und Petkau (1999) verwendet. Wir schlagen die Hypothese (4.26) vor, da sie interpretierbare Effekte betrachtet. Außerdem sind die verwerwendeten Tests für beide Hypothesen nur konsistent bezüglich der Alternative zur Hypothese (4.26).

Wie bei den vorhergehenden Verfahren impliziert die Hypothese über die Verteilungen die Hypothese mit Hilfe der relativen Effekte. Eine Beziehung zu den Hypothesen der nichtparametrischen Kovarianzanalyse läßt sich nur bei randomisierten Studien

durchführen. In diesem Fall sind unter der Hypothese (4.25) die gemeinsamen Verteilungen der Gruppen gleich. Dies ist die stärkste Hypothese, die man stellen kann. Mithin sind unter dieser Hypothese auch die Marginalverteilungen zum zweiten Zeitpunkt gleich und es gilt sowohl $H_0 : \ F_{12} = F_{22}$ als auch $H_0 : \ p_2 = 1/2$.

4.5 Parametrische faktorielle Analyse

In diesem Kapitel stellen wir die Möglichkeit einer parametrischen Analyse der Daten vor. Diese basiert auf der Theorie der verallgemeinerten Schätzgleichungen, welche in der Literatur als GEE (*Generalized Estimating Equations*) bezeichnet werden. Die grundlegende Arbeit wurde dazu von Liang und Zeger (1986) beziehungsweise Zeger und Liang (1986) geschrieben. Eine etwas ausführlichere anwendungsbezogene Beschreibung findet man bei Ziegler et al. (1996) und Fahrmeir und Tutz (1994). Sehr theoretisch sind hingegen die Ausführungen zu den GEE in Kapitel 5.4 von Shao (1999). Während Ziegler et al. (1996) die GEE aus dem biometrischen Blickwinkel betrachten und die GEE als Verallgemeinerung der verallgemeinerten linearen Modelle GLM (*Generalized Linear Models*) beschreiben, beleuchtet Shao (1999) vor allem den Standpunkt der mathematischen Statistik. Außerdem werden dort die GEE nicht nur als Verallgemeinerung der GLM, für die das Buch von McCullagh und Nelder (1989) eine gute Übersicht zu der grundlegenden Arbeit von Nelder und Wedderburn (1972) gibt, dargestellt, sondern auch als Verallgemeinerung von anderen Konzepten wie Kleinste-Quadrate-, (Quasi-)Maximum-Likelihood- oder M-Schätzern beschrieben.

Wie bei der nichtparametrischen faktoriellen Analyse werden die beiden festen Faktoren Gruppe G und Zeit T mit ihrer Wechselwirkung W betrachtet. Ferner ist die ordinale Skala der Zufallsvariablen wichtig für das Modell. Die Zufallsvariablen X_{ikj} können, wie bei der Schichtungsanalyse im vorhergehenden Abschnitt, nur die natürlichen Zahlen zwischen 1 und r annehmen, welche die Kategorien der ordinalen Skala beschreiben.

Um die Möglichkeiten der GEE verwenden zu können, führen wir nach einem Vorschlag von Clayton (1992) beziehungsweise Kenward et al. (1994) neue, binäre Zufallsvariablen Y_{ikjs} ein, mit deren Randverteilungen die ursprünglichen Randverteilungen der Zufallsvariablen X_{ikj} beschrieben werden können. Diese Zufallsvariablen werden wie folgt erzeugt:

$$Y_{ikjs} \quad = \quad \begin{cases} 1 & \text{falls } X_{ikj} \leq s \\ 0 & \text{falls } X_{ikj} > s \end{cases} \qquad i, j = 1, 2, \quad k = 1, \ldots, n_i, \quad s = 1, \ldots, r-1.$$

Es werden also für jede ordinale Zufallsvariable X_{ikj} mit r Kategorien $r-1$ binäre Zufallsvariablen gebildet. Damit gilt der Zusammenhang zwischen den Marginalverteilungen

dieser Zufallsvariablen

$$P\left(X_{ikj} \leq s\right) \;=\; P\left(Y_{ikjs} = 1\right).$$

Diese Wahrscheinlichkeiten werden mit $\pi_{ijs} = P\left(Y_{ikjs} = 1\right)$ bezeichnet. Die π_{ijs} bilden die Grundlage, um die Effekte im parametrischen faktoriellen Modell zu beschreiben. An sie wird die folgende Modellgleichung mit der Link-Funktion g gestellt:

$$g(\pi_{ijs}) \;=\; \alpha_s + \beta_{ij}, \quad i,j = 1,2, \quad s = 1,\ldots,r-1. \tag{4.27}$$

Der Term $\eta_{ijs} = \alpha_s + \beta_{ij}$ wird auch als linearer Prädiktor oder lineare Komponente des Modells bezeichnet. Die Link-Funktion g bildet vom offenen Einheitsintervall auf die reellen Zahlen ab. Des Weiteren ist sie monoton steigend, damit invertierbar, und zweimal stetig differenzierbar. Aufgrund der Modellgleichung (4.27) sind die Marginalverteilungen stochastisch geordnet. Dies folgt aus der Unabhängigkeit der α_s von der Gruppe i und dem Zeitpunkt j und der strengen monotonen Steigung von g (Fahrmeir und Tutz, 1994, Abschnitt 3.3.5).

Die Einschränkung der Menge der möglichen Verteilungsfunktionen auf eine Untermenge mit stochastischer Ordnung ist eine wesentliche Eigenschaft des Modells der parametrischen faktoriellen Analyse. Aufgrund dieser Einschränkung ist es überhaupt erst möglich zwei Marginalverteilungen mit Hilfe des Parameters β_{ij} zu vergleichen. Die nichtparametrische Kovarianzanalyse schränkt die Menge der möglichen Verteilungsfunktionen hingegen nur unwesentlich ein. Das Modell der nichtparametrischen Kovarianzanalyse ist mithin wesentlich allgemeiner als bei der parametrischen faktoriellen Analyse.

Eine Interpretation der Cutpoints α_s in der Modellgleichung (4.27) ist mittels des *Threshold*-Ansatzes möglich. Dieser Ansatz, der schon von Edwards und Thurstone (1952) beschrieben wird, geht davon aus, dass der beobachtbaren Zufallsvariablen X auf der ordinalen Skala eine nicht beobachtbare latente Zufallsvariable U auf einer stetigen Skala zu Grunde liegt. Diese stetige Skala ist durch $r-1$ Cutpoints α_s in r Intervalle aufgeteilt. Die stetige Verteilung der latenten Zufallsvariablen U bestimmt nun die Verteilung von X durch

$$X \;=\; s \;\Leftrightarrow\; \alpha_{s-1} < U \;<\; \alpha_s$$

für $s = 1,\ldots,r$ mit $\alpha_0 = -\infty$ und $\alpha_r = \infty$. Damit ist auch der Ausdruck *Cutpoint* erklärt. Jedes α_s zerschneidet die reelle Achse quasi in zwei Stücke. Ist U außerdem nach einer Verteilungsfunktion $F(u+\eta)$ mit einem Lokationsparameter η verteilt, so gilt:

$$\mathrm{P}(X \leq s) \;=\; \mathrm{P}(U < \alpha_s)$$

$$= \quad F(\alpha_s + \eta).$$

Mit der Umkehrfunktion von F erhält man so eine Modellgleichung

$$F^{-1}(\mathrm{P}(X \leq s)) \quad = \quad \alpha_s + \eta,$$

die der Gleichung (4.27) entspricht. Für eine noch ausführlichere Beschreibung dieses Ansatzes mit einigen Beispielen für F sei auf das Kapitel 3.3.1 in Fahrmeir und Tutz (1994) verwiesen.

Um die Identifizierbarkeit der Parameter zu sichern und damit die Wahrscheinlichkeiten π_{ijs} auch wirklich im offenen Einheitsintervall liegen, muss an die Marginalverteilungen der Zufallsvariablen X_{ikj} die Voraussetzung

$$\mathrm{P}(X_{ikj} = s) \quad > \quad 0, \quad i, j = 1, 2, \quad s = 1, \ldots, r, \tag{4.28}$$

gestellt werden. Die Wahrscheinlichkeit, eine Kategorie s zu beobachten, soll also für alle Kategorien und alle Faktorstufenkombinationen positiv sein. Diese Voraussetzung ist im Vergleich zu den beiden Voraussetzungen (4.2) von Seite 16 und (4.3) von Seite 16 der nichtparametrischen Kovarianzanalyse sehr stark, da sie beide impliziert.

Für die Link-Funktion g gibt es verschiedene Vorschläge. Die gebräuchlichste Link-Funktion bei der Bernoulliverteilung ist die logit-Funktion:

$$g_l(z) \quad = \quad \log\left(\frac{z}{1-z}\right).$$

Dabei bezeichnet log den natürlichen Logarithmus. Es werden aber auch die probit-Funktion

$$g_p(z) \quad = \quad \Phi^{-1}(z)$$

mit Φ als Verteilungsfunktion der Standardnormalverteilung oder die komplementäre log-log-Funktion

$$g_c(z) \quad = \quad \log(-\log(1-z))$$

verwendet. Die Umkehrfunktionen dieser Link-Funkionen sind die Verteilungsfunktionen der logistischen Verteilung

$$g_l^{-1}(z) \quad = \quad \frac{e^z}{1+e^z},$$

der Standardnormalverteilung

$$g_p^{-1}(z) \;=\; \Phi(z)$$

und der extremen Minimalwertverteilung

$$g_c^{-1}(z) \;=\; 1 - e^{-e^z}.$$

Sie werden auch als Response-Funktionen bezeichnet und entsprechen beim Threshold-Ansatz der Verteilung F der latenten Zufallsvariablen U. Zur Verdeutlichung der Unterschiede zwischen diesen Funktionen sind die Graphen der verschiedenen Verteilungen in der Abbildung 4.6 dargestellt.

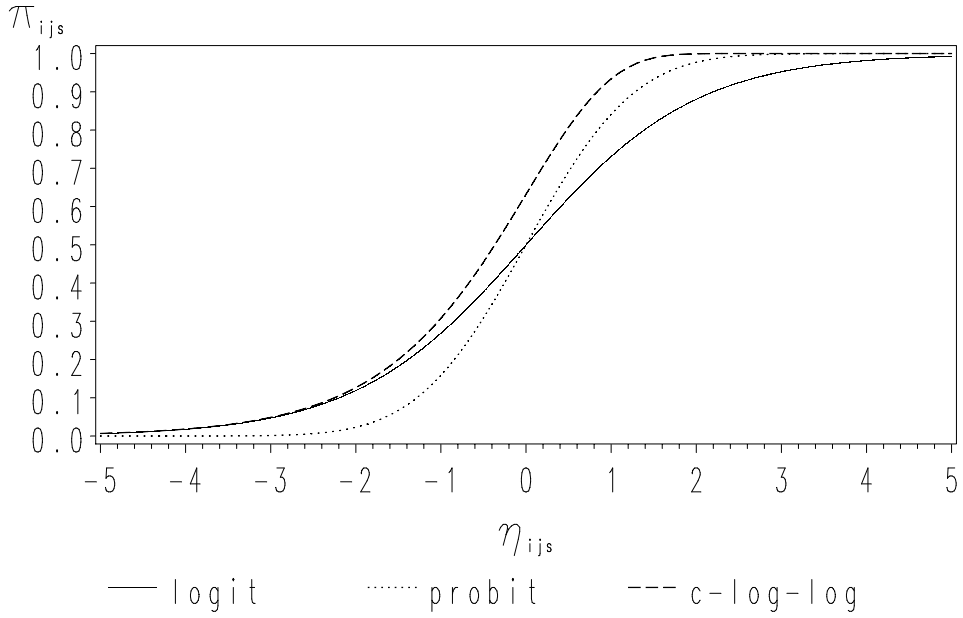

Abbildung 4.6: Graphen der logistischen Verteilungsfunktion, der Standardnormalverteilung und der extremen Minimalwertverteilung

Die Zufallsvariablen Y_{ikjs} sind Bernoulliverteilt mit der Erfolgswahrscheinlichkeit π_{ijs}. Diese unbekannten Wahrscheinlichkeiten charakterisieren also vollständig die gesuchten Marginalverteilungen. Man sagt auch, dass diese sie parametrisieren, weshalb das Modell parametrisch genannt wird. Die Bernoulliverteilung gehört zu einer größeren Klasse von parametrischen Verteilungen, der exponentiellen Familie (Shao, 1999; Fahrmeir und Tutz,

1994). Für die exponentielle Familie spielen die natürlichen Parameter eine wichtige Rolle. Im Fall der Bernoulliverteilung ist der natürliche Parameter

$$\theta_{ijs} = \log\left(\frac{\pi_{ijs}}{1-\pi_{ijs}}\right).$$

Daher wird die logit-Funktion auch als natürliche Link-Funktion bezeichnet. Sie verbindet den natürlichen Parameter θ_{ijs} mit dem Parameter π_{ijs}.

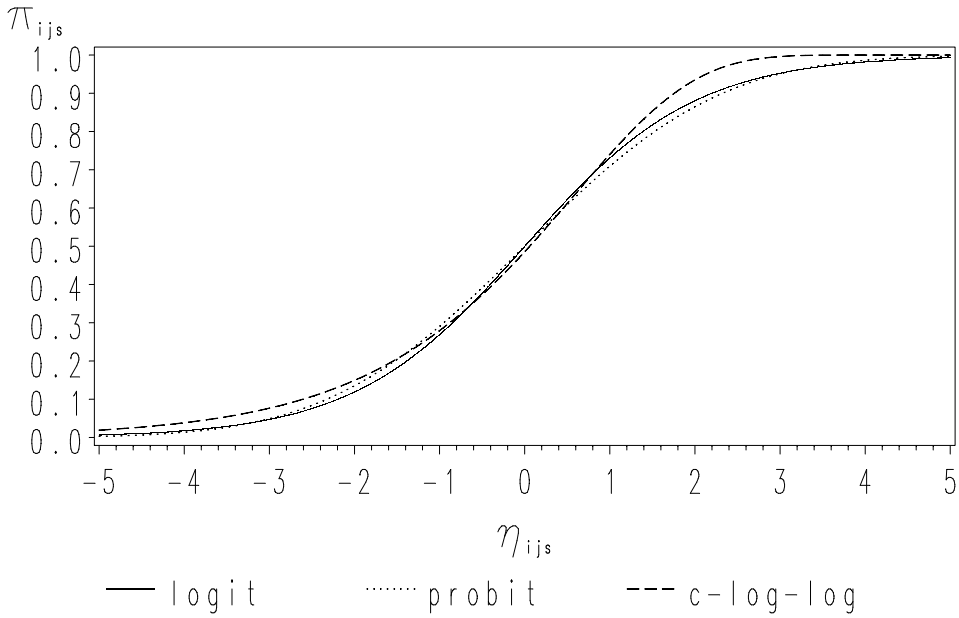

Abbildung 4.7: Graphen der logistischen Verteilungsfunktion, der Standardnormalverteilung und der extremen Minimalwertverteilung bei gleichem Erwartungswert und gleicher Varianz

Da sich durch die natürliche Link-Funktion vor allem bei den Effekten einige Vereinfachungen ergeben, wird sie im Folgenden ausschließlich betrachtet. Außerdem unterscheidet sich die zugehörige Response-Funktion, die logistische Verteilung, nicht so sehr von den anderen häufig verwendeten Response-Funktionen. Die Unterschiede in der Abbildung 4.6 werden vor allem durch die unterschiedlichen Erwartungswerte und Varianzen erzeugt. So hat die logistische Verteilung zwar wie die Standardnormalverteilung den Erwartungswert 0, die Varianz ist jedoch mit $\pi^2/3$ mehr als drei mal so hoch wie die Varianz der Standardnormalverteilung. Bei der extremen Minimalverteilung unterscheidet sich mit $\pi^2/6$ nicht nur die Varianz, sondern mit circa $-0,5772$ auch der Erwartungswert. Werden Erwartungswerte und Varianzen aller drei Verteilungen wie die der logistischen Verteilung

gewählt, so ist der Unterschied zwischen den Verteilungen recht gering, wie man in der Abbildung 4.7 sieht.

Nachdem nun die Link-Funktion g ausführlich diskutiert wurde, werden wir im Folgenden die Effekte, welche mit den Wahrscheinlichkeiten π_{ijs} verknüpft sind, näher betrachten. Die π_{ijs} sind die Erwartungswerte der Bernoulli-verteilten Zufallsvariablen Y_{ikjs}. Lägen nicht zwei Zeitpunkte und mehrere Kategorien vor, so müßte man nur die dichotomen Daten aus zwei Gruppen betrachten. Bei diesen wird als Kennzahl häufig das Verhältnis der Erfolgswahrscheinlichkeit π_i zur Wahrscheinlichkeit für den Misserfolg $1 - \pi_i$

$$\omega_i = \frac{\pi_i}{1 - \pi_i}$$

herangezogen. Dieses Verhältnis ω_i wird als *Odds*, Quote oder Chance in der Gruppe i bezeichnet (Guggenmoos-Holzmann und Wernecke, 1995; Kreienbrock und Schach, 1997). Das Verhältnis der beiden Odds der Gruppen i und i'

$$\kappa_{i,i'} = \frac{\omega_i}{\omega_i'}$$

heißt *Odds Ratio* oder Quoten-Quotient der Gruppen i und i' (Guggenmoos-Holzmann und Wernecke,1995). Da sich auch in der deutschsprachigen Literatur die englischen Ausdrücke durchgesetzt haben, werden diese im Folgenden verwendet. Anschaulich können wir das Odds Ratio folgendermaßen verstehen. Ist es größer als eins, so ist die Chance für einen Erfolg in der Gruppe i um den Faktor $\kappa_{i,i'}$ größer als in der Gruppe i'. Im vorliegenden Modell gibt es neben den beiden Gruppen auch noch die beiden Zeitpunkte. Dann wird die Gruppe i zum Zeitpunkt j mit der Gruppe i' zum Zeitpunkt j' durch das Odds Ratio

$$\kappa_{ij,i'j'} = \frac{\omega_{ij}}{\omega_{i'j'}}$$

verglichen. Bei ordinalen Skalen geht man ähnlich vor, indem man die ordinale Skala dichotomisiert. Das heißt, man betrachtet für jede Faktorstufenkombination die $r - 1$ Odds

$$\omega_{ijs} = \frac{\pi_{ijs}}{1 - \pi_{ijs}}.$$

Die Odds Ratios zwischen zwei Gruppen i und i' mit den Zeitpunkten j und j' werden dann für jedes $s = 1, \ldots, r - 1$ gebildet:

$$\kappa_{ijs,i'j's} = \frac{\omega_{ijs}}{\omega_{i'j's}}.$$

Diese Effekte finden sich in der Modellgleichung (4.27) von Seite 40 wieder. Da die natürliche Link-Funktion, der logit-Link, verwendet wird, ergibt sich

$$\begin{aligned} \omega_{ijs} &= \frac{\pi_{ijs}}{1-\pi_{ijs}} \\ &= \exp\left(\alpha_s + \beta_{ij}\right). \end{aligned} \tag{4.29}$$

Diese Modellgleichung wird auch als die des Proportional Odds Modells bezeichnet (Agresti, 1990, Seite 322; McCullagh, 1980), denn die Odds zerfallen in die Faktoren $\exp(\alpha_s)$ und $\exp(\beta_{ij})$. Des Weiteren wird das Modell auch als ordinales logistisches Modell (Scott et al., 1997), kumulatives logit Modell (Lee, 1992; Ananth und Kleinbaum, 1997), kumulatives Odds Modell (Armstrong und Sloan, 1989; Greenland, 1994) oder als McCullagh's gruppiert stetiges Modell (Greenwood und Farewell, 1988) bezeichnet. Die Odds Ratios zwischen den Gruppen i und i' zu den Zeitpunkten j und j' hängen auf Grund der Faktorisierung (4.29) nicht mehr von den Kategorien s der ordinalen Skala ab:

$$\kappa_{ij,i'j'} = \frac{\omega_{ijs}}{\omega_{i'j's}} = \exp\left(\beta_{ij} - \beta_{i'j'}\right)$$

Die Unabhängigkeit der Einflüsse der Faktoren von den Kategorien wird besonders deutlich, wenn beispielhaft die Logarithmen der Odds $L_{ijs} = \log(\omega_{ijs})$ wie in der Abbildung 4.8 aufgetragen werden. Für diese Abbildung wurde eine Skala mit $r = 4$ Kategorien gewählt. Damit ergeben sich die drei Cutpoints α_1, α_2 und α_3, welche die Einflüsse der Kategorien beschreiben. Da diese für beide Gruppen und beide Zeitpunkte gleich sind, sind diejenigen Pfeile, die zu einer Gruppe gehören, immer parallel. Des Weiteren sind die Unterschiede zwischen den L_{ijs} für jeden Cutpoint gleich. Damit wiederholen sich die Unterschiede zwischen den schwarzen und weißen Pfeilen für alle drei Cutpoints. In der Praxis könnte man anhand einer solchen Darstellung überprüfen, inwieweit die Modellgleichung des Proportional Odds Modells gerechtfertigt ist.

Die Einflüsse der Faktoren G und T und ihrer Wechselwirkung wurden bisher zusammenfassend durch die Parameter β_{ij} beschrieben. Um die Haupteffekte und die Wechselwirkung einzeln zu untersuchen, wird β_{ij} additiv zerlegt:

$$\beta_{ij} = \beta_{G,i} + \beta_{T,j} + \beta_{W,ij}, \quad i,j = 1,2.$$

Dabei können wir wie im nichtparametrischem faktoriellen Modell die Effekt- und die Referenzkodierung verwenden. Es ergeben sich die Reparametrisierungsbedingungen

$$\sum_{i=1}^{2} \beta^e_{G,i} = \sum_{j=1}^{2} \beta^e_{T,j} = \sum_{i=1}^{2} \beta^e_{W,ij} = \sum_{j=1}^{2} \beta^e_{W,ij} = 0$$

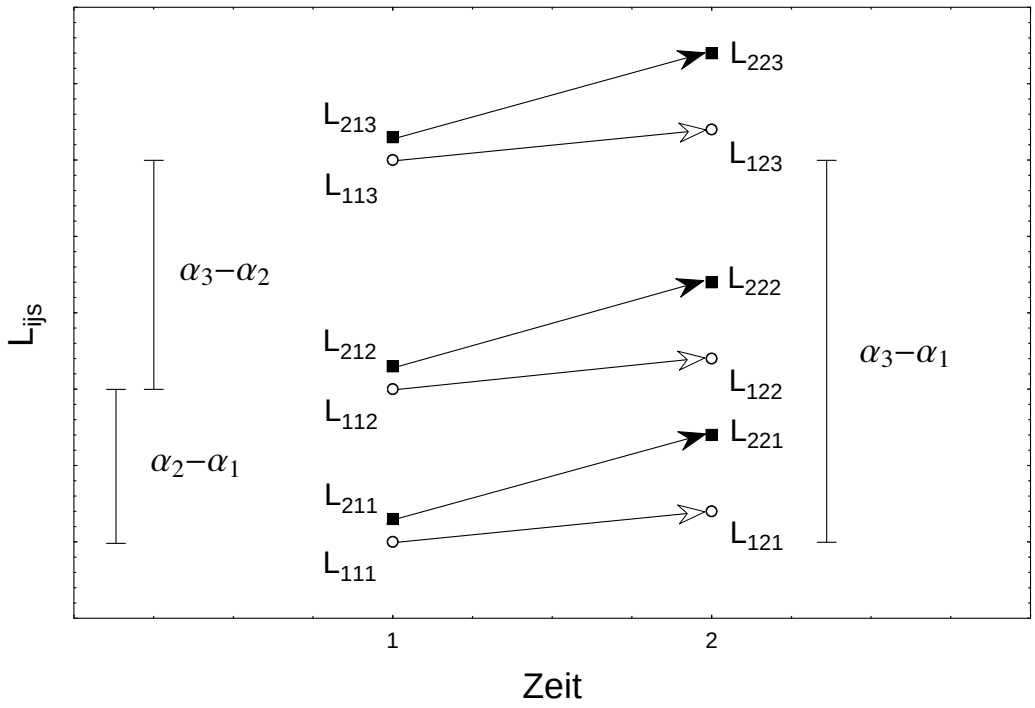

Abbildung 4.8: Die Logarithmen der Odds L_{ijs} im Proportional Odds Modell (4.27) bei vier Kategorien auf der ordinalen Skala

im Fall der Effekt- und

$$\beta^r_{G,1} = \beta^r_{T,1} = \beta^r_{W,11} = \beta^r_{W,21} = \beta^r_{W,12} = 0$$

im Fall der Referenzkodierung. Für die verschiedenen Kodierungen ergeben sich aus der Modellgleichung (4.29) von Seite 45 die folgenden Odds Ratios:

$$\begin{aligned}
\kappa_{21,11} &= \exp(\beta^r_{G,2}) & \kappa_{21,11} &= \exp(2\beta^e_{G,2} - 2\beta^e_{W,22}) \\
\kappa_{12,11} &= \exp(\beta^r_{T,2}) & \kappa_{12,11} &= \exp(2\beta^e_{T,2} - 2\beta^e_{W,22}) \\
\kappa_{22,21} &= \exp(\beta^r_{T,2} + \beta^r_{W,22}) & \kappa_{22,21} &= \exp(2\beta^e_{T,2} + 2\beta^e_{W,22}) \\
\kappa_{22,12} &= \exp(\beta^r_{G,2} + \beta^r_{W,22}) & \kappa_{22,12} &= \exp(2\beta^e_{G,2} + 2\beta^e_{W,22}).
\end{aligned}$$

Um die Notation zu vereinfachen, können die Indizes für die Gruppen und Zeitpunkte bei den Parametern $\beta_{G,2}$, $\beta_{T,2}$ und $\beta_{W,22}$ vernachlässigt werden. Es soll sich im folgenden immer um die Parameter der zweiten Gruppe oder des zweiten Zeitpunktes handeln.

Nun sollen die Abhängigkeiten der Odds Ratios $\kappa_{ij,i'j'}$ von den Parametern β_G, β_T und β_W untersucht werden, falls von einem randomisierten Versuchsplan ausgegangen werden

kann. In diesem unterscheiden sich die beiden Gruppen zum ersten Zeitpunkt nicht, und so sollte $\kappa_{21,11} = 1$ gelten. Daraus folgt für die Referenzkodierung $\beta_G^r = 0$. Bei der Effektkodierung hängt das Odds Ratio hingegen auch vom Parameter der Wechselwirkung β_W^e ab. Dort muß $\beta_G^e = \beta_W^e$ gelten.

Für die Odds Ratios, welche die beiden Gruppen zum zweiten Zeitpunkt, also nach der Behandlung, vergleichen, ergibt sich $\kappa_{22,12} = \exp(\beta_W^r)$ und $\kappa_{22,12} = \exp(4\beta_W^e)$. Damit ist der Parameter für die Wechselwirkung im Fall der Referenzkodierung viermal so groß wie im Fall der Effektkodierung. Bei der Abhängigkeit der Parameter β_T^r und β_T^e voneinander spielen auch die Parameter für die Wechselwirkung eine Rolle. Es ergeben sich ähnliche Probleme wie bei der additiven Zerlegung der relativen Effekte im nichtparametrischen faktoriellen Modell.

Dies macht eine Interpretation des Zeiteffektes schwierig. Im Fall der Effektkodierung ergibt sich der Parameter für den Zeiteffekt das dem Mittel der Logarithmen der Odds Ratios der beiden Gruppen $\beta_T^e = (\log(\kappa_{12,11}) + \log(\kappa_{22,21}))/4$. Im Fall der Referenzkodierung ist es hingegen der Logarithmus des Odds Ratios der Referenzgruppe $\beta_T^r = \log(\kappa_{12,11})$. Daher kann es bei vorhandener Wechselwirkung passieren, dass der Parameter für den Zeiteffekt im Fall der einen Kodierung gleich null, im Fall der anderen Kodierung jedoch deutlich ungleich null ist.

Diese Probleme sind beim Testen von Hypothesen über diese Parameter zu berücksichtigen. Die Hypothesen über die Haupteffekte

$$\begin{aligned} H_0: \quad \beta_G &= 0 \quad \text{und} \\ H_0: \quad \beta_T &= 0 \end{aligned}$$

sind unter Umständen schwer zu interpretieren, falls die Hypothese über die Wechselwirkung

$$H_0: \quad \beta_W = 0$$

nicht erfüllt ist. Diese Hypothesen basieren auf einem Modell, dass die Marginalverteilungen betrachtet. Daher ist es sinnvoll, diese Hypothesen mit den Hypothesen der nichtparametrischen Kovarianzanalyse und der nichtparametrischen faktoriellen Verfahren zu vergleichen.

Die Hypothesen über die Odds Ratios beziehungsweise über die Parameter β_G, β_T und β_W sind im Allgemeinen nicht äquivalent zu den entsprechenden nichtparametrischen Hypothesen. Die nichtparametrischen Hypothesen zerlegen die Wahrscheinlichkeiten additiv.

So kann die Hypothese, dass keine Wechselwirkung bezüglich der Verteilungsfunktionen vorliegt, mit der Notation der parametrischen faktoriellen Modelle als

$$H_0: \quad \pi_{11s} - \pi_{12s} - \pi_{21s} + \pi_{22s} \;=\; 0 \quad \text{für alle} \quad s = 1, \ldots, r-1$$

formuliert werden. Im parametrischen Fall hingegen werden die Logits $L_{ijs} = \log(\pi_{ijs}/(1-\pi_{ijs}))$ für die additive Zerlegung benutzt. Aus

$$H_0: \quad L_{11s} - L_{12s} - L_{21s} + L_{22s} \;=\; 0 \quad \text{für alle} \quad s = 1, \ldots, r-1,$$

also der parametrischen Hypothese auf keine Wechselwirkung, folgt somit nicht die entsprechende nichtparametrische Hypothese.

Um eine Äquivalenz der Hypothesen auf keine Wechselwirkung zu zeigen, müssen zusätzliche Annahmen getroffen werden. Eine solche Voraussetzung kann sein, dass kein parametrischer und kein nichtparametrischer Gruppen- oder Zeiteffekt vorliegt. Kann also bei einem dieser Effekte angenommen werden, dass er in beiden Modellen nicht vorhanden ist, so sind die parametrischen und die nichtparametrischen Hypothesen bezüglich der Wechselwirkung äquivalent. Dies folgt mit Hilfe der Modellgleichung im parametrischen Modell und ist im folgenden Satz zusammengefasst.

Satz 4.5.1
Gilt zusätzlich zum parametrischen faktoriellen Modell eine der folgenden beiden Voraussetzungen

$$\begin{aligned} \pi_{11s} - \pi_{12s} + \pi_{21s} - \pi_{22s} &= 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1 \quad \textit{und} \\ L_{11s} - L_{12s} + L_{21s} - L_{22s} &= 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1 \end{aligned}$$

oder

$$\begin{aligned} \pi_{11s} + \pi_{12s} - \pi_{21s} - \pi_{22s} &= 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1 \quad \textit{und} \\ L_{11s} + L_{12s} - L_{21s} - L_{22s} &= 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1 \end{aligned}$$

so sind die parametrische und die nichtparametrische Hypothese bezüglich der Wechselwirkung äquivalent, das heißt:

$$L_{11s} - L_{12s} - L_{21s} + L_{22s} \;=\; 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1 \tag{4.30}$$

$$\iff \quad \pi_{11s} - \pi_{12s} - \pi_{21s} - \pi_{22s} \;=\; 0 \quad \textit{für alle} \quad s = 1, \ldots, r-1. \tag{4.31}$$

Beweis: Siehe Abschnitt B.1. □

Man könnte auch voraussetzen, dass beispielsweise der parametrische Gruppeneffekt und der nichtparametrische Zeiteffekt null sind. In diesem Fall ist die Äquivalenz (4.30) ebenfalls gegeben. Es macht allerdings mehr Sinn entsprechende parametrische und nichtparametrische Effekte zu verwenden. Im Fall einer randomisierten Studie kann man davon ausgehen, dass sowohl der parametrische als auch der nichtparametrische Gruppeneffekt nicht vorhanden ist. Bei der parametrischen Analyse sollte man in diesem Fall die Referenzkodierung verwenden. Bei ihr folgt aus $\beta_W = 0$, dass das Odds Ratio der beiden Gruppen zum ersten Zeitpunkt $\kappa_{11,21} = 1$ ist. Bei randomisierten Studien sind also die Hypothesen über die parametrischen und nichtparametrischen Wechselwirkungen äquivalent.

Als nächstes sollen die Hypothesen im Modell der nichtparametrischen Kovarianzanalyse zu den gerade betrachteten Hypothesen über die Wechselwirkungen in Bezug gesetzt werden. Dazu gehen wir von einer randomisierten Studie aus. Es gelte also $F_{11} = F_{21}$. In diesem Fall sind offensichtlich die Hypothesen bezüglich der Marginalverteilungen

$$H_0: \quad F_{11} - F_{12} - F_{21} - F_{22} \;=\; 0$$

und

$$H_0: \quad F_{12} - F_{22} \;=\; 0 \tag{4.32}$$

äquivalent. Dabei entsprechen die Marginalverteilungen zum zweiten Zeitpunkt den Verteilungen der Zielvariablen und die Marginalverteilungen zum ersten Zeitpunkt den Verteilungen der Kovariablen. Ebenso sind die Hypothesen mit Hilfe der relativen Effekte,

$$H_0: \quad p_{11} - p_{12} - p_{21} - p_{22} \;=\; 0$$

und

$$H_0: \quad p^* \;=\; \frac{1}{2} \tag{4.33}$$

äquivalent. Bei einer stochastischen Ordnung der Verteilungsfunktionen gilt auch hier die Äquivalenz zwischen den Hypothesen (4.32) und (4.33).

Nimmt man zusätzlich das parametrische faktorielle Modell an, in dem der parametrische Gruppeneffekt Null ist, so sind die Hypothesen (4.32) und (4.33) auch äquivalent zur Hypothese

$$H_0: \quad L_{11s} - L_{12s} - L_{21s} + L_{22s} \;=\; 0 \quad \text{für alle} \quad s = 1, \ldots, r-1,$$

dass keine parametrische Wechselwirkung vorliegt. Dies zeigt man analog zu dem Beweis von Satz 4.5.1.

Nach der Betrachtung der Erwartungswerte der Y_{ikjs} soll nun das Augenmerk auf die Abhängigkeitsstruktur gerichtet werden. Zwei Zufallsvariablen Y_{ikjs} und $Y_{i'k'j's'}$ sind wie die X_{ikj} unabhängig für $(i,k) \neq (i',k')$. Sie bilden für jede Versuchseinheit sogenannte Cluster von abhängigen Zufallsvariablen. Die Größe der Cluster, das heißt die Anzahl der abhängigen Zufallsvariablen, hängt im vorliegenden Versuchsplan nicht von der Versuchseinheit ab, wie es in der Literatur häufig erlaubt ist. Sie ist für alle Versuchseinheiten konstant $2(r-1)$. Jeder Cluster wird in einem Zufallsvektor der Länge $2(r-1)$ zusammengefasst:

$$\mathbf{Y}_{ik} = \left(Y_{ik11}, \ldots, Y_{ik1(r-1)}, Y_{ik21}, \ldots, Y_{ik2(r-1)}\right)^t.$$

Damit ergeben sich n unabhängige Zufallsvektoren. Die Kovarianzmatrix dieser Zufallsvektoren $\mathbf{Y}_{ik}$ wird mit $\boldsymbol{V}_i$ bezeichnet. Sie wird durch

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{A}_i^{\frac{1}{2}}$$

modelliert. Dabei ist $\boldsymbol{A}_i$ eine $2(r-1) \times 2(r-1)$-Diagonalmatrix mit den Einträgen $\pi_{ijs}(1-\pi_{ijs})$ und $\boldsymbol{R}$ ist die sogenannte *Working Correlation Matrix*. Durch sie wird die Abhängigkeit zwischen den Einträgen der Vektoren $\mathbf{Y}_{ik}$ berücksichtigt. Wie bei jeder Korrelationsmatrix sind die Diagonalelemente 1. Für die Berechnung der anderen Komponenten gibt es verschiedene Ansätze. Hier werden die fünf Verfahren beschrieben, die von der SAS-Prozedur PROC GENMOD unterstützt werden (SAS, 1999).

Zunächst ist es möglich, $\boldsymbol{R}$ als Einheitsmatrix zu wählen. Einerseits hat dieser einfache Ansatz für die Berechnung der Schätzer der Effekte und der Teststatistiken einige Vorteile. Andererseits wird die Abhängigkeitsstruktur völlig vernachlässigt. Diese wird bei den anderen Verfahren modelliert. Dafür kann dort die iterative Berechnung zu Schwierigkeiten führen. Für die Beschreibung dieser anderen vier Verfahren werden zunächst die Indices j und s, die bei der Zufallsvariablen Y_{ikjs} und deren Erwartungswerten π_{ijs} auftreten, zu einem Index $u = 1, \ldots, 2(r-1)$ zusammengefasst. Dies wird im Folgenden die Erklärungen vereinfachen. Damit bezeichnet r_{12} die Korrelation zwischen $Y_{ik11} = Y_{ik1}$ und $Y_{ik12} = Y_{ik2}$, r_{13} die Korrelation zwischen $Y_{ik11} = Y_{ik1}$ und $Y_{ik13} = Y_{ik3}$ und so weiter.

Nun kann das zweite Verfahren beschrieben werden, welches darauf beruht, alle Elemente außerhalb der Diagonalen als identisch anzusehen:

$$r_{uu'} = \delta, \quad u, u' = 1, \ldots, 2(r-1) \quad \text{mit} \quad u \neq u'.$$

Eine solche Korrelationsstruktur tritt bei Zufallsvektoren auf, deren Komponenten austauschbar sind. Gilt ferner, dass die Diagonalelemente von $\boldsymbol{A}_i$ identisch sind, so spricht man in diesem Fall auch von einer *Compound-Symmetry*-Struktur der Kovarianzmatrix $\boldsymbol{V}_i$.

Beim dritten Verfahren werden jeweils diejenigen Korrelationen von zwei Komponenten als gleich betrachtet, bei denen der Abstand der beiden Komponenten im Vektor, das heißt die Anzahl der Einträge, die zwischen den beiden Komponenten liegen, gleich ist. Ist der Abstand größer als ein vorgegebener Wert m, so wird angenommen, dass die Komponenten unkorreliert sind. Damit erhält man die folgende Struktur:

$$r_{u(u+v)} = \begin{cases} 1 & v = 0 \\ \delta_v & v = 1, \ldots, m \\ 0 & v > m \end{cases}, \quad u = 1, \ldots, 2(r-1).$$

Korrelationsmatrizen dieser Art treten zum Beispiel bei Zeitreihen auf, bei denen man davon ausgehen kann, dass ab einem bestimmten Zeitintervall die zugehörigen Zufallsvariablen unkorreliert sind.

Das vierte Verfahren beruht auf einem autoregressiven Ansatz. Die Einträge der Korrelationsmatrix $\boldsymbol{R}$ haben dann die Form

$$r_{u(u+v)} = \delta^v, \quad u = 1, \ldots, 2(r-1) - v.$$

Durch dieses Verfahren nehmen die Korrelationen immer mehr ab, je weiter sie von der Diagonalen entfernt sind. Wie im vorhergehenden Fall tauchen solche Korrelationsstrukturen zum Beispiel bei Zeitreihen auf.

Das letzte Verfahren kann als das allgemeinste Verfahren betrachtet werden, da an die Komponenten $r_{uu'}$ keine Voraussetzungen gestellt werden. Eine solche Korrelationsmatrix wird als unstrukturiert bezeichnet. Diese Korrelationsstruktur stellt zwar die geringsten Annahmen an die Kovarianzmatrix, es kommt aber auch am häufigsten zu Problemen bei der Berechnung der Schätzer und Teststatistiken. Nach Fahrmeir und Tutz (1994) sollte die Wahl der Working Correlation Matrix $\boldsymbol{R}$ daher einen Kompromiss zwischen Einfachheit und Effizienz darstellen. Außerdem muß sie davon abhängen, wie viele Versuchseinheiten verfügbar sind.

Kenward et al. (1994) schlagen hingegen vor, die besondere Struktur der Zufallsvariablen Y_{ikjs}, die durch ihre künstliche Erzeugung entsteht, zu verwenden. Aufgrund der Definition der Y_{ikjs} kann die Korrelation zwischen zwei Zufallsvariablen in der gleichen

Gruppe i zum gleichen Zeitpunkt j direkt bestimmt werden. Für $s < s'$ gilt

$$\mathrm{Corr}(Y_{ikjs}, Y_{ikjs'}) = \frac{\pi_{ijs}(1-\pi_{ijs'})}{\sqrt{\pi_{ijs}(1-\pi_{ijs})\pi_{ijs'}(1-\pi_{ijs'})}}.$$

Da dieser gute Ansatz jedoch nicht in SAS implementiert ist, wird er nicht weiter betrachtet. Letztendlich ist von Fall zu Fall zu entscheiden, welche Korrelationsstruktur die *beste* ist.

Diese numerischen Probleme sind ein wichtiger Nachteil gegenüber der nichtparametrischen Kovarianzanalyse. Bei dieser hängen die Ergebnisse außerdem nicht davon ab, welche Korrelationsmatrix oder Link-Funktion man wählt. Gerade bei kleinen Stichprobenumfängen können die Ergebnisse sehr von dieser Wahl abhängen.

4.6 Parametrische Kovarianzanalyse

Nachdem im vorhergehenden Abschnitt ein parametrisches Modell mit der Zeit als Faktor betrachtet wurde, wird nun eine parametrische Kovarianzanalyse vorgestellt. Bei dieser werden wie in der nichtparametrischen Kovarianzanalyse die Zufallsvariablen zum ersten Zeitpunkt als Kovariablen der Zufallsvariablen zum zweiten Zeitpunkt betrachtet. Die Abhängigkeit zwischen den Zufallsvariablen, die an einer Versuchseinheit beobachtet werden, wird durch eine Regressionsgleichung berücksichtigt. Somit wird das Modell eingebettet in die Theorie der GLM (McCullagh und Nelder, 1989; Fahrmeir und Tutz, 1994) vorgestellt.

Wie im vorhergehenden Kapitel ist es wichtig, dass alle Zufallsvariablen auf einer ordinalen Skala mit den Kategorien $1, \ldots, r$ beobachtet werden. Um die Theorie der GLM verwenden zu können, müssen diese Zufallsvariablen kodiert werden. Es ergeben sich neue Zufallsvariablen Z_{ikjs} auf die folgende Weise:

$$Z_{ikjs} = \begin{cases} 1 & \text{falls } X_{ikj} = s \\ 0 & \text{falls } X_{ikj} \neq s \end{cases}, \; i,j = 1,2, \; k = 1, \ldots, n_i, \; s = 1, \ldots, r-1. \tag{4.34}$$

Die hier verwendete Kodierung soll nicht verwechselt werden mit der Kodierung, welche im vorhergehenden Abschnitt verwendet wurde. Daher wird hier der Buchstabe Z für die Zufallsvariablen verwendet.

Wie im vorhergehenden Abschnitt soll

$$\mathrm{P}(X_{ikj} = s) > 0, \quad i,j = 1,2, \quad s = 1, \ldots, r,$$

also die Annahme (4.28) von Seite 41, gelten. Damit wird gesichert, dass die Erwartungswerte und Varianzen aller Z_{ikjs} positiv sind. Die Verteilung der Zielvariablen wird bedingt auf die Kovariablen betrachtet. Damit interessiert die Wahrscheinlichkeit, bei der Zielvariablen X_{ik2} die Kategorie s zu beobachten unter der Bedingung, dass die Kovariablen X_{ik1} den Wert s' hat. Diese Wahrscheinlichkeit wird mit $\pi_{is}(s')$ bezeichnet. Um die Notation zu vereinfachen, wird das Argument der Wahrscheinlichkeiten häufig nicht erwähnt. Es handelt sich aber immer um bedingte Wahrscheinlichkeiten. Diese werden für jede Gruppe i in einem Vektor $\boldsymbol{\pi}_i = \left(\pi_{i1}, \ldots, \pi_{i(r-1)}\right)^t$ angeordnet. Für diese Vektoren werden die Regressionsgleichungen

$$g_s(\boldsymbol{\pi}_i(s')) \quad = \quad \alpha_{2s} + \mathbf{1}_{(i=2)}\, \beta_{kov} + \alpha_{1s'}, \quad i = 1, 2, \tag{4.35}$$

definiert. Dabei ist g_s die s-te Komponente der Link-Funktion

$$g(\boldsymbol{\pi}_i) \quad = \quad \begin{pmatrix} \log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right) \\ \vdots \\ \log\left(\frac{\pi_{i1}+\ldots+\pi_{is}}{1-\pi_{i1}-\ldots-\pi_{is}}\right) \\ \vdots \\ \log\left(\frac{\pi_{i1}+\ldots+\pi_{i(r-1)}}{1-\pi_{i1}-\ldots-\pi_{i(r-1)}}\right) \end{pmatrix}$$

und $\mathbf{1}_{(i=2)}$ ist eine Indikatorfunktion, die für die erste Gruppe null und für die zweite Gruppe eins wird. Die Link-Funktion, die auch kumulativer logit-Link genannt wird, ist nicht die natürliche Link-Funktion, wie dies bei der parametrischen faktoriellen Analyse der Fall war. Da die Z_{ikjs} als Vektoren zusammengefasst multinomial verteilt sind, hat die natürliche Link-Funktion in der Gruppe i die Komponenten

$$\theta_{is}(\boldsymbol{\pi}_i) \quad = \quad \log\left(\frac{\pi_{is}}{1 - \pi_{i1} - \ldots - \pi_{i(r-1)}}\right), \quad s = 1, \ldots, r-1.$$

Diese Funktion berücksichtigt jedoch nicht die Anordnung der Kategorien, wie es beim kumulativen logit-Link der Fall ist. Anstatt des kumulativen logit-Links wären wie im vorhergehenden Kapitel auch noch andere Links, wie

$$\Phi^{-1}\left(\pi_{k1} + \ldots + \pi_{ks}\right) \quad \text{oder} \quad \log\left(-\log\left(1 - \pi_{k1} - \ldots - \pi_{ks}\right)\right)$$

möglich. Zum einen unterscheiden sie sich aber nicht besonders vom kumulativen logit-Link und zum anderen ergeben sich für diesen einige Vereinfachungen bei den Effekten. Daher wird im Folgenden ausschließlich der kumulative logit-Link betrachtet.

Nach der Betrachtung der Link-Funktion soll nun näher auf die Summanden des linearen Prädiktors

$$\eta_{iss'} \quad = \quad \alpha_{2s} + \mathbf{1}_{(i=2)}\, \beta_{kov} + \alpha_{1s'}$$

eingegangen werden. Für die Kodierung der Gruppen wurde hier die Referenzkodierung mit der ersten Gruppe als Referenz gewählt. Sie ist in diesem Fall äquivalent zur Effektkodierung, bei der man $(\mathbf{1}_{(i=2)} - \mathbf{1}_{(i=1)})$ als Faktor vor β_{kov} wählen würde (Oelerich, 1998).

Die Cutpoints α_{2s} bestimmen den Einfluss der Kategorie der Zielvariablen. Die α_{1s} sind hingegen Regressionsparameter der Kovariablen. Da die Abstände zwischen den r Kategorien nicht gleich sind, müssen $r-1$ Parameter verwendet werden, um diese zu definieren. Sie fungieren als Regressionsparameter. Ansonsten werden keine weiteren Bedingungen an sie gestellt. Damit unterscheiden sie sich von den α_{2s}, denn für diese gilt auf Grund der Gleichungen (4.28) von Seite 41 und (4.35) von Seite 53

$$\alpha_{21} \;<\; \ldots \;<\; \alpha_{2r-1}. \tag{4.36}$$

Wie im vorhergehenden Abschnitt sind sie angeordnet und können wiederum mit dem Threshold-Ansatz interpretiert werden (Edwards und Thurstone, 1952).

Für die Kovariablen wurden hier die $r-1$ Parameter α_{1s} gewählt. Häufig wird nur ein Parameter α_1 verwendet und der Term $\alpha_1 x$ bei einem Wert x der Kovariablen X_{ik1} in die Modellgleichung aufgenommen. Ein solcher Ansatz geht davon aus, dass die Abstände zwischen den Kategorien der Skala, auf der die Kovariablen beobachtet werden, immer gleich sind. Da dies bei einer ordinalen Skala nicht der Fall ist, wird auf die Betrachtung eines solchen Modelles, welches nicht invariant unter streng monotonen Transformationen der Skala ist, verzichtet.

Die Interpretation der Parameter α_{js} und β_{kov} ist mit Hilfe der Odds Ratios

$$\frac{(\pi_{11}(s') + \ldots + \pi_{1s}(s'))(1 - \pi_{21}(s'') - \ldots - \pi_{2s}(s''))}{(1 - \pi_{11}(s') - \ldots - \pi_{1s}(s'))(\pi_{21}(s'') + \ldots + \pi_{2s}(s''))}, \quad s, s', s'' = 1, \ldots, r-1,$$

möglich. Werden in beiden Gruppen die gleichen Werte der Kovariablen betrachtet, dass heißt $s' = s''$, so hängen die Odds Ratios nicht von der Kategorie s der Zielvariablen oder den Kategorien s' der Kovariablen ab. Aufgrund der Regressionsgleichungen (4.35) von Seite 53 gilt

$$\kappa \quad = \quad \frac{(\pi_{11}(s') + \ldots + \pi_{1s}(s'))(1 - \pi_{21}(s') - \ldots - \pi_{2s}(s'))}{(1 - \pi_{11}(s') - \ldots - \pi_{1s}(s'))(\pi_{21}(s') + \ldots + \pi_{2s}(s'))} \quad = \quad \exp(-\beta_{kov}).$$

Damit ist κ das Verhältnis der Odds in den beiden Gruppen einen kleineren als einen vorgegebenen Wert zu beobachten. Aufgrund dieser Beziehung $\beta_{kov} = -\log(\kappa)$ wird β_{kov} auch log Odds Ratio bezeichnet. Mit Hilfe dieses Parameters kann die Nullhypothese, dass beide Behandlungen der Gruppen identisch sind, gestellt werden:

$$H_0: \quad \beta_{kov} = 0.$$

Diese Hypothese ist äquivalent zu

$$H_0: \quad \kappa = 1.$$

Nun soll betrachtet werden, was mit den Odds Ratios geschieht, wenn unterschiedliche Kategorien in der Ziel- oder Kovariablen betrachtet werden. Im ersten Fall wird bei einer Versuchseinheit aus der ersten Gruppe weiterhin die Kategorie s in der Zielvariablen beobachtet, während bei einer Versuchseinheit aus der zweiten Gruppe die Zielvariable den Wert s' annimmt. Die Kovariablen sollen für beide Versuchseinheiten den gleichen Wert annehmen. Die Argumente der Wahrscheinlichkeiten wurden daher vernachlässigt. Dann gilt

$$\frac{(\pi_{11} + \ldots + \pi_{1s})(1 - \pi_{21} - \ldots - \pi_{2s'})}{(1 - \pi_{11} - \ldots - \pi_{1s})(\pi_{21} + \ldots + \pi_{2s'})} = \exp\left(-\beta_{kov} + \alpha_{2s} - \alpha_{2s'}\right).$$

Ist nun $s > s'$, so steigt das Verhältnis wegen der Positivität der Wahrscheinlichkeiten an. Dies wird auch an der wegen der Ungleichungen (4.36) von Seite 54 positiven Differenz $\alpha_{2s} - \alpha_{2s'}$ deutlich.

Im zweiten Fall nehmen die Zielvariablen der betrachteten Versuchseinheiten aus den beiden Gruppen den gleichen Wert k an, während die Kovariablen den Wert s beziehungsweise s' annehmen. Wieder wird das Verhältnis

$$\begin{aligned} &\frac{(\pi_{11}(s) + \ldots + \pi_{1k}(s))(1 - \pi_{21}(s') - \ldots - \pi_{2k}(s'))}{(1 - \pi_{11}(s) - \ldots - \pi_{1k}(s))(\pi_{21}(s')) + \ldots + \pi_{2k}(s')} \\ &= \exp(\beta_{kov} + \alpha_{1s} - \alpha_{1s'}) \end{aligned}$$

betrachtet. An die Regressionsparameter α_{1s} wurden jedoch keine Bedingungen gestellt. Damit kann $\alpha_{1s} < \alpha_{1s'}$ auch bei $s > s'$ gelten. Es wird also kein monotoner Zusammenhang zwischen Ziel- und Kovariablen gefordert. Mithin kann also auch nicht gesagt werden, ob das Verhältnis in der Gleichung (4.37) größer oder kleiner als $\exp(-\beta_{kov})$ geworden ist.

Die Annahmen der parametrischen Kovarianzanalyse sind im Gegensatz zur nichtparametrischen Kovarianzanalyse sehr restriktiv. So wird beispielsweise vorausgesetzt, dass

für jeden Wert der Kovariablen die bedingten Verteilungen der beiden Gruppen einer bestimmten stochastischen Ordnung unterliegen. Diese Ordnung wird durch die Parameter α_{2s} bestimmt. Für unterschiedliche Werte s und s' der Kovariablen verschieben sich diese Parameter um den gleichen Wert $\alpha_{1s} - \alpha_{1s'}$. Durch diese starken Annahmen wird die Menge der möglichen Verteilungsfunktionen erheblich eingeschränkt. Daher ergibt sich hier für die nichtparametrische Kovarianzanalyse ein großer Vorteil, da sie nur sehr schwache Voraussetzungen stellt. Mithin ist die Menge der möglichen Verteilungsfunktionen wesentlich größer als bei der parametrischen Kovarianzanalyse. Dies kann in der Praxis dazu führen, dass die Interpretation der Auswertungsergebnisse bestimmter Daten bei der parametrischen Kovarianzanalyse fragwürdig ist.

Kapitel 5

Test- und Schätzverfahren der nichtparametrischen Kovarianzanalyse

5.1 Randomisierte Versuchsanlage

Nachdem die Modelle, die Effekte und die Hypothesen vorgestellt und verglichen wurden, wenden wir uns nun den Schätzern der Effekte und den Teststatistiken für die Tests der Hypothesen zu. Da die nichtparametrische Kovarianzanalyse im Zentrum dieser Arbeit steht, wird für sie die Herleitung der Schätzer und ihrer asymptotischen Verteilung im Gegensatz zu den anderen Verfahren ausführlich behandelt. Dabei gehen wir in diesem Abschnitt von einer randomisierten Studie aus. Es gilt also die Annahme $F_{11} = F_{21}$. Im nächsten Abschnitt werden dann die Änderungen erläutert, die man für nicht randomisierte Studien braucht.

Die grundlegenden Effekte in der nichtparametrischen Kovarianzanalyse sind die relativen Effekte $p_j = \int F_{1j} dF_{2j}$ zu den beiden Zeitpunkten. Diese werden geschätzt, indem die Marginalverteilungen F_{ij} durch die empirischen Marginalverteilungen $\widehat{F}_{ij}$ ersetzt werden. Mithin erhält man

$$\widehat{p}_j \quad = \quad \int \widehat{F}_{1j}\, d\widehat{F}_{2j}, \quad j = 1, 2.$$

Dabei können die empirischen Marginalverteilungen mit Hilfe der Zählfunktion c bestimmt werden:

$$\widehat{F}_{ij}(x) \quad = \quad \frac{1}{n_i} \sum_{i=1}^{n_i} c(x, X_{ikj}), \quad i, j = 1, 2.$$

Wie man leicht nachweisen kann, sind die Schätzer $\widehat{p}_j$ erwartungstreu und konsistent bezüglich der L_2-Norm:

$$\begin{aligned} \mathrm{E}\left(\widehat{p}_j\right) &= p_j, \quad j = 1,2 \quad \text{und} \\ \mathrm{E}\left(\widehat{p}_j - p_j\right)^2 &\overset{n\to\infty}{\longrightarrow} 0, \quad j = 1,2. \end{aligned}$$

Für die praktische Berechnung der Schätzer der relativen Effekte ist die folgende Darstellung mit Hilfe von Rängen hilfreich, denn Ränge können durch gute Sortieralgorithmen sehr schnell und einfach berechnet werden:

$$\widehat{p}_j = \frac{1}{n_1}\left(\overline{R}_{2\cdot j(j)} - \frac{n_2+1}{2}\right), \quad j = 1,2.$$

Dabei ist $\overline{R}_{2\cdot j(j)}$ das Mittel aller Ränge $R_{2kj(j)}$, die über die Beobachtungen zum Zeitpunkt j gebildet wurden. Die Ränge werden also für die Ziel- und die Kovariable getrennt vergeben. Dies ist in der Kovarianzanalyse natürlich, da im Allgemeinen die Ziel- und die Kovariablen nicht vergleichbar sein müssen.

Damit ergibt sich ein Schätzer $\widehat{p}_2$ für den interessierenden relativen Effekt p_2. Bei diesem werden die Zufallsvariablen zum ersten Zeitpunkt nicht berücksichtigt. Um einen verbesserten Schätzer zu erhalten, wird daher auch der Effekt p_1 geschätzt, obwohl dieser in randomisierten Studien bekannt ist. Aufgrund der Annahme $F_{11} = F_{21}$ gilt nämlich $p_1 = 1/2$. Wir betrachten nun den Schätzer

$$\widehat{p}^*(\gamma_{nkov}) = \widehat{p}_2 - \gamma_{nkov}(\widehat{p}_1 - \frac{1}{2}),$$

welcher erstmals von Siemer (1999) vorgestellt wurde. Allerdings ist er kein Schätzer im streng mathematischen Sinn, da γ_{nkov} unbekannt ist. Dies kennzeichnen wir durch die Tilde anstatt des Daches in der Notation. Der Parameter γ_{nkov} wurde von Siemer (1999) mit Hilfe eines Regressionsansatzes bestimmt. Dieser entsprach dem Ansatz von Langer (1998). Zur Erläuterung dieser beiden Verfahren werden die folgenden nicht beobachtbaren Zufallsvariablen verwendet:

$$\begin{aligned} Y_{n,ikj(j)} &= \frac{1}{n}\left(n_1 F_{1j}(X_{ikj}) + n_2 F_{2j}(X_{ikj})\right), \quad i,j = 1,2, \quad k = 1,\ldots,n_i, \\ Y_{ikj(j)} &= \frac{1}{2}\left(F_{1j}(X_{ikj}) + F_{2j}(X_{ikj})\right), \quad i,j = 1,2, \quad k = 1,\ldots,n_i. \end{aligned}$$

Diese Zufallsvariablen werden auch als gewichtete beziehungsweise ungewichtete asymptotische Rangtransformation (ART) bezeichnet. Ersetzt man die Verteilungsfunktionen durch ihre Schätzer, die empirischen Verteilungsfunktionen, so erhält man

$$\widehat{Y}_{n,ikj(j)} = \frac{1}{n}\left(n_1 \widehat{F}_{1j}(X_{ikj}) + n_2 \widehat{F}_{2j}(X_{ikj})\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\left(R_{ikj(j)} - \frac{1}{2}\right) \quad \text{und} \\
\widehat{Y}_{ikj(j)} &= \frac{1}{2}\left(\widehat{F}_{1j}(X_{ikj}) + \widehat{F}_{2j}(X_{ikj})\right) \\
&= \frac{1}{2n_{i'}}\left(R_{ikj(j)} - \frac{1}{2}\right) + \frac{n_{i'} - n_i}{2n_1 n_2}\left(R_{ikj(ij)} - \frac{1}{2}\right) \quad \text{mit} \quad i \neq i'.
\end{aligned}
$$

Dabei ist $R_{ikj(ij)}$ der Rang von X_{ikj} unter allen Zufallsvariablen der Gruppe i zum Zeitpunkt j. Dieser wird auch als Internrang von X_{ikj} bezeichnet.

Um die asymptotische Verteilung der Schätzer zu bestimmen, betrachtet man nicht die Ränge sondern die Zufallsvariablen $Y_{n,ikj(j)}$ und $Y_{ikj(j)}$, weil diese im Gegensatz zu den Rängen die gleiche Abhängigkeitsstruktur aufweisen wie die ursprünglichen Zufallsvariablen X_{ikj}. Andererseits sind sie nicht beobachtbar, da die Verteilungsfunktionen F_{ij} nicht bekannt sind. Die Regressionsgleichungen werden nun mit diesen nicht beobachtbaren Zufallsvariablen aufgestellt:

$$
Y_{n,ik2(2)} = \gamma_{nkov,nR}\left(Y_{n,ik1(1)} - \frac{1}{2}\right) + Y^*_{n,ik}, \quad i = 1,2, \quad k = 1,\ldots,n_i, \quad \text{und} \tag{5.1}
$$

$$
Y_{ik2(2)} = \gamma_{nkov,R}\left(Y_{ik1(1)} - \frac{1}{2}\right) + Y^*_{ik}, \quad i = 1,2, \quad k = 1,\ldots,n_i. \tag{5.2}
$$

Der Index R für die Regressionskoeffizienten soll darin erinnern, dass diese durch die Regressionsgleichungen definiert wurden. Zusätzlich wird bei dem ersten Parameter der Index n verwendet, um die gewichtete Version zu kennzeichnen. In beiden Fällen werden in den jeweils n Regressionsgleichungen n neue Zufallsvariablen, nämlich die $Y^*_{n,ik}$ beziehungsweise Y^*_{ik} definiert. Zusammen mit dem Regressionskoeffizienten erhält man daher $n+1$ neue Größen bei n Gleichungen. Für eine eindeutige Definition müssen daher weitere Bedingungen gestellt werden:

$$
\mathrm{E}\left(Y^*_{n,ik} - \overline{Y}^*_{n,i\cdot} | \boldsymbol{Y}_{n,i1} = \boldsymbol{y}_{n,i1}\right) = 0 \quad \text{und} \tag{5.3}
$$

$$
\mathrm{E}\left(Y^*_{ik} - \overline{Y}^*_{i\cdot} | \boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}\right) = 0. \tag{5.4}
$$

Dabei gelten die Bezeichnungen

$$
\boldsymbol{Y}_{n,i1} = \left(Y^{(1)}_{n,i1}, \ldots Y^{(1)}_{n,in_i}\right)^t \quad \text{und} \quad \boldsymbol{Y}_{i1} = \left(Y^{(1)}_{i1}, \ldots Y^{(1)}_{in_i}\right)^t.
$$

Des Weiteren ist $\overline{Y}^*_{n,i\cdot}$ das Mittel der $Y^*_{n,ik}$ und entsprechend $\overline{Y}^*_{i\cdot}$ das Mittel der Y^*_{ik}. Die Bedingungen (5.3) und (5.4) werden von Langer (1998) mit (R2) bezeichnet. Sie besagen,

dass für jede Realisation der transformierten Kovariablen die Abweichungen der transformierten Zielvariablen um die Regressionsgerade in der Weise streuen, dass die Erwartungswerte der Zielvariablen nur durch die fest vorgegebenen Kovariablen bestimmt sind. Dies ist nach Langer (1998) eine Annahme, die häufig in Regressionsmodellen getroffen wird.

Um die Annahme von Regressionsgleichungen zu umgehen, wurde von Domhof (2001) eine neue Idee zur Bestimmung von γ_{nkov} vorgestellt. Sie besteht darin, γ_{nkov} derart zu wählen, dass die Varianz von $\widetilde{p}^*(\gamma_{nkov})$ minimal wird. Um den Parameter γ_{nkov} bei diesem Ansatz von den Parametern der Regressionsgleichungen unterscheiden zu können, wird von nun an der zusätzliche Index V verwendet. Dieser soll bei $\gamma_{nkov,V}$ darin erinnern, dass der Parameter durch die Varianzminimierung und nicht mit Hilfe von Regressionsgleichungen bestimmt ist. Für die Definition von $\gamma_{nkov,V}$ zerlegen wir die asymptotische Varianz folgendermaßen:

$$\begin{aligned}\lim_{n\to\infty} \mathrm{Var}\left(\sqrt{n}\widetilde{p}^*(\gamma_{nkov,V})\right) &= \lim_{n\to\infty} \left[\mathrm{Var}\left(\sqrt{n}\widehat{p}_2\right) + \gamma^2_{nkov,V}\mathrm{Var}\left(\sqrt{n}\widehat{p}_1\right)\right. \\ &\qquad \left. -2\gamma_{nkov,V}\mathrm{Cov}\left(\sqrt{n}\widehat{p}_2, \sqrt{n}\widehat{p}_1\right)\right].\end{aligned}$$

Sie wird also minimal, falls man

$$\gamma_{nkov,V} = \lim_{n\to\infty} \frac{\mathrm{Cov}\left(\sqrt{n}\widehat{p}_2, \sqrt{n}\widehat{p}_1\right)}{\mathrm{Var}\left(\sqrt{n}\widehat{p}_1\right)} \tag{5.5}$$

wählt. Diese Gleichung definiere daher $\gamma_{nkov,V}$. Nun müssen die asymptotischen Varianzen und Kovarianzen der Schätzer bestimmt werden. Diese ergeben sich aus dem nächsten Satz, für den zunächst einige Abkürzungen definiert werden. Es gelte:

$$\begin{aligned} Z_{ikj} &= F_{i'j}(X_{ikj}), \quad i \neq i', \quad i, i', j = 1, 2, \quad k = 1, \ldots, n_i, \quad \text{und} \\ Z^*_{ik} &= Z_{ik2} - \gamma_{nkov,V} Z_{ik1}, \quad i = 1, 2, \quad k = 1, \ldots, n_i. \end{aligned}$$

Nun wird die asymptotische Verteilung der Schätzer formuliert.

Satz 5.1.1

Unter den Annahmen (4.2), (4.3), (4.4) und (4.5) ist

$$\sqrt{n}\frac{\widetilde{p}^*(\gamma_{nkov,V}) - p_2}{\sqrt{\sigma^p_{nkov}}}$$

im Fall einer randomisierten Versuchsanlage asymptotisch standardnormalverteilt. Dabei bezeichnet σ^p_{nkov} die Varianz

$$\sigma^p_{nkov} = \frac{n_1}{n} Var(Z^*_{11}) + \frac{n_2}{n} Var(Z^*_{21}).$$

Beweis: Siehe B.2. □

Um die Varianz σ^p_{nkov} zu schätzen, wird sie zunächst additiv in kleinere Teile zerlegt, die dann einzeln betrachtet werden. Mithin erhält man

$$\begin{aligned}\sigma^p_{nkov} &= \frac{n}{n_1}\left(\sigma^p_{nkov,1,22} + \gamma^2_{nkov,V}\sigma^p_{nkov,1,11} - 2\gamma_{nkov,V}\sigma^p_{nkov,1,21}\right) + \\ &\quad \frac{n}{n_2}\left(\sigma^p_{n,kov,2,22} + \gamma^2_{nkov,V}\sigma^p_{nkov,2,11} - 2\gamma_{nkov,V}\sigma^p_{nkov,2,21}\right) \qquad (5.6)\end{aligned}$$

mit den Bezeichnungen

$$\begin{aligned}\sigma^p_{nkov,i,jj} &= \mathrm{Var}\left(Z_{i1j}, Z_{i1j}\right), \quad i,j = 1,2 \quad \text{und} \\ \sigma^p_{nkov,i,21} &= \mathrm{Cov}\left(Z_{i12}, Z_{i11}\right), \quad i = 1,2.\end{aligned}$$

Diese Varianzen und Kovarianzen können nun kanonisch und konsistent geschätzt werden. Dazu werden die Verteilungsfunktionen als Transformationen der Zufallsvariablen durch die empirischen Verteilungsfunktionen ersetzt. Es gelten die Abkürzungen:

$$\widehat{Z}_{ikj} = \widehat{F}_{i'j}\left(X_{ikj}\right), \quad i \neq i', \quad i,i',j = 1,2, \quad k = 1,\ldots,n_i.$$

Diese Zufallsvariablen können auch mit Hilfe von Rängen darstellt werden.

$$\widehat{Z}_{ikj} = \frac{1}{n_{i'}}\left(R_{ikj(j)} - R_{ikj(ij)}\right), \quad i \neq i', \quad i,i',j = 1,2, \quad k = 1,\ldots,n_i.$$

Damit können nun die kanonischen Schätzer mit ihren Eigenschaften angegeben werden.

Satz 5.1.2
Unter den Annahmen (4.2), (4.3), (4.4) und (4.5) sind die Schätzer

$$\begin{aligned}\widehat{\sigma}^p_{nkov,i,jj} &= \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Z}_{ikj} - \widehat{\overline{Z}}_{i\cdot j}\right)^2 \\ &= \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\frac{1}{n_{i'}}\left(R_{ikj(j)} - R_{ikj(ij)} - \overline{R}_{i\cdot j(j)} + \frac{n_i+1}{2}\right)\right)^2.\end{aligned}$$

für die entsprechenden Varianzen und die Schätzer

$$\widehat{\sigma}^p_{nkov,i,21} = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Z}_{ik2} - \widehat{\overline{Z}}_{i\cdot 2}\right)\left(\widehat{Z}_{ik1} - \widehat{\overline{Z}}_{i\cdot 1}\right)$$

$$= \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \frac{1}{n_{i'}^2} \left(R_{ik2} - R_{ik2(i2)} - \overline{R}_{i\cdot 2} + \frac{n_i + 1}{2} \right) \left(R_{ik1} - R_{ik1(i1)} - \overline{R}_{i\cdot 1} + \frac{n_i + 1}{2} \right).$$

mit $i \neq i'$ für die entsprechenden Kovarianzen L_2-konsistent.

Beweis: Unter mehrfacher Verwendung der Lemmata A.2, A.3 und A.4 in Siemer (1999) ergeben sich die Beweise. □

Nun wurden sowohl für die relativen Effekte als auch für die einzelnen Varianzen und Kovarianzen Schätzer angegeben. Nur der Regressionsparameter $\gamma_{nkov,V}$ muss noch geschätzt werden. Ersetzt man nun die Varianzen und Kovarianzen in der Definition von $\gamma_{nkov,V}$ durch ihre Schätzer, so erhält man den Schätzer $\widehat{\gamma}_{nkov,V}$ für den Regressionskoeffizienten:

$$\widehat{\gamma}_{nkov,V} = \frac{\widehat{\sigma}^p_{nkov,1,21}/n_1 + \widehat{\sigma}^p_{nkov,2,21}/n_2}{\widehat{\sigma}^p_{nkov,1,11}/n_1 + \widehat{\sigma}^p_{nkov,2,11}/n_2}.$$

Sind die Schätzer der Kovarianzen von null weg beschränkt, so folgt aus der Konsistenz der Schätzer der Varianzen und Kovarianzen die Konsistenz von $\widehat{\gamma}_{nkov,V}$.

Werden die Schätzer der einzelnen Varianzen und Kovarianzen zusammengesetzt, so erhält man den Schätzer der Varianz für die Teststatistik:

$$\widehat{\sigma}^p_{nkov} = \frac{n}{n_1} \left(\widehat{\sigma}^p_{nkov,1,22} + \widehat{\gamma}^2_{nkov} \widehat{\sigma}^p_{nkov,1,11} - 2\widehat{\gamma}_{nkov} \widehat{\sigma}^p_{nkov,1,21} \right) + \frac{n}{n_2} \left(\widehat{\sigma}^p_{nkov,2,22} + \widehat{\gamma}^2_{nkov} \widehat{\sigma}^p_{nkov,2,11} - 2\widehat{\gamma}_{nkov} \widehat{\sigma}^p_{nkov,2,21} \right).$$

Des Weiteren ergibt sich mit dem Schätzer für $\gamma_{nkov,V}$ nun der folgende verbesserte Schätzer von p_2 im randomisierten Fall:

$$\widehat{p}^*(\gamma_{nkov,V}) = \widehat{p}_2 - \widehat{\gamma}_{nkov,V} \left(\widehat{p}_1 - \frac{1}{2} \right).$$

Nun wurden für alle Parameter, die für die Teststatistik benötigt werden, konsistente Schätzer vorgestellt. Damit erhalten wir die asymptotische Normalität der Teststatistik, welche im folgenden Satz formuliert wird.

Satz 5.1.3

Unter den Annahmen (4.2), (4.3), (4.4) und (4.5) und der Hypothese $p_2 = 1/2$ ist die Teststatistik

$$T^p_{nkov} = \sqrt{n} \frac{\widehat{p}^*(\gamma_{nkov,V}) - 1/2}{\sqrt{\widehat{\sigma}^p_{nkov}}}$$

bei randomisierten Studien asymptotisch standardnormalverteilt.

Ein ähnliches Ergebniss kann man auch mittels der Regressionsgleichungen zeigen. Dazu stellen wir zunächst die Schätzer von $\gamma_{nkov,nR}$ und $\gamma_{nkov,R}$ vor. Die kleinste-Quadrate-Schätzer sind

$$\begin{aligned}
\widehat{\gamma}_{nkov,nR} &= \frac{\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(\widehat{Y}_{n,ik2(2)} - \widehat{\overline{Y}}_{n,i\cdot 2(2)}\right)\left(\widehat{Y}_{n,ik1(1)} - \widehat{\overline{Y}}_{n,i\cdot 1(1)}\right)}{\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(\widehat{Y}_{n,ik1(1)} - \widehat{\overline{Y}}_{n,i\cdot 1(1)}\right)^2} \quad \text{und} \\
\widehat{\gamma}_{nkov,R} &= \frac{\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(\widehat{Y}_{ik2(2)} - \widehat{\overline{Y}}_{i\cdot 2(2)}\right)\left(\widehat{Y}_{ik1(1)} - \widehat{\overline{Y}}_{i\cdot 1(1)}\right)}{\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(\widehat{Y}_{ik1(1)} - \widehat{\overline{Y}}_{i\cdot 1(1)}\right)^2}.
\end{aligned}$$

Langer (1998) zeigt in Theorem 4.8 die Konvergenz des ersten Schätzers in Wahrscheinlichkeit. Die Konvergenz des zweiten Schätzers ergibt sich jedoch genauso, da die gewichteten mittleren Verteilungsfunktionen einfach durch die ungewichteten ersetzt werden müssen.

Die Unterschiede der beiden Schätzer $\widehat{\gamma}_{nkov,nR}$ und $\widehat{\gamma}_{nkov,R}$ zu dem Schätzer $\widehat{\gamma}_{nkov,V}$ werden deutlich, wenn $\widehat{\gamma}_{nkov,nR}$ und $\widehat{\gamma}_{nkov,R}$ mit Hilfe von Varianz- und Kovarianzschätzern ausgedrückt werden. Dazu seien

$$\begin{aligned}
\widehat{\sigma}^F_{nkov,n,i,jj} &= \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Y}_{n,ikj(j)} - \widehat{\overline{Y}}_{n,i\cdot j(j)}\right)^2 \\
&= \frac{1}{(n_i-1)n^2}\sum_{k=1}^{n_i}\left(R_{ikj(j)} - \overline{R}_{i\cdot j(j)}\right)^2 \quad \text{und} \\
\widehat{\sigma}^F_{nkov,i,jj} &= \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Y}_{ikj(j)} - \widehat{\overline{Y}}_{i\cdot j(j)}\right)^2 \\
&= \frac{1}{4(n_i-1)n_{i'}^2}\sum_{k=1}^{n_i}\left(R_{ikj(j)} - \overline{R}_{i\cdot j(j)} + \frac{n_{i'} - n_i}{n_i}\left(R_{ikj(ij)} - \frac{n_i+1}{2}\right)\right)^2
\end{aligned}$$

mit $i \neq i'$ die Schätzer der Varianzen

$$\begin{aligned}
\sigma^F_{nkov,n,i,jj} &= \operatorname{Var}\left(Y_{n,i1j(j)}\right), \quad i,j = 1,2, \quad \text{und} \\
\sigma^F_{nkov,i,jj} &= \operatorname{Var}\left(Y_{i1j(j)}\right), \quad i,j = 1,2,
\end{aligned}$$

und es seien

$$\widehat{\sigma}^F_{nkov,n,i,21} = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\widehat{Y}_{n,ik2(2)} - \widehat{\overline{Y}}_{n,i\cdot 2(2)}\right)\left(\widehat{Y}_{n,ik1(1)} - \widehat{\overline{Y}}_{n,i\cdot 1(1)}\right)$$

$$
\begin{aligned}
&= \frac{1}{(n_i-1)n^2}\sum_{k=1}^{n_i}\left(R_{ik2(2)}-\overline{R}_{i\cdot 2(2)}\right)\left(R_{ik1(1)}-\overline{R}_{i\cdot 1(1)}\right) \quad \text{und} \\
\widehat{\sigma}^F_{nkov,i,21} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(\widehat{Y}_{ik2(2)}-\widehat{\overline{Y}}_{i\cdot 2(2)}\right)\left(\widehat{Y}_{ik1(1)}-\widehat{\overline{Y}}_{i\cdot 1(1)}\right) \\
&= \frac{1}{4(n_i-1)n_{i'}^2}\sum_{k=1}^{n_i}\left(R_{ik2(2)}-\overline{R}_{i\cdot 2(2)}+\frac{n_{i'}-n_i}{n_i}\left(R_{ik2(i2)}-\frac{n_i+1}{2}\right)\right) \\
&\qquad \left(R_{ik1(1)}-\overline{R}_{i\cdot 1(1)}+\frac{n_{i'}-n_i}{n_i}\left(R_{ik1(i1)}-\frac{n_i+1}{2}\right)\right)
\end{aligned}
$$

mit $i \neq i'$ die Schätzer der Kovarianzen

$$
\begin{aligned}
\sigma^F_{nkov,n,i,21} &= \operatorname{Cov}\left(Y_{n,i12(2)};Y_{n,i11(1)}\right), \quad i=1,2, \quad \text{und} \\
\sigma^F_{nkov,i,21} &= \operatorname{Cov}\left(Y_{i12(2)};Y_{i11(1)}\right), \quad i=1,2.
\end{aligned}
$$

Dann erhält man die folgende Darstellung der Schätzer der Regressionskoeffizienten:

$$
\begin{aligned}
\widehat{\gamma}_{nkov,nR} &= \frac{(n_1-1)\widehat{\sigma}^F_{nkov,n,1,21}+(n_2-1)\widehat{\sigma}^F_{nkov,n,2,21}}{(n_1-1)\widehat{\sigma}^F_{nkov,n,1,11}+(n_2-1)\widehat{\sigma}^F_{nkov,n,2,11}} \\
\widehat{\gamma}_{nkov,R} &= \frac{(n_1-1)\widehat{\sigma}^F_{nkov,1,21}+(n_2-1)\widehat{\sigma}^F_{nkov,2,21}}{(n_1-1)\widehat{\sigma}^F_{nkov,1,11}+(n_2-1)\widehat{\sigma}^F_{nkov,2,11}}.
\end{aligned}
$$

Wie man sieht, unterscheiden sich diese Schätzer von $\widehat{\gamma}_{nkov,V}$ nur durch die unterschiedliche Gewichtung und durch die Verwendung von etwas anderen Varianzen und Kovarianzen. Mit diesen Schätzern aus dem Regressionsansatz erhält man nun unter der Hypothese $H_0:\ F_{12}=F_{22}$ die folgende asymptotische Verteilungsaussage.

Satz 5.1.4

Außer den Annahmen (5.1), (5.2), (5.3) und (5.4) des Regressionsmodells seien (4.2), (4.3), (4.4) und (4.5) erfüllt. Dann sind unter der Hypothese $H_0:\ F_{12}=F_{22}$ die Teststatistiken

$$
\begin{aligned}
T^F_{n,nkov} &= \sqrt{n}\frac{\widehat{p}^*(\gamma_{nkov,nR})-1/2}{\sqrt{\widehat{\sigma}^F_{nkov,n}}} \quad \textit{und} \\
T^F_{nkov} &= \sqrt{n}\frac{\widehat{p}^*(\gamma_{nkov,R})-1/2}{\sqrt{\widehat{\sigma}^F_{nkov}}}
\end{aligned}
$$

bei randomisierten Studien asymptotisch standardnormalverteilt. Dabei gelten die Bezeichnungen

$$
\widehat{\sigma}^F_{nkov,n} = \frac{n}{n_1}\left(\widehat{\sigma}^F_{nkov,n,1,22}+\widehat{\gamma}^2_{nkov,nR}\widehat{\sigma}^F_{nkov,n,1,11}-2\widehat{\gamma}_{nkov,nR}\widehat{\sigma}^F_{nkov,n,1,21}\right)+
$$

$$\begin{aligned}
& \frac{n}{n_2}\left(\widehat{\sigma}^F_{nkov,n,2,22} + \widehat{\gamma}^2_{nkov,nR}\widehat{\sigma}^F_{nkov,n,2,11} - 2\widehat{\gamma}_{nkov,nR}\widehat{\sigma}^F_{nkov,n,2,21}\right) \quad \textit{und} \\
\widehat{\sigma}^F_{nkov} \;=\; & \frac{n}{n_1}\left(\widehat{\sigma}^F_{nkov,1,22} + \widehat{\gamma}^2_{nkov,R}\widehat{\sigma}^F_{nkov,1,11} - 2\widehat{\gamma}_{nkov,R}\widehat{\sigma}^F_{nkov,1,21}\right) + \\
& \frac{n}{n_2}\left(\widehat{\sigma}^F_{nkov,2,22} + \widehat{\gamma}^2_{nkov,R}\widehat{\sigma}^F_{nkov,2,11} - 2\widehat{\gamma}_{nkov,R}\widehat{\sigma}^F_{nkov,2,21}\right)
\end{aligned}$$

Beweis: Siehe Theorem 4.5 von Langer (1998) und Satz 4.13 von Siemer (1999). □

Um eine bessere Einhaltung des Niveaus bei kleinen Stichproben zu erreichen, können für die drei verschiedenen Teststatistiken motiviert durch Smith (1936), Welch (1938) und Satterthwaite (1946) t-Approximation verwendet werden. Die Schätzer der Varianzen spalten sich bei allen drei Teststatistiken in zwei Summanden, einer für jede Gruppe, auf. Also erhält man mit

$$\begin{aligned}
\widehat{\tau}^F_{nkov,n,i} &= \frac{1}{n_i}\left(\widehat{\sigma}^F_{nkov,n,i,22} + \widehat{\gamma}^2_{nkov,nR}\widehat{\sigma}^F_{nkov,n,i,11} - 2\widehat{\gamma}_{nkov,nR}\widehat{\sigma}^F_{nkov,n,i,21}\right), \\
\widehat{\tau}^F_{nkov,i} &= \frac{1}{n_i}\left(\widehat{\sigma}^F_{nkov,i,22} + \widehat{\gamma}^2_{nkov,R}\widehat{\sigma}^F_{nkov,i,11} - 2\widehat{\gamma}_{nkov,R}\widehat{\sigma}^F_{nkov,i,21}\right) \quad \text{und} \\
\widehat{\tau}^p_{nkov,i} &= \frac{1}{n_i}\left(\widehat{\sigma}^p_{nkov,i,22} + \widehat{\gamma}^2_{nkov,V}\widehat{\sigma}^p_{nkov,i,11} - 2\widehat{\gamma}_{nkov,V}\widehat{\sigma}^p_{nkov,i,21}\right)
\end{aligned}$$

die folgenden Freiheitsgrade für die t-Approximationen

$$\begin{aligned}
\widehat{\nu}^F_{nkov,n} &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^F_{nkov,n,i}\right)^2}{\sum_{i=1}^2 \left(\widehat{\tau}^F_{nkov,n,i}\right)^2/(n_i-1)}, \\
\widehat{\nu}^F_{nkov} &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^F_{nkov,i}\right)^2}{\sum_{i=1}^2 \left(\widehat{\tau}^F_{nkov,i}\right)^2/(n_i-1)} \quad \text{und} \\
\widehat{\nu}^p_{nkov} &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^p_{nkov,i}\right)^2}{\sum_{i=1}^2 \left(\widehat{\tau}^p_{nkov,i}\right)^2/(n_i-1)}.
\end{aligned}$$

Mit den t-Approximationen ergeben sich also sechs verschiedene Teststatistiken. Diese wurden für randomisierte Studien und unter der Hypothese $H_0 : F_{12} = F_{22}$ mittels Simulationen bezüglich der Niveaueinhaltung verglichen. Für jede Simulation wurden 10000 Schritte durchlaufen. In jedem Schritt wurden zunächst unabhängige normalverteilte Zufallsvariablen erzeugt. Durch eine Linearkombination dieser Zufallsvariablen erhält man die gewünschte Abhängigkeitsstruktur. Danach werden die Beobachtungen derart in fünf Kategorien zusammengefasst, dass die Marginalverteilungen für beide Gruppen zu beiden

Tabelle 5.1: Niveausimulation für die nichtparametrische Kovarianzanalyse im balancierten Fall zum Niveau 5%

n_1	$H_0: F_{12} = F_{22}$		$H_0: p_2 = 1/2$	
	n	t	n	t
7	10,10	6,40	9,09	5,90
10	7,89	5,88	7,32	5,24
15	7,27	5,97	6,55	5,39
20	6,02	5,32	5,40	4,55
30	5,90	5,35	5,47	4,86
50	5,63	5,38	5,42	5,12

Zeitpunkten Gleichverteilungen sind. Damit ist die Voraussetzung $F_{11} = F_{21}$ erfüllt. Außerdem gilt die Hypothese $H_0: F_{12} = F_{22}$ und mithin auch $H_0: p_2 = 1/2$. Es wurden die Stichprobenumfänge 7, 10, 15, 20, 30 und 50 Versuchseinheiten pro Gruppe gewählt. Exemplarisch wurde das 5% Niveau gewählt und die empirischen Niveaus in Tabelle 5.1 aufgenommen. Da zunächst ein balanciertes Design gewählt wurde, sind die ungewichteten und die gewichteten Statistiken unter $H_0: F_{12} = F_{22}$ gleich.

Dabei bezeichnet n in den Simulationstabellen die Normal-Approximation, während t die t-Approximation bezeichnet. Alle Tests sind zunächst antikonservativ. Die t-Approximationen weichen für die Hypothese $H_0: p_2 = 1/2$ schon ab sieben und für $H_0: F_{12} = F_{22}$ ab zehn Versuchseinheiten pro Gruppe nicht mehr als ein Prozent vom nominellen Niveau ab. Sie sind bei kleinen Stichprobengrößen also gut geeignet. Die normal-Approximationen brauchen hingegen mindestens einen Stichprobenumfang von 20 bis 30 Versuchseinheiten pro Gruppe. Insgesamt schneiden die Teststatistiken unter der Hypothese $H_0: p_2 = 1/2$ besser ab.

Um den Unterschied zwischen den gewichteten und den ungewichteten Statistiken zu dokumentieren, wurden weitere Simulationen durchgeführt. Dabei wurden die Stichprobenumfänge für die beiden Gruppen unterschiedlich gewählt. Diese Werte sind in der Tabelle 5.2 festgehalten.

Der Vergleich zwischen den Teststatistiken unter der einen oder der anderen Hypothese fällt hier genauso aus wie im balancierten Fall. Ebenso haben die t-Approximationen bessere Werte als die Normalapproximationen. Der Unterschied zwischen den gewichteten und den ungewichteten Statistiken unter der Hypothese $F_{12} = F_{22}$ fällt sehr gering aus. Es ist daher schwierig eins dieser beiden Verfahren auf Grund der Simulationen zu

Tabelle 5.2: Niveausimulation für die nichtparametrische Kovarianzanalyse im unbalancierten Fall zum Niveau 5%

$n_1; n_2$	$H_0: F_{12} = F_{22}$				$H_0: p_2 = 1/2$	
	gewichtet		ungewichtet			
	n	t	n	t	n	t
6; 8	10,39	6,44	10,54	6,45	9,74	6,41
7;13	8,44	5,82	8,61	6,05	8,32	5,86
10;20	7,05	5,44	7,11	5,58	6,85	5,39
14;26	6,37	5,20	6,42	5,18	6,24	5,29
22;38	6,37	5,60	6,44	5,59	6,03	5,32
30;70	5,65	5,26	5,78	5,29	5,65	5,11

bevorzugen.

Außerdem wurde eine Powersimulation durchgeführt, deren Ergebnisse in der Tabelle 5.3 stehen. Es wurde der balancierte Fall mit 50 Versuchseinheiten pro Gruppe angenommen. Die Zufallsvariablen wurden dabei zunächst wie unter der Hypothese erzeugt. Die Alternative wurde mittels eines Parameters δ parametrisiert. Bei $\delta = 0$ sind alle Hypothesen erfüllt, da alle Wahrscheinlichkeiten für die verschiedenen Kategorien in beiden Gruppen sowohl für die Ziel- als auch für die Kovariable gleich sind. Für größere Werte von δ werden die Wahrscheinlichkeiten für die kleinen Kategorien in der zweiten Gruppe zum zweiten Zeitpunkt größer. Die Hypothese ist dann nicht mehr erfüllt. Die Wahrscheinlichkeiten $q_{22s} = \mathrm{P}(X_{2k2} = s)$, $s = 1, \ldots 5$, die sich ebenfalls in der Tabelle 5.3 befinden, geben die verschiedenen Alternativen an.

Man erkennt, dass die Unterschiede zwischen den Gütefunktionen der vier Teststatistiken sehr gering sind. Unter Alternative setzen sich im Grunde nur die geringen Differenzen fort, die schon unter Hypothese vorhanden waren. Insgesamt kann man daher die t-Approximationen empfehlen, da sie das Niveau besser einhalten als die Normalapproximationen und im Vergleich der Macht ebenbürtig sind.

5.2 Nicht randomisierte Versuchsanlage

Nachdem wir uns im vorhergehenden Abschnitt mit randomisierten Versuchsanalagen beschäftigt haben, sollen nun die Änderungen betrachtet werden, welche sich bei nicht randomisierten Studien ergeben. In diesem Fall können wir nicht mehr von der Annahme

Tabelle 5.3: Powersimulation für die nichtparametrische Kovarianzanalyse im balancierten Fall zum Niveau 5% und bei $n_i = 50$

n_1						$H_0: F_{12} = F_{22}$		$H_0: p_2 = 1/2$	
δ	q_{221}	q_{222}	q_{223}	q_{224}	q_{225}	n	t	n	t
0	20,00	20,00	20,00	20,00	20,00	5,20	4,95	5,00	4,75
1	39,35	24,02	16,19	11,65	8,79	9,18	8,75	8,97	8,39
2	41,09	23,95	15,68	11,06	8,22	17,65	17,07	17,42	16,72
3	43,08	23,79	15,08	10,42	7,63	29,60	28,57	28,71	27,83
4	44,16	23,67	14,76	10,08	7,32	55,10	54,12	53,97	53,08
5	47,49	23,20	13,75	9,098	6,46	69,47	68,71	68,60	67,75
6	48,29	23,06	13,51	8,870	6,27	81,77	81,11	80,72	80,10
7	49,83	22,76	13,04	8,450	5,92	87,78	87,33	87,18	86,59

$F_{11} = F_{21}$ ausgehen, da diese Annahme aus der zufälligen Zuteilung der Versuchseinheiten zu den beiden Gruppen resultiert. Gilt also nicht $F_{11} = F_{21}$, so unterscheiden sich die beiden Gruppen bezüglich der Baseline-Werte. Diese unterschiedlichen Ausgangswerte müssen bei der Betrachtung der Zufallsvariablen nach der Behandlung berücksichtigt werden. Folglich ist es nun nicht mehr sinnvoll die Hypothese $H_0: F_{12} = F_{22}$ zu stellen, da diese die ungleichen Verteilungen der Baseline-Werte in den beiden Gruppen nicht berücksichtigt. Auch der Effekt p_2 berücksichtigt nur die Marginalverteilungen zum zweiten Zeitpunkt. Daher ist es notwendig eine neue Größe zu definieren, welche die Marginalverteilungen zu beiden Zeitpunkten und deren Zusammenhang, das heißt die gemeinsame Verteilung, berücksichtigt. Ein solcher Effekt wurde in der Modellgleichung (4.9) auf Seite 17 vorgestellt. Der Effekt $p^*(\gamma_{nkov,V})$ ist eine natürliche Erweiterung des Effektes p_2, da im randomisierten Fall $p^*(\gamma_{nkov,V}) = p_2$ gilt. Daher wird nun die Hypothese

$$H_0: \quad p^*(\gamma_{nkov,V}) \quad = \quad \frac{1}{2} \tag{5.7}$$

gestellt. Sie reduziert sich im randomisierten Fall auf $p_2 = 1/2$, und ist somit eine Verallgemeinerung dieser Hypothese. Mit der Notation aus den vorhergehenden Kapiteln kann die Verteilung einer Teststatistik für die Hypothese (5.7) angegeben werden.

Satz 5.2.1

Unter den Annahmen (4.2), (4.3), (4.4), (4.5) und (4.9) ist die Teststatistik

$$T^p_{nkov} \quad = \quad \sqrt{n}\frac{\widehat{p}^*(\gamma_{nkov,V}) - 1/2}{\sqrt{\widehat{\sigma}^p_{nkov}}}$$

unter der Hypothese (5.7) asymptotisch standardnormalverteilt.

Tabelle 5.4: Niveausimulation für die nichtparametrische Kovarianzanalyse im balancierten und nicht randomisierten Fall zum Niveau 5%

n_1	$H_0: F_{12} = F_{22}$		$H_0: p^* = 1/2$	
	n	t	n	t
7	11,22	8,19	12,35	9,07
10	9,07	6,91	9,66	7,48
15	7,48	6,24	7,82	6,54
20	6,62	5,59	6,80	5,91
30	6,19	5,62	6,28	5,67
50	5,77	5,62	5,92	5,61
100	5,49	5,29	5,32	5,20

Dieser Satz ist die natürliche Erweiterung des Satzes 5.1.3 von Seite 62 im randomisierten Fall. Wie dort kann für kleine Stichproben eine t-Approximation verwendet werden. Um zu überprüfen, ab welchen Fallzahlen diese Tests das Niveau einhalten, wurde eine Niveausimulation durchgeführt. Die verwendeten gemeinsamen Verteilungen sind in der Tabelle 5.5 angegeben. Aus ihnen ergeben sich die relativen Effekte

$$p_1 = 0,555 \quad \text{und} \quad p_2 = 0,5169.$$

Die Marginalverteilungen der Gruppen unterscheiden sich also innerhalb der Zeitpunkte. Die gemeinsamen Verteilungen wurden jedoch so gewählt, dass sich mit dem Parameter

$$\gamma_{nkov,V} = 0,3068$$

die Gültigkeit der Hypothese (5.7) ergibt. Bei einem balancierten Versuchsdesign wurden die empirischen Niveaus für die Stichprobenumfänge $n_1 = 7, 10, 15, 20, 30, 50$ und 100 bestimmt. Um zu überprüfen, ob die Teststatistiken für den randomisierten Fall und die Hypothese $H_0: F_{12} = F_{22}$ diese Alternative aufdecken, wurden diese ebenfalls berechnet. Die Werte, welche sich bei 10000 Simulationsschritten für jeden Stichprobenumfang für ein nominelles Niveau von 5% ergaben, sind in der Tabelle 5.4 zusammengefasst.

Zunächst zeigt sich, dass die Teststatistiken für die Hypothese $H_0: F_{12} = F_{22}$ diese bei zunehmendem Stichprobenumfang immer seltener verwerfen. Die Macht konvergiert

Tabelle 5.5: Auf die vierte Nachkommastelle gerundete Wahrscheinlichkeiten $q_{iuv} = \mathrm{P}\left((X_{ik1}, X_{ik2})^t = (u, v)^t\right)$ für die Niveausimulation der nichtparametrischen Kovarianzanalyse im balancierten und nicht randomisierten Fall

Gruppe	Kategorie erster Zeitp.	Kategorie zweiter Zeitp.			
		$v = 1$	$v = 2$	$v = 3$	$v = 4$
$i = 1$	$u = 1$	0,0625	0,0625	0,0625	0,0626
	$u = 2$	0,0625	0,0625	0,0625	0,0626
	$u = 3$	0,0625	0,0625	0,0625	0,0626
	$u = 4$	0,0625	0,0625	0,0625	0,0626
$i = 2$	$u = 1$	0,1000	0,0600	0,0000	0,0000
	$u = 2$	0,0700	0,1375	0,0625	0,0000
	$u = 3$	0,0000	0,0725	0,0875	0,1000
	$u = 4$	0,0000	0,0725	0,0875	0,1500

nicht gegen eins sondern gegen das nominelle Niveau. Die Tests sind als nicht konsistent gegen diese Alternative.

Des Weiteren hält die t-Approximation des Tests für die Hypothese (5.7) von Seite 68 das Niveau ab einem Stichprobenumfang von etwa 20 Versuchseiheiten pro Gruppe ein. Ab dann weicht das empirische Niveau nicht mehr als 1% vom vorgegebenen ab. Die Normalapproximation braucht mit etwa 50 Versuchseinheiten etwas mehr Stichprobenumfang, um dieses Kriterium zu erreichen. Insgesamt ist ein höherer Stichprobenumfang als im randomisierten Fall notwendig, um das Niveau einzuhalten. Dies kann damit erklärt werden, dass nun die langsame Konvergenz von $\gamma_{nkov,V}$ eine wichtige Rolle spielt.

Kapitel 6

Anwendung der nichtparametrischen Kovarianzanalyse

6.1 Neurologische Beeinträchtigung bei MS

Die vorgestellten nichtparametischen kovarianzanalytischen Verfahren werden in diesem Abschnitt anhand der MS-Studie illustriert. Die Zufallsvariablen nach der Behandlung werden als Ziel- und die Zufallsvariablen vor der Behandlung als Kovariablen betrachtet. Damit sei F_{22} die Marginalverteilung der Zufallsvariablen nach der Behandlung in der Behandlungsgruppe und F_{21} die Marginalverteilung der zugehörigen Kovariablen. Ferner sind F_{12} und F_{11} die Marginalverteilungen der Ziel- und Kovariablen in der Kontrollgruppe.

Um diese Marginalverteilungen zu vergleichen werden die relativen Effekte p_j innerhalb der Ziel- und Kovariablen betrachtet. Für die Zielvariable kann der relative Effekt durch $\widehat{p}_2 = 0{,}4085$ geschätzt werden. Mit Hilfe der Gleichung (4.6) von Seite 17 kann der relative Effekt interpretiert werden. Da mit $i = 2$ die Behandlungsgruppe bezeichnet wurde, bedeutet ein relativer Effekt, der kleiner als 0,5 ist, dass die Werte der Zielvariablen in der Behandlungsgruppe zu kleineren Werten tendieren als in der Kontrollgruppe. Für die Kovariable wurde ein relativer Effekt von $\widehat{p}_1 = 0{,}5670$ geschätzt. Mithin tendieren die Kovariablen der Kontrollgruppe zu kleineren Werten als die der Behandlungsgruppe. Damit ist festzustellen, dass die Behandlungsgruppe, obwohl sie vor der Behandlung schlechter war als die Kontrollgruppe, nach der Behandlung zu besseren Werten neigt.

Die beiden Effekte $\widehat{p}_1$ und $\widehat{p}_2$ addieren sich in dem Schätzer für p^*. Dabei setzen sich die geringen Unterschiede zwischen den vorgeschlagenen Regressionskoeffizienten $\widehat{\gamma}_{nkov,V}$, $\widehat{\gamma}_{nkov,nR}$ und $\widehat{\gamma}_{nkov,R}$ auf den geschätzten relativen Effekt $\widehat{p}^*$ fort, denn $\widehat{p}^*(\gamma_{nkov,V})$,

Tabelle 6.1: Übersicht der Testergebnisse der nichtparametrischen Kovarianzanalyse für die MS-Studie

Hypothese	$H_0: F_{12} = F_{22}$		$H_0: p_2 = 1/2$
	gewichtet	ungewichtet	
$\widehat{p}^*$	0,3612	0,3612	0,3617
Koeffizient ($\widehat{\gamma}_{nkov,nR}$, $\widehat{\gamma}_{nkov,R}$,$\widehat{\gamma}_{nkov,V}$)	0,7055	0,7055	0,6985
Betrag der Teststatistik	2,6867	2,6826	2,7065
p-Wert bei Normalverteilung	0,0072	0,0073	0,0068
p-Wert bei t-Verteilung	0,0098	0,0100	0,0093
Freiheitsgrad der t-Verteilung	48,6829	48,6454	50,0015

$\widehat{p}^*(\gamma_{nkov,nR})$ und $\widehat{p}^*(\gamma_{nkov,R})$ unterscheiden sich erst ab der vierten Nachkommastelle (siehe Tabelle 6.1). Dieser geschätzte relative Effekt von ungefähr 0,36 kann folgendermaßen interpretiert werden. Er schätzt im vorliegenden Fall den relativen Effekt p_2 in der Zielvariablen erwartungstreu, da bei einer randomisierten Studie von der Homogenität der Kovariablen $F_{11} = F_{21}$ ausgegangen werden kann. Ein Wert von 0,36 bedeutet mithin, dass sich die Wahrscheinlichkeit, dass die Zufallsvariablen in der Behandlungsgruppe zu größeren Werten tendieren als in der Kontrollgruppe, von 0,5 um $-0{,}14$ unterscheidet. Die Wahrscheinlichkeit $\mathrm{P}\,(X_{112} < X_{212}) + 1/2\mathrm{P}\,(X_{112} = X_{212})$ wird damit durch 0,36 geschätzt. Bei gleichen Verteilungen in den beiden Gruppen wäre diese aber 0,5 und damit deutlich größer. Dieser Unterschied von $-0{,}14$ spricht also für die Behandlung, da größere und damit schlechtere Werte in der Kontrollgruppe wahrscheinlicher sind als in der Behandlungsgruppe.

Es stellt sich die Frage, inwieweit sich ein geschätzter relativer Effekt von 0,36 deutlich von 0,5 unterscheidet. Dies kann einerseits durch die Hypothese über die Verteilungen $H_0: F_{12} = F_{22}$ oder über den relativen Effekt $H_0: p_2 = 1/2$ formuliert werden. Diese sollen nun zu einem Niveau von 5% getestet werden. Bei der Hypothese über die Verteilungen kann das gewichtete von dem ungewichteten Verfahren unterschieden werden. Da sowohl Approximationen mit der Normal- als auch der t-Verteilung vorgestellt wurden, ergeben sich damit 6 p-Werte, die letztlich gleich interpretiert werden. Diese in Tabelle 6.1 aufgelisteten Werte liegen alle zwischen 0,5% und 1%, womit die Hypothesen verworfen werden können. Die Simulationen für einen solchen Stichprobenumfang haben gezeigt, dass die t-Approximationen das nominelle Niveau nicht mehr als 1% überschreiten. Die Normalapproximation ist bei den vorliegenden Anzahlen noch mit etwas Vorsicht zu behandeln.

Bei ihr liegt das simulierte Niveau zwischen 1% und 2% zu hoch.

6.2 Behandlung von Schlaflosigkeit

Wie bei der MS-Studie wurden auch bei der Insomnia-Studie die Patienten den Gruppen randomisiert zugeteilt. Dass diese Randomisierung hinsichtlich der verwendeten Skala mit den vier geordneten Kategorien erfolgreich war, ist an einem geschätzten relativen Effekt in der Kovariablen von $\widehat{p}_1 = 0{,}5052$ zu erkennen. Er unterscheidet sich kaum von 0,5. Mithin können die empirischen Verteilungen der Kovariablen als tendenziell gleich angesehen werden. Dabei gelte wie im vorhergehenden Beispiel für die Behandlungsgruppe $i = 2$ und für die Kontrollgruppe $i = 1$.

In der Zielvariablen zeigt sich hingegen mit $\widehat{p}_2 = 0{,}3855$ ein recht starker relativer Effekt. Da der Schätzer für den relativen Effekt kleiner als 0,5 ist, spricht das für die Behandlungsgruppe, da auch in diesem Beispiel die großen Werte die ungünstigen Werte sind. Ein solcher Effekt bedeutet, dass die Wahrscheinlichkeit für hohe Werte auf der ordinalen Skala in der Kontrollgruppe größer ist als in der Behandlungsgruppe. Mit anderen Worten heißt dies, dass der Schätzer des relativen Effektes dafür spricht, dass die Verteilung F_{12} tendenziell kleiner als F_{22} ist.

Tabelle 6.2: Übersicht der Testergebnisse der nichtparametrischen Kovarianzanalyse für die Insomnia-Studie

Hypothese	$H_0: F_{12} = F_{22}$		$H_0: p_2 = 1/2$
	gewichtet	ungewichtet	
$\widehat{p}^*$	0,3879	0,3879	0,3879
Koeffizient ($\widehat{\gamma}_{nkov,nR}$, $\widehat{\gamma}_{nkov,R}$,$\widehat{\gamma}_{nkov,V}$)	0,4615	0,4615	0,4587
Betrag der Teststatistik	3,5883	3,5875	3,5923
p-Wert bei Normalverteilung	0,0003	0,0003	0,0003
p-Wert bei t-Verteilung	0,0004	0,0004	0,0004
Freiheitsgrad der t-Verteilung	236,9797	236,9801	234,0597

Aus den beiden Schätzern $\widehat{p}_1$ und $\widehat{p}_2$ der relativen Effekte wird nun der Schätzer $\widehat{p}^*$ des relativen Effektes gebildet. Da sich die Schätzer der Regressionsparameter kaum unterscheiden und zudem der geschätzte relative Effekt in der Kovariablen fast 0,5 ist, unterscheiden sich die Schätzer des relativen Effektes nicht wesentlich(siehe Tabelle 6.2). Wie

bei der MS-Studie schätzt $\widehat{p}^*$ die Wahrscheinlichkeit $\mathrm{P}\left(X_{112} < X_{212}\right) + 1/2\mathrm{P}\left(X_{112} = X_{212}\right)$ erwartungstreu, weil die Kovariablen homogen sind. Dieser Schätzer ist hier 0,39, was für die Behandlungsgruppe spricht, da dort die Wahrscheinlichkeit für lange Einschlafzeiten geringer ist als in der Kontrollgruppe.

Dieses Ergebnis wird durch die Tests untermauert. Sowohl die Hypothese $H_0 : F_{12} = F_{22}$ bezüglich der Verteilungen der Zielvariablen als auch die Hypothese über den relativen Effekt $H_0 : p_2 = 1/2$ kann auf einem Niveau von 5% verworfen werden. Dabei unterscheiden sich die gewichteten Verfahren von den ungewichteten hier kaum, weil die Stichprobenumfänge fast balanciert sind. Ferner sind die geschätzten Freiheitsgrade so groß, dass sich die p-Werte für die Normalverteilung kaum von den p-Werten der t-Verteilung unterscheiden. Zuletzt ist noch zu bemerken, dass bei den vorliegenden Stichprobenumfängen von $n_1 = 119$ und $n_2 = 120$ das Niveau gut eingehalten wird. Schon bei einem Stichprobenumfang, der nicht einmal halb so groß ist, unterscheidet sich das simulierte Niveau nicht mehr als 1% vom nominellen Niveau.

Kapitel 7

Vergleich der Verfahren

7.1 Definition der Vergleichspunkte

In diesem Kapitel wird die nichtparametrische Kovarianzanalyse mit den vorgestellten Verfahren aus der Literatur verglichen. Ein solcher Vergleich kann anhand verschiedener Gesichtspunkte durchgeführt werden. Diese unterschiedlichen Gesichtspunkte werden in diesem Abschnitt vorgestellt.

Ein grundlegender Aspekt eines jeden Verfahrens ist das Modell, auf dem es beruht. Die Annahmen an dieses Modell stellen damit den ersten Gesichtspunkt dar, anhand dessen der Vergleich der Verfahren durchgeführt wird. Dabei wird insbesondere untersucht, inwiefern sich die Annahmen an die Verteilungen und an die Skalen, auf denen die Zufallsvariablen beobachteten werden, unterscheiden. Dabei sollten die Annahmen an die Skala hier nicht überbewertet werden, da die in dieser Arbeit vorgestellten Verfahren gerade nach ihrer Anwendbarkeit bei ordinalen Skalen ausgewählt wurden. Andererseits erleichtert ein vielseitiges Verfahren die Arbeit des Anwenders und ist damit von Vorteil, da der Anwender weniger Verfahren verstehen und anwenden können muß. Außerdem können wir die verschiedenen Modelle hinsichtlich ihrer symmetrischen oder asymmetrischen Betrachtungsweise der Komponenten der Zufallsvektoren unterscheiden. Eine solche Unterscheidung wird zum Beispiel auch von Fahrmeir und Tutz (1994) im Abschnitt 3.5.1 vorgeschlagen.

Aus den betrachteten Verteilungen der verschieden Verfahren ergeben sich die unterschiedlichsten Effekte. Die Interpretation dieser Effekte ist ein weiterer Aspekt des Vergleiches. Da die Hypothesen meist über die Effekte gestellt werden, ergibt sich mithin ein Vergleich der Testprobleme. Eine Betrachtung der Testprobleme ist besonders wichtig, um die unterschiedliche Güte der Tests zu untersuchen. Nur bei gleichen Testproblemen ist die Macht vergleichbar.

Für die Berechnung der Schätzer und Teststatistiken ergeben sich bei einigen Verfahren zum Teil erhebliche Probleme. Dies hängt vor allem mit der Anzahl der Kategorien auf der ordinalen Skala zusammen. Die Verfahren lassen sich dabei in zwei Klassen einteilen. Während die Verfahren der ersten Klasse weitgehend unabhängig von der Anzahl der Kategorien sind, spielt diese Anzahl bei den Verfahren der zweiten Klasse eine wichtige Rolle. Des Weiteren ist die Einhaltung des Niveaus bei kleinen Stichproben eine wichtige Eigenschaft, da alle betrachteten Verfahren asymptotischer Natur sind.

7.2 Naive Verfahren

Das Modell der naiven Verfahren unterscheidet sich deutlich vom Modell der nichtparametrischen Kovarianzanalyse, denn es betrachtet im Gegensatz zu den Marginalverteilungen F_{ij} nur die Wahrscheinlichkeiten q_{is}. An diese Wahrscheinlichkeiten wird von den naiven Verfahren nur eine simple Annahme gestellt, welche besagt, dass die Wahrscheinlichkeiten q_{is} positiv sein müssen. Praktisch bedeutet dies in der Regel kaum eine Einschränkung, da sich meist ein mindestens geringer Anteil der Versuchseinheiten in beiden Gruppen verbessert, verschlechtert oder gleich bleibt. Allerdings treten Probleme bei sehr effektiven Behandlungen der Versuchseinheiten auf, da dann die Wahrscheinlichkeiten für eine Verschlechterung null sind. Außerdem können sehr feine ordinale Skalen die Wahrscheinlichkeit für einen gleichbleibenden Zustand der Versuchseinheiten minimieren. Falls jedoch in beiden Gruppen beispielsweise überhaupt keine Versuchseinheiten zweimal die gleiche Kategorie haben, könnte ein exakter Fisher-Test angewendet werden. Insgesamt kann diese Voraussetzung daher als schwach angesehen werden. Werden die Versuchseinheiten mit keiner Änderung aus der Analyse herausgenommen, so kann der exakte Test von Fisher ohnehin angewendet werden.

Weiterhin unterscheiden sich die naiven Verfahren von der Kovarianzanalyse in ihrer symmetrischen Sichtweise der Komponenten der Zufallsvektoren. Während die Kovarianzanalyse zwischen Ziel- und Kovariable unterscheidet, betrachten die naiven Verfahren beide Zufallsvariablen gleich. Die Pearson-Statistik ist invariant unter Permutationen der Komponenten der Zufallsvektoren; bei der Cochran-Armitage-Statistik ändert sich nur das Vorzeichen, wenn die beiden Komponenten vertauscht werden.

Keine Unterschiede zeigen sich hingegen bei den Annahmen an die Skala, auf der die Zufallsvariablen beobachtet werden. Sowohl bei der nichtparametrischen Kovarianzanalyse als auch bei den naiven Verfahren muss nur eine Ordnungsstruktur auf der Skala gegeben sein. Die naiven Verfahren brauchen diese Ordnung für die Vergleiche der Werte

innerhalb jeder Versuchseinheit. Bei der nichtparametrischen Kovarianzanalyse werden die Vergleiche hingegen ausschließlich innerhalb der Zeitpunkte durchgeführt.

Aufgrund dieser unterschiedlichen Vergleichsebenen sind die Effekte und folglich auch die Hypothesen der beiden Verfahren nicht zu vergleichen. Ein Gütevergleich der Tests erübrigt sich daher. In Bezug auf das Niveau bei kleinen Stichprobenumfängen kann man sagen, dass der Cochran-Armitage-Test etwas besser das Niveau einhält, als die t-Approximation für die Hypothese $H_0 : \ p_2 = 1/2$ (siehe C.1). In beiden Fällen weicht das empirischen Niveau allerdings schon bei $n_i = 7$ nicht mehr als 1% vom vorgebenen Niveau von 5% ab. Dies gilt unabhängig von der Anzahl der Kategorien für beide Verfahren. Folglich gehören sowohl die nichtparametrische Kovarianzanalyse als auch die naiven Verfahren bezüglich der Abhängigkeit von der Anzahl der Kategorien der ordinalen Skala in die erste Klasse.

7.3 Nichtparametrische faktorielle Analyse

Ganz im Gegensatz zu den naiven Verfahren hat die nichtparametrische faktorielle Analyse mit der nichtparametrischen Kovarianzanalyse viel gemein. Beide betrachten vor allem die Marginalverteilungen. An diese stellen beide Modelle fast die gleichen Voraussetzungen. Die Voraussetzungen der nichtparametrischen faktoriellen Analyse sind etwas stärker, da die Träger aller Marginalverteilungen sich überlappen sollen. Bei der nichtparametrischen Kovarianzanalyse muß dies nur innerhalb der Zeitpunkte der Fall sein.

Des Weiteren stellen beide Verfahren die gleichen schwachen Anforderungen an die Skala auf der die Beobachtungen liegen. Es genügt jeweils eine Ordnungsstrukur. Diese wird verwendet, um die Effekte definieren zu können. Bei beiden Verfahren werden relative Effekte verwendet. Die Schätzung dieser Effekte beruht in beiden Fällen auf Rängen, weshalb die Schätzer und Teststatistiken einfach zu berechnen sind. Der Unterschied besteht hier in der unterschiedlichen Sichtweise der Zufallsvariablen zu den beiden Zeitpunkte. Die faktorielle Analyse hat im Gegensatz zur Kovarianzanalyse diesbezüglich eine symmetrische Sichtweise. Daher fließen in den relativen Effekt p_W auch die Marginalverteilungen zum ersten Zeitpunkt ein und beim Schätzer $\widehat{p}_W$ werden die Ränge über alle Beobachtungen verwendet. Dies erschwert die Interpretation von p_W im Gegensatz zur sehr einfachen Interpretation von p_2. Der Effekt p_W ähnelt also dem Effekt p_2 der Kovarianzanalyse; die Hypothesen über diese Effekte sind aber nicht äquivalent.

Betrachtet man hingegen die Hypothesen über die Marginalverteilungen, so ergibt sich unter der Annahme $F_{11} = F_{21}$ das gleiche Testproblem. Daher ist es möglich, die Macht der beiden Verfahren bei randomisierten Studien zu vergleichen. Da außerdem beide Verfahren

unabhängig von der Anzahl der Kategorien der ordinalen Skala das Niveau gut einhalten, ist für die Simulation der Güte kein sehr großer Stichprobenumfang nötig.

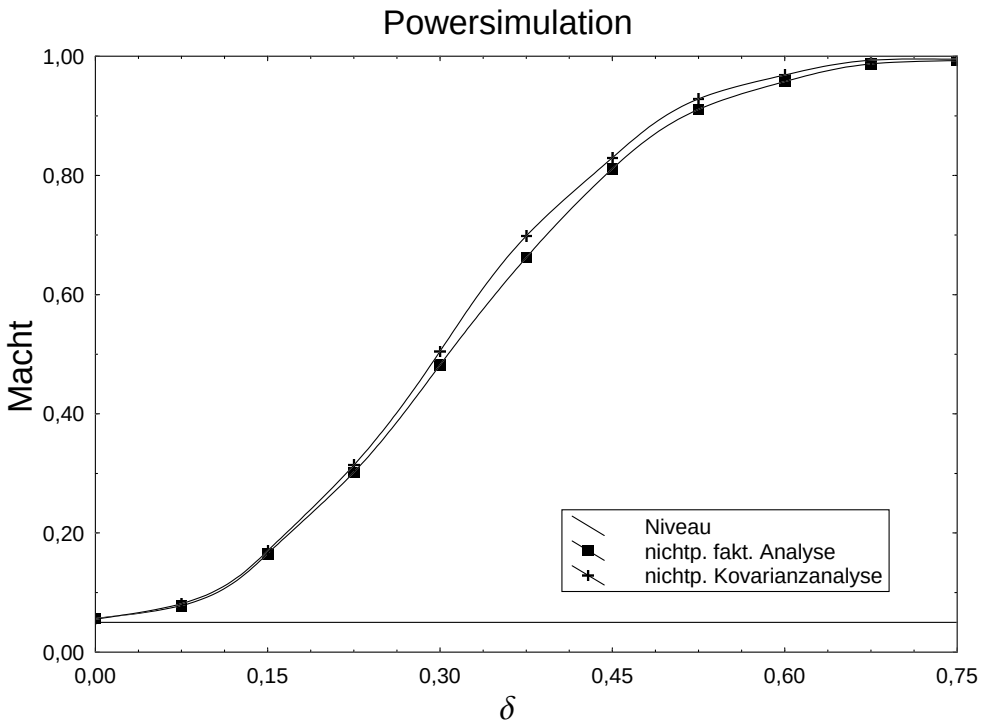

Abbildung 7.1: Vergleich der Macht zwischen der nichtparametrischen faktoriellen Analyse und der nichtparametrischen Kovarianzanalyse bei $\alpha = 0,05$ und $n_i = 50$

Wie man in der Abbildung 7.1 erkennt, ist die Macht der nichtparametrischen Kovarianzanalyse größer als die Macht der faktoriellen Analyse. Dabei gibt der Parameter δ die Alternative an. Für $\delta = 0$ gilt die Hypothese $H_0 : \ F_{12} = F_{22}$ und für $\delta > 0$ entfernt sich die Marginalverteilung F_{22} immer weiter von F_{12}.

7.4 Nichtparametrische Analyse durch Schichtung

Bei der Schichtungsanalyse werden die Follow-Up-Beobachtungen nach den Baseline-Werten geschichtet. Daraufhin werden die einzelnen Schichten für sich betrachtet. Mithin werden nicht wie in der nichtparametrischen Kovarianzanalyse die Marginalverteilungen sondern die auf die Baseline-Werte bedingten Verteilungen betrachtet. Jedoch werden an die bedingten Verteilungen die gleichen Annahmen gestellt, wie bei der nichtparametrischen Kovarianzanalyse an die Marginalverteilungen. Die Vorraussetzungen der Schich-

tungsanalyse sind aber stärker als die Annahmen der nichtparametrischen Kovarianzanalyse, da sie für jede Schicht gelten sollen.

Um diese Schichten bilden zu können, muß eine kategorielle Struktur der Daten vorliegen. Um weiterhin die Follow-Up-Werte innerhalb der Schichten miteinander vergleichen zu können, ist eine Ordnungsstruktur der Werte erforderlich. Mithin kann diese Analyse nur bei ordinalen Daten durchgeführt werden, womit die Annahmen an die Skala restriktiver sind als bei der nichtparametrischen Kovarianzanalyse.

Wird zunächst jede Schicht für sich betrachtet, so ergeben sich wie bei der nichtparametrischen Kovarianzanalyse relative Effekte für die Beschreibung der Unterschiede zwischen den Gruppen. Bei der Schichtungsanalyse werden jedoch nicht die Marginalverteilungen sondern die bedingten Verteilungen für die Definition der Effekte verwendet. Für jede einzelne Schicht sind diese relativen Effekte genauso gut zu interpretieren wie der relative Effekt p_2 bei der nichtparametrischen Kovarianzanalyse. Um aber einen Gesamteffekt zu erhalten, werden die relativen Effekte der Schichten gewichtet und summiert. Diese Zusammenfassung führt zu einem Verlust an Interpretierbarkeit. Diesbezüglich hat der Effekt p_2 der nichtparametrischen Kovarianzanalyse wesentlich bessere Eigenschaften.

In Bezug auf die Sichtweise der Komponenten der Zufallsvektoren unterscheiden sich die beiden Verfahren hingegen kaum. Sowohl bei der nichtparametrischen Kovarianzanalyse als auch bei der Schichtungsanalyse liegt der Schwerpunkt der Betrachtung bei den Follow-Up-Werten. Die Baseline-Werte spielen in beiden Verfahren die Rolle einer Kovariablen.

Ein wesentliches Problem der Schichtungsanalyse ergibt sich aus der Art, wie mit diesen Kovariablen, also den Baseline-Werte, umgegangen wird. Da für jede Kategorie der Baseline-Werte eine eigene Schicht gebildet wird, hängt die Schichtungsanalyse sehr von der Anzahl der verwendeten Kategorien ab. Bei vielen Kategorien ergeben sich erhebliche Probleme, da dann immer weniger Beobachtungen in die einzelnen Schichten fallen. Mithin gehört die Schichtungsanalyse bezüglich der Abhängigkeit von der Anzahl der Kategorien in die zweite Klasse der Verfahren, was einen erheblichen Nachteil gegenüber der nichtparametrischen Kovarianzanalyse bedeutet.

Die starke Abhängigkeit von der Anzahl der Kategorien spiegelt sich auch bei der Betrachtung der Niveausimulationen wieder (siehe C.3). Schon bei nur fünf Kategorien ist ein erheblicher Stichprobenumfang von 50 Versuchseinheiten pro Gruppe nötig, um das nominelle Niveau einzuhalten. Damit sind die Schichtungsverfahren bei kleinen und mittleren Stichprobenumfängen im Gegensatz zur nichtparametrischen Kovarianzanalyse völlig ungeeignet. Im Gegensatz zum Niveau kann die Güte der beiden Verfahren nicht miteinander verglichen werden, weil sich die Testprobleme grundsätzlich unterscheiden.

7.5 Parametrische faktorielle Analyse

Nach den Vergleichen mit den nichtparametrischen Verfahren folgen in diesem und dem nächsten Abschnitt die Vergleiche bezüglich der parametrischen Verfahren. Das Modell der parametrischen faktoriellen Analyse basiert wie in der nichtparametrischen Kovarianzanalyse auf den Marginalverteilungen. Um diese jedoch mit Hilfe von Parametern beschreiben zu können, wird eine Modellgleichung für die Marginalverteilungen angenommen. Aufgrund dieser Modellgleichung unterliegen die betrachteten Marginalverteilungen einer stochastischen Ordnung. Diese Ordnung spiegelt sich in einem Parameter wieder, der nur von den beiden Faktoren Gruppe und Zeit abhängt. Dieser Parameter beschreibt, ob die Marginalverteilung der einen Faktorstufenkombination stochastisch größer, gleich oder kleiner als die Marginalverteilung einer anderen Faktorstufenkombination ist. Mit Hilfe von anderen Parametern, den sogenannten Cutpoints, wird festgelegt, welche Marginalverteilungen in dem Modell überhaupt möglich sind. Damit wird die sehr allgemeine Menge der möglichen Marginalverteilungen aus der nichtparametrischen Kovarianzanalyse deutlich eingeschränkt. Die Annahmen an die Marginalverteilung sind in der parametrischen faktoriellen Analyse also erheblich stärker als in der nichtparametrischen Kovarianzanalyse.

Ein weiterer Nachteil der parametrischen faktoriellen Analyse ist die Festlegung auf die ordinale Skala. Nichtkategorielle Daten können im Gegensatz zur nichtparametrischen Kovarianzanalyse nicht ausgewertet werden. Des Weiteren ist die parametrische faktorielle Analyse sehr abhängig von der Anzahl der verwendeten Kategorien auf der ordinalen Skala. Vor allem bei geringen Annahmen an die Working Correlation Matrix ist ein erheblicher Stichprobenumfang notwendig, um das Niveau einzuhalten. Ferner kann es passieren, dass die iterativen Verfahren, welche zur Berechnung der Schätzer verwendet werden, bei vielen Kategorien nicht konvergieren. In diesem Fall kann die parametrische faktorielle Analyse also nicht verwendet werden.

Bezüglich der Effekte hat die parametrische faktorielle Analyse den Vorteil, dass die von ihr verwendeten Odds Ratios sehr weit verbreitet sind, auch wenn der relative Effekt p_2 mindestens ebenso einfach zu interpretieren ist. Probleme bezüglich der Interpretation liefert die faktorielle Sichtweise des Modells. Diese entsprechen den Problemen der nichtparametrischen faktoriellen Analyse, da diese ebenfalls die Auswahl zwischen der Effekt- und der Referenzkodierung hat. Je nachdem welche Kodierung man wählt, sind die Ergebnisse unterschiedlich zu interpretieren.

Die Hypothese $H_0 : \beta_W = 0$ bezüglich des Wechselwirkungseffektes der parametrischen faktoriellen Analyse entspricht im Fall der stochastischen Ordnung der Marginalverteilun-

gen und bei randomisierten Studien genau der Hypothese $H_0 : p_2 = 1/2$ beziehungsweise der Hypothese $H_0 : F_{12} = F_{22}$ aus der nichtparametrischen Kovarianzanalyse. Damit stellt sich das gleiche Testproblem und ein Gütevergleich zwischen den beiden Verfahren ist möglich.

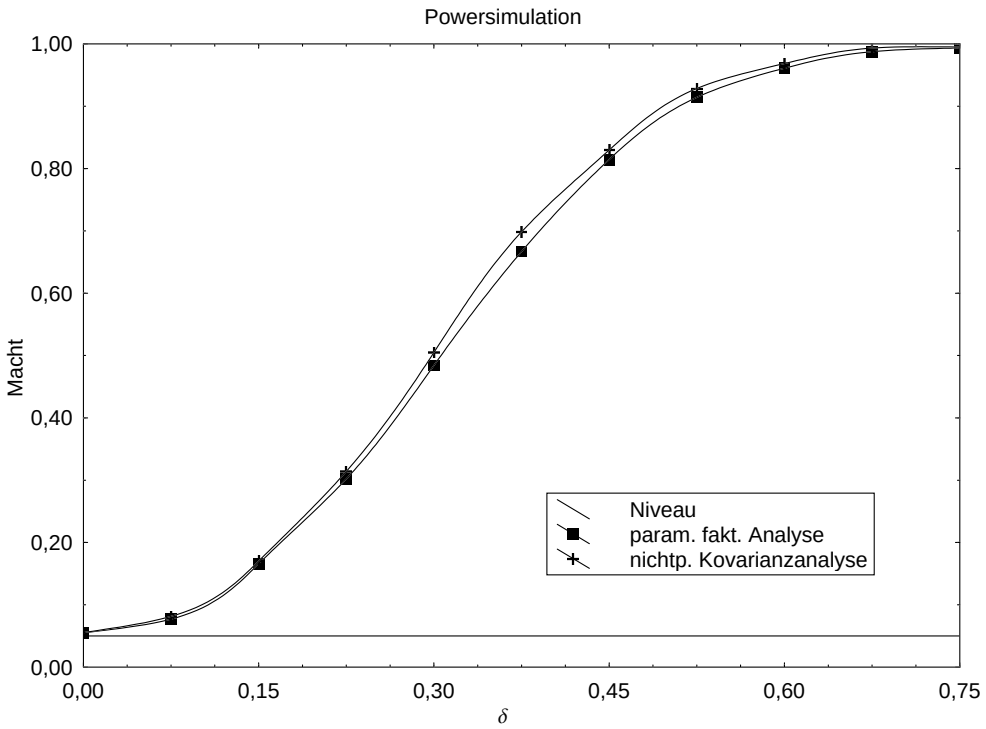

Abbildung 7.2: Vergleich der Macht zwischen der parametrischen faktoriellen Analyse und der nichtparametrischen Kovarianzanalyse bei $\alpha = 0,05$ und $n_i = 50$

Die Ergebnisse dieses Gütevergleichs sind in der Abbildung 7.2 dargestellt. Die Zufallsvariablen wurden wie bei den Niveausimulationen erzeugt. Der Parameter δ stellt die Wechselwirkung im Proportional Odds Modell dar. Im Fall von $\delta = 0$ sind alle vier Marginalverteilunge F_{ij} gleich. In diesem Fall sind alle Hypothesen erfüllt. Wie schon beim Vergleich der nichtparametrischen Kovarianzanalyse zur nichtparametrischen faktoriellen Analyse zeigen sich die Vorzüge der Kovarianzanalyse. Diese hat auch hier eine höhere Macht als das faktoriellen Verfahren.

7.6 Parametrische Kovarianzanalyse

Im letzten Abschnitt dieses Kapitels sollen nun die nichtparametrische und die parametrische Kovarianzanalyse verglichen werden. Im Gegensatz zur nichtparametrischen Kovarianzanalyse betrachtet die parametrische Kovarianzanalyse die auf die Baseline-Werte bedingten Verteilungen und nicht die Marginalverteilungen und ähnelt damit sehr der Schichtungsanalyse. Sie hat aber die gleiche unsymmetische Sichtweise der Komponenten der Zufallsvektoren wie die nichtparametrische Kovarianzanalyse. Die bedingten Wahrscheinlichkeiten in der Gruppe i den Follow-Up-Wert s zu haben bei gegebenem Baseline-Wert s' werden mit Hilfe einer Regressionsgleichung modelliert. Mithin sind die bedingten Verteilungen stochastisch geordnet, was eine wesentliche Einschränkung der möglichen Verteilungsfunktionen bedeutet.

Für die Regressionsgleichung werden jeweils $r - 1$ Parameter verwendet, um die Einflüsse der Kategorien zum ersten Zeitpunkt zu modellieren. Der Unterschied zwischen den beiden Gruppen wird durch einen Parameter modelliert, welcher sich als Odds Ratio interpretieren läßt. Die Verwendung dieses weit verbreiteten Vergleichsmaßes ist der einzige Vorteil dieses Verfahrens gegenüber der nichtparametrischen Kovarianzanalyse. Insgesamt überwiegen hingegen die Nachteile. Zunächst ist das Modell sehr restriktiv, da die Regressionsgleichung die Menge der möglichen Verteilungsfunktionen stark einschränkt. Bei der nichtparametrischen Kovarianzanalyse werden hingegen nur sehr schwache Annahmen an das Modell gestellt. Des Weiteren kann die parametrische Kovarianzanalyse im Gegensatz zur nichtparametrischen Kovarianzanalse nur ordinale Daten auswerten.

Wie alle anderen hier betrachteten Verfahren ist auch die parametrische Kovarianzanalyse von asymptotischer Natur. Um das eingehaltene Niveau bei geringen Stichprobenumfängen zu überprüfen, wurden Simulationen durchgeführt. Die Ergebnisse dieser Simulationen sind in der Tabelle C.24 zusammengefasst. Man erkennt, dass das empirische Niveau der parametrischen Kovarianzanalyse stark von der Anzahl der Kategorien abhängt. Je mehr Kategorien verwendet werden, desto mehr Stichprobenumfang wird benötigt, um das nominelle Niveau einzuhalten. Aber selbst bei nur drei Kategorien hält die parametrische Kovarianzanalyse das Niveau schlechter ein, als die nichtparametrische Analyse. Ein Vergleich der Macht der beiden kovarianzanalytischen Verfahren erübrigt sich, da sie unterschiedliche Testprobleme betrachten.

Kapitel 8

Zusammenfassung und Ausblick

Zum Abschluss dieser Arbeit sollen nun die wichtigsten Ergebnisse zusammengefasst werden. Es hat sich gezeigt, dass die Modellannahmen der nichtparametrischen Kovarianzanalyse sehr schwach sind. Sie ähneln den Annahmen der anderen nichtparametrischen Verfahren und sind wesentlich schwächer als die restriktiven Voraussetzungen der parametrischen Verfahren. Des Weiteren sind die Annahmen an die Skala so schwach, dass zum einen die Anzahl der Kategorien keine Rolle spielt und zum anderen auch stetige Skalen betrachtet werden können.

Auch in Bezug auf die Interpretierbarkeit der Parameter hat sich ergeben, dass die nichtparametrische Kovarianzanalyse zu den besten Verfahren gehört. Einzig das Odds Ratio der parametrischen Verfahren kann sich in der Interpretierbarkeit mit dem relativen Effekt p_2 messen. Dabei besteht der Vorteil des Odds Ratios vor allem in der weiteren Verbreitung und nicht in der besseren Interpretierbarkeit. Wichtig ist hierbei zu bemerken, dass die nichtparametrischen Verfahren sich verbessert haben. Seit den Arbeiten von Siemer (1999) und Domhof (2001) ist es möglich, die Hypothesen der nichtparametrischen Kovarianzanalyse nicht mehr über die Verteilungsfunktionen sondern über die interessierenden Parameter, die relativen Effekte, zu stellen. Bei nicht randomisierten Studien ist es nun zudem möglich, die Verzerrung durch die Baseline-Werte zu berücksichtigen.

Besonders gute Eigenschaften hat die nichtparametrische Kovarianzanalyse auch bei der praktischen Umsetzung. Sie hält das Niveau schon bei geringen Stichprobenumfängen ein und hat eine hohe Macht. Weiterhin ist die Berechnung der Schätzer und Teststatistiken sehr einfach und fast unabhängig von der Anzahl der Kategorien. Dies ist ein großer Vorteil gegenüber der Schichtungsanalyse und den parametrischen Verfahren. Es ergeben sich im Gegensatz zu den parametrischen Verfahren keine numerischen Probleme bei der Verwendung der nichtparametrischen Kovarianzanalyse, da keine iterativen Verfahren notwendig sind. Dies ist von besonderer Bedeutung bei der Verwendung einer Skala

mit vielen Kategorien wie zum Beispiel der Panik- und Agoraphobieskala von Bandelow (1997). Bei dieser Skala versagen die parametrischen Verfahren völlig.

Abschließend sollen nun einige Anregungen für Forschungen gegeben werden, die an das Thema dieser Arbeit beziehungsweise an die diskutierten Verfahren anknüpfen können. So wurde beim nichtparametrischen faktoriellen Ansatz der Parameter p_W^M vorgestellt. Dieser berücksichtigt in einem randomisierten Versuchsplan die Verteilung der zu Grunde liegenden Population. Entsprechend der vorhandenen Theorie können Schätzer unter Testverfahren bezüglich dieses Parameters entwickelt werden. Dabei sind einige Verbesserungen zu erwarten, da man die Information $F_{11} = F_{21}$ bei der Schätzung verwenden kann.

Des Weiteren wurde in dieser Arbeit nur der Fall von zwei Gruppen betrachtet. Mit Hilfe der dargestellten Theorie könnten die Schichtungsverfahren auf den Fall mehrerer Gruppen erweitert werden. Da dann mehr Versuchseinheiten pro Schicht vorhanden sind, könnten sich die Eigenschaften der Verfahren verbessern. Andererseits müssen dann in jeder Schicht auch Beobachtungen aus jeder Gruppe vorhanden sein, was hingegen zu Problemen führen könnte. Außerdem ist dabei zu untersuchen, inwiefern die Voraussetzung der Randomisierung bei der Bildung von Schichten notwendig ist, da die Versuchseinheiten innerhalb der Schichten auf Grund der gleichen Ausgangswerte vergleichbar sind.

Insgesamt werden die Rangverfahren im Vergleich zu den parametrischen Verfahren wenig angewendet. Eine vergleichende Auswertung bei praktischen Beispielen könnte die Akzeptanz der nichtparametrischen Verfahren gerade bei kleinen Stichprobenumfängen erhöhen. Dabei sollte die gute Interpretierbarkeit der relativen Effekte, vor allem im Zwei-Stichproben-Fall, herausgestellt werden. Ein zusätzliches Problem bei der Akzeptanz der Rangverfahren ist die fehlende Standardsoftware. Dort sind noch erhebliche Mühen erforderlich. Auch in Bezug auf einen theoretischen Vergleich der Macht der verschiedenen Verfahren sind noch viele Fragen offen. Als Ansatzpunkt könnte eventuell die Theorie aus der Arbeit von Bajorski und Petkau (1999) verwendet werden. Ein wichtiges Problem besteht dabei zunächst in der sinnvollen Formulierung von Alternativen.

Besondere Aufmerksamkeit ist auf die Berücksichtigung von nichtparametrischen Verfahren im Zulassungsprozess von pharmazeutischen Produkten zu richten. In der Einleitung wurde bereits darauf hingewiesen, dass zur Zeit an einem *Points to Consider Paper* gearbeitet wird. Auch wenn die Empfehlungen dieser Richtlinie sehr allgemein sein werden, so wird die weitere Diskussion der ICH-Richtlinien trotzdem die Akzeptanz der verschiedenen Verfahren besonders beeinflussen.

Anhang A

Test- und Schätzverfahren

A.1 Nichtparametrische faktorielle Analyse

Die Schätzer der relativen Effekte erhält man mit Hilfe des Einsetzungsprinzips. Der gewichtete relative Behandlungseffekt $p_{N,ij}$ wird mithin durch

$$\widehat{p}_{N,ij} = \int \widehat{H}_N d\widehat{F}_{ij} \quad , i,j = 1,2,$$

und der ungewichtete relative Behandlungseffekt $\widehat{p}_{ij}$ durch

$$\widehat{p}_{ij} = \int \widehat{H} d\widehat{F}_{ij} \quad , i,j = 1,2,$$

geschätzt. Wie die relativen Effekte und die Verteilungsfunktionen selbst werden auch die Schätzer in Vektoren angeordnet:

$$\begin{aligned}
\boldsymbol{p}_N &= (p_{N,11}, p_{N,12}, p_{N,21}, p_{N,22})^t, \\
\boldsymbol{p} &= (p_{11}, p_{12}, p_{21}, p_{22})^t, \\
\boldsymbol{F} &= (F_{11}, F_{12}, F_{21}, F_{22})^t, \\
\widehat{\boldsymbol{p}}_N &= (\widehat{p}_{N,11}, \widehat{p}_{N,12}, \widehat{p}_{N,21}, \widehat{p}_{N,22})^t, \\
\widehat{\boldsymbol{p}} &= (\widehat{p}_{11}, \widehat{p}_{12}, \widehat{p}_{21}, \widehat{p}_{22})^t \quad \text{und} \\
\widehat{\boldsymbol{F}} &= \left(\widehat{F}_{11}, \widehat{F}_{12}, \widehat{F}_{21}, \widehat{F}_{22}\right)^t
\end{aligned}$$

angeordnet. Die Konsistenz dieser Schätzer für die relativen Effekte wird im folgenden Satz formuliert.

Satz A.1.1
Unter der Annahme (4.4) gilt für $i, j = 1, 2$

$$E(\widehat{p}_{N,ij} - p_{N,ij})^2 \quad \rightarrow \quad 0 \quad \textit{und} \quad E(\widehat{p}_{ij} - p_{ij})^2 \quad \rightarrow \quad 0.$$

Beweis: Der Beweis folgt aus den Beweisen zu den Sätzen 3.3 und 4.5 in Siemer (1999).

□

Aus der Konsistenz folgt die asymptotische Erwartungstreue der Schätzer. Um zu zeigen, wieso die Schätzer im finiten Fall nicht erwartungstreu sind, betrachten wir zunächst die Zerlegungen

$$\begin{aligned}
\widehat{p}_{N,ij} &= \frac{1}{N}\left[\sum_{v=1}^{2}\frac{1}{n_i}\sum_{k=1}^{n_i} c(X_{ikj}; X_{ikv}) + \sum_{v=1}^{2}\frac{1}{n_i}\sum_{k=1}^{n_i}\sum_{s=1\ s\neq k}^{n_i} c(X_{ikj}; X_{isv})\right. \\
&\quad \left. + \sum_{v=1}^{2}\frac{1}{n_i}\sum_{k=1}^{n_i}\sum_{s=1}^{n_t} c(X_{ikj}; X_{tsv})\right] \quad \text{und} \\
\widehat{p}_{ij} &= \frac{1}{4}\left[\sum_{v=1}^{2}\frac{1}{n_i^2}\sum_{k=1}^{n_i} c(X_{ikj}; X_{ikv}) + \sum_{v=1}^{2}\frac{1}{n_i^2}\sum_{k=1}^{n_i}\sum_{s=1\ s\neq k}^{n_i} c(X_{ikj}; X_{isv})\right. \\
&\quad \left. + \sum_{v=1}^{2}\frac{1}{n_i n_t}\sum_{k=1}^{n_i}\sum_{s=1}^{n_t} c(X_{ikj}; X_{tsv})\right]
\end{aligned}$$

mit $i \neq t$. Bildet man nun den Erwartungswert dieser Summanden, so erhält man für die Zählfunktionen c, in denen die Zufallsvariablen unabhängig sind, die gewünschten Integrale:

$$\begin{aligned}
\mathrm{E}(c(X_{ikj}; X_{isv})) &= \int F_{iv} dF_{ij} \quad \text{mit} \quad k \neq s \quad \text{und} \\
\mathrm{E}(c(X_{ikj}; X_{tsv})) &= \int F_{tv} dF_{ij} \quad \text{mit} \quad i \neq t.
\end{aligned}$$

Die Erwartungswerte der Zählfunktionen, in denen jedoch abhängige Zufallsvariablen stehen, ergeben den Vorzeicheneffekt:

$$\Delta_{ijv} \quad = \quad \mathrm{E}(c(X_{ikj}; X_{ikv})) \quad = \quad P(X_{ikj} < X_{ikv}) + \frac{1}{2}\, P(X_{ikj} = X_{ikv}).$$

Dies ist der Effekt, welcher im Vorzeichentest (siehe Dixon und Mood,1946; und die Literaturangaben dort) verwendet wird. Da er aber in der Regel ungleich des Integrals $\int F_{ij} dF_{iv}$

ist, kommt es zu einer Verzerrung des relativen Effektes. Köhler (2001) bezeichnet Δ_{ijv} als Teil-Individualeffekt bezogen auf zwei Situationen j und v und eine Versuchseinheit der Gruppe i. Daher ist der Schätzer $\widehat{p}_{N,ij}$ um $1/N\sum_{v=1}^{2}\Delta_{ijv} - \int F_{iv}dF_{ij}$ und $\widehat{p}_{ij}$ um $1/(4n_i)\sum_{v=1}^{2}\Delta_{ijv} - \int F_{iv}dF_{ij}$ verzerrt. Die Verzerrung ist damit im gewichteten Fall maximal $2/N$ und im ungewichteten Fall maximal $1/(2n_i)$. Köhler (2001) schlägt daher vor, dass man für einen unverzerrten Schätzer die Zählfunktionen mit abhängigen Zufallsvariablen nicht berücksichtigen sollte. Damit erhält man die unverzerrten Schätzer

$$\widehat{p}^{u}_{N,ij} = \frac{1}{Nn_i}\sum_{k=1}^{n_i}\left(\sum_{v=1}^{2}\sum_{s=1}^{n_t} c(X_{ikj};X_{tsv}) + \sum_{v=1}^{2}\frac{n_i}{n_i-1}\sum_{s=1\ s\neq k}^{n_i} c(X_{ikj};X_{isv})\right)$$

und

$$\widehat{p}^{u}_{ij} = \frac{1}{4n_i}\sum_{k=1}^{n_i}\left(\sum_{v=1}^{2}\frac{1}{n_t}\sum_{s=1}^{n_t} c(X_{ikj};X_{tsv}) + \sum_{v=1}^{2}\frac{1}{n_i-1}\sum_{s=1\ s\neq k}^{n_i} c(X_{ikj};X_{isv})\right)$$

mit $i \neq t$ für $i,j = 1,2$. Da jedoch die Unterschiede zwischen den Schätzern sehr gering sind, werden im Weiteren die Standardschätzer $\widehat{p}_{N,ij}$ und $\widehat{p}_{ij}$ verwendet. Außerdem verschwinden die Verzerrungen asymptotisch, wie man an den Abschätzungen leicht erkennt.

Für die Berechnung der relativen Behandlungseffekte ist eine Darstellung mit Hilfe von Rängen sehr nützlich, da diese durch effiziente Sortieralgorithmen sehr schnell berechnet werden können. Daher werden zunächst die Ränge definiert. Der Gesamtrang R_{ikj}, oder kurz Rang, von X_{ikj} unter allen N Zufallsvariablen X_{tsv} ist

$$\begin{aligned} R_{ikj} &= \frac{1}{2} + N\widehat{H}_N(X_{ikj}) \\ &= \frac{1}{2} + \sum_{t=1}^{2}\sum_{v=1}^{2}\sum_{s=1}^{n_t} c\left(X_{ikj};X_{tsv}\right). \end{aligned}$$

Mit

$$R_{ikj(ij)} = \frac{1}{2} + \sum_{s=1}^{n_i} c\left(X_{ikj};X_{isj}\right)$$

wird der Internrang von X_{ikj} unter allen n_i Zufallsvariablen X_{isj} in der Gruppe i zum Zeitpunkt j bezeichnet. Zuletzt definiert

$$R_{ikj(-uw)} = \frac{1}{2} + \sum_{(t,v)\neq(u,w)}\sum_{s=1}^{n_t} c\left(X_{ikj};X_{tsv}\right)$$

den Teilrang von X_{ikj} unter allen $N - n_u$ Zufallsvariablen ohne die Gruppe u zum Zeitpunkt w. Der Summand $\frac{1}{2}$ in den Definitionen rührt von der Definition der Zählfunktion $c(x; y)$ an der Stelle $x = y$ her. So ergeben sich die bekannten Mittelränge.

Nun können die Schätzer der relativen Effekte mit Hilfe der Ränge dargestellt werden. Es gelten die Zusammenhänge

$$\begin{aligned}
\widehat{p}_{N,ij} &= \frac{1}{N}\left(\frac{1}{n_i}\sum_{k=1}^{n_i} R_{ikj} - \frac{1}{2}\right) \quad \text{und} \\
\widehat{p}_{ij} &= \frac{1}{4n_i}\sum_{k=1}^{n_i}\left[\frac{1}{n_i}\left(R_{ikj(ij)} - \frac{1}{2}\right) + \sum_{(t,v)\neq(i,j)} \frac{1}{n_t}\left(R_{ikj} - R_{ikj(-tv)}\right)\right].
\end{aligned}$$

Hier wird noch einmal deutlich, warum der gewichtete relative Effekt verbreiteter ist. Seine Berechnung über die Ränge ist wesentlich einfacher. Daher führten Kulle (1999) und Siemer (1999) die harmonischen Ränge ein. Sei $1/\widetilde{n} = (1/n_1 + 1/n_2)/2$, dann heißt

$$R_{(h)ikj} = \frac{\widetilde{n}}{2n_i} + \sum_{t=1}^{2}\sum_{v=1}^{2}\sum_{s=1}^{n_t} \frac{\widetilde{n}}{n_t}\, c\left(X_{ikj}; X_{tsv}\right)$$

harmonischer Gesamtrang, oder kurz harmonischer Rang, von X_{ikj} unter allen N Zufallsvariablen X_{tsv}. Mithin ergibt sich der Schätzer für den ungewichteten relativen Effekt aus

$$\widehat{p}_{ij} = \frac{1}{4\widetilde{n}}\left(\frac{1}{n_i}\sum_{k=1}^{n_i} R_{(h)ikj} - \frac{\widetilde{n}}{2n_i}\right).$$

Wie der gewichtete relative Effekt sich im Wesentlichen aus dem Mittel der Ränge in der entsprechenden Faktorstufenkombination ergibt, ist der ungewichtete relative Effekt im Grunde das Mittel der harmonischen Ränge. Der Zusammenhang zwischen den Rängen und den harmonischen Rängen wird in Kulle (1999) und Siemer (1999) durch

$$R_{(h)ikj} = \sum_{(t,v)\neq(i,j)} \frac{\widetilde{n}}{n_t}\left(R_{ikj} - R_{ikj(-tv)}\right) + \frac{\widetilde{n}}{n_i} R_{ikj(ij)}$$

gegeben. Nun wurden also Punktschätzer für die relativen Effekte vorgestellt. Um asymptotische Verteilungsaussagen über diese Schätzer treffen zu können, werden die folgenden asymptotischen Äquivalenzen benötigt.

Satz A.1.2
Unter den Annahmen (4.4) und (4.5) gilt

$$\begin{aligned}\sqrt{N}\int \widehat{H}_N d\left(\widehat{\boldsymbol{F}}-\boldsymbol{F}\right) &\doteqdot \sqrt{N}\int H_N d\left(\widehat{\boldsymbol{F}}-\boldsymbol{F}\right) \quad \textit{und}\\ \sqrt{N}\left(\widehat{\boldsymbol{p}}-\boldsymbol{p}\right) &\doteqdot \sqrt{N}\left(\int H d\widehat{\boldsymbol{F}}+\int \widehat{H} d\boldsymbol{F}-2\boldsymbol{p}\right).\end{aligned}$$

Sei $\boldsymbol{C}$ ein Kontrastvektor, dann gilt insbesondere unter der Hypothese $\boldsymbol{CF} = \boldsymbol{0}$ beziehungsweise $\boldsymbol{Cp} = \boldsymbol{0}$

$$\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}_N \doteqdot \sqrt{N}\boldsymbol{C}\int H_N d\widehat{\boldsymbol{F}} \quad \textit{beziehungsweise} \tag{A.1}$$

$$\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}} \doteqdot \sqrt{N}\boldsymbol{C}\left(\int H d\widehat{\boldsymbol{F}}-\int \boldsymbol{F} d\widehat{H}\right). \tag{A.2}$$

Beweis: Siehe die Sätze 4.12 und 4.14 in Siemer (1999). □

Unter der Hypothese $\boldsymbol{CF} = \boldsymbol{0}$ beziehungsweise $\boldsymbol{Cp} = \boldsymbol{0}$ sind die Rangstatistiken also äquivalent zu Summen von unabhängigen und beschränkten Zufallsvariablen. Denn für die Integrale auf der rechten Seite der asymptotischen Äquivalenzen gilt im ersten Fall

$$\int H_n d\widehat{F}_{ij} = \frac{1}{n_i}\sum_{k=1}^{n_i} H_N\left(X_{ikj}\right).$$

Des Weiteren gilt im zweiten Fall

$$\begin{aligned}&\int H d\widehat{F}_{ij}-\int F_{ij} d\widehat{H}\\ &= \frac{1}{n_i}\sum_{k=1}^{n_i}\left[H\left(X_{ikj}\right)-\frac{1}{4}\sum_{j'=1}^{2}F_{ij}\left(X_{ikj'}\right)\right]+\frac{1}{n_{i'}}\sum_{k=1}^{n_{i'}}\left[\frac{-1}{4}\sum_{j'=1}^{2}F_{ij}\left(X_{i'kj'}\right)\right]\end{aligned}$$

mit $i \neq i'$. Für diese unabhängigen und beschränkten Zufallsvariablen gilt nach dem zentralen Grenzwertsatz die asymptotische Normalität, womit sich auch die asymptotischen Verteilungen der Rangstatistiken ergeben. Dieses Ergebnis wird im folgenden Satz formuliert.

Satz A.1.3
Sei $\boldsymbol{C}$ ein Kontrastvektor. Unter den Annahmen (4.2), (4.4), (4.5), (4.15) und unter der Hypothese $\boldsymbol{CF} = \boldsymbol{0}$ beziehungsweise $\boldsymbol{Cp} = \boldsymbol{0}$ gilt:

1. $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}_N/\sqrt{\boldsymbol{C}\boldsymbol{\Sigma}^F_{rm}\boldsymbol{C}^t}$ *ist asymptotisch standardnormalverteilt mit*

$$\boldsymbol{\Sigma}^F_{rm} = Cov\left(\sqrt{N}\int H_N d\widehat{\boldsymbol{F}}\right).$$

2. *Ist* $\widehat{\boldsymbol{\Sigma}}^F_{rm}$ *ein konsistenter Schätzer von* $\boldsymbol{\Sigma}^F_{rm}$*, so ist* $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}_N/\sqrt{\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}^F_{rm}\boldsymbol{C}^t}$ *asymptotisch standardnormalverteilt.*

3. $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}/\sqrt{\boldsymbol{C}\boldsymbol{\Sigma}^p_{rm}\boldsymbol{C}^t}$ *ist asymptotisch standardnormalverteilt mit*

$$\boldsymbol{\Sigma}^p_{rm} = Cov\left(\sqrt{N}\left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)\right).$$

4. *Ist* $\widehat{\boldsymbol{\Sigma}}^p_{rm}$ *ein konsistenter Schätzer von* $\boldsymbol{\Sigma}^p_{rm}$*, so ist* $\sqrt{N}\boldsymbol{C}\widehat{\boldsymbol{p}}/\sqrt{\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}^p_{rm}\boldsymbol{C}^t}$ *asymptotisch standardnormalverteilt.*

□

Bei diesem Satz wird davon ausgegangen, dass für $\boldsymbol{C}$ die Kontrastvektoren

$$\begin{aligned}
\boldsymbol{C}_G &= (+1,+1,-1,-1),\\
\boldsymbol{C}_T &= (+1,-1,+1,-1) \quad \text{und}\\
\boldsymbol{C}_W &= (+1,-1,-1,+1)
\end{aligned}$$

verwendet werden. Werden hingegen mehrzeilige Kontrastmatrizen mit Rängen größer als eins verwendet, gelten die analogen Aussagen für die multivariate Normalverteilung. Des Weiteren ist zu bemerken, dass die beiden Kovarianzmatrizen $\boldsymbol{\Sigma}^F_{rm}$ und $\boldsymbol{\Sigma}^p_{rm}$ sich grundsätzlich in ihrer Struktur unterscheiden. Dies liegt an den unterschiedlichen Integratoren. Bei $\boldsymbol{\Sigma}^F_{rm}$ wird nur $\widehat{\boldsymbol{F}}$ verwendet, somit bleibt die Abhängigkeitsstruktur erhalten. Hingegen kommt bei $\boldsymbol{\Sigma}^p_{rm}$ noch $\widehat{H}$ dazu und die Abhängigkeitsstruktur zwischen den Komponenten der Vektoren ändert sich. Damit erhält man

$$\boldsymbol{\Sigma}^F_{rm} = \begin{pmatrix} \sigma^F_{rm;11,11} & \sigma^F_{rm;11,12} & 0 & 0 \\ \sigma^F_{rm;11,12} & \sigma^F_{rm;12,12} & 0 & 0 \\ 0 & 0 & \sigma^F_{rm;21,21} & \sigma^F_{rm;21,22} \\ 0 & 0 & \sigma^F_{rm;21,22} & \sigma^F_{rm;22,22} \end{pmatrix}$$

und

$$\mathbf{\Sigma}_{rm}^{p} = \begin{pmatrix} \sigma_{rm;11,11}^{p} & \sigma_{rm;11,12}^{p} & \sigma_{rm;11,21}^{p} & \sigma_{rm;11,22}^{p} \\ \sigma_{rm;11,12}^{p} & \sigma_{rm;12,12}^{p} & \sigma_{rm;12,21}^{p} & \sigma_{rm;12,22}^{p} \\ \sigma_{rm;11,21}^{p} & \sigma_{rm;12,21}^{p} & \sigma_{rm;21,21}^{p} & \sigma_{rm;21,22}^{p} \\ \sigma_{rm;11,22}^{p} & \sigma_{rm;12,22}^{p} & \sigma_{rm;21,22}^{p} & \sigma_{rm;22,22}^{p} \end{pmatrix}.$$

Für die Komponenten von $\mathbf{\Sigma}_{rm}^{F}$ können sofort die folgenden konsistenten Schätzer angegeben werden. Diese ergeben sich kanonisch aus den Rängen. Die Varianzen werden durch

$$\widehat{\sigma}_{rm;ij,ij}^{F} = \frac{1}{Nn_i(n_i-1)} \left[\sum_{k=1}^{n_i} \left(R_{ikj} - \overline{R}_{i\cdot j} \right)^2 \right]$$

geschätzt. Hingegen ergeben sich die Schätzer der Kovarianzen aus

$$\widehat{\sigma}_{rm;ij,ij'}^{F} = \frac{1}{Nn_i(n_i-1)} \left[\sum_{k=1}^{n_i} \left(R_{ikj} - \overline{R}_{i\cdot j} \right) \left(R_{ikj'} - \overline{R}_{i\cdot j'} \right) \right].$$

Dabei bezeichnet $\overline{R}_{i\cdot j} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{ikj}$ das Mittel aller Ränge in der Gruppe i zum Zeitpunkt j. Um die Schätzer für $\mathbf{\Sigma}_{rm}^{p}$ herzuleiten, wird zunächst die Zerlegung einer Komponente des Vektors $\sqrt{N}\left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)$ in unabhängige Zufallsvariablen betrachtet. Dafür sei $i \neq i'$ und $j \neq j'$:

$$\begin{aligned} &\sqrt{N}\left(\int H d\widehat{F}_{ij} - \int F_{ij} d\widehat{H}\right) \\ &= \frac{\sqrt{N}}{4}\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(-F_{ij}\left(X_{ikj'}\right) + \sum_{(t,v)\neq(i,j)} F_{tv}\left(X_{ikj}\right)\right)\right] \\ &\quad -\frac{\sqrt{N}}{4}\left[\frac{1}{n_{i'}}\sum_{k=1}^{n_{i'}}\sum_{j'=1}^{2} F_{ij}\left(X_{i'kj'}\right)\right] \\ &= \frac{\sqrt{N}}{4}\left[\frac{1}{n_i}\sum_{k=1}^{n_i} C_{ikj} + \frac{1}{n_{i'}}\sum_{k=1}^{n_{i'}} D_{ikj}\right]. \end{aligned}$$

Für die Zerlegung gelten die Bezeichnungen

$$\begin{aligned} C_{ikj} &= -F_{ij}\left(X_{ikj'}\right) + \sum_{(t,v)\neq(i,j)} F_{tv}\left(X_{ikj}\right) \quad \text{und} \\ D_{ikj} &= -\sum_{j'=1}^{2} F_{ij}\left(X_{i'kj'}\right). \end{aligned}$$

Es ist zu beachten, dass C_{ikj} und D_{ikj} unabhängig sind, da $i \neq i'$ vorausgesetzt wurde. Damit ergeben sich zwei Summen von unabhängigen Zufallsvariablen und man kann die Komponenten von $\mathbf{\Sigma}^p_{rm}$ wie folgt zerlegen. Für die Varianzen, also die Diagonalelemente der Kovarianzmatrix gilt

$$\begin{aligned}
\sigma^p_{rm;ij,ij} &= \frac{N}{16}\left[\frac{1}{n_i}\mathrm{Var}\left(C_{ikj}\right) + \frac{1}{n_{i'}}\mathrm{Var}\left(D_{ikj}\right)\right] \\
&= \frac{N}{16}\left[\frac{1}{n_i}\sigma^p_{rm;ij,ij;CC} + \frac{1}{n_{i'}}\sigma^p_{rm;ij,ij;DD}\right] \quad \text{mit} \\
\sigma^p_{rm;ij,ij;CC} &= \mathrm{Var}\left(C_{ikj}\right) \quad \text{und} \\
\sigma^p_{rm;ij,ij;DD} &= \mathrm{Var}\left(D_{ikj}\right).
\end{aligned}$$

Bei den Kovarianzen wird der Fall $i = i'$

$$\begin{aligned}
\sigma^p_{rm;ij,ij'} &= \frac{N}{16}\left[\frac{1}{n_i}\mathrm{Cov}\left(C_{ikj}, C_{ikj'}\right) + \frac{1}{n_{i'}}\mathrm{Cov}\left(D_{ikj}, D_{ikj'}\right)\right] \\
&= \frac{N}{16}\left[\frac{1}{n_i}\sigma^p_{rm;ij,ij';CC} + \frac{1}{n_{i'}}\sigma^p_{rm;ij,ij';DD}\right] \quad \text{mit} \\
\sigma^p_{rm;ij,ij';CC} &= \mathrm{Cov}\left(C_{ikj}, C_{ikj'}\right) \quad \text{und} \\
\sigma^p_{rm;ij,ij';DD} &= \mathrm{Cov}\left(D_{ikj}, D_{ikj'}\right)
\end{aligned}$$

von dem Fall $i \neq i'$

$$\begin{aligned}
\sigma^p_{rm;ij,i'v} &= \frac{N}{16}\left[\frac{1}{n_i}\mathrm{Cov}\left(C_{ikj}, D_{i'kv}\right) + \frac{1}{n_{i'}}\mathrm{Cov}\left(D_{ikj}, C_{i'kv}\right)\right] \\
&= \frac{N}{16}\left[\frac{1}{n_i}\sigma^p_{rm;ij,i'v;CD} + \frac{1}{n_{i'}}\sigma^p_{rm;ij,i'v;DC}\right] \quad \text{mit} \\
\sigma^p_{rm;ij,i'v;CD} &= \mathrm{Cov}\left(C_{ikj}, D_{i'kv}\right) \quad \text{und} \\
\sigma^p_{rm;ij,i'v;DC} &= \mathrm{Cov}\left(D_{ikj}, C_{i'kv}\right)
\end{aligned}$$

unterschieden. Die Zufallsvariablen C_{ikj} und D_{ikj} sind nicht beobachtbar, da Transformationen der Zufallsvariablen X_{ikj} mit den nicht bekannten Marginalverteilungen verwendet werden. Daher benutzt man zur Schätzung der Varianzen und Kovarianzen die empirischen Gegenstücke. Man ersetzt also die Marginalverteilungen durch die empirischen Marginalverteilungen. Mithin ergeben sich

$$\begin{aligned}
\widehat{C}_{ikj} &= -\widehat{F}_{ij}\left(X_{ikj'}\right) + \sum_{(t,v)\neq(i,j)} \widehat{F}_{tv}\left(X_{ikj}\right) \\
&= \frac{-1}{n_i}\left(R_{ikj'} - R_{ikj'(-ij)}\right) + \sum_{(t,v)\neq(i,j)} \frac{1}{n_{i'}}\left(R_{ikj} - R_{ikj(-i'j')}\right) \quad \text{und}
\end{aligned}$$

$$\begin{aligned}\widehat{D}_{ikj} &= -\sum_{j'=1}^{2} \widehat{F}_{ij}\left(X_{i'kj'}\right) \\ &= \frac{-1}{n_i}\sum_{j'=1}^{2}\left(R_{i'kj'} - R_{i'kj'(-ij)}\right).\end{aligned}$$

Nun können kanonisch die empirischen Varianzen beziehungsweise Kovarianzen berechnet werden. Mit

$$\begin{aligned}\widehat{\sigma}^p_{rm;ij,ij;CC} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(\widehat{C}_{ikj} - \widehat{\overline{C}}_{i\cdot j}\right)^2, \\ \widehat{\sigma}^p_{rm;ij,ij;DD} &= \frac{1}{n_{i'}-1}\sum_{k=1}^{n_{i'}}\left(\widehat{D}_{ikj} - \widehat{\overline{D}}_{i\cdot j}\right)^2, \\ \widehat{\sigma}^p_{rm;ij,ij';CC} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(\widehat{C}_{ikj} - \widehat{\overline{C}}_{i\cdot j}\right)\left(\widehat{C}_{ikj'} - \widehat{\overline{C}}_{i\cdot j'}\right), \\ \widehat{\sigma}^p_{rm;ij,ij';DD} &= \frac{1}{n_{i'}-1}\sum_{k=1}^{n_{i'}}\left(\widehat{D}_{ikj} - \widehat{\overline{D}}_{i\cdot j}\right)\left(\widehat{D}_{ikj'} - \widehat{\overline{D}}_{i\cdot j'}\right), \\ \widehat{\sigma}^p_{rm;ij,i'v;CD} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(\widehat{C}_{ikj} - \widehat{\overline{C}}_{i\cdot j}\right)\left(\widehat{D}_{i'kv} - \widehat{\overline{D}}_{i'\cdot v}\right) \quad \text{und} \\ \widehat{\sigma}^p_{rm;ij,i'v;DC} &= \frac{1}{n_{i'}-1}\sum_{k=1}^{n_{i'}}\left(\widehat{D}_{ikj} - \widehat{\overline{D}}_{i\cdot j}\right)\left(\widehat{C}_{i'kv} - \widehat{\overline{C}}_{i'\cdot v}\right)\end{aligned}$$

erhält man

$$\begin{aligned}\widehat{\sigma}^p_{rm;ij,ij} &= \frac{N}{16}\left[\frac{1}{n_i}\widehat{\sigma}^p_{rm;ij,ij;CC} + \frac{1}{n_{i'}}\widehat{\sigma}^p_{rm;ij,ij;DD}\right], \\ \widehat{\sigma}^p_{rm;ij,ij'} &= \frac{N}{16}\left[\frac{1}{n_i}\widehat{\sigma}^p_{rm;ij,ij';CC} + \frac{1}{n_{i'}}\widehat{\sigma}^p_{rm;ij,ij';DD}\right] \quad \text{und} \\ \widehat{\sigma}^p_{rm;ij,i'v} &= \frac{N}{16}\left[\frac{1}{n_i}\widehat{\sigma}^p_{rm;ij,i'v;CD} + \frac{1}{n_{i'}}\widehat{\sigma}^p_{rm;ij,i'v;DC}\right].\end{aligned}$$

Die Konsistenz der so berechneten Schätzer der Kovarianzmatrizen sichert der folgende Satz.

Satz A.1.4
Es gilt unter der Annahme (4.4) für $i, i', j, j' = 1, 2$

$$E\left(\widehat{\sigma}^F_{rm;ij,ij'} - \sigma^F_{rm;ij,ij'}\right)^2 \quad \rightarrow \quad 0 \quad \textit{und} \quad E\left(\widehat{\sigma}^p_{rm;ij,i'j'} - \sigma^p_{rm;ij,i'j'}\right)^2 \quad \rightarrow \quad 0.$$

Beweis: Unter mehrfacher Verwendung der Lemmata A.2, A.3 und A.4 in Siemer (1999) ergeben sich die Konsistenzen. □

Häufig halten die Teststatistiken das Niveau bei kleinen Stichprobenumfängen nicht ein. So schlagen Brunner und Langer (1999) Approximationen der Verteilungen der Teststatistiken mit Hilfe der t-Verteilung vor. Dies geschieht mit der Technik aus Smith (1936), Welch (1938) und Satterthwaite (1946). Dazu werden die Varianzen der Zufallsvariablen

$$\sqrt{N}\boldsymbol{C}\int H_N d\widehat{\boldsymbol{F}} \quad \text{und} \quad \sqrt{N}\boldsymbol{C}\left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)$$

jeweils für die Kontrastmatrizen $\boldsymbol{C}_G$, $\boldsymbol{C}_T$ und $\boldsymbol{C}_W$ betrachtet. Es gilt für die Kontrastmatrix $\boldsymbol{C}_G$ der Hypothese über den Gruppeneffekt mit $Y_{ikj} = H_N(X_{ikj})$

$$\begin{aligned}
\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_G \int H_N d\widehat{\boldsymbol{F}}\right) &= \frac{N}{n_1}\operatorname{Var}\left(Y_{111}+Y_{112}\right) + \frac{N}{n_2}\operatorname{Var}\left(Y_{211}+Y_{212}\right) \\
&= \sum_{i=1}^{2}\frac{N}{n_i}\tau_{G,i}^F \quad \text{mit} \\
\tau_{G,i}^F &= \operatorname{Var}\left(Y_{i11}+Y_{i12}\right), \quad i = 1,2.
\end{aligned}$$

Damit wurde die Varianz $\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_G \int H_N d\widehat{\boldsymbol{F}}\right)$ in zwei Varianzen $\tau_{G,1}^F$ und $\tau_{G,2}^F$ zerlegt. Dieses wird auch mit den anderen Statistiken durchgeführt. Dabei ist zu bemerken, dass die Varianz für die Teststatistik beim Zeiteffekt mit der für die Wechselwirkung übereinstimmt. Man erhält die Summe von zwei Varianzen durch

$$\begin{aligned}
\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_T \int H_N d\widehat{\boldsymbol{F}}\right) &= \frac{N}{n_1}\operatorname{Var}\left(Y_{111}-Y_{112}\right) + \frac{N}{n_2}\operatorname{Var}\left(Y_{211}-Y_{212}\right) \\
&= \sum_{i=1}^{2}\frac{N}{n_i}\tau_{T,i}^F \\
&= \operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_W \int H_N d\widehat{\boldsymbol{F}}\right) \quad \text{mit} \\
\tau_{T,i}^F &= \operatorname{Var}\left(Y_{i11}-Y_{i12}\right), \quad i = 1,2.
\end{aligned}$$

Im ungewichteten Fall für die Hypothese über den relativen Gruppeneffekt erhält man

$$\begin{aligned}
&\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_G\left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)\right) \\
&= \frac{N}{16}\left[\frac{1}{n_1}\operatorname{Var}\left(C_{111}+C_{112}-D_{211}-D_{212}\right)\right. \\
&\quad \left. + \frac{1}{n_2}\operatorname{Var}\left(D_{111}+D_{112}-C_{211}-C_{212}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{2} \frac{N}{16 n_i} \tau_{A,i}^{p} \quad \text{mit} \\
\tau_{G,1}^{p} &= \operatorname{Var}\left(C_{111} + C_{112} - D_{211} - D_{212}\right) \quad \text{und} \\
\tau_{G,2}^{p} &= \operatorname{Var}\left(D_{111} + D_{112} - C_{211} - C_{212}\right).
\end{aligned}
$$

Analog ergibt sich für die Hypothese über den relativen Zeiteffekt

$$
\begin{aligned}
&\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_T \left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)\right) \\
&= \frac{N}{16}\left[\frac{1}{n_1}\operatorname{Var}\left(C_{111} - C_{112} + D_{211} - D_{212}\right)\right. \\
&\quad \left. + \frac{1}{n_2}\operatorname{Var}\left(D_{111} - D_{112} + C_{211} - C_{212}\right)\right] \\
&= \sum_{i=1}^{2} \frac{N}{16 n_i} \tau_{T,i}^{p} \quad \text{mit} \\
\tau_{T,1}^{p} &= \operatorname{Var}\left(C_{111} - C_{112} + D_{211} - D_{212}\right) \quad \text{und} \\
\tau_{T,2}^{p} &= \operatorname{Var}\left(D_{111} - D_{112} + C_{211} - C_{212}\right),
\end{aligned}
$$

und für die Hypothese über den relativen Wechselwirkungseffekt

$$
\begin{aligned}
&\operatorname{Var}\left(\sqrt{N}\boldsymbol{C}_W \left(\int H d\widehat{\boldsymbol{F}} - \int \boldsymbol{F} d\widehat{H}\right)\right) \\
&= \frac{N}{16}\left[\frac{1}{n_1}\operatorname{Var}\left(C_{111} - C_{112} - D_{211} + D_{212}\right)\right. \\
&\quad \left. + \frac{1}{n_2}\operatorname{Var}\left(D_{111} - D_{112} - C_{211} + C_{212}\right)\right] \\
&= \sum_{i=1}^{2} \frac{N}{16 n_i} \tau_{W,i}^{p} \quad \text{mit} \\
\tau_{W,1}^{p} &= \operatorname{Var}\left(C_{111} - C_{112} - D_{211} + D_{212}\right) \quad \text{und} \\
\tau_{W,2}^{p} &= \operatorname{Var}\left(D_{111} - D_{112} - C_{211} + C_{212}\right).
\end{aligned}
$$

Indem die Marginalverteilungen in den C_{ikj} und D_{ikj} durch die empirischen Verteilungen ersetzt werden, erhält man die Schätzer

$$
\begin{aligned}
\widehat{\tau}_{G,i}^{F} &= \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(\widehat{Y}_{ik1} + \widehat{Y}_{ik2} - \widehat{\overline{Y}}_{i\cdot 1} - \widehat{\overline{Y}}_{i\cdot 2}\right)^2, \\
\widehat{\tau}_{T,i}^{F} &= \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(\widehat{Y}_{ik1} - \widehat{Y}_{ik2} - \widehat{\overline{Y}}_{i\cdot 1} + \widehat{\overline{Y}}_{i\cdot 2}\right)^2,
\end{aligned}
$$

$$\begin{aligned}
\widehat{\tau}^p_{G,i} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\Big(\widehat{C}_{ik1}+\widehat{C}_{ik2}-\widehat{D}_{i'k1}-\widehat{D}_{i'k2} \\
&\quad -\widehat{\overline{C}}_{i\cdot 1}-\widehat{\overline{C}}_{i\cdot 2}+\widehat{\overline{D}}_{i'\cdot 1}+\widehat{\overline{D}}_{i'\cdot 2}\Big)^2, \\
\widehat{\tau}^p_{T,i} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\Big(\widehat{C}_{ik1}-\widehat{C}_{ik2}+\widehat{D}_{i'k1}-\widehat{D}_{i'k2} \\
&\quad -\widehat{\overline{C}}_{i\cdot 1}+\widehat{\overline{C}}_{i\cdot 2}-\widehat{\overline{D}}_{i'\cdot 1}+\widehat{\overline{D}}_{i'\cdot 2}\Big)^2
\end{aligned}$$

und

$$\begin{aligned}
\widehat{\tau}^p_{W,i} &= \frac{1}{n_i-1}\sum_{k=1}^{n_i}\Big(\widehat{C}_{ik1}-\widehat{C}_{ik2}-\widehat{D}_{i'k1}+\widehat{D}_{i'k2} \\
&\quad -\widehat{\overline{C}}_{i\cdot 1}+\widehat{\overline{C}}_{i\cdot 2}+\widehat{\overline{D}}_{i'\cdot 1}-\widehat{\overline{D}}_{i'\cdot 2}\Big)^2
\end{aligned}$$

mit $i, i' = 1, 2$ und $i \neq i'$. Damit erhält man die Freiheitsgrade

$$\begin{aligned}
\widehat{\nu}^F_G &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^F_{G,i}/n_i\right)^2}{\sum_{i=1}^2 \frac{\left(\widehat{\tau}^F_{G,i}/n_i\right)^2}{n_i-1}}, \\
\widehat{\nu}^F_T &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^F_{T,i}/n_i\right)^2}{\sum_{i=1}^2 \frac{\left(\widehat{\tau}^F_{T,i}/n_i\right)^2}{n_i-1}}, \\
\widehat{\nu}^p_G &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^p_{G,i}/n_i\right)^2}{\sum_{i=1}^2 \frac{\left(\widehat{\tau}^p_{G,i}/n_i\right)^2}{n_i-1}}, \\
\widehat{\nu}^p_T &= \frac{\left(\sum_{i=1}^2 \widehat{\tau}^p_{T,i}/n_i\right)^2}{\sum_{i=1}^2 \frac{\left(\widehat{\tau}^p_{T,i}/n_i\right)^2}{n_i-1}}
\end{aligned}$$

und

$$\widehat{\nu}^p_W = \frac{\left(\sum_{i=1}^2 \widehat{\tau}^p_{W,i}/n_i\right)^2}{\sum_{i=1}^2 \frac{\left(\widehat{\tau}^p_{W,i}/n_i\right)^2}{n_i-1}}.$$

A.2 Nichtparametrische Analyse durch Schichtung

In diesem Abschnitt werden auf zwei unterschiedlichen Wegen Statistiken hergeleitet. Der erste Weg beruht auf der Theorie des asymptotischen Äquivalenzsatzes und seinen

Folgerungen, wie sie bereits in den vorhergehenden beiden Kapiteln vorgestellt wurde. Der zweite Weg beruht auf der Theorie der U-Statistiken. Dieser wurde von Bajorski und Petkau (1999) vorgeschlagen.

Es wird zunächst auf den ersten Fall eingegangen. In diesem werden für jede Schicht die Teststatistiken und deren Verteilungen bestimmt. Anschließend werden diese dann kombiniert, um die Hypothesen testen zu können. Es werden zunächst die Schätzer für die relativen Effekte $p_{S|u}$ angegeben. Diese werden im Schichtungsmodell durch

$$\widehat{p}_{S|u} = \int \widehat{F}_{1|u} d\widehat{F}_{2|u}$$

geschätzt. Dabei werden die empirischen Verteilungsfunktionen $\widehat{F}_{i|u}$ kanonisch mit Hilfe der Zufallsvariablen der Gruppe i gebildet, die in die Schicht u fallen. Diese Zufallsvariablen werden mit $X_{ik|u}$, $i = 1, 2$, $k = 1, \ldots, m_{iu\cdot}$ bezeichnet. Die Eigenschaften der Schätzer $\widehat{p}_{S|u}$ ergeben sich analog zu den Eigenschaften der relativen Effekte im vorhergehenden Kapitel.

Satz A.2.1
Unter den Annahmen (4.23) für alle Schichten und (4.4) sind die Schätzer $\widehat{p}_{S|u}$ konsistent und erwartungstreu für $p_{S|u}$.

Die Schätzer können auch einfacher berechnet werden, indem man die Ränge innerhalb jeder Schicht bildet. Sei also $R_{ik|u}$ der Rang von $X_{ik|u}$ in der Schicht u und $\overline{R}_{i\cdot|u}$ deren Mittel in der Gruppe i, dann gilt:

$$\widehat{p}_{S|u} = \frac{1}{m_{1u\cdot}} \left(\overline{R}_{2\cdot|u} - \frac{m_{2u\cdot} + 1}{2} \right).$$

Um Teststatistiken herleiten zu können, formulieren wir zunächst das asymptotische Verhalten der geschätzten relativen Effekte $\widehat{p}_{S|u}$ innerhalb jeder Schicht unter der Hypothese in jeder einzelnen Schicht.

Satz A.2.2
Es gelte (4.4) und (4.23) für alle Schichten. Dann ist unter der Hypothese $H_0 : F_{1|u} = F_{2|u}$

$$T_{S,F|u} = \sqrt{m_{\cdot u\cdot}} \frac{\widehat{p}_{S|u}}{\sqrt{\widehat{\sigma}_{S,F|u}}}$$

für alle $u = 1, \ldots, r$ asymptotisch standardnormalverteilt, falls die Varianz

$$\sigma_{S,F|u} = \lim_{m_{\cdot u\cdot} \to \infty} Var \left(\sqrt{m_{\cdot u\cdot}} \int \frac{1}{2} \left(F_{1|u} + F_{2|u} \right) d \left(\widehat{F}_{2|u} - \widehat{F}_{1|u} \right) \right)$$

größer als null ist. Dabei ist

$$\widehat{\sigma}_{S,F|u} = \sum_{i=1}^{2} \frac{1}{m_{\cdot u\cdot} m_{iu\cdot}(m_{iu\cdot}-1)} \sum_{k=1}^{m_{iu\cdot}} \left(R_{ik|u} - \overline{R}_{i\cdot|u}\right)^2$$

ein konsistenter Schätzer für $\sigma_{S,F|u}$. *Des Weiteren ist unter der Hypothese* $H_0 : p_{S|u} = 1/2$

$$T_{S,p|u} = \sqrt{m_{\cdot u\cdot}} \frac{\widehat{p}_{S|u}}{\sqrt{\widehat{\sigma}_{S,p|u}}}$$

für alle $u = 1, \ldots, r$ *asymptotisch normalverteilt, falls die Varianz*

$$\sigma_{S,p|u} = \lim_{m_{\cdot u\cdot}\to\infty} Var\left(\frac{\sqrt{m_{\cdot u\cdot}}}{2}\left(\int F_{1|u} d\widehat{F}_{2|u} - \int F_{2|u} d\widehat{F}_{1|u}\right)\right)$$

größer als Null ist. Der Schätzer

$$\widehat{\sigma}_{S,p|u} = \sum_{i'\neq i} \frac{m_{\cdot u\cdot}}{{m_{i'u\cdot}}^2 m_{iu\cdot}(m_{iu\cdot}-1)} \sum_{k=1}^{m_{iu\cdot}} \left(R_{ik|u} - R_{ik(i)|u} - \overline{R}_{i\cdot|u} + \frac{m_{iu\cdot}+1}{2}\right)^2$$

ist dabei konsistent für $\sigma_{S,p|u}$ *und* $R_{ik(i)|u}$ *ist der Internrang von* $X_{ik|u}$ *in der Gruppe* i *und der Schicht* u.

Beweis: Zunächst gilt analog wie im Kapitel über die nichtparametrische Kovarianzanalyse die asymptotische Äquivalenz

$$\begin{aligned}
&\sqrt{m_{\cdot u\cdot}}\left(\int \widehat{H}_{n|u} d\left(\widehat{F}_{2|u} - F_{2|u}\right) - \int \widehat{H}_{n|u} d\left(\widehat{F}_{1|u} - F_{1|u}\right)\right) \\
&\quad \doteqdot \sqrt{m_{\cdot u\cdot}}\left(\int H_{n|u} d\left(\widehat{F}_{2|u} - F_{2|u}\right) - \int H_{n|u} d\left(\widehat{F}_{1|u} - F_{1|u}\right)\right)
\end{aligned}$$

mit

$$\begin{aligned}
H_{n|u} &= \frac{1}{m_{\cdot u\cdot}} \sum_{i=1}^{2} m_{iu\cdot} F_{i|u} \quad \text{und} \\
\widehat{H}_{n|u} &= \frac{1}{m_{\cdot u\cdot}} \sum_{i=1}^{2} m_{iu\cdot} \widehat{F}_{i|u}
\end{aligned}$$

und die asymptotische Äquivalenz

$$\sqrt{m_{\cdot u\cdot}}\left(\widehat{p}_{S|u} - p_u\right) \doteqdot \sqrt{m_{\cdot u\cdot}}\left(\int F_{1|u} d\widehat{F}_{2|u} - \int F_{2|u} d\widehat{F}_{1|u} + 1 - 2p_{S|u}\right).$$

Unter den Hypothesen über die bedingten Verteilungen beziehungsweise über $p_{S|u}$ ist dann $\sqrt{m_{\cdot u \cdot}}\widehat{p}_{S|u}$ asymptotisch äquivalent zu

$$\sqrt{m_{\cdot u \cdot}} \int H_{n|u} d\left(\widehat{F}_{2|u} - \widehat{F}_{1|u}\right)$$

beziehungsweise

$$\sqrt{m_{\cdot u \cdot}} \left(\int F_{1|u} d\widehat{F}_{2|u} - \int F_{2|u} d\widehat{F}_{1|u} \right).$$

Für diese Summen von unabhängigen und beschränkten Zufallsvariablen kann dann ein zentraler Grenzwertsatz angewendet werden. Damit ergibt sich die Behauptung des Satzes.

□

Im Folgenden betrachten wir zwei Möglichkeiten, wie man die r Statistiken $T_{S,F|u}$ beziehungsweise $T_{S,p|u}$ zusammenfassen kann. Zum einen kann man die gewichtete Summe der Statistiken betrachten

$$\begin{aligned} S_{S,F|n} &= \sum_{u=1}^{r} a_u T_{S,F|u} \quad \text{und} \\ S_{S,p|n} &= \sum_{u=1}^{r} a_u T_{S,p|u}. \end{aligned}$$

Da die Statistiken unter den Hypothesen asymptotisch unabhängig standardnormalverteilt sind, sind die Summen asymptotisch normalverteilt mit Erwartungswert null und Varianz $\sum_{u=1}^{r} a_u^2$. Zum anderen kann die gewichtete Summe der Quadrate

$$\begin{aligned} S_{S,F|\chi} &= \sum_{u=1}^{r} a_u {T_{S,F|u}}^2 \quad \text{und} \\ S_{S,p|\chi} &= \sum_{u=1}^{r} a_u {T_{S,p|u}}^2 \end{aligned}$$

betrachtet werden, womit sich gegensätzliche Unterschiede in den Gruppen nicht aufheben. Die Verteilung einer gewichteten Summe von χ^2-verteilten Zufallsvariablen kann nach einer Idee von Box (1954) durch eine mit g gestreckte χ^2_f-Verteilung approximiert werden. Dabei werden die Parameter g und f so bestimmt, dass die beiden ersten asymptotischen Momente der Verteilung der gewichteten Summe und der gestreckten χ^2_f-Verteilung übereinstimmen. So ergeben sich die Parameter

$$g = \frac{\sum_{u=1}^{r} a_u^2}{\sum_{u=1}^{r} a_u} \quad \text{und}$$

$$f = \frac{\left(\sum_{u=1}^{r} a_u\right)^2}{\sum_{u=1}^{r} a_u^2}.$$

Nachdem nun die Teststatistiken, die auf dem asymptotischen Äquivalenzsatz beruhen, hergeleitet wurden, wird auf die Verfahren von Bajorski und Petkau (1999) eingegangen. Sie betrachten Zufallsvariablen der Form

$$S_{S,F|U} = \sqrt{n} \sum_{u=1}^{r} a_u \widehat{p}_{S|u}$$

mit Hilfe der Theorie der U-Statistiken. Dafür muss n_1/n für $n \to \infty$ gegen $\lambda \in (0;1)$ streben. Die asymptotische Varianz dieser Teststatistik

$$\sigma_{S,F|U} = \frac{1}{12\lambda(1-\lambda)} \sum_{u=1}^{r} a_u^2 \frac{1 - \sum_{v=1}^{r} {q_{v|u}}^3}{q_u}$$

kann dann konsistent geschätzt werden, indem die unbekannten Parameter λ, $q_{v|u}$ und q_u durch ihre konsistenten Schätzer

$$\widehat{\lambda} = \frac{n_1}{n}, \quad \widehat{q}_{v|u} = \frac{m_{\cdot uv}}{m_{\cdot u \cdot}} \quad \text{und} \quad \widehat{q}_u = \frac{m_{\cdot u \cdot}}{n}$$

ersetzt werden. Man erhält dann

$$\widehat{\sigma}_{S,F|U} = \frac{n^2}{12 n_1 n_2} \sum_{u=1}^{r} a_u^2 \frac{1 - \sum_{v=1}^{r} {\widehat{q}_{v|u}}^3}{\widehat{q}_u}.$$

Damit kann im folgenden Satz die Verteilung der Teststatistik formuliert werden.

Satz A.2.3
Falls $n \to \infty$ mit $n_1/n \to \lambda \in (0;1)$, so ist

$$T_{S,F|U} = \sqrt{n} \frac{\sum_{u=1}^{r} a_u \left(\widehat{p}_{S|u} - 1/2\right)}{\sqrt{\widehat{\sigma}_{S,F|U}}}$$

unter $H_0: \; F_{1|u} = F_{2|u}$ für alle $u = 1, \ldots, r$ asymptotisch standardnormalverteilt.

Beweis: Siehe Bajorski und Petkau (1999). □

Nun bleibt nur noch die Aufgabe, die Wahl der Gewichte a_u möglichst günstig zu treffen. Dies betrifft alle Teststatistiken. Da die Gewichte die Macht der Verfahren erhöhen sollen, werden sie im Allgemeinen vom Stichprobenumfang abhängen. Daher sollen sie

gegen Konstanten konvergieren, wodurch die Verteilungsaussagen erhalten bleiben. Die einfachste Wahl der Gewichte ist, sie konstant und gleich zu lassen $a_u = 1$. In diesem Fall sind $T_{S,F|\chi}$ und $T_{S,p|\chi}$ asymptotisch χ^2-verteilt mit r Freiheitsgraden. Es erscheint aber vernünftig, dass Schichten mit einer größeren Anzahl von Beobachtungen mehr Gewicht bekommen sollten, da sie mehr Informationen enthalten. Daher könnte man die Anzahlen der Beobachtungen pro Schicht verwenden $a_u = m_{\cdot u \cdot}$. Des Weiteren könnte man berücksichtigen, dass die Schichten, in denen fast gleich viele Beobachtungen aus beiden Gruppen stammen, gegenüber den Schichten mit großen Unterschieden zwischen diesen Anzahlen bevorzugt werden. In nahezu balancierten Schichten werden mehr Vergleiche zwischen den beiden Versuchsgruppen durchgeführt als in stark unbalancierten Gruppen. Gewichte, die dies berücksichtigen, wären zum Beispiel

$$a_u = \frac{m_{1u\cdot} m_{2u\cdot}}{n(m_{1u\cdot} + m_{2u\cdot} + 1)}. \tag{A.3}$$

Diese Gewichte wurden von van Elteren (1960) vorgeschlagen. Mit Hilfe von Aussagen über benachbarte Alternativen gelangen Bajorski und Petkau (1999) zu den Gewichten

$$a_u = \frac{m_{1\cdot u} + m_{2\cdot u}}{n(1 - \sum_{v=1}^{r} \widehat{q}_{v|u}{}^3)}. \tag{A.4}$$

Bevor auf den Vergleich der Teststatistiken eingegangen wird, sollen einige Aspekte der Berechnung dieser beleuchtet werden. Wie schon angesprochen, müssen in jeder Schicht, die berücksichtigt werden soll, Beobachtungen aus beiden Gruppen sein. Es ergeben sich jedoch noch weitergehende Einschränkungen, die aus den Gewichten von Bajorski und Petkau (1999) und aus den Varianzen folgen. Im asymptotischen Fall mögen sie unproblematisch sein; im finiten Fall hingegen gehen durch sie unter Umständen Informationen verloren.

Bei der Wahl der Gewichte von Bajorski und Petkau (1999) muss beachtet werden, dass

$$(1 - \sum_{v=1}^{r} \widehat{q}_{v|u}{}^3) \neq 0$$

für jede Schicht u gilt. Dies wird durch die einfach zu überprüfende äquivalente Bedingung

$$\exists\, (i;k), (i';k') \text{ mit } (i;k) \neq (i';k') \text{ und } X_{ik|u} \neq X_{i'k'|u} \tag{A.5}$$

gesichert. Die Bedingung (A.5) bedeutet, dass nicht alle Beobachtungen in einer Schicht in die gleiche Kategorie fallen dürfen. In diesem Fall kann die Gewichtung dieser Schicht

nicht berechnet werden und die Schicht wird aus der Analyse genommen. Dies passiert jedoch nicht bei den Gewichten von van Elteren (1960). Allerdings trägt auch dort eine solche Schicht weder zur Veränderung des Zählers noch des Nenners der Statistik $T_{S,F|u}$ bei, denn aus $(1 - \sum_{v=1}^{r} \widehat{q}_{v|u}{}^3) = 0$ folgt $\widehat{p}_{S|u} = 0$.

Eine weitergehende Bedingung als (A.5) wird für die Schätzung der Varianz $\sigma_{S,F|u}$ benötigt:

$$\exists\, k; k' \text{ mit } k \neq k' \text{ und } X_{ik|u} \neq X_{ik'|u}. \tag{A.6}$$

Damit ist die empirische Varianz $\widehat{\sigma}_{S,F|u}$ positiv. Bei dieser Bedingung müssen im Gegensatz zur Bedingung (A.5) die ungleichen Beobachtungen aus der gleichen Gruppe stammen.

Damit die empirische Varianz $\widehat{\sigma}_{S,p|u}$ positiv ist, muss außer der Bedingung (A.6) auch

$$\min_k \{X_{1ku}\} \leq \max_k \{X_{2k'u}\} \quad \text{oder} \quad \min_k \{X_{2k'u}\} \leq \max_k \{X_{1ku}\} \tag{A.7}$$

gelten. Dies bedeutet, dass die beiden empirischen Verteilungsfunktionen in den Schichten überlappen. Asymptotische reicht es zwar aus, dass die genannten Bedingungen fast sicher eintreten. Im finiten Fall müssen sie jedoch einzeln überprüft werden. Bei vielen Kategorien und wenigen Beobachtungen kann dies im Extremfall dazu führen, dass die Verfahren gar nicht angewendet werden können, weil in keiner Schicht die Bedingungen erfüllt sind.

Anhang B

Beweise

B.1 Beweis zu Satz 4.5.1

Als Erstes gelte die Hypothese auf keinen nichtparametrischen Gruppeneffekt

$$\pi_{11s} - \pi_{21s} = -(\pi_{12s} - \pi_{22s}) \; \forall \; s. \tag{B.1}$$

Gilt nun zusätzlich die Hypothese über die nichtparametrischen Wechselwirkung

$$\pi_{11s} - \pi_{21s} = \pi_{12s} - \pi_{22s} \; \forall \; s, \tag{B.2}$$

so sind die linken Seiten der beiden Gleichungen identisch. Die beiden rechten Seiten können aber nur dann gleich sein, wenn sie null sind. Damit folgen die paarweisen Gleichheiten der Wahrscheinlichkeiten

$$\begin{aligned} \pi_{11s} &= \pi_{21s} \; \forall \; s \text{ und} \\ \pi_{12s} &= \pi_{22s} \; \forall \; s. \end{aligned}$$

Da die logit-Transformation bijektiv ist, folgt

$$\begin{aligned} L_{11s} &= L_{21s} \; \forall \; s \text{ und} \\ L_{12s} &= L_{22s} \; \forall \; s. \end{aligned}$$

Somit ist auch die Hypothese auf keine parametrische Wechselwirkung

$$L_{11s} - L_{21s} = L_{12s} - L_{22s} \; \forall \; s \tag{B.3}$$

erfüllt. Es folgt also aus den Gleichungen (B.1) und (B.2) die Gleichung (B.3). Verwendet man anstatt der Hypothese auf keinen nichtparametrischen Gruppeneffekt (B.1) die Hypothese auf keinen nichtparametrischen Zeiteffekt (B.4)

$$\pi_{11s} - \pi_{12s} \;=\; -(\pi_{21s} - \pi_{22s}) \;\forall\; s, \tag{B.4}$$

so kann man mit der umformulierten Hypothese auf keine nichtparametrische Wechselwirkung

$$\pi_{11s} - \pi_{12s} \;=\; \pi_{21s} - \pi_{22s} \;\forall\; s$$

genauso schließen.

Für die Rückrichtung kann man π_{ijs} und L_{ijs} vertauschen. Die Argumentation ist dann analog.

□

B.2 Beweis zu Satz 5.1.1

Zunächst gilt unter den Annahmen bezüglich der Stichprobenumfänge die asymptotische Äquivalenz

$$\begin{aligned}
&\sqrt{n}\left(\widetilde{p}^*(\gamma_{nkov,V}) - p^*(\gamma_{nkov,V})\right) \\
&\;\doteqdot\; \sqrt{n}\left(\int F_{12} d\widehat{F}_{22} - \int F_{22} d\widehat{F}_{12} + 1 - 2p_2\right. \\
&\qquad \left. -\gamma_{nkov,V}\left[\int F_{11} d\widehat{F}_{21} - \int F_{21} d\widehat{F}_{11} + 1 - 2p_1\right]\right),
\end{aligned}$$

da die Differenz

$$\sqrt{n}\left[\int \widehat{F}_{12} - F_{12}\left(d\widehat{F}_{22} - F_{22}\right) - \gamma_{nkov,V}\left(\int \widehat{F}_{11} - F_{11}\left(d\widehat{F}_{21} - F_{21}\right)\right)\right]$$

zwischen der rechten und der linken Seite dieser Gleichung bezüglich der $\mathcal{L}_2$-Norm gegen Null konvergiert (Satz 4.17 Siemer, 1999). Mit

$$\mathrm{E}\left(\sqrt{n}\left(\widetilde{p}^*(\gamma_{nkov,V}) - p^*(\gamma_{nkov,V})\right)\right) \;=\; 0$$

kann nun ein zentraler Grenzwertsatz wie zum Beispiel aus Loève (1977) von Seite 186 angewendet werden. Dazu kann man aufgrund der asymptotischen Äquivalenz die unabhängigen Zufallsvariablen Z_{ik}^* betrachten. Diese sind unabhängig und beschränkt.

$\square$

Anhang C

Niveau- und Gütesimulationen

C.1 Naive Verfahren

Die Aussagen der naiven Verfahren beruhen auf der Annahme (4.4) von Seite 16 und gelten damit nur asymptotisch. Um das Verhalten bei kleinen Stichproben zu bewerten, wurden Niveausimulationen für das nominelle Niveau $\alpha = 0{,}05$ durchgeführt. Die Ergebnisse sind in den Tabellen C.1 bis C.5 zusammengefasst. Es wurden jeweils 10000 mal n Zufallsvariablen gemäß der Wahrscheinlichkeitsverteilungen $\boldsymbol{q}_i$ erzeugt. Die Stichprobenumfänge n_i wurden für beide Gruppen gleich gewählt. Der Anteil der Testentscheidungen zu Gunsten der Alternative wurde angegeben. In der Spalte *Pearson* befinden sich die simulierten Niveaus für die Pearson-Statistik, in der Spalte *C.-A.* die der Cochran-Armitage-Statistik.

Bei der ersten Simulation wurden alle Zellwahrscheinlichkeiten gleich groß gewählt. Die Ergebnisse dieser Simulation finden sich in der Tabelle C.1. Sowohl die Pearson- als auch die Cochran-Armitage-Statistik halten das Niveau in diesem Fall schon bei einem Stichprobenumfang von 7 Versuchseinheiten pro Gruppe gut ein. Die Pearson-Statistik ist dabei etwas konservativer als die Cochran-Armitage-Statistik. Die Abweichungen vom nominellen Niveau von 5% betragen in keinem Fall mehr als 1%. Das bedeutet, dass bei

Tabelle C.1: Niveausimulation mit $\boldsymbol{q}_i = (1/3; 1/3; 1/3)^t$ zum Niveau 5%

n_i	Pearson	C.-A.	n_i	Pearson	C.-A.
7	4,60	4,86	20	5,23	5,39
10	4,83	5,66	30	5,00	4,95
15	4,86	5,40	50	4,83	4,77

Tabelle C.2: Niveausimulation mit $\boldsymbol{q}_i = (9/20; 1/10; 9/20)^t$ zum Niveau 5%

n_i	Pearson	C.-A.	n_i	Pearson	C.-A.
7	3,00	5,50	20	4,42	5,31
10	2,72	4,96	30	5,58	4,97
15	4,03	5,58	50	4,95	5,29

einem Experiment, bei dem die Wahrscheinlichkeiten für eine Verbesserung, eine Verschlechterung oder einen gleichbleibenden Zustand der Versuchseinheit gleich groß sind, eine sehr geringe Stichprobengröße ausreicht, um das Niveau einzuhalten. Damit ist die Annahme (4.4) von Seite 16 in diesem Fall keine starke Einschränkung.

Bei ordinalen Skalen mit vielen Kategorien ist es in der Regel sehr unwahrscheinlich zweimal die gleiche Kategorie an derselben Versuchseinheit zu beobachten. Daher wurde in Tabelle C.2 für eine symmetrische Verteilung die Wahrscheinlichkeit für die Gleichheit der Beobachtungen mit 1/10 sehr niedrig gewählt. Die Konvergenz der Cochran-Armitage-Statistik wird dadurch nicht beeinflusst. Hingegen braucht die Pearson-Statistik einen deutlich größeren Stichprobenumfang. Die asymptotische Verteilung wird erst ab einem Stichprobenumfang von ungefähr 20 Versuchseinheiten pro Gruppe erreicht. Das Niveau wird zwar immer eingehalten, bei kleinen Stichprobenumfängen jedoch nicht ausgeschöpft. Für die Praxis bedeutet dies, dass für die Pearson-Statistik mehr Versuchseinheiten benötigt werden, falls die ordinale Skala sehr fein ist, und die Wahrscheinlichkeit für einen gleichbleibenden Wert bei einer Versuchseinheit sehr gering ist. Dies hängt auch mit der Annahme (4.11) von Seite 21 zusammen. Offensichtlich ist die Pearson-Statistik im vorliegenden Fall nicht so robust gegen Beinaheverletzungen dieser Annahme wie die Cochran-Armitage-Statistik. Diese Probleme bei geringen Erwartungswerten in einigen Zellen der Kontingenztafel sind bekannt. In Hartung (1995) auf Seite 439 wird als Faustregel angegeben, dass in jede Zelle mindestens 5 Beobachtungen fallen müssen, beziehungsweise unter Hypothese erwartet werden müssen, damit die Approximation der χ^2-Verteilung adäquat ist. Cochran (1954) und Yarnold (1970) geben noch genauere Regeln für die Güte der Approximation an.

In den Tabellen C.3 und C.4 wurden schiefe Verteilungen simuliert. Aber auch wenn q_{i1} recht klein wird, ändert sich an den guten Konvergenzeigenschaften der Cochran-Armitage-Statistik nichts. Die Pearson-Statistik braucht auch hier einen größeren Stichprobenumfang, da sie auf die kleinen Zellwahrscheinlichkeiten empfindlich reagiert. Außerdem ist die Verteilung der Pearson-Statistik invariant unter Permutationen der Kompo-

Tabelle C.3: Niveausimulation mit $\boldsymbol{q}_i = (1/4; 1/10; 13/20)^t$ zum Niveau 5%

n_i	Pearson	C.-A.	n_i	Pearson	C.-A.
7	2,34	4,52	20	3,91	5,05
10	2,86	5,40	30	4,82	4,85
15	4,04	4,79	50	4,94	5,03

Tabelle C.4: Niveausimulation mit $\boldsymbol{q}_i = (1/10; 1/3; 13/15)^t$ zum Niveau 5%

n_i	Pearson	C.-A.	n_i	Pearson	C.-A.
7	3,01	5,39	20	4,43	5,54
10	2,98	4,96	30	4,49	4,96
15	4,01	5,04	50	4,95	4,94

nenten des Vektors $\boldsymbol{q}_i$, womit es keine Rolle spielt, wo die kleinen Zellwahrscheinlichkeiten auftreten.

Zuletzt ist in Tabelle C.5 das Ergebniss einer Niveausimulation der Cochran-Armitage-Statistik bei unterschiedlichen Varianzen und symmetrischen Verteilungen zusammengefasst. Auch hier wird die asymptotische Verteilung früh, das heißt ab einem Stichprobenumfang von 7 Versuchseinheiten pro Gruppe, erreicht. Die Pearson-Statistik wird hier nicht tabelliert, da im vorliegenden Fall die zur Pearson-Statistik gehörende Hypothese (4.13) nicht erfüllt ist. Diese Wahrscheinlichkeiten sind ein Beispiel dafür, wann die zur Cochran-Armitage-Statistik gehörende Hypothese (4.14) erfüllt ist, während die Hypothese (4.13) nicht gilt.

Insgesamt kann man sagen, dass die Pearson-Statistik konservativer ist als die Cochran-

Tabelle C.5: Niveausimulation mit $\boldsymbol{q}_1 = (1/3; 1/3; 1/3)^t$ und $\boldsymbol{q}_2 = (9/20; 1/10; 9/20)^t$ zum Niveau 5%

n_i	C.-A.	n_i	C.-A.
7	5,43	20	5,20
10	5,49	30	5,31
15	5,50	50	5,36

Armitage-Statistik. Auch reagiert die Pearson-Statistik empfindlicher auf geringe Zellwahrscheinlichkeiten. Allerdings wird sie dann konservativ und folglich die Nullhypothese mit geringerer Wahrscheinlichkeit als das vorgegebene Niveau ablehnen, wenn sie gilt. Bezüglich der Niveausimulationen ist abschließend zu bemerken, dass das Niveau α von beiden Statistiken nie deutlich überschritten wird.

Es ist noch zu bemerken, dass bei kleinen Stichprobenumfängen anstatt der asymptotischen Verfahren auch Permutationsverfahren eingesetzt werden können. Diese sollten noch bessere Ergebnisse liefern als die asymptotischen Verfahren. Da die Ergebnisse der Cochran-Armitage-Statistik aber schon sehr zufriedenstellend sind, wird diesbezüglich hier kein Vergleich durchgeführt.

Nachdem die Teststatistiken bezüglich des empirischen Niveaus bei kleinen Stichprobenumfängen untersucht wurden, soll nun ein Vergleich bezüglich der Macht vorgestellt werden. Dazu ist in der Abbildung C.1 die simulierte Macht der beiden naiven Statistiken bei 10000 Simulationsläufen pro Stützpunkt aufgetragen. Zwischen den Stützpunkten wurde interpoliert. Unter der Hypothese wurden die Wahrscheinlichkeiten $\boldsymbol{q}_i = (0{,}4; 0{,}2; 0{,}4)^t$ verwendet. Um die Alternative zu erzeugen, wurde die Wahrscheinlichkeit q_{21} im gleichen Maße verringert, wie die Wahrscheinlichkeit q_{23} erhöht wurde, das heißt, die Summe dieser beiden Wahrscheinlichkeiten wurde konstant gehalten. Aufgetragen in der Abbildung C.1 ist die relative Häufigkeit der Ablehnungen der Nullhypothese in Abhängigkeit von der Wahrscheinlichkeit q_{21}. Für $q_{21} = 0,4$ gelten die Hypothesen (4.13) und (4.14). Je kleiner die Wahrscheinlichkeit q_{21} wird, desto deutlicher ist der Unterschied zwischen den beiden Verteilungen. Für den Stichprobenumfang wurden 30 Versuchseinheiten pro Gruppe gewählt, da dann beide Statistiken das Niveau von $\alpha = 0{,}05$ sehr gut einhalten (Pearson: 0,048 und Cochran-Armitage: 0,052). Die Abbildung C.1 zeigt, dass die Macht des Cochran-Armitage-Tests in weiten Teilen über der Macht des Pearson-Tests liegt. Damit deckt der Cochran-Armitage-Test die simulierten

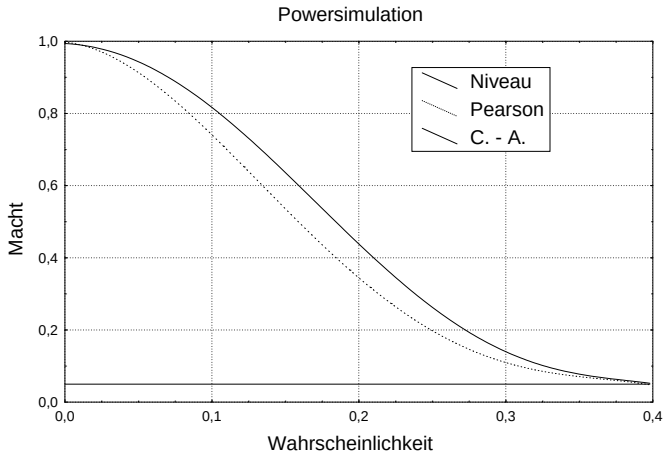

Abbildung C.1: Powersimulation für die naiven Verfahren mit $\alpha = 0{,}05$ und $n_i = 30$

Alternativen mit höherer Wahrscheinlichkeit auf.

Nimmt man an, dass die Wahrscheinlichkeit für zwei gleiche Werte an derselben Versuchseinheit sehr groß ist ($\boldsymbol{q}_i = (0{,}1; 0{,}8; 0{,}1)^t$), so stellt sich heraus, dass die Macht des Pearson-Tests für sehr kleine Wahrscheinlichkeiten q_{21} besser wird als die Macht des Cochran-Armitage-Tests. Zu beachten ist allerdings, dass die Wahrscheinlichkeiten dann mit weniger als 0,02 schon sehr gering sind. Dies zeigt die Abbildung C.2, bei der wieder die Wahrscheinlichkeit q_{21} auf der horizontalen Achse aufgetragen ist. Hier wurde ein recht großer Stichprobenumfang von 100 Versuchseinheiten pro Gruppe verwendet, um das Niveau einzuhalten.

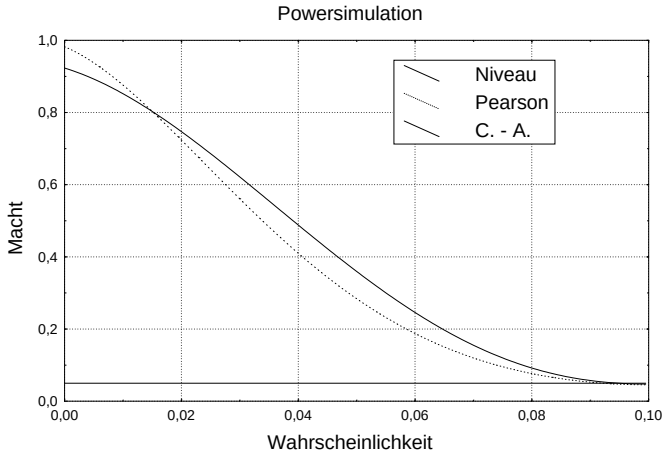

Abbildung C.2: Powersimulation für die naiven Verfahren mit $\alpha = 0{,}05$ und $n_i = 100$

Die meist höhere Macht des Cochran-Armitage-Tests kann durch die Eingrenzung der Alternative erklärt werden. Da im Gegensatz zum Pearson-Test nur die interessierenden Alternativen betrachtet werden, ist der Fehler zweiter Art bei diesen geringer.

C.2 Nichtparametrische faktorielle Analyse

Tabelle C.6: Niveausimulation der nichtparametrischen faktoriellen Analyse für den Gruppeneffekt

	$H_0: \boldsymbol{C}_G\boldsymbol{F} = 0$		$H_0: \boldsymbol{C}_G\boldsymbol{p} = 0$	
	Normalvert.	t-Verteilung	Normalvert.	t-Verteilung
7	7,34	4,94	8,40	5,61
10	7,10	5,37	7,74	5,87
15	5,91	5,00	6,38	5,38
20	5,66	4,99	5,99	5,15
30	5,80	5,28	6,05	5,53

Tabelle C.7: Niveausimulation der nichtparametrischen faktoriellen Analyse für den Zeiteffekt

	$H_0: \boldsymbol{C}_T\boldsymbol{F} = 0$		$H_0: \boldsymbol{C}_T\boldsymbol{p} = 0$	
	Normalvert.	t-Verteilung	Normalvert.	t-Verteilung
7	6,99	4,39	3,56	1,81
10	6,81	4,87	4,03	2,60
15	6,10	4,99	4,44	3,52
20	5,84	4,99	4,38	3,60
30	5,31	4,82	4,60	4,05

Um die Teststatistiken der nichtparametrischen faktoriellen Analyse bezüglich des Niveaus bei kleinen Stichprobenumfängen zu vergleichen, wurden Simulationen mit jeweils 10000 Schritten durchgeführt. Dazu wurden zunächst unabhängige und normalverteilte Zufallsvariablen erzeugt. Durch eine Linearkombination dieser Zufallsvariablen erhält man abhängige normalverteilte Zufallsvariablen. Diese werden dann in fünf Kategorien zusammengefasst. Dabei wurden die Wahrscheinlichkeiten für alle Kategorien gleich gewählt. Mithin sind alle Marginalverteilungen Gleichverteilungen und alle Hypothesen der nichtparametrischen faktoriellen Analyse sind erfüllt. Für die Stichprobenumfänge pro Gruppe wurden 7, 10, 15, 20 und 30 Versuchseinheiten gewählt. Für die Hypothesen auf keinen Einfluss der Gruppe ergaben sich die empirischen Niveaus in Tabelle C.6 bei einem nomi-

Tabelle C.8: Niveausimulation der nichtparametrischen faktoriellen Analyse für die Wechselwirkung

	$H_0: \boldsymbol{C}_W\boldsymbol{F} = 0$		$H_0: \boldsymbol{C}_W\boldsymbol{p} = 0$	
	Normalvert.	t-Verteilung	Normalvert.	t-Verteilung
7	6,92	4,03	3,62	1,95
10	6,87	5,04	4,14	2,73
15	6,50	5,46	4,82	3,87
20	5,69	4,93	4,41	3,73
30	5,29	4,74	4,43	4,01

nellen Niveau von 5%. In der nächsten Tabelle C.7 wurden die empirischen Niveaus für die Tests auf keinen Einfluss der Zeit erfasst. Schließlich sind in der Tabelle C.8 die Werte für die Wechselwirkung aufgetragen.

In allen Tabellen sind die t-Approximationen weniger antikonservativ als die Normalapproximationen. Dadurch halten die t-Approximationen das nominelle Niveau wesentlich besser ein. Dies gilt für beide Hypothesen $\boldsymbol{CF} = 0$ und $\boldsymbol{Cp} = 0$. Vergleicht man jeweils die normal- und die t-Approximationen, so wird das Niveau unter der Hypothese über die Verteilungen besser eingehalten als unter der Hypothese über die relativen Effekte. Während bei der Hypothese auf keinen Einfluss der Gruppe die Werte alle antikonservativ sind, sind dies bei den anderen Hypothesen nur die Werte der normal-Approximationen unter der Hypothese über die Verteilungsfunktionen. Unter der Hypothese über die relativen Effekte sind im Fall des Zeiteffektes und der Wechselwirkung die normal-Approximationen den t-Approximationen vorzuziehen. Die t-Approximationen sind dort gerade bei sehr geringen Stichprobengrößen sehr konservativ. Insgesamt findet man für jede Situation, sei es für die Hypothese über die Verteilungen oder über die relativen Effekte, sei es für den Einfluss der Gruppe, der Zeit oder der Wechselwirkung, eine adäquate Statistik ab einem Stichprobenumfang von 7 bis 10.

Da die Hypothesen auf keinen Einfluss der Wechselwirkung im Vordergrund dieses Abschnitts stehen, wurde für die entsprechenden Teststatistiken zusätzlich ein Vergleich der Power durchgeführt. Da die Unterschiede zwischen den Gütefunktionen der einzelnen Teststatistiken sehr gering sind, wurden sie nicht in einer Abbildung dargestellt, sondern in der Tabelle C.9 zusammengefasst.

Die Zufallszahlen wurden wie bei den Niveausimulationen erzeugt. Es wurde ein Stichprobenumfang von 50 Versuchseinheiten pro Gruppe gewählt, damit das Niveau vergleich-

bar gut eingehalten wird. Die Alternative wurde durch eine Größe δ parametrisiert. Für $\delta = 0$ gelten die Hypothesen. In diesem Fall sind in allen Gruppen zu allen Zeitpunkten alle Kategorien gleich wahrscheinlich. Um ein Alternativen zu erzeugen wurde die Marginalverteilung der zweiten Gruppe zum zweiten Zeitpunkt geändert, während alle anderen Marginalverteilungen gleich blieben. Je größer δ wird, desto wahrscheinlicher sind die kleinen Kategorien in der zweiten Gruppe zu dem zweiten Zeitpunkt. Diese Wahrscheinlichkeiten $q_{22s} = \mathrm{P}(X_{2k2} = s)$ für $s = 1, \ldots, 5$ wurden ebenfalls in die Tabelle C.9 aufgenommen.

Die Simulationsergebnisse zeigen, dass sich die geringen Unterschiede zwischen den Statistiken unter der Hypothese auch in der Alternative fortsetzen. Die verschiedenen Gütefunktionen der Statistiken unterscheiden sich kaum mehr als 2% voneinander. Damit decken die vier Teststatistiken eine vorhandene Wechselwirkung mit nahezu gleich großer Wahrscheinlichkeit auf.

Tabelle C.9: Powersimulation der nichtparametrischen faktoriellen Analyse für die Wechselwirkung

						$H_0: \boldsymbol{C}_W\boldsymbol{F} = 0$		$H_0: \boldsymbol{C}_W\boldsymbol{p} = 0$	
δ	q_{221}	q_{222}	q_{223}	q_{224}	q_{225}	n	t	n	t
0	20,00	20,00	20,00	20,00	20,00	5,13	4,88	4,72	4,41
1	21,65	20,77	19,95	19,18	18,45	9,64	9,12	8,76	8,39
2	23,39	21,49	19,81	18,32	16,99	25,93	24,96	24,28	23,49
3	25,23	22,13	19,57	17,43	15,63	48,33	47,32	46,29	45,31
4	27,16	22,70	19,25	16,53	14,35	71,13	70,37	69,53	68,71
5	29,19	23,17	18,85	15,63	13,17	87,88	87,26	86,69	86,00
6	31,30	23,55	18,37	14,72	12,06	95,79	95,62	95,35	95,08

C.3 Nichtparametrische Analyse durch Schichtung

Die Annahmen (A.5), (A.6) und (A.7) mögen im praktischen Einzelfall nicht auftreten, erschweren allerdings Simulationen für kleine Stichprobenumfänge. Bei diesen wurden bei einem solchen Fall als p-Wert der entsprechenden Statistik 1 angenommen, weil die Alternative nicht nachgewiesen werden kann. Im Folgenden sollen unter dieser Prämisse nun die fünf verschiedenen Teststatistiken

$$\begin{aligned}
T_{S,F|n} &= \frac{\sum_{u=1}^{r} a_\xi T_{S,F|u}}{\sqrt{\sum_{u=1}^{r} a_u^2}}, \\
T_{S,p|n} &= \frac{\sum_{u=1}^{r} a_u T_{S,p|u}}{\sqrt{\sum_{u=1}^{r} a_u^2}}, \\
T_{S,F|\chi} &= \frac{\sum_{u=1}^{r} a_u T_{S,F|u}^2}{\sum_{u=1}^{r} a_u^2}, \\
T_{S,p|\chi} &= \frac{\sum_{u=1}^{r} a_u T_{S,p|u}^2}{\sum_{u=1}^{r} a_u^2} \quad \text{und} \\
T_{S,F|U} &= \sqrt{n}\frac{\sum_{u=1}^{r} a_u \left(\widehat{p}_{S|u} - 1/2\right)}{\sqrt{\widehat{\sigma}_{S,F|U}}}
\end{aligned}$$

für die aus von van Elteren (1960) und von Bajorski und Petkau (1999) mittels Simulationen verglichen werden. Dafür wurden die Zufallszahlen wie in den vorhergehenden Kapiteln erzeugt. Ebenfalls wie bisher wurden jeweils 10000 Simulationschritte durchlaufen und das 5% Niveau für die Tests verwendet. Des Weiteren wurde ein balanciertes Design mit den Stichprobenumfängen $n_1 = n_2 = 7$, 10, 15, 20, 30, 50 und 100 verwendet.

Für $r = 3$, $r = 5$ und $r = 10$ Kategorien der ordinalen Skala sind in der Tabelle C.10 die Raten angegeben, wie häufig die Bedingungen (A.5), (A.6) und (A.7) jeweils erfüllt waren. Da die Bedingung (A.7) restriktiver ist als (A.6) und diese wiederum restriktiver ist als die Bedingung (A.5), nehmen die Raten in den drei Blöcken pro Zeile von links nach rechts ab. Besonders deutlich ist dabei der Unterschied zwischen der jeweils ersten Spalte und den jeweils anderen Spalten. Ab einem Stichprobenumfang von 15 Versuchseinheiten pro Gruppe in den ersten beiden Blöcken und ab 20 Versuchseinheiten pro Gruppe im dritten Block ist die Wahrscheinlichkeit, dass die Daten überhaupt nicht mit den Verfahren auszuwerten sind, zu vernachlässigen. Ab diesen Stichprobenumfängen hat die künstliche Ersetzung des p-Wertes durch eine Eins im Falle der Nichterfüllung der jeweiligen Bedingung auch keinen Einfluss mehr auf das Niveau.

Die Bedingung (A.5) scheint für alle Anzahlen von Kategorien nur bei Stichprobenumfängen von weniger als 10 Versuchseinheiten pro Gruppe ein Problem zu sein. Hingegen

Tabelle C.10: Raten für keinen Abbruch der Durchführung des Schichtungsverfahrens bei unterschiedlich vielen Kategorien auf Grund der Bedingungen (A.5), (A.6) und (A.7)

	3 Kategorien			5 Kategorien			10 Kategorien		
n_1	(A.5)	(A.6)	(A.7)	(A.5)	(A.6)	(A.7)	(A.5)	(A.6)	(A.7)
7	93,44	79,78	79,56	97,26	60,01	58,38	95,24	19,46	17,31
10	98,19	95,71	95,67	99,79	91,90	91,24	99,69	51,31	46,91
15	99,83	99,67	99,66	99,99	99,84	99,81	100,00	93,12	90,81
20	99,99	99,98	99,98	100,00	100,00	100,00	100,00	99,78	99,62
30	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
50	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
100	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

sind die beiden anderen Bedingungen bei zunehmender Anzahl der Kategorien immer weniger erfüllt. Bei 10 Kategorien und wenigen Versuchseinheiten sind sie nur noch in einem Fünftel bis der Hälfte der Simulationsschritte erfüllt. Dies macht deutlich, dass diese Verfahren bei vielen Kategorien völlig ungeeignet sind, denn die Daten sind dann häufig nicht einmal auswertbar.

Tabelle C.11: Niveausimulation für die Schichtungsverfahren bei drei Kategorien

	van Elteren					Bajorski und Petkau				
n_1	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$
7	5,14	4,33	5,81	4,95	6,67	4,45	4,25	4,89	4,90	6,47
10	6,83	6,15	7,80	8,16	6,84	5,20	5,09	6,16	6,46	6,38
15	7,36	6,68	8,27	8,44	6,91	5,46	4,59	6,31	5,91	6,40
20	7,00	6,55	7,74	8,44	6,34	5,18	4,36	5,69	5,75	5,77
30	6,14	5,88	6,56	7,08	5,63	4,88	4,58	5,05	5,14	5,14
50	5,79	6,21	5,99	6,82	5,32	5,31	5,73	5,43	6,03	5,08
100	5,38	5,65	5,46	5,92	5,23	5,43	5,43	5,48	5,52	5,18

Die Simulationsergebnisse für ein Niveau von 5% und bei $r = 3$ Kategorien sind in der Tabelle C.11 aufgelistet. Auf Grund der hohen Fehlerraten bei den Stichprobenumfängen 7 und 10 sind die Werte in den ersten beiden Zeilen nicht valide. Daher werden im Folgenden nur die Werte für einen Stichprobenumfang von mindestens 15 Versuchseinheiten pro

Gruppe diskutiert. Als Vergleichsmerkmal für die Geschwindigkeit der Konvergenz der asymptotischen Verfahren wird der minimale Stichprobenumfang verwendet, ab dem das simulierte, empirische Niveau nicht mehr als 1% von dem vorgegebenen Niveau von 5% abweicht, es also im Intervall $[4\%, 6\%]$ liegt.

Zunächst kann festgestellt werden, dass die Statistiken, bei denen die Gewichte von Bajorski und Petkau (1999) verwendet werden, den Statistiken mit den Gewichten von van Elteren (1960) bezüglich der Konvergenz unter Hypothese, das heißt beim Einhalten des Niveaus, vorzuziehen sind. Während bei den Ersteren ein Stichprobenumfang von 15 bis 20 Versuchseinheiten pro Gruppe ausreicht, brauchen die Letzteren einen Umfang von 30 bis 100.

Vergleicht man die jeweils nebeneinanderstehenden Approximationen der Normal- und der χ^2-Verteilung, so zeigt die Normalapproximation in der Regel bessere Werte als die Approximation mit Hilfe der χ^2-Verteilung. Die Statistik $T_{S,F|U}$ schneidet im Vergleich zu den anderen Statistiken im Fall der Gewichte von van Elteren (1960) besser und im Fall der Gewichte von Bajorski und Petkau (1999) schlechter ab. Insgesamt kann man jedoch sagen, dass die Unterschiede bezüglich des Einhaltens des Niveaus bei drei Kategorien sehr gering sind.

Tabelle C.12: Niveausimulation für die Schichtungsverfahren bei fünf Kategorien

	van Elteren					Bajorski und Petkau				
n_1	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$
7	6,24	6,17	7,07	6,75	10,21	6,05	6,17	6,82	6,75	10,33
10	9,59	9,74	10,58	10,94	9,81	8,61	9,35	9,48	10,63	10,07
15	11,47	14,61	12,50	17,49	9,00	9,39	12,21	10,49	14,46	9,56
20	10,46	15,04	12,04	18,54	7,71	8,62	12,27	9,92	15,24	8,09
30	8,37	12,92	10,12	16,73	6,72	6,92	9,64	7,86	13,53	6,96
50	6,38	8,92	7,41	11,43	5,61	5,94	7,45	6,51	9,30	5,56
100	5,47	7,03	5,79	8,01	5,24	5,66	6,72	5,74	7,33	5,47

Für fünf Kategorien sind die Simulationsergebnisse in der Tabelle C.12 zusammengefasst. Im Vergleich zur vorhergehenden Tabelle sind die Werte wesentlich schlechter. Hinzu kommt, dass die Werte für die Statistik $T_{S,F|U}$ ab 10 Versuchseinheiten und die Werte der anderen Statistiken erst ab 15 Versuchseinheiten valide sind. Wieder liefern die Gewichte von Bajorski und Petkau (1999) außer für die Statistik $T_{S,F|U}$ bessere empirische Niveaus als die Gewichte von van Elteren (1960). Gerade bei der Statistik $T_{S,F|U}$, die von

Bajorski und Petkau (1999) vorgeschlagen wurde, ist es umgekehrt. Für beide Gewichte hält die Statistik $T_{S,F|U}$ das Niveau am besten ein. Allerdings ist der Unterschied zur Statistik $T_{S,F|n}$ bei den Gewichten von Bajorski und Petkau (1999) unwesentlich und zur Statistik $T_{S,p|n}$ gering. Die Approximationen mit der χ^2-Verteilung sind wie im Fall mit drei Kategorien schlechter als die Normalapproximationen.

Tabelle C.13: Niveausimulation für die Schichtungsverfahren bei zehn Kategorien

	van Elteren					Bajorski und Petkau				
n_1	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$
7	3,00	3,10	2,71	2,72	13,55	2,96	3,09	2,71	2,72	13,91
10	8,46	8,92	7,16	7,07	12,85	8,20	8,82	7,09	7,08	13,02
15	15,92	19,74	12,69	14,11	10,95	14,74	19,00	11,69	13,84	11,47
20	16,25	27,65	15,71	23,50	9,70	14,80	26,13	14,21	22,35	10,75
30	15,09	32,07	17,25	32,53	8,66	13,56	29,83	15,10	30,66	10,01
50	10,03	27,36	13,79	33,74	6,82	9,66	25,08	12,69	32,10	7,68
100	6,75	13,26	8,18	17,89	5,90	6,91	12,32	7,97	16,48	6,01

Zuletzt wurde das Niveau auch für 10 Kategorien simuliert. In diesem Fall konvergieren alle Statistiken erst sehr spät. Selbst die beste Statistik $T_{S,F|U}$ braucht mindestens 100 Versuchseinheiten pro Gruppe um das Niveau einigermaßen einzuhalten. Vor allem die χ^2-Approximationen $T_{S,F|\chi}$ und $T_{S,p|\chi}$ brechen völlig ein. Sogar bei 100 Versuchseinheiten pro Gruppe liegen sie noch deutlich über 10% und nicht bei 5%. Hier erübrigt sich auf Grund der schlechten Konvergenz ein Vergleich.

Insgesamt kann man die Verfahren entweder nur bei sehr wenigen Kategorien oder bei einem extrem großen Stichprobenumfang verwenden. Ansonsten sind die Statistiken zum Teil erheblich antikonservativ. Überhaupt nähern sich die empirischen Niveaus dem nominellen Niveau fast immer von der antikonservativen Seite.

Für drei Kategorien wurden zusätzlich zu den Niveau- auch Powersimulationen durchgeführt. Für fünf oder zehn Kategorien erübrigt sich dieses auf Grund der schlechten Ergebnisse der Niveausimulationen. Die Ergebnisse der Powersimulation sind in der Tabelle C.14 und den Abbildungen C.3 und C.4 dargestellt. Für beide Gewichte wurde somit jeweils eine Abbildung erstellt. In diesen Abbildungen sind mit F die Statistiken gekennzeichnet, welche die Hypothese über die Verteilungsfunktionen testen. Dies sind $T_{S,F|n}$, $T_{S,F|\chi}$ und $T_{S,F|U}$. Da die anderen beiden Statistiken $T_{S,p|n}$ und $T_{S,p|\chi}$ die Hypothese über die relativen Effekte testen, wurden diese mit p gekennzeichnet. Des Weiteren gilt die

Tabelle C.14: Powersimulation für die Schichtungsverfahren

	van Elteren					Bajorski und Petkau				
δ	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$	$T_{S,F\|n}$	$T_{S,F\|\chi}$	$T_{S,p\|n}$	$T_{S,p\|\chi}$	$T_{S,F\|U}$
0	5,51	6,16	5,73	6,97	5,22	5,18	5,40	5,23	5,75	4,90
1	18,91	13,49	19,52	14,73	18,83	14,94	10,65	15,36	11,40	16,46
2	52,57	37,76	53,51	39,60	53,18	38,47	26,81	39,25	28,56	45,55
3	83,40	70,84	83,97	72,41	84,16	62,38	52,22	63,42	54,35	75,51
4	96,77	91,64	96,97	92,08	96,97	76,82	72,80	77,85	74,93	92,26
5	99,71	98,54	99,71	98,61	99,77	84,50	85,87	85,76	87,65	97,95
6	99,97	99,82	99,95	99,81	99,97	90,41	93,59	91,63	94,59	99,53

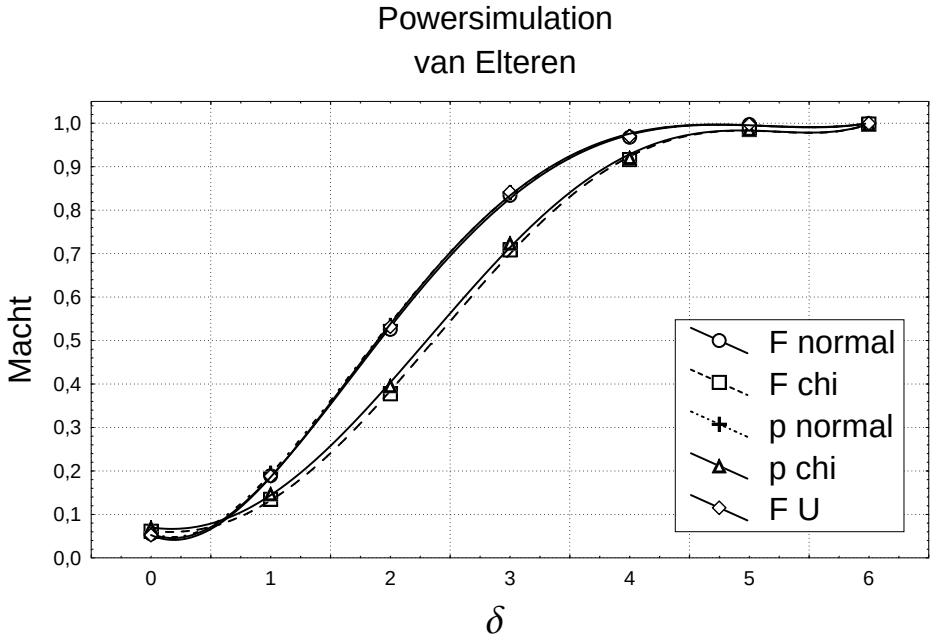

Abbildung C.3: Powersimulation für die Schichtungsverfahren mit den Gewichten von van Elteren

Kennzeichnung *normal* für die Statistiken $T_{S,F|n}$ und $T_{S,p|n}$ mit der Normalapproximation. *chi* wurde hingegen bei $T_{S,F|\chi}$ und $T_{S,p|\chi}$ wegen der χ^2-Approximation verwendet. Schließlich wird in der Abbildung mit $F\ U$ die Statistik $T_{S,F|U}$ bezeichnet.

Die Simulation zeigt für die Gewichte von van Elteren (1960) und von Bajorski und Petkau (1999), dass die χ^2 Approximationen trotz der schlechteren Einhaltung des Niveaus

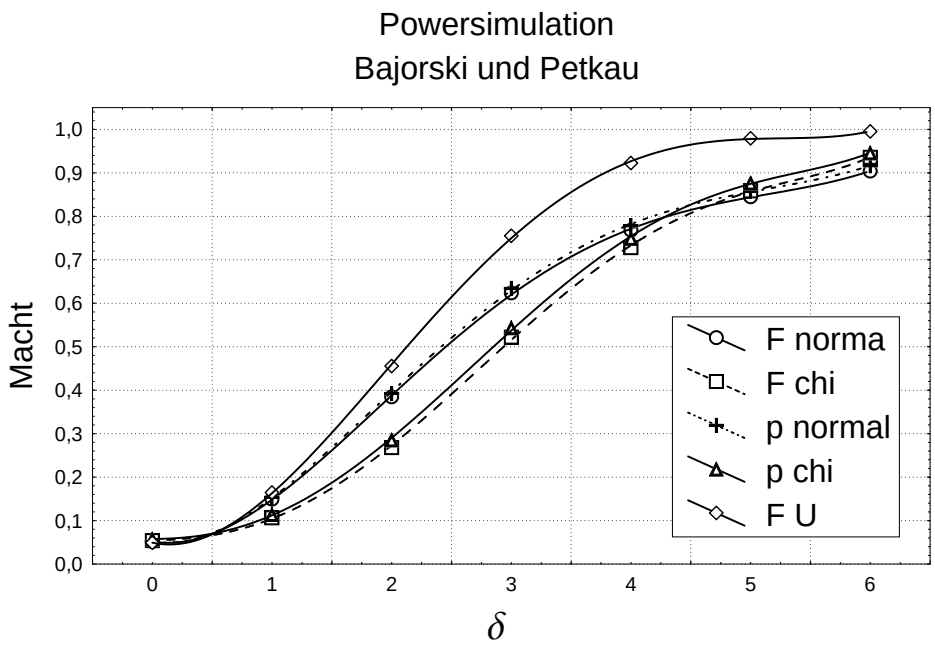

Abbildung C.4: Powersimulation für die Schichtungsverfahren mit den Gewichten von Bajorski und Petkau

eine schlechtere Power haben. Die anderen drei Statistiken unterscheiden sich bei den Gewichten von van Elteren (1960) bezüglich der Güte hingegen kaum. Hingegen ist bei den Gewichten von Bajorski und Petkau (1999) die Statistik $T_{S,F|U}$ besser als die Statistiken $T_{S,F|n}$ und $T_{S,p|n}$. Vergleicht man die Macht der Statistiken mit den unterschiedlichen Gewichten, so stellt man fest, dass die Güte bei den Gewichten von van Elteren (1960) deutlich besser ist. Der Unterschied beträgt teilweise mehr als 20%.

Abschließend kann man bezüglich der Vergleiche der Gewichte feststellen, dass bei kleinen Stichprobenumfängen zunächst die Gewichte von Bajorski und Petkau (1999) vorzuziehen sind. So ist sicherzustellen, dass das nominelle Niveau eingehalten wird. Bei größeren Stichproben jedoch, bei denen auch die Statistiken mit den Gewichten von van Elteren (1960) verwendet werden können, sollte man auf diese zurückgreifen, da sie eine bessere Macht haben. Werden die Teststatistiken untereinander verglichen, so ist fast immer die Statistik $T_{S,F|U}$ den anderen vorzuziehen. Zwar ist sie in der Tabelle C.11 den anderen Statistiken bei den Gewichten von Bajorski und Petkau (1999) leicht unterlegen, gleicht dies jedoch in allen anderen Fällen deutlich aus. Allerdings gilt dies alles nur mit der Einschränkung, dass nur wenige Kategorien verwendet werden. Ansonsten ist zu prüfen, ob der Stichprobenumfang wirklich so groß ist, dass das Niveau eingehalten wird.

C.4 Parametrische faktorielle Analyse

Tabelle C.15: Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 3 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	0,87	0,09	0,00	0,00	0,00	0,00
	$\widehat{\alpha}$	5,72	5,70	5,36	5,49	5,02	5,38
$\boldsymbol{R}_{ar}$	f	3,17	0,38	0,01	0,00	0,00	0,00
	$\widehat{\alpha}$	5,18	5,89	5,37	5,40	4,94	5,54
$\boldsymbol{R}_{cs}$	f	5,97	1,87	0,31	0,08	0,01	0,00
	$\widehat{\alpha}$	4,73	5,85	5,69	5,76	5,16	5,55
$\boldsymbol{R}_{mde}$	f	73,81	70,16	62,50	55,21	46,47	30,48
	$\widehat{\alpha}$	8,36	7,17	4,24	3,08	2,20	2,30
$\boldsymbol{R}_{un}$	f	61,01	42,44	24,39	9,29	1,02	0,00
	$\widehat{\alpha}$	6,28	7,14	6,94	6,91	5,87	5,74

In diesem Abschnitt soll überprüft werden, ab wann die Asymptotik für die parametrische faktorielle Analyse greift. Dazu wurden die Verfahren für die Stichprobenumfänge $n_i = 7, 10, 15, 20, 30$ und 50 simuliert. Für jede Simulation wurden wie bei den vorhergehenden Verfahren 10000 Datensätze nach dem gleichen Prinzip erstellt. Tabelliert sind die Fehlerraten f der verschiedenen Verfahren, das heißt, wie häufig die Iteration nicht konvergierte, und das empirische Niveau $\widehat{\alpha}$, welches sich aus den Simulationsschritten berechnet, bei denen die Iteration konvergierte. Als nominelles Niveau wurde $\alpha = 5\%$ gewählt. Die Angaben in den Tabellen sind ebenfalls in Prozent. Getestet wurden die Hypothesen auf keinen Gruppeneffekt $H_0 : \beta_G = 0$, auf keinen Zeiteffekt $H_0 : \beta_T = 0$ und auf keine Wechselwirkung $H_0 : \beta_W = 0$. Die Ergebnisse der dazugehörigen Simulationen sind in unterschiedlichen Tabellen zusammengefasst. Des Weiteren unterscheiden sich die Tabellen hinsichtlich der Anzahl der Kategorien auf der ordinalen Skala, die bei den Simulationen verwendet wurde. Exemplarisch wurden wie bei den Schichtungsverfahren $r = 3, 5$ und 10 Kategorien verwendet.

Zunächst wird das Iterationsverhalten bei unterschiedlichen Wahlen der Matrix $\boldsymbol{R}$ diskutiert. Diese hängt nicht von der Wahl der Hypothese ab, da alle Parameter simultan iterativ berechnet werden. Falls also die Iteration nicht konvergiert, kann keine der drei Hypothesen getestet werden. Hingegen hängt die Fehlerrate der Iteration sehr von $\boldsymbol{R}$ ab.

Tabelle C.16: Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 3 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	0,87	0,09	0,00	0,00	0,00	0,00
	$\widehat{\alpha}$	4,83	5,68	6,23	5,76	5,18	5,70
$\boldsymbol{R}_{ar}$	f	3,17	0,38	0,01	0,00	0,00	0,00
	$\widehat{\alpha}$	3,79	4,98	5,80	5,43	4,89	5,55
$\boldsymbol{R}_{cs}$	f	5,97	1,87	0,31	0,08	0,01	0,00
	$\widehat{\alpha}$	4,52	5,30	6,13	5,66	5,10	5,67
$\boldsymbol{R}_{mde}$	f	73,81	70,16	62,50	55,21	46,47	30,48
	$\widehat{\alpha}$	2,18	2,85	4,00	4,60	5,79	7,93
$\boldsymbol{R}_{un}$	f	61,01	42,44	24,39	9,29	1,02	0,00
	$\widehat{\alpha}$	4,18	5,00	4,88	4,32	4,70	5,55

Denn je komplexer die Matrix $\boldsymbol{R}$ ist, desto mehr Parameter müssen geschätzt werden. Nun soll im Einzelnen beschrieben werden, welcher Stichprobenumfang bei gegebener Anzahl der Kategorien notwendig ist, damit die Fehlerrate nur wenige Prozent beträgt.

Tabelle C.17: Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 3 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	0,87	0,09	0,00	0,00	0,00	0,00
	$\widehat{\alpha}$	4,86	5,82	6,29	5,60	5,22	5,34
$\boldsymbol{R}_{ar}$	f	3,17	0,38	0,01	0,00	0,00	0,00
	$\widehat{\alpha}$	9,05	7,24	6,28	5,79	5,53	5,35
$\boldsymbol{R}_{cs}$	f	5,97	1,87	0,31	0,08	0,01	0,00
	$\widehat{\alpha}$	4,53	5,65	6,13	5,52	5,20	5,34
$\boldsymbol{R}_{mde}$	f	73,81	70,16	62,50	55,21	46,47	30,48
	$\widehat{\alpha}$	4,43	2,95	2,96	1,88	2,35	2,57
$\boldsymbol{R}_{un}$	f	61,01	42,44	24,39	9,29	1,02	0,00
	$\widehat{\alpha}$	4,85	4,92	4,55	4,22	4,92	5,14

Für die Bezeichnungen der verschiedenen Working Correlation Matrizen wird ein Index

benutzt. Bei der Wahl der Einheitsmatrix wird *ind* für $\boldsymbol{R}$ verwendet. Für eine autoregressive Korrelation wird *ar* und bei Austauschbarkeit *cs* an $\boldsymbol{R}$ angehängt. Des Weiteren bezeichnet $\boldsymbol{R}_{mde}$ die Verwendung einer Korrelationsmatrix, bei der die Einträge, die weiter als zwei Einträge von der Hauptdiagonalen entfernt liegen, mit null geschätzt werden. Dies entspricht dem dritten vorgestellten Verfahren. Zuletzt gebrauchen wir $\boldsymbol{R}_{un}$ bei der Wahl einer unstrukturierten Matrix $\boldsymbol{R}$. Diese Kürzel werden fast alle auch von der Prozedur GENMOD in SAS verwendet, mit der die Simulationen durchgeführt wurden.

Tabelle C.18: Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 5 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	5,05	0,92	0,07	0,00	0,00	0,00
	$\widehat{\alpha}$	6,79	6,05	5,13	5,35	5,67	5,36
$\boldsymbol{R}_{ar}$	f	6,03	1,11	0,09	0,00	0,00	0,00
	$\widehat{\alpha}$	7,06	6,45	5,80	5,46	5,66	5,52
$\boldsymbol{R}_{cs}$	f	22,67	11,54	4,57	2,07	0,57	0,04
	$\widehat{\alpha}$	6,34	6,91	7,24	6,87	7,23	6,29
$\boldsymbol{R}_{mde}$	f	27,73	17,09	11,36	6,91	3,54	0,86
	$\widehat{\alpha}$	8,84	8,25	6,99	6,43	6,45	5,99
$\boldsymbol{R}_{un}$	f	90,98	88,71	83,39	76,23	51,50	6,41
	$\widehat{\alpha}$	5,43	7,26	5,78	6,18	7,81	7,69

Bei drei Kategorien ergeben sich für die drei einfachsten Working Correlation Matrizen $\boldsymbol{R}_{ind}$, $\boldsymbol{R}_{ar}$ und $\boldsymbol{R}_{cs}$ bei einem Stichprobenumfang von $n_i = 7$ Fehlerraten von eins bis sechs Prozent. Ab zehn Versuchseinheiten pro Gruppe können die Fehlerraten dann vernachlässigt werden. Hingegen konnten bei $\boldsymbol{R}_{mde}$ und $\boldsymbol{R}_{un}$ für $n_i = 7$ die Parameter in weniger als der Hälfte der Fälle berechnet werden. Die Wahl dieser Matrizen benötigt wesentlich mehr Stichprobenumfang. Im Fall von $\boldsymbol{R}_{mde}$ reichen 50 Versuchseinheiten gerade aus, um nur in jedem dritten Simulationsschritt einen Abbruch zu erzeugen. Bei $\boldsymbol{R}_{un}$ ist $n_i = 20$ notwendig, um die Fehlerrate unter 10% zu drücken. Erst ab ungefähr 30 Versuchseinheiten pro Gruppe ist sie dann zu vernachlässigen.

Wird die Anzahl der Kategorien auf fünf erhöht, so ändert sich bei den einfachen Matrizen $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{ar}$ nicht viel. Sie zeigen gute Fehlerraten ab $n_i = 10$. Bei $\boldsymbol{R}_{cs}$ werden hingegen schon 15 Versuchseinheiten pro Gruppe gebraucht, um nur noch eine Fehlerrate von ungefähr 5% zu erhalten. Bei $\boldsymbol{R}_{mde}$ wird die Fehlerrate seltsamerweise besser, wenn die

Anzahl der Kategorien sich von drei auf fünf erhöht. Nun sind die Fehlerraten akzeptabel ab $n_i = 30$ bis 50. Für $\boldsymbol{R}_{un}$ wächst der Anteil der Abbrüche hingegen stark an. Noch bei $n_i = 50$ ist dieser etwas größer als 6%.

Tabelle C.19: Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 5 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	5,05	0,92	0,07	0,00	0,00	0,00
	$\widehat{\alpha}$	7,74	6,25	5,84	5,92	5,09	4,79
$\boldsymbol{R}_{ar}$	f	6,03	1,11	0,09	0,00	0,00	0,00
	$\widehat{\alpha}$	5,32	5,12	5,29	5,20	4,91	4,46
$\boldsymbol{R}_{cs}$	f	22,67	11,54	4,57	2,07	0,57	0,04
	$\widehat{\alpha}$	6,81	5,80	5,40	5,67	5,03	4,76
$\boldsymbol{R}_{mde}$	f	27,73	17,09	11,36	6,91	3,54	0,86
	$\widehat{\alpha}$	5,15	4,67	4,30	4,61	4,70	4,28
$\boldsymbol{R}_{un}$	f	90,98	88,71	83,39	76,23	51,50	6,41
	$\widehat{\alpha}$	4,43	5,23	5,96	8,33	6,76	6,51

Eine weitere Verfeinerung der ordinalen Skala auf zehn Kategorien lässt alle Fehlerraten deutlich steigen. Die Eigenschaften von $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{ar}$ sind dabei sehr ähnlich. Dies ist dadurch zu erklären, dass bei $\boldsymbol{R}_{ar}$ nur der Parameter δ geschätzt werden muss. Dass bei der Wahl einer dieser beiden Matrizen die Verfahren bei ansteigendem r schlechter werden, hängt hauptsächlich mit der größeren Anzahl an Cutpoints zusammen. Bei den anderen drei Working Correlation Matrizen kommt hingegen noch die Schätzung der Parameter für die Matrizen hinzu. Bei zehn Kategorien hat dies die Auswirkung, dass für $n_i = 50$ die Fehlerrate bei $\boldsymbol{R}_{cs}$ circa 5%, bei $\boldsymbol{R}_{mde}$ 50% und bei $\boldsymbol{R}_{un}$ sogar mehr als 80% beträgt. Es ist also bei vielen Kategorien ein sehr großer Stichprobenumfang nötig, damit die Verfahren mit komplexen Working Correlation Matrizen überhaupt angewendet werden können.

Nach den Fehlerraten sollen nun die empirischen Niveaus betrachtet werden. Dabei werden nur diejenigen empirischen Niveaus verwendet, bei denen die zugehörigen Fehlerraten unter 10% liegen. Ansonsten erübrigt sich jede Diskussion des Niveaus. Als Qualitätsmaß für die Konvergenz der Teststatistiken gegen die χ^2_1-Verteilung soll der Stichprobenumfang dienen, ab dem die empirischen Niveaus $\widehat{\alpha}$ im Intervall $(4\%, 6\%)$ liegen. Damit werden Abweichungen von 1% toleriert.

Tabelle C.20: Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 5 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	5,05	0,92	0,07	0,00	0,00	0,00
	$\widehat{\alpha}$	6,97	6,69	5,94	5,70	5,52	5,50
$\boldsymbol{R}_{ar}$	f	6,03	1,11	0,09	0,00	0,00	0,00
	$\widehat{\alpha}$	9,93	8,31	7,05	6,04	5,89	5,55
$\boldsymbol{R}_{cs}$	f	22,67	11,54	4,57	2,07	0,57	0,04
	$\widehat{\alpha}$	6,32	6,06	5,53	5,55	5,46	5,51
$\boldsymbol{R}_{mde}$	f	27,73	17,09	11,36	6,91	3,54	0,86
	$\widehat{\alpha}$	9,30	8,29	6,55	5,83	5,71	5,55
$\boldsymbol{R}_{un}$	f	90,98	88,71	83,39	76,23	51,50	6,41
	$\widehat{\alpha}$	3,10	2,57	1,81	2,78	4,68	5,33

Bei fünf Kategorien ändert sich das Verhalten der Tests bezüglich des eingehaltenen Niveaus etwas. Bei den Tests der Hypothese $H_0 : \beta_G = 0$ halten die Verfahren mit $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{ar}$ das Niveau ab 15 Versuchseinheiten ein. Bei $\boldsymbol{R}_{cs}$ und $\boldsymbol{R}_{un}$ wird das Niveau hingegen bei den betrachteten Versuchsumfängen immer um ein bis drei Prozent überschritten. Mit 5,99% liegt das empirische Niveau für $\boldsymbol{R}_{mde}$ bei $n_i = 50$ gerade im geforderten Intervall. Daher kommen für die Hypothese bezüglich des Gruppeneffektes nur $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{ar}$ bei kleinen Stichprobenumfängen in Frage. Für den Zeiteffekt sind die Ergebnisse hingegen besser. Dort hält das Verfahren mit $\boldsymbol{R}_{ar}$ das Niveau ab sieben Versuchseinheiten pro Gruppe ein. Bei $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{cs}$ sind es mit $n_i = 15$ etwas mehr und für $\boldsymbol{R}_{mde}$ werden 20 Versuchseinheiten pro Gruppe benötigt. Wird $\boldsymbol{R}_{un}$ als Working Correlation Matrix gewählt, so ist erst ab ungefähr $n_i = 50$ die Fehlerrate unter 10%. Aber selbst dann ist das empirische Niveau noch größer als 6%. Beim Wechsel zur nächsten Tabelle, in der die empirischen Niveaus für die Tests der Hypothese $H_0 : \beta_T = 0$ stehen, ändert sich für $\boldsymbol{R}_{ind}$, $\boldsymbol{R}_{cs}$ und $\boldsymbol{R}_{mde}$ nichts. Bei $\boldsymbol{R}_{ar}$ tritt hingegen eine deutliche Verschlechterung ein. Dort wird das Niveau erst ab $n_i = 30$ eingehalten. Vorher ist es zum Teil deutlich antikonservativ. Für $\boldsymbol{R}_{un}$ reichen nun 50 Beobachtungen pro Faktorstufenkombination aus, um das Niveau einzuhalten und eine nicht zu hohe Fehlerrate zu haben.

Wir beginnen mit der Diskussion bei der Niveausimulation für die Tests der Hypothese auf keinen Gruppeneffekt bei drei Kategorien. Dort liegen die empirischen Niveaus $\widehat{\alpha}$ für $\boldsymbol{R}_{ind}$, $\boldsymbol{R}_{ar}$ und $\boldsymbol{R}_{cs}$ ab $n_i = 7$ im geforderten Intervall. Bei $\boldsymbol{R}_{mde}$ sind die Fehlerraten zu hoch, als dass eine Betrachtung des empirischen Niveaus sinnvoll wäre, und für $\boldsymbol{R}_{un}$

Tabelle C.21: Niveausimulation für den Gruppeneffekt der faktoriellen parametrischen Analyse bei 10 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	29,73	14,31	3,68	1,14	0,06	0,00
	$\widehat{\alpha}$	6,39	5,93	5,37	5,27	5,31	4,84
$\boldsymbol{R}_{ar}$	f	31,84	14,62	3,69	1,14	0,06	0,00
	$\widehat{\alpha}$	9,33	7,73	6,63	6,34	6,07	5,27
$\boldsymbol{R}_{cs}$	f	66,01	49,79	33,29	23,74	13,45	4,92
	$\widehat{\alpha}$	4,18	3,58	4,23	5,38	6,18	7,13
$\boldsymbol{R}_{mde}$	f	91,52	87,83	80,47	74,67	62,92	48,08
	$\widehat{\alpha}$	22,52	13,31	6,81	6,00	4,31	3,56
$\boldsymbol{R}_{un}$	f	97,28	92,07	92,98	91,24	89,21	86,63
	$\widehat{\alpha}$	29,04	2,27	5,98	6,51	9,18	7,40

sind ungefähr 30 Versuchseinheiten pro Gruppe notwendig, um das Niveau einzuhalten. Für die Tests der Hypothese auf keinen Zeiteffekt ist bei $\boldsymbol{R}_{ind}$, $\boldsymbol{R}_{ar}$ und $\boldsymbol{R}_{cs}$ ein größerer Stichprobenumfang für die Einhaltung des Niveaus erforderlich. Während $\boldsymbol{R}_{ar}$ mit $n_i = 10$ auskommt, brauchen $\boldsymbol{R}_{ind}$ und $\boldsymbol{R}_{cs}$ doppelt so viele Versuchseinheiten. Bei $\boldsymbol{R}_{un}$ sind ebenfalls $n_i = 20$ Versuchseinheiten pro Gruppe notwendig. Ebenfalls ab dieser Anzahl von Versuchseinheiten pro Gruppe erfüllen alle, bis auf $\boldsymbol{R}_{mde}$, bei der Hypothese auf keine Wechselwirkung das Qualitätskriterium, dass $\widehat{\alpha}$ nicht mehr als 1% vom nominellen Niveau abweicht. Dabei können die Tests bei weniger Versuchseinheiten sowohl konservativ als auch antikonservativ sein.

Zuletzt wird nun das empirische Niveau für die verschiedenen Wahlen von $\boldsymbol{R}$ bei zehn Kategorien betrachtet. Hier scheiden $\boldsymbol{R}_{mde}$ und $\boldsymbol{R}_{un}$ aus, weil die Fehlerraten inakzeptabel sind. Sobald diese bei $\boldsymbol{R}_{ind}$ unter zehn Prozent liegen, dies ist ab $n_i = 15$ der Fall, wird das Niveau für alle drei Hypothesen mit $\boldsymbol{R}_{ind}$ eingehalten. Bei $\boldsymbol{R}_{ar}$ ist ebenfalls ab $n_i = 15$ die Fehlerrate im geforderten Bereich, das empirische Niveau liegt hingegen nur für die Hypothese $H_0 : \beta_T = 0$ schon bei diesem Stichprobenumfang im Intervall $(4\%, 6\%)$. Beim Test auf keinen Gruppeneffekt liegt erst ab 50 Versuchseinheiten das empirische Niveau unterhalb von 6% und für die Wechselwirkung ist selbst dann diese Schranke noch nicht erreicht. Zuletzt kann noch $\boldsymbol{R}_{cs}$ betrachtet werden. Bei dieser Working Correlation Matrix ist die Fehlerrate erst bei $n_i = 50$ unter 10%. Dann liegt das empirische Niveau bei den Hypothesen $H_0 : \beta_T = 0$ und $H_0 : \beta_W = 0$ im geforderten Intervall. Für die Hypothese auf keinen Gruppeneffekt ist der Test bei $n_i = 50$ hingegen mit $\widehat{\alpha} = 7{,}13\%$ noch zu

Tabelle C.22: Niveausimulation für den Zeiteffekt der faktoriellen parametrischen Analyse bei 10 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	29,73	14,31	3,68	1,14	0,06	0,00
	$\widehat{\alpha}$	7,33	6,80	5,45	5,51	5,29	5,09
$\boldsymbol{R}_{ar}$	f	31,84	14,62	3,69	1,14	0,06	0,00
	$\widehat{\alpha}$	6,03	5,32	4,60	4,94	4,67	4,89
$\boldsymbol{R}_{cs}$	f	66,01	49,79	33,29	23,74	13,45	4,92
	$\widehat{\alpha}$	6,88	6,55	5,23	4,97	5,19	4,99
$\boldsymbol{R}_{mde}$	f	91,52	87,83	80,47	74,67	62,92	48,08
	$\widehat{\alpha}$	4,01	4,60	3,23	3,43	3,32	3,68
$\boldsymbol{R}_{un}$	f	97,28	92,07	92,98	91,24	89,21	86,63
	$\widehat{\alpha}$	2,94	2,02	2,56	4,45	5,19	6,73

Tabelle C.23: Niveausimulation für die Wechselwirkung der faktoriellen parametrischen Analyse bei 10 Kategorien

	n_i	7	10	15	20	30	50
$\boldsymbol{R}_{ind}$	f	29,73	14,31	3,68	1,14	0,06	0,00
	$\widehat{\alpha}$	6,79	6,97	5,80	5,36	5,80	5,27
$\boldsymbol{R}_{ar}$	f	31,84	14,62	3,69	1,14	0,06	0,00
	$\widehat{\alpha}$	10,58	10,59	8,56	8,14	7,01	6,25
$\boldsymbol{R}_{cs}$	f	66,01	49,79	33,29	23,74	13,45	4,92
	$\widehat{\alpha}$	6,47	6,07	5,38	5,00	5,67	5,26
$\boldsymbol{R}_{mde}$	f	91,52	87,83	80,47	74,67	62,92	48,08
	$\widehat{\alpha}$	0,94	0,58	0,05	0,16	0,35	2,21
$\boldsymbol{R}_{un}$	f	97,28	92,07	92,98	91,24	89,21	86,63
	$\widehat{\alpha}$	6,62	2,14	0,43	0,91	1,20	0,60

antikonservativ.

Abschließend ist zu bemerken, dass man bei vielen Kategorien im Einzelfall Glück haben muss, damit die Verfahren bei wenigen Annahmen an die Struktur von $\boldsymbol{R}$ überhaupt Ergebnisse liefern. Es sei denn, der Stichprobenumfang ist sehr groß. Um das Niveau einzuhalten, ist aber unter Umständen eine noch weitergehende Erhöhung des Stichpro-

benumfanges notwendig. Häufig wird man sich mit der Wahl von $\boldsymbol{R}_{ind}$ zufrieden geben müssen. Diese hält auch bei mäßig großen Stichprobenumfängen das Niveau ein. Bei sehr wenigen Kategorien können hingegen die Restriktionen an $\boldsymbol{R}$ minimiert werden. Ab 20 oder 30 Versuchseinheiten pro Gruppe zeigen die Ergebnisse hier zufriedenstellende Werte für die Fehlerraten und das empirische Niveau.

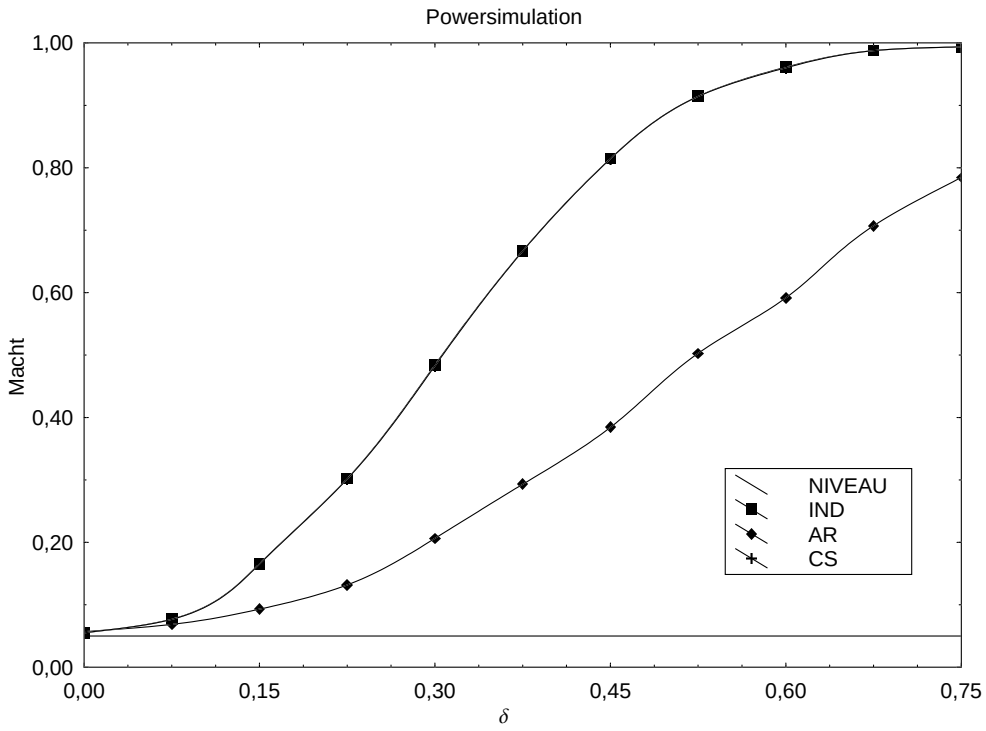

Abbildung C.5: Powersimulation für die parametrische faktorielle Analyse mit verschiedenen Working Correlation Matrizen

Für die drei Working Correlation Matrizen $\boldsymbol{R}_{ind}$, $\boldsymbol{R}_{ar}$ und $\boldsymbol{R}_{cs}$ wurde eine Powersimulation für den Effekt der Wechselwirkung durchgeführt. Die beiden anderen Working Correlation Matrizen wurden aufgrund der schlechten Eigenschaften beim Niveau nicht betrachtet. Die Erzeugung der Simulationsdaten erfolgte dabei wie bei dem Machtvergleich zur nichtparametrischen Kovarianzanalyse. In Abbildung C.5 ist zu erkennen, dass die Teststatistiken mit der Working Correlation Matrix $\boldsymbol{R}_{ind}$ nicht nur das Niveau gut einhalten und gut zu berechnen sind, sie haben auch eine vergleichsweise gute Power in dieser Simulationsstudie.

C.5 Parametrische Kovarianzanalyse

Tabelle C.24: Niveausimulation für die parametrische Kovarianzanalyse bei 3, 5 oder 10 Kategorien und einem nominellen Niveau von 5%

Anzahl r der Kategorien	Stichprobenumfang n_i pro Gruppe						
	7	10	15	20	30	50	100
3	3,21	5,51	5,99	5,50	6,01	5,38	5,25
5	5,97	8,44	7,77	7,03	6,29	5,59	5,49
10	9,20	15,51	12,94	10,12	8,48	6,91	6,04

Um zu untersuchen, ab welchem Stichprobenumfang die Abweichungen von der asymptotischen Verteilung zu vernachlässigen sind, wurde eine Niveausimulation durchgeführt. Dabei wurden für die Stichprobenumfänge $n_i = 7, 10, 15, 20, 30, 50$ und 100 mit $r = 3, 5$ oder 10 Kategorien jeweils 10000 simulierte Datensätze ausgewertet. Diese entstanden nach dem Modell der parametrischen Kovarianzanalyse und die Ergebnisse sind in der Tabelle C.24 aufgelistet. Dort ist zu erkennen, dass bei drei Kategorien das nominelle Niveau nicht wesentlich überschritten wird. Hingegen ist bei kleinen Stichprobenumfängen der Test konservativ. Im Gegensatz dazu wird bei fünf und auch bei zehn Kategorien der Test für die meisten Stichprobenumfänge sehr antikonservativ. Erst bei über 100 Versuchseinheiten pro Gruppe sind die Abweichungen des empirischen vom nominellen Niveau bei zehn Kategorien zu tolerieren.

Anhang D

Auswertungen

In diesem Kapitel sollen die verschiedenen Verfahren, mit denen die nichtparametrische Kovarianzanalyse verglichen wurde, erläutert werden. Dies geschieht, indem die bereits vorgestellten Beispiele mit den verschiedenen Verfahren ausgewertet werden. So erhält man einen Eindruck, welche Auswirkungen beispielsweise die Modellannahmen haben.

D.1 Naive Verfahren

D.1.1 Neurologische Beeinträchtigung bei MS

Im vorliegenden Beispiel sollen zwei Behandlungsmethoden verglichen werden. Dazu wurden den Gruppen *Behandlung* und *Kontrolle* Patienten randomisiert zugewiesen. Die naiven Verfahren untersuchen, ob die Wahrscheinlichkeiten für eine Verbesserung, Verschlechterung oder einen gleichbleibenden Zustand der Patienten in den beiden Gruppen identisch sind. Diese Wahrscheinlichkeiten werden mit q_{is} bezeichnet, welche durch die relativen Häufigkeiten geschätzt werden können. Diese sind in Prozent zusammen mit den absoluten Häufigkeiten, welche in Klammern angegeben sind, in der Kontingenztafel D.1 zusammengefasst.

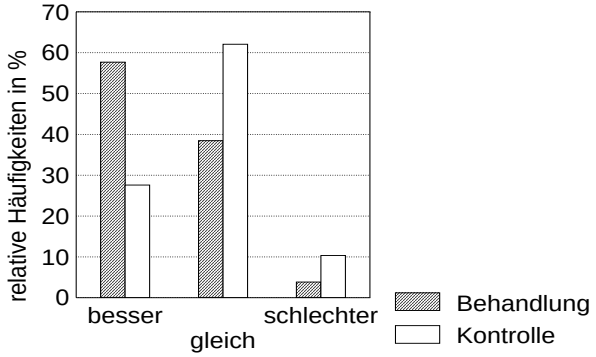

Abbildung D.1: Die relativen Häufigkeiten der beiden Gruppen bei der MS-Studie

Die Abbildung D.1 verdeutlicht die Zahlen noch einmal. Zunächst ist zu erkennen, dass

Tabelle D.1: Kontingenztafel der MS-Studie mit den relativen Häufigkeiten in Prozent und den dazugehörenden absoluten Häufigkeiten

	besser	gleich	schlechter	n_i
Behandlung	57,69 (15)	38,46 (10)	3,85 (1)	26
Kontrolle	27,59 (8)	62,07 (18)	10,34 (3)	29
	41,82 (23)	50,91 (28)	7,27 (4)	55

sich in beiden Gruppen der Zustand der Patienten eher verbessert als verschlechtert hat. Diese Verbesserung hängt jedoch deutlich von der betrachteten Gruppe ab. Während sich unter Behandlung nur ungefähr 4% verschlechtert haben, sind es unter Kontrolle mit circa 10% mehr als doppelt so viele. Außerdem haben in der Behandlungsgruppe mit 58% ungefähr doppelt so viele Patienten nach der Behandlung verbesserte Werte als in der Kontrollgruppe mit 28%. Zuletzt ist zu bemerken, dass sich bei den meisten Patienten (62%) in der Kontrollgruppe der Zustand nicht messbar verändert hat. In der Behandlungsgruppe sind dies 38%. Die Effekte fassen diese Informationen zusammen. So beschreibt der Pearson-Effekt χ_P, der hier durch $\widehat{\chi}_P = 0{,}0327$ geschätzt wird, die Summe der quadratischen Unterschiede zwischen den Wahrscheinlichkeiten für Verbesserung, Verschlechterung und Konstanz der Werte. Er ist im Gegensatz zum Cochran-Armitage-Effekt schwer zu interpretieren. Dieser wird hier durch $\widehat{u}_{CA} = 0{,}3660$ geschätzt. Er setzt sich zusammen aus den 30%, die in der Behandlungsgruppe mehr Verbesserungen auftreten, und den gut 6%, die in der Behandlungsgruppe weniger Verschlechterungen auftreten. Die Verbesserung, die durch die Behandlung erreicht wird, muss also nicht durch mehr Verschlechterung bezahlt werden. Im Gegenteil gewinnt auch dort die Behandlung gegenüber der Kontrolle.

Abschließend wurden die beiden Hypothesen (4.13) und (4.14) zu einem Niveau von 5% getestet. Die Hypothese (4.13) bedeutet dabei, dass sich die beiden Gruppen hinsichtlich der Wahrscheinlichkeiten für eine Verbesserung oder Verschlechterung nicht unterscheiden. Die Wahrscheinlichkeit, dass der Zustand eines Patienten konstant bleibt, ist dann ebenfalls in den beiden Gruppen identisch. Die Hypothese (4.14) kann wie folgt interpretiert werden. Eine Erhöhung der Wahrscheinlichkeit für eine Verbesserung wird durch eine gleich große Erhöhung bei der Wahrscheinlichkeit für eine Verschlechterung erkauft. Mit den Teststatistiken $T_{\chi_P} = 5{,}2682$ und $T_{u_{CA}} = 2{,}2234$ ergeben sich die p-Werte 7,18% und 2,62%. Damit ergibt sich eine absurde Situation. Obwohl die Hypothese (4.13) die

Hypothese (4.14) impliziert, wird (4.14) im Gegensatz zu (4.13) verworfen. Während also beim Cochran-Armitage-Test ein Unterschied zwischen den Behandlungsgruppen nachgewiesen werden kann, liefert der Pearson-Test keinen Hinweis auf einen Unterschied. Diese Diskrepanz kann zwei Gründe haben. Zum einen ist der Pearson-Test sehr konservativ, falls nur wenige Beobachtungen in einige Zellen der Kontingenztafel fallen. Dies ist hier der Fall, da sich in der Behandlungsgruppe nur die Werte eines und in der Kontrollgruppe nur die Werte von drei Patienten verschlechtern. Zum anderen unterscheiden sich die beiden Tests hinsichtlich der Macht, was ebenfalls zu inkonsistenten Entscheidungen führen kann.

D.1.2 Behandlung von Schlaflosigkeit

Im zweiten Beispiel wird eine aktive Behandlung der Schlaflosigkeit mit einer Placebobehandlung verglichen. Die Gruppe mit der aktiven Behandlung wird als Behandlungsgruppe und die andere als Kontrollgruppe bezeichnet. Die Schätzer der Wahrscheinlichkeiten q_{is}, also die relativen Häufigkeiten, sind für dieses Beispiel in der Tabelle D.2 aufgelistet und in Abbildung D.2 anschaulich dargestellt. Es ist zu erkennen, dass sich die beiden Gruppen bezüglich der Verschlechterung kaum unterscheiden. In beiden Fällen liegt bei gut 8% der Patienten nach der Behandlung eine längere Einschlafzeit vor als vor der Behandlung. Deutlich wird der Unterschied zwischen den beiden Gruppen hingegen bei den anderen beiden Kategorien. Während sich in der Kontrollgruppe die restlichen Patienten ungefähr zu gleichen Teilen auf die beiden Kategorien *besser* und *gleich* aufteilen, ist das Verhältnis dieser beiden Kategorien in der Behandlungsgruppe fast 3 : 1.

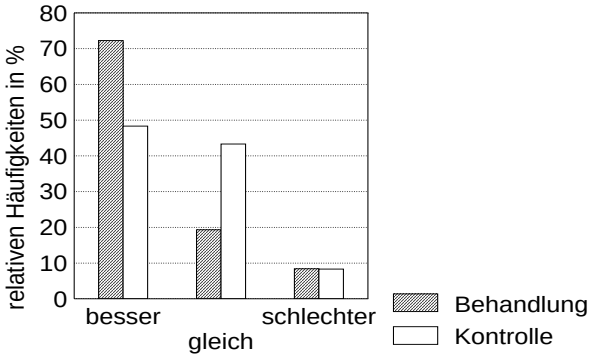

Abbildung D.2: Die relativen Häufigkeiten der beiden Gruppen bei der Insomnia-Studie

Dieser deutliche Unterschied ist auch an dem Schätzer des Cochran-Armitage-Effektes von $\widehat{u}_{CA} = 0{,}3463$ zu erkennen. Dieser ist ähnlich groß wie bei der MS-Studie. Der Schätzer des Pearson-Effektes ist mit $\widehat{\chi}_p = 0{,}0265$ etwas kleiner als im vorhergehenden Beispiel. Für die Tests der Nullhypothesen (4.13) und (4.14), dass diese Effekte null sind, ergeben

Tabelle D.2: Kontingenztafel der Insomnia-Studie mit den relativen Häufigkeiten in Prozent und den absoluten Häufigkeiten in Klammern

	besser	gleich	schlechter	n_i
Behandlung	72,27 (86)	19,33 (23)	8,40 (10)	119
Kontrolle	48,33 (58)	43,33 (52)	8,33 (10)	120
	60,25 (144)	31,38 (75)	8,37 (20)	239

sich mit den Teststatistiken $T_{\chi_P} = 19{,}5433$ und $T_{u_{CA}} = 3{,}9632$ p-Werte, welche kleiner als 0,0001 sind. Die Approximation an die χ^2_2-Verteilung kann im Gegensatz zum vorhergehenden Beispiel bei den vorliegenden Stichprobenumfängen und den geschätzten Wahrscheinlichkeiten der Kontingenztafel als ausreichend angesehen werden Damit fällt in beiden Tests die Entscheidung zu Gunsten der Alternative. Wie im vorhergehenden Beispiel kann man also von unterschiedlichen Behandlungswirkungen in den Gruppen ausgehen.

D.2 Nichtparametrische faktorielle Analyse

D.2.1 Neurologische Beeinträchtigung bei MS

In diesem Abschnitt wird die MS-Studie mit Hilfe der nichtparametrischen faktoriellen Verfahren ausgewertet. Dabei können der feste Faktor *Behandlung* mit den Stufen *Behandlung* und *Kontrolle* und der feste Faktor *Zeit* mit den Stufen *vorher* und *nachher* betrachtet werden. Vor allem geht es jedoch um die Frage, ob sich die beiden Gruppen über die Zeit hinweg unterschiedlich entwickeln. Mithin ist die primäre Fragestellung, ob eine Wechselwirkung zwischen den beiden festen Faktoren vorliegt.

Um die Einflüsse der Faktoren zu beschreiben, wurden zunächst die Effekte mit Hilfe der Verteilungsfunktionen vorgeschlagen (Akritas und Arnold, 1994). Da diese aber nicht zu interpretieren sind, gehen wir auf die relativen Effekte ein. Diese werden für jede Faktorstufenkombination (i, j) gebildet. Die mit den Stichprobenumfängen gewichteten und die ungewichteten relativen Effekte werden durch $\widehat{p}_{N,ij}$ und $\widehat{p}_{ij}$ geschätzt. Diese Schätzer sind in der Tabelle D.3 für die MS-Studie aufgelistet. Dabei kennzeichnet der Index $i = 1$ die Behandlungs- und $i = 2$ die Kontrollgruppe. Des Weiteren wird für den ersten Zeitpunkt $j = 1$ und für den zweiten Zeitpunkt $j = 2$ verwendet. Da die Stichprobenumfänge mit $n_1 = 26$ Versuchseinheiten in der Behandlungsgruppe und $n_2 = 29$ Versuchseinhei-

Tabelle D.3: Die gewichteten und ungewichteten relativen Effekte im faktoriellen Design für die MS-Studie

		$\widehat{p}_{N,ij}$	$\widehat{p}_{ij}$
Behandlung $i = 1$	vorher $j = 1$	0,5822	0,5826
	nachher $j = 2$	0,4010	0,4015
Kontrolle $i = 2$	vorher $j = 1$	0,5171	0,5175
	nachher $j = 2$	0,4980	0,4984

ten in der Kontrollgruppe fast gleich sind, unterscheiden sich die ungewichteten relativen Effekte nur wenig von den gewichteten relativen Effekten. Daher wurden nur die gewichteten relativen Effekte in der Abbildung D.3 dargestellt. Es ist zu erkennen, dass sich die relativen Effekte trotz der Randomisierung vor der Behandlung unterscheiden. Die Behandlungsgruppe hat schlechtere Anfangswerte als die Kontrollgruppe, was an dem größeren relativen Effekt zu erkennen ist. Je größer der relative Effekt, desto mehr tendieren die Zufallsvariablen der zugehörigen Faktorstufenkomination zu höheren Werten auf der ordinalen Skala. Da in der hier verwendeten DSS-Skala die hohen Werte einen schlechten Zustand des Patienten bewerten, deuten große relative Effekte auf einen schlechten Zustand der Patienten in der betrachteten Faktorstufenkombination hin.

Zum zweiten Zeitpunkt, also nach der Behandlung, ist der relative Effekt in der Kontrollgruppe etwas geringer als zum ersten Zeitpunkt. In der Behandlungsgruppe nimmt er hingegen stark ab. Während der relative Effekt der Behandlungsgruppe zum ersten Zeitpunkt größer als der der Kontrollgruppe ist, liegt er zum zweiten Zeitpunkt deutlich unter dem relativen Effekt der Kontrollgruppe. Das Verhältnis von Kontroll- und Behand-

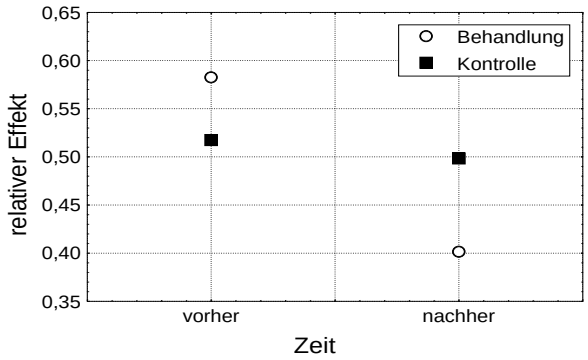

Abbildung D.3: Die gewichteten relativen Effekte der MS-Studie

lungsgruppe hat sich damit umgekehrt. Dies spricht für die Behandlung und gegen die Kontrolle.

Um die Einflüsse der beiden festen Faktoren und ihrer Wechselwirkung zu untersuchen, können die Effekte

$$\begin{aligned} p_G &= \int \frac{1}{2}\left(F_{11}+F_{12}\right) d\left(\frac{1}{2}\left(F_{21}+F_{22}\right)\right) \\ &= \frac{1}{2}-\frac{1}{8}\left(p_{11}+p_{12}-p_{21}-p_{22}\right), \\ p_T &= \int \frac{1}{2}\left(F_{11}+F_{21}\right) d\left(\frac{1}{2}\left(F_{12}+F_{22}\right)\right) \\ &= \frac{1}{2}-\frac{1}{8}\left(p_{11}-p_{12}+p_{21}-p_{22}\right) \end{aligned}$$

und

$$\begin{aligned} p_W &= \int \frac{1}{2}\left(F_{11}+F_{22}\right) d\left(\frac{1}{2}\left(F_{11}+F_{22}\right)\right) \\ &= \frac{1}{2}-\frac{1}{8}\left(p_{11}-p_{12}-p_{21}+p_{22}\right) \end{aligned}$$

herangezogen werden. Deren Schätzer ergeben die Werte $\widehat{p}_G = 0{,}5040$, $\widehat{p}_T = 0{,}4750$ und $\widehat{p}_W = 0{,}4798$. Da der Schätzer für den relativen Effekt der Wechselwirkung einen von 0,5 unterschiedlichen Wert hat, ist die Interpretation der anderen beiden Schätzer problematisch. Für die Interpretation von $\widehat{p}_W$ kann die Darstellung mit Hilfe von Wahrscheinlichkeiten (4.20) verwendet werden. Anhand der Abbildung D.3 ist gut zu erkennen, dass die Zufallsvariable Y_1, welche mit gleicher Wahrscheinlichkeit nach $\widehat{F}_{11}$ oder $\widehat{F}_{22}$ verteilt ist, zu höheren Werten tendiert als die Zufallsvariable Y_2, die mit gleicher Wahrscheinlichkeit nach $\widehat{F}_{12}$ oder $\widehat{F}_{21}$ verteilt ist.

Um diese Ergebnisse zu untermauern, können nun die verschiedenen Hypothesen auf keinen Einfluss der festen Faktoren und ihrer Wechselwirkung getestet werden. Dabei ist zu unterscheiden, ob die Hypothesen bezüglich der Marginalverteilungen

$$\begin{aligned} H_0: \quad F_G &= 0, \\ H_0: \quad F_T &= 0 \quad \text{und} \\ H_0: \quad F_W &= 0 \end{aligned}$$

mit

$$F_G = F_{11}+F_{12}-F_{21}-F_{22},$$

$$F_G = F_{11} + F_{12} - F_{21} - F_{22} \quad \text{und}$$
$$F_G = F_{11} + F_{12} - F_{21} - F_{22}$$

oder bezüglich der relativen Effekte

$$H_0: \quad p_G = 0,$$
$$H_0: \quad p_T = 0 \quad \text{und}$$
$$H_0: \quad p_W = 0$$

gestellt werden. In beiden Fällen sollen die Tests zum Niveau 5% durchgeführt werden. Für die Teststatistiken werden sowohl die Approximation mittels der Normalverteilung als auch mittels der t-Verteilung betrachtet. Es ist dabei zu bedenken, dass einige Tests bei den vorliegenden Stichprobenumfängen etwas antikonservativ sind, das Niveau also nicht ganz einhalten.

Tabelle D.4: Übersicht der Testergebnisse für die MS-Studie

Hypothese	Teststatistik	p-Wert bei χ^2_1-Verteilung	p-Wert bei t-Verteilung	Freiheitsgrad der t-Verteilung
$H_0: F_G = 0$	0,0540	0,8162	0,8171	52,9247
$H_0: F_T = 0$	12,9261	0,0003	0,0008	47,3314
$H_0: F_W = 0$	8,4600	0,0036	0,0055	47,3314
$H_0: p_G = 0$	0,0536	0,8169	0,8178	52,8294
$H_0: p_T = 0$	12,8961	0,0003	0,0008	48,5461
$H_0: p_W = 0$	8,5678	0,0034	0,0052	47,5747

Die Ergebnisse aller Tests sind in der Tabelle D.4 aufgelistet. In der ersten Spalte stehen die zu testenden Hypothesen. Danach sind die Quadrate der asymptotisch normalverteilten Statistiken angegeben. Durch einen Vergleich mit den Quantilen der χ^2_1-Verteilung erhält man dann die p-Werte, die in der dritten Spalte stehen. Verwendet man hingegen die Wurzeln der Teststatistiken und berechnet dann die p-Werte mit Hilfe der t-Approximation, so erhält man die p-Werte in der vierten Spalte. Die dabei verwendeten geschätzten Freiheitsgrade der t-Verteilung sind in der letzten Spalte angegeben. Da sie alle recht hoch sind und somit die Unterschiede zwischen der Normal- und der t-Verteilung gering sind, unterscheiden sich die p-Werte pro Zeile kaum voneinander.

Die Hypothesen auf keinen Einfluss der Behandlung können sowohl im Fall der Marginalverteilungen als auch im Fall der relativen Effekte nicht verworfen werden. Hingegen

liegen die p-Werte für die Hypothesen auf keinen Einfluss der Zeit alle unterhalb von 5%. Damit können die Hypothesen $H_0 : F_T = 0$ und $H_0 : p_T = 0$ verworfen werden. Da jedoch auch die Hypothesen $H_0 : F_W = 0$ und $H_0 : p_W = 0$ verworfen werden, muss von einer vorliegenden Wechselwirkung ausgegangen werden. Damit sind die Hypothesen über die Haupteffekte kaum zu interpretieren. Verwendet man die Referenzkodierung, so hängt es hier von der Wahl der Referenz ab, ob ein Zeiteffekt vorliegt oder nicht. Denn bei der Kontrollgruppe ändert sich der relative Effekt im Gegensatz zur Behandlungsgruppe kaum. Beim Gruppeneffekt ergeben sich ähnliche Probleme.

Die Simulationen haben gezeigt, dass bei den vorliegenden Stichprobenumfängen die Tests für die Wechselwirkung konservativ sind. Nur im Fall der Hypothese über die Marginalverteilungen und bei Verwendung der Normalverteilung wird das nominelle Niveau leicht überschritten (siehe Tabelle C.8). Da aber alle p-Werte recht klein sind, kann insgesamt davon ausgegangen werden, dass die vorgegebene Irrtumswahrscheinlichkeit für einen Fehler erster Art nicht überschritten wird.

D.2.2 Behandlung von Schlaflosigkeit

Wie im vorhergehenden Abschnitt werden zunächst die relativen Effekte betrachtet. Die Bezeichnung mittels der Indices wird dabei von der MS-Studie übernommen. In der Tabelle D.5 sind die gewichteten und die ungewichteten relativen Effekte aufgelistet. Da sich das Verhältnis der beiden Stichprobenumfänge, in der Behandlungsgruppe sind 119 und in der Kontrollgruppe sind 120 Versuchseinheiten, kaum von eins unterscheidet, ist auch zwischen den ungewichteten und den gewichteten relativen Effekten erst in der vierten Nachkommastelle ein Unterschied festzustellen. Die gewichteten relativen Effekte sind in der Abbildung D.4 dargestellt. Für die ungewicheten relativen Effekte ergibt sich auf Grund der sehr feinen Unterschiede das gleiche Bild.

Tabelle D.5: Die relativen Effekte im faktoriellen Design für die Insomnia-Studie

		$\widehat{p}_{N,ij}$	$\widehat{p}_{ij}$
Behandlung $i = 1$	vorher $j = 1$	0,6012	0,6014
	nachher $j = 2$	0,3413	0,3414
Kontrolle $i = 2$	vorher $j = 1$	0,6022	0,6023
	nachher $j = 2$	0,4548	0,4549

Die Schätzer $\widehat{p}_{11} = 0{,}6014$ und $\widehat{p}_{21} = 0{,}6023$ der relativen Effekte p_{11} und p_{21}, also die

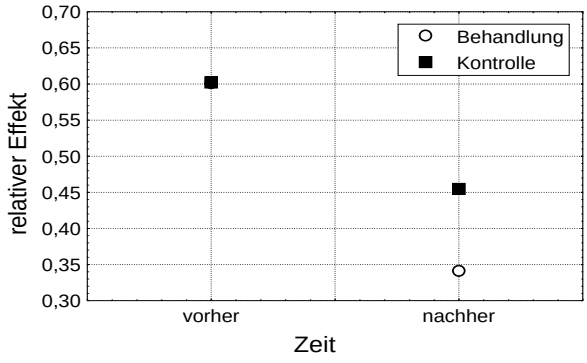

Abbildung D.4: Die gewichteten relativen Effekte der Insomnia-Studie

relativen Effekte vor der Behandlung, unterscheiden sich kaum. Daher ist in der Abbildung D.4 zum ersten Zeitpunkt nur der relative Effekt der Kontrollgruppe zu erkennen. Diese Gleichheit sollte auch vorhanden sein, da es sich um eine randomisierte Studie handelt. Man kann mithin davon ausgehen, dass die Zufallsvariablen zum ersten Zeitpunkt nicht in der einen Gruppe zu höheren Werten tendieren als in der anderen Gruppe. Mit anderen Worten sind die empirischen Marginalverteilungen der beiden Gruppen zum ersten Zeitpunkt $\widehat{F}_{11}$ und $\widehat{F}_{12}$ mit Bezug auf die mittlere empirische Marginalverteilung $\widehat{H}$ tendenziell nahezu gleich.

Zum zweiten Zeitpunkt fällt in beiden Gruppen der relative Effekt deutlich ab. Dies spricht für einen starken Zeiteffekt, der als Placeboeffekt interpretiert werden kann. Allerdings ist der Unterschied vom ersten zum zweiten Zeitpunkt in der Behandlungsgruppe deutlich größer als in der Kontrollgruppe. Dies spricht für die Behandlung, denn die Marginalverteilung $\widehat{F}_{12}$ ist tendenziell kleiner als $\widehat{F}_{22}$ mit Bezug auf die mittlere Verteilungsfunktion $\widehat{H}$.

Um festzustellen, inwieweit solche Ergebnisse unter den verschiedenen Hypothesen wahrscheinlich sind, werden die Verfahren der nichtparametrischen faktoriellen Analyse angewendet. Die Ergebnisse dieser Tests sind in der Tabelle D.6 zusammengefasst. Alle Hypothesen können zu einem Niveau von 5% verworfen werden. Dabei liegt der p-Wert der Tests auf keinen Einfluss der Gruppen, $H_0 : \ F_G = 0$ und $H_0 : \ p_G = 0$, mit 4,39% beziehungsweise 4,50% nur knapp unterhalb des Niveaus. Insgesamt unterscheiden sich die p-Werte der beiden Approximationen kaum, weil die geschätzten Freiheitsgrade der t-Verteilung sehr groß sind. Dass beide Verfahren das Niveau einhalten, kann bei dem vorliegenden Stichprobenumfang mit Blick auf die Simulationen als gesichert betrachtet werden.

In Hinblick auf die Abbildung D.4 kann zunächst das Verwerfen der Hypothese auf

Tabelle D.6: Übersicht der Testergebnisse für die Insomnia-Studie

Hypothese	Teststatistik	p-Wert bei χ^2_1-Verteilung	p-Wert bei t-Verteilung	Freiheitsgrad der t-Verteilung
$H_0: F_G = 0$	4,0620	0,0439	0,0450	230,7764
$H_0: F_T = 0$	135,2884	0,0000	0,0000	235,3218
$H_0: F_W = 0$	10,3103	0,0013	0,0015	235,3218
$H_0: p_G = 0$	4,0615	0,0439	0,0450	229,4750
$H_0: p_T = 0$	133,1717	0,0000	0,0000	236,3623
$H_0: p_W = 0$	10,3834	0,0013	0,0015	235,8476

keinen Zeiteffekt gut interpretiert werden. Da in beiden Gruppen die relativen Effekte nach der Behandlung wesentlich kleiner sind als vor der Behandlung, ist dieser Zeiteffekt als Placeboeffekt zu interpretieren. Die Signifikanz der Wechselwirkung deutet zudem auf eine bessere Absenkung der Werte in der Behandlungs- als in der Kontrollgruppe hin. Einzig der Gruppeneffekt ist hier schwierig zu interpretieren, da sich die Gruppen zum ersten Zeitpunkt bezüglich der relativen Effekte nicht unterscheiden. Ein Gruppeneffekt sollte aber auch hier schon sichtbar sein. Die Signifikanz kann also nur durch den Unterschied zwischen den Gruppen zum zweiten Zeitpunkt erzeugt worden sein. Diesen Unterschied haben wir jedoch der Wechselwirkung zugeschrieben. Diese Diskussion führt zurück auf das Problem der Reparametrisierung in additiven Modellen.

D.3 Nichtparametrische Analyse durch Schichtung

D.3.1 Neurologische Beeinträchtigung bei MS

Die nichtparametrische Schichtungsanalyse beruht auf der Schichtung der Beobachtungen zum zweiten Zeitpunkt nach den Beobachtungen zum ersten Zeitpunkt. Nur innerhalb der Schichten werden die beiden Gruppen miteinander verglichen. In Hinsicht auf diese Betrachtungsweise sind die Daten der MS-Studie in der Tabelle D.7 zusammengefasst.

Dort sind die verschiedenen Schichten als Zeilen zu erkennen. Diese werden mit den DSS-Scores zu Beginn bezeichnet. Da nur die Scores 3 bis 9 zum ersten Zeitpunkt beobachtet wurden, gibt es auch nur diese 6 Schichten. Die Schichtungsverfahren können aber nicht die gesamte Information in dieser Tabelle verarbeiten. Die schwächsten Einschränkungen benötigen die Verfahren, welche auf den U-Statistiken beruhen. Aber selbst

Tabelle D.7: Die Daten der MS-Studie in geschichtet nach den Beobachtungen zum ersten Zeitpunkt

DSS-Score zu Beginn	DSS Score nach 20 Wochen	
	Behandlungsgruppe	Kontrollgruppe
3	0	5
4		3, 6
5	3, 3, 5	3, 3, 4, 4
6	2, 3, 6, 6, 6, 6, 6, 6, 7	6, 6, 6, 6, 6, 6, 6, 6, 6, 9
7	5, 6, 6, 6, 6, 7, 7	5, 6, 7, 7, 7, 7, 7
8	6, 6, 8	8, 8, 8
9	6, 6, 9	7, 9

bei diesen Verfahren können die Beobachtungen der Kontrollgruppe in der Schicht 4 nicht verwendet werden. Diese beiden Beobachtungen fallen aus der Analyse heraus, da in der Behandlungsgruppe keine vergleichbaren Beobachtungen liegen. Der Stichprobenumfang reduziert sich damit in der Kontrollgruppe von 29 auf 27 Patienten. Für die anderen Verfahren gehen zwei weitere Beobachtungen verloren. Da in der Schicht 3 jeweils nur eine Beobachtung vorliegt, kann die Varianz der Statistik in dieser Schicht nicht geschätzt werden.

Die Sensibilität der Verfahren, welche die Hypothesen über die relativen Effekte pro Schicht stellen, ist noch größer. Wäre bei dem Patienten in der Schicht 8 der Behandlungsgruppe, welcher auch zum zweiten Zeitpunkt in die Kategorie 8 fällt, nur eine Verbesserung um einen Skalenpunkt beobachtet worden, so wäre die Bedingung (4.24) nicht erfüllt gewesen. Dies hätte zum Verlust von 3 Beobachtungen in jeder Gruppe geführt, was mehr als 10% der Daten entspricht. Ein derart großer Verlust an Daten erscheint inakzeptabel.

In der Tabelle D.7 ist nicht nur zu erkennen, dass die schwache Besetzung einiger Zeilen zum Verlust von Information führt, die stark unterschiedlichen Stichprobenumfänge in den Zeilen müssen sich auch in einer unterschiedlichen Gewichtung niederschlagen. Dafür wurden zwei Verfahren vorgeschlagen. Die Gewichte nach van Elteren (1960) und nach Bajorski und Petkau (1999) sind in der Tabelle D.8 zusammengefasst. Da die Gewichte stark unterschiedliche Größenordnungen haben, sind auch die prozentualen Anteile der Schichten am Gesamtgewicht angegeben. Des Weiteren ergeben sich für die Verfahren, denen die Theorie der U-Statistiken zu Grunde liegt, andere Gewichte, da diese auch die Schicht 3 berücksichtigen.

Tabelle D.8: Die Gewichte der Schichten bei der MS-Studie

Schicht u	van Elteren		Bajorski und Petkau	
	absolut	prozentual	absolut	prozentual
3	0,00629	2,8045	0,0503	3,3031
5	0,02830	12,6202	0,1678	11,0148
6	0,08491	37,8606	0,7066	46,3857
7	0,06164	27,4840	0,3196	20,9806
8	0,02426	10,8173	0,1698	11,1478
9	0,01887	8,4135	0,1092	7,1681
5	1,50000	12,9843	8,8926	11,3911
6	4,50000	38,9530	37,4486	47,9701
7	3,26667	28,2770	16,9383	21,6972
8	1,28571	11,1294	9,0000	11,5286
9	1,00000	8,6562	5,7870	7,4130

Zunächst sollen nun die Gewichte diskutiert werden, bei denen auch die Schicht 3 einen positiven Anteil am Gesamtgewicht erhält. Es fällt auf, dass ungefähr zwei Drittel des Gewichtes sowohl bei van Elteren (1960) als auch bei Bajorski und Petkau (1999) sich auf die Schichten 6 und 7 konzentriert. Dies ist verständlich, da sie mit 33 Beobachtungen auch ungefähr diesen Anteil an den insgesamt 53 verwendeten Beobachtungen haben. Bei diesen beiden Schichten unterscheiden sich die beiden Gewichtungsverfahren aber erheblich. Während bei Bajorski und Petkau (1999) das Verhältnis des Gewichtes von Schicht 6 zum Gewicht von Schicht 7 mehr als 2,2 beträgt, ist dieses Verhältnis bei der Gewichtung nach van Elteren (1960) nur knapp 1,4. Dies entspricht relativ genau dem Verhältnis der Stichprobenumfänge in den beiden Schichten von knapp 1,4. Das ist verständlich, da die Gewichte bei van Elteren (1960) nur von den Stichprobenumfängen abhängen. Die Gewichte von Bajorski und Petkau (1999) hängen außerdem umgekehrt proportional von $\widehat{b}_u = 1 - \sum_{v=1}^{r} \widehat{q}_{v|u}^3$ ab. Der Parameter $\widehat{b}_u$ ist in den Schichten u besonders klein, in denen ein großer Anteil der Beobachtungen in nur eine Kategorie fällt und somit die Varianz der Wilcoxon-Statistik in dieser Schicht besonders gering ist. Da in der Schicht 6 von 19 Beobachtungen 15 in die Kategorie 6 fallen, bekommt diese Schicht mehr als 46% des gesamten Gewichtes, obwohl nur knapp 36% der Beobachtungen in dieser Schicht liegen.

Ansonsten fällt bei den geschätzten relativen Effekten $\widehat{p}_{S,u}$ auf, dass kein Schätzer größer als $0,5$ ist. Dies spricht für die Behandlungsgruppe. Nur in der Schicht 5 sind die empirischen Verteilungsfunktionen tendenziell gleich. Dies ist leicht an den Daten zu

Tabelle D.9: Die relativen Effekte in den Schichten bei der MS-Studie

Schicht u	3	5	6	7	8	9
$\widehat{p}_{S,u}$	0,0	0,5	0,4	0,3163	0,1667	0,25

erkennen. Wenn alle Paare von Beobachtungen aus der einen und der anderen Gruppe betrachtet werden, dann findet man jeweils genau vier Paare, die für oder gegen die Behandlung sprechen. Bei den restlichen Paaren sind beide Beobachtungen gleich.

Nun sollen die Tests der Hypothese über die Verteilungsfunktionen $F_{i|u}$

$$H_0: \quad F_{1|u} \;=\; F_{2|u} \quad \text{für alle} \quad u \;=\; 1,\ldots,r$$

beziehungsweise über die relativen Behandlungseffekte $p_{S|u}$

$$H_0: \quad p_{S|u} \;=\; 1/2 \quad \text{für alle} \quad u \;=\; 1,\ldots,r$$

zum Niveau 5% durchgeführt werden. Dabei ist zu bedenken, dass bei dem vorliegenden Stichprobenumfang und bei 11 Kategorien die Tests das Niveau nicht einhalten. Man kann allerdings anführen, dass nur 5 beziehungsweise 6 Schichten besetzt sind und daher eher die Tabelle C.12 als C.13 zu betrachten ist. Doch auch in diesem Fall wird der Fehler erster Art nicht eingehalten. Besonders die Approximationen mit der gestreckten χ^2-Verteilung sind in dem vorliegenden Fall sehr antikonservativ. Um so mehr verwundert es, dass diese Statistiken $T_{S,F|\chi}$ und $T_{S,p|\chi}$ für beide Gewichte die Hypothesen nicht verwerfen, wie man in der Tabelle D.10 sieht. Dies kann nur damit erklärt werden, dass die Alternative, die von diesen Tests aufgedeckt wird, größer ist als bei den anderen Statistiken. Dafür kann es vorkommen, dass die Macht bei der vorliegenden Alternative geringer ist.

Bei den Statistiken $T_{S,F|n}$ und $T_{S,p|n}$ hängt es von der Wahl der Gewichte ab, ob sie verwerfen. Im Fall der Gewichte von Bajorski und Petkau liegen die zugehörigen p-Werte knapp oberhalb und im Fall von van Elteren (1960) knapp unterhalb des Niveaus. Die p-Werte von $T_{S,F|U}$ liegen für beide Gewichte bei ungefähr 1,5%, womit die Hypothesen auf Gleichheit der Verteilungsfunktionen in diesem Fall verworfen werden. Aber auch hier ist die Testentscheidung zu bezweifeln, da nicht klar ist, wie antikonservativ die Verfahren an dieser Stelle sind. Es steht wohl nur außer Frage, dass sie das Niveau nicht einhalten.

D.3.2 Behandlung von Schlaflosigkeit

Bei der Auswertung der Insomnia-Studie ergeben sich gegenüber der MS-Studie wesentlich weniger Probleme, da die ordinale Skala nur vier Kategorien hat und insgesamt 239

Tabelle D.10: Die Teststatistiken, deren Freiheitsgrade und die entsprechenden p-Werte für die MS-Studie

	van Elteren			Bajorski und Petkau		
	Statistik	p-Wert	Freiheitsgrad	Statistik	p-Wert	Freiheitsgrad
$T^2_{S,F\|n}$	4,2316	0,0397	1,0000	3,6826	0,0550	1,0000
$T_{S,F\|\chi}$	5,2049	0,2348	3,7253	4,4708	0,2445	3,2368
$T^2_{S,p\|n}$	4,0765	0,0435	1,0000	3,5662	0,0590	1,0000
$T_{S,p\|\chi}$	5,0342	0,2504	3,7253	4,3450	0,2571	3,2368
$T^2_{S,F\|U}$	5,9125	0,0150	1,0000	5,9311	0,0149	1,0000

Patienten betrachtet wurden. Die Verteilung dieser Patienten auf die vier Schichten ist in der Tabelle D.11 angegeben. Des Weiteren sind dort die verschiedenen Gewichte a_u in absoluten und prozentualen Werten angegeben.

Tabelle D.11: Die Gewichte und Stichprobenumfänge der Schichten bei der Insomnia-Studie

Schicht u	$m_{1u\cdot}$	$m_{2u\cdot}$	van Elteren		Bajorski und Petkau	
			absolut	prozentual	absolut	prozentual
1	12	14	0.02603	10.6160	0.1338	11.5488
2	20	20	0.04082	16.6454	0.2263	19.5320
3	40	35	0.07708	31.4291	0.3550	30.6420
4	47	51	0.10131	41.3095	0.4434	38.2772

Es ist zu erkennen, dass die Schichten von 1 bis 4 ansteigend besetzt sind. Ungefähr im gleichen Maße steigen auch die Gewichte an. Dabei ist der Unterschied zwischen den Gewichten von van Elteren (1960) und Bajorski und Petkau (1999) recht gering, wenn die prozentualen Werte betrachtet werden. In keiner Schicht tritt ein ähnliches Phänomen auf wie bei der MS-Studie.

Tabelle D.12: Die relativen Effekte in den Schichten bei der Insomnia-Studie

Schicht u	1	2	3	4
$\widehat{p}_{S\|u}$	0,4315	0,5963	0,2761	0,3277

Bei der Betrachtung der relativen Effekte ergibt sich jedoch im Gegensatz zur MS-Studie ein uneinheitliches Bild. Während in den Schichten 1, 3 und 4 die Behandlungsgruppe zu besseren Werten neigt als die Kontrollgruppe, ist das Umgekehrte in der Schicht 2 der Fall. Dort spricht der geschätzte relative Effekt gegen die Behandlungsgruppe. Insgesamt tendiert jedoch die Behandlungsgruppe zu den besseren Werten, da gerade bei den Patienten, die länger als eine halbe Stunde zum Einschlafen brauchen (Kategorie 3 und 4), die relativen Effekte sehr für sie sprechen. In diesen beiden Schichten liegen zudem mit 173 Patienten über 72% Prozent des Stichprobenumfanges.

Tabelle D.13: Die Teststatistiken, deren Freiheitsgrade und die entsprechenden p-Werte für die Insomnia-Studie

	van Elteren			Bajorski und Petkau		
	Statistik	p-Wert	Freiheitsgrad	Statistik	p-Wert	Freiheitsgrad
$T^2_{S,F\|n}$	18,4161	0,0000	1,0000	12,7037	0,0000	1,0000
$T_{S,F\|\chi}$	29,3845	0,0000	3,2425	29,7539	0,0000	3,4259
$T^2_{S,p\|n}$	19,0187	0,0000	1,0000	17,5454	0,0000	1,0000
$T_{S,p\|\chi}$	30,2580	0,0000	3,2425	30,6201	0,0000	3,4259
$T^2_{S,F\|U}$	14,5371	0,0001	1,0000	12.7037	0,0004	1,0000

Bei den diskutierten deskriptiven Ergebnissen ist es nicht verwunderlich, dass die Hypothesen über die Verteilungen und über die relativen Effekte von allen Statistiken auf einem Niveau von 5% verworfen werden. Die Statistiken mit den Freiheitsgraden der χ^2-Verteilung und die entsprechenden p-Werte sind für beide Gewichtungen in der Tabelle D.13 zusammengestellt. Bis auf die Tests mit den Statistiken $T_{S,F|\chi}$ und $T_{S,p|\chi}$, die hier leicht antikonservativ sind, halten alle Tests das Niveau bei einem Stichprobenumfang von 239 Versuchseinheiten und nur vier Kategorien ein.

D.4 Parametrische faktorielle Analyse

D.4.1 Neurologische Beeinträchtigung bei MS

Nach den nichtparametrischen Verfahren werden nun die parametrischen Verfahren angewendet. Wie zuvor beginnen wir mit der Analyse der MS-Studie. Die DSS-Skala, die bei dieser Studie verwendet wurde, hat die elf Kategorien 0 bis 10. Im Modell können hier die Kategorien 1 und 10 jedoch nicht berücksichtigt werden, da sie in der Studie nicht

beobachtet wurden. Daher wird im Folgenden mit neun Kategorien gearbeitet.

Für die weitere Modellbildung ist die Wahl der Reparametrisierungsbedingung wichtig. Aufgrund der einfacheren Interpretation, gerade bei randomisierten Studien, wird hier die Referenzkodierung verwendet. Die Wahl der Kontrollgruppe ($i = 2$) zum ersten Zeitpunkt ($j = 1$) als Referenz erscheint sinnvoll. Dabei werden die Indices i und j wie in den vorhergehenden Auswertungen verwendet. Mithin ist die additive Zerlegung der Logarithmen der odds

$$\begin{aligned}
\text{Behandlungsgruppe, vorher:} \quad L_{11s} &= \alpha_s + \beta_G, \\
\text{Kontrollgruppe, vorher} \quad L_{21s} &= \alpha_s, \\
\text{Behandlungsgruppe, nacher} \quad L_{12s} &= \alpha_s + \beta_G + \beta_T + \beta_W \quad \text{und} \\
\text{Kontrollgruppe, nachher} \quad L_{22s} &= \alpha_s + \beta_T.
\end{aligned}$$

Um die cutpoints α_s, den Parameter des Gruppeneffektes β_G, den Parameter des Zeiteffektes β_T und den Parameter der Wechselwirkung β_W zu schätzen, werden die GEE verwendet. Dies beruht auf einem Iterationsprozess, der von der SAS Prozedur GENMOD durchgeführt wird. Er konvergiert bei dem vorliegenden Beispiel nicht, wenn die Korrelationsmatrix $\boldsymbol{R}$ unstrukturiert ist. Dies ist damit zu erklären, dass zu wenig unabhängige Zufallsvariablen pro zu schätzendem Parameter vorliegen. Bei den anderen Korrelationsmatrizen wird die Anzahl der Parameter durch eine Strukturierung von $\boldsymbol{R}$ reduziert. Damit ergeben sich für jeden Parameter vier Schätzer, je nach Wahl der Korrelationsmatrix. Diese Schätzer sind in der Tabelle D.14 zusammengefasst. Insgesamt unterscheiden sich die verschiedenen Zeilen in dieser Tabelle kaum. Vor allem beim interessantesten Parameter β_W, der die Wechselwirkung beschreibt, sind die Unterschiede sehr gering. Nur der Schätzer des autoregressiven Ansatzes fällt etwas aus der Reihe.

Tabelle D.14: Parameterschätzer der parametrischen faktoriellen Analyse bei der MS-Studie

$\boldsymbol{R}$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\alpha}_5$	$\widehat{\alpha}_6$	$\widehat{\alpha}_7$	$\widehat{\alpha}_8$	$\widehat{\beta}_G$	$\widehat{\beta}_T$	$\widehat{\beta}_W$
$\boldsymbol{R}_{ind}$	-4,99	-4,29	-2,58	-2,20	-1,47	0,37	1,60	2,46	0,44	-0,22	-1,17
$\boldsymbol{R}_{ar}$	-4,91	-4,18	-2,46	-2,08	-1,38	0,43	1,65	2,48	0,61	-0,15	-1,21
$\boldsymbol{R}_{cs}$	-4,60	-3,98	-2,38	-2,01	-1,30	0,53	1,77	2,67	0,64	-0,22	-1,16
$\boldsymbol{R}_{mde}$	-5,04	-4,34	-2,52	-2,15	-1,48	0,35	1,56	2,49	0,49	-0,23	-1,16

Vorausgesetzt, dass das logistische Modell die Daten gut beschreibt, können mit diesen Schätzern die erwarteten Häufigkeiten der Kategorien für die verschiedenen Faktorstufenkombinationen geschätzt werden. Durch einen Vergleich mit den beobachteten Häufigkeiten kann beurteilt werden, ob das logistische Modell die Wirklichkeit ausreichend gut beschreibt. Für diesen Vergleich sind die geschätzten und die beobachteten Häufigkeiten in der Tabelle D.15 aufgelistet.

Tabelle D.15: Beobachtete und auf Grund der Schätzer erwartete Häufigkeiten der MS-Studie

Faktor		beobachtet	Kategorie								
Gruppe	Zeit	oder $\boldsymbol{R}$	0	2	3	4	5	6	7	8	9
		beobachtet	0	0	1	0	3	9	7	3	3
		$\boldsymbol{R}_{ind}$	0,1	0,1	1,0	0,5	1,6	9,2	7,2	3,2	3
$i = 1$	$j = 1$	$\boldsymbol{R}_{ar}$	0,1	0,1	0,9	0,5	1,5	8,7	7,4	3,3	3,4
		$\boldsymbol{R}_{cs}$	0,1	0,1	1,0	0,5	1,5	9,0	7,4	3,3	3,0
		$\boldsymbol{R}_{mde}$	0,1	0,1	1,0	0,5	1,5	8,9	7,3	3,5	3,1
		beobachtet	1	1	3	0	2	14	4	0	1
		$\boldsymbol{R}_{ind}$	0,4	0,4	3,4	1,6	3,9	10,8	3,6	1,1	0,8
$i = 1$	$j = 2$	$\boldsymbol{R}_{ar}$	0,4	0,4	3,2	1,5	3,6	10,9	3,9	1,2	1,0
		$\boldsymbol{R}_{cs}$	0,5	0,4	3,2	1,5	3,7	10,8	3,8	1,1	0,8
		$\boldsymbol{R}_{mde}$	0,4	0,4	3,4	1,5	3,6	10,9	3,8	1,2	0,9
		beobachtet	0	0	1	2	4	10	7	3	2
		$\boldsymbol{R}_{ind}$	0,2	0,2	1,6	0,9	2,5	11,7	6,9	2,6	2,3
$i = 2$	$j = 1$	$\boldsymbol{R}_{ar}$	0,2	0,2	1,8	0,9	2,6	11,8	6,8	2,4	2,2
		$\boldsymbol{R}_{cs}$	0,3	0,2	1,9	1,0	2,8	12,0	6,6	2,3	1,9
		$\boldsymbol{R}_{mde}$	0,2	0,2	1,8	0,9	2,4	11,6	7,0	2,8	2,2
		beobachtet	0	0	3	2	2	11	6	3	2
		$\boldsymbol{R}_{ind}$	0,2	0,2	2,0	1,0	2,9	12,2	6,3	2,2	1,9
$i = 2$	$j = 2$	$\boldsymbol{R}_{ar}$	0,2	0,3	2,1	1,1	2,9	12,1	6,3	2,2	1,9
		$\boldsymbol{R}_{cs}$	0,4	0,3	2,3	1,2	3,2	12,3	5,9	1,9	1,5
		$\boldsymbol{R}_{mde}$	0,2	0,2	2,2	1,0	2,8	12,1	6,3	2,4	1,8

Es ist zu erkennen, dass die geschätzten Häufigkeiten bei ungefähr zwei Dritteln der Zellen recht nahe an den beobachteten Häufigkeiten liegen. Dort beträgt die Differenz zwischen geschätzter und beobachteter Häufigkeit weniger als 1. Vor allem in der Be-

handlungsgruppe zum ersten Zeitpunkt passt das Modell gut. Doch bei den mittleren Kategorien von 3 bis 6 entstehen deutliche Unterschiede zwischen den beobachteten und den geschätzten Häufigkeiten. Vor allem in der am meisten beobachteten Kategorie 6 ist die Diskrepanz bedeutend.

Ein weiterer wichtiger Aspekt des Modells ist die Proportional Odds Annahme. Um diese zu überprüfen, wurde vorgeschlagen die beobachteten log Odds ähnlich der Abbildung 4.8 darzustellen. Dies wirft bei der vorliegenden Studie Probleme auf. Die niedrigen Kategorien wurden in mehreren Faktorstufen nicht beobachtet, womit dort die log Odds nicht definiert sind, da die Odds null sind. Außerdem wird die Abbildung bei zu vielen Kategorien sehr unübersichtlich. Daher wurde eine solche Abbildung hier nicht bereitgestellt.

Diese Probleme bezüglich der Proportional Odds Annahme beim Modell sind auch bei der nun folgenden Schätzung der Effekte zu berücksichtigen. Die Effekte haben nur dann Aussagekraft, falls das Modell die Wirklichkeit gut beschreibt. Mit Hilfe der Parameterschätzer $\widehat{\beta}_G$, $\widehat{\beta}_T$ und $\widehat{\beta}_W$ können die Odds Ratios, die interessierenden Parameter im logistischen Modell, geschätzt werden. Es gilt

$$\begin{aligned}
\widehat{\kappa}_{21,11} &= \exp(-\widehat{\beta}_G), \\
\widehat{\kappa}_{21,22} &= \exp(-\widehat{\beta}_T), \\
\widehat{\kappa}_{11,12} &= \exp(-\widehat{\beta}_T - \widehat{\beta}_W) \quad \text{und} \\
\widehat{\kappa}_{22,12} &= \exp(-\widehat{\beta}_G - \widehat{\beta}_W).
\end{aligned}$$

Die mit diesen Gleichungen berechneten Schätzer sind der Tabelle D.16 zu entnehmen. Da die Parameterschätzer sich für die unterschiedlichen Korrelationsmatrizen nicht stark unterscheiden, sind auch die Schätzer der Odds Ratios für alle Verfahren ähnlich. Für die Betrachtung der Effekte ist besonders wichtig, ob die Schätzer für ein Odds Ratio alle in die gleiche Richtung tendieren. Dies ist bei allen Schätzern der Fall, denn entweder sind alle Werte in einer Zeile der Tabelle D.16 kleiner oder größer als eins.

$\widehat{\kappa}_{11,21}$ ist in allen vier Fällen kleiner als eins. Das bedeutet, dass die Odds zum ersten Zeitpunkt in der Behandlungsgruppe, also in der Gruppe $i = 1$, kleiner und damit schlechter als in der Kontrollgruppe sind. Ein Wert von ungefähr 0,5 bedeutet, dass die Odds in der Behandlungsgruppe circa halb so groß wie in der Kontrollgruppe sind. Zum zweiten Zeitpunkt kehrt sich dieses Verhältniss mit $\widehat{\kappa}_{12,22} \approx 2$ um. Dann sind die Odds in der Behandlungsgruppe fast zweimal so groß wie in der Kontrollgruppe. Dies spricht für die Behandlung.

Äquivalent kann auch die zeitliche Verbesserung in den beiden Gruppen verglichen werden. In der Kontrollgruppe ist das geschätzte Odds Ratio $\widehat{\kappa}_{22,21}$ mit ungefähr 1,2 etwas

Tabelle D.16: Geschätzte Odds Ratios im parametrischen faktoriellen Modell bei der MS-Studie

	$\boldsymbol{R}_{ind}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{mde}$
$\widehat{\kappa}_{11,21}$	0,6447	0,5451	0,5259	0,6099
$\widehat{\kappa}_{22,21}$	1,2450	1,1658	1,2425	1,2581
$\widehat{\kappa}_{12,11}$	4,0217	3,9001	3,9785	3,9944
$\widehat{\kappa}_{12,22}$	2,0826	1,8236	1,6839	1,9363

größer als eins, was für eine leichte Verbesserung spricht. Eine deutliche Verbesserung kann man hingegen in der Behandlungsgruppe feststellen. Dort beträgt das geschätzte Odds Ratio $\widehat{\kappa}_{12,11}$ ungefähr vier. Die Verbesserung bezüglich der Odds Ratios ist in der Behandlungsgruppe mithin fast viermal so groß wie in der Kontrollgruppe.

Tabelle D.17: Teststatistiken und p-Werte im parametrischen faktoriellen Modell bei der MS-Studie

Hypothese		$\boldsymbol{R}_{ind}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{mde}$
$H_0: \beta_G = 0$	χ^2_1-Statistik	0,82	1,53	1,50	1,08
	p-Wert	0,3638	0,2155	0,2200	0,2976
$H_0: \beta_T = 0$	χ^2_1-Statistik	0,70	0,31	0,71	0,67
	p-Wert	0,4036	0,5760	0,3978	0,4126
$H_0: \beta_W = 0$	χ^2_1-Statistik	8,72	6,66	8,86	7,04
	p-Wert	0,0032	0,0098	0,0029	0,0080

Zuletzt werden die Testergebnisse für die Hypothesen $H_0: \beta_G = 0$, $H_0: \beta_T = 0$ und $H_0: \beta_W = 0$ vorgestellt. Für jede Wahl von $\boldsymbol{R}$ ergeben sich unterschiedliche Teststatistiken und p-Werte. Diese sind in der Tabelle D.17 zusammengefasst. Auch wenn die p-Werte gerade bei den Hypothesen über die Haupteffekte für die unterschiedlichen $\boldsymbol{R}$ deutlich differieren, ist die Testentscheidung doch immer die gleiche. Zum 5% Niveau werden die Hypothesen $H_0: \beta_G = 0$ und $H_0: \beta_T = 0$ nicht verworfen. Für die Wechselwirkung ergibt sich hingegen ein signifikanter Unterschied auf dem 5% Niveau, da alle p-Werte kleiner als 1% sind. Dabei ist allerdings zu beachten, dass nur bei $\boldsymbol{R}_{ind}$ und bei $\boldsymbol{R}_{cs}$ das empirische Niveau für den Test auf Wechselwirkung ungefähr beim nominellen Niveau liegt. Die anderen Working Correlation Matrizen brauchen bei so vielen Kategorien deutlich mehr

Stichprobenumfang.

D.4.2 Behandlung von Schlaflosigkeit

Bei der Auswertung der Insomnia-Studie ergeben sich keine Probleme mit der Skala, da alle Kategorien in allen Faktorstufenkombinationen häufig beobachtet wurden. Das Verhältnis der Anzahl der Kategorien zum Stichprobenumfang ist bei der Insomnia-Studie wesentlich günstiger als bei der MS-Studie. Daher ergaben sich bei der Insomnia-Studie auch keine Probleme bei der Berechnung der Schätzer. Das Iterationsverfahren konvergierte für jede Wahl der Working Correlation Matrix $\boldsymbol{R}$. Die sich ergebenden Schätzer sind in der Tabelle D.18 zusammengefasst. Dabei sind die Schätzer für die Parameter β_G, β_T und β_W wie bei der MS-Studie zu verwenden, da die gleiche Kodierung und Bezeichnung verwendet wird. Die Schätzer in der Tabelle D.18 unterscheiden sich für die unterschiedlichen $\boldsymbol{R}$ etwas. Dabei treten die größten Unterschiede bei den Schätzern für die Gruppe $\widehat{\beta}_G$ und die Wechselwirkung $\widehat{\beta}_W$ auf.

Tabelle D.18: Parameterschätzer der parametrischen faktoriellen Analyse bei der Insomnia-Studie

$\boldsymbol{R}$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\beta}_G$	$\widehat{\beta}_T$	$\widehat{\beta}_W$
$\boldsymbol{R}_{ind}$	-2,23	-0,92	0,37	0,00	-1,01	-0,78
$\boldsymbol{R}_{ar}$	-2,32	-1,01	0,29	-0,17	-1,12	-0,53
$\boldsymbol{R}_{cs}$	-2,27	-0,95	0,33	-0,02	-1,01	-0,78
$\boldsymbol{R}_{mde}$	-2,34	-1,03	0,25	-0,21	-1,15	-0,49
$\boldsymbol{R}_{un}$	-2,28	-0,98	0,32	-0,06	-1,07	-0,67

Aus den Schätzern können wie bei der MS-Studie die erwarteten Häufigkeiten für die verschiedenen Kategorien in den vier Faktorstufenkombinationen geschätzt werden. Diese können in der Tabelle D.19 mit den beobachteten Häufigkeiten verglichen werden. Es ist festzustellen, dass die Unterschiede in einigen Zellen nicht unwesentlich sind. Besonders zum zweiten Zeitpunkt in der Behandlungsgruppe ($i = 1$, $j = 2$) unterscheiden sich die beobachteten und aufgrund des Modells erwarteten Häufigkeiten in den Kategorien 1 und 2 deutlich. In den meisten Zellen geben jedoch die für das logistische Modell erwarteten Häufigkeiten die beobachteten Häufigkeiten gut wieder.

Um die Proportional Odds Annahme zu überprüfen, werden die beobachteten log Odds wie in Abbildung 4.8 graphisch dargestellt. Die dort eingezeichneten Pfeile sollten

Tabelle D.19: Beobachtete und auf Grund der Schätzer erwartete Häufigkeiten der Insomnia-Studie

Faktor		beobachtet	Kategorie			
Gruppe	Zeit	oder $\boldsymbol{R}$	1	2	3	4
$i = 1$	$j = 1$	beobachtet	12	20	40	47
		$\boldsymbol{R}_{ind}$	11,6	22,5	36,4	48,6
		$\boldsymbol{R}_{ar}$	12,3	23,4	37,1	46,1
		$\boldsymbol{R}_{cs}$	11,3	22,3	36,3	49,2
		$\boldsymbol{R}_{mde}$	12,7	23,7	36,6	46,0
		$\boldsymbol{R}_{un}$	11,7	22,2	36,7	48,3
$i = 1$	$j = 2$	beobachtet	40	49	19	11
		$\boldsymbol{R}_{ind}$	46,5	37,4	22,8	12,3
		$\boldsymbol{R}_{ar}$	44,8	37,5	23,9	12,9
		$\boldsymbol{R}_{cs}$	45,8	37,6	23,0	12,6
		$\boldsymbol{R}_{mde}$	45,2	37,3	23,5	13,0
		$\boldsymbol{R}_{un}$	45,8	37,0	23,5	12,7
$i = 2$	$j = 1$	beobachtet	14	20	35	51
		$\boldsymbol{R}_{ind}$	11,7	22,7	36,7	49,0
		$\boldsymbol{R}_{ar}$	10,7	21,3	36,7	51,3
		$\boldsymbol{R}_{cs}$	11,2	22,2	36,5	50,1
		$\boldsymbol{R}_{mde}$	10,6	20,9	35,9	52,5
		$\boldsymbol{R}_{un}$	11,2	21,6	36,8	50,5
$i = 2$	$j = 2$	beobachtet	31	29	35	25
		$\boldsymbol{R}_{ind}$	27,3	35,5	33,1	24,1
		$\boldsymbol{R}_{ar}$	27,7	35,6	33,2	23,5
		$\boldsymbol{R}_{cs}$	26,4	35,3	33,4	24,9
		$\boldsymbol{R}_{mde}$	28,0	35,4	32,8	23,8
		$\boldsymbol{R}_{un}$	27,7	35,1	33,4	23,9

innerhalb jeder Gruppe parallel sein, damit die Proportional Odds Annahme erfüllt ist. Wie man in der Abbildung D.5 sieht, ist dies bei der Insomnia-Studie tatsächlich der Fall. Sowohl die transparenten Pfeile, welche von den log Odds zum ersten auf den zweiten Zeitpunkt in der Behandlungsgruppe zeigen, als auch die nichttransparenten Pfeile, die zur Kontrollgruppe gehören, liegen jeweils weitgehend parallel zueinander.

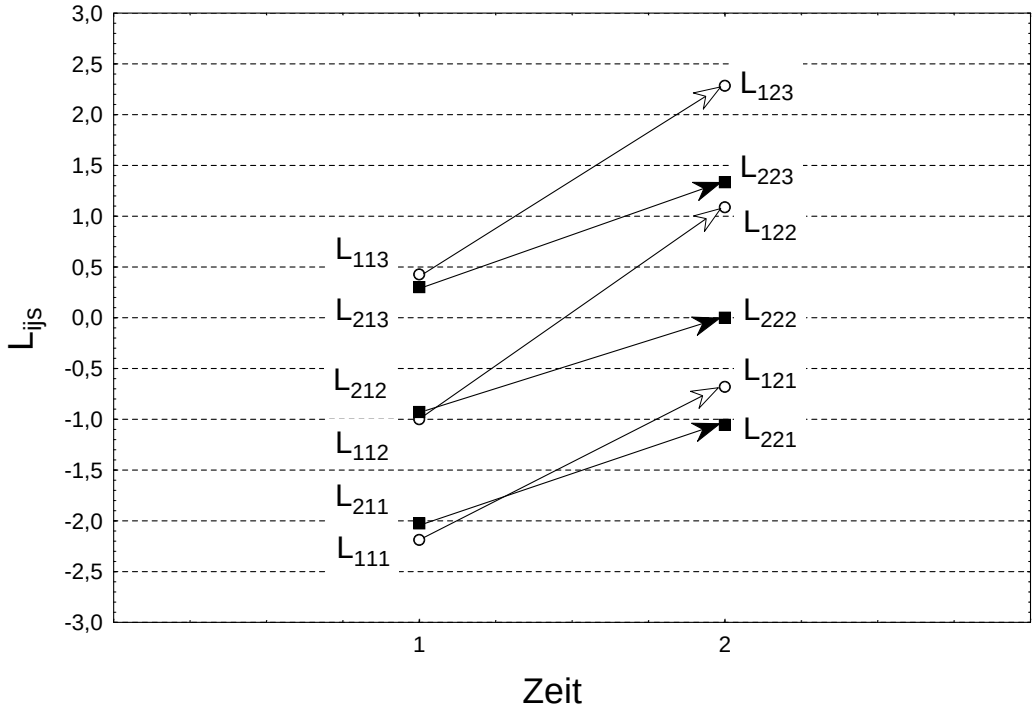

Abbildung D.5: Die Logarithmen der beobachteten Odds bei der Insomnia-Studie

Mit den geschätzten Parametern $\widehat{\beta}_G$, $\widehat{\beta}_T$ und $\widehat{\beta}_W$ des Proportional Odds Modells können die besser zu interpretierenden Odds Ratios $\kappa_{ij,i'j'}$ berechnet werden. Diese sind in der Tabelle D.20 zusammengefasst. Die geringen Unterschiede zwischen den Schätzern $\widehat{\beta}_G$, $\widehat{\beta}_T$ und $\widehat{\beta}_W$ für die unterschiedlichen Wahlen von $\boldsymbol{R}$ führen auf Ebene der Schätzer der Odds Ratios zum Teil zu deutlichen Unterschieden. So schwanken beispielsweise die Schätzer der Odds Ratios $\kappa_{12,11}$ zwischen 5 und 6. Dies liegt darin begründet, dass geringe Unterschiede bei den log Odds durch die Abbildung mit der Exponentialfunktion sich unter Umständen stark vergrößern. Dabei hängt es davon ab, wie groß die log Odds sind. Je größer sie vom Betrag her sind, desto mehr wirken sich geringe Unterschiede aus. Mit anderen Worten spielt es bei großen Effekten eine größere Rolle, welche Working Correlation Matrix verwendet wird, als bei kleinen Effekten.

Tabelle D.20: Geschätzte Odds Ratios im parametrischen faktoriellen Modell bei der Insomnia-Studie

	$\boldsymbol{R}_{ind}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{mde}$	$\boldsymbol{R}_{un}$
$\widehat{\kappa}_{11,21}$	0,9996	1,1805	1,0176	1,2365	1,0633
$\widehat{\kappa}_{22,21}$	2,7409	3,0670	2,7456	3,1509	2,9177
$\widehat{\kappa}_{12,11}$	5,9614	5,2179	5,9691	5,1233	5,7225
$\widehat{\kappa}_{12,22}$	2,1741	2,0083	2,2122	2,0105	2,0855

Insgesamt zeigen die Schätzer der Odds Ratios jedoch eine einheitliche Tendenz. So ist das Verhältnis der Odds von der Kontroll- zur Behandlungsgruppe zum ersten Zeitpunkt ungefähr eins. Dies spricht für eine gelungene Randomisierung, da keine der beiden Gruppen zu deutlich höheren Werten tendiert. Hingegen ist der Schätzer $\widehat{\kappa}_{12,22}$ des Odds Ratios zum zweiten Zeitpunkt für alle Ansätze von $\boldsymbol{R}$ in etwa zwei. Damit sind die geschätzten Odds in der Behandlungsgruppe ungefähr doppelt so groß wie in der Kontrollgruppe nach der Behandlung.

Beide Gruppen verbessern sich über die Zeit gesehen. Das Odds Ratio in der Kontrollgruppe wird durch $\widehat{\kappa}_{22,21} \approx 3$ und in der Behandlungsgruppe durch $\widehat{\kappa}_{12,11} \approx 5{,}5$ geschätzt. Damit ist die Verbesserung in der Behandlungsgruppe ungefähr doppelt so groß wie in der Kontrollgruppe. Da diese beiden Schätzer jedoch relativ groß sind, sind auch die Unterschiede bei den verschiedenen Ansätzen für $\boldsymbol{R}$ recht groß. Eine genaue Angabe der Quantität der Verbesserung ist damit nicht möglich.

Tabelle D.21: Teststatistiken und p-Werte im parametrischen faktoriellen Modell bei der Insomnia-Studie

Hypothese		$\boldsymbol{R}_{ind}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{mde}$	$\boldsymbol{R}_{un}$
$H_0: \beta_G = 0$	χ^2_1-Statistik	0,00	0,46	0,01	0,71	0,07
	p-Wert	0,9987	0,4962	0,9422	0,3980	0,7962
$H_0: \beta_T = 0$	χ^2_1-Statistik	34,39	40,12	34,35	40,65	41,62
	p-Wert	$< 0,0001$	$< 0,0001$	$< 0,0001$	$< 0,0001$	$< 0,0001$
$H_0: \beta_W = 0$	χ^2_1-Statistik	9,46	4,29	9,40	3,08	7,78
	p-Wert	0,0021	0,0383	0,0022	0,0795	0,0053

Nach der Interpretation der Schätzer werden nun die Hypothesen bezüglich der zu

schätzenden Parameter zum 5% Niveau getestet. Die Statistiken, welche χ_1^2-verteilt sind, sind zusammen mit den p-Werten in der Tabelle D.21 aufgelistet. Für den Test auf keinen Gruppeneffekt $H_0 : \beta_G = 0$ ergeben sich immer p-Werte, die deutlich größer als das Niveau sind. Hingegen wird die Hypothese auf keinen Zeiteffekt $H_0 : \beta_T = 0$ immer verworfen. Bei der wichtigsten Hypothese $H_0 : \beta_W = 0$ gibt es hingegen Unterschiede in den Testentscheidungen. Bei $\boldsymbol{R}_{mde}$ kann die Hypothese im Gegensatz zu den anderen Working Correlation Matrizen nicht verworfen werden. Insgesamt unterscheiden sich die p-Werte erheblich, wenn man $\boldsymbol{R}$ unterschiedlich wählt. Insgesamt wird man sich jedoch gegen die Hypothese entscheiden, da sie bei $\boldsymbol{R}_{ind}$ verworfen wird und dort das Niveau am besten eingehalten wird.

D.5 Parametrische Kovarianzanalyse

D.5.1 Neurologische Beeinträchtigung bei MS

Für die parametrische Kovarianzanalyse der MS-Studie ergeben sich ähnliche Probleme mit der Skala wie bei der parametrischen faktoriellen Analyse. Für die Zielvariable werden die Kategorien 1 und 10 und für die Kovariable zusätzlich die Kategorien 0 und 2 nicht beobachtet. Damit muss in der Zielvariablen mit einer reduzierten Skala mit neun Kategorien gearbeitet werden. Für die Kovariable reduziert sich die Skala um weitere zwei Kategorien. Damit erhalten wir für die Behandlungsgruppe die Modellgleichung

$$\text{logit}(\text{P}(X_{1k2} = s | X_{1k1} = s')) \quad = \quad \alpha_{2s} + \beta_{kov}\alpha_{1s'}.$$

Dabei bezeichnet $s = 0, 2, 3, 4, 5, 6, 7, 8, 9$ die Kategorien der Skala der Zielvariablen und $s' = 3, 4, 5, 6, 7, 8, 9$ die der Kovariablen. Für die Kontrollgruppe fällt der Parameter β_{kov} weg. Damit ergibt sich in der Kontrollgruppe die Modellgleichung

$$\text{logit}(\text{P}(X_{2k2} = s | X_{2k1} = s')) \quad = \quad \alpha_{2s} + \alpha_{1s'}.$$

Mit Hilfe der Prozedur LOGISTIC von SAS können die Parameterschätzer berechnet werden. Für die α_{js} sind die Schätzer in der Tabelle D.22 zusammengefasst. Es ist zu erkennen, dass die Schätzer der Regressionsparameter α_{1s} nahezu der Größe nach sortiert sind. Dies spiegelt wider, dass die Zielvariable zu kleineren Werten tendiert, wenn die Kovariable kleiner wird. Dieser Zusammenhang ist plausibel, denn gute Anfangswerte sollten auf gute Endwerte hindeuten.

Tabelle D.22: Schätzer der α_{js} in der MS-Studie

Kategorie (s, s')	0	2	3	4	5	6	7	8
Zielvariable $(j = 2)$	-10,20	-9,28	-7,15	-6,66	-5,66	-2,06	-0,16	0,81
Kovariable $(j = 1)$			7,87	6,32	6,87	3,2082	1,78	0,23

Der interessierende Parameter β_{pkov}, über den die Hypothese gestellt wird, wird durch $\widehat{\beta}_{pkov} = -1{,}6205$ geschätzt. Dies liefert ein geschätztes odds ratio von $\widehat{\kappa} = 5{,}056$. Damit ist die Chance, in der Behandlungsgruppe eine kleinere Kategorie als eine vorgegebene Kategorie zu beobachten, fünfmal so groß wie in der Kontrollgruppe, wenn von der gleichen Kovariablen ausgegangen wird. Der Wald-Test der Hypothese $H_0 : \beta_{pkov} = 0$ liefert mit der Statistik 7,5376 einen p-Wert von 0,006. Mithin wird die Hypothese auf keine Wechselwirkung bei einem Niveau von 5% verworfen.

Es ist aber zu beachten, dass für viele Werte der Kovariablen nur wenige Werte der Zielvariablen vorliegen. Des Weiteren beschränken sich die Zielvariablen bei vorgegebener Kategorie der Kovariablen auf einige wenige Kategorien, was zu sehr vielen Nullhäufigkeiten führt. Diese sind problematisch für das Modell. Damit ist auch bei der Interpretation der Parameterschätzer und der Teststatistik Vorsicht geboten. Außerdem haben die Simulationen gezeigt, dass der Wald-Test bei einem so geringen Stichprobenumfang und so vielen Kategorien das Niveau nicht einhält. Damit sind Zweifel an der Testentscheidung angebracht.

D.5.2 Behandlung von Schlaflosigkeit

Die Probleme bezüglich der Kategorien, wie sie bei der MS-Studie auftreten, ergeben sich bei der Insomnia-Studie nicht, da mehr Versuchseinheiten auf weniger Kategorien verteilt werden. Mit den vier Kategorien 1 bis 4, sowohl für die Ziel- als auch für die Kovariable, ergibt sich in der Behandlungsgruppe die Modellgleichung

$$\text{logit}(\text{P}(X_{1k2} = s | X_{1k1} = s')) \quad = \quad \alpha_{2s} + \beta_{kov} + \alpha_{1s'}$$

mit $s, s' = 1, \ldots, 3$. Für die Kontrollgruppe verwenden wir hingegen

$$\text{logit}(\text{P}(X_{1k2} = s | X_{1k1} = s')) \quad = \quad \alpha_{2s} + \alpha_{1s'}.$$

Die Schätzer der Cutpoints beziehungsweise der Regressionsparameter finden sich in der Tabelle D.23. Wieder sind die Schätzer der Regressionsparameter nahezu der Größe nach

geordnet, was die ordinale Struktur der Kovariablen widerspiegelt. Nur von der ersten zur zweiten Kategorie der Kovariablen scheinen sich die Werte nicht zu verschlechtern, sondern zu verbessern.

Tabelle D.23: Schätzer der α_{js} in der Insomnia-Studie

Kategorie (s, s')	1	2	3
Zielvariable $(j = 2)$	-2,1570	-0,4260	1,0533
Kovariable $(j = 1)$	2,3068	2,6726	1,1524

Der Einfluss der Gruppe ist im vorliegenden Modell durch den Parameter β_{kov} gegeben. Dieser wird hier mit $\widehat{\beta}_{kov} = -0{,}4553$ geschätzt, was einem odds ratio von 2,486 entspricht. Die odds in der Behandlungsgruppe sind also ungefähr 2,5 mal so hoch wie in der Kontrollgruppe, wenn von gleichen Kovariablen ausgegangen wird. Der Wald-Test bezüglich der Hypothese $H_0 : \beta_{kov} = 0$ liefert bei einer Statistik von 13,5571 einen p-Wert von 0,0002. Damit wird die Hypothese auf dem 5%-Niveau verworfen. Die Simulationen haben gezeigt, dass dieser Test bei so wenigen Kategorien und so vielen Versuchseinheiten das Niveau einhält.

Da der Test sehr von der Validität des Modells abhängt, sollte dies überprüft werden. Eine Möglichkeit dazu ist der Vergleich der beobachteten mit den aufgrund der Schätzer und des Modells erwarteten Häufigkeiten. Diese Häufigkeiten sind in der Tabelle D.24 zusammengestellt. Dort fällt auf, dass die beobachteten und die erwarteten Häufigkeiten in vielen Zellen übereinstimmen. Hat die Kovariable jedoch den Wert 3, so ergeben sich deutliche Differenzen zwischen den beiden Häufigkeiten in beiden Gruppen. In der Kontrollgruppe weichen darüberhinaus auch bei einem Baseline-Wert von 4 die erwarteten von den beobachteten Häufigkeiten ab. Die meisten Versuchseinheiten haben jedoch den Wert 3 oder 4 als Baseline-Wert. Damit sollte gerade in diesen Fällen das Modell gut passen. Die vorliegenden Differenzen zwischen den beiden Häufigkeiten erzeugen mithin Zweifel an der Gültigkeit des Modells und der darauf basierenden Schlussfolgerungen.

Tabelle D.24: Erwartete und beobachtete Häufigkeiten bei der Insomnia-Studie

		Kovariable							
		1		2		3		4	
	Zielvariable	erw.	beob.	erw.	beob.	erw.	beob.	erw.	beob.
Behandlung $i = 1$	1	7,8	7	14,5	11	14,6	13	7,2	9
	2	3,2	4	4,2	5	16,0	23	16,6	17
	3	0,8	1	1,0	2	6,8	3	14,6	13
	4	0,3	0	0,3	2	2,6	1	8,5	8
Kontrolle $i = 2$	1	5,9	7	10,3	14	6,6	6	3,5	4
	2	5,3	4	6,8	5	13,3	9	11,5	11
	3	2,0	2	2,1	1	10,0	18	18,0	14
	4	0,7	1	0,7	0	5,2	2	18,1	22

Literaturverzeichnis

(1999). *SAS/STAT User's Guide, Version 8*. SAS Institute Inc., Cary, NC.

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

Akritas, M. G. and Arnold, S. F. (1994). Fully Nonparametric Hypotheses for Factorial Designs I: Multivariate Repeated Measures Designs. *Journal of the American Statistical Association*, 89:336–343.

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

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26:1323–1333.

Anderson, J. A. (1984). Regression and Ordered Categorical Variables. *Journal of the Royal Statistical Society*, B 46:1–30.

Armitage, P. (1955). Test for linear trend in proportions and frequencies. *Biometrics*, 11:375–386.

Armstrong, B. and Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129:191–204.

Bajorski, P. and Petkau, J. (1999). Nonparametric Two-Sample Comparisons of Changes on Ordinal Responses. *Journal of the American Statistical Association*, 94:970–978.

Bandelow, B. (1997). *Panik- und Agoraphobieskala (PAS)*. Hofgrefe, Göttingen.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics*, volume 1. Prentice Hall, Upper Saddle River, 2 edition.

Box, G. E. P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *The Annals of Mathematical Statistics*, 25:290–302.

Brunner, E. and Langer, F. (1999). *Nichtparametrische Analyse longitudinaler Daten.* R. Oldenbourg, München.

Clayton, D. (1992). Repeated ordinal measurements: A generalized estimating equations approach. Technical report, Medical Research Council Biostatistics Unit, Cambridge, U.K.

Cliff, N. (1993). Dominance Statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114:494–509.

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ^2-Test. *Biometrics*, 10:417–451.

Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41:557–566.

Domhof, S. (1999). Rangverfahren mit unbeschränkten Scorefunktionen in faktoriellen Versuchsplänen. Diplomarbeit, Institut für Mathematische Stochastik, Universität Göttingen, Göttingen.

Domhof, S. (2001). *Nichtparametrische relative Effekte.* Dissertation, Mathematisch-naturwissenschaftliche Fakultäten der Universität Göttingen.

Edwards, A. L. and Thurstone, L. L. (1952). An Internal Consistency Check for Scale Values Determined by the Method of Successive Intervals. *Psychometrika*, 17:169–180.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer Series in Statistics. Springer-Verlag, New York.

Francom, S. F., Chuang-Stein, C., and Landis, J. R. (1989). A log-linear model for ordinal data to characterize differential change among treatments. *Statistics in Medicine*, 8:571–582.

Gerthsen, C. und Vogel, H. (1993). *Physik.* Springer, Berlin, 17. Auflage.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13:1665–1677.

Greenwood, C. and Farewell, V. (1988). A comparison of regression models for ordinal data in an analysis fo transplant-kidney function. *Canadian Journal of Statistics*, 16:325–335.

Guggenmoos-Holzmann, I. und Wernecke, K.-D. (1995). *Medizinische Statistik*. Blackwell, Berlin, Wien.

Hartung, J. (1995). *Statistik: Lehr- und Handbuch der angewandten Statistik*. Oldenbourg, München, 10. Auflage.

Kenward, M. G., Lesaffre, D., and Molenberghs, G. (1994). An Application of Maximum Likelihood and Generalized Estimating Equations to the Analysis fo Ordinal Data from a Longitudinal Study with Cases Missing at Random. *Biometrics*, 50:945–953.

Khatri, B. O., McQuillen, M. P., Harrington, G. J., Schmoll, D., and Hoffman, R. G. (1985). Plasmapheresis in Progressive Multiple Sclerosis. *Neurology*, 35:614.

Köhler, K. (2001). Unverzerrte Schätzer für die relativen Effekte im nichtparametrischen gemischten Modell. Diplomarbeit, Institut für Mathematische Stochastik, Universität Göttingen, Göttingen.

Kirk, R. E. (1982). *Experimental Designs*. Brooks/Cole, Pacific Grove, second edition.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement*, volume I. Academic Press, New York.

Kreienbrock, L. und Schach, S. (1997). *Epidemiologische Methoden*. Gustav Fischer, Stuttgart, 2. Auflage.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621.

Kulle, B. (1999). Nichtparametrisches Behrens-Fisher Problem im Mehr-Stichprobenfall. Diplomarbeit, Institut für Mathematische Stochastik, Universität Göttingen, Göttingen.

Kurtzke, J. F. (1961). On the Evaluation of Disability in Multiple Sclerosis. *Neurology*, 11:686–694.

Langer, F. (1998). *Berücksichtigung von Kovariablen im nichtparametrischen gemischten Modell*. Dissertation, Institut für Mathematische Stochastik, Universität Göttingen, Göttingen.

Lee, J. (1992). Cumulative logit modelling for ordinal response variables: Applications to biomedical research. *Computer Applications in Biosciences*, 8:555–562.

Lewis, J. A. (1999). Statistical Principles for Clinical Trials (ICH E9) an Introductory Note on an International Guideline. *Statistics in Medicine*, 18:1903–1904.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Analysis using Generalized Linear Models. *Biometrika*, 73:13–22.

Loève, M. (1977). *Probability Theory I.* Springer, Berlin, 4th edition.

Luce, R. D., Krantz, D. H., Suppes, P., and Tversky, A. (1990). *Foundations of measurement*, volume III. Academic Press, New York.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 17:50–60.

McCullagh, P. (1980). Regressions models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, B 42:109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall, London, second edition.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the American Statistical Association*, 135:370–384.

Oelerich, A. (1998). Teststatistiken zur Analyse ordinaler Daten bei kleinen Stichprobenumfängen. Diplomarbeit, Institut für Mathematische Stochastik, Universität Göttingen.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2:110–114.

Scott, S. C., Goldberg, M. S., and Mayo, N. E. (1997). Statistical assessment of ordinal outcomes in comparative studies. *Journal of Clinical Epidemiology*, 50:45–55.

Shao, J. (1999). *Mathematical Statistics.* Springer Texts in Statistics. Springer-Verlag, New York.

Siemer, A. (1999). Die Berücksichtigung von heterogen verteilten Kovariablen in einem nichtparametrischen Modell. Diplomarbeit, Institut für Mathematische Stochastik, Universität Göttingen, Göttingen.

Smith, H. F. (1936). The Problem of Comparing the Results of Two Experiments with Unequal Errors. *Journal of the Council for Scientific and Industrial Research (Australia)*, 9:211–212.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In Stevens, S. S., editor, *Handbook fo experimental psychology*, pages 1–49. Wiley, New York.

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1989). *Foundations of measurement*, volume II. Academic Press, New York.

van Elteren, P. (1960). On the combination of independent two-sample tests of wilcoxon. *Bulletin of the International Statistical Institute*, 37:351–161.

Vargha, A. and Delaney, H. D. (1998). The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2):170–192.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, 3:119–122.

Yarnold, J. K. (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, 65:864–886.

Ziegler, A., Kastner, C., Grömping, U. und Blettner, M. (1996). Die Generalized Estimating Equations: Herleitung und Anwendung. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, 27(2):69–91.

Lebenslauf

Am 21. November 1973 wurde ich in Göttingen als erstes Kind von Brigitte Siemer, geborene Bothe, und Konrad Siemer geboren.

Die ersten vier Jahre der Schullaufbahn verbrachte ich an der Grundschule Ankum bis 1984. Danach wechselte ich auf das Gymnasium Leoninum in Handrup, welches ich bis 1991 besuchte. Die Allgemeine Hochschulreife erwarb ich daraufhin 1993 nach zweijährigem Aufenthalt am Gymnasium Bersenbrück. Den danach folgenden zwölfmonatigen Grundwehrdienst leistete ich bei den Fallschirmjägern in Oldenburg bis Ende Juni 1994.

Zum Wintersemester 1994/95 nahm ich das Studium der Mathematik mit Nebenfach Physik an der Georg-August-Universität Göttingen auf. Nach vier Semestern legte ich im Oktober 1996 die Diplom-Vorprüfung ab. Während des Wintersemesters 1997/98 war ich als studentische Hilfskraft am Mathematischen Institut beschäftigt und betreute eine Übungsgruppe der Vorlesung „Differential- und Integralrechnung I“. Danach nahm ich meine Tätigkeit als studentische Hilfskraft in der Abteilung Medizinische Statistik auf, die ich bis zum Ende meines Studiums fortführte. Dort arbeitete ich hauptsächlich im Rahmen des DFG-Forschungsprojektes „Test- und Schätzverfahren für ordinale Daten“. Während dieser Zeit fertigte ich außerdem unter Anleitungen von Prof. Dr. Edgar Brunner meine Diplomarbeit mit dem Titel „Die Berücksichtigung von heterogen verteilten Kovariablen in einem nichtparametrischen Modell“ an. Mit dieser Arbeit erwarb ich im Oktober 1999 das Diplom in Mathematik.

Seit dem November 1999 bin ich als wissenschaftlicher Mitarbeiter im Rahmen des oben genannten DFG-Forschungsprojektes an der Abteilung Medizinische Statistik angestellt. Des Weiteren arbeitete ich seither an der vorliegenden Arbeit. Diese betreute der Leiter der Abteilung, Prof. Dr. Edgar Brunner.