Trendtests für geordnete kategoriale Daten bei sehr kleinen Fallzahlen

Vom Fachbereich Gartenbau
der Universität Hannover
zur Erlangung
des Grades eines

Doktors der Gartenbauwissenschaften
- Dr. rer. hort. -

genehmigte

Dissertation

von

Dipl.-Math. **Dirk Seidel**
geboren am 08. 04. 1967 in Rostock

2001

Referent: Prof. Dr. L. A. Hothorn (Hannover)
Koreferent: Dr. H. Bleiholder (Limburgerhof)
Tag der Promotion: 15.12.2000

Kurzfassung

In der Landwirtschaft und im Gartenbau werden kategoriale Daten oft als *Bonituren* (lat. Abschätzung, Einstufung) bezeichnet. Die Skala, auf der die Werte geschätzt werden, ist zum Teil auf sehr wenige Kategorien eingeschränkt. Zwischen diesen einzelnen Kategorien besteht jedoch eine Ordnung, die einen Vergleich der einzelnen Kategorien untereinander zuläßt. Bonituren können daher als geordnete kategoriale Daten aufgefaßt werden. Allgemeine Auswertungsmethoden für geordnete kategoriale Daten sind somit sowohl für landwirtschaftliche als auch für gartenbauliche Versuche von hohem Interesse.

Gerade bei Feldversuchen in der Landwirtschaft und im Gartenbau, aber auch bei pharmakologischen Versuchen in der Medizin werden häufig Experimente mit sehr kleinen Fallzahlen durchgeführt. Zum Teil beschränkt sich der Umfang der Stichproben auf drei, vier oder fünf. Da die Güte eines statistischen Tests vom Stichprobenumfang abhängt, ist es sinnvoll, zusätzliche Informationen zu nutzen, die zu einer höheren Güte führen. Werden z. B. zwei oder mehrere Dosen miteinander verglichen, so ist oft intuitiv mit einer größeren Wirkung der höheren Dosen zu rechnen. Dieses Vorwissen kann genutzt werden, um einseitige (gerichtete) Hypothesen aufzustellen und effiziente Tests zu wählen.

Der statistische Nachweis von Wirkungsunterschieden auf der Basis geordneter kategorialer Daten ist mittels vieler, sehr verschiedener Verfahren möglich. In Abhängigkeit davon, welches Modell den zu analysierenden Daten unterstellt wird, können unterschiedliche Tests zum Prüfen aufgestellter Hypothesen genutzt werden. Die Auswahl eines Modells und eines Tests erfolgt häufig willkürlich. Vor allem bei sehr kleinen Fallzahlen ist jedoch große Vorsicht geboten. Zum einen können die Modellannahmen kaum realistisch überprüft werden, zum anderen treten bei sehr kleinen Fallzahlen zusätzliche technische Probleme auf. Technische Schwierigkeiten bestehen z. B. in der nicht gesicherten Existenz von Maximum-Likelihood-Schätzern, in der extremen Konservativität oder Antikonservativität von Tests oder in Konvergenzproblemen numerischer Verfahren.

In der vorliegenden Arbeit werden Tests für zwei und mehr unabhängige Stichproben vorgestellt, die zum Prüfen einseitiger bzw. gerichteter Alternativen geeignet sind. Aus der Vielzahl der bekannten Verfahren wurden diejenigen ausgewählt, die zum einen für zwei und mehr Stichproben definierbar sind und zum anderen einen guten Kompromiß hinsichtlich Nutzerakzeptanz und einfacher Anwendbarkeit darstellen. Einen Schwerpunkt bilden unter-

schiedliche Verteilungsapproximationen für multiple Kontraste und Teststatistiken, die auf isotonen Schätzern basieren. Es wurden sowohl Permutationstests und Bootstraptests als auch Tests, die auf einer t- bzw. Normalverteilungsapproximation basieren studiert. Anhand von Simulationen wurden die Tests hinsichtlich ihrer Eignung bei der Auswertung geordneter kategorialer Daten unter der Bedingung untersucht, daß nur sehr kleine Stichproben zur Verfügung stehen. Bei diesen Simulationen zeigte sich, daß viele Verfahren bei diesen kleinen Stichprobenumfängen nicht geeignet sind. Gerade das Schätzen der Varianz ist mit großen Problemen verbunden. Als besonders geeignet erwiesen sich einfache Teststatistiken in Verbindung mit Mid-p-Werten. Diese werden daher für den Einsatz in der Praxis empfohlen.

Oftmals werden auf Basis der Daten eines Experiments mehrere Tests durchgeführt. Für bestimmte Aussagen ist es dann notwendig, die Fehlerwahrscheinlichkeit 1. Art zu adjustieren. Da für ein kleines Signifikanzniveau ein hoher Stichprobenumfang erforderlich ist, sollten Informationen, die zur einer günstigeren Adjustierung führen, genutzt werden. Neben den kleinen Fallzahlen wirkt sich dann nicht auch noch ein kleines Signifikanzniveau für den einzelnen Test limitierend aus. Für die Bestimmung ausgewählter Dosen wurden daher Prozeduren beschrieben, die unter der Anwendung der empfohlenen Tests auch bei diesen kleinen Fallzahlen mit einer hohen Güte verbunden sind.

Für die Durchführung der Tests wurden C-Programme erstellt. Anhand der beschriebenen Macros wird gezeigt, wie sich diese Programme auch in Verbindung mit anderen Programmen, wie z. B. SAS und Excel, nutzen lassen.
Schlagworte: geordnete kategoriale Daten, Trendtest, kleine Fallzahlen

Abstract

In the agriculture and horticulture categorical data are frequently called rating scale data. The scale, on which the rates are estimated, has sometimes only very few categories. But between these categories one assumes an order, which enables a comparison between the different categories. Hence, rating scale data can be regarded as ordered categorical data. General data analysis methodologies are therefore of interest for both agricultural and horticultural experiments.

Experiments with very small sample sizes are very common in the agriculture and horticulture research areas, but they occur also in medical and pharmacological studies. Sometimes the groups sample sizes is restricted to three, four or five observations. Since the power of a statistical test depends on the sample size, it is mandatory to include possible additional information, which increase the power. If, for example, two doses are compared with each other, it is natural to assume that the higher dose leads to a higher response. This preliminary knowledge can be used to construct one-sided (ordered) hypotheses and chose efficient tests.

The statistical assessment of differences in response based on ordered categorical data can be performed with several, very different procedures. Depending on the model which is assumed to hold for the underlying data, different statistics can be chosen to test the constructed hypotheses. The choice of a model and of a test is rather arbitrary. But in particular for small sample sizes caution is needed. On the one hand it is difficult to check the model assumptions realistically. On the other hand, small sample sizes may generate quickly technical problems. Such technical problems are, for example, the existence of maximum likelihood estimators, the extreme conservativeness or liberality of some tests, or the convergence problems of numerical approaches.

The present thesis introduces adequate tests for two or more independent sample groups for testing one-sided or ordered alternatives. Form the many tests available we have chosen only some particular tests statistics. On the one hand, they should be defined for two and more sample groups. On the other hand, they should give a good compromise between easy application and acceptance of the practitioner. One main topic of the thesis are different distribution approximations for multiple contrasts and test statistics based on isotonic estimates. Permutation tests, bootstrap tests and tests based on a t- or normal approximation

are investigated in detail. With the help of simulations the tests are studied with respect to there suitability of analyzing ordered categorical data under the restriction of very small sample sizes. These simulations show that many procedures are not adequate for small sample sizes. In particular the estimation of the variance leads large problems. Simple test statistics in combination with mid-p values turned out to give the best results. Therefore they are recommended for the use in practice.

Usually several tests are performed on the basis of the data of an experiment. For particular assessments it is therefore important to adjust the type I error. Since a smaller significance level requires higher sample sizes, one should use any information which leads to a less stringent adjustment. A small significance level is one of the limiting factors beside the small sample sizes. For the estimation of certain doses, procedures are described, which still perform good in terms of power, even under the severe sample size restrictions.

For the conduction of the tests C programs have been written. With the use of the accompanied macros it is shown, how that programs can be used in connection with other software packages, such as SAS or Excel.

Keywords: ordered categorical data, trendtest, small samples

Inhaltsverzeichnis

1. Einleitung 1

1.1. Dosis-Wirkungs-Versuche 1

1.2. Geordnete kategoriale Daten 11

1.3. Gliederung der Arbeit 19

2. Permutationstests 21

2.1. Exakte bedingte Permutationstests 21

2.2. Exakte unbedingte Permutationstests 34

2.3. Modifizierte Permutationstests 38

2.4. Approximative Permutationstests 39

3. Bootstraptests 41

3.1. Parametrischer und nichtparametrischer Bootstraptest 41

3.2. Double-Bootstraptest 52

4. Parametrische Tests 54

4.1. Trendtests für normalverteilte Daten auf der Basis von isotonen Schätzern 54

4.2. Parametrische Kontrasttests für normalverteilte Daten 69

4.3. Likelihood-Quotienten-Test für Multinomialverteilungen 85

4.4. GSK-Methode 87

4.5. Scorestatistiken 88

4.6. Geordnete kategoriale Regressionsmodelle 93

5. Nichtparametrische Verfahren 96

5.1. Nichtparametrische Tests auf der Basis von isotonen Schätzern 97

5.2. Nichtparametrische Kontraststatistiken 105

5.3. Adaptive Tests 108

6. Simulationen 113

6.1. Allgemeine Aussagen 113

6.2. Ungeeignete Verfahren 117

6.3. Geeignete Tests 130

6.3.1. Zweistichprobentests 130

6.3.1.1. Güte unter der Nullhypothese 132

6.3.1.2. Güte unter der Alternativhypothese 140

6.3.2. Drei- und Vierstichprobentests 149

6.3.2.1. Güte unter der Nullhypothese 156

6.3.2.2. Güte unter der Alternativhypothese 164
7. Multiples Testen 181
8. Anwendungsbeispiel 189
9. Zusammenfassung 195
Literaturverzeichnis 199
Anhang A Exakte Mid-p-Tests 208
Anhang B SAS/IML-Routinen 213

Tabellenverzeichnis

1.1 Fungizidversuch mit 3 Dosen (D_1, D_2, D_3), einem Standard (D_S) und einer Kontrolle (D_0) 2

1.2 Minimal nachweisbarer Unterschied des einseitigen t-Tests ($\alpha = 0{,}05$, $\beta = 0{,}2$) in Abhängigkeit von der Varianz und dem Stichprobenumfang 4

1.3 Symmetrische diskrete Verteilungen mit Varianz $\sigma^2 = 1{,}2$ ($\pi_s = P(X = s)$, $s = 1,\ldots,r$) 4

1.4 Güte des einseitigen t-Tests in Abhängigkeit vom Stichprobenumfang, dem Träger der Verteilung ($r = 3,4,5,6,\infty$) und der Differenz der Erwartungswerte $\delta = \mu_2 - \mu_1$ 5

1.5 Beispiele für die Intervallbreite und die maximale Abweichung (in %) in Abhängigkeit von wahrer Güte, Simulationsanzahl und Irrtumswahrscheinlichkeit 20

2.1 Beispielversuch ohne Bindungen 22

2.2 Daten nach Permutieren der Originaldaten aus Tabelle 2.1 23

2.3 Beispielversuch mit vielen Bindungen 24

2.4 Originalkontingenztafel für die Beispieldaten aus Tabelle 2.3 25

2.5 Empfehlung für M beim approximativen Permutationstest in Abhängigkeit von p 40

3.1 Beispiele für die Anzahl der möglichen Resamplingaufteilungen bei $n = n_1 = n_2$ 45

4.1 Wahrscheinlichkeit für einen Varianzschätzer mit dem Wert Null (linker Eintrag) und Wahrscheinlichkeit, daß mindestens für eine von 10.000 Resamplingstichproben ein Varianzschätzer mit dem Wert Null auftritt 69

4.2 Beispiele für optimale Kontrastvektoren: $k = 3$ und gleiche Stichprobenumfänge 71

4.3 Beispiele für optimale Kontrastvektoren: $k = 3$ und $\boldsymbol{n} = (10,5,5,5)'$ 72

4.4 Beispiele für optimale Kontrastvektoren: $k = 3$ und $\boldsymbol{n} = (5,10,10,5)'$ 72

4.5 Beispiel einer doppelt geordeneten Kontingenztafel mit zugeordneten Scores 89

6.1 Güte der Kontrasttests PK, HK, RHK und TK ohne Varianzschätzer auf der Basis von: Rängen, Bootstrapverteilung, binomialverteilten Zufallszahlen, konkaven Profilen ($m = 4$ und $\alpha = 0{,}05$) 117

6.2 Güte des TK-Tests für binomialverteilte Zufallszahlen bei unterschiedlichen Verteilungsapproximationen und den Parametern: $\alpha = 0{,}05$, $k = 2$, $m = 3$, $n = 3$ 120

6.3 Güte des TK-Tests für binomialverteilte Zufallszahlen bei unterschiedlichen Verteilungsapproximationen und den Parametern: $\alpha = 0{,}05$, $k = 2$, $m = 3$, $n = 5$ 120

6.4 Bezeichner für die im Abschnitt 6.3.1 beschriebenen Zweistichprobentests 131
6.5 Vergleich „exakter“ Bootstraptest versus Permutationstests ($\alpha = 0{,}05$, $k = 2$, $n = 3$, $\theta = 0$ bzw. $\theta = 5$, $\omega = 0$, konkave Profile, Verteilungstyp = GV) 150
6.6 Vergleich Bootstraptest (TknpbomoV) versus Double-Bootstraptest (TknpdbovoV) auf der Basis von Rängen ($k = 3$, konkave Profile) 150
6.7 Bezeichner für die im Abschnitt 6.3.2 beschriebenen k-Stichprobentests 151
6.8 Chacko-Test in Abhängigkeit von r und n ($\alpha = 0{,}05, k = 3, \theta = 15, \omega = 0{,}5$, konkave Profile, Verteilungstyp = LSV) 167
6.9 Vergleich finiter und infiniter Verteilungen anhand der Batholomew-Statistik und der Chacko- Statistik ($\alpha = 0{,}05, k = 2, n = 3, \omega = 0{,}5$, konkave Profile, Verteilungstyp = RSV) 168
6.10 Vergleich finiter und infiniter Verteilungen anhand der Batholomew- Statistik und der Chacko- Statistik ($\alpha = 0{,}05, k = 3, n = 3, \omega = 0{,}5$, konkave Profile, Verteilungstyp = RSV) 169
6.11 Vergleich finiter Verteilungen anhand der Batholomew-Statistik und der Chacko- Statistik ($\alpha = 0{,}01, k = 2, n = 3, \omega = 0{,}5$, konkave Profile, Verteilungstyp = RSV) 170
7.1 Beispieldaten aus einem Fungizidversuch der BASF AG 1995 189
7.2 p-Werte für die *MED*-Bestimmung am 1. Zeitpunkt 190
7.3 p-Werte für die *HEDS*-Bestimmung am 1. Zeitpunkt 191
7.4 p-Werte für die *MÄD*-Bestimmung am 1. Zeitpunkt 191
7.5 p-Werte für die *MED*-Bestimmung am 2. Zeitpunkt 191
7.6 p-Werte für die *HEDS*-Bestimmung am 2. Zeitpunkt 192
7.7 p-Werte für die *MÄD*-Bestimmung am 2. Zeitpunkt 192
7.8 p-Werte für die *MED*-Bestimmung am 3. Zeitpunkt 193
7.9 p-Werte für die *HEDS*-Bestimmung am 3. Zeitpunkt 193
7.10 p-Werte für die *MÄD*-Bestimmung am 3. Zeitpunkt 193

Abbildungsverzeichnis

1.1 Güte des einseitigen t-Tests in Abhängigkeit von der Differenz der Erwartungswerte (δ) und dem Stichprobenumfang ($n = n_1 = n_2$) (Varianz $\sigma^2 = 2$) 3

2.1 Ausschnitt aus einem auf den Daten der Tabelle 2.4 beruhenden Netzwerks 28

6.1 Vergleich Bootstrapverteilung versus Permutationsverteilung anhand des *TK*-Tests ($\alpha = 0{,}05$, $k = 3$, $n = 4$, diskretisierte exponentialverteilte Zufallszahlen, konkave Erwartungswertprofile, μ = Lageparameter der Exponentialverteilung) 123

6.2 Vergleich des Bartholomew-Tests und der Tests von Roth (jeweils parametrische Verteilung; $\alpha = 0{,}05$, $k = 3$, $n = 5$, $r = 5$, $\omega = 0$ und konkave Dosis-Wirkungs-Profile) 126

6.3 Güte des Linearen Kontrasttests bei Normalverteilungsapproximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen) 126

6.4 Güte des Linearen Kontrasttests bei Normalverteilungs-, Cochran- oder Welch-Approximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen) 128

6.5 Güte des Helmert-Kontrasttests bei Normalverteilungs-, Cochran- oder Welch-Approximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen) 128

6.6 Güte des TK-Kontrastes bei Normalverteilungs- bzw. t-Verteilungsapproximation und den auf S_I^2 bzw. $\hat{\sigma}_I^2$ basierenden Varianzschätzern ($\alpha = 0{,}05$, $k = 3$, $n = 5$, $r = 4$, $\omega = 1$ und linearen Dosis-Wirkungs-Profilen) 129

6.7 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (*r*) ($\alpha = 0{,}01$, $n = 3$, $\theta = 0$, $\omega = 0{,}5$, Verteilungstyp = RSV) 133

6.8 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (*r*) ($\alpha = 0{,}01$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 134

6.9 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (*r*) ($\alpha = 0{,}01$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 134

6.10 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (*r*) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 135

6.11 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der

Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 4,\ \omega = 0{,}5,\ \theta = 0$, Verteilungstyp = RSV) 135

6.12 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 5,\ \omega = 0{,}5,\ \theta = 0$, Verteilungstyp=RSV) 136

6.13 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 3,\ \theta = 0,\ \omega = 0,$ Verteilungstyp = GV) 137

6.14 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 4,\ \omega = 0,\ \theta = 0$, Verteilungstyp = GV) 138

6.15 Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 5,\ \omega = 0,\ \theta = 0$, Verteilungstyp = GV) 139

6.16 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01,\ n = 3,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 141

6.17 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01,\ n = 4,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 142

6.18 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01,\ n = 5,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 142

6.19 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 3,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 143

6.20 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 4,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 143

6.21 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 5,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = RSV) 144

6.22 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 3,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = GV) 145

6.23 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 4,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = GV) 146

6.24 Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05,\ n = 5,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = GV) 147

6.25 Power von Zweistichprobentests in Abhängigkeit vom Stichprobenumfang (n) $\alpha = 0{,}05,\ r = 7,\ \theta = 5,\ \omega = 0{,}5$, Verteilungstyp = LSV 148

6.26 Güte von Dreistichprobentests (infinite Verteilungen) für a) konkave bzw. b) konvexe Profile ($\alpha = 0{,}05, n = 4, r = 5, \omega = 0{,}5$, Verteilungstyp = RSV) 154

6.27 Güte von Vierstichprobentests (infinite Verteilungen) für a) konkave bzw. b)

konvexe Profile ($\alpha = 0{,}05$, $n = 4$, $r = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV) 155

6.28 Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 158

6.29 Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 159

6.30 Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 160

6.31 Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 161

6.32 Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 162

6.33 Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV) 163

6.34 Power von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 172

6.35 Power von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 173

6.36 Power von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 174

6.37 Power von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 175

6.38 Power von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 176

6.39 Power von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV) 177

6.40 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 3$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 178

6.41 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 3$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 178

6.42 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 4$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 179

6.43 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in

Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 4$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 179

6.44 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 5$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 180

6.45 Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 5$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV) .. 180

7.1 Mittelwerte der Versuchsglieder zu den drei verschiedenen Zeitpunkten 190

Abkürzungen

MED	Minimale effektive Dosis
HEDS	Höchster effektiver Dosisschritt
MÄD	Minimale äquivalente Dosis
MED&MÄD	Minimale effektive und äquivalente Dosis
MERT	Maximaler effizienter robuster Test
StatXact	Statistikprogramm der Cytel Software Corporation, Cambridge
SAS	Statistikprogramm des SAS Institute Inc., North Carolina
IML	Programmiersprache unter SAS
!	Fakultät
$\oplus$	Kronecker-Summe für Matrizen
α	Fehler 1. Art
$\mathbb{R}^a$	a-dimensionaler euklidischer Vektorraum
$\mathbf{A}'$	zu $\mathbf{A}$ gehörende(r) transformierte(r) Matrix (Vektor)
$\mathbf{A}^{-1}$	Inverse der Matrix $\boldsymbol{A}$
$[\mathbf{A}]^-$	verallgemeinerte Inverse der Matrix $\boldsymbol{A}$
$\lvert B \rvert$	Mächtigkeit der Menge B, sonst Betrag
$\lfloor x \rfloor$	kleinste ganze Zahl kleiner gleich x
$X_N \sim F$	(asymptotische) Wahrscheinlichkeitsverteilung von X_N ist F
$P_{X_N} \Rightarrow P_X$	Wahrscheinlichkeitsverteilung der Zufallsvariablen X_N konvergiert schwach gegen die Wahrscheinlichkeitsverteilung der Zufallsvariablen X
$\hat{X}_n \xrightarrow[n\to\infty]{} X$	Schätzer $\hat{X}$ konvergiert fast sicher gegen die Zufallsvariable X, falls der Stichprobenumfang gegen unendlich strebt
EX	1. Moment der Zufallsvariablen X (Erwartungswert)
cov($\boldsymbol{X}$)	Kovarianzmatrix des Vektors $\boldsymbol{X}$
$\lVert\ \rVert_m$	euklidische Norm des $\mathbb{R}^m$
$\forall$	für alle
∞	unendlich
F_i	Verteilungsfunktion des i-ten Versuchsgliedes
P_X	Wahrscheinlichkeitsverteilung der Zufallsvariable X

1 Einleitung

1.1 Dosis-Wirkungs-Versuche

Ihren Anstoß fand die vorliegende Arbeit im Bestreben der BASF AG, die statistische Auswertung ihrer Pflanzenschutzmittelversuche weiter zu verbessern. Als Beispielversuche stehen Dosis-Wirkungs-Versuche, die mit Fungiziden oder Herbiziden durchgeführt werden. Diese Versuche werden sowohl an verschiedenen Orten als auch zu mehreren Zeitpunkten und in Zusammenhang mit verschiedenen Kulturpflanzen durchgeführt. Für jedes der drei bis fünf Versuchsglieder beträgt die Anzahl der Wiederholungen nur drei, vier oder fünf. Ursprünglich wurden der Schädlingsbefall und die Wirkung bzw. der Schaden auf einer von 0 bis 100 Prozent reichenden Skala mit Schrittweite eins visuell geschätzt. Als Ergebnis einer intensiven Studie der Daten [1] wurde die Skala aber auf eine kleinere eingeschränkt. Die Kategorien, die in einem weniger wichtigen und zudem schwer differenzierbaren Bereich lagen, wurden zusammengefaßt. Im Bereich 0-15 und 85-100 wurde die 1%-Schritt-Skala beibehalten. Ansonsten beträgt die Schrittweite fünf Prozent. Es liegt demnach eine nichtlineare Skala mit 45 Meßpunkten (Kategorien) vor. Im fortgeschrittenen Versuchsstadium sind die Beobachtungen der Dosen jedoch beim Wert 0 bzw. 100 geclustert, wobei die Variabilität an den Grenzen abnimmt.

Bei Versuchen in der Landwirtschaft und im Gartenbau werden häufig gröbere Skalen genutzt. Little [2] beschreibt z. B. unterschiedliche Bonitурskalen, die ebenfalls auf geschätzten Prozentwerten basieren. Dabei treten sowohl Drei-, Vier- oder auch Fünfpunktskalen auf. Relativ häufig wird eine Neunpunktskala empfohlen. Zum Beispiel kann folgende Skala für Herbizidversuche genutzt werden [3]:

1	=	kein	Unkraut			6	=	15	- 25%	Unkrautbefall im
2	=	0	- 2,5%	Unkrautbefall im		7	=	25	- 35%	Vergleich zur
3	=	2.5	- 5%	Vergleich zur		8	=	35	- 67,5%	unbehandelten
4	=	5	- 10%	unbehandelten		9	=	67,5	- 100%	Parzelle
5	=	10	- 15%	Parzelle						

Weitere Beispiele für Bonitурskalen befinden sich in der Methodenbeschreibung der BASF AG [4].

Wiederholung	D_0	D_1	D_2	D_3	D_S
1	20	3	1	0	2
2	20	1	1	0	5
3	15	3	2	1	7
4	20	0	0	0	0

Tabelle 1.1: Fungizidversuch mit 3 Dosen (D_1, D_2, D_3), einem Standard (D_S) und einer Kontrolle (D_0)

Der in Tabelle 1.1 dargestellte Datensatz ist Teil eines Fungizidversuchs, bei dem für die Kulturpflanze Winterweizen der Befall mit Erysighe graminis bonitiert wurde. Die 20 Beobachtungen dieses Versuches sind auf 8 Punkte konzentriert. Stetigkeitsannahmen (mit Wahrscheinlichkeit 1 keine gleichen Werte) sind offensichtlich nicht vertretbar.
Eine Bonitur ist stets subjektiv und daher mit Fehlzuordnungen verbunden. Wird ein Versuch jedoch von *einer* Person bewertet, so werden die Fehler eher so geartet sein, daß entweder alle Werte etwas zu hoch oder alle Werte etwas zu niedrig geschätzt werden. Die Schätzungen verschiedener geschulter Personen werden daher im allgemeinen gut korrelieren. In der Medizin oder in der Qualitätskontrolle treten ähnliche Probleme auf[5]. Zum Beispiel beurteilt ein Arzt anhand bestimmter vorliegender Informationen den Schweregrad einer Krankheit (leicht, mittel, schwer, ...). Auch hier sind Zuordnungsfehler denkbar. Thöni [6] zeigt, wie Zuordnungsfehler modelliert werden können. In dieser Arbeit werden sie jedoch nicht weiter untersucht.

Was sind die Ursachen der kleinen Fallzahlen und wie wirken sich diese Fallzahlen aus?
Wirtschaftliche Aspekte (Kosten), Platzprobleme (Mangel an genügend homogenen Versuchsflächen) bzw. organisatorische Schwierigkeiten (z. B. nicht genügend Patienten mit einer gewissen Krankheit) führen meist zu Fallzahlbeschränkungen. Bei Versuchen an Mensch und Tier kommen ethische Gründe hinzu[7; 8; 9]. „Kleine“ Experimente können zudem leichter und schneller unter verschiedenen Bedingungen (z. B. an mehreren Orten) durchgeführt werden. Falls sich die Zusammenhänge unter verschiedenen Bedingungen reproduzieren lassen, sind dann allgemeinere Aussagen ableitbar. Der Einzelversuch kann aufgrund der kleinen Fallzahlen jedoch vermehrt zu Fehlentscheidungen führen, da beim Testen der Fehler 2. Art nicht in angemessener Weise beschränkt werden kann. Klar ist, daß bei sehr kleinen Fallzahlen nur relativ deutliche Unterschiede mit hoher Wahrscheinlichkeit aufgedeckt werden können. Letzteres soll das folgende Beispiel verdeutlichen.

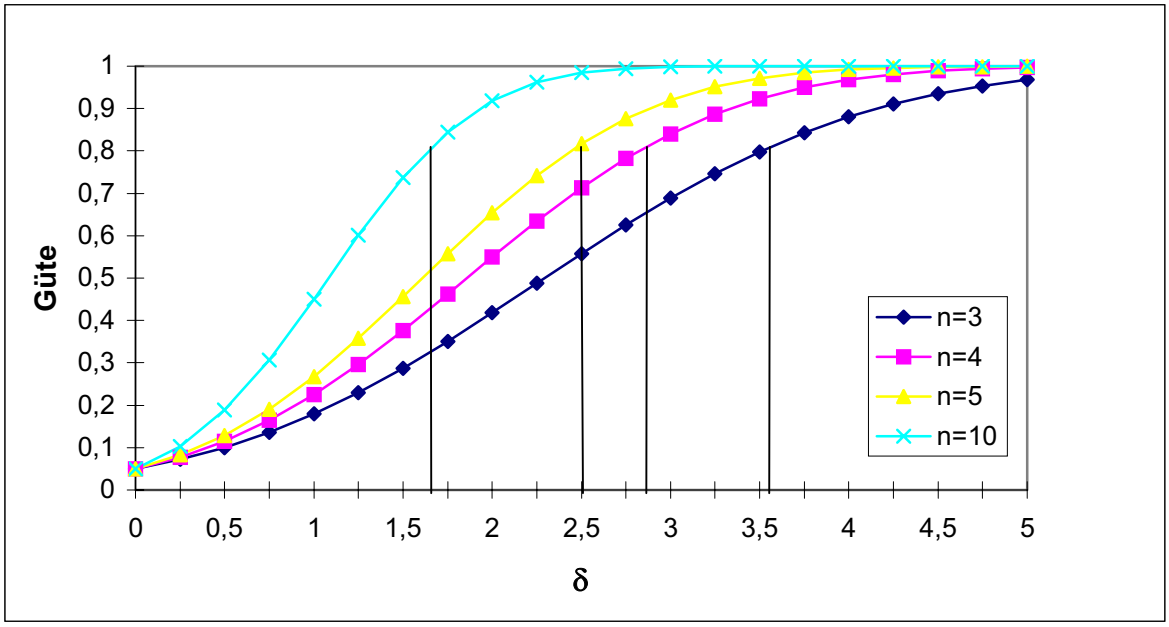


Abbildung 1.1: Güte des einseitigen t-Tests in Abhängigkeit von der Differenz der Erwartungswerte (δ) und dem Stichprobenumfang ($n = n_1 = n_2$) (Varianz $\sigma^2 = 2$)

Es seien zwei normalverteilte Stichproben mit gleichen Varianzen und gleichen Stichprobenumfängen gegeben ($F_1 = N(\mu_1, \sigma^2), F_2 = N(\mu_2, \sigma^2)$). Die Hypothese, daß der Erwartungswert der zweiten Stichprobe μ_2 größer ist als der Erwartungswert μ_1 der ersten, wird mit dem einseitigen t-Test getestet. In Abbildung 1.1 ist in Abhängigkeit von $\delta = \mu_2 - \mu_1$ die Güte des einseitigen t-Tests für 2 Stichproben bei fester Varianz dargestellt.
An diesem Beispiel ist zu erkennen, daß der Unterschied zwischen $n = 3$ und $n = 5$ deutlich ausfallen kann. Noch gravierender ist der Unterschied zwischen $n = 3$ und $n = 10$. Für den Fall, daß der Fehler 2. Art (β) bei einem vorgegebenen Signifikanzniveau von $\alpha = 0{,}05$ durch 0,2 beschränkt werden soll, sind in Abhängigkeit von der Varianz und vom Stichprobenumfang in Tabelle 1.2 die notwendigen Unterschiede für die Erwartungswerte beschrieben. Auch hier sind deutliche Unterschiede zwischen $n = 3$ und $n = 5$ zu erkennen. Ähnlich sieht es aus, wenn statt der zwei Normalverteilungen zwei diskrete Verteilungen betrachtet werden, die zumindest in einigen Merkmalen mit der Normalverteilung übereinstimmen. Die in Tabelle 1.3 dargestellten diskreten Verteilungen besitzen dieselbe Varianz und Schiefe wie $F_1^\infty = N(0, 1{,}2)$ und $F_2^\infty = N(1, 1{,}2)$. Werden zwei Stichproben, deren Verteilungen durch F_1^r bzw. F_2^r ($r = 3,4,5,6,\infty$) gegeben sind, mittels des t-Tests verglichen, so ergeben sich die in Tabelle 1.4 dargestellten simulierten Gütewerte (die Differenz der Erwartungswerte ist jeweils eins). Zusätzlich ist jeweils der Vergleich F_1^r gegen F_1^r (entspricht der Nullhypothese;

Varianz (σ^2)	n = 3	n = 4	n = 5	n = 10
1	2,48	1,99	1,72	1,16
2	3,51	2,82	2,44	1,64
3	4,30	3,45	2,99	2,00
4	4,97	3,99	3,45	2,31

Tabelle 1.2: Minimal nachweisbarer Unterschied des einseitigen t-Tests ($\alpha = 0{,}05$, $\beta = 0{,}2$) in Abhängigkeit von der Varianz und dem Stichprobenumfang

	π_1	π_2	π_3	π_4	π_5	π_6	π_7	μ_i	σ^2
F_1^3	0,15	0	0,7	0	0,15	0	0	3	1,2
F_2^3	0	0,15	0	0,7	0	0,15	0	4	1,2
F_1^4	0,2375	0,2625	0,2625	0,2375	0	0	0	2,5	1,2
F_2^4	0	0,2375	0,2625	0,2625	0,2375	0	0	3,5	1,2
F_1^5	0,1	0,2	0,4	0,2	0,1	0	0	3	1,2
F_2^5	0	0,1	0,2	0,4	0,2	0,1	0	4	1,2
F_1^6	0,05	0,0875	0,3625	0,3625	0,0875	0,05	0	3,5	1,2
F_2^6	0	0,05	0,0875	0,3625	0,3625	0,0875	0,05	4,5	1,2

Tabelle 1.3: Symmetrische diskrete Verteilungen mit Varianz $\sigma^2 = 1{,}2$ ($\pi_s = P(X = s)$, $s = 1,\ldots,r$)

$r = 3,4,5,6,\infty$) aufgenommen worden (oberer Eintrag). Wie im allgemeinen erwartet, sinkt mit zunehmender Diskretisierung die Güte des t-Tests ($r = 4, 5, 6$). Der Einfluß der Diskretheit ist jedoch bei weitem nicht so groß wie der Einfluß des Stichprobenumfangs. Die simulierten Güten des Vergleichs F_1^∞ gegen F_2^∞ können in diesem Beispiel sogar als guter Richtwert für die Vergleiche der diskreten Verteilungen benutzt werden. An den simulierten Gütewerten des Vergleichs F_1^3 gegen F_2^3 ist allerdings auch erkennbar, daß zum Teil nicht so offensichtliche Effekte bei der Kombination von Meßskala und Stichprobenumfang zu beachten sind (Differenz der Powerwerte für F_1^3 vs. F_2^3 und F_1^4 vs. F_2^4 in Abhängigkeit von n).

		Vergleich				
Fallzahl n	Differenz $\delta = \mu_2 - \mu_1$	F_1^3 vs. F_2^3	F_1^4 vs. F_2^4	F_1^5 vs. F_2^5	F_1^6 vs. F_2^6	F_1^∞ vs. F_2^∞
3	0	0,004	0,043	0,036	0,028	0,050
	1	0,262	0,188	0,205	0,216	0,240
4	0	0,041	0,045	0,052	0,047	0,050
	1	0,276	0,277	0,301	0,334	0,308
5	0	0,052	0,049	0,049	0,052	0,050
	1	0,348	0,338	0,364	0,388	0,372
6	0	0,056	0,049	0,046	0,054	0,048
	1	0,453	0,397	0,420	0,447	0,429
7	0	0,052	0,054	0,048	0,054	0,050
	1	0,466	0,470	0,476	0,503	0,495
8	0	0,054	0,053	0,049	0,047	0,049
	1	0,551	0,529	0,531	0,544	0,532
9	0	0,052	0,051	0,053	0,050	0,051
	1	0,594	0,568	0,583	0,592	0,575
10	0	0,049	0,051	0,054	0,049	0,049
	1	0,611	0,612	0,622	0,629	0,624

Tabelle 1.4: Güte des einseitigen t-Tests in Abhängigkeit vom Stichprobenumfang, dem Träger der Verteilung ($r = 3,4,5,6,\infty$) und der Differenz der Erwartungswerte ($\delta = \mu_2 - \mu_1$)

Während beim Vergleich der beiden Normalverteilungen stets durch eine hinreichende Erhöhung von $\delta = \mu_2 - \mu_1$ (n konstant) eine Güte von eins erreicht werden kann, ist dies bei diskreten Verteilungen nicht immer möglich (δ kann nicht unbeschränkt erhöht werden). Zusätzlich können extreme jedoch eigentlich wünschenswerte Situationen (z. B. keine Variabilität in den Gruppen) dazu führen, daß die Berechnung der Teststatistik unmöglich ist (Varianzschätzer gleich Null).

Wie kann das auszuwertende Versuchsdesign allgemein b eschrieben werden?

In vorliegender Arbeit werden einfaktorielle Anlagen untersucht. Im Sinne einer einfachen Formulierung werden diese Experimente als Dosis-Wirkungs-Versuche bezeichnet. Die Va-

suchsglieder werden stets als Dosen eines Präparats interpretiert. Zum Vergleich wird zusätzlich eine Kontrollgruppe (unbehandelte Gruppe; in klinischen Studien auch als Placebogruppe bezeichnet) ins Design aufgenommen. Weiterhin wird ein Standardprodukt (z. B. in Fungizidversuchen) bzw. eine sogenannte positive Kontrolle (z. B. in toxikologischen Studien[10]) betrachtet. Kurz zusammengefaßt soll im weiteren folgendes Design untersucht werden:

1. Das Versuchsdesign stellt eine univariate einfaktorielle vollständig randomisierte Anlage dar.
2. Mit $\boldsymbol{X}_0 = (X_{01},...,X_{0n_0})'$ seien stets die Zufallsgrößen der Kontrolle D_0 und mit $\boldsymbol{X}_i = (X_{i1},...,X_{in_i})'$ die der ausgewählten Dosen D_i, $i = 1,...,k$ $(D_1 < \cdots < D_k)$ bezeichnet. Mit $\boldsymbol{X}_S = (X_{S1},...,X_{Sn_S})'$ werden die Zufallsgrößen des Standards D_S bezeichnet. Die Realisierungen (Beobachtungen) der Zufallsgrößen werden durch Kleinbuchstaben gekennzeichnet. Der Stichprobenumfang der Stichprobe i $(i = 0, \ldots, k, S)$ wird stets mit n_i abgekürzt.
3. Die Elemente innerhalb einer Stichprobe i seien identisch verteilt. Ihre Verteilungsfunktion wird mit F_i $(F_{ij} = F_i, j = 1, \ldots, n_i)$ und der Erwartungswert der Zufallsvariablen X_{ij} wird mit $\mu_{ij} = \mu_i = \int x\, dF_i(x)$ bezeichnet.
4. Alle Zufallsgrößen $(X_{ij})_{\substack{i=0,\ldots,k,S \\ j=1,\ldots,n_i}}$ sind voneinander unabhängig.
5. Die Wirkung wird auf einer ordinalen Skala mit einer endlichen Anzahl (hier stets durch r gekennzeichnet) von Kategorien K_1 bis K_r gemessen .

Wie lauten mögliche Versuchsfragen?

A) Globale Aussage

Mit einem globalen Test wird geprüft, ob überhaupt eine nachweisbare Abhängigkeit zwischen Dosis und Wirkung besteht. Hierbei werden bestimmte Parameter der Wahrscheinlichkeitsverteilungen verglichen, die den Daten der einzelnen Gruppen unterstellt werden. In vielen Fällen ist dieser Parameter der Erwartungswert (asymptotisch, d. h., für sehr große Fallzahlen (Stichprobenumfänge) ist dies der Mittelwert). Intuitiv ist bei Dosis-Wirkungs-Studien (z. B. bei Versuchen mit Pflanzenschutzmitteln) mit hoher Wahrscheinlichkeit ein monotoner Zusammenhang (zumindest in bestimmten Dosisbereichen) zwischen Dosis und Wirkung zu erwarten. Um die Nullhypothese „die Wirkung der unterschiedlichen Dosen sind gleich“ mit hoher Wahrscheinlichkeit zu verwerfen, sollten daher Trendtests benutzt werden.

Sie besitzen speziell für die Alternative „mit steigender Dosis nimmt die Wirkung zu“ eine hohe Güte. Liegt real ein monotoner Zusammenhang zwischen Dosis und Wirkung vor, so ist die Wahrscheinlichkeit, die Nullhypothese abzulehnen, größer, als wenn ein Test benutzt wird, der nur auf beliebige Unterschiede testet. Mit Hilfe der Erwartungswerte kann die globale Aussage (Gleichheit versus Trend) im Lokationsmodell $F_i(x) = F(x - \mu_i)$ durch folgendes Hypothesenpaar (hier als Wirkungsanstieg formuliert) beschrieben werden:

Nullhypothese: $H_{0k}^{\mu} : \mu_0 = \mu_1 = \cdots = \mu_k$ versus

Alternativhypothese: $H_{Ak}^{\mu} : \mu_0 \le \mu_1 \le \cdots \le \mu_k , \mu_0 < \mu_k .$

Mittels der Verteilungsfunktionen F_i kann das Problem nichtparametrisch durch die stochastische Ordnung der Verteilungsfunktionen beschrieben werden:

Nullhypothese: $H_{0k}^{F} : F_0 = F_1 = \cdots = F_k$ versus

Alternativhypothese: $H_{Ak}^{F} : F_0 \le F_1 \le \cdots \le F_k , F_0 < F_k ,$

($F_i \le F_{i+1}$ gilt genau dann, wenn $1 - F_i(x) \le 1 - F_{i+1}(x) \; \forall x$). Speziell im Fall geordneter kategorialer Daten mit beschränkter Kategorienanzahl kann dies mittels der Punktwahrscheinlichkeiten $\pi_{is} = P(X_{ij} = K_s)$ durch folgende Ungleichungen beschrieben werden:

$$\begin{aligned} \pi_{i1} &\ge \pi_{(i+1)1} \\ \pi_{i1} + \pi_{i2} &\ge \pi_{(i+1)1} + \pi_{(i+1)2} \\ &\vdots \\ \pi_{i1} + \pi_{i2} + \cdots + \pi_{i(r-1)} &\ge \pi_{(i+1)1} + \pi_{(i+1)2} + \cdots + \pi_{(i+1)(r-1)} . \end{aligned} \qquad (i = 0, \ldots, k-1)$$

Die Formulierung der Alternative als Anstieg ist hier vollkommen willkürlich, genauso hätte sie als abfallend definiert werden können. Die Verfahren sind analog definierbar. Eine strenge Monotonie der Versuchsglieder, d. h., alle Wirkungszuwächse sind positiv, wird allerdings nicht gefordert. Hierfür sei auf Grömping [11] verwiesen.

B) Minimale effektive Dosis

Wird die globale Hypothese H_{0k}^{μ} bzw. H_{0k}^{F} abgelehnt, sind auch die nachfolgenden Hypothesenpaare von Interesse (können analog für die Verteilungsfunktionen formuliert werden):

$$
\begin{array}{ll}
H_{0k-1}^{\mu}: \mu_0 = \mu_1 = \cdots = \mu_{k-1} & H_{Ak-1}^{\mu}: \mu_0 \le \mu_1 \le \cdots \le \mu_{k-1} \quad (\mu_0 < \mu_{k-1}) \\
H_{0k-2}^{\mu}: \mu_0 = \mu_1 = \cdots = \mu_{k-2} & H_{Ak-2}^{\mu}: \mu_0 \le \mu_1 \le \cdots \le \mu_{k-2} \quad (\mu_0 < \mu_{k-2}) \\
\vdots & \vdots \\
H_{02}^{\mu}: \mu_0 = \mu_1 = \mu_2 & H_{A2}^{\mu}: \mu_0 \le \mu_1 \le \mu_2 \qquad (\mu_0 < \mu_2) \\
H_{01}^{\mu}: \mu_0 \ge \mu_1 & H_{A1}^{\mu}: \mu_0 < \mu_1
\end{array}
\tag{1.1}
$$

Als **minimale effektive Dosis** (***MED***) wird die kleinste Dosis *i* bezeichnet, für die H_{0i}^{μ} (H_{0i}^{F}) abgelehnt werden kann:

$$MED = \min(D_i, 1 \le i \le k : \mu_0 < \mu_i) \text{ bzw. } MED = \min(D_i, 1 \le i \le k : F_0 < F_i).$$

Ist die Monotonieannahme der Alternativhypothese nicht sicher (d. h., es können auch andere Dosis-Wirkungs-Zusammenhänge auftreten, z. B. erst Anstieg, dann Abfall der Wirkung), so ist die Verwendung von Trendtests problematisch. Aus der Ablehnung von H_{0i}^{μ} kann dann nicht mehr auf $\mu_0 < \mu_i$ geschlossen werden, da ein Trendtest durchaus auch bei anderen real vorliegenden Alternativen zur Ablehnung von H_{0i}^{μ} führen kann. Soll dennoch eine minimale effektive Dosis bestimmt werden, so sollten paarweise Vergleiche zu den folgenden Hypothesen durchgeführt werden:

$$
\begin{array}{ll}
H_{00k}^{\mu}: \mu_0 \ge \mu_k & H_{A0k}^{\mu}: \mu_0 < \mu_k \\
H_{00(k-1)}^{\mu}: \mu_0 \ge \mu_{k-1} & H_{A0(k-1)}^{\mu}: \mu_0 < \mu_{k-1} \\
\vdots & \vdots \\
H_{002}^{\mu}: \mu_0 \ge \mu_2 & H_{A02}^{\mu}: \mu_0 < \mu_2 \\
H_{001}^{\mu}: \mu_0 \ge \mu_1 & H_{A01}^{\mu}: \mu_0 < \mu_1 \ .
\end{array}
\tag{1.2}
$$

Sollen die Dosen und die *MED* die Forderung erfüllen „Alle Dosen, die größer als $D_i = MED$ sind, sind ebenfalls effektiv“, so ist die *MED* wie folgt zu definieren:

$$
\begin{aligned}
&MED = \min(D_i, 1 \le i \le k : \mu_0 < \mu_j \text{ für } j = i, \ldots, k) \\
&\text{bzw. } MED = \min(D_i, 1 \le i \le k : F_0 < F_j \text{ für } j = i, \ldots, k).
\end{aligned}
\tag{1.3}
$$

C) Höchster effektiver Dosisschritt

Liegt global ein Unterschied zwischen den Versuchgliedern vor, so ist von Interesse, ob auch je Dosisschritt ein Unterschied nachgewiesen werden kann, d. h., unterscheidet sich die Wirkung der Dosis D_i von Dosis D_{i+1}. Soll z. B. abhängig von bekannten Störgrößen (u. a. Bodenfeuchtigkeit) nicht die *MED*, sondern die nächsthöhere Dosis eingesetzt werden, sollte gesichert sein, daß diese Dosis auch wirklich eine höhere (und nicht gleiche) Wirkung besitzt. Dies ist auch für alle nachfolgenden Schritte interessant. Der höchste Dosisschritt, der noch nachweislich mit einer größeren Wirkung verbunden ist, wird als **höchster effektiver Dosisschritt** bezeichnet: ***HEDS*.** Die Bestimmung erfolgt mittels Testens folgender Hypothesen:

$$\begin{array}{ll} H^{\mu}_{0,(k-1)k}: \mu_{k-1} \geq \mu_k & H^{\mu}_{A,(k-1)k}: \mu_{k-1} < \mu_k \\ H^{\mu}_{0,(k-2)(k-1)}: \mu_{k-2} \geq \mu_{k-1} & H^{\mu}_{A,(k-2)(k-1)}: \mu_{k-2} < \mu_{k-1} \\ \vdots & \vdots \\ H^{\mu}_{0,12}: \mu_1 \geq \mu_2 & H^{\mu}_{A12}: \mu_1 < \mu_2 \ . \end{array} \qquad (1.4)$$

Existiert eine *MED*, sind die Dosisschritte oberhalb der *MED* besonders interessant. Ist aber z. B. die Variabilität der Kontrollgruppe sehr groß (als *MED* wird wahrscheinlich eher eine hohe Dosis erkannt) oder existiert überhaupt keine Kontrollgruppe im Design, sind die Testergebnisse bezüglich aller Dosisschritte von Interesse.

D) Minimale äquivalente Dosis

Für alle (effektiven) Dosen stellt sich die Frage, ob sie eine Wirkung besitzen, die der Wirkung eines Standards äquivalent ist. Die Dosis D_i wird dabei als äquivalent zum Standard bezeichnet, falls das Folgende gilt:

i) $\mu_S - \varepsilon < \mu_i$ (Wirkung von Dosis D_i ist höchstens um ε schlechter oder gar besser)

ii) $\mu_S - \delta\,\mu_S < \mu_i$ (Wirkung von Dosis D_i ist höchstens um $\delta\,\mu_S$ schlechter oder gar besser).

Die Konstanten ε bzw. δ müssen vor dem Versuch festgelegt werden. Die minimale Dosis, welche eine zum Standard äquivalente Wirkung besitzt wird mit ***MÄD*** **(= minimale äquivalente Dosis)** bezeichnet:

$$M\ddot{A}D = \min(D_i, 1 \leq i \leq k: \mu_i > \mu_S - \varepsilon) \text{ bzw. } M\ddot{A}D = \min(D_i, 1 \leq i \leq k: F_{S\varepsilon} < F_i)$$

$$M\ddot{A}D = \min(D_i, 1 \leq i \leq k: \mu_i > (1-\delta)\mu_S) \text{ bzw. } M\ddot{A}D = \min(D_i, 1 \leq i \leq k: F_{S\delta} < F_i),$$

wobei $\varepsilon \geq 0, \delta \geq 0$, $F_{S\varepsilon} = F_S(x+\varepsilon)$ und $F_{S\delta}(x) = F_S(x+\delta\,\mu_S)$ gilt. Beschrieben werden kann dieses Problem durch folgende Hypothesen:

$$\begin{array}{ll} H^{\mu}_{0,Sk}(\varepsilon): \mu_S - \varepsilon \geq \mu_k & H^{\mu}_{A,Sk}(\varepsilon): \mu_S - \varepsilon < \mu_k \\ H^{\mu}_{0,S(k-1)}(\varepsilon): \mu_S - \varepsilon \geq \mu_{k-1} & H^{\mu}_{A,S(k-1)}(\varepsilon): \mu_S - \varepsilon < \mu_{k-1} \\ \vdots & \vdots \\ H^{\mu}_{0,S1}(\varepsilon): \mu_S - \varepsilon \geq \mu_1 & H^{\mu}_{A,S1}(\varepsilon): \mu_S - \varepsilon < \mu_1 \end{array} \qquad (1.5)$$

bzw.

$$\begin{array}{ll} H^{\mu}_{0,Sk}(\delta): (1-\delta)\mu_S \geq \mu_k & H^{\mu}_{A,Sk}(\delta): (1-\delta)\mu_S < \mu_k \\ H^{\mu}_{0,S(k-1)}(\delta): (1-\delta)\mu_S \geq \mu_{k-1} & H^{\mu}_{A,S(k-1)}(\delta): (1-\delta)\mu_S < \mu_{k-1} \\ \vdots & \vdots \\ H^{\mu}_{0,S1}(\delta): (1-\delta)\mu_S \geq \mu_1 & H^{\mu}_{A,S1}(\delta): (1-\delta)\mu_S < \mu_1 . \end{array} \qquad (1.6)$$

Eine Verschärfung der *MÄD*-Definition wie im *MED*-Problem ist auch hier möglich:

$$MÄD = \min(D_i, 1 \leq i \leq k: \mu_j > \mu_S - \varepsilon, j = i,\ldots,k) \text{ bzw.}$$
$$MÄD = \min(D_i, 1 \leq i \leq k: F_{S\varepsilon} < F_j, j = i,\ldots,k)$$

oder

$$MÄD = \min(D_i, 1 \leq i \leq k: \mu_j > (1-\delta)\mu_S, j = i,\ldots,k)$$
$$\text{bzw. } MÄD = \min(D_i, 1 \leq i \leq k: F_{S\delta} < F_j, j = i,\ldots,k).$$

E) Minimale effektive und äquivalente Dosis

Für die Anwendung sind vor allem die Dosen interessant, die sich von der Kontrolle unterscheiden und eine zum Standard äquivalente Wirkung besitzen. Die minimale dieser Dosen wird im weiteren als ***MED&MÄD*** **(= minimale effektive und äquivalente Dosis)** bezeichnet:

$$MED\&\,MÄD = \min(D_i, 1 \leq i \leq k: \mu_i > \mu_S - \varepsilon \wedge \mu_i > \mu_0) \text{ bzw.}$$
$$MED\&\,MÄD = \min(D_i, 1 \leq i \leq k: F_{S\varepsilon} < F_i \wedge F_0 < F_i)$$

oder

$$MED\&\,MÄD = \min(D_i, 1 \leq i \leq k: \mu_i > \mu_S - \delta\,\mu_S \wedge \mu_i > \mu_0) \text{ bzw.}$$
$$MED\&\,MÄD = \min(D_i, 1 \leq i \leq k: F_{S\delta} < F_i \wedge F_0 < F_i).$$

Die zu testenden Hypothesen ergeben sich aus den Problemen B) und D).

Die Punkte A) bis E) behandeln Fragen, die zur Routine bei vielen Dosis-Wirkungs-Studien gehören [12; 13; 14; 15]. Zur Lösung der Probleme B, C, D und E werden mehrere Tests sequentiell durchgeführt.

Ziel der vorliegenden Arbeit

Es werden zum einen einseitge Zweistichprobentests, zum anderen Trendtests (Mehrstichprobentests) gesucht, die für die Auswertung geordneter kategorialer Daten sowohl bei kleinen Stichprobenumfängen als auch bei einer kleinen Anzahl beobachteter Kategorien geeignet sind. Da die Bedingungen, die an die Daten gestellt werden, recht allgemein gehalten werden sollen, besteht das Ziel nicht im Finden eines (gleichmäßig) besten Verfahrens (das es nicht gibt). Vielmehr sollen existierende Tests bzw. Verfahren bezüglich ihrer Anwendbarkeit verglichen und wenn möglich, verbessert werden. Einen Schwerpunkt bilden hierbei bekannte parametrische Tests und nichtparametrische Rangstatistiken. Als Ergebnis sollen Tests bzw. Verfahren vorgeschlagen werden, die technisch stabil, d. h. mit hoher Wahrscheinlichkeit technisch durchführbar, sind und die ein stabiles Güteverhalten aufweisen. Es wird angestrebt, möglichst einfache bzw. leicht verständliche Tests bzw. Verfahren vorzuschlagen, da deren Akzeptanz wesentlich höher ist.

1.2 Geordnete kategoriale Daten

Im Mittelpunkt dieser Arbeit stehen Verfahren zur Auswertung geordneter kategorialer Daten. Daher wird zunächst eine kurze Übersicht zu diesen Daten und ihrer Auswertung gegeben. Zum weiterführenden Studium werden die Bücher von Agresti [16; 17] empfohlen. Sie geben einen ausführlichen Überblick zu diesem Thema. Dabei ist letztgenanntes Werk vor allem für den Praktiker geeignet. Stokes et al. [18] beschreiben Möglichkeiten der Auswertung von kategorialen Daten mit Hilfe des Statistikprogrammes SAS. In bezug auf sehr kleine Stichprobenumfänge kann auch das zum Programm StatXact gehörende Handbuch [19] empfohlen werden. In einer Vielzahl von Artikeln eines breiten Zeitschriftenspektrums wird ebenfalls auf die Auswertung geordneter kategorialer Daten eingegangen. Hervorzuheben ist hierbei ein Artikel von Chuang-Stein und Agresti [13], der sich hauptsächlich mit monotonen Dosis-Wirkungs-Beziehungen ordinaler Daten befaßt. Ebenso ist ein Artikel von Agresti [20]

zu empfehlen, der speziell einen Überblick über die modellgebundene Auswertung ordinaler Daten gibt und ein sehr umfangreiches Literaturregister besitzt.

Modellierung geordneter kategorialer Daten

Agresti [17, S.2] definiert eine kategoriale Variable wie folgt: *„A categorical variable is one for which the measurement scale consist of a set of categories."*

Oft werden diese Variablen auch als nominale Variablen bezeichnet. Ein einfaches Beispiel stellen die Farbkategorien einer Pflanzenblüte (z. B. bei einer Tulpe: rot, rot-weiß, ...) dar. Eine geordnete kategoriale Variable definiert Agresti [16, S.2] hingegen folgendermaßen:

„A categorical variable is referred to as „ordinal" rather than „interval" when there is a clear ordering of the categories but the absolute distances among them are unknown."

Beispiele hierfür sind Bonituren in der Landwirtschaft und im Gartenbau, z. B. Qualitätsklassen (gute, mittlere, schlechte, sehr schlechte) oder Befallsklassen (kein, schwacher, mittelstarker, starker), aber auch Beschreibungen von Krankheiten in der Medizin (ohne Symptome, leichte Symptome, mittlere Symptome, schwere Symptome).

Eine Zufallsvariable, die ein geordnetes kategoriales Merkmal beschreibt, kann unter der Einschränkung, daß die Anzahl der Kategorien endlich sei, wie folgt mathematisch modelliert werden:

1. Möglichkeit: Die Realisierungen der Zufallsvariablen liegen in der Menge $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_r\}$, wobei $\boldsymbol{x}_s$ ein r-dimensionaler Spaltenvektor ist, der aus Nullen und einer Eins (an s-ter Stelle) besteht. Wird die s-te Kategorie beobachtet, so wird dies durch den Vektor $\boldsymbol{x}_s$ dargestellt. Die Verteilung dieser Vektoren wird durch eine Multinomialverteilung

$$\left\{ \mathrm{M}(1,\boldsymbol{\pi}):\ \boldsymbol{\pi} = (\pi_1,\ldots,\pi_r),\ \pi_s \geq 0, \sum_{s=1}^{r} \pi_s = 1 \right\},\ P(\boldsymbol{X}_{ij} = \boldsymbol{x}_s) = \pi_s$$

beschrieben. Vor allem bei der Auswertung von Kontingenztafeln wird diese Modellierung häufig genutzt. Trotz der komplizierten multivariaten Darstellung können sowohl asymptotische als auch exakte Verteilungsaussagen hergeleitet werden. Dabei ist es in der Regel nicht notwendig, daß den Kategorien Scores zugeordnet werden.

2. Möglichkeit: Die Realisierungen der Zufallsvariablen liegen in der Menge $\{a_1,..,a_r\}$, wobei a_s reelle (oft natürliche) Zahlen sind, die den Kategorien K_s zugeordnet werden. Die

Scores $\{a_1,\ldots,a_r\}$ erfüllen die Beziehung $a_1 < a_2 < \ldots < a_r$. Die Verteilung der Daten wird durch Angabe von $\pi_s = P(X = a_s),\ s = 1,\ldots,r$ definiert. Diese Verteilungsfamilie steht oft mit nichtparametrischen Modellierungen und verallgemeinerten linearen Modellen in Verbindung.

Für die asymptotischen Tests, wie sie in Kapitel 4 vorgestellt werden, ist es unwesentlich, welche Form gewählt wird. Durch geeignete lineare Transformationen ergeben sich dieselben Tests. Zum einen wird die asymptotische Normalität vom Vektor der absoluten Häufigkeiten, zum anderen die asymptotische Normalität der einzelnen Gruppenmittelwerte genutzt.

Entstehungsursachen für geordnete kategoriale Daten

Die Kategorisierung von stetigen Merkmalen, deren exakte Bestimmung zu aufwendig bzw. nicht möglich ist, ist eine häufige Entstehungsursache geordneter kategorialer Daten. So könnte der Befall einer Pflanze durch einen Pilz mit hohem technischen Aufwand auch genauer bestimmt werden. Der Aufwand würde aber oft in keinem Verhältnis zum Nutzen stehen. In vielen Fällen werden kategoriale Daten auch durch eine gewollte Datenreduktion (Verdichtung) generiert. Ein Beispiel ist die Zuordnung des Gewichts in Gewichtsklassen. Mathematisch können diese gruppierten stetigen Daten auch durch folgendes Modell beschrieben werden:

Es existieren eine beobachtbare oder auch eine unbeobachtbare Zufallsvariable *Y*, deren Wertebereich in *r* disjunkte Bereiche (B_1, ..., B_r) unterteilt ist, und eine beobachtbare kategoriale Zufallsvariable *X*, die abhängig von *Y* *r* Werte annehmen kann. Die Variable *X* nimmt gerade dann den Wert x_i an, wenn die Variable *Y* im *i*-ten Bereich ihres Wertebereichs liegt:

$$X = x_i \quad \Leftrightarrow Y \in B_i \ \forall\, i\,.$$

Anders ist es bei den Daten, bei denen subjektiv eine vorhandene Menge von Informationen bewertet wird. Die Qualität einer Gemüsesorte kann z. B. in Abhängigkeit von Größe, Gewicht, Farbe oder Geschmack bewertet werden. Der Schweregrad einer Krankheit wird durch einen Arzt, z. B. anhand von Information über äußeres Auftreten, persönliches Wohlbefinden und klinische Befunde eingeschätzt. Hier treten also auch Größen auf, die nicht gemessen werden können. Bei den gruppierten stetigen Daten ist die Zuordnung zu einer Kategorie oft einfach, und es treten wenige Fehlzuordnungen auf. Bei den subjektiv festgelegten Daten

kann es hingegen in Abhängigkeit vom Umfang und von der Qualität der Informationen häufiger zu Fehlklassifizierungen kommen [5].

Die Klassifizierung von Bonituren des Typs Prozentbefall oder prozentuale Wirkung (z. B. 10% oder 20%) als geordnete kategoriale Daten ist im Gegensatz zu den Boniturnoten (u. a. 1, 2, 3, 4,...) nicht ganz offensichtlich, da der Abstand zwischen den Kategorien noch gut interpretierbar ist (z. B. 1%). Treten jedoch vielen Bindungen in den Daten auf und ist die Prozentskala auf weniger als 100 Schritte eingeschränkt worden, ist die Zuordnung zu geordneten kategorialen Daten eher gerechtfertigt als die Zuordnung zu den intervallskalierten Daten.

Statistische Auswertungsverfahren

Die statistische Auswertung geordneter kategorialer Daten kann mittels vieler verschiedener Methoden erfolgen. Aufgrund der im vorherigen Abschnitt definierten Ziele werden hier nur die Testverfahren vorgestellt, die das Testen von gerichteten Hypothesen ermöglichen. Chuang-Stein und Agresti [13] beschreiben u. a. folgende Möglichkeiten der Auswertung:

1. Tests auf der Grundlage von einfachen Zusammenhangsmaßen [16, S. 157ff]; z. B. Goodman und Kruskal's Gamma [21];
   diese Maße sind zwar leicht bestimmbar, eine klare Klassifizierung zwischen abhängiger und unabhängiger Variable ist meist jedoch nicht notwendig.
2. Tests, die die Wirkung als eine stetig verteilte Variable behandeln; z. B. Jonckheere-Terpstra-Test [22; 23], Mittelwertvergleiche beruhend auf Normalverteilungsannahmen [24]
   ➔ siehe Abschnitt 4.2
3. Kategoriale Regressionsmodelle; z. B. weighted least squares nach Grizzle et al. [25]
   ➔ siehe Abschnitt 4.4
4. Likelihood-Quotienten-Tests, bei denen die Wirkung als stetiges Merkmal betrachtet wird; z. B. Test von Bartholomew [26], Test von Chacko [27]
   ➔ siehe Abschnitt 4.1 und Abschnitt 5.1
5. Modellgebundene Auswertung, z. B. ordinale Regressionsmodelle [16; 28]
   ➔ siehe Abschnitt 4.6.

Die Komplexität der Annahmen und die Komplexität der Verfahren nimmt von 1. bis 5. zu. Auch bei den Tests der ersten vier Gruppen wird mit Modellen gearbeitet. Eine genaue Modellierung des funktionalen Dosis-Wirkungs-Zusammenhanges wird jedoch nur bei der modellgebundenen Auswertung (5.) angestrebt. Wenn das Modell den Zusammenhang adäquat beschreibt, können Maße für die Stärke des Zusammenhangs angeben werden, und es

können Vorhersagen für andere Dosen getroffen werden. Mit Hilfe von Tests bzgl. der zum Modell gehörenden Parameter kann ebenfalls wie bei den ersten vier Punkten geprüft werden, ob Wirkungsunterschiede zwischen den Dosen vorliegen. Das Schätzen der Modellparameter ist bei sehr kleinen Fallzahlen jedoch problematisch. So ist z. B. nur unter sehr strengen Voraussetzungen die Existenz von geeigneten Schätzern gesichert [5; 29]. Desweiteren können die sehr kleinen Fallzahlen zu Konvergenzproblemen bei den numerischen Verfahren führen [18, S.199]. Ein zusätzlicher Kritikpunkt ist die willkürliche Modellwahl. Oft ist ein Kompromiß zwischen Einfachheit und Angemessenheit des Modells notwendig. Eine optimale Strategie für die Modellwahl gibt es nicht [30, S.230ff]. Der Jonckheere-Terpstra Test ist einer der bekanntesten nichtparametrischen Tests für geordnete Alternativen [31, S.232] und wird in Kapitel 2 kurz beschrieben. Zusätzlich zu den Tests der 4. Gruppe können auch Tests für geordnete Alternativen betrachtet werden, wenn die Zielvariable als diskret angenommen wird. Zum Beispiel beschreiben Robertson und Wright [32] Tests für den Vergleich zweier Multinomialverteilungen. Patefield [33] und Agresti und Coull [34] untersuchen Trendtests für geordnete $r \times c$-Kontingenztafeln. Diese letztgenannten Verfahren haben zwar den Vorteil, daß keine Scores für die Kategorien benötigt werden, oft sind die Verteilungen der Teststatistiken aber sehr kompliziert. Die Maximierung der Likelihoodfunktion unter Ordnungsrestriktionen stellt zudem häufig ein sehr aufwendiges Problem dar. Nicht immer existieren hierfür geschlossene Lösungen. Daher werden oft numerische Verfahren zur Lösung des Problems eingesetzt. Sind schon in wenigen Stichproben einige Kategorien nicht besetzt, treten große technische Probleme auf (z. B. Division durch Null). Auf diese Punkte wird im Kapitel 4 ausführlich eingegangen.

Ein wesentliches Unterscheidungsmerkmal zwischen allen Verfahren besteht darin, ob Scores für die Zielvariablen benötigt werden oder nicht. In der Literatur wird die Wahl der Scores umfassend diskutiert. Wie aus obiger Definition ersichtlich ist, ist der eigentliche Abstand zwischen den Kategorien unbekannt. Durch die Vergabe von Scores wird aber indirekt ein Abstand eingeführt, der dann auf viele Tests einen Einfluß hat. Cochran schreibt:

„...any set of scores gives a valid test, provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive.The scores should therefor embody the best insight available about the way in which the classification was constructed and used.“ [35]

Als einfachste Variante wird die Skala durch den Anwender a priori codiert, d. h., jeder Kategorie wird ein Score (eine Zahl) zugeordnet (z. B. kein Befall =>1 = a_1, schwacher Befall =>

$2 = a_2$, mittelstarker Befall => $3 = a_3$, starker Befall => $4 = a_4$). Diese Vergabe ist aber willkürlich. Im Gegensatz zu Cochran betrachten Agresti et al.[36] die Scores als zusätzliche Parameter eines zugrundeliegenden Modells und schätzen diese aus den Daten. Agresti[37, S.294] verweist darauf, daß diese Verfahren in einigen Fällen eine schlechtere Güte besitzen können, als ein Verfahren, das a priori festgelegte Scores nutzt (z. B. aufgrund des kleineren Freiheitsgrades). Bei sehr kleinen Fallzahlen können auch Schwierigkeiten beim Schätzen der Parameter auftreten. Kimeldorf und Sampson[38] beschreiben ein Verfahren zur Bestimmung der Scores für ein Zweistichprobenproblem. Sie geben sich die Teststatistik des t-Tests vor und suchen diejenigen Scores, welche diese Statistik maximieren bzw. minimieren. Damit erhalten sie eine obere und eine untere Schranke für die Teststatistik der real vorliegenden Daten. Führt ein Vergleich der oberen Schranke mit dem Quantil der t-Verteilung (Voraussetzung ist, daß der Stichprobenumfang eine Approximation der Verteilung der Statistik durch eine t-Verteilung zuläßt) dann nicht zur Ablehnung, gilt dies auch für alle anderen Scores. Kommt es hingegen durch den Vergleich von unterer Schranke und Quantil zur Ablehnung, ergibt sich als Testergebnis, unabhängig von der Wahl der Scores, stets eine Ablehnung der Hypothese. Wenn weder das eine noch das andere gilt, d. h., wenn sowohl Scores existieren, die zur Ablehnung führen, aber auch Scores, die nicht zur Ablehnung führen, erleichtert dieses Verfahren die Wahl der Scores jedoch nicht. Die wahre Verteilung der auf den maximierenden bzw. minimierenden Scores basierenden Tests ist zudem keine einfache t-Verteilung. Das führt zu der Möglichkeit, daß zwar der auf den maximierenden Scores beruhende Test (mit der wahren Verteilung) die Nullhypothese nicht ablehnt, hingegen ein Test mit a priori festgelegten Scores (mit t-Verteilung oder Normalverteilung) zur Ablehnung führt. Diese Vorgehensweise ist daher nur sinnvoll, wenn die wahre Verteilung vernachlässigt wird. Lin und Tang[39] beschreiben ein sehr aufwendiges Verfahren zur Bestimmung optimaler Scores in einem linearen Modell. Auf die Verteilung möglicher Teststatistiken gehen sie jedoch nicht ein. Da die Scores aus den Daten bestimmt werden, sind sie Zufallsgrößen. Ihre Verteilung ist aufgrund des komplexen Berechnungsalgorithmus nur schwer handhabbar. In einer persönlichen Mitteilung schreibt Lin (1998): *„One way of using optimal scores in statistical inference without knowing its distribution is to use exact test procedure such as Patefield*." Dies führt auf Probleme, wie sie in Abschnitt 4.3 beschrieben werden.

Graubard und Korn[40] beschreiben Nachteile von Scores, die in Abhängigkeit der realen Daten erzeugt werden. Ein wesentlicher Kritikpunkt ist die Abhängigkeit der Scores von den absoluten Häufigkeiten der Kategorien im realen Datensatz. Bei Rangverfahren unterscheiden sich die Ränge für zwei benachbarte Kategorien, die jeweils nur einmal im Datensatz vorhan-

den sind, kaum. Anders sieht es bei zwei benachbarten Kategorien aus, von denen die eine sehr häufig und die andere sehr selten im Datensatz auftritt. Bei den hier betrachteten Fallzahlen und der Anzahl der Gruppen scheint dieses Argument jedoch nicht zu limitierend zu sein. Während Graubard und Korn [40] die a-priori-Vergabe von Scores empfehlen, verweisen Fleiss [41, S.83-84] und Stiger et al. [42] auf die Anwendung von Ridit- bzw. Rangscores (da zwischen den Mittelwerten der Riditscores und der Rangscores eine Bijektion besteht, wird auf Riditscores hier nicht weiter eingegangen). Zur Wahl der Scores schreiben Chuang-Stein und Agresti [13]:

„When uncertain about this choice, one should perform a sensitivity analysis, selecting two or three 'sensible' choices and checking that the conclusions are similar for each; ... Equally-spaced scores often provide a reasonable compromise when the category labels do not suggest any obvious choices, such as the response categories (worse, no change, better).“

Statt mehrere Tests mit unterschiedlichen Scores durchzuführen (hier wäre eigentlich eine Adjustierung des Signifikanzniveaus nötig) zeigen Podgor et al. [43], wie das von Gastwirth [44] vorgeschlagene Prinzip eines „Maximin Efficiency Robust Test“ (MERT) zur Generierung von Scores verwendet werden kann. Dieses Verfahren hat jedoch den Nachteil, daß eine feste Anzahl von Modellen bzw. eine ganz spezielle Modellklasse vorgegeben werden muß. Der MERT besitzt dann nur in dieser Klasse die maximale minimale Effizienz. Der kritische Punkt ist die Vorgabe der Modelle. Dies erfolgt in vielen Fällen eher willkürlich. Ähnlich sieht es bei vielen adaptiven Verfahren aus [45], bei denen die Wahl der Scores abhängig von einem a priori festgelegten Kriterium erfolgt (d. h., statt mehrere Tests durchzuführen, wird anhand der Daten ohne eine Adjustierung des Signifikanzniveaus ein Test ausgewählt; siehe Kapitel 5.3). Die Empfehlung für äquidistante Scores [13] ist auch bei Armitage [46], Graubard und Korn [40] und Chakraborti und Schaafsma [47] zu finden. StatXact [19] benutzt als Defaultscores bei zweifach geordneten Kontingenztafeln ebenfalls äquidistante Scores.

Zu den Problemen Modellwahl (Wahl der Teststatistik) und Wahl von Scores kommt bei kleinen Fallzahlen noch das Problem der Wahl eines adäquaten Verteilungsmodells hinzu. Viele der zu den Punkten 1. bis 5. gehörenden Teststatistiken sind mit stetigen Verteilungsfunktionen verknüpft. Diese Verteilungsaussagen sind zum Teil exakt und zum Teil nur asymptotisch gültig. Wie robust die Verfahren gegen Abweichung von Modellannahmen sind und ab wann eine approximative Verteilung eine gute Annäherung an die wahre Verteilung ist, kann selten explizit ausgedrückt werden. Einige Autoren geben Faustregeln für die Anwendung asymptotischer Verfahren in Abhängigkeit von den Stichprobenumfängen [18, S.305;48] an. Andere untersuchen mittels Simulationen die Güte parametrischer oder asymptotischer

Verfahren [49; 50; 51; 52]. Basierend auf den Ergebnissen geben sie Empfehlungen für die Anwendbarkeit der Verfahren.

Die Bestimmung der exakten Verteilung mit Hilfe von Permutationsargumenten stellt hier eine geeignete Alternative dar. Der Vorteil von exakten Permutationstests liegt gerade darin, daß sie ein breites Einsatzgebiet besitzen und nur auf wenigen Voraussetzungen basieren [53, S.2]. Hervorzuheben ist, daß die Verteilungsaussagen sowohl für kleine als auch für große Stichprobenumfänge gelten. Im Gegensatz zu asymptotischen Tests überschreiten exakte Permutationstests ein vorgegebenes Signifikanzniveau nicht. Aber auch Permutationstests besitzen Anwendungsgrenzen bei sehr kleinen Fallzahlen. Dies ist z. B. der Fall, wenn die Anzahl aller möglichen Permutationen so klein ist, daß theoretisch kein Unterschreiten eines vorgegebenen Signifikanzniveaus möglich ist. Aufgrund der oft sehr diskreten Verteilung sind Permutationstests zudem häufig sehr konservativ (schöpfen das vorgegebene Signifikanzniveau nicht vollständig aus). Desweiteren steigt der rechentechnische Aufwand bei wachsenden Fallzahlen schnell an. Modifikationen können dann zum Teil zwar Abhilfe schaffen, dies geht aber in der Regel auf Kosten der Exaktheit [19, S.50]. Aufgrund der häufig extremen Konservativität empfehlen einige Autoren daher robuste asymptotische Tests[54].

Analog zu Permutationstests kann mit Bootstrapverfahren das Problem der unbekannten Verteilung gelöst werden. Allerdings sind diese Verteilungsaussagen oft nur asymptotisch gültig. Der Vergleich von parametrischen, asymptotischen bzw. permutativen Verfahren und Bootstrapverfahren erfolgt in zahlreichen Büchern und Artikeln gerade für stetige Daten umfangreich, z. B. [53; 55; 56]. In den zitierten Büchern wird gleichzeitig ein guter Einblick in Permutations- und Bootstrapverfahren gegeben. Für diskrete Daten (ordinale Daten) existiert deutlich weniger Literatur, z. B. [57; 58].

Angesichts der kleinen Fallzahlen sind für die Auswertung auch noch asymptotische Verfahren höherer Ordnung (small sample asymptotic [59]) interessant. Booth et al. [60] beschreiben z. B. Edgeworth- und Sattelpunkt-Approximationen für Summen diskreter Variablen und zeigen, daß diese Approximationen zum Teil besser sind als die Normalverteilungsapproximation. Agresti et al. [61] zeigen anhand von Simulationen, daß Sattelpunkt-Approximationen eine exzellente Approximation der exakten Verteilung darstellen können. Die „forbidding mathematics“ und „lack of software“ sieht Brazzale [62] als Hauptursachen dafür, daß diese Verfahren selten in der Praxis genutzt werden. Auch in der vorliegenden Arbeit sollen diese Verfahren nicht betrachtet werden.

Für die Anwendung von Bayesverfahren zur Auswertung ordinaler Daten sei auf Evans et al. [63] verwiesen.

1.3 Gliederung der Arbeit

Im folgenden werden die Verfahren vorgestellt, die dann in umfangreichen Simulationen hinsichtlich ihres Verhaltens bei sehr kleinen Fallzahlen untersucht werden. Vor der Analyse konkreter Tests werden in Kapitel 2 und 3 allgemeine Konzepte für Permutations- bzw. Bootstraptests vorgestellt. Das Verhalten der in Kapitel 4 und 5 vorgestellten parametrischen und nichtparametrischen Tests wird im Kapitel 6 mit den analogen Permutations- bzw. Bootstraptests verglichen. Ordinale Regressionsmodelle sind zwar schwer interpretierbar[20] und mit technischen Schwierigkeiten bei kleinen Fallzahlen verbunden; da sie sonst jedoch ein gutes und oft beschriebenes Werkzeug zur Auswertung ordinaler Daten darstellen, wird in Abschnitt 4.6 kurz auf sie eingegangen. In Kapitel 7 und 8 folgt schließlich die Anwendung der vorgestellten Tests zur Lösung ausgewählter Versuchsfragen und Beispiele.

Simulationen

Simulationen werden häufig, so auch in der vorliegenden Arbeit, genutzt, um das Gütevorhalten von Tests sowohl unter der Nullhypothese als auch unter der Alternativhypothese zu untersuchen. Die im Rahmen der vorliegenden Arbeit erstellten Simulationsprogramme wurden in den Programmiersprachen IML (SAS 6.12) bzw. C geschrieben. Da Bootstrap- und Permutationstests sehr rechen- und zeitintensiv sind, wurde ein Großteil der Simulationen auf einem Parallelrechner ausgeführt. Hierzu wurde dann die Programmiersprache C und die Parallelsoftware MPI genutzt. Die Nutzung eines Parallelrechners zur Berechnung statistischer Probleme und speziell von Resamplingverfahren wird u. a. bei Adams et al.[64] und Kaufman et al. [65] beschrieben.

Bei diskreten Daten halten viele Tests ein vorgegebenes Signifikanzniveau in der Regel nicht exakt ein. Ein Vergleich ist daher theoretisch nicht korrekt. Trotzdem werden Simulationen sehr häufig zur Bewertung von Testverfahren eingesetzt. Die Güte der einzelnen Tests unter der Alternativhypothese wird jedoch nur dann miteinander verglichen, wenn kein Test das Signifikanzniveau unter der Nullhypothese deutlich überschreitet. Schöpft ein Test das Signifikanzniveau schlecht aus, so wird sich dies auch auf die Güte unter der Alternativhypothese auswirken. Der Test wird im Vergleich demnach eher schlecht abschneiden. Asymptotische Güteaussagen zum Vergleich der untersuchten Tests sind zwar zum Teil möglich, aber bei den betrachteten Fallzahlen erscheint ein asymptotischer Vergleich nicht als das Mittel der Wahl [19, S.12].

Als ein asymptotisches $(1-\alpha)$-Konfidenzintervall für die Güte eines Tests (Π) kann

$$\hat{\Pi} \pm z_{1-\alpha/2} \sqrt{\Pi(1-\Pi)/M}$$

genutzt werden [66] ($M =$ Simulationsanzahl, $z_{1-\alpha} = (1-\alpha) -$ Quantil der Standardnormalverteilung). Dieses Konfidenzintervall beruht auf dem zentralen Grenzwertsatz für binomialverteilte Zufallsvariablen und erreicht seine größte Breite, wenn die wahre Güte den Wert 0,5 annimmt. Cochran [67] bezeichnet einen Test als robust, wenn sein aktuelles Signifikanzniveau das nominale Signifikanzniveau nicht um 20% übersteigt. Heeren und D'Agostino[49] fordern sogar, daß diese Differenz kleiner 10% sei. Bei einer Irrtumswahrscheinlichkeit von 0,99 für das Konfidenzintervall sind 10.000 Simulationsschritte demnach für kleine Signifikanzniveaus zu wenig. Trotzdem wurde die Anzahl der Simulationsläufe aus Zeitgründen auf 10.000 beschränkt. Bei der Diskussion der Niveaueinhaltung der einzelnen Tests muß dies berücksichtigt werden. Für die Simulation der Güte unter der Alternativhypothese sind 10.000 Simulationsläufe ausreichend, da die maximale Abweichung vernachlässigbar ist (siehe Tabelle 1.5).

wahre Güte	**Simulationsanzahl**	**0.95-Konfidenz-intervall**		**0.99-Konfidenz-intervall**	
		halbe Intervall-breite	maximale Ab-weichung in %	halbe Intervall-breite	maximale Ab-weichung in %
0,01	1.000	0,0062	61,669	0,0081	81,047
0,05	1.000	0,0135	27,016	0,0178	35,505
0,10	1.000	0,0186	18,594	0,0244	24,436
0,50	1.000	0,0310	6,198	0,0407	8,146
0,80	1.000	0,0248	3,099	0,0326	4,073
0,01	10.000	0,0020	19,50	0,0026	25,629
0,05	10.000	0,0043	8,543	0,0056	11,228
0,10	10.000	0,0059	5,880	0,0077	7,727
0,50	10.000	0,0098	1,960	0,0129	2,576
0,80	10.000	0,0078	0,980	0,0103	1,288

Tabelle 1.5: Beispiele für die Intervallbreite und die maximale Abweichung (in %) in Abhängigkeit von wahrer Güte, Simulationsanzahl und Irrtumswahrscheinlichkeit

2 Permutationstests

2.1 Exakte bedingte Permutationstests

Was ist ein Permutationstest? Was ist ein Randomisationstest? Wo liegt der Unterschied zwischen beiden (oder gibt es gar keinen)? Auf diese Fragen geben verschiedene Autoren unterschiedliche Antworten. So schreibt Gibbons [68, S.690]: *„Permutation tests are a special kind of randomization tests."* In derselben Reihe, nur in Band 7 statt 6, schreibt Edgington [69, S.530]: *„A randomization test is a permutation test which based on randomization,..."*. Kempthorne [70, S.524] weist auf einen deutlichen Unterschied zwischen beiden hin: *„The distinction between randomization tests and permutation tests is important. The latter based on the assumption of random sampling, an assumption that is often patently false or unverifiable, ..."*. Manly [71, S.3] schreibt zu diesem Thema: *„Randomisation tests are most easily justified if either the samples being analysed are random or the experimental design itself justifies randomization testing. This has led some authors to use the description 'permutation test' for situations where random samples justify the calculations, and 'randomization tests' for situations where the experimental design provides the justification."*
Technisch besteht kein Unterschied zwischen Permutationstests und Randomisationstests, d. h., es wird dieselbe Teststatistik und dieselbe Wahrscheinlichkeitsverteilung genutzt. Für den Vergleich zweier Versuchsgruppen betrachtet Lehmann [72, S. 64] z. B. zum einen ein Randomisationsmodell ($N = m + n$ Objekte sind gegeben, wobei m Objekte der Behandlung 1 und n Objekte der Behandlung 2 zufällig zugeordnet werden.) und zum anderen ein Populationsmodell ($N = m + n$ Objekte werden zufällig aus zwei Populationen gezogen). Um Unterschiede zwischen den Versuchsgruppen aufzudecken, kann jeweils z. B. derselbe asymptotische Rangtest genutzt werden. Die Bezeichnung des Tests erfolgt hier jedoch unabhängig vom betrachteten Modell. Gegen den Begriff Randomisationstest spricht die leichte Verwechslungsgefahr mit sogenannten randomisierten Tests, deren Testentscheidung von einem zusätzlichen Randomisationsschritt abhängt. Die Bezeichnung „Rerandomisationstest", wie sie z. B. von Petrondas und Gabriel [73] benutzt wird, drückt den Sachverhalt klarer aus. In der vorliegenden Arbeit wird der Begriff Permutationstest für alle Tests benutzt, dessen p-Wert (Quantil) auf dem Permutieren der Daten beruht.
Das Prinzip eines Permutationstests ist recht einfach. Die Durchführung erfolgt stets mittels folgender Schritte [53]:

1. Analyse des Problems und Festlegen eines Hypothesenpaares
2. Wahl einer für das Hypothesenpaar geeigneten Teststatistik
3. Berechnung der Teststatistik für die Originalwerte
4. Erzeugen aller dem Randomisationsschema bzw. der Nullhypothese entsprechenden Permutationen der Daten
5. Berechnung der Teststatistiken für jede erzeugte Permutation
6. Berechnung des p-Wertes mittels der erzeugten Permutationsverteilung.

Folgendes fiktives Beispiel verdeutliche dieses Vorgehen:
Angenommen, es wurde ein Feldversuch mit vier ansteigenden Dosen eines Präparats als balancierte vollständig randomisierte Anlage mit 3 Wiederholungen durchgeführt. Mit Hilfe des Jonckheere-Terpstra-Tests [22; 23] soll die Nullhypothese „kein Behandlungsunterschied“ gegen die Alternativhypothese „Wirkungsanstieg“ getestet werden. Die Teststatistik des Jonckheere-Terpstra-Tests hat für Daten, in denen Bindungen auftreten, folgende Form:

$$J = \sum_{i=0}^{k-1} \sum_{j=i+1}^{k} U_{ij} \quad \text{mit} \quad U_{ij} = \sum_{s=1}^{n_i} \sum_{t=1}^{n_j} \phi(X_{is}, X_{jt}) \text{ und}$$

$$\phi(X_{is}, X_{jt}) = \begin{cases} 1 & \text{falls } X_{is} < X_{jt} \\ 0.5 & \text{falls } X_{is} = X_{jt} \\ 0 & \text{falls } X_{is} > X_{jt}. \end{cases}$$

Eine Standardisierung der Statistik J (wie bei asymptotischen Tests) ist überflüssig, da sowohl der Erwartungswertschätzer als auch der Varianzschätzer permutationsinvariant sind. Die folgenden Ergebnisse (Tabelle 2.1) seien beobachtet worden.

Dosis	D_0	D_0	D_0	D_1	D_1	D_1	D_2	D_2	D_2	D_3	D_3	D_3
Parzelle	8	4	10	5	12	2	11	1	7	3	6	9
Wert	9	3	4	1	7	11	2	5	12	6	8	10

Tabelle 2.1: Beispielversuch ohne Bindungen

Für diese Aufteilung beträgt der Wert der Teststatistik $J(\boldsymbol{x}) = J(x_{01}, \ldots, x_{kn_k}) = j_0 = 33$. Hat die Dosishöhe keinen Einfluß, so sollten die Werte auf den Parzellen invariant bezüglich eines

Vertauschens der Dosen (d. h., einer anderen zufälligen Zuordnung) sein. Die Teststatistik nimmt für die in Tabelle 2.2 dargestellten zufälligen Anordnung den Wert 32 an.

Dosis	D_0	D_0	D_0	D_1	D_1	D_1	D_2	D_2	D_2	D_3	D_3	D_3
Parzelle	5	12	2	8	4	10	11	1	7	3	6	9
Wert	1	7	11	9	3	4	2	5	12	6	8	10

Tabelle 2.2: Daten nach Permutieren der Originaldaten aus Tabelle 2.1

Insgesamt existieren $N!$ verschiedene Zuordnungen, die durch Permutieren der Daten erzeugt werden können. Da Vertauschungen innerhalb der Gruppen keinen Einfluß auf die Teststatistik *J* haben, brauchen „nur“ die

$$\binom{N}{n_0\, n_1\, n_2\, n_3} = \frac{(n_0+n_1+n_2+n_3)!}{n_0!\, n_1!\, n_2!\, n_3!} = \frac{12!}{3!\ 3!\ 3!3!} = 369.600$$

Zuordnungen (Permutationen) und die zugehörigen Werte der Teststatistik bestimmt werden. Weil ein extrem großer Wert für j_0 gegen die Nullhypothese sprechen würde, wird anhand der Häufigkeitsverteilung der einzelnen Werte (der Wertebereich der Teststatistik sind in diesem Beispiel die ganzen Zahlen zwischen 0 und 54) die Wahrscheinlichkeit bestimmt, einen Wert größer als j_0 anzunehmen. Allgemein würde gelten:

$$p = P(J \geq j_0) = \sum_{\{\boldsymbol{x}: J(\boldsymbol{x}) \geq j_0\}} \frac{n_0!\, n_1!\, n_2!\, n_3!}{(n_0+n_1+n_2+n_3)!},$$

wobei die Summation über alle zulässigen Aufteilungen läuft, die zu einem Wert für*J* führen, der größer oder gleich j_0 ist. In diesem einfachen Beispiel (es treten keine gleichen Werte auf) kann dies durch

$$p = \sum_{j=33}^{54} P(J = j \mid H_0) = 0{,}222$$

ausgedrückt werden. Die Nullhypothese könnte also nicht abgelehnt werden. Würden statt der drei Wiederholungen vier (fünf) Wiederholungen im Versuch angelegt, würden 63.063.000 (11.732.745.024) verschiedene Zuordnungen existieren. Ein Problem liegt also im Finden

aller möglichen Aufteilungen. Treten innerhalb der beobachteten Daten keine Bindungen (gleiche Werte) auf, können wie bei einer parametrischen Verteilung die Quantile $j_{1-\alpha} = \min(j: P(J \geq j|H_0) \leq \alpha)$ in Abhängigkeit der Stichprobenumfänge für J vertafelt werden. Für obiges Beispiel und $\alpha = 0{,}05$ wäre $j_{1-\alpha} = 40$, wobei $P(J \geq j_{1-\alpha}|H_0) = 0{,}0374$ gelten würde. Das aktuelle Signifikanzniveau des Tests wäre also deutlich kleiner als das vorgegebene Niveau. Dies ist eine bekannte Eigenschaft von Permutationstests und wird im allgemeinen auch als Konservativität bezeichnet. Die Ursache dieser Konservativität liegt in der Diskretheit der Wahrscheinlichkeitsverteilung der Teststatistik, die besonders bei kleinen Fallzahlen zum Tragen kommt. Treten Bindungen in den beobachteten Werten auf, reduziert sich die Anzahl der verschiedenen Aufteilungen zum Teil beträchtlich. Dies heißt im allgemeinen aber nicht, daß die Verteilung der Teststatistik auch weniger Werte annähme. Würden statt der Werte 6 und 10 bei D_3 jeweils der Wert 8 beobachtet, so würde zwar der Mittelwert der Gruppe gleich bleiben, die Verteilung der Jonckheere-Terpstra Statistik würde aber auf 103 Punkte konzentriert sein (d. h. fast doppelt so viele). Für $\alpha = 0{,}05$ gilt: $j_{1-\alpha} = 39$ und $P(J \geq j_{1-\alpha}|H_0) = 0{,}0464$, d. h., es ist sogar eine bessere Niveauausschöpfung möglich. Allgemein wird mit steigender Anzahl an Bindungen die Verteilung der Teststatistik auf weniger Punkte konzentriert sein. Treten viele Bindungen auf, können die Daten auch mittels einer Kontingenztafel dargestellt werden.

Dosis	D_0	D_0	D_0	D_1	D_1	D_1	D_2	D_2	D_2	D_3	D_3	D_3
Parzelle	8	4	10	5	12	2	11	1	7	3	6	9
Wert	3	3	3	7	8	8	7	8	10	8	10	10

Tabelle 2.3: Beispielversuch mit vielen Bindungen

Die in Tabelle 2.3 gegebenen Werte können z. B. durch die in Tabelle 2.4 dargestellte Kontingenztafel beschrieben werden. Die Daten je Dosis i werden ohne Informationsverlust auf einen Häufigkeitsvektor (Zeile i) reduziert. Kontingenztafeln eignen sich für die komprimierte Darstellung von mehreren Merkmalen, die nur wenige Ausprägungen besitzen. Dabei können durchaus mehrere zufällige Merkmale dargestellt werden, wobei nicht zwischen abhängigen und unabhängigen Merkmalen (Variablen) unterschieden werden muß. Die Zeilen- bzw. Randsummen können sich jeweils zufällig ergeben. Wird ein Dosis-Wirkungs-Versuch in eine Kontingenztafel überführt, so existiert eine klare Trennung zwischen abhängigen und unabhängigen Variablen.

	Wert				
Dosis	3	7	8	10	Randsummen
D_0	3	0	0	0	$3 = n_0$
D_1	0	1	2	0	$3 = n_1$
D_2	0	1	1	1	$3 = n_2$
D_3	0	0	1	2	$3 = n_3$
Randsummen	$3 = t_1$	$2 = t_2$	$4 = t_3$	$3 = t_4$	$12 = N$

Tabelle 2.4: Originalkontingenztafel für die Beispieldaten aus Tabelle 2.3

Wird, wie in Tabelle 2.4, das Merkmal Dosis durch die Zeilen und das Merkmal Wirkung, repräsentiert durch die verschiedenen Ausprägungen, mittels der Spalten dargestellt, so sind die Zeilensummen fest durch den Versuch (Stichprobenumfänge) vorgeben. Die Spaltensummen ergeben sich zufällig. Viele Verfahren und Techniken, die für die Auswertung von Kontingenztafeln konstruiert wurden, können zur Auswertung von Dosis-Wirkungs-Versuchen genutzt werden. Das Finden aller verschiedenen Aufteilungen (Permutationen) der Daten für einen Permutationstest ist demnach äquivalent zum Finden aller verschiedenen Tafeln mit denselben Randsummen wie die Originaltafel. Wird mit

$$\Gamma_{\boldsymbol{n},\boldsymbol{t}} = \left\{ \boldsymbol{z}: \sum_{i=0}^{3} z_{ij} = t_j, \sum_{j=1}^{4} z_{ij} = n_i, i = 0,\ldots,3, j = 1,\ldots,4 \right\}$$

die Menge aller 4×4-Tafeln mit denselben Randsummen wie die Originaltafel bezeichnet und mit $\boldsymbol{Z}$ eine Zufallsvariable, die Werte in dieser Menge annimmt, so kann der p-Wert des permutativen Jonckheere-Terpstra-Tests durch folgende Summe dargestellt werden:

$$p = \sum_{\{z: z \in \Gamma_{\boldsymbol{n},\boldsymbol{t}}, J(z) \geq j_0\}} P(\boldsymbol{Z} = \boldsymbol{z} | H_0, \boldsymbol{n}, \boldsymbol{t}) = \sum_{\{z: z \in \Gamma_{\boldsymbol{n},\boldsymbol{t}}, J(z) \geq j_0\}} \frac{\prod_{j=1}^{4} \binom{t_j}{z_{1j}\, z_{2j}\, z_{3j}\, z_{4j}}}{\binom{N}{n_0\, n_1\, n_2\, n_3}}.$$

Ein allgemeiner Zugang zu diesem p-Wert führt über den Vergleich von Multinomialverteilungen. Sind die Stichproben, repräsentiert durch ihre Häufigkeitsvektoren

$\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{ir})'$, $i = 0,\ldots,k$, unabhängig und jeweils nach $M(n_i,\pi_{i1},\ldots,\pi_{ir})$ verteilt, so ist die gemeinsame Verteilung der $k+1$ Stichproben durch

$$P(Y_{01} = y_{01},\ldots,Y_{kr} = y_{kr}) = \prod_{i=0}^{k} \binom{n_i}{y_{i1}\ldots y_{ir}} \pi_{i1}^{y_{i1}} \ldots \pi_{ir}^{y_{ir}}$$

gegeben. Diese Dichte kann durch folgende Umformungen in Exponentialgestalt gebracht werden:

$$\begin{aligned}
&P(\boldsymbol{Y} = \boldsymbol{y}) = P(Y_{01} = y_{01},\ldots,Y_{kr} = y_{kr}) \\
&= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r}\sum_{i=0}^{k} y_{ij} \ln(\pi_{ij})\Big) \quad \text{mit} \quad h_i(y_i) = \binom{n_i}{y_{i1}\ldots y_{ir}} \\
&= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r}\sum_{i=0}^{k-1} y_{ij} \ln(\pi_{ij}) + \sum_{j=1}^{r} y_{kj} \ln(\pi_{kj})\Big) \\
&= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r}\sum_{i=0}^{k-1} y_{ij} \ln(\pi_{ij}) + \sum_{j=1}^{r} (y_{\cdot j} - \sum_{i=0}^{k-1} y_{ij}) \ln(\pi_{kj})\Big) \\
&= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r}\sum_{i=0}^{k-1} y_{ij} \ln(\pi_{ij}/\pi_{kj}) + \sum_{j=1}^{r} y_{\cdot j} \ln(\pi_{kj})\Big) \\
&\text{durch analoge Abspaltungen führt dies zu} \\
&= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r-1}\sum_{i=0}^{k-1} y_{ij} \ln\Big(\frac{\pi_{ij}\pi_{kr}}{\pi_{kj}\pi_{ir}}\Big) + \sum_{j=1}^{r-1} y_{\cdot j} \ln\Big(\frac{\pi_{kj}}{\pi_{kr}}\Big) + \sum_{i=0}^{k-1} y_{i\cdot} \ln\Big(\frac{\pi_{ir}}{\pi_{kr}}\Big) + y_{\cdot\cdot} \ln(\pi_{kr})\Big).
\end{aligned}$$

Falls die Nullhypothese $\pi_j = \pi_{0j} = \cdots = \pi_{kj}$ für $j = 1,\ldots,r$ gilt, vereinfacht sich die Formel zu

$$\begin{aligned}
P(\boldsymbol{Y} = \boldsymbol{y}|H_0) &= \prod_{i=0}^{k} h_i(y_i) \exp\Big(\sum_{j=1}^{r-1} y_{\cdot j} \ln\Big(\frac{\pi_j}{\pi_r}\Big) + y_{\cdot\cdot} \ln(\pi_r)\Big) \\
&= \pi_1^{y_{\cdot 1}} \ldots \pi_r^{y_{\cdot r}} \prod_{i=0}^{k} \binom{n_i}{y_{i1}\cdots y_{ir}}.
\end{aligned}$$

Aufgrund der Exponentialgestalt ist $\boldsymbol{T} = (T_j = \sum_{i=0}^{k} Y_{ij})_{j=1,\ldots,r}$ eine suffiziente Statistik für diese Verteilungsfamilie (d. h., $\boldsymbol{T}$ schöpft alle Informationen zu $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_r)'$, die in der Stichprobe $\boldsymbol{Y} = (\boldsymbol{Y}_0',\ldots,\boldsymbol{Y}_k')'$ enthalten sind, aus). Die bedingte Verteilung von $\boldsymbol{Y}$ hängt zudem bei festem $\boldsymbol{T} = \boldsymbol{t}$ nicht vom Parametervektor $\boldsymbol{\pi}$ ab. Zur Elimination der selbst unter der Nullhypothese unbekannten Parameter $\pi = (\pi_1,\ldots,\pi_r)'$ bietet es sich daher an, dieses Problem bedingt auf T zu betrachten. Nach Lehmann [72, S.384] ist die Verteilung von $\boldsymbol{T}$ eine

Multinomialverteilung $M(N,\pi_1,...,\pi_r)$, so daß die bedingte Verteilung von folgender Gestalt ist:

$$P(\boldsymbol{Y}=\boldsymbol{y}|H_0,\boldsymbol{n},\boldsymbol{t}) = P(Y_{01}=y_{11},...,Y_{kr}=y_{kr}|H_0,\boldsymbol{n},\boldsymbol{t})$$
$$= \frac{\prod_{i=0}^{k}\binom{n_i}{y_{i1}\dots y_{ir}}\pi_1^{y_{i1}}\dots\pi_r^{y_{ir}}}{\binom{N}{t_1\dots t_r}\pi_1^{t_1}\dots.\pi_r^{t_r}} = \frac{\prod_{i=0}^{k}\binom{n_i}{y_{i1}\dots y_{ir}}}{\binom{N}{t_1\dots t_r}}.$$

Für eine Statistik $S = S(Y)$ ist die bedingte Verteilung durch

$$P(S(\boldsymbol{y})=s|H_0,\boldsymbol{n},\boldsymbol{t}) = \sum_{\{z:z\in\Gamma_{\boldsymbol{n},\boldsymbol{t}},S(z)=s\}} \frac{\prod_{i=0}^{k}\binom{n_i}{z_{i1}\dots z_{ir}}}{\binom{N}{t_1\dots t_r}} \qquad (2.1)$$

mit $\Gamma_{\boldsymbol{n},\boldsymbol{t}} = \{\boldsymbol{z}:\sum_{i=0}^{k} z_{ij} = t_j, \sum_{j=1}^{r} z_{ij} = n_i, i=0,\dots,k, j=1,\dots,r\}$ definiert. Die Wahrscheinlichkeit, größer oder gleich einem beobachteten Wert zu sein, ergibt sich gerade, indem in obiger Gleichung $S(y)=s$ und $S(z)=s$ durch $S(y)\geq s$ bzw. $S(z)\geq s$ ersetzt wird. Wird mit r die Anzahl der verschieden Werte in der gepoolten Stichprobe bezeichnet, kann jeder Permutationstest für einfaktorielle Anlagen auf diese Weise auf ein Tafelproblem zurückgeführt und als exakter bedingter Permutationstest bezeichnet werden.

Worin besteht nun der Vorteil der Betrachtungsweise mit Hilfe von Kontingenztafeln?

Wie schon angedeutet, ist die Anzahl der möglichen Permutationen sehr groß. Eine systematische Aufzählung aller Permutationen ist daher kein rationelles Verfahren. Effiziente Algorithmen zur Bestimmung der gesamten exakten Permutationsverteilung bzw. zur Bestimmung eines p-Wertes sind daher notwendig. Einen guten Überblick dazu geben Verbeck und Kroonenberg [74]. Mehta und Patel [75] beschreiben einen effizienten Algorithmus (Netzwerkalgorithmus) zur Bestimmung aller verschiedenen Tafeln bzw. eines p-Wertes. Dieser Algorithmus beruht auf der Idee, jede Tafel als einen Pfad eines Netzwerkes (Menge von Knoten und Kanten) anzusehen. In den Knoten werden die Häufikeiten der r verschiedenen Werte gespeichert, die noch zur Aufteilung zur Verfügung stehen. Im Startknoten werden somit die Häufigkeiten der r verschiedenen Werte der gepoolten Stichprobe gespeichert. Ausgehend von einem Knoten an dem schon i Spalten festgelegt sind, kann ein nachfolgender Knoten generiert werden, indem von den noch zur Verfügung stehenden Werten, die Werte

abgezogen werden, die der (i+1)-ten Spalten zugeordnet werden. Die Tafel aus Tabelle 2.4 sowie zwei zusätzliche Tafeln sind beispielhaft in Abbildung 2.1 dargestellt.

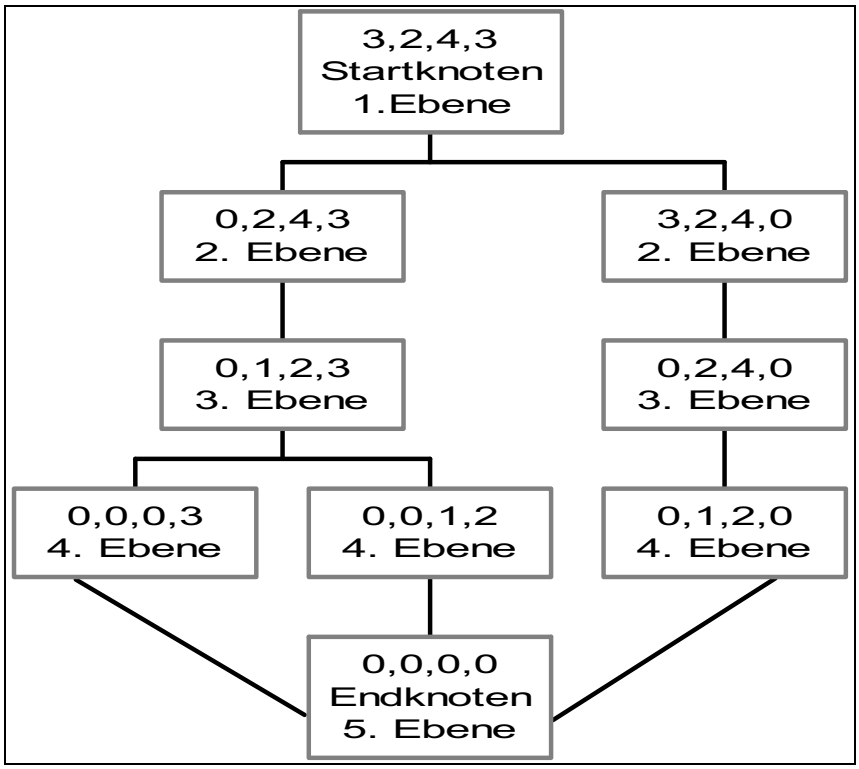


Abbildung 2.1: Ausschnitt aus einem auf den Daten der Tabelle 2.4 beruhenden Netzwerks

Beginnend beim Startknoten (1. Ebene) wird das Netzwerk entlang der Kanten zum Endknoten (5. Ebene) durchlaufen. Die 1. Zeile der Tafel aus Tabelle 2.4 ergibt sich aus der komponentenweisen Subtraktion der Knoten: (3,2,4,3)-(0,2,4,3)=(3,0,0,0). Die 2. Zeile ergibt sich analog aus: (0,2,4,3)-(0,1,2,3)=(0,1,2,0) usw. Tafeln, die zum Teil aus gleichen Zeilen bestehen, besitzen auch einen gemeinsamen Teil eines Pfades. Dies wirkt sich positiv auf den Speicherbedarf aus. Soll nur ein p-Wert berechnet werden, muß nur ein Pfad gleichzeitig im Speicher gehalten werden. Der Speicherbedarf ist im Gegensatz zur Erzeugung der gesamten Verteilung daher sehr gering. Ist der Anteil einer Zeile i an der Gesamtstatistik nur von den Zeilen 0 bis i (Bedingung 1) abhängig und kann der Anteil der restlichen $k-i$ Zeilen durch eine untere und obere Schranke abgeschätzt werden (Bedingung 2), kann zusätzlich der Rechenaufwand minimiert werden. Am Beispiel des Jonckheere-Terpstra-Tests (Test auf Anstieg) kann dies leicht verdeutlicht werden. Im ersten Schritt wird die erste Zeile (z_0) festgelegt. Für diese Zeile wird ihr Anteil an der Gesamtstatistik (s_0) bestimmt. Der Anteil der restlichen Zeilen, unabhängig davon, wie die weitere Aufteilung der Daten auf die restlichen Zeilen erfolgt, wird durch eine obere Schranke ($s^o_{1,\ldots,k}$) und eine untere Schranke ($s^u_{1,\ldots,k}$) abge-

schätzt. Für den Jonckheere-Terpstra-Test kann die obere Schranke bestimmt werden, indem die noch nicht zugeordneten Daten der Größe nach aufsteigend sortiert werden und die ersten n_1 der 2. Zeile, die nächsten n_2 der 3. Zeile usw. zugeordnet werden. Ist j_0 der beobachtete Wert und gilt $s_0 + s^o_{1,\ldots,k} < j_0$, so führt keine Tafel, deren 1. Zeile durch $\boldsymbol{z}_0$ gegeben ist, zu einem Wert, der größer als j_0 ist. Keine dieser Tafeln hat daher einen Anteil am p-Wert und muß daher nicht weiter bestimmt werden. Die untere Schranke $s^u_{1,\ldots,k}$ kann bestimmt werden, indem die Daten der Größe nach absteigend sortiert und zugeordnet werden. Gilt $s_0 + s^u_{1,\ldots,k} > j_0$, führen alle Tafeln, deren 1. Zeile durch $\boldsymbol{z}_0$ gegeben ist, zu einem Wert, der größer als j_0 ist. Sie haben demnach alle einen Anteil am p-Wert (2.1). Die Wahrscheinlichkeit all dieser Tafeln kann durch einen geschlossenen Ausdruck in Abhängigkeit von $\boldsymbol{z}_0$ (allgemein auch für alle bereits bestimmten Zeilen) und den Randsummen berechnet werden [75]. Gilt eine der Relationen, wird eine neue 1. Zeile erzeugt. Gilt weder die eine noch die andere Relation, wird eine 2. Zeile ($\boldsymbol{z}_1$) erzeugt und eine obere Schranke ($s^o_{2,\ldots,k}$) und eine untere Schranke ($s^u_{2,\ldots,k}$) für den Anteil der restlichen $k-1$ Zeilen bestimmt. Gilt $s_0 + s_1 + s^u_{2,\ldots,k} > j_0$, wird der Anteil am p-Wert für alle Tafeln, deren erste zwei Zeilen durch $\boldsymbol{z}_0$ und $\boldsymbol{z}_1$ gegeben sind, und eine neue 2. Zeile bestimmt. Gilt $s_0 + s_1 + s^u_{2,\ldots,k} < j_0$, wird ebenfalls eine neue 2. Zeile erzeugt. Gilt keine von beiden Relationen, wird eine 3. Zeile erzeugt. Ist bei gegebener 1. Zeile $\boldsymbol{z}_0$ die Bildung einer 2. von $\boldsymbol{z}_1$ verschiedenen Zeile nicht mehr möglich, wird eine neue 1. Zeile bestimmt. Das Vorgehen kann analog auf eine beliebige Zeile *i* fortgesetzt werden. Ist mittels der Schranken keine Entscheidung möglich, wird eine weitere Zeile erzeugt. Ansonsten wird zur nächstniedrigeren Ebene, auf der noch eine neue Zeile konstruierbar ist, zurückgesprungen. Alle notwendigen Tafeln sind erzeugt worden, wenn keine neue 1. Zeile mehr gebildet werden kann. Ein Großteil von (linearen) Rang- bzw. Scorestatistiken erfüllen Bedingung 1 und 2. Schwierig ist allerdings meist das Finden einer oberen und unteren Schranke, die auch in wenigen Schritten berechnet werden kann. Nicht immer ist die minimale obere bzw. die maximale untere Schranke die beste Wahl. Einfach berechenbare Schranken können durchaus schneller zum Ziel führen, auch wenn eventuell mehrere Tafeln erzeugt werden müssen [76].

Die Rechenzeit kann selbst bei diesem Verfahren schnell ansteigen. Sie hängt nicht nur von den Fallzahlen, sondern auch von der Struktur der Tafel und vom Rechenaufwand zur Bestimmung der Schranken und der Statistiken ab.

Wird mit $s_{\boldsymbol{n},\boldsymbol{t},1-\alpha} = \min(s: P(S \geq s|H_0,\boldsymbol{n},\boldsymbol{t}) \leq \alpha)$ das von $\boldsymbol{t}$ und $\boldsymbol{n}$ abhängige $1-\alpha$-Quantil bezeichnet, so kann die Güte dieses bedingten Tests unter der Nullhypothese bzw. unter der Alternativhypothese durch

$$P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0,\boldsymbol{n},\boldsymbol{t}) = \sum_{\{z:z\in\Gamma_{\boldsymbol{n},\boldsymbol{t}},S(z)\geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}\}} \frac{\prod_{i=0}^{k}\binom{n_i}{z_{i1}\ldots z_{ir}}}{\binom{N}{t_1\ldots t_r}}$$

bzw.

$$P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_A,\boldsymbol{n},\boldsymbol{t}) = \sum_{\{\boldsymbol{u}:\boldsymbol{u}\in\Gamma_{\boldsymbol{n},\boldsymbol{t}},S(\boldsymbol{u})\geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}\}} \frac{\prod_{i=0}^{k}\binom{n_i}{u_{i1}\ldots u_{ir}}\pi_{i1}^{u_{i1}}\ldots\pi_{ir}^{u_{ir}}}{\sum_{\{z:z\in\Gamma_{\boldsymbol{n},\boldsymbol{t}}\}}\prod_{i=0}^{k}\binom{n_i}{z_{i1}\ldots z_{ir}}\pi_{i1}^{z_{i1}}\ldots\pi_{ir}^{z_{ir}}}$$

berechnet werden. Um die Abhängigkeit der Güte von $\boldsymbol{t}$ zu eliminieren, schlagen Hilton und Mehta [77] vor, die unbedingte Güte zu betrachten, indem der Erwartungswert der bedingten Güte betrachtet wird:

$$\begin{aligned}\Pi &= \sum_{\{\boldsymbol{t}|\boldsymbol{t}\in\boldsymbol{\Omega}_N\}} P(\boldsymbol{T}=\boldsymbol{t}|H_A)\, P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_A,\boldsymbol{n},\boldsymbol{t}) \\ &= \sum_{\{\boldsymbol{t}|\boldsymbol{t}\in\boldsymbol{\Omega}_N\}} \sum_{\{\boldsymbol{u}:\boldsymbol{u}\in\Gam_{\boldsymbol{n},\boldsymbol{t}},S(\boldsymbol{u})\geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}\}} \prod_{i=0}^{k}\binom{n_i}{u_{i1}\ldots u_{ir}}\pi_{i1}^{u_{i1}}\ldots\pi_{ir}^{u_{ir}}\end{aligned}$$

mit $\boldsymbol{\Omega}_N = \left\{\boldsymbol{t}|\boldsymbol{t}=(t_1,\ldots,t_r)', \sum_{s=1}^{r} t_s = N\right\}$. Die Mächtigkeit der Menge $\boldsymbol{\Omega}_N$ ist nach Nijenhuis und Wilf [78, S.47] gerade $(N+r-1)!/(N!(r-1)!)$, so daß der Aufwand für diese Berechnung sehr hoch ist. Mehta et al. [79] beschreiben die technische Lösung am Beispiel des Cochran-Armitage-Trendtests [46; 80] mit Hilfe des Netzwerkalgorithmus. Für kleine Fallzahlen und eine beliebige Statistik kann dies auch allgemein programmiert werden, ohne daß alle Vorteile vom Netzwerkalgorithmus genutzt werden. Zu unterscheiden sind dabei zwei Fälle:

a) Die Statistik (*S*) nimmt nur ganze Zahlen zwischen zwei bekannten Grenzen $(s_{\min}, s_{\max})$ an, und die daher bekannte maximale Anzahl von Werten ($a = s_{\max} - s_{\min} + 1$), welche die Statistik annehmen kann, ist nicht zu groß. Ein Beispiel hierfür ist der Jonckheere-Terpstra-Test. Der minimal mögliche Wert ist Null (treten Bindungen in den Daten auf, ist er größer). Der

maximale Wert ist $J_{\max} = \sum_{i=0}^{k-1} n_i \left(N - \sum_{j=0}^{i} n_i \right)$ (treten Bindungen in den Daten auf, ist er kleiner). Werden alle Realisierungen von J mit zwei multipliziert, nimmt die Teststatistik auch im Fall von Bindungen nur ganzzahlige Werte an.

b) Die Anzahl der verschiedenen Werte, die die Statistik annehmen kann, ist groß oder schlecht abschätzbar.

Da der Algorithmus für den Fall a) wesentlich leichter erklärbar ist, soll nur er kurz beschrieben werden.

Algorithmus 2.1: Berechnung der unbedingten Güte eines bedingten Tests

1. Eingabe:
   - 1.1 $k+1$ =Anzahl der Behandlungen
   - 1.2 r = Anzahl der verschiedenen Atome der Verteilung
   - 1.3 $\boldsymbol{n} = (n_0,\ldots,n_k)'$ Vektor der Stichprobenumfänge
   - 1.4 $\boldsymbol{\pi}_i = (\pi_{i1},\ldots,\pi_{ir})'$ $(i = 0,\ldots,k)$ Vektoren der Zellwahrscheinlichkeiten
   - 1.5 α = Signifikanzniveau
2. Berechne: $a = s_{\max} - s_{\min} + 1$
3. Initialisiere: $\Pi = 0$
4. REPEAT
   - 4.1 Erzeuge ein Element aus $\Omega_N = \{\boldsymbol{t} | \boldsymbol{t} \in \mathbb{N}^r, \sum_{j=1}^{r} t_j = N\}$ (z. B. mit Algorithmus „Nexcom“ von Nijenhuis und Wilf [78, S.49])
   - 4.2 Initialisiere: $pbed[i] = 0,\ punbed[i] = 0,\ i = 1,\ldots,a$
   - 4.3 REPEAT
     - 4.3.1 Erzeuge eine Tafel $\boldsymbol{z}$ (z. B. mit dem Netzwerkalgorithmus) aus $\Gamma_{n,t} = \{\boldsymbol{z}: \sum_{i=0}^{k} z_{ij} = t_j, \sum_{j=1}^{r} z_{ij} = n_i, i = 0,\ldots,k, j = 1,\ldots,r\}$
     - 4.3.2 Berechne:
       - 4.3.2.1 $s = S(\boldsymbol{z})$,
       - 4.3.2.2 $p(\boldsymbol{z}|H_0) = \prod_{i=0}^{k} \binom{n_i}{z_{i1}\ldots z_{ir}}$
       - 4.3.2.3 $p(\boldsymbol{z}|H_A) = \prod_{i=0}^{k} \binom{n_i}{z_{i1}\ldots z_{ir}} \pi_{i1}^{z_{i1}} \ldots \pi_{ir}^{z_{ir}}$
       - 4.3.2.4 $pbed[s - s_{\min} + 1] = pbed[s - s_{\min} + 1] + p(\boldsymbol{z}|H_0)$
       - 4.3.2.5 $punbed[s - s_{\min} + 1] = punbed[s - s_{\min} + 1] + p(\boldsymbol{z}|H_A)$

UNTIL alle Tafeln erzeugt

4.4 Initialisiere: $pbed_{temp} = 0$, $punbed_{temp} = 0$, $s_{temp} = a$, $grenze = (1-\alpha)\binom{N}{t_1 \ldots t_r}$,

ende=0

4.5 REPEAT

4.5.1 Berechne:

4.5.1.1 $pbed_{temp} = pbed_{temp} + pbed[s_{temp}]$

4.5.1.2 $punbed_{temp} = punbed_{temp} + punbed[s_{temp}]$

4.5.1.3 IF $pbed_{temp} \geq grenze$ THEN ende=1

ELSE $s_{temp} = s_{temp} - 1$

UNTIL ende=1

4.6 $punbed_{temp} = punbed_{temp}$ - $pbed[s_{temp}]$

4.7 $\Pi = \Pi + punbed_{temp}$

UNTIL alle $\boldsymbol{t}$ erzeugt

5. Ausgabe Π

Im Fall b) wird statt der Felder in Punkt 4.2 eine dynamische Liste initialisiert. Diese verkettet geordnete Elemente, die s, $P(S = s|H_0, \boldsymbol{n}, \boldsymbol{t})$ und $P(S = s|H_A, \boldsymbol{n}, \boldsymbol{t})$ aufnehmen können. In Punkt 4.3.2 müssen die neu berechneten Werte in die Liste einsortiert werden, was den Algorithmus entscheidend langsamer macht. Im Abschnitt 4.5 werden die entsprechenden Elemente der geordneten Liste durchlaufen. Die Punkte 4.3 - 4.5 beschreiben, wie das Quantil $s_{\boldsymbol{n},\boldsymbol{t},1-\alpha} = \min(s: P(S \geq s|H_0, \boldsymbol{n}, \boldsymbol{t}) \leq \alpha)$ gefunden werden kann.

Iterativ können so die nötigen Fallzahlen berechnet werden, um eine vorgegebene unbedingte Power eines bedingten Tests zu erreichen. Da die Beziehungen

$$\min_{\{\boldsymbol{t}|\boldsymbol{t}\in\boldsymbol{\Omega}_N\}} P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_A, \boldsymbol{n}, \boldsymbol{t}) \leq \Pi \leq \max_{\{\boldsymbol{t}|\boldsymbol{t}\in\boldsymbol{\Omega}_N\}} P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_A, \boldsymbol{n}, \boldsymbol{t})$$

gelten, ist es möglich, daß für ein festes $\boldsymbol{t}$ dies zu einer Über- bzw. Unterschätzung der Fallzahlen führt.

Als alternativen Schätzer für die Güte eines bedingten Tests schlagen Hilton und Mehta[77] einen Monte-Carlo-Schätzer vor. Dazu werden M zufällige Elemente aus $\boldsymbol{\Omega}_N$ gezogen, für die die bedingte Power berechnet und anschließend gemittelt wird:

$$\hat{\Pi} = M^{-1} \sum_{a=1}^{M} P(\boldsymbol{T} = \boldsymbol{t}_a | H_A)\, P(S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha} | H_A, \boldsymbol{n}, \boldsymbol{t}) \,.$$

Die einfachste Variante zur Schätzung der Güte eines bedingten Tests stellt der übliche Monte-Carlo-Schätzer für Tests dar, der durch Simulationen erhalten werden kann.

Fazit:

Grundidee eines Permutationstests:

Die gemeinsame Verteilung (*F*) aller Beobachtungen ist unter der Nullhypothese invariant bzgl. einer Gruppe von Permutationen (G), d. h. $F(g_u(x_{01},...,x_{kn_k})) = F(g_v(x_{01},...,x_{kn_k}))$ mit $g_u, g_v \in G$. Die Definition dieser Gruppe G ergibt sich aus der Nullhypothese und dem Randomisationsschema.

Vorteile von exakten bedingten Permutationstests

1. Es werden keine speziellen Verteilungsvoraussetzungen benötigt.
2. Selbst Stetigkeitsannahmen sind nicht notwendig.
3. Die Komplexität der Teststatistik spielt keine Rolle.
4. Der Fehler 1. Art ist durch das Signifikanzniveau beschränkt.
5. Die unbedingte Power kann berechnet werden und somit können Fallzahlschätzungen durchgeführt werden.

➔ Diese Tests sind eigentlich ideal für ordinale Daten. Mehta und Patel [19, S. 22] schreiben: *„Ideally one would use exact p-values all the time. They are, after all, the gold standard."*

Nachteile von bedingten Permutationstests

1. Aufgrund der diskreten Verteilung der Statistik wird das Signifikanzniveau in vielen Fällen nicht ausgeschöpft [19, S.743], d. h., die Tests sind meist konservativ.
2. In Extremfällen kann die Diskretheit der Permutationsverteilungsfunktion dazu führen, daß ein vorgegebenes Signifikanzniveau theoretisch gar nicht unterschritten werden kann. Werden z. B. zwei Stichproben verglichen und ist der Umfang je drei, so ist der minimal mögliche p-Wert gerade 0,05.
3. Der technische Aufwand zur Berechnung ist zum Teil enorm und ohne Computer kaum durchführbar (und somit auch schwer validierbar).
4. Der Bekanntheitsgrad von Permutationstests ist nicht sehr hoch, so daß sie auf der einen Seite seltener angewendet werden und andererseits die Akzeptanz geringer ist. Ludbrock

und Dudley [81] beantworten die Frage, warum kaum Permutationstests in der Medizin und der Biologie eingesetzt werden, wie folgt: *„A trivial reason is that the editors of biomedical journals might not understand permutation tests and their statistical advisers might not accept the arguments we have put forward. Our personal experience is that it is much easier to get a manuscript published if one stays with classical tests under the population model.“*

Die ersten beiden Punkte sind dabei die entscheidenden. Gerade bei den in dieser Arbeit betrachteten Fallzahlen können sie nicht unbeachtet gelassen werden. Der 4. Punkt sowie die rechentechnischen Probleme werden sich bei der rasanten Entwicklung der Computerhardware irgendwann verlieren.

2.2 Exakte unbedingte Permutationstests

Einen wesentlichen Punkt bei bedingten Tests stellt das Eliminieren von unbekannten (störenden) Modellparametern durch suffiziente Statistiken dar. Das Bedingen auf feste Randsummen in den oben beschriebenen Tests schränkt den Wertebereich der Teststatistik und damit ihre Verteilung aber sehr ein. Außerdem ist das Bedingen auf feste Randsummen nicht immer angebracht. Ob bedingte oder unbedingte Tests die „richtigen“ Tests sind, soll hier nicht diskutiert werden. In der Literatur zu 2x2-Tafeln und zu dichotomen Daten wird dies hinreichend diskutiert. Eine kleine Zusammenfassung zu diesem Thema gibt Neuhäuser [82, S.93-95]. Hier sollen stattdessen diese Tests hinsichtlich ihrer Güte bei kleinen Fallzahlen untersucht werden.
Wird mit einer Statistik S die Gleichheit von $k+1$-Multinomialverteilungen geprüft, so kann ihre Verteilung durch

$$P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{\pi}) = \sum_{\{z | z \in \Gamma_N ; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \pi_1^{z_{.1}} \ldots \pi_r^{z_{.r}}$$

bei Gültigkeit der Nullhypothese und durch

$$P(S(\boldsymbol{y}) \geq s_0 | H_A, \boldsymbol{n}, \boldsymbol{\pi}) = \sum_{\{z | z \in \Gamma_N ; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \pi_{i1}^{z_{i1}} \ldots \pi_{ir}^{z_{ir}}$$

bei Gültigkeit der Alternativhypothese mit $\Gamma_N = \{z: \sum_{j=1}^{r} z_{ij} = n_i, i = 0,\ldots,k\}$ beschrieben werden. Diese Verteilung hängt nun selbst unter der Nullhypothese noch von dem unbekannten Parametervektor $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)'$ ab. Für den Spezialfall „Vergleich von $k+1$ Binomialverteilungen $(r = 2,\ \boldsymbol{\pi} = (\pi_1, 1-\pi_1)')$“ schlagen Mehta und Hilton [83] für den globalen Vergleich und Koch [84] für den many-to-one Vergleich vor, den Nuisance-Parameter mittels Supremumbildung zu eliminieren:

$$P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}) = \sup_{\boldsymbol{\pi}} \sum_{\{z | z \in \Gamma_N; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \pi_1^{z_{.1}} \ldots \pi_r^{z_{.r}}.$$

Mit $s_{1-\alpha}(\boldsymbol{n}, \boldsymbol{\pi}) = \min(s: P(S \geq s | H_0, \boldsymbol{n}, \boldsymbol{\pi}) \leq \alpha)$ und $s_{1-\alpha} = \sup_{\boldsymbol{\pi}} \{s_{1-\alpha}(\boldsymbol{n}, \boldsymbol{\pi})\}$ ist die Güte unter der Nullhypothese $\pi_j = \pi_{0j} = \cdots = \pi_{kj},\ j = 1,\ldots,r$

$$\Pi(s_{1-\alpha} | H_0, \boldsymbol{n}, \boldsymbol{\pi}) = \sum_{\{z | z \in \Gamma_N; S(z) \geq s_{1-\alpha}\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \pi_1^{z_{.1}} \ldots \pi_r^{z_{.r}}$$

und für eine beliebige Alternative $\boldsymbol{\pi} = (\pi_{01}, \ldots, \pi_{0r}, \ldots, \pi_{k1}, \ldots, \pi_{kr})'$

$$\Pi(s_{1-\alpha} | H_A, \boldsymbol{n}, \boldsymbol{\pi}) = \sum_{\{z | z \in \Gamma_N; S(z) \geq s_{1-\alpha}\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \pi_{i1}^{z_{i1}} \ldots \pi_{ir}^{z_{ir}}$$

gegeben. Der Aufwand zur Bestimmung dieser beiden Größen ist allerdings enorm. Ein Problem stellt schon das Finden von $s_{1-\alpha}$ dar. Anders als im Fall der Binomialverteilungen hängt das Supremum nicht von einem, sondern von $r-1$ Parametern ab. Das Supremum bestimmen Mehta und Hilton mittels Optimierungssoftware, während Koch [84] es in Anlehnung an Habers [85] Vorschlag für 2×2-Tafeln iterativ bestimmt. Beides scheint für $(k+1) \times r$-Tafeln zu aufwendig. Daher wird analog zu Koch [84] vorgeschlagen, die Parameter mittels der suffizienten Statistik $T_j = \sum_{i=0}^{k} Y_{ij},\ j = 1,\ldots r$ unter der Nullhypothese durch $\hat{\pi}_j = \frac{t_j}{N}$ zu schätzen:

$$P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \hat{\boldsymbol{\pi}}) = \sum_{\{z | z \in \Gamma_N ; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \hat{\pi}_1^{z_{i1}} \ldots \hat{\pi}_r^{z_{ir}}$$

Die Darstellung

$$\begin{aligned} & P(S(x) \geq s_0 | H_0, \boldsymbol{n}, \hat{\boldsymbol{\pi}}) \\ &= \sum_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} \hat{\pi}_1^{t_1} \ldots \hat{\pi}_r^{t_r} \sum_{\{z | z \in \Gamma_{n,t} ; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \\ &= \sum_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} \hat{\pi}_1^{t_1} \ldots \hat{\pi}_r^{t_r} \binom{N}{t_1 \ldots t_r} P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) \end{aligned} \tag{2.2}$$

legt wieder nahe, den Netzwerkalgorithmus zu benutzen, da der dritte Faktor innerhalb der Summe nichts anderes als ein p-Wert eines bedingten Permutationstests ist. Da die Menge $\boldsymbol{\Omega}_N$ iterativ bestimmt werden kann, können diese unbedingten Tests für jede Statistik programmiert werden. Diese Tests sind zwar nur asymptotisch exakt, die Ergebnisse von Koch [84] und Neuhäuser [82] lassen aber hoffen, daß kaum Niveauverletzungen auftreten und daß die Power zumindest für kleine Fallzahlen weit besser als die eines bedingten Gegenparts ist. Aus Formel (2.2) folgt:

1. Analog zur unbedingten Güte eines bedingten Permutationstests kann der p-Wert dieser unbedingten Permutationstests als unbedingte Wahrscheinlichkeit eines bedingten Permutationstests, unter der Nullhypothese $\pi_j = \pi_{0j} = \cdots = \pi_{kj} = \frac{t_j}{N}$, $j=1,\ldots,r$ größer oder gleich s_0 zu sein, interpretiert werden.
2. Da $\sum_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} \hat{\pi}_1^{t_1} \ldots \hat{\pi}_r^{t_r} \binom{N}{t_1 \ldots t_r} = N^{-N} (t_1 + \cdots + t_r)^N = 1$ gilt, folgt:

   $$\min_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) \leq P_{\hat{\boldsymbol{\pi}}}(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}) \leq \max_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}).$$

   Somit kann ein exakter bedingter Permutationstest eine niedrigere, aber auch eine höhere Güte als der vorgestellte unbedingte Test besitzen.
3. Der vorgestellte unbedingte Test ist im strengen Sinne ein parametrischer Bootstraptest (siehe Kapitel 3).

Diese Tests werden im folgenden kurz als unbedingte Permutationstests bezeichnet. Der Algorithmus 2.3 kann zur Berechnung der p-Werte genutzt werden.

Algorithmus 2.3: Berechnung des p-Wertes eines unbedingten Permutationstests (auf Basis der geschätzten Zellwahrscheinlichkeiten)

1. Eingabe:
   1.1 $k+1$ =Anzahl der Behandlungen
   1.2 r =Anzahl der verschiedenen Atome der Verteilung
   1.3 $\boldsymbol{n} = (n_0,\ldots,n_k)'$ Vektor der Stichprobenumfänge
   1.4 $\boldsymbol{y}_i = (y_{i1},\ldots,y_{ir})'$ $(i = 0,\ldots,k)$ Vektoren der Zellwahrscheinlichkeiten
2. Berechne:
   2.1 $s_0 = S(y)$
   2.2 $\hat{\pi}_j = \frac{y_{.j}}{N}, j=1,\ldots,r$
3. Initialisiere: $p = 0$
4. REPEAT
   4.1 Erzeuge ein Element aus $\Omega_N = \{\boldsymbol{t} | \boldsymbol{t} \in \mathbb{N}^r, \sum_{j=1}^{r} t_j = N\}$ (z. B. mit Algorithmus „Nexcom“ Nijenhuis und Wilf[78, S.49])
   4.2 Berechne:
      4.2.1 $p_{bed} = P(S(\boldsymbol{y}) \geq s_0 | H_0, T_j = t_j, j = 1,\ldots,r)$ (z. B. mit Netzwerkalgorithmus)
      4.2.2 $p = p + \hat{\pi}_1^{t_1} \ldots \hat{\pi}_r^{t_r} \binom{N}{t_1 \ldots t_r} p_{bed}$

   UNTIL alle $\boldsymbol{t}$ erzeugt
5. Ausgabe: p

Ist in Punkt 4.2.1 die Bestimmung des bedingten p-Wertes mit dem Netzwerkalgorithmus nicht möglich, so kann analog zu Punkt 4.3 aus Algorithmus 2.1 der Netzwerkalgorithmus zumindest zur Erzeugung der Tafeln genutzt werden. Der bedingte p-Wert wird dann nach Formel (2.1) berechnet.

2.3 Modifizierte Permutationstests

Für eine Statistik S, dessen Permutationsverteilungsfunktion auf nur wenige Punkte konzentriert ist, wird es oft keinen Punkt s geben, für den gilt: $P(S \geq s|H_0, \boldsymbol{\pi}) = \alpha$. Ist $s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}$ der kleinste Wert, für den $P(S \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0, \boldsymbol{n}, \boldsymbol{t}) = \alpha_{\text{exp}} < \alpha$ gilt, so kann ein konservativer α-Test beruhend auf S definiert werden durch

$$\Psi(\boldsymbol{y}|\boldsymbol{n},\boldsymbol{t}) = \begin{cases} 1: & S(\boldsymbol{y}) \geq s_{\boldsymbol{n},\boldsymbol{t},1-\alpha} \\ 0: & S(\boldsymbol{y}) < s_{\boldsymbol{n},\boldsymbol{t},1-\alpha} \end{cases} .$$

Das Risiko, eine gültige Nullhypothese fälschlicherweise abzulehnen (Fehler 1. Art), beträgt für diesen Test jedoch nicht α sondern α_{exp}. Ist die Differenz $\alpha - \alpha_{\text{exp}}$ groß, wirkt sich dies negativ auf die Güte des Test unter der Alternative aus. Für einen Punkt $q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}$ mit

$$P(S \geq q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0, \boldsymbol{n}, \boldsymbol{t}) > \alpha > P(S > q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0, \boldsymbol{n}, \boldsymbol{t})$$

können Tests definiert werden, die das Niveau besser ausschöpfen und somit in der Regel eine höhere Güte besitzen. Vom theoretischen Standpunkt (siehe z. B. Schrage [86] für einfache Kontrasttests) ist z. B. der Test

$$\Psi_\gamma(\boldsymbol{y}) = \begin{cases} 1 & S(\boldsymbol{y}) > q_{\boldsymbol{n},\boldsymbol{t},1-\alpha} \\ \gamma & S(\boldsymbol{y}) = q_{\boldsymbol{n},\boldsymbol{t},1-\alpha} \\ 0 & S(\boldsymbol{y}) < q_{\boldsymbol{n},\boldsymbol{t},1-\alpha} \end{cases} \quad \text{mit} \quad \gamma = \frac{\alpha - P(S(\boldsymbol{Y}) > q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0, \boldsymbol{n}, \boldsymbol{t})}{P(S(\boldsymbol{Y}) = q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}|H_0, \boldsymbol{n}, \boldsymbol{t})}$$

eine bessere Wahl. Das zusätzliche Zufallsexperiment, welches nötig ist, falls $S(\boldsymbol{y}) = q_{\boldsymbol{n},\boldsymbol{t},1-\alpha}$ gilt, führt jedoch oft zu einer Ablehnung des Verfahrens in der Praxis. Zudem kann eine Signifikanz nicht durch einen üblichen p-Wert beschrieben werden. Dies führt zu den modifizierten p-Werten. Jeder p-Wert eines bedingten Tests kann wie folgt zerlegt werden:

$$\begin{aligned} P(S(\boldsymbol{y}) \geq s_0|H_0, \boldsymbol{n}, \boldsymbol{t}) &= P(S(\boldsymbol{y}) > s_0|H_0, \boldsymbol{n}, \boldsymbol{t}) + P(S(\boldsymbol{y}) = s_0|H_0, \boldsymbol{n}, \boldsymbol{t}) \\ &= p_> + p_= . \end{aligned}$$

Durch eine Gewichtung des Terms $p_=$ mit einer Zahl zwischen 0 und 1 kann der p-Wert verkleinert werden. Unter H_0 wird der Test somit weniger konservativ. Die Power des Tests verbessert sich ebenfalls. Aber es sind auch Niveauüberschreitungen möglich, d. h., es liegt kein exakter Niveau-α-Test mehr vor. Lancaster [87] schlägt als Gewicht ½ vor. Dieser sogenannte Mid-p-Wert ist der bekannteste dieser modifizierten p-Werte und wird in der Literatur recht häufig empfohlen [19; 57; 88; 89]. Barnard [89] zeigt, daß der Erwartungswert des Mid-p-Wertes bei ½ und die Varianz nahe 1/12 liegt. Erwartungswert und Varianz entsprechen also fast den Werten eines *p*-Werts bei stetigen Verteilungen, womit wenigstens zum Teil eine theoretische Rechtfertigung für diese empirische Modifikation existiert [19].
Cohen und Sackrowitz [90] , Berger und Sackrowitz [91] und Berger et al. [92] schlagen weitere randomisierte Tests vor. Auf diese soll hier jedoch nicht eingegangen werden.

2.4 Approximative Permutationstests

Ist die Berechnung der Statistik aufwendig oder existieren sehr viele verschiedene Permutationen, führt dies zu langen Berechnungszeiten der Permutationsverteilung bzw. des p-Wertes. Dwass [94] schlug daher 1957 vor, statt aller Permutationen nur eine bestimmte Anzahl (*M*) der möglichen Permutationen zufällig zu erzeugen und auf Basis dieser, den Permutationstest durchzuführen. Für die Erzeugung solcher zufälligen Permutationen existieren unterschiedliche Algorithmen.
Eine sehr einfache Variante zum Permutieren eines Feldes $(x_1,...,x_N)$ der Länge N beschreibt Berry [66]. Dazu werden N auf den Intervall (0,1) gleichverteilte Zufallszahlen $(z_1,...,z_N)$ erzeugt und anschließend in Ränge $(R(z_1),...,R(z_N))$ transformiert. Dieser Rangvektor wird als Indexvektor benutzt. Das permutierte Feld ergibt sich aus $y_i = x_{R(z_i)}, i = 1,...,N$. Berry bezeichnet diese approximativen Permutationstests als simulation-based-Tests. Für einen Test auf Anstieg werden *M* Permutationen erzeugt und jeweils die Teststatistik berechnet. Der p-Wert wird definiert als Anteil der Permutationen, die zu einem Wert führen, der nicht kleiner ist als der Originalwert (t_0). Mit der Indikatorfunktion I (I(E) = 1, falls logischer Ausdruck E wahr, I(E) = 0, falls E falsch) und $V_m = I(T(g_m(\boldsymbol{x})) \geq t_0)$ ist der p-Wert bestimmt durch:

$$p^* = \frac{1}{M}\sum_{m=1}^{M} V_m .$$

Die unabhängigen Zufallsvariablen V_i sind dabei binomialverteilt $B(1, p)$, wobei p der wahre p-Wert des exakten Tests ist. Mit Hilfe des Zentralen Grenzwertsatzes kann daher ein asymptotisches Konfidenzintervall für den wahren p-Wert konstruiert werden:

$$[p^* - z_{1-\alpha/2}\sqrt{p(1-p)/M}, p^* + z_{1-\alpha/2}\sqrt{p(1-p)/M}].$$

Die Länge dieses Intervalls ist vom unbekannten p und von M abhängig. Für $0 \le p \le 0{,}5$ ist $z_{1-\alpha/2}\sqrt{p(1-p)/M}$ eine monoton fallende Funktion. Somit kann M für eine angestrebte Intervallbreite bestimmt werden, wenn Vorstellungen über den p-Wert existieren (z. B. kleiner 0,5). Soll z. B. die Breite eines 0,99-Konfidenzintervalls nicht größer als 0,01 sein (das Standardnormalverteilungsquantil ist $z_{1-\alpha/2} = 2{,}576$), so werden für *M* die in Tabelle 2.5 dargestellten Zahlen empfohlen.

p	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1
M >	2627	5201	7723	10.191	12.606	14.968	17.277	19.533	21.735	23.885

Tabelle 2.5: Empfehlung für *M* beim approximativen Permutationstest in Abhängigkeit von p

Durch die Wahl von *M* kann der exakte p-Wert beliebig gut approximiert werden. Ein Nachteil ist die Abhängigkeit des p-Wertes von den benutzten Permutationen. Diese werden mittels Zufallsgeneratoren erzeugt und sind somit vom Generator und dessen Startwert abhängig. Liegt der berechnete p-Wert sehr nahe dem vorgegebenen Niveau, so kann es vorkommen, daß der Test in manchen Fällen zur Ablehnung und in anderen zur Annahme führt. Streng genommen sind hier demnach genauso zusätzliche Zufallsexperimente (Auswahl der zufälligen Permutation) notwendig wie beim randomisierten exakten Permutationstest. Allerdings kann hier wenigstens ein p-Wert definiert werden.

Diese approximativen Permutationstests können in die größere Klasse von Monte-Carlo-Tests eingegliedert werden. Viele Ergebnisse, die allgemein für Monte-Carlo-Tests hergeleitet wurden, können so auf Permutationstests übertragen werden. Jöckel [95] zeigt, wie z. B. die asymptotische relative Pitman-Effizienz eines approximativen Tests zum exakten Test berechnet werden kann. Ebenso zeigt er, wie mit Hilfe der asymptotischen relativen Pitman-Effizienz zwei approximative Permutationstests verglichen werden können. Da hierfür jedoch große Fallzahlen benötigt werden, soll hier nicht weiter darauf eingegangen werden.

3 Bootstraptests

3.1 Parametrischer und nichtparametrischer Bootstraptest

Bootstrapverfahren wurden erstmalig 1979 von Efron [96] zur Schätzung der Varianz und der Verzerrung eines Schätzers eingeführt. Seitdem wurden Bootstrapverfahren für fast alle Teilgebiete der Statistik entwickelt. Für die Testtheorie sind Bootstrapverfahren aufgrund der Möglichkeit, Quantile und Verteilungsfunktionen zu schätzen, besonders interessant. Das Prinzip eines Bootstraptests und dessen Vor- und Nachteile bei kleinen Stichprobenumfängen werden zunächst an einem einfachen Beispiel diskutiert. Die Übertragung auf den Mehrstichprobenfall ist dann leicht nachvollziehbar.

Beispiel 3.1: Einstichprobentest für das Problem $H_0: \mu \le \mu_0$ vs. $H_A: \mu > \mu_0$ (μ sei der Erwartungswert einer Zufallsgröße X, μ_0 sei ein bekannter Wert)

Basierend auf einer unabhängig identisch verteilten Stichprobe $X_1,\ldots,X_N$ können diese Hypothesen, z. B. mittels der einfachen Statistik $T_N = \overline{X}_\cdot - \mu_0$ (große Werte sprechen gegen die Nullhypothese), geprüft werden. Die im allgemeinen unbekannte Verteilung dieser Statistik kann durch Simulation der Verteilungsfunktion $F(x) = P(X < x)$ approximiert werden, indem

$$G_N(t|F) = P(T_N < t|F) \;\approx\; \hat{G}_N(t|F) = P(T_N < t|\hat{F}) \tag{3.1}$$

gesetzt wird. Wird unterstellt, daß die Verteilungsfunktion F zu einer parametrischen Verteilungsfamilie ($F \in \{F_\theta: \theta \in \mathbb{R}^a\}$) gehört, so kann $\hat{F} = F_{\hat{\theta}}$ gewählt werden. Wird diese Annahme nicht getroffen, so wird F nichtparametrisch, z. B. durch die empirische Verteilungsfunktion $\hat{F}_N$, geschätzt. Die Schätzung $\hat{G}_N(t|F) = P(T_N < t|\hat{F})$ kann im einfachsten Fall (ordinary bootstrap) mittels Simulationen bestimmt werden, indem M zufällige Stichproben (Resamples, Resamplingstichproben) $\boldsymbol{X}_m^* = (X_{m1}^*,\ldots,X_{mN}^*)$ ($m = 1,\ldots,M$) entsprechend $\hat{F}$ (z. B. durch Ziehen aus der Originalstichprobe) generiert und jeweils die Werte $T(\boldsymbol{X}_m^*)$ berechnet werden. Die Schätzung für $\hat{G}_N(t|F)$ kann dann mittels $V(\boldsymbol{X}_m^*, t) = I(T(\boldsymbol{X}_m^*) < t)$ beschrieben werden:

$$\hat{G}_N(t|F) = \frac{1}{M}\sum_{m=1}^{M} V(\boldsymbol{X}_m^*, t)\,.$$

Bei der Konstruktion eines Tests ist zu beachten, daß die Verteilung der Statistik unter Gültigkeit der Nullhypothese zu bestimmen ist. Es wird also nicht $\hat{G}_N(t|F)$ sondern $\hat{G}_N(t|F,H_0)$ benötigt. Ist z. B. $\hat{F} = \hat{F}_N$ die empirische Verteilungsfunktion und sollen die Resamples durch Ziehen mit Zurücklegen aus der Originalstichprobe erzeugt werden, so sind die Daten so zu transformieren, daß sie den Erwartungswert μ_0 besitzen. Beispielsweise kann folgende Transformation gewählt werden:

$$Y_j = X_j - \overline{X}_. + \mu_0 \quad (j = 1,\ldots,N).$$

Für die Teststatistik bzw. für den theoretischen Erwartungswert ergibt sich dann:

$$T(\boldsymbol{Y}_m^*) = \overline{Y}_{m.}^* - \mu_0 = \frac{1}{N}\sum_{j=1}^{N} X_{mj}^* - \overline{X}_. + \mu_0 - \mu_0 = \overline{X}_{m.}^* - \overline{X}_.$$

bzw.

$$E\,T(\boldsymbol{Y}_m^*) = \frac{1}{N}\sum_{j=1}^{N} \underbrace{EX_{mj}^*}_{=\overline{X}_.} - \overline{X}_. + \mu_0 - \mu_0 = 0\,.$$

Wird mit $t_0 = \overline{x}_. - \mu_0$ der beobachtete Wert von T bezeichnet, so ist

$$p(\boldsymbol{y}_1^*,\ldots,\boldsymbol{y}_M^*) = 1 - \frac{1}{M}\sum_{m=1}^{M} V(\boldsymbol{y}_m^*, t_0)$$

ein Schätzer für die Wahrscheinlichkeit, größer oder gleich dem beobachteten Wert zu sein. Würden die nicht transformierten Daten genutzt , so ergäbe sich:

$$T(\boldsymbol{X}_m^*) = \frac{1}{N}\sum_{j=1}^{N} X_{mj}^* - \mu_0 = \overline{X}_{m.}^* - \mu_0\,, \quad E^*\,T_N(X_m^*) = \overline{X}_. - \mu_0 \quad \text{und}$$

$$p(\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_M^*) = 1 - \frac{1}{M}\sum_{m=1}^{M} V(\boldsymbol{x}_m^*, t_0)\,.$$

Ist die Nullhypothese falsch und ist der wahre Erwartungswert μ_1 deutlich größer als μ_0, so gilt: $T(\boldsymbol{Y}_m^*) = \overline{X}_{m.}^* - \mu_1 + \mu_1 - \overline{X}_.$ bzw. $T(\boldsymbol{X}_m^*) = \overline{X}_{m.}^* - \mu_1 + \mu_1 - \mu_0$. Da der Mittelwert mit

wachsendem Stichprobenumfang gegen den Erwartungswert strebt (Gesetz der Großen Zahlen), wird bei genügend großem N stets $T(\boldsymbol{Y}_m^*) \leq T(\boldsymbol{X}_m^*)$ gelten (der p-Wert ist somit für die transformierten Daten kleiner). Die Verwendung der nicht transformierten Daten würde demnach dazu führen, daß erstens die Verteilung der Statistik nicht unter der Nullhypothese simuliert wird und zweitens, daß der zugehörige Test eine geringere Power besitzt. Beran[97] und Hall und Wilson [98] verweisen darauf, daß die Verteilungen von standardisierten Statistiken besser approximiert werden können. Beran verweist dabei auf den Zusammenhang zwischen Standardisierung und Prepivotstatistiken, die im allgemeinen zu einer höheren Güte unter der Nullhypothese führen (siehe Abschnitt 3.2). Für das obige Beispiel kann daher die folgende Statistik gewählt werden:

$$T_N(\boldsymbol{X}) = \frac{\sqrt{N}(\overline{X}_{\cdot} - \mu_0)}{\hat{\sigma}} \text{ bzw. } T_N(\boldsymbol{Y}_m^*) = \frac{\sqrt{N}(\overline{Y}_{m\cdot}^* - \mu_0)}{\hat{\sigma}^*} = \frac{\sqrt{N}(\overline{X}_{m\cdot}^* - \overline{X}_{\cdot})}{\hat{\sigma}^*}.$$

Der Varianzschätzer wird dabei ebenfalls auf der Basis der Bootstrapstichprobe berechnet. Zum Beispiel wäre

$$S_N^2(\boldsymbol{X}) = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \overline{X}_{\cdot})^2 \text{ bzw. } S_N^2(\boldsymbol{Y}_m^*) = \frac{1}{N-1}\sum_{j=1}^{N}(Y_{mj}^* - \overline{Y}_{m\cdot}^*)^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_{mj}^* - \overline{X}_{\cdot}^*)^2$$

ein möglicher Schätzer für die Varianz. Werden die Daten aus einer Stichprobe mit N verschiedenen Werten gezogen, so beträgt die Wahrscheinlichkeit dafür, daß der Varianzschätzer den Wert Null annimmt (die Teststatistik ist dann nicht berechenbar),

$$P(S_N^2(\boldsymbol{Y}_m^*) = 0) = N^{-(N-1)} \overset{N=5}{=} 0{,}0016.$$

Werden M Resamples gezogen, so ist die Wahrscheinlichkeit, daß zumindest einmal dieses Ereignis auftritt

$$P(\min_m(S_N^2(\boldsymbol{Y}_m^*)) = 0) = 1 - (1 - N^{-(N-1)})^M \overset{N=5, M=10.000}{\approx} 1.$$

Treten Bindungen in den Daten auf, verschärft sich dieses Problem weiter. Bei sehr kleinen Fallzahlen stellt daher das Schätzen der Varianz und die Berechnung standardisierter

Statistiken ein Problem dar. Es stellt sich die Frage, wie die Resamplingstichproben mit nichtpositivem Varianzschätzer zu behandeln sind. Folgende Varianten sind denkbar:

1. Sie werden einfach ignoriert (Wert der Statistik wird als nicht extrem gewertet);
   ==> dies führt zu liberalen Tests.
2. Es werden jeweils Ersatzstichproben gezogen;
   ==> dies kann zu endlos langen Schleifen führen.
3. Der Test wird bestraft, indem der Wert als extrem gewertet wird;
   ==> dies führt zu konservativen Tests.

Bei der Wahl eines Varianzschätzers ist daher darauf zu achten, wie robust er gegen den Wert Null ist. In extremen Situationen (sehr hoher Bindungsanteil) sollte überlegt werden, ob es nicht sogar besser ist, ohne Varianzschätzer zu arbeiten. Hall und Wilson [98] beschreiben ebenfalls Situationen, bei denen keine „guten“ Varianzschätzer existieren und daher auf sie verzichtet werden sollte. Allgemein wird im Falle einer parametrischen Schätzung für F vom parametrischen und sonst vom nichtparametrischen Bootstrapschätzer gesprochen [99, S.148]. Ist die Schätzung für F eine diskrete Verteilungsfunktion mit endlichem Träger (z. B. die empirische Verteilungsfunktion), so können theoretisch alle möglichen Stichproben entsprechend der Verteilungsfunktion $\hat{F}$ generiert und ihre Wahrscheinlichkeiten berechnet werden. In den meisten Fällen wird dies allerdings numerisch zu aufwendig sein.
Bei der Verallgemeinerung auf das Mehrstichprobenproblem ergeben sich zusätzliche Schwierigkeiten. Schon die Wahl des Resamplingraumes ist nicht immer einfach. So können die Resamples aus der gepoolten Stichprobe oder auch separat aus den einzelnen Stichproben gezogen werden. Zusätzlich können Zentrierung, Normierung oder gar Standardisierung angezeigt sein [100, S.187; 101]. Insgesamt sind daher im allgemeinen zwei Approximationsschritte und die Wahl des Resamplingraumes notwendig. Im Gegensatz zu exakten Permutationstests sind Bootstraptests approximative Tests. Permutations- und Bootstraptests haben jedoch auch viele Gemeinsamkeiten und sind in manchen Fällen asymptotisch äquivalent [102]. Die Schritte zur Durchführung eines Bootstraptests sind daher denen eines Permutationstests ähnlich. Der wesentliche Unterschied zum Permutationstest liegt im vierten Punkt (siehe S. 22). Werden z. B. zwei Behandlungen (Nullhypothese: gleiche Wirkung) verglichen, so kann der Unterschied zwischen Bootstrap- und Permutationstests anhand des Urnenmodells leicht erklärt werden. Bei einem Permutationstest werden aus der Urne, in der sich $N = n_1 + n_2$ Kugeln (Verkörpern die Testergebnisse) befinden, n_1 Kugeln für die erste Behandlung und

n_2 Kugeln für die zweite Behandlung ohne Zurücklegen der Kugeln gezogen. Falls die empirische Verteilung der gepoolten Stichprobe der Schätzer für die wahre Verteilung unter der Nullhypothese ist, wird beim Bootstraptest nach jedem Ziehen die Kugel wieder zurückgelegt. Ist die gewählte Teststatistik unabhängig von der Anordnung der Werte innerhalb der Behandlungsstichproben und existieren keine Bindungen in den Daten, so können beim Permutationstest

$$\frac{N!}{n_1!n_2!} \qquad \text{(Kombinationen ohne Wiederholung)}$$

und beim Bootstraptest

$$\frac{(N+n_1-1)!}{n_1!(N-1)!}\frac{(N+n_2-1)!}{n_2!(N-1)!} \qquad \text{(Kombinationen mit Wiederholung)}$$

unterschiedliche Stichproben gezogen werden (weitere ergeben sich aus Permutieren der Daten innerhalb der Gruppen).

	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 8$	$n = 9$
Permutations-test	20	70	252	924	12.870	48.620
Bootstraptest	3.136	108.900	4.008.004	153.165.376	2,4041E11	9,7628E12

Tabelle 3.1: Beispiele für die Anzahl der möglichen Resamplingaufteilungen bei $n = n_1 = n_2$

Entspricht die gepoolte empirische Verteilung nicht einem adäquaten Verteilungsmodell, so können beim Bootstraptest die Resamplingstichproben auch jeweils innerhalb der einzelnen Behandlungsgruppen gezogen werden. Dies ist z. B. angezeigt, wenn Lokationsunterschiede getestet werden sollen, sich die Verteilungen aber hinsichtlich ihrer Varianzen unterscheiden [101] (im Fall $n = 3$ existieren für obiges Beispiel dann immer noch 100 verschiedene Resamplingstichproben).

Fisher und Hall [103] beschreiben für den Einstichprobenfall einen nichtparametrischen Bootstrapansatz ($\hat{F} = \hat{F}_N$), der auf der vollständigen Enumeration aller Resamplingstichproben beruht. $\hat{G}_N(t|F)$ wird in diesem Fall nicht simuliert, sondern exakt berechnet. Für $N = 6$ ist die Bootstrapverteilung $\hat{G}_N(t|F)$ auf 462 Punkte konzentriert. Ein „exakter“ Bootstraptest

kann dann einem Bootstraptest, der auf mehreren hundert Resamplingstichproben (z. B. 10.000) beruht, vorgezogen werden, da er weniger Berechnungen erfordert und nur auf einem Approximationsschritt beruht [103]. Die Erzeugung aller Resamplingstichproben der gepoolten Stichprobe ist für kleine Stichprobenumfänge, auch für Mehrstichprobentests (Einweganlage), mit Hilfe des Netzwerkalgorithmus möglich. Der Aufwand entspricht dem des beschriebenen unbedingten Permutationstests (siehe Algorithmus 2.3). Wird mit r die Anzahl der verschiedenen Werte im Resamplingraum und mit S die Statistik bezeichnet, kann der p-Wert beschrieben werden durch

$$
\begin{aligned}
&P(S(\boldsymbol{x}) \geq s_0 | H_0, \boldsymbol{n}) \\
&= r^{-N} \sum_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} \sum_{\{z | z \in \Gamma_{\boldsymbol{t}}; S(z) \geq s_0\}} \prod_{i=0}^{k} \binom{n_i}{z_{i1} \ldots z_{ir}} \\
&= r^{-N} \sum_{\{\boldsymbol{t} | \boldsymbol{t} \in \boldsymbol{\Omega}_N\}} \binom{N}{t_1 \ldots t_r} P(S(\boldsymbol{x}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}).
\end{aligned}
$$

Treten die r verschiedenen Werte in der gepoolten Originalstichprobe mit derselben Häufigkeit h auf, so gilt: $\hat{\pi}_1^{t_1} \hat{\pi}_2^{t_2} \cdots \hat{\pi}_r^{t_r} = (\frac{h}{N})^{t_1} (\frac{h}{N})^{t_2} \cdots (\frac{h}{N})^{t_r} = (\frac{h}{hr})^N = r^{-N}$. In diesem Fall stimmt der Bootstraptest mit dem in Abschnitt 2.2 vorgestellten asymptotisch exakten unbedingten Permutationstest überein. Der p-Wert kann ebenfalls wie in Abschnitt 2.2 als unbedingte Wahrscheinlichkeit eines bedingten Permutationstests interpretiert werden (Nullhypothese: $\pi_j = \pi_{0j} = \cdots = \pi_{cj} = \frac{1}{r}$, $j = 1, \ldots, r$). Da der Aufwand aber vor allem für komplizierte Statistiken zu groß ist, wird im Mehrstichprobenproblem fast immer eine Monte-Carlo-Variante des Bootstraps angebracht sein.

Fazit:

Grundidee:

Es existiert ein Schätzer, der die gemeinsame Verteilung aller Beobachtungen unter der Nullhypothese gut approximiert.

Vorteile von Bootstraptests

1. Es werden kaum einschränkende Verteilungsannahmen vorausgesetzt.
2. Bootstraptests können in manchen Fällen selbst dann noch genutzt werden, wenn Permutationstests versagen (extrem kleine Fallzahlen) bzw. nicht gerechtfertigt sind (z. B. Interaktionsmodelle) [53, S.59].

3. Stetigkeitsannahmen werden auch hier nicht benötigt.
4. Durch den größeren Resamplingraum ist eine höhere Güte als bei Permutationstests möglich.

Nachteile von Bootstraptests

1. Streng genommen ist die Konsistenz der Bootstrapschätzung (im Prinzip die Rechtfertigung für Formel 3.1) zu zeigen, da nur dann Bootstraptests genutzt werden dürfen (dies ist in vielen Fällen ein schwieriges Problem).
2. Bootstraptests hängen von der Güte der Approximation der wahren Verteilung unter der Nullhypothese ab.
3. In vielen Fällen ist die Wahl eines geeigneten Resamplingraumes unter der Nullhypothese nicht offensichtlich [100, S.187].
4. Das Einhalten eines Signifikanzniveaus kann nur asymptotisch abgesichert werden (z. B. mittels Landausymbolik). In einigen Spezialfällen kann aber theoretisch bewiesen werden, daß diese Approximation besser als eine Normalverteilungsapproximation ist[100, S. 93].
5. Der Bekanntheitsgrad dieser Tests ist noch geringer als der von Permutationstests.
6. Für die Resamplinganzahl gibt es in der Regel nur Empfehlungen. Zudem beeinflußt sie hauptsächlich die Approximation der exakten Bootstrapverteilung. Konvergiert die exakte Bootstrapverteilung nicht gegen die exakte Verteilung der Originalstatistik, ist die Erhöhung der Resamplinganzahl nur mit einer besseren Approximation der exakten Bootstrapverteilung verbunden. Eine Erhöhung der Resamplinganzahl geht daher nicht immer konform mit einem Gütegewinn für das eigentliche Problem.

Soll das vorgegebene Signifikanzniveau auf jeden Fall eingehalten werden, sind exakte Permutationstests den Bootstraptests vorzuziehen. Werden jedoch kleine Überschreitungen des Niveaus zugunsten einer höheren Power toleriert, so stellt sich die Frage, ob die Bootstraptests zumindest bei sehr kleinen Fallzahlen die bessere Wahl sind. Mittels Simulationen wird versucht, für ausgewählte Statistiken eine Antwort auf diese Frage zu geben.
Die Konsistenz der Bootstrapschätzung kann zwar für Spezialfälle [100; 101; 102] gezeigt werden, darauf soll hier aber aufgrund der kleinen Fallzahlen und der damit verbundenen Frage der Gültigkeit der asymptotischen Aussagen [106] nur kurz eingegangen werden. Efron [107, S.34] schreibt:
„The rationale for the bootstrap method is particulary evident when the sample space H is finite, say H={1,2,3,...,L}.“

Efron zeigt, daß für diskrete unabhängig und identisch verteilte Zufallsvariablen die Bootstrapverteilung unter sehr schwachen Voraussetzungen asymptotisch korrekt ist (d. h., die mittels der Resamplingstichproben simulierte Verteilung konvergiert schwach gegen die wahre Verteilung der Statistik). Die Aussagen beruhen auf dem folgenden Satz für Multinomialverteilungen.

Satz 3.1: *Es seien* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ *unabhängig und identisch multinomialverteilte r–dimensionale Zufallsvektoren (*$P_{X_i} = \mathrm{M}(1, \boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_r)'$*) und* $\hat{\boldsymbol{\pi}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$. *Gilt zusätzlich* $0 < \zeta_0 \le \hat{\pi}_s \le 1 - \zeta_0 < 1$, $s = 1, \ldots, r$, *dann folgt:*

1. $P_{\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})} \Rightarrow N(\mathbf{0}, \boldsymbol{\Sigma})$ *mit* $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1,\ldots,r}$ *und* $\sigma_{ij} = \begin{cases} \pi_i (1 - \pi_i) & : i = j \\ -\pi_i \pi_j & : i \neq j \end{cases}$

2. $P_{\sqrt{n}(\hat{\boldsymbol{\pi}}^* - \hat{\boldsymbol{\pi}})} \Rightarrow N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$ *mit* $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{i,j=1,\ldots,r}$ *und* $\hat{\sigma}_{ij} = \begin{cases} \hat{\pi}_i (1 - \hat{\pi}_i) & : i = j \\ -\hat{\pi}_i \hat{\pi}_j & : i \neq j \end{cases}$

3. $\hat{\boldsymbol{\Sigma}} \xrightarrow[n \to \infty]{} \boldsymbol{\Sigma}$ *mit Wahrscheinlichkeit* 1.

Hierbei sind $\boldsymbol{X}_j^*$ *die entsprechend der empirischen Verteilung von* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ *gezogenen Resamples und* $\hat{\boldsymbol{\pi}}^* = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j^*$.

Beweis: zu 1. (Bishop et al. [104, S.470]): Es gilt

$$E\,\boldsymbol{X}_j = \pi_1 \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \pi_r \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_r \end{pmatrix},$$

$$\begin{aligned} \operatorname{cov}(\boldsymbol{X}_j) &= E(\boldsymbol{X}_j - \boldsymbol{\pi})(\boldsymbol{X}_j - \boldsymbol{\pi})' = E\,(\boldsymbol{X}_j \boldsymbol{X}_j' - \boldsymbol{X}_j \boldsymbol{\pi}' - \boldsymbol{\pi} \boldsymbol{X}_j' + \boldsymbol{\pi}\boldsymbol{\pi}') \\ &= \boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}' \end{aligned}$$

mit

$$\boldsymbol{D}_{\boldsymbol{\pi}} = (d_{uv})_{1 \le u,v \le r} = \begin{cases} \pi_u & \text{falls u = v} \\ 0 & \text{falls u} \neq \text{v} \end{cases}.$$

Für die momentenerzeugende Funktion gilt

$$M_{n\hat{\boldsymbol{\pi}}}(\boldsymbol{t}) = (M_{\boldsymbol{X}_1}(\boldsymbol{t}))^n = (E\, e^{\boldsymbol{t}'\boldsymbol{X}_1})^n = \left(\sum_{s=1}^{r} \pi_s e^{t_s}\right)^n .$$

Somit gilt für den Vektor $\boldsymbol{U}_n = \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$

$$M_{\boldsymbol{U}_n}(\boldsymbol{t}) = E\, e^{\boldsymbol{t}'\boldsymbol{U}_n} = e^{-\sqrt{n}\boldsymbol{t}'\boldsymbol{\pi}}\, M_{n\hat{\boldsymbol{\pi}}}(n^{-\frac{1}{2}}t) = e^{-\sqrt{n}\,\boldsymbol{t}'\boldsymbol{\pi}}\left(\sum_{s=1}^{r} \pi_s e^{n^{-\frac{1}{2}}t_s}\right)^n = \left(\sum_{s=1}^{r} \pi_s e^{n^{-\frac{1}{2}}(t_s - t'\pi)}\right)^n .$$

Da für die Exponentialfunktion in der Umgebung des Punktes Null

$$e^x = 1 + x + 0{,}5x^2 + o(x^2) \quad (x \to 0)$$

gilt, folgt

$$\begin{aligned} M_{\boldsymbol{U}_n}(\boldsymbol{t}) &= \left(\sum_{s=1}^{r} \pi_s (1 + n^{-\frac{1}{2}}(t_s - \boldsymbol{t}'\boldsymbol{\pi}) + \tfrac{1}{2n}(t_s - \boldsymbol{t}'\boldsymbol{\pi})^2 + o(n^{-1})\right)^n \\ &\qquad \left(\text{da:} \sum_{s=1}^{r} \pi_s (1 + n^{-\frac{1}{2}}(t_s - \boldsymbol{t}'\boldsymbol{\pi}) = 1\right) \\ &= \left(1 + \sum_{s=1}^{r} \pi_s (\tfrac{1}{2n}(t_s - \boldsymbol{t}'\boldsymbol{\pi})^2 + o(n^{-1})\right)^n \\ &= \left(1 + \tfrac{1}{2n}\boldsymbol{t}'(\boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}')\boldsymbol{t} + o(n^{-1})\right)^n . \end{aligned}$$

Daraus folgt, daß die momentenerzeugende Funktion von $\boldsymbol{U}_n$ gegen die momentenerzeugende Funktion eines multivariat normalverteilten Vektors mit Erwartungswert Null und Kovarianzmatrix $\boldsymbol{\Sigma} = (\boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}')$ konvergiert:

$$\lim_{n\to\infty} M_{\boldsymbol{U}_n}(\boldsymbol{t}) = e^{\frac{1}{2}\boldsymbol{t}'(\boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}')\boldsymbol{t}} \quad \left(\text{da:} \lim_{n\to\infty}(1 + a/n)^{\frac{n}{a}} = e^a\right).$$

Hieraus folgt die schwache Konvergenz der Verteilungen (1. Punkt des Satzes).
Die zweite Aussage des Satzes folgt aus $P_{\boldsymbol{X}_i^*} = \mathrm{M}(1, \hat{\boldsymbol{\pi}}),\ \hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_r)'$. Die dritte Aussage folgt aus dem starken Gesetz der Großen Zahlen.

Der folgende Satz wird sich später als sehr nützlich erweisen, um die Güte eines Tests unter Alternativhypothesen zu betrachten. Da die Beweisführung analog der von Satz 3.1 verläuft, soll er schon an dieser Stelle zitiert werden.

Satz 3.2: *Gilt statt der Verteilungsaussage in Satz 3.1* $\boldsymbol{X}_i \sim \mathrm{M}(1, \ddot{\boldsymbol{\pi}})$ $\ddot{\boldsymbol{\pi}} = (\ddot{\pi}_1, \ldots, \ddot{\pi}_r)'$ *mit den lokalen Alternativen*

$$\ddot{\pi}_s = \pi_s + \tfrac{1}{\sqrt{n}}\theta_s,\ s = 1, \ldots, r \text{ und } \sum_{s=1}^{r} \theta_s = 0,$$

dann folgt $P_{\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})} \Rightarrow N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. *Die Kovarianzmatrix ist hierbei dieselbe wie in Satz 3.1.*

Beweis (siehe Bishop et al.[104, S.471]).

Satz 3.3: *Es seien k+1 unabhängige, multinomialverteilte Stichproben* $\boldsymbol{X}_{01}, \ldots, \boldsymbol{X}_{0n_0}, \ldots,$ $\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k}$ $(\boldsymbol{X}_{ij} \sim M(1, \boldsymbol{\pi}_i), \boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ir})')$ *gegeben. Falls*

(A0) $0 < \zeta_0 \le \hat{\pi}_{is} \le 1 - \zeta_0 < 1,\ s = 1, \ldots, r,\ i = 0, \ldots, k$

(A1) $\min_{i=0,\ldots,k} n_i \to \infty$

(A2) $0 < \lambda_0 \le n_i / N \le 1 - \lambda_0 < 1,\ i = 0, \ldots, k$ *gilt,*

folgt für $\boldsymbol{U}_N = (\boldsymbol{U}'_{N0}, \ldots, \boldsymbol{U}'_{Nk})'$ *mit* $\boldsymbol{U}_{Ni} = \sqrt{N}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i)$, $\boldsymbol{\pi} = (\boldsymbol{\pi}_0', \ldots, \boldsymbol{\pi}_k')'$ *und* $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_0', \ldots, \hat{\boldsymbol{\pi}}_k')'$*:*

1. $P_{\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})} \Rightarrow N(\boldsymbol{0}, \boldsymbol{V})$ *mit* $\boldsymbol{V} = \bigoplus_{i=0}^{k} \frac{N}{n_i}(\boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i \boldsymbol{\pi}_i')$
2. $P_{\sqrt{N}(\hat{\boldsymbol{\pi}}^* - \hat{\boldsymbol{\pi}})} \Rightarrow N(\boldsymbol{0}, \hat{\boldsymbol{V}})$ *mit* $\hat{\boldsymbol{V}} = \bigoplus_{i=0}^{k} \frac{N}{n_i}(\boldsymbol{D}_{\hat{\boldsymbol{\pi}}_i} - \hat{\boldsymbol{\pi}}_i \hat{\boldsymbol{\pi}}_i')$ *(Resamples jeweils aus den einzelnen Stichproben ziehen)*
3. $\hat{\boldsymbol{V}} \xrightarrow[n \to \infty]{} \boldsymbol{V}$ *mit Wahrscheinlichkeit* 1.

Beweis:

Aufgrund der Cramer-Wold-Technik [105, S.18] ist nur zu zeigen, daß für jeden beliebigen Vektor $\boldsymbol{b} \in \mathbb{R}^{(k+1)r}$ gilt: $P_{\boldsymbol{b}'\boldsymbol{U}_N} \Rightarrow N(\boldsymbol{0}, \boldsymbol{b}'\boldsymbol{V}\boldsymbol{b})$.

Der Vektor $\boldsymbol{b}$ kann unterteilt werden in $\boldsymbol{b} = (\underbrace{b_1, \ldots, b_r}_{\boldsymbol{d}_0'}, \ldots, \underbrace{b_{rk+1}, \ldots, b_{r(k+1)}}_{\boldsymbol{d}_k'})'$. Somit folgt:

$$\boldsymbol{b}'\boldsymbol{U}_N = \sum_{i=0}^{k} \boldsymbol{d}_i'\boldsymbol{U}_{Ni} = \sum_{i=0}^{k} \boldsymbol{d}_i' \tfrac{\sqrt{N}}{\sqrt{n_i}} \sqrt{n_i}(\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_i).$$

Die Verteilung der einzelnen Summanden strebt aufgrund von Satz 3.1 und Cramer-Wold gegen $N(\boldsymbol{0}, \boldsymbol{d}_i' \tfrac{N}{n_i}(\boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i\boldsymbol{\pi}_i')\boldsymbol{d}_i)$. Aus der Unabhängigkeit der einzelnen Summanden folgt die erste Behauptung. Werden die Resamplingstichproben jeweils in den einzelnen Stichproben gezogen, kann die zweite Aussage genau wie die erste bewiesen werden. Die dritte Behauptung folgt wieder aus dem starken Gesetz der Großen Zahlen.

Bemerkung 3.1: Gilt die Hypothese $\pi_s = \pi_{0s} = \cdots = \pi_{ks}, s = 1,\ldots,r$, so kann auch aus der gepoolten Stichprobe gezogen werden. Es gilt dann allerdings mit

$$\boldsymbol{X}_i^* \sim \mathrm{M}(1, \hat{\boldsymbol{\pi}}_{H_0}), \hat{\boldsymbol{\pi}}_{H_0} = (\hat{\boldsymbol{\pi}}', \ldots, \hat{\boldsymbol{\pi}}')', \hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_r)',$$

$$\hat{\tilde{\boldsymbol{\pi}}}_0 = \tfrac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i, \ \hat{\boldsymbol{V}} = \bigoplus_{i=0}^{k} \tfrac{N}{n_i}(\boldsymbol{D}_{\hat{\boldsymbol{\pi}}} - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}').$$

Bemerkung 3.2: Da sich diskrete Verteilungen stets durch endlich viele Parameter beschreiben lassen (z. B. immer durch $\pi_s = P(X = K_s)$ bzw. $\pi_s = P(X = a_s)$, $s = 1,\ldots,r$), können Bootstraptests, angewandt auf diskrete Daten, als parametrische Bootstrapverfahren interpretiert werden.

Bemerkung 3.3: Für eine stetige Funktion g gilt somit, daß $g(\hat{\boldsymbol{\pi}})$ und $g(\hat{\boldsymbol{\pi}}^*)$ asymptotisch dieselbe Verteilung haben.

Beispiel 3.2

Gegeben sei folgende Scorematrix $\boldsymbol{A} = \bigoplus_{i=0}^{k} \boldsymbol{a}', \boldsymbol{a} = (a_1, \ldots, a_r)'$ (g ist dann eine Abbildung vom $\mathbb{R}^{(k+1)r}$ in den $\mathbb{R}^{(k+1)}$). Dann folgt aus Satz 3.2 (Bezeichnungen wie in Satz 3.2):

$$P_{\boldsymbol{A}\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})} \Rightarrow N(\boldsymbol{0}, \boldsymbol{AVA}') \text{ und } P_{\boldsymbol{A}\sqrt{N}(\hat{\boldsymbol{\pi}}^*-\hat{\boldsymbol{\pi}})} \Rightarrow N(\boldsymbol{0}, \boldsymbol{A}\hat{\boldsymbol{V}}\boldsymbol{A}').$$

3.2 Double-Bootstraptest

Efron und Tibshirani [55, S.338-357] beschreiben Methoden zur Verbesserung von Bootstrapverfahren. Die dort genannten Techniken führen zur Reduktion der Verzerrung oder der Varianz eines Schätzers. Auch kann die benötigte Resamplinganzahl durch effektivere Algorithmen vermindert werden. Der Double-Bootstrap ermöglicht eine bessere Approximation der Verteilung unter der Nullhypothese und benötigt zumindest bei Konfidenzintervallschätzungen keinen Varianzschätzer (bei gleicher Approximationsgüte [108]). Dies macht ihn gerade bei den in der vorliegenden Arbeit betrachteten sehr kleinen Fallzahlen interessant. Auf ihn wird daher nun näher eingegangen.

Es sei $\hat{G}_N(t|F) = P(T_N < t|\hat{F})$, und $H(u|F_0) = P(\hat{G}_N(t|F) < u|F_0)$ sei die Verteilung der Zufallsvariablen $\hat{G}_N(t|F)$ (hängt über die Schätzung $\hat{F}$ von der zufälligen Stichprobe ab), falls die Nullhypothese ($F = F_0$) gilt. Ist $\hat{G}_N(t|F)$ eine stetige Verteilung, so ist $H(u|F_0)$ annähernd auf dem Intervall [0,1] gleichverteilt ($P(H(u|F_0) \le v|F_0) = v$). Da aus einem großen Wert für $T_N = t_0$ ein großer Wert für $\hat{G}_N(t|F) = \hat{g}_N(t_0|F)$ und somit für $H(\hat{g}_N(t_0|F)|F_0) = P(\hat{G}_N(t|F) < \hat{g}_N(t_0|F)|F_0)$ folgt, stellt $H(u|F_0)$ eine alternative Teststatistik dar. Der Vorteil dieser Statistik liegt darin, daß ihre Verteilung von keinem unbekannten Parameter abhängt (ist demnach eine Pivotstatistik) und leicht mittels einer Bootstrapschätzung simuliert werden kann. Dazu werden Resamples entsprechend der Verteilung F_0 erzeugt und jeweils $H(u|F_0)$ bestimmt. $H(u|F_0)$ kann z. B. mittels einer parametrischen Verteilung oder wie beim Double-Bootstrap mittels eines weiteren Bootstrapexperiments bestimmt werden. Das heißt, aus der erzeugten Resamplingstichprobe werden erneut Resamplingstichproben gezogen und $H(u|F_0)$ geschätzt. Es werden also 2 Resamplingstufen durchlaufen. Ein Algorithmus für dieses Verfahren ist bei Davison und Hinkley [99, S.177] zu finden. Der Rechenaufwand ist enorm. Werden für die innere und für die äußere Schleife jeweils 1.000 Resamplingstichproben gezogen, so sind insgesamt 1.000.000 Resamples notwendig. Da über die Güte dieses Verfahrens bei diskreten Daten und sehr kleinen Fallzahlen bislang nichts bekannt ist, sind zur Beurteilung Simulationen notwendig. Bei 10.000 Simulationsschritten sind dies 10.000.000.000 Resamplingstichproben. Ein derartiges Simulationsexperiment ist selbst bei Verwendung von 128 Prozessoren (jeweils mit 300 Megaherz) auf einem Parallelrechner nur in beschränktem Maße durchführbar (die Rechenzeit für die Güteschätzung des U-Tests und des t-Tests beträgt für eine Parameterkonstellation abhängig von n_i ca. 40 Minuten (d. h. ca. 3,3 Tage bei einen gleichwertigen PC mit einem

Prozessor). Auf einem leistungsfähigen PC (Pentium II mit 300 Megaherz) in Verbindung mit SAS/IML ist eine solche Simulation zur Zeit zu aufwendig.

Algorithmus 3.1: Durchführung eines Double-Bootstraptests

5. Eingabe:
   - 5.1 k+1 = Anzahl der Behandlungen
   - 5.2 r = Anzahl der verschiedenen Atome der Verteilung
   - 5.3 $\boldsymbol{n} = (n_0,\ldots,n_c)'$ Vektor der Stichprobenumfänge
   - 5.4 $M_1 = 10.000$ Resamplinganzahl für die äußere Schleife
   - 5.5 $M_2 = 10.000$ Resamplinganzahl für die innere Schleife
6. Berechne: t_0, den Wert der Statistik für die Originalstichprobe
7. Initialisieren: Zähler1 = 1, Zähler2 = 1
8. FOR m_1 = 1 TO $M_1 - 1$
   - 8.1 Ziehe eine Bootstrapstichprobe $\boldsymbol{y}^*_{m_1}$ aus $\hat{F}_{H_0}$ und berechne $t^*_{m_1}$
   - 8.2 IF $t^*_{m_1} \geq t_0$ THEN Zähler = Zähler+1
   - 8.3 Schätze $\hat{F}_{m_1 H_0}$ aus $y^*_{m_1}$
   - 8.4 FOR m_2 = 1 TO $M_2 - 1$
     - 8.4.1 Ziehe eine Bootstrapstichprobe $\boldsymbol{y}^{**}_{m_1 m_2}$ aus $\hat{F}_{m_1 H_0}$ und berechne $t^{**}_{m_1 m_2}$
     - 8.4.2 IF $t^{**}_{m_1 m_2} \geq t^*_{m_1}$ THEN Zähler2 = Zähler2+1

     END
   - 8.5 Berechne $p^*_{m_1} = Z\ddot{a}hler2 \,/\, M_2$

   END
9. Berechne: $p = Z\ddot{a}hler1 \,/\, M_1$
10. Berechne: $p_{adj} = (1 + \sum_{m_1=1}^{M_1-1} I(p^*_{m_1} \leq p)) \,/\, M_1$

4 Parametrische Tests

4.1 Trendtests für normalverteilte Daten auf Basis von isotonen Schätzern

In diesem Kapitel werden ausgewählte parametrische Verfahren vorgestellt. Die Verfahren der beiden ersten Abschnitte beruhen auf Zufallsvariablen $X_{i1},...,X_{in_i}$ $(i=0,...,k)$, die voneinander unabhängig und normalverteilt $(P_{X_{ij}} = N(\mu_i, \sigma_i^2))$ sind. Diese klassischen parametrischen Verfahren sind zwar für normalverteilte Daten hergeleitet worden, sie sind aber bei geringen Abweichungen von den Normalverteilungsannahmen relativ robust[49; 109]. Aufgrund ihres hohen Bekanntheitsgrades und zum Teil fehlender Alternativen in einigen Softwarepaketen werden sie daher auch zur Auswertung ordinaler Daten genutzt. Chuang-Stein und Agresti [13] schreiben dazu: „*From our experience, treating ordinal data as continuous with constant variance can provide a useful approximation when the number of response categories is large, but may be inadequate when that number is less than five.*“
Quebe-Fehling [110, S.28] verweist darauf, daß Modelle, die auf einem Parameter basieren, aufgrund möglicher Fehlzuordnungen robuster sein können als Verfahren, die z. B. auf Vektoren von Wahrscheinlichkeiten (z. B. Multinomialverteilungsmodelle) beruhen. Er sieht darin sogar eine Rechtfertigung für die Auswertung diskreter Mittelwerte. Werden in einem landwirtschaftlichen oder gartenbaulichen Versuch pro Parzelle mehrere Einzelpflanzen bonitiert, so bezeichnet es Thöni [111] als naheliegend, die Mittelwerte der Parzellen varianzanalytisch auszuwerten, wenn die notwendigen Voraussetzungen erfüllt sind. Liegt nur ein Wert pro Parzelle (eine bonitierte Pflanze bzw. ein Schätzwert für die gesamte Parzelle) vor, so ist dies zwar eine entartete Mehrfachbonitur, das ändert jedoch nichts an der Aussagekraft der Gesamtmittelwerte je Versuchsglied. Thöni[112] schreibt außerdem: *„Ein Vorteil der Verwendung der Parzellen-Lokationswerte in einer varianzanalytischen Auswertung besteht im weiteren darin, daß für die individuellen Sortenvergleiche die üblichen Verfahren zur Berechnung von Standardfehlern und multiplen Vergleichen unmittelbar zur Verfügung stehen.“*
Ein Ziel der vorliegenden Arbeit besteht darin, diese Aussagen zu untermauern oder zu widerlegen. Die Simulationen sollen zu Erkenntnissen über das Risiko führen, welches bei der Anwendung einer nicht „angemessenen“ Methode (hier das parametrische Verfahren) besteht.
Ausgewählt wurden zum einen der Likelihood-Quotienten-Test von Bartholomew [26] und seine Modifikation nach Wright [113] aufgrund der hohen Güte bei normalverteilten Daten. Zum anderen wurden einfache und multiple Kontrasttests gewählt. Sie besitzen zum Teil eine annähernd gleich hohe Güte wie der Test von Bartholomew [114; 115] und bieten sich aufgrund

ihrer einfachen linearen Struktur in Hinblick auf Bemerkung 3.3 (der Konsistenz von Bootstrapverteilungen) geradezu an. Diese Verfahren setzen allerdings homogene Varianzen in den Gruppen voraus. Gerade bei ordinalen Daten ist diese Annahme kritisch zu beurteilen, denn eine hohe Dosis mit einer großen Wirkung wird oft zu Werten am unteren bzw. oberen Ende der Skala führen. Die Varianz in dieser Gruppe wird daher gegen Null tendieren. Als Konsequenz werden auch parametrische Verfahren für normalverteilte Daten vorgestellt, die ohne die Annahme homogener Varianzen auskommen. Hierbei handelt es sich um Tests, die bei Roth [116] bzw. Grimes und Federer [117] zu finden sind. Die zu diesen Verfahren zugehörigen Teststatistiken unterscheiden sich nur im Nenner vom Bartholomew-Test bzw. von den Kontraststatistiken. Ihre Verteilungen sind komplizierter und zum Teil nur asymptotisch gültig. Zusätzlich ist zu erwarten, daß diese Verfahren noch anfälliger gegen Varianzschätzer mit dem Wert Null sind, da die geschätzten Gruppenvarianzen zum Teil als einzelne Nenner auftreten.
Für die Anwendung dieser ausgewählten parametrischen Verfahren sind Scores notwendig. Die Ergebnisse sind daher von der Wahl der Scores abhängig.

Test von Bartholomew

Für das Hypothesenpaar

$$H_0 : \mu_0 = \mu_1 = \cdots = \mu_k \quad \text{vs.} \quad H_A : \mu_0 \le \mu_1 \le \cdots \le \mu_k \,(\mu_0 < \mu_k) \tag{4.1}$$

kann mittels der unter H_0 und H_A bekannten Wahrscheinlichkeitsdichten ein Likelihood-Quotient (Λ)

$$\Lambda = \frac{\max_{\boldsymbol{\mu} \in H_0} L(x_{01}, \ldots, x_{kn_k}, \boldsymbol{\mu}, \boldsymbol{\sigma})}{\max_{\boldsymbol{\mu} \in H_A} L(x_{01}, \ldots, x_{kn_k}, \boldsymbol{\mu}, \boldsymbol{\sigma})}$$

und damit ein Test konstruiert werden [26]. Sind die Varianzen homogen und bekannt, führen das Maximieren der logarithmierten Normalverteilungsdichten und weitere Umrechnungsschritte zu

$$\bar{\chi}^2_{01} = \sum_{i=0}^{k} w_i \,(\mu_i^* - \hat{\mu})^2 \,. \tag{4.2}$$

Hierbei sind μ_i^* die Schätzer für μ_i unter der Alternative (sie werden als isotone Schätzer bezeichnet), $\hat{\mu}$ der Schätzer für das gemeinsame μ unter der Nullhypothese und $w_i = n_i / \sigma_i^2$

die Gewichte. Die Schätzer lassen sich unter oben angegebenen Hypothesen wie folgt bestimmen:

$$\mu_i^* = \max_{0\le u\le i}\min_{i\le v\le k}\frac{\sum_{j=u}^{v} w_j\,\overline{X}_{j.}}{\sum_{j=u}^{v} w_j} \qquad \text{und} \qquad \hat{\mu} = \frac{\sum_{i=0}^{k} w_i\,\overline{X}_{i.}}{\sum_{i=0}^{k} w_i}\ . \tag{4.3}$$

Sind die Varianzen unbekannt und unterscheiden sie sich nur hinsichtlich bekannter Faktoren ($\sigma_i^2 = f_i\sigma^2$, $i = 0,\ldots,k$ und somit $w_i = n_i / f_i$), so führt dies zu folgender Statistik:

$$\overline{E}_{01}^2 = \frac{\sum_{i=0}^{k} w_i\,(\mu_i^* - \hat{\mu})^2}{\sum_{i=0}^{k} f_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{..})^2}\ . \tag{4.4}$$

Die Verteilungen der Statistiken (4.2) und (4.4) sind bekannt und werden im folgenden Satz beschrieben.

Satz 4.1: *Sind die Daten unabhängig und in den Gruppen identisch normalverteilt* $X_{ij} \sim N(\mu_i, \sigma_i^2)$, *so gilt für das Hypothesenpaar (4.1), daß die exakten parametrischen Verteilungen der Statistiken* $\overline{\chi}_{01}^2$ *und* $\overline{E}_{01}^2$ *für* $t > 0$ *durch*

$$P(\overline{\chi}_{01}^2 \ge t) = \sum_{l=2}^{k+1} P(l, k+1, \boldsymbol{w})\,P(\chi_{l-1}^2 \ge t)\,, \qquad w_i = \frac{n_i}{\sigma_i^2}\ ,$$

bzw.

$$P(\overline{E}_{01}^2 \ge t) = \sum_{l=2}^{k+1} P(l, k+1, \boldsymbol{w})\,P(B_{(l-1)/2,(N-l)/2} \ge t)\,, \qquad w_i = \frac{n_i}{f_i^2},$$

gegeben sind. Für $t = 0$ *gilt* $P(\overline{\chi}_0^2 = 0) = P(1, k+1, \boldsymbol{w})$ *bzw.* $P(\overline{E}_{01}^2 = 0) = P(1, k+1, \boldsymbol{w})$. *Die Größen* χ_{l-1}^2 *bzw.* $B_{(l-1)/2,(N-l)/2}$ *sind hierbei Chi-Quadrat- bzw. betaverteilte Zufallsvariablen. Die* $P(l, k+1, \boldsymbol{w})$ *werden als Levelwahrscheinlichkeiten bezeichnet. Sie drücken die Wahrscheinlichkeit aus, daß bei den gegebenen Gewichten die* $(k+1)$ *Mittelwerte zu l verschiedenen isotonen Schätzern führen.*

Beweis: siehe Robertson et al. [114, S. 70ff].

Problematisch ist hierbei die Berechnung der $P(l, k+1, \boldsymbol{w})$. Die folgenden zwei Fälle können dabei unterschieden werden:

a) Balancierter Fall (alle Gewichte gleich)
   Im balancierten Fall erfolgt die Berechnung mittels der Stirlingschen Zahlen 1. Art [114, S. 82]: $P(l, k+1, w) = \frac{|S_{k+1}^{l}|}{(k+1)!}$.
   Für die Berechnung der Zahlen S_{k+1}^{l} sei auf Nijenhuis und Wilf [78, S.174] verwiesen. Für $k = 3$ gilt beispielsweise $S_4^1 = -6, S_4^2 = 11, S_4^3 = -6, S_4^4 = 1$.

b) Unbalancierter Fall (ungleiche Stichprobenumfänge oder $\sigma_i^2 = f_i \sigma^2$)
   Im unbalancierten Fall gilt folgende allgemeine Rekursionsformel[114, S.77]:

$$P(m, k+1, w) = \sum_{(B_1, B_2, \ldots, B_m) \in L_{m(k+1)}} P(m, m; W_{B_1}, W_{B_2}, \ldots, W_{B_m}) \prod_{i=1}^{m} P(1, C_{B_i}; \boldsymbol{w}(B_i)) \,. \quad (4.5)$$

   $L_{\mathrm{m(k+1)}}$ stellt dabei die Menge aller möglichen Zerlegungen der $(k+1)$ Mittelwerte in m Teilmengen (B_i) dar. C_{B_i} entspricht der Mächtigkeit der Menge B_i, $w(B_i)$ dem Gewichtsvektor (bestehend aus den zu den Mittelwerten der Menge B_i zugehörigen Gewichten) und W_{B_i} der Summe dieser Gewichte. $P(m, m; W_{B_1}, W_{B_2}, \ldots, W_{B_m})$ werden als Orthantwahrscheinlichkeiten bezeichnet und können für $k+1 = 2, 3, 4$ [114, S.77ff] exakt berechnet werden. Für größere k müssen numerische Integrationsverfahren genutzt werden, wie sie Bretz[115] beschreibt.

Ein Algorithmus zur Berechnung der Levelwahrscheinlichkeiten ist u. a. von Bohrer und Chow [118] beschrieben worden. Für die Simulationen wurde ein SAS-Programm erstellt, welches die exakten Formeln nutzt und die Fälle balanciert und unbalanciert unterscheidet. Die Bestimmung von (4.5) wurde durch die folgenden Schritte realisiert:

- Erzeugung von $L_{\mathrm{m(k+1)}}$: Wie schon in Kapitel 2 und 3 wird wieder davon ausgegangen, daß jede natürliche Zahl $k+1$ in $m \le k+1$ ganzzahlige Summanden zerlegt werden kann. Zum Beispiel kann die Zahl 4 in folgende Summanden zerlegt werden:
  $4 = 4\,(m = 1),\ 4 = 1 + 3\,(m = 2),\ 4 = 1 + 2 + 1\,(m = 3),\ 4 = 1 + 1 + 1 + 1\,(m = 4)$

Weitere Zerlegungen ergeben sich durch Permutieren der angegebenen Zerlegungen. Ein Element $(B_1,\ldots,B_m)$ aus $L_{\mathrm{m(k+1)}}$ kann erzeugt werden, indem entsprechend einer Zerlegung der Zahl die Mittelwerte der Reihe nach in die Mengen B_l einsortiert werden (z. B.: $1,2,1 \Longrightarrow B_1 = \{\overline{X}_{0.}\}, B_2 = \{\overline{X}_{1.}, \overline{X}_{2.}\}, B_3 = \{\overline{X}_{3.}\}$). Zur Erzeugung der Zerlegungen kann wieder der Algorithmus Nexcom von Nijenhuis und Wilf[78, S. 49] genutzt werden. Beginnend mit $t = 2$ wird $L_{\mathrm{it}}(j,..,j+t)$ für $t = 2,\ldots, k+1$, $i = 2,\ldots,t$ und $j = 1,\ldots,k-t+1$ bestimmt, wobei der zusätzliche Index j angibt, daß L_{it} für die Mittelwerte $\overline{X}_{j.},\ldots,\overline{X}_{(j+t).}$ bestimmt wird. Für jedes Element aus $L_{\mathrm{it}}(j,..,j+t)$ wird nach der Erzeugung das Produkt aus Formel (4.5) berechnet. Dies ist möglich, da in vorhergehenden Schritten die benötigten Levelwahrscheinlichkeiten niedrigerer Ordnung bereits berechnet und gespeichert wurden und

$$P(1,1,(W_{B_1},W_{B_2})) = P(1,2,(W_{B_1},W_{B_2})) = P(2,2,(W_{B_1},W_{B_2})) = 0.5$$

gilt.

- Bestimmung der Orthantwahrscheinlichkeiten: Die $P(m,m;W_{B_1},W_{B_2},\ldots,W_{B_m})$ können in Abhängigkeit von der Dimension k und der gewünschten Genauigkeit mittels bestimmter Module berechnet werden, die von Bretz[115] beschrieben werden.

Test von Wright

Für unbekannte, aber homogene Gruppenvarianzen schlägt Wright[113] vor, die unbekannte Varianz mittels der Statistik

$$S_I^2 = \frac{1}{N-k-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i.})^2$$

zu schätzen. Dies führt zu einer Teststatistik, die sich von $\overline{E}_{01}^2$ nur im Nenner unterscheidet.

Satz 4.2: *Unter den Voraussetzungen von Satz 4.1 und dem Hypothesenpaar (4.1) ist die exakte parametrische Verteilung der Statistik*

$$\overline{T}_{01}^2 = \frac{\sum_{i=0}^{k} w_i(\mu_i^* - \hat{\mu})^2}{S_I^2}, \quad w_i = \frac{n_i}{f_i^2},$$

durch

$$P(\overline{T}_{01}^2 \geq t) = \sum_{l=2}^{k+1} P(l, k+1, \boldsymbol{w})\, P((l-1)\, F_{(l-1),(N-l)} \geq t)$$

gegeben. Die Größen $F_{(l-1),(N-l)}$ *sind hierbei F-verteilte Zufallsvariablen.* $P(l, k+1, \boldsymbol{w})$ *sind wieder die Levelwahrscheinlichkeiten aus Satz 4.1 .*

Beweis: siehe Wright [113].

Wright zeigt, daß dieser Test asymptotisch äquivalent zu $\overline{E}_{01}^2$ ist. Desweiteren zeigen seine geschätzten Powerwerte, daß dieser Test robuster gegen Abweichungen von den Ordnungsrestriktionen ist. Seine Bedeutung wird im Abschnitt 4.2 noch einmal erläutert.

Die Beschreibung der Gütefunktionen sowohl für den Bartholomew-Test als auch für den Wright-Test, ist ein sehr komplexes Problem, für das es noch keine allgemeine Lösung gibt. Auf die Darstellung von Ergebnissen für die wenigen Spezialfälle ($k+1=3$ [114, S.87ff], $k+1=4$ [26]) wird verzichtet.

Bemerkung 4.1: Die Alternativhypothese kann geometrisch durch einen polyhydralen Kegel *K* des $\mathbb{R}^{k+1}$ beschrieben werden. Dabei ist ein polyhydraler Kegel eine Teilmenge des $\mathbb{R}^{k+1}$, dessen Punkte eine endliche Anzahl von homogenen linearen Ungleichungen erfüllen. Der Vektor der isotonen Schätzer stellt dann die Projektion des Mittelwertvektors auf den Kegel*K* dar [114, S.109ff].

Aufgrund dieser Bemerkung soll hier noch etwas genauer auf den Bartholomew-Test und seine geometrische Interpretation eingegangen werden. Es wird sich zeigen, daß auch ohne Normalverteilungsannahmen die Teststatistik des Bartholomew-Tests sehr gut interpretierbar ist.

Bei genauer Betrachtung der Formel 4.3 ist zu erkennen, daß die isotonen Schätzer gewichtete Mittel benachbarter Mittelwerte sind. Zum Beispiel kommen für die isotonen Schätzer im Fall $k=2$ nur folgende Werte in Frage:

a) $$\boldsymbol{\mu}^* = (\mu_0^*, \mu_1^*, \mu_2^*)' = \left(\frac{n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}}{n_0 + n_1}, \frac{n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}}{n_0 + n_1}, \overline{X}_{2.} \right)'$$

b) $$\boldsymbol{\mu}^* = (\mu_0^*, \mu_1^*, \mu_2^*)' = \left(\overline{X}_{0.}, \frac{n_1 \overline{X}_{1.} + n_2 \overline{X}_{2.}}{n_1 + n_2}, \frac{n_1 \overline{X}_{1.} + n_2 \overline{X}_{2.}}{n_1 + n_2} \right)'$$

c) $\boldsymbol{\mu}^* = (\mu_0^*, \mu_1^*, \mu_2^*)' = \left(\overline{X}_{0.}, \overline{X}_{1.}, \overline{X}_{2.}\right)'$

oder d)

$$\boldsymbol{\mu}^* = (\mu_0^*, \mu_1^*, \mu_2^*)' = \left(\frac{n_0\,\overline{X}_{0.} + n_1\,\overline{X}_{1.} + n_2\,\overline{X}_{2.}}{n_0+n_1+n_2}, \frac{n_0\,\overline{X}_{0.} + n_1\,\overline{X}_{1.} + n_2\,\overline{X}_{2.}}{n_0+n_1+n_2}, \frac{n_0\,\overline{X}_{0.} + n_1\,\overline{X}_{1.} + n_2\,\overline{X}_{2.}}{n_0+n_1+n_2} \right)'.$$

Allgemein kann dieser Prozeß der Mittelung durch Matrixoperationen beschrieben werden. Das heißt, für die isotonen Schätzer gibt es stets eine $(k+1)$-dimensionale quadratische Matrix $(\boldsymbol{C}_l)$ mit $\boldsymbol{\mu}^* = \boldsymbol{C}_l \overline{\boldsymbol{X}}$. Für $\boldsymbol{C}_l$ bieten sich im Fall $k=2$ z. B. die folgenden Matrizen an:

$$\boldsymbol{C}_1 = \begin{pmatrix} \frac{n_0}{n_1+n_0} & \frac{n_1}{n_1+n_0} & 0 \\ \frac{n_0}{n_1+n_0} & \frac{n_1}{n_1+n_0} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; \boldsymbol{C}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{n_1}{n_2+n_1} & \frac{n_2}{n_2+n_1} \\ 0 & \frac{n_1}{n_2+n_1} & \frac{n_2}{n_2+n_1} \end{pmatrix}, \; \boldsymbol{C}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{C}_0 = \frac{1}{N}\boldsymbol{n}\boldsymbol{l}_3' = \frac{1}{N}\begin{pmatrix} n_0 & n_1 & n_2 \\ n_0 & n_1 & n_2 \\ n_0 & n_1 & n_2 \end{pmatrix}.$$

Die Suche nach einer Matrix $\boldsymbol{C}_l$ mit $\boldsymbol{\mu}^* = \boldsymbol{C}_l \overline{\boldsymbol{X}}$ kann stets auf ein System M von 2^k Matrizen eingeschränkt werden, das im folgenden vorgestellt wird.

<u>Erzeugung von M</u>

Zunächst werden analog zu Abschnitt 4.1 alle Zerlegungen einer natürlichen Zahl in positive ganze Zahlen generiert. Auf der Basis jeder dieser Zerlegung wird anschließend eine Matrix definiert. Ist z. B. $\boldsymbol{h}_l = (h_{l0}, \ldots, h_{lm_l})'$ eine Zerlegung für die Zahl $k+1$ $\left(k+1 = \sum_{s=0}^{m_l} h_{ls}, h_{ls} > 0\right)$ und wird mit $\boldsymbol{D}_{\boldsymbol{n}}$ die Diagonalmatrix mit den Hauptdiagonalelementen $D_{ii} = n_i \; (i = 0, \ldots, k)$ bezeichnet, so kann mittels des Vektors $\boldsymbol{h}_l$, den Zahlen $g_{l(-1)} = 0, \quad g_{ls} = g_{l(s-1)} + h_{ls}$, $(s = 0, \ldots, m_l)$, $\widetilde{n}_{ls} = \sum_{w=g_{l(s-1)}}^{g_{ls}-1} n_w, s = 0, \ldots, m_l$ und der Matrix $\boldsymbol{M}_l$

$$M_{luv} = \begin{cases} \left(\widetilde{n}_{ls}\right)^{-1} & g_{ls-1} \le u < g_{ls} \text{ und } g_{l(s-1)} \le v < g_{ls} \\ 0 & \text{sonst} \end{cases} \qquad s = 0, \ldots, m_l, \; 0 \le u, v \le k$$

die Matrix $\boldsymbol{C}_l = \boldsymbol{M}_l \boldsymbol{D}_n$ definiert werden. Das Ergebnis der Matrixmultiplikation $\boldsymbol{C}_l \overline{\boldsymbol{X}}$ ist dann ein Vektor der Form

$$\boldsymbol{C}_l \overline{\boldsymbol{X}} = \begin{pmatrix} \frac{n_0 \overline{X}_{0.} + \ldots + n_{g_{l0}-1} \overline{X}_{g_{l0}-1.}}{n_0 + \ldots + n_{g_{l0}-1}} \boldsymbol{1}_{g_{l0}-1} \\ \vdots \\ \frac{n_{g_{lm_l-1}} \overline{X}_{g_{lm_l-1}.} + \ldots + n_{g_{lm_l}-1} \overline{X}_{g_{lm_l}-1.}}{n_{g_{lm_l-1}} + \ldots + n_{g_{lm_l}-1}} \boldsymbol{1}_{g_{lm_l}-1} \end{pmatrix}.$$

Alle auf diese Weise erzeugten Matrizen seien im System M zusammengefaßt. Wird ein Mittelwertpaar $(\overline{X}_{i.}, \overline{X}_{i+1.})$ mit $o_{li} = 0$ bzw. $o_{li} = 1$ gekennzeichnet, falls für die Indizes $g_{ls-1} \le i, i+1 < g_{ls}$ bzw. $g_{ls-1} \le i < g_{ls} \le i+1$ gilt, so stellt $o_{l0} \ldots o_{lk-1}$ eine Dualzahl dar, die einer natürlichen Zahl zwischen 0 und $2^k - 1$ entspricht. Jeder dieser Dualzahlen entspricht genau eine Möglichkeit der Mittelung (sind gerade die Variationen mit Wiederholungen der zwei Elemente $(<, =)$ der Ordnung k) und somit eine Matrix $\boldsymbol{C}_l$. Die Mächtigkeit des Systems M ist demnach 2^k.

<u>Interpretation des Systems M</u>

Wird das Hypothesenpaar 4.1 durch die Räume $\boldsymbol{H}_0$ und $\boldsymbol{H}_A = \bigcup_{s=1}^{2^k-1} H_{A_i}$ ausgedrückt, hier beispielhaft für $k = 3$ dargestellt:

Teilraum von $\boldsymbol{H}_A$	Dualzahl	natürliche Zahl (Index für $\boldsymbol{C_l}$)
$\boldsymbol{H}_{A_1} = \{\boldsymbol{\mu}: \ \mu_0 = \mu_1 = \mu_2 < \mu_3\}$	0 0 1	1
$\boldsymbol{H}_{A_2} = \{\boldsymbol{\mu}: \ \mu_0 = \mu_1 < \mu_2 = \mu_3\}$	0 1 0	2
$\boldsymbol{H}_{A_3} = \{\boldsymbol{\mu}: \ \mu_0 = \mu_1 < \mu_2 < \mu_3\}$	0 1 1	3
$\boldsymbol{H}_{A_4} = \{\boldsymbol{\mu}: \ \mu_0 < \mu_1 = \mu_2 = \mu_3\}$	1 0 0	4
$\boldsymbol{H}_{A_5} = \{\boldsymbol{\mu}: \ \mu_0 < \mu_1 = \mu_2 < \mu_3\}$	1 0 1	5
$\boldsymbol{H}_{A_6} = \{\boldsymbol{\mu}: \ \mu_0 < \mu_1 < \mu_2 = \mu_3\}$	1 1 0	6
$\boldsymbol{H}_{A_7} = \{\boldsymbol{\mu}: \ \mu_0 < \mu_1 < \mu_2 < \mu_3\}$	1 1 1	7
$\boldsymbol{H_0}$		
$\boldsymbol{H}_0 = \{\mu: \ \mu_0 = \mu_1 = \mu_2 = \mu_3\}$	0 0 0	0,

so gelten die folgenden Aussagen.

Hilfssatz 4.1:

1. *Liegt der Mittelwertvektor im Alternativraum* $(\overline{\boldsymbol{X}} \in \boldsymbol{H}_A)$, *so folgt* $\boldsymbol{C}_l \overline{\boldsymbol{X}} \in \boldsymbol{H}_A$ $(l = 1,\ldots,2^k - 1)$.
2. *Für* $l = 1,\ldots,2^k - 1$ *und* $\overline{\boldsymbol{X}} \in \boldsymbol{H}_{A_l}$ *folgt* $\boldsymbol{C}_l \overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}$ *(und somit* $\boldsymbol{C}_l \overline{\boldsymbol{X}} \in \boldsymbol{H}_{A_l}$*).*
3. *Für* $l = 1,\ldots,2^k - 1$ *folgt* $(\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{\boldsymbol{X}}, \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{1}_{k+1} \overline{X}_{..}) = 0$.
4. $||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{X}_{..} \boldsymbol{1}_{k+1}||^2_{k+1} = ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}}||^2_{k+1} + ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{X}_{..} \boldsymbol{1}_{k+1}||^2_{k+1}$.

Beweis:

Die erste Implikation resultiert aus $\overline{X}_{0.} \leq \cdots \leq \overline{X}_{k.}$ und den Ungleichungen

$$\frac{n_{g_{l(s-1)}} \overline{X}_{g_{l(s-1)}.} + \ldots + n_{g_{ls}-1} \overline{X}_{g_{ls}-1.}}{n_{g_{l(s-1)}} + \ldots + n_{g_{ls}-1}} \leq \overline{X}_{g_{ls}-1.} \leq \overline{X}_{g_{ls}.} \leq \frac{n_{g_{ls}} \overline{X}_{g_{ls}.} + \ldots + n_{g_{l(s+1)}-1} \overline{X}_{g_{l(s+1)}-1.}}{n_{g_{ls}} + \ldots + n_{g_{l(s+1)}-1}}.$$

Aufgrund von

$$\frac{n_{g_{l(s-1)}} \overline{X}_{g_{l(s-1)}.} + \ldots + n_{g_{ls}-1} \overline{X}_{g_{ls}-1.}}{n_{g_{l(s-1)}} + \ldots + n_{g_{ls}-1}} = \frac{n_{g_{l(s-1)}} \overline{X}_{g_{l(s-1)}.} + \ldots + n_{g_{ls}-1} \overline{X}_{g_{ls-1}.}}{n_{g_{l(s-1)}} + \ldots + n_{g_{ls}-1}} = \overline{X}_{g_{l(s-1)}.}$$

ergibt sich die zweite Aussage. Die dritte Aussage besagt, daß die beiden Vektoren $(\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}})$ und $(\boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}} \overline{X}_{..} \boldsymbol{1}_{k+1})$ zueinander orthogonal sind (und somit im Fall normalverteilter Daten voneinander unabhängig sind). Sie folgt aus $\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l = \boldsymbol{C}_l \boldsymbol{D}_{\boldsymbol{n}}$, $\boldsymbol{C}_l \boldsymbol{C}_l = \boldsymbol{C}_l$, $\boldsymbol{C}_l \boldsymbol{1}_{k+1} = \boldsymbol{1}_{k+1}$,

$$\begin{aligned}
&(\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \overline{\boldsymbol{X}}, \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..}) \\
&= (\boldsymbol{D}_{\boldsymbol{n}} \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}})' (\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..}) \\
&= \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' \boldsymbol{C}_l' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..} + \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..} \\
&= \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' (\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l)' \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l' \boldsymbol{1}_{k+1} \overline{X}_{..} + \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..} \\
&= \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \boldsymbol{C}_l \overline{\boldsymbol{X}} - \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l' \boldsymbol{1}_{k+1} \overline{X}_{..} + \overline{\boldsymbol{X}}' \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l' \boldsymbol{1}_{k+1} \overline{X}_{..} \\
&= 0
\end{aligned}$$

und

$$\begin{aligned}
&(\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}^{-1} (\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \overline{\boldsymbol{X}}), \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}^{-1} (\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..})) \\
&= \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}^{-1} (\boldsymbol{D}_{\boldsymbol{n}} \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} C_l \overline{\boldsymbol{X}})' (\boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{C}_l \overline{\boldsymbol{X}} - \boldsymbol{D}_{\boldsymbol{n}} \boldsymbol{1}_{k+1} \overline{X}_{..}) \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}^{-1} \\
&= 0.
\end{aligned}$$

Die vierte Aussage folgt aus der dritten und

$$
\begin{aligned}
&||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} = ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} + \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} \\
&= ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}}||^2_{k+1} + ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} + 2(\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}}, \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}) \\
&= ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}}||^2_{k+1} + ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} \ .
\end{aligned}
$$

Die Matrizen $\boldsymbol{C}_l$ bilden somit einen Vektor $\boldsymbol{Z} \in \mathbb{R}^{k+1}$ (hier beispielhaft für $\boldsymbol{Z} = \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}}$ gezeigt) orthogonal in den Raum $\boldsymbol{K}_{A_l} = \{\boldsymbol{\mu}: \mu_0 = \cdots = \mu_{g_{ls}-1}, \ldots, \mu_{g_{lmj-1}} = \cdots = \mu_k\} \supset \boldsymbol{H}_{A_l}$ ab. Ist der Abstand von $\overline{\boldsymbol{X}}$ zu $\boldsymbol{H}_{A_l}$ nicht „zu groß“, gilt $\boldsymbol{C}_l\overline{\boldsymbol{X}} \in \boldsymbol{H}_{A_l}$. Bei „zu großem“ Abstand folgt für einige bzw. sogar für alle Matrizen $\boldsymbol{C}_l$ $(l = 1,\ldots,2^k - 1)$ hingegen $\boldsymbol{C}_l\overline{\boldsymbol{X}} \notin \boldsymbol{H}_{A_l}$. Im Gegensatz dazu bildet die Matrix $\boldsymbol{C}_0$ den Mittelwertvektor stets in den Raum der Nullhypothese ab ($\boldsymbol{C}_0\overline{\boldsymbol{X}} = \overline{X}_{..}\boldsymbol{1}_{k+1} \in \boldsymbol{H}_0$). In diesem Sinne können die Matrizen $\boldsymbol{C}_l$ $(l = 1,\ldots,2^k - 1)$ als „Pseudoprojektionsmatrizen“ bezeichnet werden. Wird mit $M(\overline{\boldsymbol{X}}) = \{\boldsymbol{C}_l: \boldsymbol{C}_l\overline{\boldsymbol{X}} \in \boldsymbol{H}_A\}$ die Menge der Matrizen bezeichnet, die den Mittelwertvektor $\overline{\boldsymbol{X}}$ in den Raum der Alternativhypothese abbilden, so kann der Bartholomew-Test auch in folgender Form dargestellt werden:

$$
\begin{aligned}
\overline{E}^2_{01} &= \max_{\boldsymbol{C}_l \in M(\overline{\boldsymbol{X}})} \frac{||\boldsymbol{D}_{\sqrt{n}}(\boldsymbol{C}_l\overline{\boldsymbol{X}} - N^{-1}\boldsymbol{n}\boldsymbol{1}'_{k+1}\overline{\boldsymbol{X}})||^2_{k+1}}{\sum_{i=0}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{..})^2} \\
&= \left(\max_{\widetilde{\boldsymbol{C}}_l \in \widetilde{M}(\overline{\boldsymbol{X}})} \frac{||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}'_N\widetilde{\boldsymbol{X}}||^2_N}{||\widetilde{\boldsymbol{X}} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}'_N\widetilde{\boldsymbol{X}}||^2_N} \right).
\end{aligned}
$$

$$
\overline{E}^2_{01} = 0, \text{ falls } M(\overline{\boldsymbol{X}}) \text{ bzw. } \widetilde{M}(\overline{\boldsymbol{X}}) \text{ die leere Menge ist}
$$

$$
(\widetilde{\boldsymbol{C}}_l = \bigoplus_{s=0}^{m_l} \tfrac{1}{\widetilde{n}_{ls}}\boldsymbol{1}_{\widetilde{n}_{ls}}\boldsymbol{1}'_{\widetilde{n}_{ls}},\ \widetilde{\boldsymbol{X}} = (\boldsymbol{X}'_0,\ldots,\boldsymbol{X}'_k)' \in \mathbb{R}^N, \quad \widetilde{M}(\overline{\boldsymbol{X}}) = \{\widetilde{\boldsymbol{C}}_l: \boldsymbol{C}_l \in M(\overline{\boldsymbol{X}})\}\ ).
$$

Aus der Gleichung

$$
||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} = ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{X}_{..}\boldsymbol{1}_{k+1}||^2_{k+1} - ||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\overline{\boldsymbol{X}} - \boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}}||^2_{k+1}
$$

folgt, daß $\overline{E}^2_{01}$ genau dann maximal ist, wenn $\boldsymbol{C}_l\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}$ gilt. Falls $\boldsymbol{\mu}^* = \boldsymbol{C}_l\overline{\boldsymbol{X}}$ ($\widetilde{\boldsymbol{\mu}}^* = \widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}$) gilt und der Bartholomew-Test zur Ablehnung der Nullhypothese führt, deutet dies darauf hin,

daß der Vektor $\overline{\boldsymbol{X}}$ ($\widetilde{\boldsymbol{X}}$) nicht weit vom Raum entfernt oder gar im Raum $\boldsymbol{H}_{A_l}$ ($\widetilde{\boldsymbol{H}}_{A_l}$) liegt. Die geometrische Darstellung beruht nur auf dem Mittelwertvektor (Schätzer für den wahren Erwartungswertvektor) und ist somit auch für andere Verteilungsmodelle gut interpretierbar. Zumindest mittels Permutations- bzw. Bootstraptests ist die Verteilung stets verfügbar, woraus die Relevanz der Bartholomew-Statistik auch für andere Verteilungsmodelle folgt.
Die komplexe Verteilung von $\overline{E}_{01}^2$ (folgt aus der Einschränkung der Maximumbildung auf $M(\overline{\boldsymbol{X}})$) und die technisch aufwendige Programmierung haben dazu geführt, daß immer wieder Tests vorgeschlagen wurden, die eine annähernd gleich hohe Güte wie $\overline{E}_{01}^2$ besitzen. Zum einen wurde dies erreicht, indem die Alternative so geändert wurde, daß sie eine gute Überdeckung mit der ursprünglichen Alternative besitzt, ihre geometrische Darstellung aber einfacher ist und somit die Verteilung der zugehörigen Likelihood-Quotienten-Tests von einfacherer Gestalt sind [119;120, S.114ff]. Diese Tests werden in der vorliegenden Arbeit nicht weiter untersucht. Zum anderen wurden einfachere Tests, z. B. Kontrasttests, vorgeschlagen [114, S. 115], auf die in Kapitel 4.2 genauer eingegangen wird.

Tests von Roth

Roth [116] schlägt für heterogene Varianzen zwei parametrische Tests vor, die dem Test von Bartholomew sehr ähneln. Die erste Variante ist eine Verallgemeinerung des k-Stichprobentests von Welch [122]:

$$WT = \frac{\sum_{s=1}^{l} \widetilde{w}_s (\mu_s^* - \hat{\mu})^2}{(N-l)[1+2((l-2)/(l^2-1))\sum_s h_s]}.$$

Hierbei sind μ_i^* wieder die isotonen Schätzer für μ_i unter der Alternative. Die Anzahl der verschieden μ_i^* wird dabei durch *l* bezeichnet. Die weiteren Größen berechnen sich nach

$$w_i = n_i / S_i^2 \;,\; h_s = (1-\widetilde{w}_s / \widetilde{w}_\cdot)/(\widetilde{n}_s - 1) \;,\; \widetilde{n}_s = \sum_{i \in B_s} n_i \;,\; \widetilde{w}_s = \sum_{i \in B_s} w_i \;.$$

Als zweite Variante schlägt Roth eine Verallgemeinerung des k-Stichprobentests von Brown-Forsythe [123] auf die Trendsituation vor:

$$BFT = \frac{\sum_{s=1}^{l} \tilde{n}_s (\mu_s^* - \hat{\mu})^2}{(l-1)(N-l)^{-1} \sum_{i=0}^{k} b_i}$$

mit $b_i = S_i^2 \left(\frac{n_i}{\tilde{n}_s} - \frac{n_i}{N} \right)$.

Hilfssatz 4.2: *Eine approximative parametrische Verteilung für WT bzw. BFT ist gegeben durch*

$$P(WT > t) \approx \sum_{\{B:|B|\geq 2\}} P(B, k+1)\, P((l-1) F_{(l-1), f_B} \geq (N-l)\, t)$$

$$P(BFT > t) \approx \sum_{\{B:|B|\geq 2\}} P(B, k+1)\, P((l-1) F_{(l-1), g_B} \geq (N-l)\, t) .$$

Die Freiheitsgrade der F-verteilten Zufallsgröße werden dabei mittels

$$1/f_B = 3\sum_{s=1}^{l} h_s / (l^2 - 1) \text{ bzw.} 1/g_B = \sum_{s=1}^{l} q_s^2 / (\tilde{n}_s - 1),\ q_s = d_s / \sum_{s=1}^{l} d_s,\ d_s = \sum_{j \in B_s} b_j$$

berechnet.

Beweis: siehe Roth [116].

Die Mengen des Systems $\{B : |B| > 1\}$ stellen die Zerlegungen der (k+1) Mittelwerte in l Teilmengen dar. Die Summation aller $P(B, k+1)$ für ein festes l ergibt gerade die Level-wahrscheinlichkeiten $P(l, k+1, \boldsymbol{w})$, die in diesem Fall auch von den Varianzschätzern abhängen .

Roth [116] vergleicht diese beiden Tests mit $\overline{E}_{01}^2$ für normalverteilte Daten und sehr kleine Fallzahlen (z. B. $k = 3, n_i = 3$). Die Ergebnisse und Empfehlungen bei kleinen Fallzahlen können wie folgt zusammengefaßt werden:

1. WT ist sehr liberal.
2. $\overline{E}_{01}^2$ ist leicht liberal bei heterogenen Varianzen, BFT hingegen nicht.
3. BFT hat eine sehr viel geringere Power als $\overline{E}_{01}^2$.

Daher schlägt Roth die Nutzung von $\overline{E}_{01}^2$ vor. Meng et al. [124] empfehlen für den Fall heterogener Varianzen und sehr kleiner Fallzahlen ebenfalls, die Varianzheterogenität zu ignorieren und $\overline{T}_{01}^2$ zu nutzen.

Die Durchführung eines exakten Permutationstests ist am einfachsten für $\overline{E}_{01}^2$, da diese Statistik im wesentlichen auf den Zähler (*Z*)

$$\overline{E}_{01}^2(Z) = \sum_{i=0}^{k} w_i \, \mu_i^{*2}$$

reduziert werden kann (die weiteren Anteile der eigentlichen Teststatistik sind permutationsinvariant). Die vollständige Ausnutzung vom Netzwerkalgorithmus konnte allerdings auch für diese einfache Statistik nicht programmiert werden, weil kein Abbruchkriterium gefunden wurde (der Anteil von μ_i^* an der Teststatistik hängt nicht nur von den Stichproben 0 bis *i* ab; siehe Bedingung 1 und 2). Zumindest für kleine Stichproben wurde ein C-Programm erstellt, das die Durchführung eines exakten Permutationstests auf Basis von $\overline{E}_{01}^2(Z)$ und $\overline{T}_{01}^2$ und die Berechnung der unbedingten Power auch im Rahmen von Simulationen zuläßt.

Shi und Meng [121] konstruieren mittels $\overline{E}_{01}^2$ Bootstraptests, wobei sie die Fälle „homogene Varianzen“ (Fall 1) und „heterogene Varianzen“ (Fall 2) unterscheiden. Die Resamplingstichproben werden im Fall 1 für jede Gruppe aus

$$RS_Z = \{X_{ij} - \overline{X}_{i.}, i = 0,\ldots,k, j = 1,\ldots,n_i\}$$

gezogen und im Fall 2 für Gruppe *i* aus

$$RS_{Zi} = \{X_{ij} - \overline{X}_{i.}, j = 1,\ldots,n_i\}, \; i = 0,\ldots,k.$$

Satz 4.3: *Es werde angenommen, daß die Verteilungsfunktion der i-ten Gruppe durch* $F_i(x) = F(\frac{x-\mu_i}{\sigma_i})$ *gegeben ist und die zweiten Momente existieren (*$EX_{ij}^2 < \infty$*). Sind zusätzlich die Bedingungen*

a) $\min(n_0,\ldots,n_k) \to \infty$ *und*

b) $n_i / N \to \lambda_i > 0, \; i = 0,\ldots,k$

erfüllt, so gilt für $\overline{E}_{01}^2(B_s) = \sum_{i=0}^{k} w_i (\mu_i^* - \hat{\mu})^2$ unter $H_0^{\mu} : \mu_0 = \mu_1 = \cdots = \mu_k$ *:*

1. $P(\overline{E}_{01}^2 > t) \to \sum_{l=2}^{k+1} P(l, k+1, \boldsymbol{w})\, P(\chi_{l-1}^2 > t)\ \ (t > 0)$

Fall 1: $w_i = \lambda_i / S_I^2, i = 0,\ldots,k$ *) bzw. Fall 2* $w_i = \lambda_i / S_i^{\,2}, i = 0,\ldots,k$ *.*

2. Die Bootstrapverteilung P^* *von* $\overline{E}_{01}^{2*}$ *strebt fast sicher gegen dieselbe Verteilung wie die Verteilung von* $\overline{E}_{01}^2$ *.*

3. Für das $(1-\alpha)$ *Quantil* $q_{1-\alpha}$ *der Bootstrapverteilung gilt* $P(\overline{E}_{01}^2 \geq q_{1-\alpha}) \to \alpha$ *.*

Beweis: Shi und Meng [121].

Sind die Voraussetzungen des Satzes erfüllt, sind auf (4.2) und (4.4) basierende Bootstraptests theoretisch gerechtfertigt. Die von Shi und Meng [121] durchgeführten Simulationen für verschiedene stetige Verteilungsfamilien zeigen, daß es selbst im Fall heterogener Varianzen besser ist, die Stichproben aus RS_Z zu ziehen und den gepoolten Varianzschätzer zu benutzen, da diese Prozedur zu höherer Power führt. Für heterogene Varianzen zeigen ihre Ergebnisse, daß die Tests eher konservativ sind.

Der WT-Test und der BFT-Test gehen von unterschiedlichen Varianzen der Stichproben aus. Nach Fisher und Hall [101] sollten die Resamplingstichproben daher aus den einzelnen Stichproben gezogen werden (ein Permutationstest auf Basis dieser 2 Statistiken wurde daher nicht konstruiert). Da schon im stetigen Fall bei kleinen Stichproben Varianzschätzer mit dem Wert Null Probleme bereiten [121], müssen im diskreten Fall vor allem beim WT-Test noch größere Schwierigkeiten erwartet werden. Aber auch beim BFT-Test, beim Wright-Test und beim Bartholomew-Test stellen Varianzschätzer mit dem Wert Null ein Problem dar. Wird mit A^{Test} das Ereignis „Teststatistik nicht berechenbar (Division durch Null)“ bezeichnet, so wird dies anhand der berechenbaren Wahrscheinlichkeiten für dieses Ereignis deutlich. Wird mit $\boldsymbol{y}_i$ der Häufigkeitsvektor der (transformierten) Stichprobe $\boldsymbol{X}_i$ und mit $\boldsymbol{y}^{pool}$ der Häufigkeitsvektor der gepoolten (transformierten) Stichprobe bezeichnet, so ergibt sich für das Ereignis „$S_i^2(\boldsymbol{X}_{mi}^*) = 0$“ in Abhängigkeit vom Resamplingraum folgende Wahrscheinlichkeit:

Fall 1) $RS_Z = \{x_{ij} - \bar{x}_{i\cdot}, i = 0,\ldots,k, j = 1,\ldots,n_i\}$ bzw. $RS = \{x_{ij}, i = 0,\ldots,k, j = 1,\ldots,n_i\}$

$$P^*(A_i^{pool}) = P^*(S_i^2(\boldsymbol{X}_{mi}^*) = 0 | y^{pool} = (y_1^{pool},\ldots,y_{r_{pool}}^{pool})') = \sum_{s=1}^{r_{pool}} \left(\frac{y_s^{pool}}{N}\right)^{n_i} = \left(\frac{1}{N}\right)^{n_i} \sum_{s=1}^{r_{pool}} {y_s^{pool}}^{n_i}$$

$$\left(\overset{r_{pool} < N}{>} \left(\frac{1}{N}\right)^{n_i} \sum_{s=1}^{r_{pool}} y_s^{pool} = \left(\frac{1}{N}\right)^{n_i} N = P^*(S_i^2(\boldsymbol{X}_{mi}^*) = 0 | N = r_{pool}) \right)$$

Fall 2) $RS_{Zi} = \{x_{ij} - \bar{x}_{i.}, j = 1,\ldots,n_i\}$ bzw. $RS_i = \{x_{ij}, j = 1,\ldots,n_i\}$

$$P^*(A_i^{\text{single}}) = P^*(S_i^2(\boldsymbol{X}_{mi}^*) = 0 | \boldsymbol{y}_i = (y_{i1},\ldots,y_{ir_i})') = \sum_{s=1}^{r_i}\left(\frac{y_{is}}{n_i}\right)^{n_i} = \left(\frac{1}{n_i}\right)^{n_i}\sum_{s=1}^{r_i} y_{is}^{n_i}$$

$$\left(\overset{r_i<n_i}{>} \left(\frac{1}{n_i}\right)^{n_i}\sum_{s=1}^{r_i} y_{is} \quad = \left(\frac{1}{n_i}\right)^{n_i} n_i = P^*(S_i^2(\boldsymbol{X}_{mi}^*) = 0 | r_i = n_i)\right).$$

Für die einzelnen Tests beträgt die Wahrscheinlichkeit für das Ereignis A^{Test}:

$$P^*(A^{\overline{E}_{01}^2}) = \sum_{s=1}^{r_{pool}}\left(\frac{y_s^{pool}}{N}\right)^N$$

$$P^*(A^{\overline{T}_{01}^2}) = P^*\left(\bigcap_{i=0}^{k} A_i^{pool}\right) = \prod_{i=0}^{k} P^*(A_i^{\text{pool}})$$

$$P^*(A^{WT}) = P^*\left(\bigcup_{i=0}^{k} A_i^{\text{single}}\right) = 1 - \prod_{i=0}^{k}(1 - P^*(A_i^{\text{single}}))$$

$$P^*(A^{BFT}) = P^*\left(\bigcap_{i=0}^{k} A_i^{\text{single}}\right) = \prod_{i=0}^{k} P^*(A_i^{\text{single}}).$$

Aufgrund von Tabelle 4.1 ist zu erwarten, daß die Teststatistiken des WT-Tests bzw. des BFT-Tests bei sehr kleinen Stichprobenumfängen für Bootstraptests ungeeignet sind. Wird der Varianzschätzer weggelassen, so führt dies bei beiden Statistiken zum Zähler des Bartholomew-Tests. Das Zentrieren kann beim Ziehen aus der gepoolten Stichprobe sowohl zur Erhöhung als auch zur Abnahme der Wahrscheinlichkeit eines nichtpositiven Varianzschätzers führen. In den Bootstrapsimulationen wurde daher die Originalstatistik des Bartholomew-Tests und des Wright-Tests (Resamplingraum: RS_Z, RS) sowie der Zähler des Bartholomew-Tests (Resamplingraum: RS_Z, RS, RS_{Zi}, RS_i) betrachtet.

	Originalstichprobe ($n_0 = n_1 = n_2 = 3$)			
	$\boldsymbol{X}_0 = (1,2,3)'$ $\boldsymbol{X}_1 = (4,5,6)'$ $\boldsymbol{X}_2 = (7,8,9)'$		$\boldsymbol{X}_0 = (1,2,3)'$ $\boldsymbol{X}_1 = (4,5,6)'$ $\boldsymbol{X}_2 = (7,8,8)'$	
$\overline{E}_{01}^2$ RS	2,323E-8	0,0002	1,340E-6	0,013
RS_Z	1,524E-4	0,7822	5,288E-6	0,514
$\overline{T}_{01}^2$ RS	1,882E-6	0,0186	8,711E-6	0,083
RS_Z	1,371E-3	1	9,275E-5	1
WT	0,297	1	0,473	1
BFT	1,371E-3	1	4,115E-3	1

Tabelle 4.1: Wahrscheinlichkeit für einen Varianzschätzer mit dem Wert Null (linker Eintrag) und Wahrscheinlichkeit, daß mindestens für eine von 10.000 Resamplingstichproben ein Varianzschätzer mit dem Wert Null auftritt

Weitere parametrische Tests, die auf isotonen Schätzern beruhen, schlagen Williams[125] und Marcus [126] vor. Liegen keine konkreten Informationen über die spezielle Struktur der Alternative vor, empfiehlt Marcus [126] den Likelihood-Quotienten-Test von Bartholomew. Robertson et al. [114, S.109] empfehlen ebenfalls $\overline{E}_{01}^2$, falls die Alternative nicht noch konkreter eingeschränkt werden kann. Für eine weitere Diskussion des Williams-Tests und des Marcus-Tests wird auf Bretz [115] verwiesen.

4.2 Parametrische Kontrasttests für normalverteilte Daten

In diesem Abschnitt werden Kontrasttests für normalverteilte Daten ($X_{ij} \sim N(\mu_i, \sigma_i^2)$) vorgestellt, deren Verteilung sowohl unter der Nullhypothese als auch unter der Alternative berechenbar ist. Für sie können die Power und benötigten Fallzahlen zum Erreichen einer vorgegebenen Güte berechnet werden.

Gilt die Nullhypothese $H_0 : \mu_0 = \mu_1 = \cdots = \mu_k$, so folgt für jeden Vektor $\boldsymbol{c}$ mit

$$\sum_{i=0}^{k} c_i = 0 \qquad L = \sum_{i=0}^{k} c_i \mu_i = 0 .$$

Ein geschätzter Wert für L, der sich deutlich von Null unterscheidet, spricht demnach gegen die Nullhypothese. Offensichtlich wird L dann besonders groß, wenn die größten Gewichte

bei den größten Erwartungswerten stehen. Erfüllen die Erwartungswerte $(\mu_0,...,\mu_k)$ die Alternativhypothese $H_A: \mu_0 \le \mu_1 \le \cdots \le \mu_k$, so wird L mit hoher Wahrscheinlichkeit einen großen Wert annehmen, falls die Komponenten von $\boldsymbol{c}$ den Bedingungen

$$c_0 \le c_1 \le \cdots \le c_k \tag{4.6}$$

genügen. Gilt (4.6) können demnach auf L basierende Tests für das Hypothesenpaar (4.1) konstruiert werden. Als Schätzer für μ_i bietet sich im Normalverteilungsmodell der Mittelwert der Gruppe $(\overline{X}_{i.})$ an. Die Linearkombination L kann so erwartungstreu durch

$$\hat{L} = \sum_{i=0}^{k} c_i \overline{X}_{i.} \quad (E\,\hat{L} = L)$$

geschätzt werden. Die Varianz dieses Schätzers beträgt

$$Var(\hat{L}) = \sum_{i=0}^{k} \tfrac{c_i^2 \sigma_i^2}{n_i}.$$

Sind die Varianzen unbekannt, aber homogen, so kann die gemeinsame Varianz durch den erwartungstreuen Schätzer $S_I^2 = \dfrac{1}{N-k-1} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2$ geschätzt werden. Die Statistik

$$T = \frac{\sum_{i=0}^{k} c_i \overline{X}_{i.}}{S_I \sqrt{\sum_{i=0}^{k} \tfrac{c_i^2}{n_i}}}$$

wird dann als Kontraststatistik (oder kurz Kontrast) bezeichnet. Im Spezialfall $k = 1$ ist T gerade die Statistik des t-Tests. Gelten die Nullhypothese und die Bedingung $\sum_{i=0}^{k} c_i = 0$ (die letzte Bedingung soll im weiteren stets erfüllt sein), so ist T t-verteilt mit $N-k-1$ Freiheitsgraden. Unter der Alternative ist T nichtzentral t-verteilt mit dem Nichtzentralitätsparameter

$$\delta(\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{n},\sigma^2)=\frac{\sum_{i=0}^{k} c_i\,\mu_i}{\sigma\sqrt{\sum_{i=0}^{k} c_i^2/n_i}}\ .$$

Die Güte eines auf T beruhenden Tests hängt unter der Alternativhypothese also direkt vom Kontrastvektor $\boldsymbol{c}$ ab. Für feste Stichprobenumfänge, eine feste Varianz und ein fest vorgegebenes Profil der Erwartungswerte ist genau der Kontrastvektor optimal (höchste Güte), der den Nichtzentralitätsparameter maximiert. Nach wenigen Umformungen folgt

$$\delta(\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{n},\sigma^2)=\frac{\sqrt{\sum_{i=0}^{k} n_i(\mu_i-\overline{\mu})^2}}{\sigma}\,\frac{\sum_{i=0}^{k}\frac{c_i}{\sqrt{n_i}}\sqrt{n_i}(\mu_i-\overline{\mu})}{\sqrt{\sum_{i=0}^{k}\frac{c_i^2}{n_i}}\sqrt{\sum_{i=0}^{k} n_i(\mu_i-\overline{\mu})^2}}\qquad \text{mit } \overline{\mu}=\tfrac{1}{N}\sum_{i=0}^{k} n_i\mu_i\ .$$

Der erste Faktor der rechten Seite hängt nicht von $\boldsymbol{c}$ ab und spielt demnach keine Rolle beim Maximieren. Der zweite Term ist gerade der Kosinus des Winkels β, der von den Vektoren $(\frac{c_0}{\sqrt{n_0}},\ldots,\frac{c_k}{\sqrt{n_k}})'$ und $(\sqrt{n_0}(\mu_0-\overline{\mu}),\ldots,\sqrt{n_k}(\mu_k-\overline{\mu}))'$ im $\mathbb{R}^{k+1}$ aufgespannt wird. Da der Kosinus als maximalen Wert eins annehmen kann, führt der Kontrastvektor

$$\boldsymbol{c}_{opt}=(n_0(\mu_0-\overline{\mu}),\ldots,n_k(\mu_k-\overline{\mu}))'$$

zu einem optimalen Kontrasttest. Mögliche Profile und die zugehörigen optimalen Kontraste sind in den folgenden Tabellen dargestellt.

Profil $(\mu_0,\mu_1,\mu_2,\mu_3)$	**Optimaler Kontrast**	**Name**
$(0,0,0,\theta)$	(-1, -1, -1, 3)	Helmert-Kontrast [12]
$(0,\theta,\theta,\theta)$	(-3, 1, 1, 1)	Reverse–Helmert-Kontrast [127]
$(0,\theta,\theta,2\theta)$	(-1, 0, 0, 1)	Paarweiser Kontrast [128]
$(0,\theta,2\theta,3\theta)$	(-1,5, -0,5, 0,5, 1,5)	Linearer Kontrast [129]

Tabelle 4.2: Beispiele für optimale Kontrastvektoren: $k=3$ und gleiche Stichprobenumfänge

Profil $(\mu_0,\mu_1,\mu_2,\mu_3)$	**Optimaler Kontrast**
$(0,0,0,\theta)$	(-2, -1, -1, 4)
$(0,0,\theta,\theta)$	(-4, -2, 3, 3)
$(0,\theta,\theta,\theta)$	(-6, 2, 2, 2)
$(0,\theta,\theta,2\theta)$	(-8, 1, 1, 6)
$(0,\theta,2\theta,3\theta)$	(-12, -1, 4, 9)

Tabelle 4.3: Beispiele für optimale Kontrastvektoren: $k=3$ und $\boldsymbol{n}=(10,5,5,5)'$

Profil $(\mu_0,\mu_1,\mu_2,\mu_3)$	**Optimaler Kontrast**
$(0,0,0,\theta)$	$(-\frac{1}{6},-\frac{1}{3},-\frac{1}{3},\frac{5}{6})$
$(0,0,\theta,\theta)$	$(-\frac{1}{2},-1,1,\frac{1}{2})$
$(0,\theta,\theta,\theta)$	$(-\frac{5}{6},\frac{1}{3},\frac{1}{3},\frac{1}{6})$
$(0,\theta,\theta,2\theta)$	$(-1,0,0,1)$
$(0,\theta,2\theta,3\theta)$	$(-3,-1,1,3)$

Tabelle 4.4: Beispiele für optimale Kontrastvektoren: $k=3$ und $\boldsymbol{n}=(5,10,10,5)'$

Unbalancierte Fallzahlen können demnach zu Kontrastkoeffizienten führen, die die Bedingung (4.6) nicht erfüllen. Soll eine falsche Nullhypothese mit hoher Wahrscheinlichkeit abgelehnt werden, kann die Bedingung (4.6) in diesem Fall zu streng sein. Werden die Kontrasttests in der Form

$$T=\frac{\sum_{i=0}^{k} n_i c_i \overline{X}_{i.}}{S_I\sqrt{\sum_{i=0}^{k} n_i c_i^2}}$$

definiert, tritt dieses Problem nicht auf. Die Herleitung eines optimalen Kontrasts in dieser Form verläuft analog zu oben. Abelson und Tukey [130] konstruierten im balancierten Fall einen Kontrasttest, dessen minimale Power maximal unter allen Kontrasten ist. Schaafsma und Smid [131] kommen über einen anderen Weg zum gleichen Test für unbalancierte Anlagen. Wie bei der Bestimmung des optimalen Kontrastvektors kann dieser Maxmin-Kontrast mit Hilfe von Winkeln im $\mathbb{R}^{k+1}$ und dem Nichtzentralitätsparameter veranschaulicht werden. Um den Nichtzentralitätsparameter für alle zulässigen Profile groß zu halten, muß der

gesuchte Vektor gerade einen möglichst kleinen Winkel mit allen zulässigen Profilen bilden. Intuitiv ist es vorstellbar, daß dies ein Vektor ist, der im Zentrum des Kegels liegt. Die Koeffizienten sind im balancierten Fall durch

$$c_i = \sqrt{i\,(1-\tfrac{i}{k+1})} - \sqrt{(i+1)\,(1-\tfrac{i+1}{k+1})} \quad (i=0,...,k)$$

definiert. Für $k = 3$ sind die Koeffizienten beispielsweise durch

$$\boldsymbol{c}_{mm} = (\text{-}0{,}866025,\ \text{-}0{,}133975,\ 0{,}1339746,\ 0{,}8660254)'$$

bestimmt. Aufgrund der extremen Abhängigkeit von den Kontrastvektoren sind Kontraststatistiken nur zu empfehlen, wenn a priori spezielle Informationen über die Alternative vorliegen. Ohne zusätzliches Wissen ist nur der Maxmin-Kontrast zu empfehlen. Der Paarweise Kontrast ist im Prinzip ein t-Test, nur daß die Varianz durch mehr als zwei Gruppen geschätzt wird. Dies führt zu einer Erhöhung des Freiheitsgrades (somit besitzt er eine höhere Güte als der t-Test, falls wirklich homogene Varianzen vorliegen). Eine Ablehnung der Nullhypothese kann mittels des Paarweisen Kontrastes nur erfolgen, wenn $\overline{X}_{k.} > \overline{X}_{0.}$ gilt. Bei allen anderen Kontrasttests gilt dies so nicht, da auch bei $\mu_0 \le \mu_1 \le \mu_2 \ge \mu_3 \cdots \ge \mu_k\,(\mu_0 = \mu_k)$ das Testergebnis „Verwerfung der Nullhypothese“ lauten kann. Für einen Vergleich mehrerer verschiedener Kontrasttests mittels asymptotischer relativer Effizienzen sei auf Neuhäuser [82] verwiesen.

Werden die Erwartungswerte geschätzt und in den optimalen Kontrastvektor eingesetzt, so ergibt sich für den Nichtzentralitätsparameter $\delta(\boldsymbol{c}_{opt}, \boldsymbol{\mu}, \boldsymbol{n}, \sigma^2)$

$$\delta(\hat{\boldsymbol{c}}_{opt}, \hat{\boldsymbol{\mu}}, \boldsymbol{n}, \sigma^2) = \frac{\sqrt{\sum_{i=0}^{k} n_i (\hat{\mu}_i - \hat{\mu})^2}}{\sigma} \frac{\sum_{i=0}^{k} \frac{n_i(\hat{\mu}_i - \bar{\mu})}{\sqrt{n_i}} \sqrt{n_i}(\hat{\mu}_i - \hat{\mu})}{\sum_{i=0}^{k} n_i (\hat{\mu}_i - \hat{\mu})^2}$$

$$= \frac{\sqrt{\sum_{i=0}^{k} n_i (\hat{\mu}_i - \hat{\mu})^2}}{\sigma}.$$

Aufgrund der Eigenschaften der isotonen Schätzer aus Gleichung (4.3) folgt, daß $\hat{\mu}_i = \mu_i^*$ $\delta(\boldsymbol{c}_{opt}, \mu, \boldsymbol{n}, \sigma^2)$ (unter Annahme totaler Ordnung und $\sigma_i^2 = \sigma^2, i = 0,\ldots,k$) zum Maximum

führen. Wird $\hat{c}_{opt}$ in die Kontraststatistik eingesetzt, so führt dies nach wenigen Umformungen zu

$$T = \frac{\sum_{i=0}^{k} n_i (\mu_i^* - \hat{\mu})(\overline{X}_{i.} - \mu_i^* + \mu_i^* - \hat{\mu})}{S_I \sqrt{\sum_{i=0}^{k} n_i (\mu_i^* - \hat{\mu})^2}}$$

$$= \frac{\sum_{i=0}^{k} n_i (\mu_i^* - \hat{\mu})^2 + \sum_{i=0}^{k} n_i (\overline{X}_{i.} - \mu_i^*)(\mu_i^* - \hat{\mu})}{S_I \sqrt{\sum_{i=0}^{k} n_i (\mu_i^* - \hat{\mu})^2}} \text{ (der 2. Summand ist = 0 Hilfssatz 4.1)}$$

$$= \frac{\sqrt{\sum_{i=0}^{k} n_i (\mu_i^* - \hat{\mu})^2}}{S_I} = \sqrt{\overline{T}_{01}^2}\,.$$

Bei unbekannter Varianz führt diese Kontraststatistik zu der von Wright vorgeschlagenen Teststatistik, während sie bei bekannter Varianz zum Test von Bartholomew($\overline{\chi}_{01}^2$) führt.
Cohen und Sackrowitz [132; 133] zeigen, daß studentisierte Statistiken (z. B. Kontrasttests, aber auch der Test von Wright) im Gegensatz zum Test von Bartholomew für total geordnete Alternativen unzulässig sind (können gleichmäßig verbessert werden). Sie schlagen daher vor, den gepoolten Varianzschätzer durch den Schätzer für die Gesamtvarianz zu ersetzen:

$$S_{II}^2 = \frac{1}{N-1} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{..})^2\,.$$

Aufgrund der im Abschnitt 4.1 dargestellten Wahrscheinlichkeiten für einen Varianzschätzer mit dem Wert Null, ist S_{II}^2 zudem eher für Bootstrapverfahren geeignet. Bei Permutationstests kann S_{II}^2 vernachlässigt werden, da er invariant gegen Permutationen der Daten ist.
Um die (einfachen) Kontraststatistiken trotzdem nutzen zu können, werden in der Literatur unterschiedliche Wege beschrieben. Die Idee ist dabei stets, mehrere Kontrasttests gleichzeitig durchzuführen, was zu einem multiplen Testproblem führt. Um einen α-Test zu erhalten, sind α-Adjustierungen notwendig. In der einfachsten Variante werden m Kontrasttests durchgeführt, und die Nullhypothese wird verworfen, falls ein Test auf dem Niveau α / m zur Ablehnung führt (Bonferroni-Adjustierung). Dieses Verfahren nutzt aber die zum Teil hohe Korrelation zwischen den Kontraststatistiken nicht aus. Effizienter ist es, die gemeinsame

Verteilung der verschiedenen Kontraststatistiken zu berücksichtigen. Die Teststatistik eines multiplen Kontrasttests ist als Maximum über mehrere Kontraststatistiken definiert:

$$T^{\max} = \max(T_1,\ldots,T_m) \quad \text{mit } T_l = \frac{\sum_{i=0}^{k} c_{li}\overline{X}_{i.}}{S_I\sqrt{\sum_{i=0}^{k}\frac{c_{li}^2}{n_i}}},\ l=1,\ldots,m\ .$$

Die gemeinsame Verteilung der *m* Kontraststatistiken ist unter der Nullhypothese eine zentrale *m*-dimensionale t-Verteilung mit *N-k-1* Freiheitsgraden und der Korrelationsmatrix $\boldsymbol{R}$, die durch die Koeffizienten (4.7) bestimmt ist[134]:

$$R_{uv} = \frac{\sum_{i=0}^{k} c_{ui}\,c_{vi}/n_i}{\sqrt{\sum_{i=0}^{k} c_{ui}^2/n_i}\sqrt{\sum_{i=0}^{k} c_{vi}^2/n_i}} \text{ mit } u,v = 1,\ldots,m\,. \qquad (4.7)$$

Die Wahl der Kontrastvektoren stellt ein komplexes Problem dar. McDermott und Mudholkar [135] schreiben dazu: *„In the multiple-contrast approach, the choice is based on a set of a priori important alternatives.“* Und ebenda: *„In general this choice is not unique in the absence of prior information regarding the relative importance of various alternatives.“*
Im Prinzip kann jede Matrix $\boldsymbol{C}$ (im weiteren stets als Kontrastmatrix bezeichnet) vom Format $m\times(k+1)$, die der Bedingung $\boldsymbol{C}\,\boldsymbol{1}_{k+1} = \boldsymbol{0}_m$ genügt ($\boldsymbol{1}_{k+1} = (1,\ldots,1)' \in \mathbb{R}^{k+1}$, $\boldsymbol{0}_m = (0,,\ldots,0)' \in \mathbb{R}^m$), zur Bildung eines multiplen Kontrastes genutzt werden,.

Beispiele für multiple Kontrasttests

1. Neuhäuser[82] schlägt in der einfachsten Variante einen bivariaten Kontrast vor, der aus zwei Kontrasten besteht, die eine hohe Güte für extrem konvexe (Helmert-Kontrast) bzw. für extrem konkave (Reverse-Helmert-Kontrast) Profile besitzen.
   Beispiel für $k = 3$:

$$\boldsymbol{C}_{BK} = \begin{pmatrix} -1 & -1 & -1 & 3 \\ -3 & 1 & 1 & 1 \end{pmatrix}.$$

2. Eine einfache Verbesserung des bivariaten Kontrastes von Neuhäuser ist durch die Hinzunahme des linearen Kontrastes möglich.

Beispiel für $k = 3$:

$$\boldsymbol{C}_{TK} = \begin{pmatrix} -1 & -1 & -1 & 3 \\ -3 & -1 & 1 & 3 \\ -3 & 1 & 1 & 1 \end{pmatrix}.$$

3. Sugiura [136] konstruiert einen approximativen verallgemeinerten Bayes-Test, der sich letztendlich von dem durch Hirotsu et al. [137] vorgeschlagenen multiplen Kontrast nur durch eine Normierungskonstante unterscheidet (im balancierten Fall stimmen die Tests überein):

$$c_{li} = -\sqrt{\tfrac{k+1}{(k-l)(l+1)}}(k-l)/(k+1), l = 0,\ldots,k-1, i \le l$$
$$c_{li} = \sqrt{\tfrac{k+1}{(k-l)(l+1)}}(l+1)/(k+1), l = 0,\ldots,k-1, i > l.$$

Beispiel für $k = 3$:

$$\mathrm{C}_{\mathrm{SUG}} = \begin{pmatrix} \text{-0,866025} & \text{0,2886751} & \text{0,2886751} & \text{0,2886751} \\ \text{-0,5} & \text{-0,5} & \text{0,5} & \text{0,5} \\ \text{-0,2886751} & \text{-0,2886751} & \text{-0,2886751} & \text{0,866025} \end{pmatrix}.$$

Bretz [115] nutzt die theoretischen Aussagen von Abelson und Tukey [130] zur Konstruktion eines multiplen Kontrasttests mit $2^k - 1$ Kontrasten. In Analogie zum Bartholomew-Test und den isotonen Schätzern soll dieser Test hier auf einem anderen Weg eingeführt werden.
In Abschnitt 4.1 wurde gezeigt, daß der Bartholomew-Test mittels Matrizen darstellbar ist. Die Nullhypothese wird abgelehnt, wenn der Abstand von $\widetilde{\boldsymbol{X}}$ zu $\widetilde{\boldsymbol{H}}_{A_l}$ kleiner ist als der Abstand von $\widetilde{\boldsymbol{X}}$ zu $\widetilde{\boldsymbol{H}}_0$ (im weiteren wird stets die N-dimensionale Darstellung genutzt, die Herleitung mittels der $k+1$-dimensionalen Darstellung verläuft etwas aufwendiger, jedoch analog). Da die Matrizen $\widetilde{\boldsymbol{C}}_l$ in den Raum

$$\widetilde{\boldsymbol{K}}_A = \bigcup_{l=1}^{2^k-1} \widetilde{\boldsymbol{K}}_{A_l}$$

abbilden (beispielhaft für $k = 3$ dargestellt):

$$\widetilde{\boldsymbol{K}}_{A_1} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad = \quad \mu_{11} = \cdots = \mu_{1n_1} \quad = \quad \mu_{21} = \cdots = \mu_{2n_2} \quad , \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_2} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad = \quad \mu_{11} = \cdots = \mu_{1n_1} \quad , \quad \mu_{21} = \cdots = \mu_{2n_2} \quad = \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_3} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad = \quad \mu_{11} = \cdots = \mu_{1n_1} \quad , \quad \mu_{21} = \cdots = \mu_{2n_2} \quad , \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_4} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad , \quad \mu_{11} = \cdots = \mu_{1n_1} \quad = \quad \mu_{21} = \cdots = \mu_{2n_2} \quad = \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_5} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad , \quad \mu_{11} = \cdots = \mu_{1n_1} \quad = \quad \mu_{21} = \cdots = \mu_{2n_2} \quad , \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_6} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad , \quad \mu_{11} = \cdots = \mu_{1n_1} \quad , \quad \mu_{21} = \cdots = \mu_{2n_2} \quad = \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$
$$\widetilde{\boldsymbol{K}}_{A_7} = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad , \quad \mu_{11} = \cdots = \mu_{1n_1} \quad , \quad \mu_{21} = \cdots = \mu_{2n_2} \quad , \quad \mu_{31} = \cdots = \mu_{3n_3}\}$$

$$(\widetilde{\boldsymbol{H}}_0 = \{\widetilde{\boldsymbol{\mu}}: \quad \mu_{01} = \cdots = \mu_{0n_0} \quad = \quad \mu_{11} = \cdots = \mu_{1n_1} \quad = \quad \mu_{21} = \cdots = \mu_{2n_2} \quad = \quad \mu_{31} = \cdots = \mu_{3n_3}\}),$$

wird die Maximierung auf $\widetilde{\boldsymbol{C}}_l \in M(\widetilde{\boldsymbol{X}})$ eingeschränkt. Dies verkompliziert die Verteilungsaussagen zwar entscheidend, ist jedoch notwendig, um gerade ansteigende Trends mit hoher Wahrscheinlichkeit aufzudecken. Ist der Abstand zwischen $\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}$ und $\widetilde{\boldsymbol{X}}$ groß, so ist dies äquivalent zu einem großen Wert für den Kosinus des Winkels (β), der durch die Vektoren $\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}$ und $\widetilde{\boldsymbol{X}}$ im $\mathbb{R}^N$ aufgespannt wird:

$$\cos(\beta) = \frac{(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}})}{||\widetilde{\boldsymbol{X}}||_N \, ||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}||_N} \quad .$$

Allerdings ist dieser Winkel auch dann klein, wenn die Koordinaten des Vektors $\widetilde{\boldsymbol{X}}$ abfallen. Im folgenden wird aber gezeigt, daß mittels der Matrizen $\widetilde{\boldsymbol{C}}_l$ ein Test konstruiert werden kann, der im wesentlichen auf dieser Idee beruht.
Aus der Gleichung

$$||\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}||_N^2 = \min_{0 \le l < 2^k - 1} ||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}||_N^2$$

und der Umformung

$$\begin{aligned}
&||\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}||_N^2 = (\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}})'(\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}) \\
&= \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} + \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} \\
&= \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} \\
&= ||\widetilde{\boldsymbol{X}}||_N^2 - ||\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}}||_N^2 \le ||\widetilde{\boldsymbol{X}}||_N^2 - ||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}||_N^2
\end{aligned}$$

folgt

$$||\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}}||_N^2 = \max_{0\le l<2^k-1} ||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}||_N^2$$

und damit

$$||\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}}||_N^2 = \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{C}}_L\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{X}}) = \max_{0\le l<2^k-1}(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{X}})\,. \tag{4.8}$$

Es sei nun $\widetilde{\boldsymbol{c}} \in \mathbb{R}^N$ ein beliebige Vektor, der den Bedingungen

$$\widetilde{\boldsymbol{c}} = (c_0\boldsymbol{1}'_{n_0},\ldots,c_k\boldsymbol{1}'_{n_k})',\ \sum_{i=0}^{k} n_i c_i = 0\,,\ \ c_0 \le\cdots\le c_k$$

genügt. Wenn $\widetilde{\boldsymbol{X}} \in H_{A_l}$ gilt, dann sollte auch $(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{X}}) = (\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}})$ groß sein. Aus (4.8) folgt somit

$$\begin{aligned}\widetilde{\boldsymbol{c}}'\widetilde{\boldsymbol{C}}_{L_c}\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{C}}_{L_c}\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{c}}) = \max_{0\le l<2^k-1}(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{c}}) = \max_{0\le l<2^k-1}\widetilde{\boldsymbol{c}}'\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}\\ \left(\boldsymbol{c}'\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_{L_c}\overline{\boldsymbol{X}} = \max_{0\le l<2^k-1}\boldsymbol{c}'\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\overline{\boldsymbol{X}}\right).\end{aligned} \tag{4.9}$$

Die Umformungsschritte führen also ganz automatisch zu einem multiplen Kontrast. Offen ist jedoch noch die Wahl des Vektors $\widetilde{\boldsymbol{c}}\,(\boldsymbol{c})$. Wünschenswert ist, daß selbst nach einer Normierung mit $||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N^2$ $(||\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\boldsymbol{c}||_{k+1}^2)$ die Aussage der Gleichung (4.9) erhalten bleibt und $(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{X}})$ möglichst groß ist. Hier helfen die von Abelson und Tukey [130] (balancierter Fall) bzw. Schaafsma und Smid [131] (unbalancierter Fall) bewiesenen Aussagen über den Maxmin-Kontrast. Es sei dazu $\widetilde{\varphi}_l$ der durch die Vektoren $\widetilde{\boldsymbol{c}}$ und $\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}$ aufgespannte Winkel; dann folgt

$$\cos(\widetilde{\varphi}_l) = \frac{(\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}})}{||\widetilde{\boldsymbol{c}}||_N\,||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N} = \frac{(\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}})}{||\widetilde{\boldsymbol{c}}||_N\,||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N} = \frac{(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}})}{||\widetilde{\boldsymbol{c}}||_N\,||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}_N||} = \frac{||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N^2}{||\widetilde{\boldsymbol{c}}||_N\,||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N}\,.$$

Bildet der Vektor $\widetilde{\boldsymbol{c}}$ also mit allen Vektoren $\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}$ $(l = 1,\ldots,2^k-1)$ denselben Winkel, so ist der Kosinus und somit die Norm konstant: $||\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}||_N = \cos(\varphi_l)||\widetilde{\boldsymbol{c}}||_N$.

Der Maxmin-Kontrast erfüllt diese Bedingung und bildet im balancierten Fall mit den Vektoren $\boldsymbol{C}_l\boldsymbol{c}$ $(l = 0,\ldots,2^k-1)$ z. B. die Winkel 49.37° $(k = 3)$, 53.49° $(k = 4)$ und 56.11° $(k = 5)$. Zusätzlich maximiert der Maxmin-Kontrast die minimale Korrelation mit Elementen

des Alternativraumes $\widetilde{H}_A$ [130; 131]. Für die Kontraste ergeben sich somit folgende Darstellungen:

$$\|\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}\|_N^{-1}\,\widetilde{\boldsymbol{c}}'\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}} = \|\boldsymbol{D}_{\sqrt{n}}\boldsymbol{C}_l\boldsymbol{c}\|_{k+1}^{-1}\,\boldsymbol{c}'\boldsymbol{D}_{\boldsymbol{n}}\boldsymbol{C}_l\overline{\boldsymbol{X}} = \frac{\sum_{i=0}^{k} n_i d_{li}\overline{X}_{i.}}{\sqrt{\sum_{i=0}^{k} n_i d_{li}^2}} \quad \text{falls } n=n_i \quad = \frac{\sum_{i=0}^{k} d_{li}\overline{X}_{i.}}{\sqrt{\sum_{i=0}^{k} d_{li}^2 n_i^{-1}}}.$$

Dabei sind die Koeffizienten d_{li} gerade gewichtete Mittel der Koeffizienten des Maxmin-Kontrastes:

$$d_{li} = \frac{n_{g_{li}}c_{g_{li}} + \cdots + n_{g_{l(i+1)}-1}c_{g_{l(i+1)}-1}}{n_{g_{li}} + \cdots + n_{g_{l(i+1)}-1}} \qquad i = 0,\ldots,k\,.$$

Die Kontraste können also recht leicht berechnet werden. Wie bei der Bestimmung der isotonen Schätzer können hier wieder die Zerlegungen einer natürlichen Zahl genutzt werden. Zu beachten ist, daß im unbalancierten Fall der Maxmin-Test von Schaafsma und Smid[131] genutzt werden muß:

$$c_i = \frac{1}{n_i\sqrt{N}}\left(\underbrace{\left(\sum_{s=0}^{i-1} n_s\right)^{\frac{1}{2}}\left(N-\sum_{s=0}^{i-1} n_s\right)^{\frac{1}{2}}}_{=0 \text{ für } i=0} - \left(\sum_{s=0}^{i} n_s\right)^{\frac{1}{2}}\left(N-\sum_{s=0}^{i} n_s\right)^{\frac{1}{2}}\right) \quad i = 0,\ldots,k\,.$$

Weitere Adjustierungen bzgl. ungleicher Stichprobenumfänge sind dann jedoch nicht notwendig. Im weiteren wird der auf den Vektoren $\boldsymbol{d}_l, l = 1,\ldots,2^k-1$ basierende multiple Kontrast als *isotoner* Kontrast bezeichnet. Liegen vier Gruppen vor ($k+1=4$) und sind die Stichprobenumfänge homogen, so ist die Kontrastmatrix des isotonen Kontrastes von folgender Gestalt:

$$\boldsymbol{C}_{ISO} = \begin{pmatrix} -0{,}866025 & -0{,}133975 & 0{,}1339746 & 0{,}8660254 \\ -0{,}866025 & -0{,}133975 & 0{,}5 & 0{,}5 \\ -0{,}866025 & 0 & 0 & 0{,}8660254 \\ -0{,}866025 & 0{,}2886751 & 0{,}2886751 & 0{,}2886751 \\ -0{,}5 & -0{,}5 & 0{,}1339746 & 0{,}8660254 \\ -0{,}5 & -0{,}5 & 0{,}5 & 0{,}5 \\ -0{,}288675 & -0{,}288675 & -0{,}288675 & 0{,}8660254 \end{pmatrix}.$$

Auf den ersten Blick scheint der Ablehnungsbereich des isotonen Kontrastes größer zu sein als der des Bartholomew-Tests. Es ist jedoch zu beachten, daß falls $M(\overline{\boldsymbol{X}})$ nicht die leere Menge ist, der isotone Kontrast sein Maximum ebenfalls für eine Matrix aus der Menge $M(\overline{\boldsymbol{X}})$ annimmt:

$$\frac{\sum_{i=0}^{k} n_i d_{L_c i} \overline{X}_{i.}}{\sqrt{\sum_{i=0}^{k} n_i d_{L_c i}^2}} = \max_{1 \le l \le 2^k - 1} \frac{\sum_{i=0}^{k} n_i d_{li} \overline{X}_{i.}}{\sqrt{\sum_{i=0}^{k} n_i d_{li}^2}} = \max_{\boldsymbol{C}_l \in M(\overline{\boldsymbol{X}})} \frac{\sum_{i=0}^{k} n_i d_{li} \overline{X}_{i.}}{\sqrt{\sum_{i=0}^{k} n_i d_{li}^2}} .$$

Der Beweis dieser Aussage basiert auf der Ordnung der Komponenten des Maxmin-Kontrastes, $(\boldsymbol{C}_l \boldsymbol{c}, \overline{\boldsymbol{X}}) = (\boldsymbol{c}, \boldsymbol{C}_l \overline{\boldsymbol{X}})$, und kann wie folgt indirekt bewiesen werden. Angenommen, es existiert eine Matrix $\boldsymbol{C}_l$ mit $\boldsymbol{Z}_l = \boldsymbol{C}_l \overline{\boldsymbol{X}} \notin \boldsymbol{H}_A$, dann gibt es mindestens eine Komponente, für die $Z_{li} > Z_{li+1}$ gilt. Es sei ohne Beschränkung der Allgemeinheit $i = 0$, $g_{l0} = 1$, $g_{l1} = 2$, d. h. $Z_{l0} = \overline{X}_{0.} > \overline{X}_{1.} = Z_{l1}$. Wird $\overline{X}_{0.}$ und $\overline{X}_{1.}$ durch $\frac{n_0 \overline{X}_0 + n_1 \overline{X}_1}{n_0 + n_1}$ (es existiert mindestens eine Matrix $\boldsymbol{C}_l$, die dies realisiert) ersetzt, so gilt

$$n_0 c_0 \overline{X}_0 + n_1 c_1 \overline{X}_1 \le n_0 c_0 \frac{n_0 \overline{X}_0 + n_1 \overline{X}_1}{n_0 + n_1} + n_1 c_1 \frac{n_0 \overline{X}_0 + n_1 \overline{X}_1}{n_0 + n_1} . \qquad (4.10.$$

Wenn dies nicht so wäre, so würde folgendes gelten:

$$
\begin{aligned}
&0 < n_0 c_0 \overline{X}_{0.} + n_1 c_1 \overline{X}_{1.} - n_0 c_0 \tfrac{n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}}{n_0 + n_1} - n_1 c_1 \tfrac{n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}}{n_0 + n_1} \\
\Leftrightarrow\ &0 < (n_0 + n_1)(n_0 c_0 \overline{X}_{0.} + n_1 c_1 \overline{X}_{1.}) - n_0 c_0 (n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}) - n_1 c_1 (n_0 \overline{X}_{0.} + n_1 \overline{X}_{1.}) \\
&\quad = n_1 n_0 c_0 \overline{X}_{0.} + n_1 n_0 c_1 \overline{X}_{1.} - n_1 n_0 c_0 \overline{X}_{1.} - n_1 n_0 c_1 \overline{X}_{0.} \\
\Leftrightarrow\ &0 < \overline{X}_{1.}(c_1 - c_0) - \overline{X}_{0.}(c_1 - c_0) = \underbrace{(\overline{X}_{1.} - \overline{X}_{0.})}_{<0} \underbrace{(c_1 - c_0)}_{\geq 0} \leq 0.
\end{aligned}
$$

Aus diesem Widerspruch folgt, daß (4.10) stets gilt. Existieren noch weitere Inversionen, kann sukzessiv so fortgefahren werden. Somit besteht demnach eine Analogie zwischen dem Bartholomew-Test und dem isotonen Kontrast. Aufgrund der unterschiedlichen Varianzschätzer (Nenner) ergeben sich jedoch unterschiedliche Verteilungsaussagen. Der Bartholomew-Test nutzt den „besseren“ Varianzschätzer (Nenner), die Verteilung des isotonen Kontrastes ist hingegen sowohl unter der Nullhypothese als auch unter der Alternativhypothese bekannt. Beide Tests haben somit gewisse Vorteile.
Im folgenden Satz werden Verteilungsaussagen für die Kontrasttests zusammengefaßt.

Satz 4.4 : *Sind die Daten voneinander unabhängig und in den Gruppen identisch normalverteilt* $X_{ij} \sim N(\mu_i, \sigma^2)$, *so gilt:*

1. $T_l = \dfrac{\sum_{i=0}^{k} c_{li} \overline{X}_{i.}}{S_I \sqrt{\sum_{i=0}^{k} \frac{c_{li}^2}{n_i}}}$ *ist unter der Nullhypothese zentral und unter der Alternative nichtzentral t-verteilt mit N-k-1 Freiheitsgraden. Der Nichtzentralitätsparameter* δ *ist durch*

$$\delta(\boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{n}, \sigma^2) = \frac{\sum_{i=0}^{k} c_i \mu_i}{\sigma \sqrt{\sum_{i=0}^{k} c_i^2 / n_i}}$$

gegeben.

2. $\boldsymbol{T} = (T_1, \ldots, T_m)'$ *mit* $T_l = \dfrac{\sum_{i=0}^{k} c_{li} \overline{X}_{i.}}{S_I \sqrt{\sum_{i=0}^{k} \frac{c_{li}^2}{n_i}}}$ *l=1,...,m ist unter der Nullhypothese zentral und unter der Alternative nichtzentral m-dimensional multivariat t-verteilt mit N-k-1 Freiheitsgraden. Die Korrelationsstruktur ist durch (4.7) bestimmt. Der Nichtzentralitätsparameter beträgt*

$$\delta(\boldsymbol{C},\boldsymbol{\mu},\boldsymbol{n},\sigma^2)=\left(\frac{\sum_{i=0}^{k} c_{li}\,\mu_i}{\sigma\sqrt{\sum_{i=0}^{k} c_{li}^2/n_i}}\right)_{1\le l\le m}.$$

3. *Die parametrische Verteilung von* $T^{\max}=\max(T_1,\ldots,T_m)$ *ist aufgrund von Punkt 2 bekannt, so daß exakte parametrische* α*-Tests mittels* $T^{\max}$ *definiert werden können.*

Beweis: Punkt 1 ist eine einfache Verallgemeinerung der bekannten Aussagen für den t-Test. Die Punkte 2 und 3 stellen einen Schwerpunkt der Arbeit von Bretz [115] dar. Formeln und Algorithmen zur Berechnung der Verteilung können auch Genz und Bretz [138] entnommen werden. Die Herleitung eines beliebigen multiplen Kontrastes, basierend auf einer multivariaten t-Verteilung bei Verwendung des von Cohen und Sackrowitz [132; 133] empfohlenen Varianzschätzers S_{II}^2, scheitert an unterschiedlichen Freiheitsgraden des Varianzschätzers. Zum Beispiel kann für jede Matrix $\boldsymbol{C}_l \in \mathrm{M}$ ein einseitiger α-Test hergeleitet werden, der auf einem 2α-Test und einer t-Verteilung mit $N-m_l$ Freiheitsgraden beruht. Der Beweis beruht auf folgende Aussagen:

$$\begin{aligned}
P(T>t) &= P\left(\frac{\left|\sum_{i=0}^{k} n_i d_{li}\overline{X}_{i.}\right|}{S_{II}\,\|\boldsymbol{D}_{\sqrt{\boldsymbol{n}}}\boldsymbol{C}_l\boldsymbol{c}\|_{k+1}}>t\right) = P\left(\frac{\sqrt{N-1}\,|((\boldsymbol{c},\boldsymbol{C}\widetilde{\boldsymbol{X}})-\overline{X}_{..}\boldsymbol{1}_N)|}{\|\widetilde{\boldsymbol{X}}-\overline{X}_{..}\boldsymbol{1}_N\|_N\,\|\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}\|_N}>t\right)\\
&= P\left(\frac{\sqrt{N-m_l}\,|((\boldsymbol{c},\boldsymbol{C}\widetilde{\boldsymbol{X}})-\overline{X}_{..}\boldsymbol{1}_N)|}{\|\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}\|_N}>\frac{t\sqrt{(N-m_l)}}{\sqrt{N-1}\sqrt{1-t^2}}\right)\\
&= P\left(|Z|>\frac{t\sqrt{(N-m_l)}}{\sqrt{N-1}\sqrt{1-t^2}}\right).
\end{aligned}$$

Dabei ist Z eine t-verteilte Zufallsvariable mit $N-m_l$ Freiheitsgraden (statt der Transformation $\overline{E}_{01}\Rightarrow\frac{\overline{E}_{01}}{1-\overline{E}_{01}}$ beim Bartholomew-Test wird die Transformation $T\Rightarrow\frac{|\frac{T}{\sqrt{N-1}}|}{\sqrt{1-\frac{T^2}{N-1}}}$ genutzt).

Für weitere Beispiele und zur ausführlichen Diskussion von multiplen Kontrasten unter Normalverteilungsannahmen bzw. für dichotome Daten sei auf Bretz [115] verwiesen. Asymptotisch sind die Kontraststatistiken normal bzw. multivariat normalverteilt mit der

Korrelationsstruktur (4.7) (für beide Varianzschätzer). Dies bleibt auch erhalten, wenn statt der normalverteilten Daten geordnete kategoriale Daten (Scores) eingesetzt werden. Ein direkter Beweis für a priori festgelegte Scores befindet sich im Abschnitt 4.5. Hirotsu et al. [137] zeigen, daß dies auch gilt, wenn die Scores von den Daten abhängen.

Das Verhalten der parametrischen Tests bei Anwendung auf geordnete kategoriale Daten unter der Restriktion kleiner Fallzahlen ist in der Literatur kaum beschrieben. Heeren und D'Agostino [49] beschreiben das Verhalten des t-Tests für Fallzahlen größer 5. Leider haben sie weder Aussagen zur Power gemacht, noch haben sie ihn mit anderen Tests unter den gegebenen Situationen verglichen. Auf multiplen Kontrasten basierende exakte Permutationstests können mittels des Netzwerkalgorithmus programmiert werden. Dazu werden alle Kontrastkoeffizienten mittels Addition einer Konstanten in nichtnegative Werte transformiert. Für das Abbruchkriterium werden zwei Hilfsvektoren eingeführt:

Hilfsvektor 1: In der Komponente i steht das Maximum aller i-ten Komponenten der Kontrastvektoren.

Hilfsvektor 2: In der Komponente i steht das Minimum aller i-ten Komponenten der Kontrastvektoren.

Nach der Erzeugung einer Spalte wird aufgrund der Hilfsvektoren entschieden, ob die restlichen Spalten erzeugt werden müssen oder nicht. Schon für drei Kontraste ist dies Verfahren sehr zeitaufwendig. Approximative Permutationstests sind daher für große Fallzahlen die bessere Alternative. Das "Bootstrappen" von Kontraststatistiken ist eines der wenigen Beispiele, das in der Literatur für den Mehrstichprobenfall beschrieben wird. Ein Satz über die Konsistenz der Bootstrapverteilung ist bei Shao [100, S.95] zu finden. Bemerkung 3.3 sichert für die linearen Funktionen gerade die Konsistenz im diskreten Fall.

Verallgemeinerungen von Kontrasttests auf den Fall heterogener Varianzen beschreiben z. B. Grimes und Federer [117]. Statt eines gepoolten Schätzers werden dabei die Gruppenvarianzen jeweils einzeln geschätzt:

$$T_w = \frac{1}{S_{III}} \sum_{i=1}^{k} c_i \, \overline{X}_{i.} \quad \text{mit} \quad S_{III}^2 = \sum_{i=0}^{k} \frac{c_i^2 S_i^2}{n_i} \text{ und } S_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2 .$$

Die Verteilung dieser Statistiken ist allerdings nicht trivial herleitbar. Grimes und Federer geben vier Approximationen an. Zwei von ihnen seien hier vorgestellt.

a) Welch-Approximation: Statt einer zentralen t-Verteilung mit *n-k*-1 Freiheitsgraden wird eine t-Verteilung mit Freiheitsgrad df_w genutzt:

$$df_w = \frac{(\sum_{i=1}^{k} c_i^2 \, S_i^2 / n_i)^2}{\sum_{i=1}^{k} c_i^4 \, S_i^4 / (n_i^2 \, (n_i - 1))}.$$

b) Cochran-Approximation: Statt des Quantils einer zentralen t-Verteilung wird ein Quantil ($q_{1-\alpha}$), das auf den geschätzten Varianzen, den Kontrastkoeffizienten und den Quantilen der t-Verteilungen mit n_i-1 Freiheitsgraden ($t_{n_i-1,1-\alpha}$) basiert, für die Testentscheidung genutzt:

$$q_{1-\alpha} = \frac{\sum_{i=1}^{k} c_i^2 \, S_i^2 \, t_{n_i-1,1-\alpha} / n_i}{\sum_{i=1}^{k} c_i^2 \, S_i^2 / n_i}.$$

Für den Spezialfall k=1 führt die Approximation a) zum bekannten 2-Stichproben-Welch-Test:

$$T_w = \frac{\overline{X}_{1.} - \overline{X}_{0.}}{\sqrt{s_0^2 / n_0 + s_1^2 / n_1}}.$$

Grimes und Federer leiten geschlossene Formeln für die Powerfunktion der Tests her. Die Lösung der angegebenen Integralgleichungen ist aber nur mit numerisch großem Aufwand möglich.

Hochberg und Tamhane [139, S.181ff] diskutieren weitere Verfahren. In der vorliegenden Arbeit werden nur die vorgestellten Verfahren untersucht, da sie zum einen sehr leicht berechenbar sind und zum anderen den Welch-Test (dem wohl bekanntesten Zweistichprobentests für normalverteilte Daten bei heterogenen Varianzen) als Spezialfall einschließen. Bootstraptests sind für diese Statistiken ebenfalls konstruierbar. Dabei haben die Bootstraptests den Vorteil, daß sie auch auf einen Unterschied der Erwartungswerte testen können, ohne anzunehmen, daß die gesamte Verteilung gleich ist. Die Resamplingstichproben sollten dann aber jeweils in den einzelnen Stichproben gezogen werden. Für kleine Fallzahlen und geordnete kategoriale Daten erscheinen diese Statistiken wesentlich robuster als z. B. die von Roth [116] vorgeschlagenen Trendtests, da an keiner Stelle durch die einzelnen Varianzschätzer geteilt werden muß.

4.3 Likelihood-Quotienten-Test für Multinomialverteilungen

Im folgenden wird ein Likelihood-Quotienten-Test vorgestellt, der für den Vergleich von $(k+1)$ finiten Verteilungsfunktionen (hier auf r Punkte konzentriert) konstruiert wurde. Das Testproblem wird durch die Verteilungsfunktionen beschrieben:

$$H_0: F_0 = F_1 = \cdots = F_k \text{ vs. } H_A: F_0 \leq F \leq_1 \cdots \leq F_k,\ F_0 < F_k.$$

Werden die $(k+1)$ Stichproben $X_{i1},...,X_{in_i}$ $(X_{ij} \in \{K_1,...,K_r\},\ i=0,...,k,\ j=1,...,n_i)$ auf die jeweiligen Häufigkeitsvektoren $\boldsymbol{Y}_i = (Y_{i1},...,Y_{ir})$ $(Y_{is} = \sum_{j=1}^{n_i} I(X_{ij} = K_s))$ reduziert (d. h., die Stichproben werden in eine Kontingenztafel überführt), so hängt die Verteilung eines jeden Vektors $\boldsymbol{Y}_i$ $(i=0,...,k)$ nur von den Parametern n_i und den Wahrscheinlichkeiten $\pi_{is} = P(X_{ij} = K_s)(s=1,..,r)$ ab. Die Wahrscheinlichkeitsdichte ist dann durch

$$P\left(Y_{i1} = y_{i1},...,Y_{ir} = y_{ir} | n_i, \boldsymbol{\pi}_i\right) = \binom{n_i}{y_{i1}\, y_{i2}\, y_{i3} \ldots y_{ir}} \pi_{i1}^{y_{i1}}\, \pi_{i2}^{y_{i2}} \ldots \pi_{ir}^{y_{ir}}$$

gegeben. Basierend auf der gemeinsamen Dichte

$$L(\boldsymbol{y}_0,...,\boldsymbol{y}_k;\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k) = \prod_{i=0}^{k} \binom{n_i}{y_{i1}\, y_{i2}\, y_{i3} \ldots y_{ir}} \pi_{i1}^{y_{i1}}\, \pi_{i2}^{y_{i2}} \ldots \pi_{ir}^{y_{ir}}$$

kann folgender Likelihood-Quotient definiert werden:

$$\begin{aligned}\Lambda &= \frac{\max\limits_{\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k \in H_0} L(\boldsymbol{y}_0,...,\boldsymbol{y}_k;\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k)}{\max\limits_{\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k \in H_A} L(\boldsymbol{y}_0,...,\boldsymbol{y}_k;\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k)} \\ &= \frac{\max\limits_{\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k \in H_0} \prod_{i=0}^{k} \binom{n_i}{y_{i1} \ldots y_{ir}} \prod_{i=0}^{k} \pi_{i1}^{y_{i1}} \ldots \pi_{ir}^{y_{ir}}}{\max\limits_{\boldsymbol{\pi}_0,...,\boldsymbol{\pi}_k \in H_A} \prod_{i=0}^{k} \binom{n_i}{y_{i1} \ldots y_{ir}} \prod_{i=0}^{k} \pi_{i1}^{y_{i1}} \ldots \pi_{ir}^{y_{ir}}}.\end{aligned}$$

Nach dem Einsetzen von Schätzwerten für π_{is} $(i=0,...,k,\ s=1,...,r)$ sowie nach dem Logarithmieren von Λ führt dies zu

$$-2\ \ln\Lambda = -2\sum_{i=0}^{k}\sum_{s=1}^{r} y_{is}\ln(\hat{\pi}_{is}^{(H_A)}) + 2\sum_{i=0}^{k}\sum_{s=1}^{r} y_{is}\ln(\hat{\pi}_{is}^{(H_0)})$$

(ein großer Wert spricht gegen die Nullhypothese). Im Gegensatz zur Bestimmung von $\hat{\pi}_{is}^{(H_0)} = y_{.s} / N$ ist die Bestimmung von $\hat{\pi}_{is}^{(H_A)}$ sehr aufwendig. Im Fall $k = 1$ (also zwei Stichproben) beschreiben Robertson und Wright [32] die Bestimmung der Schätzer und die asymptotische Verteilung der Statistik. Das Schätzverfahren beruht wieder auf isotonen Schätzern, und die Verteilung konvergiert gegen die Verteilung von $\overline{\chi}_0^2$ (siehe Abschnitt 4.1), allerdings ist das Verfahren sehr anfällig. Schon wenn eine Zelle (es existiert ein Paar (i, j) mit $y_{ij} = 0$) nicht besetzt ist, können die Schätzer nicht bestimmt werden (bzw. die Nullen müssen durch willkürliche kleine Konstanten ersetzt werden). Allgemein kann das Schätzen mittels Optimierungssoftware durchgeführt werden. Das Maximierungsproblem mit den $(r\text{-}1)k$ nichtlinearen Ungleichungen der Alternativhypothese, den $(k+1)$ Nebenbedingungen $1 = \sum_{j=1}^{r} \hat{\pi}_{ij}^{(H_A)}$ und den $(k+1)r$ Nebenbedingungen $\hat{\pi}_{is}^{(H_A)} \geq 0$ kann z. B. mit C-Routinen des Programmpakets FSQP [140] gelöst werden.

Agresti und Coull [34] schlagen einen approximativen Permutationstest auf der Basis obiger Statistik vor. Beim Versuch, diese Tests in Simulationsexperimente einzubinden, zeigte sich für kleine Fallzahlen folgendes:

1. Es wurden nicht immer Lösungen für das Optimierungsproblem gefunden.
2. Das Maximieren, welches für jede Permutation der Daten durchgeführt werden muß, nimmt viel Zeit in Anspruch. Agresti und Coull [34] schreiben dazu: *„Power comparison of tests for r x c tables are difficult to conduct in any level of generality, and they are highly computationally intensiv for the order-restricted tests.“*

Auf Simulationsexperimente mit diesen Tests wurde daher verzichtet. Das Verfahren ist aufgrund der Instabilitäten für die Routine bei sehr kleinen Fallzahlen nur bedingt geeignet und wird auch aufgrund fehlender Güteabschätzungen nicht empfohlen. Weitere Schwierigkeiten und Lösungsansätze beschreibt Wang [141].

4.4 GSK-Methode

Grizzle et al. [25] führten folgendes lineare Modell für kategoriale Daten ein (dargestellt durch die Vektoren der absoluten Häufigkeiten):

$$E\,G(\hat{\boldsymbol{\pi}}) = G(\boldsymbol{\pi}) = \boldsymbol{\Delta\beta}\,.$$

Dabei sind $\boldsymbol{\pi} = (\boldsymbol{\pi}_0',...,\boldsymbol{\pi}_k')'$ und $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\pi}}_0',...,\hat{\boldsymbol{\pi}}_k')'$ wie in Satz 3.2 definiert, und G ist eine Funktion, die Elemente des $\mathbb{R}^{(k+1)r}$ in den $\mathbb{R}^u$ $(u \le (k+1)(r-1))$ abbildet. Wie im allgemeinen linearen Modell stellt $\boldsymbol{\Delta}$ eine $u \times v$-Designmatrix und $\boldsymbol{\beta}$ einen Parametervektor der Dimension v dar. Zu Ehren von Grizzle, Starmer und Koch wird dieses Verfahren oft als GSK-Methode bezeichnet. Da der Parametervektor $\boldsymbol{\beta}$ mittels der gewichteten kleinsten Quadrate-Methode (Weighted Least Squares (WLS)) geschätzt werden kann, ist auch die Bezeichnung WLS-Methode gebräuchlich. Gilt $u = k+1,\ t = k+1$ und ist G durch die $(k+1) \times r(k+1)$-Matrix $\boldsymbol{A} = \bigoplus_{i=0}^{k} \boldsymbol{a}',\ \boldsymbol{a} = (a_1,...,a_r)'$ definiert, so gilt z. B. für $k+1 = 5$:

$$E\,\boldsymbol{A}\hat{\boldsymbol{\pi}} = \boldsymbol{A\pi} = \begin{pmatrix} \frac{1}{n_0}\sum_{s=1}^{r} a_s \pi_{0s} \\ \vdots \\ \frac{1}{n_4}\sum_{s=1}^{r} a_s \pi_{4s} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \boldsymbol{\Delta\beta} = \begin{pmatrix} \alpha - \sum_{i=1}^{4} \beta_i \\ \alpha + \beta_1 \\ \alpha + \beta_2 \\ \alpha + \beta_3 \\ \alpha + \beta_4 \end{pmatrix} = \begin{pmatrix} \alpha + \beta_0 \\ \alpha + \beta_1 \\ \alpha + \beta_2 \\ \alpha + \beta_3 \\ \alpha + \beta_4 \end{pmatrix}.$$

Der WLS-Schätzer $\hat{\boldsymbol{\beta}}$ ist dann der Vektor, der $(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta\beta})'(\boldsymbol{A}\frac{1}{N}\hat{\boldsymbol{V}}\,\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta\beta})$ minimiert ($\hat{\boldsymbol{V}}$ ist im Satz 3.3 definiert). Mit $\boldsymbol{\Sigma} = (\boldsymbol{A}\frac{1}{N}\hat{\boldsymbol{V}}\boldsymbol{A}')$ folgt für $\hat{\boldsymbol{\beta}}$:

$$\min_{\boldsymbol{\beta}}(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta\beta})'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta\beta}) = \min_{\boldsymbol{\beta}}||\,\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta\beta})||^2_{k+1} = ||\,\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}\hat{\boldsymbol{\pi}} - \boldsymbol{\Delta}\hat{\boldsymbol{\beta}})||^2_{k+1}\ .$$

Es können demnach wie im Normalverteilungsmodell Mittelwertvektoren ($\boldsymbol{A}\hat{\boldsymbol{\pi}}$) modelliert werden. Die Schätzer werden jedoch unabhängig von den Hypothesen bestimmt. Im Gegensatz zum Bartholomew-Test wird hier das Minimum über alle möglichen Vektoren gesucht, und wie bei den Tests von Roth geht hier keine Annahme bezüglich homogener

Varianzen ein. Eine geschlossene Form kann sowohl für den WLS-Schätzer $\hat{\boldsymbol{\beta}}$ als auch für dessen Kovarianzmatrix $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ hergeleitet werden [16, S.230]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Delta}'(\boldsymbol{A}\tfrac{1}{N}\hat{\boldsymbol{V}}\boldsymbol{A})^{-1}\boldsymbol{\Delta}'(\boldsymbol{A}\tfrac{1}{N}\hat{\boldsymbol{V}}\,\boldsymbol{A}')^{-1}\boldsymbol{A}\hat{\boldsymbol{\pi}} \qquad \boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{\Delta}'(\boldsymbol{A}\tfrac{1}{N}\hat{\boldsymbol{V}}\,\boldsymbol{A}')^{-1}\boldsymbol{\Delta})^{-1}.$$

Die Hypothesen werden analog zum Normalverteilungsmodell mittels der Parameter formuliert, z. B. $H_0: \beta_0 = \cdots = \beta_k$ vs. $H_A: \beta_0 \leq \cdots \leq \beta_k, \beta_0 < \beta_k$.

Werden andere Funktionen für *G* gewählt, so ist bekannt, daß die asymptotische Verteilung des Parametervektors (eine multivariate Normalverteilung) für sehr kleine Fallzahlen zu liberalen Tests führt [142]. Desweiteren ist dieses Verfahren anfällig gegen Zellen, die mit dem Wert Null besetzt sind ($\hat{\pi}_{is} = 0$). Einige Programme bieten daher die Möglichkeit, vor der Auswertung zu jeder Zelle eine Konstante hinzuzuaddieren. Die Addition einer Konstanten, z. B. $1/r$ [143], läßt jedoch einen nicht akzeptablen Einfluß, der schwer zu spezifizieren ist, auf das Testergebnis und auf die Schätzer zu. Auch aufgrund der bekannten schlechten Asymptotik ist dieses Verfahren bei kleinen Fallzahlen nicht empfehlenswert. Werden statt der WLS-Schätzer Maximum-Likelihood-Schätzer genutzt, ist die Addition einer Konstanten nicht notwendig, da ML-Schätzer nur von den Randsummen abhängen. Auch in Hinblick auf die etwas bessere Asymptotik schreibt Agresti [37, S.462]: *„For large cell expected frequencies, ML and WLS give similar results. Both estimators are in the class of best asymptotically normal estimators. For small samples, practical considerations often give an edge to ML estimation.“* Wenn die Stichproben zu klein werden, treten allerdings verstärkt Konvergenzprobleme bei der Bestimmung der ML-Schätzer auf (die iterativen Verfahren, z. B. Newton-Raphson-Verfahren, konvergieren nur langsam oder divergieren sogar). Bei sehr kleinen Fallzahlen können somit auch diese Verfahren nicht empfohlen werden.

4.5 Scorestatistiken

Werden den Kategorien und den Gruppen (Dosen) jeweils monoton wachsende Scores zugeordnet, können technisch einfachere Tests, die mit schwächeren Modellvoraussetzungen auskommen, konstruiert werden. Durch einfache Umrechnungen kann ein Zusammenhang zu den in Kapitel 4 vorgestellten Kontraststatistiken gezeigt werden. Zunächst soll aber dargestellt werden, warum auch allgemeine Scorestatistiken genutzt werden können. Dazu seien die Daten in Form einer Kontingenztafel gegeben (z. B. $k+1=4,\ r=4$).

	K_1 mit $Score(K_1) = a_1$	K_2 mit $Score(K_2) = a_2$	K_3 mit $Score(K_3) = a_3$	K_4 mit $Score(K_4) = a_4$
D_0 mit $Score(D_0) = c_0$	$y_{01} = 2$	$y_{02} = 1$	$y_{03} = 0$	$y_{04} = 0$
D_1 mit $Score(D_1) = c_1$	$y_{11} = 0$	$y_{12} = 3$	$y_{13} = 0$	$y_{14} = 0$
D_2 mit $Score(D_2) = c_2$	$y_{21} = 0$	$y_{22} = 0$	$y_{23} = 3$	$y_{24} = 0$
D_3 mit $Score(D_3) = c_3$	$y_{31} = 0$	$y_{32} = 0$	$y_{33} = 1$	$y_{34} = 2$

Tabelle 4.5: Beispiel einer doppelt geordeneten Kontingenztafel mit zugeordneten Scores

Liegt der Trend „mit ansteigender Dosis höhere Wirkung“ vor, dann sollten die schattierten Zellen in Tabelle 4.5 am stärksten besetzt sein, d. h., zwischen den Spaltenscores und den Zeilenscores besteht eine positive Korrelation. Patefield[33] schlägt z. B. die Statistik

$$T = \frac{1}{N} \sum_{i=0}^{k} \sum_{s=1}^{r} y_{is} c_i a_s \tag{4.11}$$

vor, wobei die Scores sowohl für die Spalten als auch für die Zeilen a priori vorgegeben werden. Als Alternative zur a-priori-Vergabe betrachtet er das Supremum von T

$$T_{\sup} = \sup_{\boldsymbol{S}} \frac{1}{N} \sum_{i=0}^{k} \sum_{s=1}^{r} y_{is} c_i a_s \tag{4.12}$$

über die Menge $\boldsymbol{S}$:

$$\boldsymbol{S} = \{\boldsymbol{a}, \boldsymbol{c}: a_1 \leq \cdots \leq a_r, c_0 \leq \cdots \leq c_k, \sum_i y_{i.} c_i = 0, \sum_s y_{.s} a_s = 0, \sum_i y_{i.} c_i^2 = N, \sum_s y_{.s} a_s^2 = N\} .$$

Patefield konstruiert mittels dieser Teststatistiken exakte bzw. approximative Permutationstests. Als a-priori-Scores benutzt er äquidistante Scores. Obwohl die Nutzung von (4.12) mit

hohem technischen Aufwand verbunden ist, empfiehlt Patefield diese Statistik. Die a priori vorgegebenen Scores lehnt er mit den gleichen Argumenten ab, wie in Abschnitt 4.2 die einfachen Kontrasttests abgelehnt wurden. Der Zusammenhang zu den einfachen Kontrasten wird durch folgende Formeln deutlich:

a) für die Statistik

$$T = \tfrac{1}{N}\sum_{i=0}^{k}\sum_{s=1}^{r} Y_{is}c_i a_s = \sum_{i=0}^{k} \underbrace{\tfrac{1}{N}c_i n_i}_{b_i} \tfrac{1}{n_i}\sum_{s=1}^{r} Y_{is}a_s$$

$$= \sum_{i=0}^{k} b_i \tfrac{1}{n_i}\sum_{s=1}^{r} Y_{is}a_s$$

b) für den Erwartungswert $\quad E\,T = \sum_{i=0}^{k} b_i \tfrac{1}{n_i}\sum_{s=1}^{r} EY_{is}a_s = \sum_{i=0}^{k} b_i \sum_{s=1}^{r} \pi_{is}a_s$

c) für die Varianz

$$V(T) = \sum_{i=0}^{k} b_i^2 V\left(\tfrac{1}{n_i}\sum_{s=1}^{r} Y_{is}a_s\right) = \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\boldsymbol{a}'(\boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i\boldsymbol{\pi}_i')\boldsymbol{a}$$

$$= \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\sum_{s=1}^{r} a_s^2\pi_{is} - \sum_{s=1}^{r}\sum_{t=1}^{r} a_s a_t \pi_{is}\pi_{it}\right)$$

$$= \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\sum_{s=1}^{r} a_s^2\pi_{is} - \left(\sum_{s=1}^{r} a_s\pi_{is}\right)^2\right)$$

$$= \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\sum_{s=1}^{r}\left(a_s - \sum_{s=1}^{r} a_s\pi_{is}\right)^2 \pi_{is}\right).$$

Als Schätzer für die Varianz kann

$$\hat{V}(T) = \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\sum_{s=1}^{r}\left(a_s - \sum_{s=1}^{r} a_s\hat{\pi}_{is}\right)^2 \hat{\pi}_{is}\right)$$

$$= \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\sum_{s=1}^{r}\left(a_s - \sum_{s=1}^{r} a_s \tfrac{Y_{is}}{n_i}\right)^2 \tfrac{Y_{is}}{n_i}\right)$$

$$(z_{ij} = a_s \quad \text{für} \quad Y_{i1}+\cdots+Y_{i(s-1)} \le j < Y_{i1}+\cdots+Y_{is} \quad (Y_{i0} = 0))$$

$$= \sum_{i=0}^{k} b_i^2 \tfrac{1}{n_i}\left(\tfrac{1}{n_i}\sum_{j=1}^{n_i}\left(Z_{ij} - \tfrac{1}{n_i}\sum_{s=1}^{n_i} Z_{ij}\right)^2\right) = \sum_{i=0}^{k} \tfrac{b_i^2}{n_i}\tfrac{n_i-1}{n_i}\tfrac{1}{n_i-1}\sum_{i=1}^{n_i}(Z_{ij} - \overline{Z}_{i.})^2$$

$$= \tfrac{1}{N^2}\sum_{i=0}^{k} c_i^2 n_i \tfrac{n_i-1}{n_i} S_i^2 \quad \text{bzw. } \hat{V}_{III}(T) = \tfrac{1}{N^2}\sum_{i=0}^{k} c_i^2 n_i S_i^2$$

benutzt werden. Gilt die Nullhypothese $H_0 : \boldsymbol{\pi}_0 = \cdots = \boldsymbol{\pi}_k$, so können die Zellwahrscheinlichkeiten auch durch $\hat{\pi}_{is} = \frac{Y_s}{N}$ $(s = 1,\ldots,r)$ geschätzt werden. Im Gegensatz zum obigen ungepoolten Varianzschätzer führt dies zu einem gepoolten Varianzschätzer:

$$\hat{V}(T) = \sum_{i=0}^{k} b_i^2 \frac{1}{n_i}\left(\frac{N-1}{N}\frac{1}{N-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \frac{1}{N}\sum_{i=0}^{k}\sum_{j=1}^{n_i} Z_{ij}\right)^2\right) = \sum_{i=0}^{k}\frac{b_i^2}{n_i}\frac{N-1}{N}S_{II}^2$$

$$= \frac{1}{N^2}\sum_{i=0}^{k} c_i^2 n_i \frac{N-1}{N} S_{II}^2 \text{ bzw.} \hat{V}_{II}(T) = \frac{1}{N^2}\sum_{i=0}^{k} c_i^2 n_i S_{II}^2.$$

Die standardisierte Statistik besitzt dann z. B. im Fall $n = n_i,\ i = 0,\ldots,k$ und bei Benutzung von $\hat{V}_{II}(T)$ die Form

$$T_I = \frac{\frac{n}{N}\sum_{i=0}^{k} c_i \frac{1}{n}\sum_{s=1}^{n} Z_{is}}{\sqrt{\frac{1}{N^2}\sum_{i=0}^{k} c_i^2 n\, S_{II}^2}} = \frac{\sum_{i=0}^{k} c_i \overline{Z}_{i.}}{\sqrt{\sum_{i=0}^{k}\frac{c_i^2}{n} S_{II}^2}}.$$

Das heißt, die einfache Vorstellung der Korrelationen der Zeilen- und Spaltenscores führt automatisch zu Kontrasttests, falls $\sum_{i=0}^{k} y_{i.} c_i = \sum_{i=0}^{k} n_i c_i = 0$ gilt. Hirotsu [144] benutzt (4.11) und (4.12) für asymptotische Tests. Als Scores schlägt er u. a. äquidistante Scores vor. Er verweist jedoch darauf, daß die Verteilung von (4.12) schwer bestimmbar ist. Hirotsu überträgt daher das Konzept der multiplen Kontrasttests auch auf kategoriale Daten.

Satz 4.5: *Es seien (k+1) unabhängige multinomialverteilte Stichproben* $\boldsymbol{X}_{01},\ldots,\boldsymbol{X}_{0n_0}$, $\ldots,\boldsymbol{X}_{k1},\ldots,\boldsymbol{X}_{kn_k}$ $(\boldsymbol{X}_{ij} \sim M(1,\boldsymbol{\pi}_i), \boldsymbol{\pi}_i = (\pi_{i1},\ldots,\pi_{ir}))$ *gegeben. Falls*

(A0) $0 < \zeta \le \hat{\pi}_{is} \le 1-\zeta < 1,\ s = 1,\ldots,r,\ i = 0,\ldots,k$

(A1) $\min_{i=0,\ldots,k} n_i \to \infty$

(A2) $0 < \lambda \le n_i / N \le 1-\lambda < 1,\ i = 0,\ldots,c$ *gilt, so folgt für eine beliebige Kontrastmatrix* $\boldsymbol{C}$ *vom Format* $m \times (k+1)$*:*

1. $P_{\boldsymbol{CA}\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})} \Rightarrow N(\boldsymbol{0}, \boldsymbol{CAVA'C'})$ *mit* $\boldsymbol{V} = \bigoplus_{i=0}^{k}\frac{N}{n_i}(\boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'),\ \boldsymbol{A} = \bigoplus_{i=0}^{k}\boldsymbol{a}', \boldsymbol{a} = (a_1,\ldots,a_r)'$
2. $\boldsymbol{C}\boldsymbol{A}\hat{\boldsymbol{V}}\boldsymbol{A}'\boldsymbol{C}' \xrightarrow[\min(n_i)\to\infty]{} \boldsymbol{CAVA'C'}$ *mit Wahrscheinlichkeit* 1*,*
3. $\boldsymbol{C}\boldsymbol{A}\hat{\boldsymbol{V}}_{II}\boldsymbol{A}'\boldsymbol{C}' \xrightarrow[\min(n_i)\to\infty]{} \boldsymbol{CAVA'C'}$ *fast sicher, falls zusätzlich die Nullhypothese gilt.*

Beweis: Die Verteilung folgt aus Satz 3.3 und Beispiel 3.2.

Da für $\boldsymbol{A}\hat{\boldsymbol{V}}\boldsymbol{A}' = (\upsilon_{uw})_{0\le u,w\le k+1}$

$$(\upsilon_{uw})_{0\le u,w\le k+1} = \begin{cases} \frac{N}{n_u} S_u^2 & \text{falls u = w} \\ 0 & \text{falls u} \neq \text{w} \end{cases} \quad \text{mit} \quad S_i^2 = \sum_{s=1}^{r} (a_s - \sum_{t=1}^{r} a_s \hat{\pi}_{it})^2 \hat{\pi}_{it}$$

gilt, folgt für $\boldsymbol{CA}\hat{\boldsymbol{V}}\boldsymbol{A}'\boldsymbol{C}' = (\rho_{uw})_{1\le u,w\le m}$

$$(\rho_{uw})_{1\le \text{u},w\le m} = \sum_{i=0}^{k} c_{ui}\, c_{wi}\, \tfrac{N}{n_i} S_i^2 \,. \tag{4.13}$$

Wird der Schätzer $\hat{V}_{II}$ statt $\hat{V}$ genutzt, so ergibt sich:

$$(\rho_{uw})_{1\le \text{u},w\le m} = S_{II}^2 \sum_{i=0}^{k} c_{ui}\, c_{wi}\, \tfrac{N}{n_i} \,. \tag{4.14}$$

Die in diesem Abschnitt vorgestellten Statistiken stimmen bis auf den Varianzschätzer mit denen aus Abschnitt 4.2 überein. Statt der (multivariaten) t-Verteilung wird nun allerdings eine (multivariate) Normalverteilung benötigt. Zur Berechnung der Quantile können die durch Bretz[115] beschriebenen Algorithmen genutzt werden. Dazu ist es jedoch sinnvoll, die Statistiken zu normieren, d. h., die Kovarianzmatrix in eine Korrelationsmatrix zu überführen. Sind alle Varianzschätzer bzw. der gepoolte Varianzschätzer größer Null, kann dies mit einer Matrixmultiplikation durchgeführt werden. Dann gilt $P_{\hat{\boldsymbol{D}}\boldsymbol{CA}\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})} \Rightarrow N(\boldsymbol{0}, \hat{\boldsymbol{D}}\boldsymbol{CAVA}'\boldsymbol{C}'\hat{\boldsymbol{D}}')$, wobei $\hat{\boldsymbol{D}}$ eine Diagonalmatrix ist, deren Elemente bestimmt sind durch

$$(\tau_{uw})_{1\le \text{u},w\le k+1} = \begin{cases} (\rho_{uw})^{-1/2} & u = w \\ 0 & u \neq w \end{cases}, \ \rho_{uw} \text{ aus (4.13) bzw.(4.14)} \,.$$

Die Nullhypothese H_0 kann daher mit

$$T^{\max} = \max_{1\le u\le m}(T_u) \text{ mit } \boldsymbol{T} = \hat{\boldsymbol{D}}\boldsymbol{CA}\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$$

geprüft werden. Wie in Abschnitt 4.2 können mit Hilfe dieser Statistiken Permutationstests bzw. Bootstraptests konstruiert werden.
Für lokale Alternativen der Form

$$\ddot{\pi}_i = \pi_i + \tfrac{1}{\sqrt{N}}\theta_i \text{ mit } \sum_{s=1}^{r} \theta_{is} = 0 \tag{4.15}$$

gilt der folgende Satz.

Satz 4.6: *Sind* $\boldsymbol{X}_{01},\ldots,\boldsymbol{X}_{0n_0},\ldots,\boldsymbol{X}_{k1},\ldots,\boldsymbol{X}_{kn_k}$ *multinomialverteilte Zufallsvektoren mit* $F_{\boldsymbol{X}_{ij}} = M(1,\ddot{\boldsymbol{\pi}}_i)$, $\ddot{\boldsymbol{\pi}}_i = (\ddot{\pi}_{i1},\ldots,\ddot{\pi}_{ir})'$, *sind die Voraussetzungen von Satz 4.5 erfüllt und gilt für* $i = 0,\ldots,k$ *(4.15), so folgt für* $\boldsymbol{\pi} = (\boldsymbol{\pi}',\ldots,\boldsymbol{\pi}')'$, $\hat{\ddot{\boldsymbol{\pi}}} = (\hat{\ddot{\boldsymbol{\pi}}}'_0,\ldots,\hat{\ddot{\boldsymbol{\pi}}}'_c)'$ *und* $\boldsymbol{\theta} = (\boldsymbol{\theta}'_0,\ldots,\boldsymbol{\theta}'_c)'$

1. $P_{\boldsymbol{CA}\sqrt{N}(\hat{\ddot{\boldsymbol{\pi}}}-\boldsymbol{\pi})} \Rightarrow N(\boldsymbol{CA\theta},\boldsymbol{CAVA'C'})$ *mit* $\boldsymbol{V} = \bigoplus_{i=0}^{k} \frac{N}{n_i}(\boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi\pi'})$, $\boldsymbol{A} = \bigoplus_{i=0}^{k} \boldsymbol{a}'$, $\boldsymbol{a} = (a_1,\ldots,a_r)'$
2. $P_{\hat{\boldsymbol{D}}\boldsymbol{CA}\sqrt{N}(\hat{\ddot{\boldsymbol{\pi}}}-\boldsymbol{\pi})} \Rightarrow N(\hat{\boldsymbol{D}}\boldsymbol{CA\theta},\hat{\boldsymbol{D}}\boldsymbol{CAVA'C'}\hat{\boldsymbol{D}}')$.

Beweis: Mittels der Cramer-Wold-Technik und Satz 3.2 kann der Satz analog zu Satz 3.3 bewiesen werden.
Aufgrund von Satz 4.6 kann die Power für diese multiplen Kontrasttests auf analoge Art, wie es Bretz [115] für dichotome Daten zeigt, bestimmt werden ($N(\boldsymbol{CA\theta},\boldsymbol{CAVA'C'}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$):

$$\begin{aligned}
&P\left(\max_{1\le u\le m}(T_u) \ge z_{m,1-\alpha} \mid H_A\right) = 1 - P\left(\max_{1\le u\le m}(T_u) < z_{m,1-\alpha} \mid H_A\right)\\
&= 1 - P(T_1 < z_{m,1-\alpha} \wedge \cdots \wedge T_m < z_{m,1-\alpha} \mid H_A)\\
&= 1 - P(T_1 - \mu_1 < z_{m,1-\alpha} - \mu_1 \wedge \cdots \wedge T_m - \mu_m < z_{m,1-\alpha} - \mu_m \mid H_A)\\
&= 1 - \Phi(z_{m,1-\alpha} - \mu;\boldsymbol{\theta},\boldsymbol{\Sigma}).
\end{aligned}$$

4.6 Geordnete kategoriale Regressionsmodelle

Geordnete kategoriale Regressionsmodelle können in die große Klasse der verallgemeinerten linearen Modelle eingegliedert werden [30, S.258]. Anhand des kumulativen logistischen Modells sollen hier nur die Grundideen für einfaktorielle Versuche dargestellt werden.

Definition: *Es sei Y_{ij} eine ordinalskalierte Zufallsgröße, mit dem Wertebereich $\{a_s : s = 1,\ldots,r; a_s < a_{s+1}\}$, und x_i sei ein Kovariablenvektor. Gilt für eine bijektive, 2-mal stetig differenzierbare Verteilungsfunktion F:* $\gamma_{is} = P(Y_{ij} \le y_s | \boldsymbol{x}_i'\boldsymbol{\beta}) = F(\theta_s - \boldsymbol{x}_i'\boldsymbol{\beta}), i = 0,\ldots k$, $j = 1,\ldots,n_i$, $s = 1,\ldots,r$ *$(-\infty \equiv \theta_0 < \theta_1 < \cdots < \theta_{r-1} < \theta_r \equiv \infty)$, so wird dieses Modell als geordnet kategoriales Regressionsmodell bezeichnet. Hierbei sind die Vektoren $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_{r-1})'$ und $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_k)'$ unbekannte Parameter.*

Bei den in der vorliegenden Arbeit untersuchten Dosis-Wirkungs-Versuchen sind die Kovariablen $\boldsymbol{x}_i = (0,\ldots,\underset{i-te\,Stelle}{1},\ldots 0)'$ für $i = 1,\ldots,k$ und $\boldsymbol{x}_0 = (0,\ldots,0)'$; d. h., wie im linearen Modell werden die Faktorstufen mittels sogenannter Indikatorvariablen modelliert. Für diese Kovariablenvektoren folgt $(\boldsymbol{x}_i'\boldsymbol{\beta}) \ge (\boldsymbol{x}_j'\boldsymbol{\beta}) \Leftrightarrow \beta_i \ge \beta_j$ und damit

$$P(Y_{is} > y) = P(Y > y | \boldsymbol{x}_i'\beta) \ge P(Y > y | \boldsymbol{x}_j'\beta) = P(Y_{js} > y).$$

Dies bedeutet, daß die Zufallsvariablen bzgl. der Komponenten von $\boldsymbol{\beta}$ stochastisch geordnet sind. Zum Testen „Gleichheit der Verteilungen vs. stochastische Ordnung“ können daher Tests bzgl. der Parameter β_i konstruiert werden. Der unbekannte Interceptparametervektor $\boldsymbol{\theta}$ (Vektor der Schwellenwerte) sowie der Regressionsparametervektor $\boldsymbol{\beta}$ können z. B. mittels der Maximum-Likelihood-Methode bestimmt werden. Das am häufigsten beschriebene Modell ist das kumulative logistische Modell (proportionales Odds-Modell)[20], bei dem *F* die logistische Verteilungsfunktion darstellt:

$$\gamma_{is} = \frac{\exp(\theta_s - \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\theta_s - \boldsymbol{x}_i'\boldsymbol{\beta})}.$$

Die Bezeichnung beruht auf folgendem Zusammenhang der logarithmierten kulminierten Quotenverhältnisse (Odds-Ratio):

$$\begin{aligned}\operatorname{logit}(\gamma_{is}) - \operatorname{logit}(\gamma_{(i+1)s}) &= \ln\left(\frac{\gamma_{is}}{(1-\gamma_{is})}\right) - \ln\left(\frac{\gamma_{i+1s}}{(1-\gamma_{i+1s})}\right) \\ &= \ln\left(\frac{(1-\gamma_{i+1s})\gamma_{is}}{(1-\gamma_{is})\gamma_{i+1s}}\right) = (\boldsymbol{x}_{i+1}' - \boldsymbol{x}_i')\boldsymbol{\beta}.\end{aligned}$$

Die Chance, daß ein Element der Stichprobe i in eine der Kategorien K_1 bis K_s fällt, ist demnach nicht von der Kategorie abhängig und ist um $e^{\beta_{i+1}-\beta_i}$ größer (kleiner) als dieselbe Chance für ein Element der Stichprobe $i+1$:

$$\frac{\gamma_{is}}{(1-\gamma_{is})} = e^{(\boldsymbol{x}'_{i+1}-\boldsymbol{x}'_i)\boldsymbol{\beta}} \frac{\gamma_{i+1s}}{(1-\gamma_{i+1s})} = e^{\beta_{i+1}-\beta_i} \frac{\gamma_{i+1s}}{(1-\gamma_{i+1s})}.$$

Soll der Einfluß der Kategorie stärker in das Modell miteingehen, so kann dies durch komplexere Modelle mit einer noch größeren Anzahl von Parametern realisiert werden [30, S.253]. Allerdings kann dies bei sehr kleinen Fallzahlen schnell zu Situationen führen, bei denen mehr Parameter zu schätzen sind als Beobachtungen vorliegen. Die Existenz der Schätzer ist dabei im allgemeinen nicht gesichert. Der Vorteil gegenüber den bisher betrachteten Methoden liegt darin, daß der Dosis-Wirkungs-Zusammenhang direkt modelliert wird. Demnach sind Maße für die Stärke des Zusammenhangs ableitbar sowie Vorhersagen möglich. Das Testen der Hypothesen verläuft allerdings relativ analog. Statt Mittelwerte zu vergleichen, werden bei diesen Verfahren die geschätzten Parameter verglichen. Die Hypothesen lauten dann:

$$H_0: \beta_1 = \cdots = \beta_k = 0 (= \beta_0) \text{ vs. } H_A: \beta_1 \leq \cdots \leq \beta_k, \beta_k > 0.$$

Wie in den Abschnitten 4.2 und 4.4 können die Hypothesen mittels einfachen, als auch multiplen Kontrasten geprüft werden, da unter der Nullhypothese und bestimmten Regularitätsbedingungen der Regressionsparametervektor multivariat normalverteilt ist [30, S.273].
Oelerich [145] vergleicht nichtparametrische Verfahren mit dem proportionalen Odds-Modell (ML und WLS-Schätzungen) und zeigt, daß die Güte der Verfahren relativ gut übereinstimmt. Die von ihm untersuchten nicht gerichteten asymptotischen Verfahren zeigen, daß eine gute Approximation der wahren Verteilung durch eine asymptotische Verteilung erst ab einem Stichprobenumfang von $n = 7$ je Gruppe gewährleistet ist. Aufgrund der starken Voraussetzungen an die Daten und an das Modell, aufgrund der technischen Probleme und aufgrund der hohen Rechenzeit bei auf diesen Tests basierenden Simulationen (besonders für Resamplingverfahren) wurde auf die weitere Untersuchung dieser Verfahren verzichtet.

5 Nichtparametrische Verfahren

5.1 Nichtparametrische Tests auf der Basis von isotonen Schätzern

Rangverfahren stellen die bekannteste Klasse der nichtparametrischen Verfahren dar. Sie stellen schwächere Voraussetzungen an die Verteilung der Daten und sind in vielen Fällen robuster als parametrische Verfahren [31,S.vi]. Ein großer Vorteil von Rängen liegt in ihrer Invarianz bezüglich streng monotoner Funktionen. Auch wenn für Rangverfahren eine Zuordnung von Scores notwendig ist, sind sie dennoch unabhängig von den absoluten Größen der Scores. Ursprünglich wurden Rangtests für stetig verteilte Daten (fast sicher keine Bindungen zwischen den Daten) konstruiert. Treten doch Bindungen auf (z. B. durch ungenaue Meßgeräte, Rundungen), werden die folgenden modifizierten Ränge betrachtet:

1. Maximale Ränge: Innerhalb der Bindungsgruppen wird jeder Beobachtung der maximale Rang der für die gebundenen Daten gesamt zu vergebenen Ränge zugeordnet.
2. Zufallsränge: Innerhalb der Bindungsgruppen wird jeder Beobachtung zufällig ein Rang der für die gebundenen Daten gesamt zu vergebenen Ränge zugeordnet.
3. Durchschnittsränge: Innerhalb der Bindungsgruppen wird jeder Beobachtung das arithmetische Mittel aus den für die gebundenen Daten gesamt zu vergebenen Ränge zugeordnet.

In der Praxis werden die maximalen Ränge genauso selten wie die Zufallsränge benutzt. Mittels der Zufallsränge können jedoch viele Aussagen, die eigentlich für stetige Daten hergeleitet wurden, auch auf Daten mit Bindungen übertragen werden. Gültige Verteilungsaussagen für Durchschnittsränge wurden u. a. auch mit Hilfe asymptotischer äquivalenter Statistiken, die auf Zufallsrängen basieren, hergeleitet [93, S.171]. Verteilungsaussagen für Statistiken, die auf Durchschnittsrängen beruhen, werden seit Mitte der 90-er Jahre auch direkt für diskrete Daten hergeleitet. Dabei werden funktionalanalytische Methoden zur Definition einer normalisierten Verteilungsfunktion genutzt. Diese normalisierte Verteilungsfunktion steht dann im Mittelpunkt aller weiteren Hypothesen. Bevor die eigentlichen Tests definiert werden, sollen einige bekannte Aussagen für bestimmte Rangvektoren zusammengefaßt dargestellt werden. Als erstes werden einige Größen und Bezeichnungen beschrieben.

Es seien $(R_{01},...,R_{0n_0},...,R_{k1},...,R_{kn_0})$ die zu den Zufallsvariablen $(X_{01},...,X_{0n_i},...,$ $X_{k1},...,X_{kn_i})$ gehörigen Durchschnittsränge (im weiteren seien mit Rängen immer die Durchschnittsränge gemeint). Mit $F_i(x) = F_{ij}(x) = \frac{1}{2}(F_{ij}^{+} + F_{ij}^{-})$ wird die normalisierte Verteilungsfunktion der Zufallsvariablen der *i*-ten Stichprobe bezeichnet. Die Zerlegung in die rechtsstetige Verteilungsfunktion $F_{ij}^{+}(x) = P(X_{ij} \le x)$ und in die linksstetige Verteilungsfunktion $F_{ij}^{-}(x) = P(X_{ij} < x)$ bringt wesentliche, technische Vorteile für die Beweisbarkeit vieler folgender Aussagen. Desweiteren kann ein Zusammenhang zwischen den zugehörigen empirischen Verteilungsfunktionen und den Durchschnittsrängen genutzt werden. Die normalisierte empirische Verteilungsfunktion besitzt folgende analoge Form:

$$\hat{F}_i(x) = \hat{F}_{ij}(x) = \tfrac{1}{2}(\hat{F}_{ij}^{+}(x) + \hat{F}_{ij}^{-}(x)) = \frac{1}{n_i}\sum_{j=1}^{n_i} \tfrac{1}{2}(c^{+}(x) + c^{-}(x))$$

mit

$$c^{+}(x) = \begin{cases} 1 & : x \ge 0 \\ 0 & : x < 0 \end{cases} \text{ und } c^{-}(x) = \begin{cases} 1 & : x > 0 \\ 0 & : x \le 0 \end{cases}.$$

Die entsprechenden Verteilungsfunktionen der gepoolten Stichproben sind analog darstellbar durch

$$H(x) = \tfrac{1}{N}\sum_{i=1}^{k} n_i F_i(x) \text{ bzw. } \hat{H}(x) = \tfrac{1}{N}\sum_{i=1}^{k} n_i \hat{F}_i(x).$$

Im Zweistichprobenfall und bei stetigen Verteilungsfunktionen spielt die Wahrscheinlichkeit $P(X_{21} > X_{11})$ eine bedeutende Rolle, da unter der Nullhypothese $F_{11} = F_{21}$

$$p = P(X_{21} \ge X_{11}) = \int F_1(x)\,dF_2(x) = \tfrac{1}{2}$$

gilt. Sind die Verteilungsfunktionen jedoch nicht stetig, gilt diese Beziehung im allgemeinen weder für die linksstetige noch für die rechtsstetige Verteilungsfunktion. Wird p durch $p = P(X_{21} > X_{11}) + \frac{1}{2}P(X_{21} = X_{11})$ definiert, so gilt für die normalisierten Verteilungsfunktionen

$$p = \int F_1(x)\, dF_2(x),$$

und für $F_1 = F_2$ folgt $p = \frac{1}{2}$. Mittels der Größe p können somit Unterschiede zwischen zwei Verteilungen geordneter kategorialer Größen beschrieben werden. Wird dieses Funktional mittels der empirischen Funktionen geschätzt, führt dies zu

$$\begin{aligned}
\hat{p} &= \hat{P}(X_{21} > X_{11}) = \int \hat{F}_1(x)\, d\hat{F}_2(x) \\
&= \tfrac{1}{n_1}\left(\int (n_1\hat{F}_1(x) + n_2\hat{F}_2(x))\, d\hat{F}_2(x) - n_2 \int \hat{F}_2(x) d\hat{F}_2(x)\right) \\
&= \tfrac{1}{n_1}\left((n_1+n_2)\int \hat{H}(x)\, d\hat{F}_2(x) - n_2 \int \hat{F}_2(x) d\hat{F}_2(x)\right) \\
&= \tfrac{1}{n_1}\left(\tfrac{n_1+n_2}{n_2}\sum_{j=1}^{n_2} \hat{H}(x_{2j}) - \tfrac{n_2}{n_2}\sum_{j=1}^{n_2} \hat{F}_2(x_{2j})\right) \qquad \text{(es gilt } \hat{H}(X_{ij}) = \tfrac{1}{n_1+n_2}(R_{ij} - \tfrac{1}{2})\text{)} \\
&= \tfrac{1}{n_1}\left(\tfrac{1}{n_2}\sum_{j=1}^{n_2} R_{2j} - \tfrac{1}{2} - \tfrac{1}{2}(n_2+1) + \tfrac{1}{2}\right) = \tfrac{1}{n_1}(\overline{R}_{2\cdot} - \tfrac{n_2+1}{2}).
\end{aligned}$$

Analog gilt für die gepoolten Verteilungsfunktionen unter der Nullhypothese $H_{0k}^F : F_0 = \cdots = F_k$

$$p_s = \int H(x) dF_s(x) = \int \sum_{i=0}^{k} \tfrac{n_i}{N} F_i(x)\, dF_s(x) = \tfrac{1}{2}\ . \qquad (5.1)$$

Diese p_s können als Wahrscheinlichkeit interpretiert werden, daß eine Zufallsgröße, deren Verteilung F_s ist, größer ist als eine Zufallsgröße, deren Verteilung durch die gepoolte Verteilung H gegeben ist. Die Größe p_i (zusammengefaßt in $\boldsymbol{p} = (p_0, ..., p_k)'$) wird daher als relativer Effekt der Gruppe *i* in bezug auf alle anderen Gruppen bezeichnet [146]. Dies entspricht der Methodik der Ridit-Analyse [147], wo die Verteilungen der verschiedenen Stichproben mit einer Standardverteilung verglichen werden. Als Standardverteilung wird im obigen Fall die gepoolte Verteilung genutzt. Die Schätzer $\hat{p}_s$ können somit auch als „mean ridit" der Gruppe *s* bezeichnet werden. Ist das „mean ridit" der Gruppe *s* größer (kleiner) als 0.5, so tendieren die Elemente der Stichprobe *s* zu größeren (kleineren) Werten auf der ordinalen Skala als Elemente der Standardgruppe. Werden die Verteilungsfunktionen bzw. die empirischen Verteilungsfunktionen im Vektor $\boldsymbol{F} = (F_0, ..., F_k)'$ bzw. $\hat{\boldsymbol{F}} = (\hat{F}_0, ..., \hat{F}_k)'$ zusammengefaßt und stellt $\boldsymbol{C}$ eine Kontrastmatrix dar, so sind $H_{0k}^F : F_0 = \cdots = F_k$ und $H_{0k}^F : \boldsymbol{CF} = 0$ äquivalente Formulierungen der Nullhypothese. Gilt die Nullhypothese H_{0k}^F, so

folgt $\boldsymbol{p} = \frac{1}{2}\boldsymbol{1}_{k+1}$. Aus der Negation der letzten Aussage folgt, daß falls $\boldsymbol{p} = \frac{1}{2}\boldsymbol{1}_{k+1}$ nicht gilt, auch $F_0 = \cdots = F_k$ nicht gelten kann. Sind die Verteilungsfunktionen stochastisch geordnet, dann folgt auch $p_0 \leq \cdots \leq p_k$. Werden in p_s die normalisierten Verteilungsfunktionen durch die normalisierten empirischen Verteilungsfunktionen ersetzt, ergibt sich:

$$\hat{p}_s = \int \hat{H}(x) d\hat{F}_s(x) = \int \sum_{i=0}^{k} \tfrac{n_i}{N} \hat{F}_i(x)\, d\hat{F}_s = \tfrac{1}{n_s} \sum_{j=1}^{n_s} \sum_{i=0}^{k} \tfrac{n_i}{N} \hat{F}_i(x_{sj})$$

$$= \tfrac{1}{n_s} \sum_{j=1}^{n_s} \sum_{i=0}^{k} \sum_{t=1}^{n_i} \tfrac{1}{N} c(x_{sj} - x_{it}) = \tfrac{1}{N}(\overline{R}_{s.} - \tfrac{1}{2}).$$

Die Schätzer können somit gut interpretiert und leicht berechnet werden. Die folgenden Bedingungen sind zudem hinreichend, um Konsistenz- und Verteilungsaussagen für $\hat{\boldsymbol{p}}$ herzuleiten:

(A0) $X_{ij} \sim F_i$ *i=0,...,k und* X_{ij} *seien unabhängig voneinander*

(A1) $\min_{i=0,\ldots,k} n_i \to \infty$

(A2) $0 < \lambda_0 \leq n_i / N \leq 1 - \lambda_0 < 1$

(A3) $H_{0k}^F : F_0 = \cdots = F_k$

(A4) $\sigma_i^2 = Var(H(X_{ij})) > 0$ für $i = 0,\ldots,k$.

Hilfssatz 5.1: *Sind die Bedingungen (A0)- (A2) erfüllt, dann ist*

$$\hat{\boldsymbol{p}} = (\hat{p}_0,\ldots,\hat{p}_k)' = \int \hat{H}(x)\, d\hat{\boldsymbol{F}}(x) = \tfrac{1}{N}(\overline{R}_{0.} - \tfrac{1}{2},\ldots,\overline{R}_{k.} - \tfrac{1}{2})'$$

ein unverzerrter und konsistenter Schätzer für $\boldsymbol{p}$.
Beweis: siehe Brunner und Puri [146].

Zur Herleitung von Verteilungsaussagen für $\boldsymbol{p}$ werden die Statistik

$$\sqrt{N} \int \hat{H}(x) d(\hat{\boldsymbol{F}} - \boldsymbol{F})(x) \tag{5.2}$$

und die Statistik

$$\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})(x) \tag{5.3}$$

betrachtet. In Satz 5.1 wird gezeigt, daß dies zwei äquivalente Statistiken sind (d. h., ihre asymptotischen Verteilungen sind gleich). Es ist daher ausreichend, Verteilungsaussagen für (5.3) herzuleiten. Gilt die Nullhypothese (A3), so folgt:

$$\boldsymbol{C}\sqrt{N}\int \hat{H}(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})(x)=\boldsymbol{C}\sqrt{N}\,\hat{\boldsymbol{p}}-\sqrt{N}\int \hat{H}(x)d(\boldsymbol{CF})(x)=\boldsymbol{C}\sqrt{N}\,\hat{\boldsymbol{p}}$$

bzw.

$$\boldsymbol{C}\sqrt{N}\int \hat{H}(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})(x)=\boldsymbol{C}\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}})(x).$$

Aufgrund des Zusammenhangs von $\hat{H}$ und den Rängen werden die Komponenten $\overline{Y}_i=\sqrt{N}\,\frac{1}{n_i}\sum_{j=1}^{n_i}H(x_{ij})$ des Vektors $\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}})(x)$ als asymptotisch rangtransformiert (ART) bezeichnet. Diese Komponenten besitzen den Vorteil, daß sie voneinander unabhängig und gleichmäßig beschränkt sind. Verteilungsaussagen können somit auf der Grundlage des Zentralen Grenzwertsatzes hergeleitet werden. Alle Größen von $\boldsymbol{C}\sqrt{N}\,\hat{\boldsymbol{p}}$ sind im Gegensatz zu denen von $\boldsymbol{C}\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}})(x)$ bekannt. Die Statistik (5.3) wird daher nur zur Herleitung von Verteilungsaussagen benutzt. Die eigentlichen Tests werden auf der Basis von (5.2) konstruiert.

Satz 5.1: *Falls (A0)-(A3) gelten, sind* $\sqrt{N}\int \hat{H}(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})$ *und* $\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})$ *asymptotisch äquivalente Statistiken. Gilt zusätzlich (A4), so ist die asymptotische Verteilung von* $\sqrt{N}\int H(x)d(\hat{\boldsymbol{F}}-\boldsymbol{F})$ *eine multivariate Normalverteilung N(**0**,**V**) mit* $\boldsymbol{V}=N\bigoplus_{i=0}^{k}\frac{\sigma_i^2}{n_i}$.

Beweis: siehe Brunner und Puri [146].

Die noch von unbekannten Größen abhängende Kovarianzmatrix kann konsistent durch die Ränge geschätzt werden.

Satz 5.2: *Die Kovarianzmatrix* $\boldsymbol{V} = N \bigoplus_{i=0}^{k} \frac{\sigma_i^2}{n_i}$ *kann konsistent durch die Matrix* $\hat{\boldsymbol{V}}_{III} = N \bigoplus_{i=0}^{k} \frac{\hat{\sigma}_i^2}{n_i}$ *(*$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (R_{ij} - \overline{R}_{i.})^2$*) geschätzt werden, falls die Bedingungen (A0)-(A4) erfüllt sind. Da unter der Bedingung (A3)* $\sigma^2 = \sigma_0^2 = \cdots = \sigma_k^2$ *gilt, folgt auch:*

1. $\hat{\sigma}_I^2 = \frac{1}{(N-k-1)} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (R_{ij} - \overline{R}_{i.})^2 \xrightarrow{p} \sigma^2$

2. $\hat{\sigma}_{II}^2 = \frac{1}{(N-1)} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (R_{ij} - \overline{R}_{..})^2 \xrightarrow{p} \sigma^2$.

Beweis: siehe Brunner und Puri [146].

Bemerkung 5.1: Werden mit $(a_1, ..., a_r)$ die r unterschiedlichen Ränge in der gepoolten Stichprobe und mit $(t_1, \ldots, t_r)$ ihre Häufigkeiten bezeichnet, so gilt [72, S.329]

$$\frac{1}{N} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (R_{ij} - \overline{R}_{..})^2 = \frac{N^2 - 1}{12} - \frac{1}{12\,N} \sum_{s=1}^{r} t_s\,(t_s^2 - 1).$$

Der Varianzschätzer $\hat{\sigma}_{II}^2$ führt somit genau auf die bei Rangstatistiken häufig verwendete Bindungskorrektur:

$$\begin{aligned}
\hat{\sigma}_{II}^2 &= \frac{N}{(N-1)} \left(\frac{N^2 - 1}{12} - \frac{1}{12\,N} \sum_{s=1}^{r} t_s\,(t_s^2 - 1) \right) \\
&= \frac{N(N+1)}{12} \left(1 - \frac{1}{12\,N(N^2-1)} \sum_{s=1}^{r} t_s\,(t_s^2 - 1) \right) \\
&= \frac{N(N+1)}{12} \quad \text{(falls keine Bindungen auftreten).}
\end{aligned}$$

Satz 5.3: *Es seien die Bedingungen (A0)-(A4) erfüllt, und* $\boldsymbol{C}$ *sei eine beliebige Kontrastmatrix. Dann gilt:*

1. *Die asymptotische Verteilung der Statistik* $\sqrt{N}\boldsymbol{C}\hat{\boldsymbol{p}}$ *ist eine multivariate Normalverteilung* $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ *mit* $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'$*, wobei* $\boldsymbol{\Sigma}$ *durch* $\hat{\boldsymbol{\Sigma}} = \boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}'$ *konsistent geschätzt werden kann.*

2. *Die asymptotische Verteilung der quadratischen Form* $Q_N(\boldsymbol{C}) = N\hat{\boldsymbol{p}}'\boldsymbol{C}'[\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}']^{-}\boldsymbol{C}\hat{\boldsymbol{p}}$ *ist eine zentrale* χ^2_f*-Verteilung mit Freiheitsgrad f=Rang(**C**). Hierbei ist* $[\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}']^{-}$ *eine verallgemeinerte inverse Matrix von* $\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}'$.
3. *Hat die Matrix* $\boldsymbol{C}$ *vollen Zeilenrang, so ist die asymptotische Verteilung von* $Q_N(\boldsymbol{C}) = N\hat{\boldsymbol{p}}'\boldsymbol{C}'[\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}']^{-1}\boldsymbol{C}\hat{\boldsymbol{p}}$ *eine zentrale* χ^2_f*-Verteilung mit Freiheitsgrad f=Rang(**C**). Hierbei ist* $[\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}']^{-1}$ *dann die stets existierende inverse Matrix von* $\boldsymbol{C}\hat{\boldsymbol{V}}_{III}\boldsymbol{C}'$.

Beweis: siehe Brunner und Puri [146].

Chacko [27] führte einen zum Bartholomew-Test analogen Rangtest unter Ordnungsrestriktionen für balancierte einfaktorielle Anlagen ein. Shorack [148] verallgemeinerte diesen Test auf unbalancierte Anlagen. Allerdings gehen beide von stetigen Verteilungsfunktionen aus. Im folgenden Satz wird nun gezeigt, daß die von Chacko bzw. Shorack gezeigten Verteilungsaussagen auch bei nichtstetigen Verteilungsfunktionen erhalten bleiben.

Satz 5.4: D*ie Bedingungen (A0)-(A4) seien erfüllt. Dann ist die asymptotische Verteilung von*

$$\overline{\chi}^2_{01}(R_{II}) = \frac{\sum_{i=0}^{k} n_i\,(\overline{R}^*_{i\cdot} - (N+1)/2)^2}{(N-1)^{-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(R_{ij} - (N+1)/2)^2} \tag{5.5}$$

bzw.

$$\overline{\chi}^2_{01}(R_I) = \frac{\sum_{i=0}^{k} n_i\,(\overline{R}^*_{i\cdot} - (N+1)/2)^2}{(N-k-1)^{-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(R_{ij} - \overline{R}_{i\cdot})^2} \tag{5.6}$$

für t>0 durch

$$P(\overline{\chi}^2_{01}(R_s) \geq t) = \sum_{l=2}^{k+1} P(l, k+1, \boldsymbol{w})\, P(\chi^2_{l-1} \geq t) \quad s = I, II$$

gegeben. Hierbei sind $\overline{R}^*_{i\cdot}$ *die isotonen Schätzer bzgl. der Gewichte* $w_i = n_i$.

Beweis: Der Beweis beruht auf den oben zusammengefaßten Ergebnissen und auf dem Beweis des analogen Satzes für stetige Verteilungsfunktionen durch Robertson et al.[114, S.205]. Die Aussagen für (5.5) und (5.6) sind äquivalent beweisbar, daher wird hier nur (5.5) bewiesen.

Für $i=0,...,k$ sei $\lim_{\min\{n_i : i=0,...,k\}\to\infty} \frac{n_i}{N} = \gamma_i, \gamma_i \in (0,1)$ und $\boldsymbol{I}_{k+1}$ die $(k+1)\times(k+1)$ Einheitsmatrix. Mittels dieser Größen werde folgende Kontrastmatrix definiert: $\boldsymbol{C}_{\boldsymbol{\gamma}} = \boldsymbol{I}_{k+1} - \boldsymbol{1}_{k+1}\boldsymbol{\gamma}'$ (daß $\boldsymbol{C}_{\boldsymbol{\gamma}}$ eine Kontrastmatrix ist, folgt aus der Additivität des Grenzwertes). Analog werde mit $\boldsymbol{n}_N = (\frac{n_0}{N},...,\frac{n_k}{N})'$ die Matrix $\boldsymbol{C}_{\boldsymbol{n}_N} = \boldsymbol{I}_{k+1} - \boldsymbol{1}_{k+1}\boldsymbol{n}_N'$ eingeführt. Weiterhin seien $Z_0,...,Z_k$ unabhängige, normalverteilte Zufallsvariablen mit $P_{Z_i} = N(0,\gamma_i^{-1})$ und $\hat{\mu} = \sum_{i=0}^{k} \gamma_i Z_i / \sum_{i=0}^{k} \gamma_i$. Die Verteilung des Vektors $\boldsymbol{U} = (Z_0 - \hat{\mu},...,Z_k - \hat{\mu})'$ ist dann eine multivariate Normalverteilung $N(\boldsymbol{0}, \boldsymbol{\Sigma}^{(u)})$ mit

$$\sigma_{ij}^{(u)} = \begin{cases} \gamma_i^{-1} - 1 & : i = j \\ -1 & : i \neq j \end{cases}.$$

Sind $Z_0^*,...,Z_k^*$ die isotonen Schätzer bzgl. der Gewichte $\boldsymbol{\gamma} = (\gamma_0,...,\gamma_k)'$, dann ist die Verteilung von $T = \sum_{i=0}^{k} \gamma_i (Z_i^* - \hat{\mu})^2$ durch Satz 4.1 gegeben. Gilt die Nullhypothese (A3), so ist nach Satz 5.3 die asymptotische Verteilung von $\sqrt{N}\boldsymbol{C}_{\gamma}\hat{\boldsymbol{p}}$ eine multivariate Normalverteilung mit $\boldsymbol{\Sigma} = \boldsymbol{C}_{\boldsymbol{\gamma}}\boldsymbol{V}\boldsymbol{C}_{\boldsymbol{\gamma}}'$. Der Satz von Slutzky [149, S.6] sichert, daß $T_N = \frac{1}{\sqrt{\sigma^2}}\sqrt{N}\boldsymbol{C}_{\boldsymbol{n}_N}\hat{\boldsymbol{p}}$ sowie $\hat{T}_N = \frac{1}{\sqrt{\hat{\sigma}^2}}\sqrt{N}\boldsymbol{C}_{\boldsymbol{n}_N}\hat{\boldsymbol{p}}$ gegen $N(\boldsymbol{0}, \boldsymbol{\Sigma}^{(u)})$ $(\boldsymbol{\Sigma}^{(u)} = \sigma^{-2}\boldsymbol{\Sigma}, \text{(aufgrund von } (A3)\ \ \sigma_i = \sigma_j = \sigma))$ konvergiert. Stellt $\Pr_{\boldsymbol{n}_N}(T_N|K)$ die in Bemerkung 4.1 beschriebene Projektion von T_N auf den polyhydralen Kegel K bzgl. der Gewichte $\boldsymbol{n}_N = (\frac{n_0}{N},...,\frac{n_k}{N})'$ und der gegebenen Alternativhypothese H_{Ak}^F dar, so kann (5.5) auch durch folgende Gleichung beschrieben werden:

$$\overline{\chi}_{01}^2(R) = \sum_{i=0}^{k} \frac{n_i}{N}\left[\Pr_{\boldsymbol{n}_N}(T_N \mid K)_i\right]^2.$$

Insgesamt folgt:

a) Die Verteilung von $\boldsymbol{T}_N$ konvergiert schwach gegen die Verteilung von $\boldsymbol{U}$ ($P_{T_N} \Rightarrow P_U$).

b) $P_{\Pr_{\boldsymbol{n}_N}(T_N|K)} \Rightarrow P_{\Pr(U|K)}$, da bei Anwendung einer stetigen Funktion die schwache Konvergenz erhalten bleibt[149, S.9]. Die Stetigkeit der Projektion bzgl. ihres Argumentes und der Gewichte folgt aus (4.3), der Stetigkeit der Maximumbildung und der Stetigkeit der Minimumbildung.

c) Die Stetigkeit der Summe und der quadratischen Funktion sowie der Satz von Slutzky führen zur Aussage des Satzes.

Wird zur Konstruktion der Teststatistiken eine Scorefunktion $J_{ij} = J(\frac{1}{N}(R_{ij} - \frac{1}{2}))$ genutzt, so können Lemma 5.1 und alle nachfolgenden Sätze ebenfalls gezeigt werden, falls

a) J eine zweimal stetig differenzierbare Funktion ist, dessen 2. Ableitung beschränkt ist (Beweis: Munzel [150]) oder

b) J eine beschränkte Funktion ist
(Beweis: Bregenzer [151]).

Eine Verallgemeinerung des Chacko-Tests auf Scorefunktionen, wie es Shiraishi [152] für stetige Verteilungsfunktionen beschreibt, ist demnach ebenfalls möglich.

Satz 5.5: *Die Bedingungen (A0)-(A4) seien erfüllt. Dann ist die asymptotische Verteilung von*

$$\overline{\chi}^2_{01}(J(R_{II})) = \frac{\sum_{i=0}^{k} n_i (J_i^* - \bar{J}..)^2}{(N-1)^{-1} \sum_{i=0}^{k} \sum_{j=1}^{n_i} (J_{ij} - \bar{J}..)^2}$$

mit $\bar{J}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} J_{ij}$, $\bar{J}_{..} = \frac{1}{N} \sum_{i=0}^{k} \sum_{j=1}^{n_i} J_{ij}$ *und den isotonen Schätzern* J_i^* *durch*

$$P(\overline{\chi}^2_{01}(J(R_{II})) \geq t) = \sum_{l=2}^{k+1} P(l, k+1, w)\, P(\chi^2_{k-1} \geq t)$$

bestimmt.

Beweis: siehe oben.

Exakte und approximative Permutationstests, angewandt auf die rangtransformierten Daten, sind analog Abschnitt 4.1 ohne Probleme konstruierbar. Beim „Bootstrapen“ ist dies schon

etwas schwieriger. Prinzipiell kann mit den rangtransformierten Daten ein Bootstraptest durchgeführt werden. Diese Variante entspricht der Rangtransformationsmethode [153]: „transformiere die Daten in Ränge und werte sie parametrisch aus“. Exakter ist jedoch die Variante, bei der jeweils aus den Originaldaten gezogen wird und die Ränge neu bestimmt werden. Für diese Variante existieren unter der Annahme stetiger Verteilungsfunktionen Konsistenzaussagen [154]. In ebengenannter Arbeit werden auch Ansätze beschrieben, wie es möglich wäre, Konsistenzaussagen für diskrete Daten herzuleiten. Der aufwendige und sehr technische Weg soll hier aber nicht gegangen werden. In den Simulationen werden daher beide Varianten untersucht.

5.2 Nichtparametrische Kontraststatistiken

Analog zum parametrischen Fall können nichtparametrische einfache als auch multiple Kontrasttests durch

$$T_l(R_s) = \frac{\sqrt{N}\sum_{i=0}^{k} c_{li}\overline{R}_{i.}}{\hat{\sigma}_s\sqrt{\sum_{i=0}^{k}\frac{N\,c_{li}^2}{n_i}}},\ s=I,II\ ,\ \text{und}\quad T^{max}(R_s) = max(T_1(R_s),\ldots,T_m(R_s))$$

mit

$$\hat{\sigma}_I^2 = \frac{1}{N-k-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(R_{ij}-\overline{R}_{i.})^2\ ,\ \hat{\sigma}_{II}^2 = \frac{1}{N-1}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(R_{ij}-\overline{R}_{..})^2 \text{ und } \sum_{i=0}^{k} c_{li} = 0$$

definiert werden. Die Verteilungsaussagen sind durch die Sätze 5.1-5.3 gegeben.

Satz 5.6: *Die Bedingungen (A0)-(A4) seien erfüllt. Dann folgt:*

1. $P_{T_l(R_s)} \Rightarrow N(0,1)\ ,\ s = I, II\ .$
2. $P_{\boldsymbol{T}(R_s)} \Rightarrow N(\boldsymbol{0},\boldsymbol{\rho})$ mit $\boldsymbol{T}(R_s) = (T_1(R_s),\ldots,T_m(R_s))'$, $\boldsymbol{\rho} = (\rho_{uv})_{1\le uv\le m}$ und

$$\rho_{uv} = \frac{N\sum_{i=0}^{k} c_{ui}\,c_{vi}/n_i}{\sqrt{N\sum_{i=0}^{k} c_{ui}^2/n_i}\ \sqrt{N\sum_{i=0}^{k} c_{vi}^2/n_i}} \text{ für } u,v = 1,\ldots,m.$$

3. *Mit $T^{max}(R_s)$ kann wieder ein multipler Kontrast definiert werden. Seine Verteilung konvergiert schwach gegen die Verteilung von $Z = \max(Z_1,...,Z_m)$, wobei die gemeinsame Verteilung der $(Z_1,...,Z_m)$ die Grenzverteilung aus dem 2. Punkt ist.*

Beweis: 1. und 2. folgen direkt aus Satz 5.3, indem die dort beliebige Kontrastmatrix ***C*** entsprechend des gewählten Kontrastes bzw. der *m* gewählten Kontraste konstruiert wird. Der 3. Punkt beruht auf der Stetigkeit der Maximumbildung und dem Stetigkeitssatz für die schwache Konvergenz [149, S.9].

Bemerkung 5.2 Für kleine Fallzahlen schlagen Brunner und Puri [155] vor, empirisch die Verteilung der Kontraste (5.7) durch eine t-Verteilung mit *N-k-2* Freiheitsgraden zu approximieren. Wie die Simulationen zeigen werden, ist es auch für die multiplen Kontraste auf Basis von $\hat{\sigma}_I^2$ empfehlenswert, die multivariate Normalverteilung empirisch durch eine multivariate t-Verteilung mit *N-k*-2 Freiheitsgraden zu ersetzen.

Genau wie in der parametrischen Situation können asymptotisch optimale Kontraste definiert werden [156]. Allerdings ist dies nur im Lokationsmodell relativ leicht.
Aufgrund von Satz 5.3 können auch die folgenden Kontraste definiert werden.

Satz 5.7: *Die Bedingungen (A0)-(A4) seien erfüllt.*

1. *Die asymptotische Verteilung der Kontraststatistik*

$$T_l(R_{III}) = \frac{1}{\hat{\sigma}_{III}^2}\sqrt{N}\sum_{i=0}^{k} c_{li}\overline{R}_{i.} \tag{5.8}$$

bzw.

$$T_l(R_{IV}) = \frac{\sqrt{N}\sum_{i=0}^{k}\sqrt{\frac{n_i}{N\hat{\sigma}_i^2}}c_{li}\overline{R}_{i.}}{\sqrt{\sum_{i=0}^{k} c_{li}^2}} \tag{5.9}$$

mit $c_{l.} = \sum_{i=1}^{k} c_{li} = 0$, $\hat{\sigma}_{III}^2 = \sum_{i=0}^{k}\frac{N\hat{\sigma}_i^2 c_{li}^2}{n_i}$ *und* $\hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(R_{ij} - \overline{R}_{i.})^2$ *ist die Standardnormalverteilung.*

2. *Die gemeinsame Verteilung von m Kontraststatistiken der Form (5.8) bzw. von m Kontrasten der Form (5.9) konvergiert schwach gegen eine multivariate Normalverteilung mit dem Erwartungswert* $\boldsymbol{0}$ *und der Korrelationsmatrix* $\widetilde{\boldsymbol{\rho}}$*. Dabei sind die Koeffizienten von* $\widetilde{\boldsymbol{\rho}}$ *im Fall (5.8) durch*

$$\widetilde{\rho}_{uv} = \frac{N\sum_{i=0}^{k} \hat{\sigma}_i^2 c_{ui}\, c_{vi} / n_i}{\sqrt{N\sum_{i=0}^{k} \hat{\sigma}_i^2 c_{ui}^2 / n_i}\sqrt{N\sum_{i=0}^{k} \hat{\sigma}_i^2 c_{vi}^2 / n_i}} \text{ mit } u,v = 1,\ldots,m$$

und im Fall (5.9) durch

$$\widehat{\rho}_{uv} = \frac{\sum_{i=0}^{k} c_{ui}\, c_{vi}}{\sqrt{\sum_{i=0}^{k} c_{ui}^2}\sqrt{\sum_{i=0}^{k} c_{vi}^2}} \text{ mit } u,v = 1,\ldots,m$$

gegeben.

3. *Mit* $T^{\max}(R_{III}) = \max(T_1(R_{III}),\ldots,T_m(R_{III}))$ bzw. $T^{\max}(R_{IV}) = \max(T_1(R_{IV}),\ldots,T_m(R_{IV}))$ *kann ein multipler Kontrast definiert werden. Die Verteilung konvergiert schwach gegen die Verteilung von* $Z = \max(Z_1,\ldots,Z_m)$*, wobei die gemeinsame Verteilung der* $(Z_1,\ldots,Z_m)$ *die jeweilige Grenzverteilung aus dem 2. Punkt ist.*

Beweis:

Die Aussage für (5.8) folgt aus Satz 5.3, indem nicht wie in Satz 5.6 ein gepoolter Varianzschätzer, sondern $\hat{\boldsymbol{V}}_{III} = N\bigoplus_{i=0}^{k}\frac{\hat{\sigma}_i^2}{n_i}$ aus Satz 5.2 benutzt wird. Zum Nachweis von (5.9) seien folgende Matrizen definiert:

$$\boldsymbol{D} = \tfrac{1}{\sqrt{N}}\bigoplus_{i=0}^{k}\sqrt{\tfrac{n_i}{\sigma_i^2}} \text{ (existiert da (A4) gilt), } \hat{\boldsymbol{D}} = \tfrac{1}{\sqrt{N}}\bigoplus_{i=0}^{k}\sqrt{\tfrac{n_i}{\hat{\sigma}_i^2}},\ \boldsymbol{P}_{k+1} = \boldsymbol{I}_{k+1} - \frac{1}{k+1}\boldsymbol{1}_{k+1}\boldsymbol{1}_{k+1}'.$$

Aus Satz 5.3 folgt

$$\boldsymbol{D}\sqrt{N}\boldsymbol{P}_{k+1}\hat{\boldsymbol{p}} \sim N(\boldsymbol{0}, N\,\boldsymbol{D}\boldsymbol{P}_{k+1}\boldsymbol{V}\boldsymbol{P}_{k+1}\boldsymbol{D}).$$

Der Satz von Slutzki sichert auch die Gültigkeit von

$$\hat{\boldsymbol{D}}\sqrt{N}\boldsymbol{P}_{k+1}\hat{\boldsymbol{p}} \sim N(\boldsymbol{0}, N\,\hat{\boldsymbol{D}}\,\boldsymbol{P}_{k+1}\boldsymbol{V}\,\boldsymbol{P}_{k+1}\hat{\boldsymbol{D}}).$$

Somit folgt die Behauptung aus

$$\boldsymbol{C}\hat{\boldsymbol{D}}\sqrt{N}\boldsymbol{P}_{k+1}\hat{\boldsymbol{p}} \sim N(\boldsymbol{0}, N\boldsymbol{C}\hat{\boldsymbol{D}}\boldsymbol{P}_{k+1}\boldsymbol{V}\boldsymbol{P}_{k+1}\hat{\boldsymbol{D}}\boldsymbol{C}')\,.$$

Der 3. Punkt beruht wieder auf der Stetigkeit der Maximumbildung.
Die Kontrasttests der Form (5.8) sind das nichtparametrische Analogon zu den von Grimes [117] betrachteten parametrischen Kontrasten. Die von Akritas und Brunner [48] vorgeschlagene t-Verteilungsapproximation mit Freiheitsgrad *df*,

$$df = \frac{(\sum_{i=0}^{k} c_i^2\,\hat{\sigma}_i^2 / n_i)^2}{\sum_{i=0}^{k} c_i^4\,\hat{\sigma}_i^4 / (n_i^2\,(n_i - 1))}\,,$$

ist daher naheliegend. Die Kontraste der Form (5.9) sind genauso wie der WT-Test von Roth anfällig gegen Varianzschätzer mit dem Wert Null. Hinsichtlich Permutationstests bzw. Bootstraptests gelten die Aussagen von Abschnitt 5.1 und 4.2 auch hier.

5.3 Adaptive Tests

Bei genauer Betrachtung ist leicht zu erkennen, daß die vorgestellten nichtparametrischen Verfahren den Kategorien ebenfalls Scores zuordnen. Diese Scores sind die Durchschnittsränge der Kategorien in der gepoolten Stichprobe. Anders als a priori festgelegte Scores hängen sie jedoch von den vorliegenden Daten ab. Die einfache Darstellung und der gut interpretierbare Zusammenhang mit den Verteilungsfunktionen legt die Wahl der Durchschnittsränge zwar nahe, aber auch dies ist eine gewisse Willkür. Es könnten ebenso durch sogenannte Scorefunktionen ($J(X_{ij})$ oder auch $J(R_{ij})$) erzeugte Scores den Kategorien zugeordnet werden. Ist die Verteilung der Daten bekannt, können sowohl im stetigen Fall als auch im nicht-stetigen Fall für lineare Rangstatistiken lokal optimale Scorefunktionen (Scores) hergeleitet werden [157, S.108; 158], was die Wahl einer geeigneten Scorefunktion

erleichtert. In der Regel liegen jedoch keine Informationen über die Verteilung vor. In diesem Fall bieten sich adaptive Verfahren an. Adaptive Verfahren können in grob-adaptierende und fein-adaptierende Verfahren unterteilt werden. Bei den fein-adaptierenden Verfahren wird die optimale Scorefunktion geschätzt. Behnen und Neuhaus [93, S.208] beschreiben u. a. fein-adaptierende Trendtests für das Mehrstichprobenproblem. Die Scorefunktion wird bei diesem Verfahren mit Hilfe eines Kerndichteschätzers und der maximalen Ränge bestimmt. Aufgrund des hohen Aufwandes bei der Berechnung wurde die Güte dieses Verfahrens nur für wenige Situationen untersucht. Im wesentlichen zeigten sich kaum Gütevorteile gegenüber den multiplen Kontrasten. Aufgrund der sehr komplexen Mathematik und fehlender allgemeiner Aussagen (zu wenige Simulationen) werden diese Tests und die Simulationsergebnisse nicht dargestellt. Grob-adaptierende Verfahren sind meist Zweischrittverfahren. In einem ersten Schritt wird eine Selektorstatistik berechnet, und in Abhängigkeit von deren Größe wird aus einer vorgegebenen Menge von Scorefunktionen diejenige ausgewählt, die dann zur eigentlichen Testentscheidung (Generierung der Scores) genutzt wird. Obwohl zwei Schritte durchgeführt werden, sind im allgemeinen keine α-Adjustierungen notwendig (siehe z. B. Hogg et al. [159]). Die Festlegung der natürlichen Scores (1, 2, 3, 4...) ist naheliegend, weitere Scores lassen sich jedoch nur schwer vorgeben. Die Wahl mehrerer Scorefunktionen und somit von Scores, die von den Daten abhängen, kann jedoch aufgrund bekannter Ergebnisse für optimale Scorefunktionen gut vermittelt werden. Somit können adaptive Verfahren auch als eine Lösung für das Problem der Scorewahl betrachtet werden.

Hogg et al. [159] nutzen die Unabhängigkeit der Ordnungsstatistik und des Rangvektors zur Konstruktion eines adaptiven Zweistichprobentests ohne α-Adjustierungen. In einem ersten Schritt werden auf der Basis der Ordnungsstatistik Schätzer für die Schiefe ($\hat{Q}_1$) und für den Exzeß ($\hat{Q}_2$) bestimmt. Im zweiten Schritt wird eine lineare Rangstatistik in Abhängigkeit von $\hat{Q}_1$ und $\hat{Q}_2$ und einem fest vorgegebenen Auswahlschema berechnet. Hill et al. [160] verallgemeinern diesen Test auf das Mehrstichprobenproblem. Falls Bindungen in den Daten auftreten, empfehlen sie die Nutzung gemittelter Scores. Treten Bindungen auf, sind der Rangvektor und die Ordnungsstatistik jedoch im allgemeinen nicht unabhängig. Conover [158] zeigt, daß, wenn auf den Bindungsvektor bedingt wird, die Ordnungsstatistik und der Rangvektor unabhängig voneinander sind. Das Prinzip von Hogg et al. kann somit auch auf nichtstetige Daten übertragen werden. Beier und Bünning [161] schlagen folgende Schätzer für die Schiefe bzw. die Schwere der Verteilungsenden (Tails) vor:

$$\hat{Q}_1 = \frac{\hat{y}_{0.975} - \hat{y}_{0.5}}{\hat{y}_{0.5} - \hat{y}_{0.025}} \qquad \hat{Q}_2 = \frac{\hat{y}_{0.975} - \hat{y}_{0.025}}{\hat{y}_{0.875} - \hat{y}_{0.125}},$$

wobei gilt

$$\hat{y}_p = \begin{cases} Y_{(1)} & falls\ p < \frac{1}{2N} \\ (1-\lambda)Y_{(j)} + \lambda Y_{(j+1)} & falls\ \frac{1}{2N} \le p \le 1 - \frac{1}{2N} \\ Y_{(N)} & falls\ p > 1 - \frac{1}{2N} \end{cases} \quad \text{mit } \lambda = Np + 0{,}5 - j,\ j = \lfloor N\,p + 0{,}5 \rfloor.$$

Mittels dieser Maße legen sie folgenden adaptiven Test fest:

$$T_{ad} = \begin{cases} T_{ls} & if\ 0 \le \hat{Q}_1 \le 0{,}6; \hat{Q}_2 \ge 1 \\ T_{lt} & if\ 0{,}6 \le \hat{Q}_1 \le 2; 1 \le \hat{Q}_2 \le 1{,}5 \\ T_{ws} & if\ 0 \le \hat{S}_1 \le 2; 1{,}5 < \hat{Q}_2 \le 2 \\ T_{st} & if\ 0 \le \hat{Q}_1 \le 2; \hat{Q}_2 > 2 \\ T_{rs} & if\ \hat{Q}_1 > 2; \hat{Q}_2 \ge 2 \end{cases} .$$

Hierbei sind $T_{lt}, T_{rs}, T_{ws}, T_{st}$ und T_{lt} lineare Kontrasttests auf Basis der fünf verschiedenen Scorefunktionen. Die Scorefunktionen sind dabei effizient für:

1. rechtsschiefe Verteilungen [159]

$$a_{rs}(r) = \begin{cases} 0 & falls\ r > 0{,}5(N+1) \\ r - 0{,}5(N+1) & falls\ r \le 0{,}5(N+1) \end{cases}$$

2. linksschiefe Verteilungen [159]

$$a_{ls}(r) = \begin{cases} 0 & falls\ r \le 0{,}5(N+1) \\ r - 0{,}5(N+1) & falls\ r > 0{,}5(N+1) \end{cases}$$

3. Verteilungen mit leichten Tails [162]

$$a_{lt}(r) = \begin{cases} r - 0{,}25(N+1) & falls\ r \le 0{,}25(N+1) \\ 0 & falls\ 0{,}25(N+1) < r < 0{,}75(N+1) \\ r - 0{,}75(N+1) & falls\ r \ge 0{,}75(N+1) \end{cases}$$

4. Verteilungen mit schweren Tails [163]

$$a_{st}(r) = \begin{cases} -(\lfloor 0{,}25N \rfloor + 1) & falls\, r < \lfloor 0{,}25N \rfloor + 1 \\ r - 0{,}5(N+1) & falls \lfloor 0{,}25N \rfloor + 1 \le r \le \lfloor 0{,}75(N+1) \rfloor \\ \lfloor 0{,}25N \rfloor + 1 & falls\, r \ge \lfloor 0{,}75(N+1) \rfloor \end{cases}$$

5. symmetrische Verteilungen mit mittleren bis schweren Tails (Wilcoxon Scores) [164]

$$a_{ws}(r) = r,\ r = 1,\ldots,N\ .$$

Aufgrund der starken Profilabhängigkeit des linearen Kontrasttests schlagen Seidel et al. [165] statt dessen die Benutzung der multiplen Kontraste T_u^{max} ($u = ls, rs, ws, lt, st$) vor (es könnten auch die von Shiraishi [152] vorgeschlagenen Verallgemeinerungen des Chacko-Tests genutzt werden):

$$T_{ad5}^{hogg} = \begin{cases} T_{ls}^{\max} & if\ 0 \le \hat{Q}_1 \le 0.6; \hat{Q}_2 \ge 1 \\ T_{lt}^{\max} & if\ 0.6 \le \hat{Q}_1 \le 2; 1 \le \hat{Q}_2 \le 1.5 \\ T_{ws}^{\max} & if\ 0 \le \hat{S}_1 \le 2; 1.5 < \hat{Q}_2 \le 2 \\ T_{st}^{\max} & if\ 0 \le \hat{Q}_1 \le 2; \hat{Q}_2 > 2 \\ T_{rs}^{\max} & if\ \hat{Q}_1 > 2; \hat{Q}_2 \ge 2 \end{cases} .$$

Der somit erhaltene Test (basierend auf mehreren Scores) ist robust gegen verschiedene Verteilungen und verschiedene Profile. Sowohl auf der Basis asymptotischer Tests als auch auf der Basis von Permutations- bzw. Bootstraptests kann dieses Konzept auf ordinale Daten übertragen werden. Neuhäuser et al.[166] schlagen alternativ zu diesem Zweischrittverfahren multiple Kontrasttests vor, die mittels verschiedener Scores und verschiedener Kontrastvektoren definiert werden:

$$T_{ad3}^{max} = max(T_{ls}^{max}, T_{rs}^{max}, T_{ws}^{max})\ \text{ bzw. }\ T_{ad5}^{max} = max(T_{ls}^{max}, T_{rs}^{max}, T_{ws}^{max}, T_{lt}^{max}, T_{st}^{max})\ .$$

Mit diesen Statistiken können sowohl Permutationstests als auch Bootstraptests für beliebiges *k* auf der Basis einer beliebigen Kontrastmatrix (siehe Abschnitt 4.2) definiert werden.

Es existieren noch viele andere adaptive Verfahren. Donegani [167] beschreibt z. B. adaptive Verfahren, die auf der Unterteilung des Resamplingstichprobenraumes von Bootstraptests beruhen. Da sich in den Simulationen für die betrachteten Fallzahlen keine Gütegewinne gegenüber einfachen multiplen Kontrasten zeigten, wird auf die Beschreibung und Darstellung von Simulationsergebnissen dieser Verfahren verzichtet.

Werden bei multiplen Kontrasten keine Varianzschätzer benutzt, so können aufgrund von $(\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{X}}) = (\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}})$ auch diese Tests als adaptive Tests bezeichnet werden. Mittels $\widetilde{\boldsymbol{C}}_l\widetilde{\boldsymbol{X}}$ werden den Beobachtungen abhängig von *l* unterschiedliche Scores (Mittelwerte) zugeordnet, mit dem Vektor $\widetilde{\boldsymbol{c}}$ korreliert und anschließend das Maximum über *l* gesucht. Dies scheint auch eine Erklärung dafür zu sein, daß multiple Kontraste relativ robust gegenüber Abweichungen von speziellen Verteilungsannahmen sind.

6 Simulationen

6.1 Allgemeine Aussagen

Die in der vorliegenden Arbeit vorgestellten Teststatistiken können jeweils mit einer parametrischen Verteilung (Normalverteilung oder t-Verteilung) oder mit einer finiten Verteilung (Permutationsverteilung oder Bootstrapverteilung) zur Konstruktion eines gerichteten Tests benutzt werden. In diesem Kapitel werden die unterschiedlichen Kombinationen von Teststatistik und Verteilung mittels Simulationen untersucht und miteinander verglichen. Dazu wird das Güteverhalten der Tests sowohl unter der Nullhypothese „keine Wirkungsunterschiede" als auch unter der Alternative „mit steigender Dosis steigende Wirkung" simuliert. Das Versuchsdesign stellt stets eine vollständig randomisierte Anlage (einfaktorielles Design) dar, wobei sowohl die Anzahl der Versuchsglieder ($k+1$), die Anzahl der Wiederholungen je Versuchsglied (n_i = Fallzahl = Stichprobenumfang) und das Signifikanzniveau (α) variiert werden. Die Anzahl der Versuchsglieder wird nach oben durch $k+1=4$ (drei Dosen (niedrig, mittel, hoch) und eine Kontrolle) beschränkt. Diese Einschränkung spiegelt zum einen die Pflanzenschutzmittelversuche der BASF wider, ist jedoch häufig auch bei anderen Dosis-Wirkungs-Versuchen zu beobachten [168; 169]. In Hinblick auf die Bestimmung einer minimalen effektiven Dosis (MED, siehe Abschnitt 1.1) werden auch die Designs mit zwei ($k=1$) und drei ($k=2$) Versuchsgliedern untersucht. Die Anzahl der Wiederholungen beträgt je Versuchsglied drei, vier oder fünf. Der Einfachheit halber und um den Simulationsumfang etwas einzuschränken, werden nur balancierte Designs ($n=n_i$, $i=0,\ldots,k$) untersucht. Für die meisten Versuche in der Landwirtschaft bzw. im Gartenbau ist dies jedoch keine zu einschränkende Annahme. Zum Beispiel werden in mehreren Richtlinien [3; 170] balancierte Anlagen mit vier Wiederholungen je Versuchsglied empfohlen. Da in der Praxis das Signifikanzniveau $\alpha=0{,}05$ eine dominierende Stellung einnimmt, liegt der Schwerpunkt der Simulationen bei diesem Niveau. Zusätzlich werden das Niveau $\alpha=0{,}01$ und $\alpha=0{,}10$ in einem geringeren Umfang untersucht. Da geordneten kategorialen Daten kein allgemeines Verteilungsmodell unterstellt werden kann, werden mehrere, auf unterschiedlichen Zufallszahlen basierende Simulationsstudien durchgeführt. Den wichtigsten Parameter stellt hierbei die Anzahl der Punkte dar, auf die die den Versuchsgliedern zugeordnete Wahrscheinlichkeitsverteilung konzentriert ist. Bonituren erfolgen oft auf einer Skala, die von 1 bis r ($r=3,4,5,6,7,8,9$) reicht [2]. Zu beachten ist jedoch, daß die Skala für die Beurteilung der Wirkung vor dem Experiment festgelegt wird. Real können die Experimente bei dieser

Vorgehensweise auch Daten liefern, die auf weniger als r Punkte konzentriert sind. In den Simulationen sollen bei der Generierung der Datensätze planmäßig zwar auch auf r Punkte konzentrierte Verteilungen erzeugt werden. Werden nun jedoch trotzdem Datensätze generiert, die auf weniger als r Punkte konzentriert sind (dies kommt bei den betrachteten Fallzahlen besonders unter der Alternative relativ häufig vor), so werden sie realitätsgetreu mit in die Güteabschätzungen einbezogen.

Steht im Nenner einer Teststatistik ein Varianzschätzer, so kann es vorkommen, daß der Varianzschätzer gleich Null und die Teststatistik somit nicht berechenbar ist. Dieses Problem ist vor allem bei Bootstraptests relevant. Aber auch bei parametrischen Tests kann sich aufgrund zu geringer Variabilität in den Gruppen innerhalb der Simulationsexperimente ein Varianzschätzer mit dem Wert Null ergeben. In Abschnitt 3.1 wurde schon darauf eingegangen, wie solche Datensätze behandelt werden können, falls sie beim „Resamplen“ generiert werden. Völlig analog gelten die dort beschriebenen Möglichkeiten der Behandlung und die Folgen, falls schon der erzeugte Originaldatensatz zu einem Varianzschätzer mit dem Wert Null führt. In der vorliegenden Arbeit wird innerhalb der Simulationen jeder Datensatz, für den die Teststatistik nicht berechenbar ist, so gezählt, als wenn er nicht gegen die Nullhypothese spricht (z. B. wird dem p-Wert der Wert 1 zugewiesen). Die Nullhypothese wird demnach eher zu selten abgelehnt, die Güte unter der Alternative unterschätzt. In Anlehnung an die Definition eines konservativen Tests (lehnt bei Gültigkeit der Nullhypothese die Nullhypothese zu selten ab), soll diese Art der Powerschätzung als „konservative Powerschätzung“ bezeichnet werden.

Die Überlegenheit eines Tests gegenüber anderen Tests kann theoretisch in einigen Fällen bewiesen werden, indem gezeigt wird, daß der Test ein festgelegtes Kriterium (z. B. Minimierung des Fehlers 2. Art) erfüllt. Um dies zu zeigen, sind jedoch meist limitierende und schwer überprüfbare Bedingungen als erfüllt anzusehen. Simulationen kommen mit wenigen Annahmen aus, liefern jedoch keinen strengen Beweis für die Überlegenheit eines Tests (es werden nur endlich viele Situationen untersucht). Im folgenden wird ein Test als bester Test bezeichnet, wenn er sich durch das beste simulierte Güteverhalten auszeichnet. Ein Test wird nur dann empfohlen, wenn er folgende Bedingungen erfüllt:

1. Liegen Abweichungen von der Nullhypothese in Richtung der Alternativhypothese vor, so soll der Test dies mit einer möglichst hohen Güte erkennen.

2. Die simulierte Güte unter der Nullhypothese darf das vorgegebene Signifikanzniveau nicht wesentlich überschreiten. Als nicht wesentlich wird eine Überschreitung bewertet, wenn die simulierte Güte sich innerhalb der Grenzen des in Abschnitt 1.3 dargestellten zweiseitigen 0,99-Konfidenzintervalls befindet. Es könnte auch ein einseitiges 0,99-Konfidenzintervall benutzt werden, da hauptsächlich eine Überschreitung von Relevanz ist (die obere Grenze des Intervalls wäre dann für $\alpha = 0{,}05$ 0,0551 statt 0,0556). Um jedoch auch die Konservativität der Tests zu bewerten, wird das zweiseitige Intervall genutzt.

Zusätzlich werden folgende Eigenschaften als wünschenswert betrachtet:

3. Der Test sollte mit hoher Wahrscheinlichkeit durchführbar sein (Teststatistik berechenbar). Dieser Punkt ist vor allem für Teststatistiken mit Varianzschätzer relevant.
4. Wird die Verteilung der Dosisgruppen von der Nullhypothese in Richtung der Alternative verschoben, so soll dies mit einer höheren Güte des Tests verbunden sein. Aufgrund der oben definierten „konservativen“ Powerschätzung ist zu erwarten, daß, wenn der Abstand zwischen Nullhypothese und Alternativhypothese immer größer wird, die Variabilität in mindestens einer Gruppe gegen Null strebt. Daraus folgt zum einen eine höhere Wahrscheinlichkeit für einen nichtpositiven Varianzschätzer und zum anderen eine im allgemeinen noch diskretere finite Verteilung.
5. Der Test sollte leicht verständlich und leicht implementierbar sein.

Da gleiche Teststatistiken mit unterschiedlichen Verteilungen betrachtet werden, setzen sich die Bezeichner für die einzelnen Tests stets wie folgt zusammen:

Name|*Scores*|*Verteilung*|*Varianzschätzer.*

- *Name* stellt eine Abkürzung für die Teststatistik dar (z. B. RHK für den Reverse-Helmert-Kontrast).
- Mit *Scores* wird gekennzeichnet, ob der Test auf äquidistanten Scores (*p*; da meist parametrische Tests) oder auf Rängen (*np*; da meist nichtparametrische Tests) beruht.
- *Verteilung* benennt die Verteilungsaussagen, die für den jeweiligen Test genutzt werden. Hierbei steht *tv* für die t-Verteilung, *nv* für die Normalverteilung, *pea* für die approximative Permutationsverteilung, *pex* für die exakte Permutationsverteilung, *per* für

den randomisierten Permutationstest, *pmp* für den Mid-p-Test, *bov* für die vollständige Bootstrapverteilung („exakter" Bootstraptest), *bom* für die Bootstrapverteilung (Monte-Carlo), *db* für die Double-Bootstrapverteilung und *peu* für die vorgestellte unbedingte Permutationsverteilung (asymptotischer unbedingter Permutationstest).

- *Varianzschätzer* gibt an, auf welchem Varianzschätzer der Test beruht. Hierbei steht *V*1(*V*2, *V*3), wenn der Varianzschätzer auf S_I^2 (S_{II}^2, S_{III}^2) (falls Scores = p) bzw. auf $\hat{\sigma}_I^2$ ($\hat{\sigma}_{II}^2$, $\hat{\sigma}_{III}^2$) (falls Scores = np) beruht. Durch *oV* werden Tests gekennzeichnet, bei denen kein Varianzschätzer benutzt wird.

Zum Beispiel wird mit *TKnpnvV2* der Test bezeichnet, der auf der Kontrastmatrix $\boldsymbol{C}_{TK}$, auf Rängen, auf der Normalverteilungsapproximation und auf dem Varianzschätzer $\hat{\sigma}_{II}^2$ basiert. Allein durch die Wahl verschiedener Scores, Verteilungen und Varianzschätzer ergeben sich schon für eine Teststatistik (z. B. *TK)* mehr als 50 Kombinationsmöglichkeiten. Alle Ergebnisse können daher hier nicht dargestellt werden. In vielen Situationen zeigen die untersuchten Tests ein ähnliches Verhalten; daher werden nur ausgewählte Ergebnisse beschrieben. Die Ergebnisse werden vor allem verbal dargestellt und nur zum Teil durch Tabellen (vor allem bei geringen Unterschieden zwischen den Tests) oder Abbildungen (vor allem zur Veranschaulichung deutlicher Effekte) untermauert. Einige Kombinationen erweisen sich schon in den ersten Simulationsstudien als ungeeignet. Zum einen zeigt sich dies in deutlichen Niveauüberschreitungen, zum anderen in geringer Power (meist gekoppelt mit starker Konservativität). Im Abschnitt 6.2 werden daher anhand ausgesuchter Simulationsexperimente diejenigen Verfahren beschrieben, die bei den betrachteten Fallzahlen als weniger gut oder gar ungeeignet einzustufen sind. Schwerpunkt in Abschnitt 6.2 sind also nicht einzelne Tests oder ihr Güteverhalten bei ganz speziellen Konstellationen, sondern die den Tests zugrundeliegende Methodik (z. B. Varianzschätzer und Verteilungsapproximationen). Außerdem werden Vor- und Nachteile von Methoden beschrieben, die dann in Abschnitt 6.3 zur Konstruktion geeigneter Tests benutzt werden.

Die Simulationsergebnisse, die im folgenden beschrieben werden, basieren stets auf 10.000 Simulationen (siehe dazu Tabelle 1.4 im Abschnitt 1.3). Für die Monte-Carlo-Versionen der Bootstrap- und Permutationstests werden jeweils 10.000 Resamplingstichproben mit bzw. ohne Zurücklegen gezogen.

6.2 Ungeeignete Verfahren

In ersten Simulationsstudien wurden bekannte diskrete Verteilungen für $n \in \{3,4,5\}$ untersucht. Als Beispiel sei hier die Binomialverteilung genannt, d. h.

$$X_{ij} \sim B(m, \vartheta_i) \text{ mit } \pi_{is} = P(X_{ij} = s) = \binom{m}{s} \vartheta_i^s (1-\vartheta_i)^{m-s} \; s = 0,\ldots,m, (r = m+1).$$

Schwerpunkt dieser Studien waren die exakten und die approximativen Permutationstests sowie der Mid-p-Test. Außerdem wurden unterschiedliche Algorithmen zur Erzeugung zufälliger Kontingenztafeln untersucht. Als Statistiken wurden einfache Kontraste (Paarweiser Kontrast für $k = 1$ (tT), Paarweiser Kontrast (PK), Linearer Kontrast (LK), Helmert-Kontrast (HK) und Reverse-Helmert-Kontrast (RHK)) und multiple Kontraste (bivariater Kontrast BK=max(HK, RHK), trivariater Kontrast TK=max(HK, RHK, LK)) auf der Basis von Durchschnittsrängen jeweils ohne Varianzschätzer genutzt.

	Profil	$n = 6$, $k = 1$	$n = 4$, $k = 2$	$n = 3$, $k = 3$
tTnpbomoV	$\vartheta_0 = \vartheta_k = 0{,}25$	0,048	0,045	0,056
(stets $k = 1$)	$\vartheta_0 = 0{,}25, \vartheta_k = 0{,}55$	0,640	0,437	0,440
PKnpbomoV	$\vartheta_0 = \cdots = \vartheta_k = 0{,}25$		0,051	0,045
	$\vartheta_0 = 0{,}25, \vartheta_1 = \cdots = \vartheta_k = 0{,}55$		0,459	0,366
HKnpbomoV	$\vartheta_0 = \cdots = \vartheta_k = 0{,}25$		0,057	0,057
	$\vartheta_0 = 0{,}25, \vartheta_1 = \cdots = \vartheta_k = 0{,}55$		0,203	0,115
RHKnpbomoV	$\vartheta_0 = \cdots = \vartheta_k = 0{,}25$		0,054	0,050
	$\vartheta_0 = 0{,}25, \vartheta_1 = \cdots = \vartheta_k = 0{,}55$		0,596	0,542
TKnpbomoV	$\vartheta_0 = \cdots = \vartheta_k = 0{,}25$		0,054	0,051
	$\vartheta_0 = 0{,}25, \vartheta_1 = \cdots = \vartheta_k = 0{,}55$		0,521	0,430

Tabelle 6.1: Güte der Kontrasttests PK, HK, RHK und TK ohne Varianzschätzer auf der Basis von: Rängen, Bootstrapverteilung, binomialverteilten Zufallszahlen, konkaven Profilen ($m = 4$ und $\alpha = 0{,}05$)

Zum Vergleich wurden diese sieben Teststatistiken auch als Bootstraptests (ohne Varianzschätzer; Resamplingraum $RS(R) = \{R_{ij}, i = 0,\ldots,k, j = 1,\ldots,n_i\}$) mitgeführt. Anhand dieser Bootstraptests ist in Tabelle 6.1 beispielhaft die Abhängigkeit der Tests vom Profil der Erwartungswerte dargestellt. Besonders deutlich wird diese Abhängigkeit beim Helmert-Kontrast und beim Reverse-Helmert-Kontrast. Die Güte des trivariaten Kontrastes (TK) ist zwar deutlich schlechter als die des Reverse-Helmert-Kontrastes, allerdings ist sie bei weitem nicht so schlecht wie die des Helmert-Kontrastes. Bei konvexen Profilen ist es genau umgekehrt. Der Helmert-Kontrast besitzt die deutlich bessere Güte im Vergleich zum Reverse-Helmert-Kontrast. Der trivariate Kontrast hingegen liegt wieder deutlich näher am „optimalen" Kontrast. Im balancierten Fall sind sowohl der Helmert-Kontrast als auch der Reverse-Helmert-Kontrast aufgrund der speziellen Struktur der Koeffizienten als Zweistichprobentest interpretierbar. Beim Reverse-Helmert-Kontrast werden die letzten k Gruppen zusammengefaßt (gepoolt) und mit der Kontrolle verglichen (unbalanciertes Design mit $\tilde{n}_0 = n_0$ und $\tilde{n}_1 = n_1 + \cdots + n_k$). Beim Helmert-Kontrast werden die ersten k Gruppen zusammengefaßt und mit der höchsten Dosis verglichen ($\tilde{n}_0 = n_0 + \cdots + n_{k-1}$ und $\tilde{n}_1 = n_k$). In Tabelle 6.1 ist erkennbar, daß eine Aufteilung des Gesamtstichprobenumfangs ($12 = N = n_0 + n_1 = 6 + 6$) zu einer Güte von 0,640 führt (*tTnpbomoV*, $k = 1$). Eine Aufteilung von $3 + 9 = 12$ (*RHKnpbomoV*) ergibt hingegen 0,542. Analog erweist sich aufgrund der Ergebnisse des Helmert-Kontrastes bei konvexen Profilen eine Aufteilung von $9 + 3 = 12$ als schlecht. Bei permutativen Tests ist es ähnlich. Zum Beispiel würde für den Mid-p-Test (gleiche Situation) die Aufteilung $6 + 6 = 12$ zu einer Güte von 0,614 und die Aufteilung $3 + 9 = 12$ zu 0,520 führen. Für Permutationstests ist dies zum Teil erklärbar, da die Anzahl der verschiedenen Permutationen im balancierten Design maximal ist. Da es auch bei der Aufteilung $4 + 8 = 12$ mehr unterschiedliche Permutationen gibt als im Fall $3 + 9 = 12$, ist zumindest für den Reverse-Helmert-Kontrast der Vorteil des Dreistichprobendesigns gegenüber dem Vierstichprobendesign zu erklären. Für den trivariaten Kontrast (TK) ist dies nicht so einfach nachvollziehbar. Im Fall $k = 3$ gibt es für diesen Test mehr relevante, unterschiedliche Permutationen als im Fall $k = 2$.

Die wesentlichen Ergebnisse dieser ersten Screeningstudien können wie folgt zusammengefaßt werden:

1. Zur Erzeugung der zufälligen Kontingenztafeln wurden die Algorithmen „Rcount“ von Boyett [172] und „Rcount2“ von Patefield [173] untersucht. Beide Algorithmen erzeugen die Tafeln entsprechend ihrer exakten Wahrscheinlichkeiten unter der Nullhypothese. Als dritte Variante wurde ein primitiver Algorithmus („Permut“) untersucht, der auf einem einfachen Vertauschen der Stichprobenelemente basiert und wesentlich schneller ist. Bei der Verwendung von 10.000 zufälligen Permutationen und 10.000 durchgeführten Simulationsschritten ist nur ein kleiner, mit n abnehmender Unterschied (für $n > 3$ oft nicht größer als 0,005-0,02) zwischen den geschätzten Powerwerten des exakten Permutationstests und denen der zugehörigen approximativen Permutationstests festzustellen. Im allgemeinen sind die approximativen Verfahren jedoch etwas weniger konservativ und daher mit minimal höherer Power verbunden. Ein wesentlicher Einfluß des verwendeten Algorithmus auf die Güte konnte nicht aufgedeckt werden. Bis auf wenige extreme Situationen führen exakte Permutationstests bei den betrachteten Statistiken und Fallzahlen in einer angemessenen Zeit (unter 60 Sekunden) zu einer Entscheidung. Auf die Anwendung eines approximativen Tests sollte daher verzichtet werden. Wenn überhaupt ein approximativer Test benutzt werden soll, reicht der primitivste und schnellste Algorithmus „Permut“ aus.
2. Einfache Kontrasttests erweisen sich wie erwartet aufgrund ihrer Profilabhängigkeit auch für diskrete Daten als ungeeignet. Sie sollten daher nicht genutzt werden.
3. Basieren der Mid-p-Test und ein Bootstraptest auf derselben Teststatistik, so besitzen sie eine annähernd gleichhohe Güte. Die Bootstraptests neigen jedoch eher, wenn auch insgesamt selten, zu Niveauüberschreitungen.
4. Ähnliche Ergebnisse wurden auch in Simulationen mit poissonverteilten und negativ binomialverteilten Zufallszahlen beobachtet.

Als Beispiel sei hier das Verhalten des trivariaten Kontrasttests für drei Stichproben und konkave Dosis-Wirkungs-Profile dargestellt. Tabelle 6.2 und Tabelle 6.3 verdeutlichen zum einen den deutlichen Powergewinn, wenn der Stichprobenumfang je Gruppe von drei auf fünf erhöht wird. Zum anderen wird deutlich, daß mit steigendem Stichprobenumfang die Unterschiede zwischen den approximativen Permutationstests sowie auch die Unterschiede zum exakten Permutationstest kleiner werden.

$(\vartheta_0, \vartheta_1, \vartheta_2)$	*TKnpbom-oV*	*TKnppea-oV +Rcount2*	*TKnppea-oV +Rcount*	*TKnppea-oV +Permut*	*TKnppex-oV*	*TKnppmp-oV*
(0,25, 0,25, 0,25)	0,052	0,039	0,035	0,037	0,029	0,045
(0,25, 0,3, 0,3)	0,070	0,053	0,051	0,050	0,040	0,062
(0,25, 0,35, 0,35)	0,105	0,083	0,079	0,078	0,066	0,092
(0,25, 0,4, 0,4)	0,147	0,112	0,102	0,102	0,085	0,137
(0,25, 0,45, 0,45)	0,204	0,157	0,145	0,144	0,122	0,181
(0,25, 0,5, 0,5)	0,277	0,210	0,190	0,194	0,163	0,242
(0,25, 0,55, 0,55)	0,347	0,266	0,240	0,244	0,205	0,294

Tabelle 6.2: Güte des TK-Tests für binomialverteilte Zufallszahlen bei unterschiedlichen Verteilungsapproximationen und den Parametern: $\alpha = 0{,}05$, $k = 2$, $m = 3$, $n = 3$

$(\vartheta_0, \vartheta_1, \vartheta_2)$	TKnpbom-oV	TKnppea-oV +Rcount2	TKnppea-oV +Rcount	TKnppea-oV +Permut	TKnppex-oV	TKnppmp-oV
(0,25, 0,25, 0,25)	0,058	0,040	0,043	0,044	0,037	0,052
(0,25, 0,3, 0,3)	0,094	0,074	0,075	0,076	0,071	0,089
(0,25, 0,35, 0,35)	0,141	0,121	0,121	0,124	0,114	0,132
(0,25, 0,4, 0,4)	0,211	0,195	0,192	0,195	0,180	0,206
(0,25, 0,45, 0,45)	0,304	0,282	0,278	0,283	0,271	0,291
(0,25, 0,5, 0,5)	0,401	0,387	0,380	0,385	0,366	0,390
(0,25, 0,55, 0,55)	0,510	0,491	0,487	0,493	0,471	0,499

Tabelle 6.3: Güte des TK-Tests für binomialverteilte Zufallszahlen bei unterschiedlichen Verteilungsapproximationen und den Parametern: $\alpha = 0{,}05$, $k = 2$, $m = 3$, $n = 5$

Weitere diskrete Verteilungen wurden durch Runden generiert. Dazu wurden stetige Zufallszahlen erzeugt und anschließend auf die nächstnaheliegende ganze Zahl gerundet. Zum Beispiel wurden mittels normalverteilter Zufallszahlen symmetrische Verteilungen erzeugt:

$$X_{ij} = \lfloor Z_{ij} + 0{,}5 \rfloor \text{ mit } Z_{ij} \sim N(\mu_i, \sigma_i^2).$$

Zur Erzeugung schiefer Verteilungen wurden exponentialverteilte Zufallszahlen und Zufallszahlen, die mit Hilfe des Fleishman-Systems [174] erzeugt wurden, genutzt. Diese Art der Erzeugung ist besonders zur Modellierung einer diskretisierten stetigen Zufallsvariablen geeignet. Eigenschaften, wie z. B. die Symmetrie bzw. Schiefe der Ursprungsverteilung, übertragen sich dabei zum Teil auf die so diskretisierte Verteilungsfunktion (d. h., auch sie ist symmetrisch bzw. schief). Allerdings kann nur abgeschätzt werden, auf wieviele Punkte die erzeugte Verteilung konzentriert ist. Diese Zufallszahlen wurden zur Untersuchung der vorgestellten parametrischen und nichtparametrischen Tests sowie verschiedener Bootstrapverfahren genutzt. Approximative Permutationstests wurden mittels des Algorithmus „Permut“ ebenfalls simuliert. Sowohl die in den Abschnitten 4.2 bzw. 5.2 definierten unterschiedlichen Varianzschätzer als auch unterschiedliche Resamplingräume

$$RS = \{X_{ij}, i = 0,\ldots,k, j = 1,\ldots,n_i\} \qquad RS(R) = \{R_{ij}, i = 0,\ldots,k, j = 1,\ldots,n_i\},$$

$$RS_Z = \{X_{ij} - \overline{X}_{i.}, i = 0,\ldots,k, j = 1,\ldots,n_i\} \quad RS_Z(R) = \{R_{ij} - \overline{R}_{i.}, i = 0,\ldots,k, j = 1,\ldots,n_i\},$$

$$RS_{Zi} = \{X_{ij} - \overline{X}_{i.}, j = 1,\ldots,n_i\} \;\; i = 0,\ldots,k \quad RS_{Zi}(R) = \{R_{ij} - \overline{R}_{i.}, j = 1,\ldots,n_i\} \;\; i = 0,\ldots,k$$

wurden betrachtet. Außerdem wurden äquidistante Scores und Durchschnittsränge (auch Zufallsränge wurden untersucht) als Grundlage für die Tests gewählt. Aufgrund des hohen Rechenaufwandes wurde sich auf nur wenige Punkte sowohl unter der Null- als auch unter der Alternativhypothese beschränkt. Untersucht wurden allerdings mehrere Stichprobenumfänge ($n = 3,4,5$), mehrere Verteilungen (schiefe und symmetrische), mehrere Dosis-Wirkungs-Profile (lineare, konvexe, konkave), mehrere Anzahlen von Gruppen ($k = 1,2,3$) und mehrere Signifikanzniveaus ($\alpha = 0{,}01, 0{,}05, 0{,}1$).
Die wesentlichen Ergebnisse dieser Studien sind:

1. Das Ziehen aus den einzelnen Stichproben ($RS_{Zi}, RS_{Zi}(R)$) führt zu Bootstraptests, die das Niveau kaum überschreiten. Die Güte dieser Tests unter der Alternative ist jedoch gering. Diese Methode kann daher nicht empfohlen werden.
2. Mit steigender Anzahl von Stichproben nähern sich die Güten der asymptotischen Tests und der Bootstraptests (ohne Varianzschätzer) relativ schnell an. Bei $k = 3$ tritt kaum noch ein relevanter Unterschied (größer als 2 Hundertstel) zwischen diesen Verfahren auf. Bei $k = 1,2$ hingegen besitzen die Bootstraptests (besonders deutlich bei $k = 2$) gegenüber ihren analogen asymptotischen Tests (d. h. jeweils gleiche Teststatistik) Gütevorteile.

3. Generell sind die Bootstraptests mit einer höheren Güte verbunden als die approximativen Permutationstests (siehe z. B. Abbildung 6.1). Zwar werden die Unterschiede mit wachsendem k und wachsendem n kleiner, bei $k = 3$ und $n = 5$ zeigen sich jedoch in der Regel unter der Alternative noch Gütevorteile von 2 bis 3 Hundertstel. Auch aus diesem Grund sollten approximative Permutationstests bei den betrachteten Fallzahlen nicht benutzt werden.
4. Der auf dem Hoggschen Prinzip aufbauende permutative adaptive Test T_{ad5}^{hogg} (siehe Abschnitt 5.3) erweist sich zwar dem permutativen Test T_{ad5}^{max} (siehe Abschnitt 5.3) gegenüber als überlegen, jedoch war ein deutlicher Gütegewinn gegenüber einem nur auf Durchschnittsrängen beruhendem multiplen Kontrasttest *TKnpnvV2* selten. Erst ab $n = 10$ oder für vereinzelte Profile erweist sich der auf dem Hoggschen Prinzip aufbauende permutative bzw. der asymptotische Test als sinnvoll. Bei den analogen Bootstraptests treten ähnliche Probleme bei der Bestimmung der Selektionsstatistiken auf wie beim Schätzen der Varianz (Division durch Null). Für die Fallzahlen $n = 3,4,5$ werden diese Verfahren daher nicht empfohlen.
5. Bootstraptests auf der Basis von Zufallsrängen (um die Probleme mit nichtpositiven Varianzschätzern zu verringern, wurden nach jedem Ziehen die Zufallsränge statt der Durchschnittsränge neu vergeben) führen zu Tests mit einer geringen Güte. Diese Variante kann nicht empfohlen werden.
6. Von den vorgestellten Varianzschätzern erweisen sich S_{II}^2 bzw. $\hat{\sigma}_{II}^2$ in Verbindung mit Bootstraptests (ziehen aus RS_Z) als am geeignetsten. Bootstraptests ohne Varianzschätzer (ziehen aus RS) zeigen sich allerdings in den meisten Fällen als mächtiger, neigen aber eher zu Niveauverletzungen. Ziehen aus der gepoolten, jeweils mit den Gruppenmittelwerten zentrierten Stichprobe (RS_Z) (1. Empfehlung von Hall und Wilson [98]), führt bei Nutzung von Statistiken ohne Varianzschätzer in den meisten Fällen zu deutlichen Niveauüberschreitungen.
7. Zwischen den Bootstraptests auf der Grundlage von multiplen Kontrasten und der Bartholomew- oder Chacko-Statistik sind nur relativ kleine Vorteile zugunsten letzterer zu entdecken. Dies ist unabhängig davon, ob ein Varianzschätzer benutzt wird oder nicht.

Weitere Schlußfolgerungen aus diesen Simulationen fließen in die im folgenden ausführlicher diskutierten Ergebnisse ein.

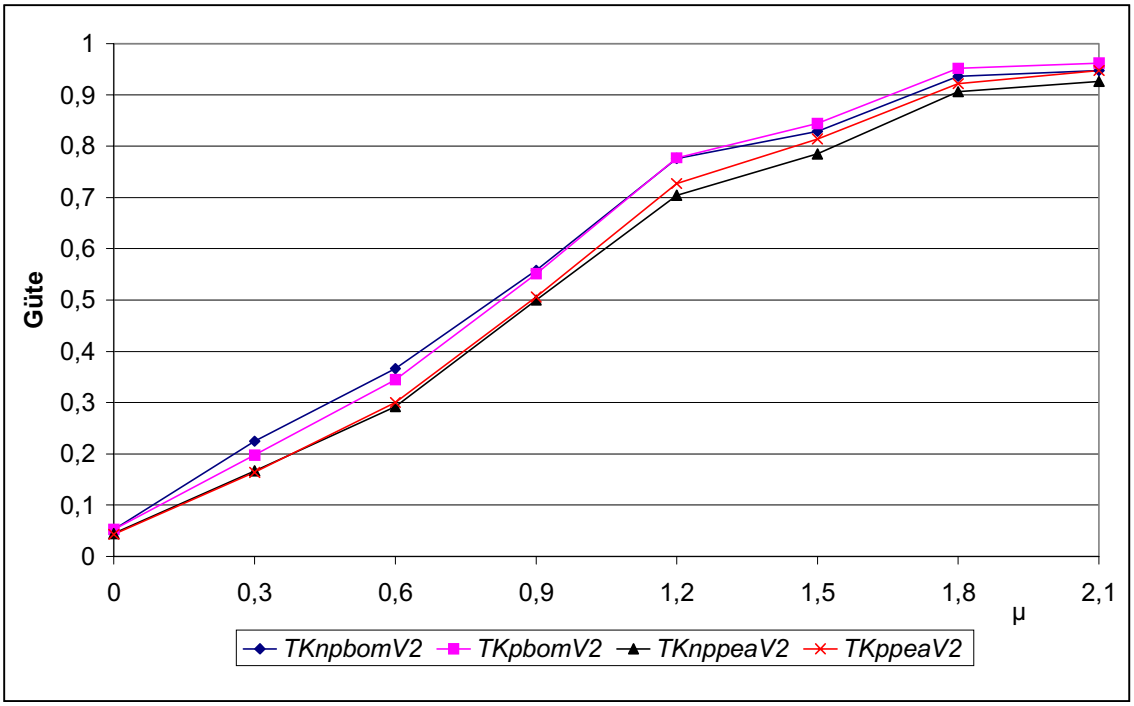


Abbildung 6.1: Vergleich Bootstrapverteilung versus Permutationsverteilung anhand des *TK*-Tests ($\alpha = 0{,}05$, $k = 3$, $n = 4$, diskretisierte exponentialverteilte Zufallszahlen, konkave Erwartungswertprofile, μ = Lageparameter der Exponentialverteilung)

Die Zufallszahlen für die weiteren Simulationen wurden mittels der Festlegung der Wahrscheinlichkeiten $\pi_{is} = P(X_{ij} = K_s), (s = 1,\ldots,r)$, auf dem Intervall $(0, 1)$, gleichverteilter Zufallszahlen (U) und mittels der Vorschrift

$$X_{ij} = s \Leftrightarrow \pi_{i1} + \cdots + \pi_{is-1} \leq U < \pi_{i1} + \cdots + \pi_{is} \; (s = 1,\ldots,r, \pi_{i0} = 0)$$

generiert. Der Einfachheit halber wurde zur Modellierung der diskreten Verteilungen analog zu Hilton [175] die logistische Verteilungsfunktion genutzt. Durch Variierung von Parametern können sowohl unter der Nullhypothese als auch unter der Alternative (stochastische Ordnung der Verteilungsfamilien) Verteilungen leicht definiert werden. Die Zellwahrscheinlichkeiten werden dazu wie bei einem ordinalen Regressionsmodell mittels der stetigen Verteilungsfunktion definiert:

$$\pi_{is} = \begin{cases} \frac{1}{1+\exp(-(\tau_s - \varsigma_i))} & falls \;\; s = 1 \\ \frac{1}{1+\exp(-(\tau_s - \varsigma_i))} - \frac{1}{1+\exp(-(\tau_{s-1} - \varsigma_i))} & falls \; 2 \leq s \leq r-1 \\ 1 - \frac{1}{1+\exp(-(\tau_{s-1} - \varsigma_i))} & falls \; s = r \end{cases} .$$

Die Parametervektoren τ und ξ wurden zur Erzeugung unterschiedlicher Verteilungsformen und unterschiedlicher Dosis-Wirkungs-Profil benutzt. Schiefe Verteilungen können durch folgende Parameterwahl erzeugt werden:

a) linksschiefe Verteilungen (Bez.: LSV): $\tau_s = -log(\frac{r-s}{s}) - \omega$, $\omega > 0$, $s = 1,\ldots,r-1$,

b) rechtsschiefe Verteilungen (Bez.: RSV): $\tau_s = -log(\frac{r-s}{s}) + \omega$, $\omega > 0$, $s = 1,\ldots,r-1$.

In Abhängigkeit davon, ob *r* eine gerade oder ungerade Zahl ist, wurden symmetrische Verteilungen (Bez.: SV) erzeugt, indem z. B. für ungerades *r* und $t = \frac{r-1}{2}$

$$\tau_s = \begin{cases} -log(\frac{r-s}{s}) - \omega & falls\ \ 1 \le s \le t \\ -log(\frac{r-s-t}{s+t}) + \omega & falls\ \ t < s \le r-1 \end{cases}$$

gewählt wurde. Zusätzlich wurden U-förmige Verteilungen (Bez.: UV; ein Beispiel für die Relevanz diskreter U-förmiger Verteilungsdichten beschreiben Lesaffre et al.[176]) mittels

$$\tau_s = \begin{cases} -\upsilon \log(\frac{r-s}{s}) & falls\ \ 1 \le s \le t \\ -\upsilon \log(\frac{r-s-t}{s+t}) & falls\ \ t < s \le r-1 \end{cases} \qquad 0 < \upsilon < 1$$

generiert. Unterschiedliche konkave, konvexe oder „fast lineare“ Dosis-Wirkungs-Profile wurden erzeugt durch:

A) konkave Profile: $$\xi_i = \begin{cases} -\frac{1}{2}\theta & falls\ \ 1 \le i \le k \\ 0 & falls\ \ i = 0 \end{cases}$$

B) konvexe Profile: $$\xi_i = \begin{cases} -\frac{1}{2}\theta & falls\ \ i = k \\ 0 & falls\ \ 0 \le i \le k-1 \end{cases}$$

C) lineare Profile: $$\xi_i = \begin{cases} -\frac{1}{10} i\,\theta & falls\ \ 1 \le i \le k \\ 0 & falls\ \ i = 0 \end{cases} .$$

In Abhängigkeit von den untersuchten Tests (asymptotisch oder Resampling) wurden für die Parameter ω bzw. υ Werte aus den Mengen $\{0,\ 0{,}5,\ 1{,}0,\ 1{,}5,\ 2,\ 0{,}25\}$ bzw. $\{0{,}875,\ 0{,}75,\ 0{,}625,\ 0{,}5,\ 0{,}375\}$ gewählt. Eine Erhöhung der Parameter ω bzw. υ wirkt sich jeweils in einer stärkeren Konzentration auf wenige der r Punkte aus. Die diskrete Gleichverteilung ($\pi_{i1} = \cdots = \pi_{ir} = \frac{1}{r}$, Bez.: GV) ergibt sich z. B. für $\omega = 0$ bei linksschiefen Verteilungen. Der Parameter θ wurde genutzt, um Zufallszahlen für die Nullhypothese ($\theta = 0$) bzw. für die Alternativhypothese ($\theta \in \{2,4,5,6,8,10,15\}$) zu erzeugen.

Die Tests von Roth [116] aus Abschnitt 4.1 erweisen sich als extrem ungeeignet. Bei den Originaltests, d. h., mit den in Abschnitt 4.1 beschriebenen parametrischen Verteilungen, treten enorme Probleme bei der Bestimmung der Levelwahrscheinlichkeiten auf (bei beiden Tests müssen die Gewichte $w_i = n_i / s_i^2$ genutzt werden (siehe Roth [116])). Bei den Bootstrapvarianten erweisen sich die Varianzschätzer wie in Abschnitt 4.1 vermutet als ungeeignet (zu oft Schätzer mit dem Wert Null). Das Güteverhalten von Bootstraptests auf der Basis von Kontrasten und des Varianzschätzers S_{III}^2 (äquidistante Scores) bzw. $\hat{\sigma}_{III}^2$ (Ränge) ist ähnlich schlecht. Eine Normalverteilungsapproximation für die von Grimes und Federer [117] vorgeschlagenen Kontrasttests (Varianzschätzer S_{III}^2) bzw. für die analogen Rangstatistiken (Varianzschätzer $\hat{\sigma}_{III}^2$) führt ebenfalls bei den hier betrachteten Problemstellungen zu ungeeigneten Tests. Sowohl die einfachen Kontrasttests ($\hat{\alpha} > 0{,}07$ bei $\alpha = 0{,}05$) als auch die multiplen Kontrasttests ($\hat{\alpha} > 0{,}09$ bei $\alpha = 0{,}05$) sind in der Regel sehr liberal. In Abbildung 6.3 ist für den linearen Kontrast der Einfluß der Varianzschätzer S_{III}^2 bzw. $\hat{\sigma}_{III}^2$ im Vergleich zu den Varianzschätzern S_{II}^2 bzw. $\hat{\sigma}_{II}^2$ dargestellt. Zusätzlich tritt in Fällen, bei denen die Varianz in den Gruppen gegen Null tendiert, ähnlich wie bei den von Roth vorgeschlagenen Tests (siehe Abbildung 6.2), ein Güteabfall trotz steigendem θ auf. Werden statt der Normalverteilungsapproximation die von Grimes und Federer vorgeschlagenen Welch- bzw. die Cochran-Approximationen oder die von Akritas und Brunner vorgeschlagene Welch-Approximation [48] genutzt, verbessert sich das Verhalten der auf S_{III}^2 bzw. $\hat{\sigma}_{III}^2$ basierenden einfachen Kontrasttests etwas. Die Cochran-Approximation führt allerdings zu konservativen Tests mit niedriger Power. Die Welch-Approximation hingegen führt abhängig von der Form des Kontrastes zum Teil zu geeigneten aber auch zu liberalen Tests (sowohl bei Rängen als auch bei äquidistanten Scores). In Abbildung 6.4 ist beispielhaft das Verhalten des Linearen Kontrasttests dargestellt.

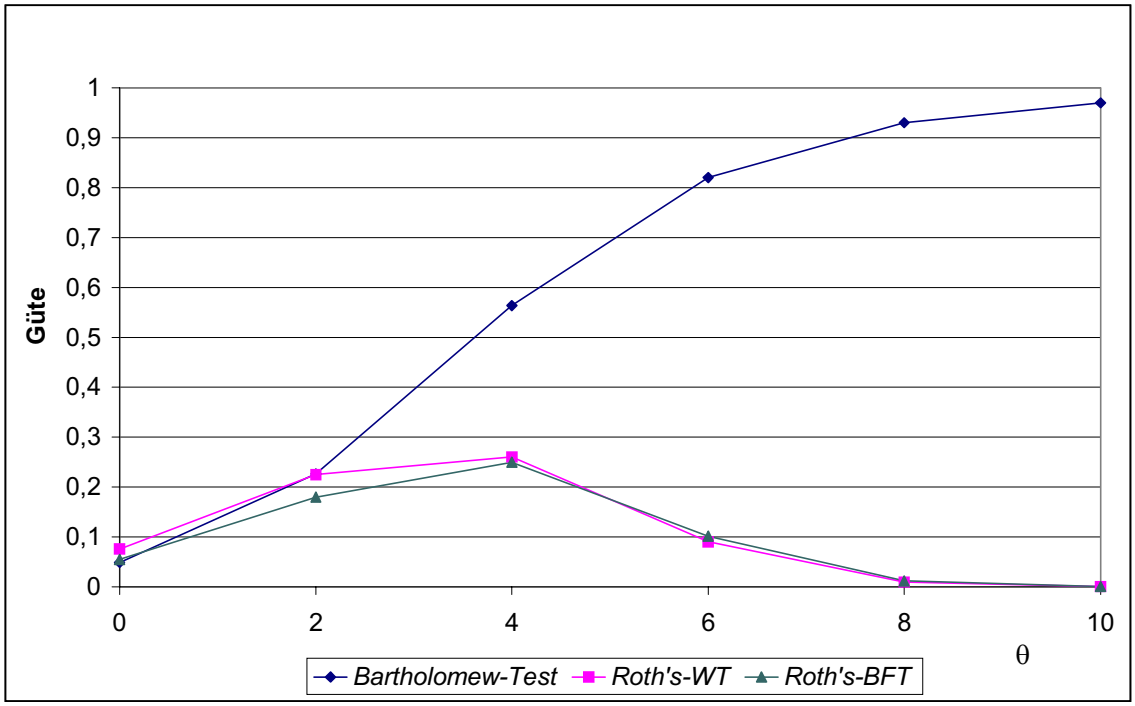


Abbildung 6.2: Vergleich des Bartholomew-Tests und der Tests von Roth (jeweils parametrische Verteilung; $\alpha = 0{,}05$, $k = 3$, $n = 5$, $r = 5$, $\omega = 0$ und konkave Dosis-Wirkungs-Profile)

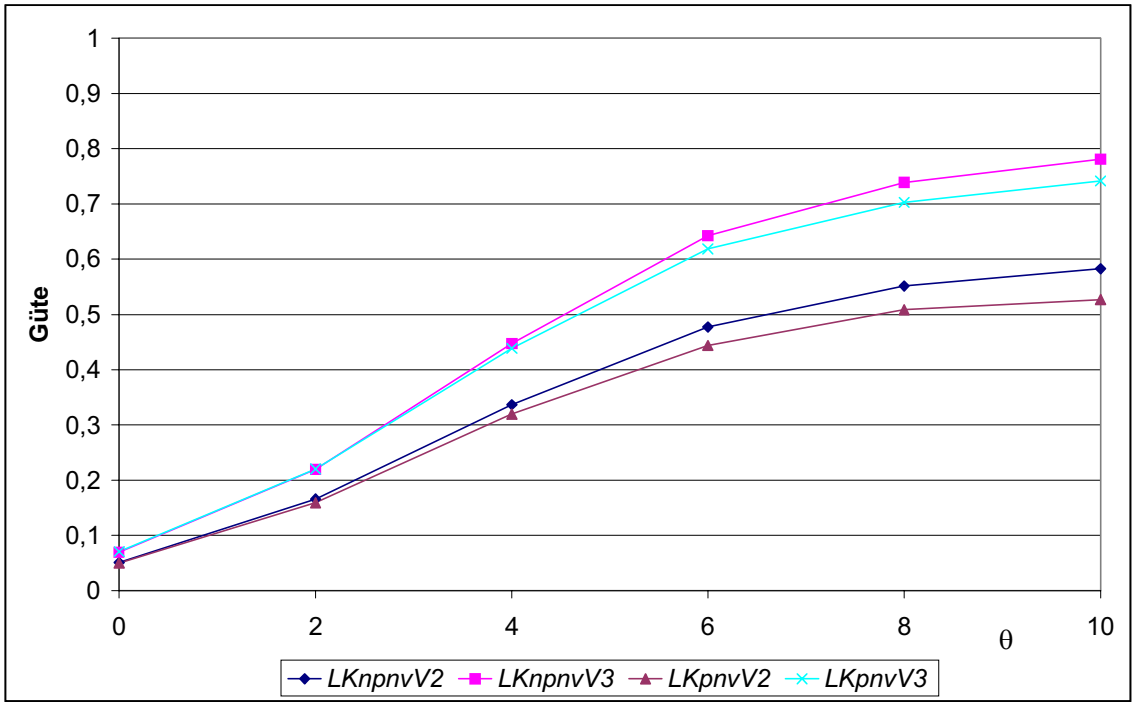


Abbildung 6.3: Güte des Linearen Kontrasttests bei Normalverteilungsapproximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen)

Wird statt des Linearen Kontrastes der Helmert-Kontrast gewählt (Abbildung 6.5), so bleibt die Konservativität der Cochran-Approximation erhalten. Die beiden Welch-Approximationen hingegen überschreiten unter denselben Voraussetzungen und bei Nutzung des Helmert-Kontrastes das Signifikanzniveau deutlich. Im Vergleich dazu zeigt sich das Verhalten der auf den Varianzschätzern S_{II}^2 bzw. $\hat{\sigma}_{II}^2$ basierenden einfachen Kontrasttests wesentlich unabhängiger von der Wahl des Kontrastes. Im Zweistichprobenfall erweisen sich sowohl die Cochran- als auch die Welch-Approximation (basierend auf äquidistanten Scores) als sehr konservativ. Zudem führen sie zu geringer Güte unter der Alternativhypothese. Die Welch-Approximation in Verbindung mit Rängen ist zwar deutlich besser, insgesamt gesehen jedoch ist der auf dieser Variante beruhende Test den anderen Tests (siehe unten) unterlegen. Da aufgrund fehlender Verteilungsaussagen weder für die Cochran- noch für die Welch-Approximation multiple Kontraste definiert werden können, werden diese Approximationen nicht für den Einsatz in der Praxis empfohlen. Einfache oder multiple Kontrasttests verbunden mit Varianzschätzern, die auf S_I^2 (bei äquidistanten Scores) bzw. $\hat{\sigma}_I^2$ (bei Rängen) beruhen, führen bei Normalverteilungsapproximation ebenfalls zu liberalen Tests. Die Niveauüberschreitungen liegen bei allen untersuchten Niveaus meist zwischen 20% und 30%, d. h., z. B. $0{,}06 \le \hat{\alpha} \le 0{,}065$ bei $\alpha = 0{,}05$. Der Schätzer $\hat{\sigma}_I^2$ ist auch in Verbindung mit dem Chacko-Test aufgrund enormer Liberalität nicht geeignet. Im Gegensatz zur Normalverteilungsapproximation kann die t-Verteilungsapproximation in Verbindung mit S_I^2 bzw. $\hat{\sigma}_I^2$ sowohl für $k = 1$ als auch für $k = 2,3$ bedingt empfohlen werden (siehe Abbildung 6.6). Die Bootstraptests, gekoppelt mit den Schätzern S_I^2 bzw. $\hat{\sigma}_I^2$, erweisen sich zwar nicht als liberal, haben jedoch in den meisten Fällen eine geringere Güte als die Bootstraptests auf der Basis von S_{II}^2 bzw. $\hat{\sigma}_{II}^2$.

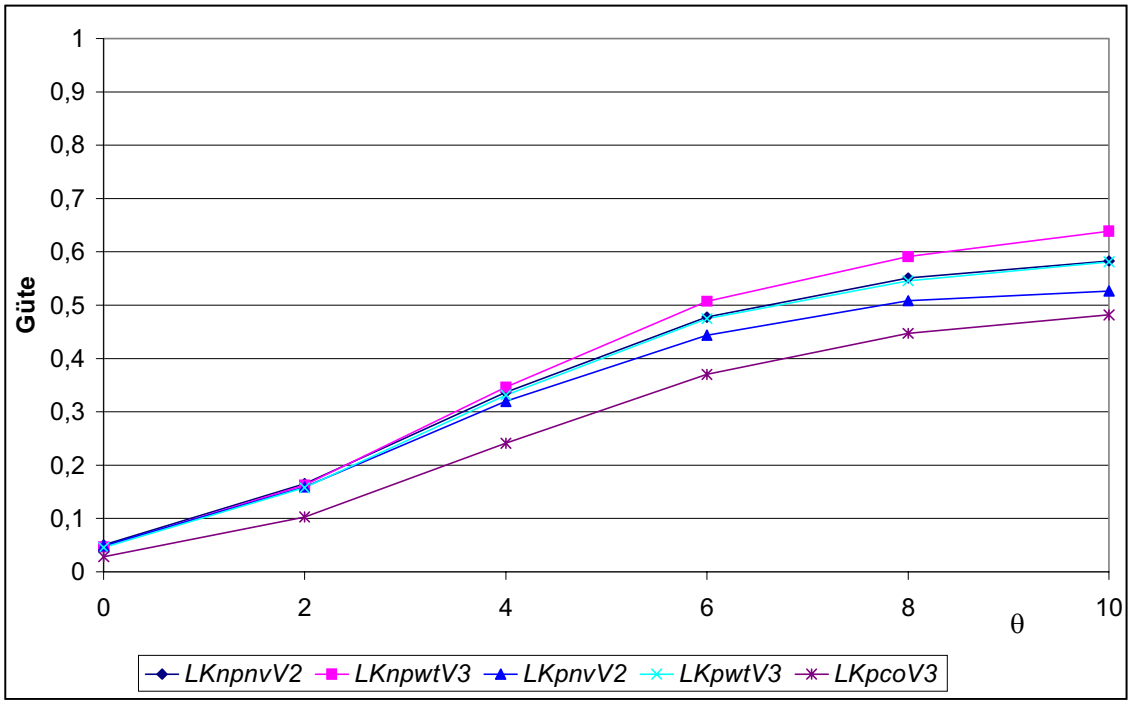


Abbildung 6.4: Güte des Linearen Kontrasttests bei Normalverteilungs-, Cochran- oder Welch-Approximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen)

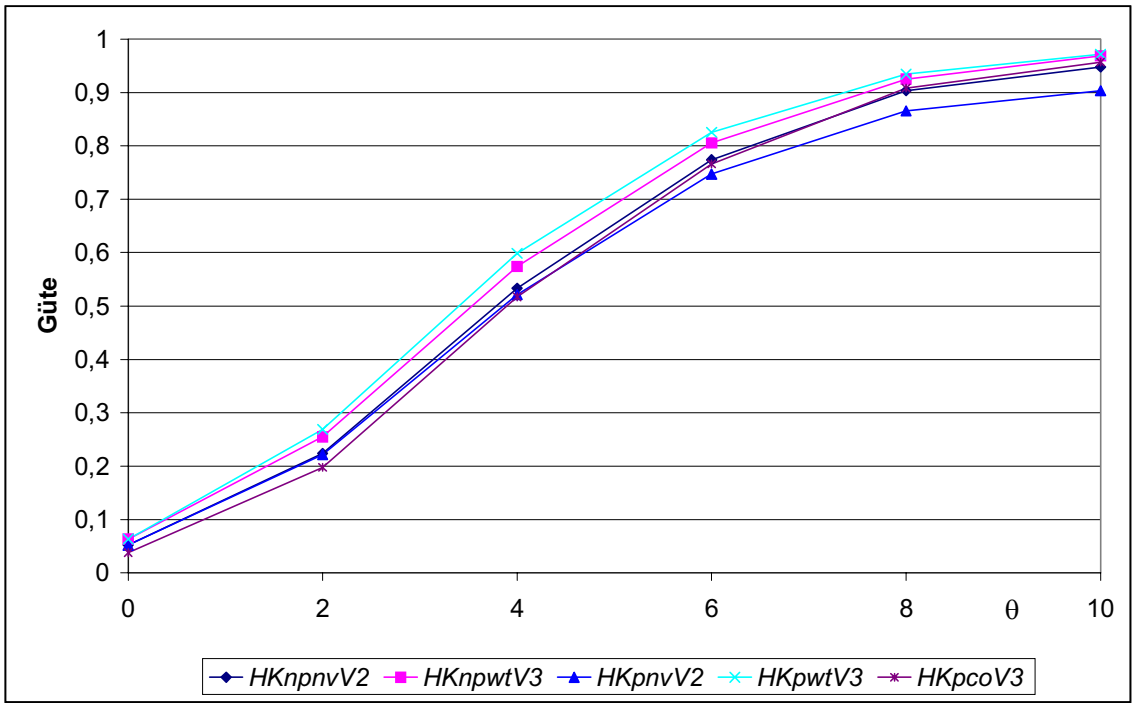


Abbildung 6.5: Güte des Helmert-Kontrasttests bei Normalverteilungs-, Cochran- oder Welch-Approximation und den Varianzschätzern $V2 \in \{S_{II}^2, \hat{\sigma}_{II}^2\}$ bzw. $V3 \in \{S_{III}^2, \hat{\sigma}_{III}^2\}$ ($\alpha = 0{,}05$, $k = 3$, $n = 4$, $r = 4$, $\omega = 0$ und konvexen Dosis-Wirkungs-Profilen)

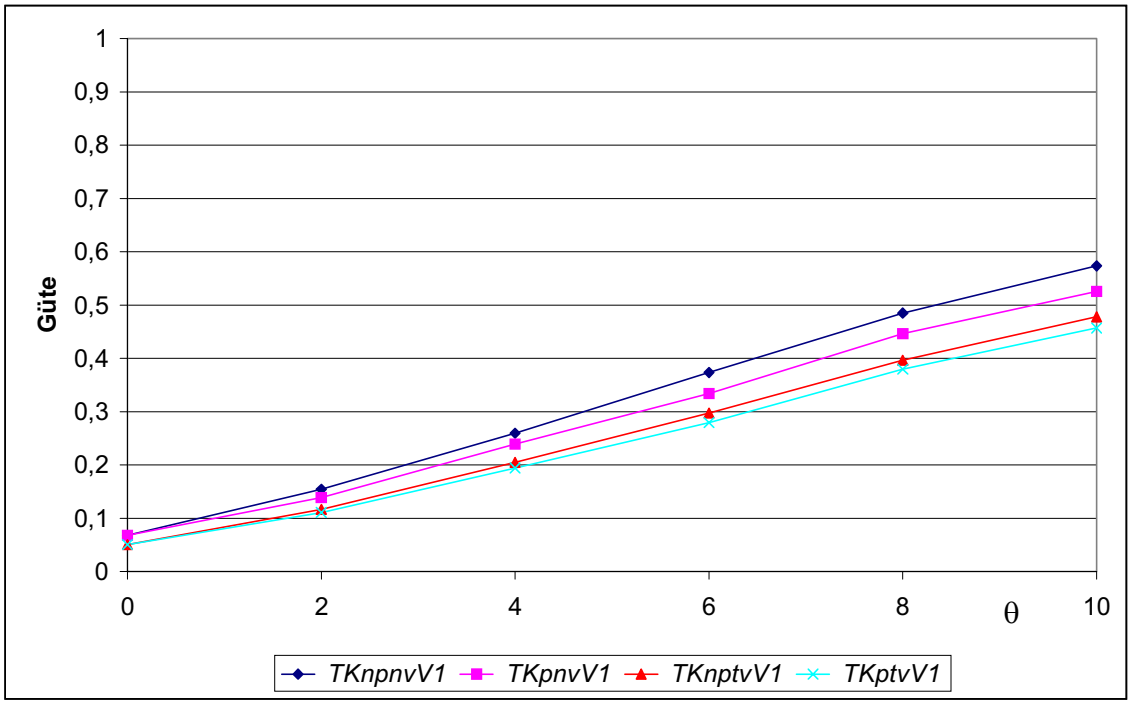


Abbildung 6.6: Güte des TK-Kontrastes bei Normalverteilungs- bzw. t-Verteilungsapproximation und den auf S_I^2 bzw. $\hat{\sigma}_I^2$ basierenden Varianzschätzern ($\alpha = 0{,}05$, $k = 3$, $n = 5$, $r = 4$, $\omega = 1$ und linearen Dosis-Wirkungs-Profilen)

Fazit:

1. Die Modellierung von Varianzheterogenitäten und den damit verbundenen Varianzschätzern führt bei den betrachteten Fallzahlen zu ungeeigneten Tests. Dies betrifft Zwei-, Drei- und Vierstichprobentests, die auf parametrischen Verteilungen bzw. Bootstrapverteilungen basieren.
2. Tests, die Varianzheterogenitäten ignorieren (Varianzschätzer auf der Basis von S_I^2 und S_{II}^2 bzw. $\hat{\sigma}_I^2$ und $\hat{\sigma}_{II}^2$) sind sensitiv für die interessierenden Hypothesen. Die Wahrscheinlichkeit für einen Fehler 1. Art ist im wesentlichen durch das vorgegebene Signifikanzniveau beschränkt. Wird als Varianzschätzer S_I^2 bzw. $\hat{\sigma}_I^2$ benutzt, so ist nur eine t-Verteilungsapproximation relevant. Mit den Varianzschätzern S_{II}^2 bzw. $\hat{\sigma}_{II}^2$ können sowohl Bootstraptests als auch auf einer Normalverteilungsapproximation basierende parametrische (asymptotische) Tests konstruiert werden (t-Verteilung führt zu konservativen Tests mit geringer Power; Bootstraptests sind aufgrund des Varianzschätzers zu anfällig).

6.3 Geeignete Tests

Aufgrund der Aussagen des Abschnitts 6.2 werden im Abschnitt 6.3 Simulationsergebnisse für Tests, die sich als geeignet erwiesen haben, ausführlicher vorgestellt. Die dargestellten Ergebnisse basieren auf den mittels der logistischen Verteilungsfunktion generierten Zufallszahlen.

6.3.2 Zweistichprobentests

Nimmt r die Werte $3,4,5,6,7,8,9$, n die Werte $3,4,5$, α die Werte $0{,}01$, $0{,}05$, ς die Werte $0{,}875$, $0{,}75$, $0{,}625$, $0{,}5$, $0{,}375$, ω die Werte $0{,}5$, $1{,}0$, $1{,}5$, $2{,}0$, $2{,}5$, θ die Werte 0, 5, 10, 15 an und werden zusätzlich die vier oben beschriebenen Verteilungstypen betrachtet, so ergeben sich $7 \times 3 \times 2 \times 5 \times 4 \times 4 + 7 \times 3 \times 2 \times 4 = 3528$ verschiedene Parameterkombinationen. Mit einem Viertel dieser Möglichkeiten ($\theta = 0$) kann das Güteverhalten der Tests unter der Nullhypothese geprüft werden. Die Ergebnisse werden hier für die in Tabelle 6.4 aufgeführten Tests dargestellt. Da bei Permutationstests die Varianzschätzer auf der Basis von $V1$ bzw. $V2$ keinen Einfluß auf die Tests haben, wurde auch beim „exakten“ Bootstraptest und beim unbedingten Test auf Varianzschätzer verzichtet (die Wahrscheinlichkeit für einen nichtpositiven Varianzschätzer ist zudem sehr hoch, was zu sehr konservativen Tests mit geringer Power führt). Ein Vergleich läßt daher eher Rückschlüsse auf die genutzten Verteilungen zu. Der Double-Bootstrap erwies sich erwartungsgemäß als sehr zeitintensiv, so daß nur wenige Simulationen für die auf Rängen bzw. äquidistanten Scores basierenden Mittelwertdifferenzen durchgeführt wurden. In diesen Simulationen zeigten sich kaum verallgemeinerbare Ergebnisse. In den meisten Fällen hatte der Double-Bootstraptest (innere und äußere Schleife mit jeweils 1.000 Resamplingstichproben) zwar sowohl unter der Null- als auch unter Alternativhypothese eine höhere Güte als der einfache Monte-Carlo-Bootstraptest, es traten jedoch zum Teil auch deutliche Niveauüberschreitungen auf. Eine intensivere Untersuchung von Double-Bootstraptests scheint zwar angezeigt, aufgrund des enormen Rechenaufwandes und des zum Teil geringen Gütegewinns (oft nicht mehr als 0,02 bis 0,03) wurde jedoch darauf verzichtet.

Statistik	Verteilung	Varianz-schätzer (V)	Bezeichner
$\frac{\overline{X}_{1.} - \overline{X}_{0.}}{\sqrt{V}}$ äqui-distante Scores	t-Verteilung aus Satz 4.4	*V1*	*tTptvV1*
	Normalverteilung aus Satz 4.5	*V2*	*tTpnvV2*
	Bootstrapverteilung;Resamplingraum *RS*	*oV*	*tTpbomoV*
	Bootstrapverteilung;Resamplingraum RS_Z	*V2*	*tTpbomV2*
	„exakte“ Bootstrapverteilung *RS*	*oV*	*tTpbovoV*
	exakte bedingte Permutationverteilung	*oV*	*tTppexoV*
	Mid-p-Test	*oV*	*tTppmpoV*
	randomisierter Permutationstest	*oV*	*tTpperoV*
	unbedingter Permutationstest	*oV*	*tTppeuoV*
$\frac{\overline{R}_{1.} - \overline{R}_{0.}}{\sqrt{V}}$ Ränge	t-Verteilung aus Satz 4.4	*V1*	*tTnptvV1*
	Normalverteilung aus Satz 4.5	*V2*	*tTnpnvV2*
	Bootstrapverteilung;Resamplingraum *RS*	*oV*	*tTnpbomoV*
	Bootstrapverteilung;Resamplingraum RS_Z	*V2*	*tTnpbomV2*
	„exakte“ Bootstrapverteilung *RS*	*oV*	*tTnpbovoV*
	exakte bedingte Permutationverteilung	*oV*	*tTnppexoV*
	Mid-p-Test	*oV*	*tTnppmpoV*
	randomisierter Permutationstest	*oV*	*tTnpperoV*
	unbedingter Permutationstest	*oV*	*tTnppeuoV*

Tabelle 6.4: Bezeichner für die im Abschnitt 6.3.1 beschriebenen Zweistichprobentests

Im balancierten Fall besitzen die Varianzschätzer folgende Form:

$$V_1 = \frac{1}{n(n-1)} \sum_{i=0}^{1} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{i.})^2 \ , \ V_2 = \frac{2}{n(2n-1)} \sum_{i=0}^{1} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{..})^2 \ \text{ bzw.}$$

$$V_1 = \frac{1}{n(n-1)} \sum_{i=0}^{1} \sum_{j=1}^{n} (R_{ij} - \overline{R}_{i.})^2 \ , \ V_2 = \frac{2}{n(2n-1)} \sum_{i=0}^{1} \sum_{j=1}^{n} (R_{ij} - \overline{R}_{..})^2 \ .$$

Der *tTptvV1*-Test ist der parametrische t-Test und der *tTnpnvV2*-Test ist der asymptotische Wilcoxon-Mann-Whitney-Test[164; 177].

6.3.2.1 Güte unter der Nullhypothese

Auch wenn die Tests, die im folgenden nur noch betrachtet werden, als geeignet angesehen werden, treten auch bei ihnen wesentliche Niveauüberschreitungen auf. Diese treten jedoch deutlich seltener auf als bei den bereits selektierten Tests und können zum Teil speziellen Konstellationen zugeordnet werden. In Abbildung 6.7. bis 6.15 ist beispielhaft das Güteverhalten aller bzw. ausgewählter Tests dargestellt. Die extrem konservativen exakten bedingten Permutationstests sind die einzigen Tests, bei denen keine Niveauüberschreitungen auftreten. Die randomisierten Permutationstests schöpfen hingegen das vorgegebene Signifikanzniveau am besten aus (theoretisch wird das Signifikanzniveau durch diese Tests stets voll ausgeschöpft und nicht überschritten; praktisch treten in den Simulationen wenige Niveauverletzungen aufgrund des wiederholten Zufallsexperiments auf). Etwas erstaunlich ist das Verhalten der Bootstraptests mit Varianzschätzer. Sie sind ähnlich konservativ wie die bedingten Permutationstests. Da dies auch bei $n = 5$ und $r = 7$ auftritt, kann dies nicht nur dem Auftreten nichtpositiver Varianzschätzer zugeordnet werden. Hier wirkt sich auch schon die geringe Variabilität in den Resamplingstichproben negativ aus (kleine Varianzschätzer führen zu großen Werten der Statistik; das Zentrieren kann zu einer Verringerung der Gesamtvariabilität in der gepoolten Stichprobe führen).
Die wesentlichen Niveauüberschreitungen der Tests sind besonders vom Stichprobenumfang und vom Signifikanzniveau abhängig. Vor allem beim Signifikanzniveau $\alpha = 0{,}01$ treten Niveauüberschreitungen auf. Der *tTptvV1*-Test z. B. überschreitet bei fast jeder vierten Parameterkonstellation das Niveau $\alpha = 0{,}01$ deutlich (und zwar bei allen Werten für *r*). Die Monte-Carlo-Bootstraptests ohne Varianzschätzer (vor allem bei den auf Rängen basierenden Tests) führen ebenfalls zu zu vielen Niveauüberschreitungen. Die „exakten“ Bootstraptests überschreiten vor allem beim Niveau $\alpha = 0{,}05$ das Signifikanzniveau zu oft. Ähnlich sieht es beim *tTnpnvV2*-Test aus. Bei $\alpha = 0{,}01$ ist dieser Test eher konservativ, bei $\alpha = 0{,}05$ (vor allem bei $n = 3$) hingegen treten deutlich mehr Niveauüberschreitungen als z. B. beim *tTptvV1*-Test auf. Die Mid-p-Tests sind bei $\alpha = 0{,}01$ eher konservativ und liegen bei $\alpha = 0{,}05$ meist unter $0{,}05$ (Niveauverletzungen treten so gut wie keine auf). Die unbedingten Permutationstests neigen eher bei großen Werten für *r* zu Niveauüberschreitungen. Die t-Verteilungsapproximation in Verbindung mit Rängen führt, ähnlich wie bei den äquidistanten Scores, vor allem bei $\alpha = 0{,}01$ zu deutlichen Niveauüberschreitungen. Aufgrund der Ergebnisse aller durchgeführten Simulationen kann zusammenfassend gesagt werden:

1. Die Bootstraptests ohne Varianzschätzer und der *tTnpnvV2*-Test führen bei den betrachteten Konstellationen für r und n und $\alpha = 0{,}05$ zu häufig zu Niveauüberschreitungen. Sie sollten daher im Zweistichprobenfall nicht genutzt werden.
2. Bei $\alpha = 0{,}01$ führt der Schätzer *V1* in Kombination mit der t-Verteilung sowohl bei den rangtransformierten Daten als auch bei den äquidistanten Scores zu oft zu Niveauüberschreitungen.
3. Mit zunehmender Variabilität in den Daten ($r > 5$) neigen sowohl die unbedingten Tests als auch die „exakten" Bootstraptests zu Niveauüberschreitungen.
4. Die exakten Permutationstests erweisen sich als extrem konservativ.
5. Bei der Verwendung von Mid-p-Tests ist im wesentlichem nicht mit Niveauüberschreitungen zu rechnen.
6. Erst ab $n = 5$ und $\alpha = 0{,}05$ zeigen die meisten Tests ein stabiles Güteverhalten.

Das Güteverhalten ist für ausgewählte Tests in den Abbildungen 6.10 bis 6. 15 dargestellt.

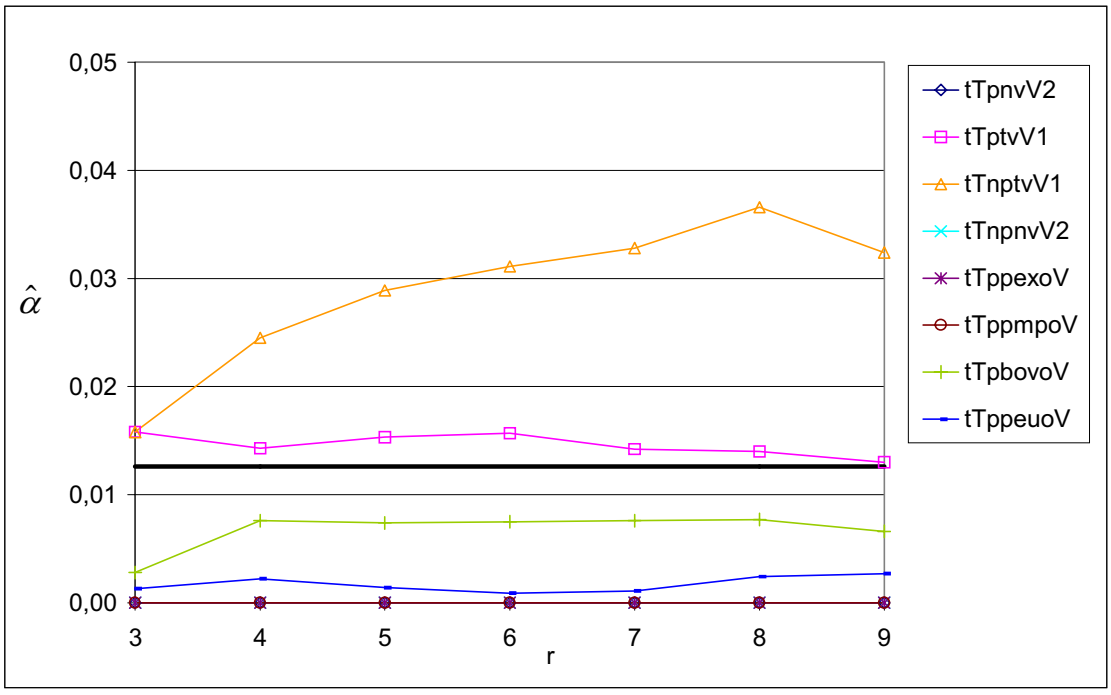


Abbildung 6.7: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 3$, $\theta = 0$, $\omega = 0{,}5$, Verteilungstyp = RSV)

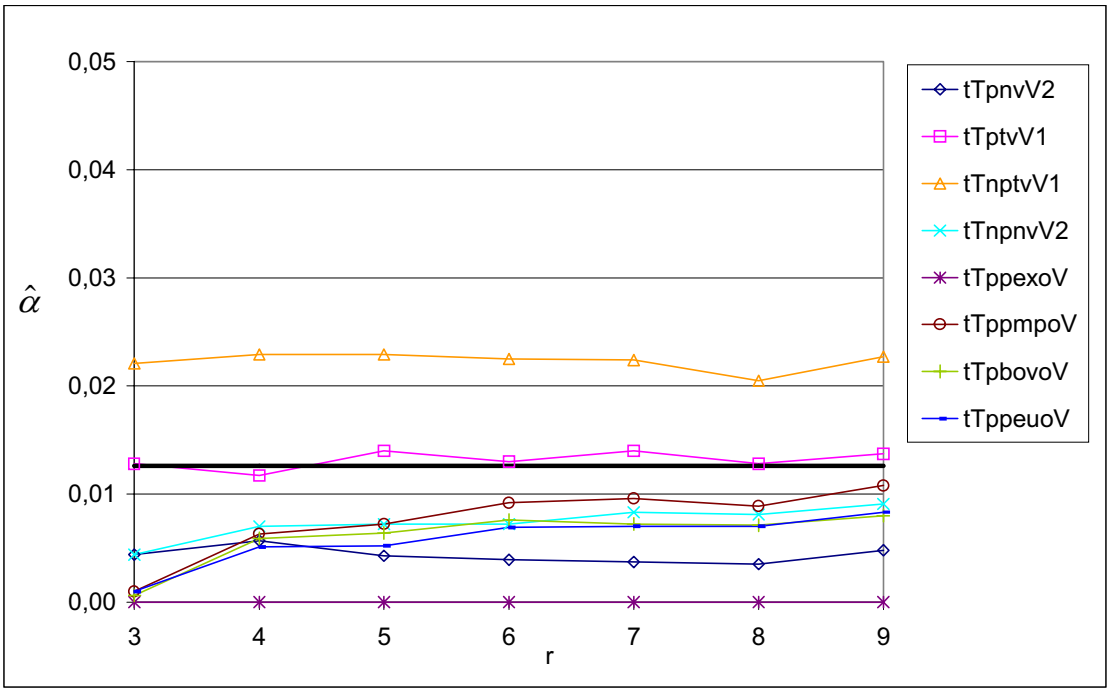


Abbildung 6.8: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

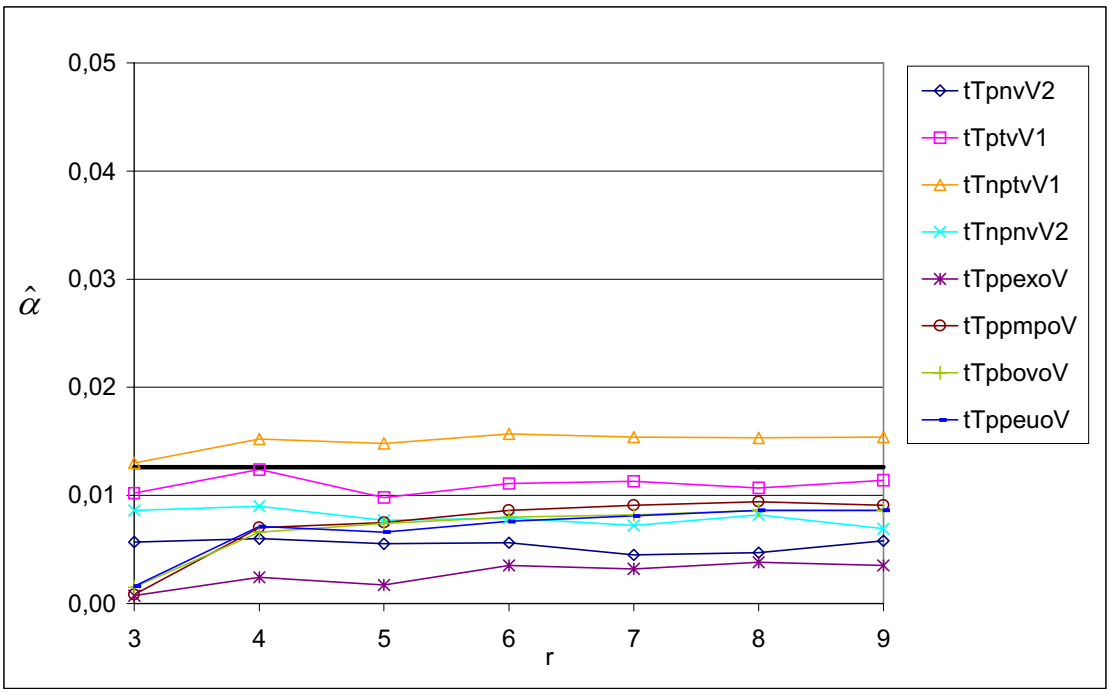


Abbildung 6.9: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

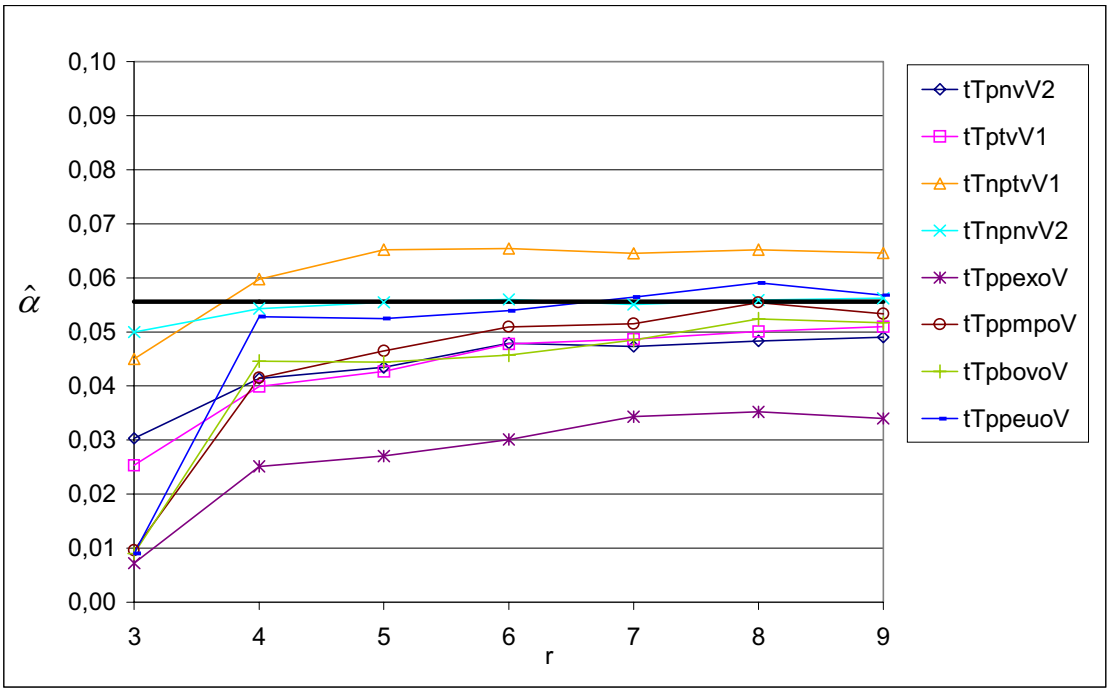


Abbildung 6.10: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

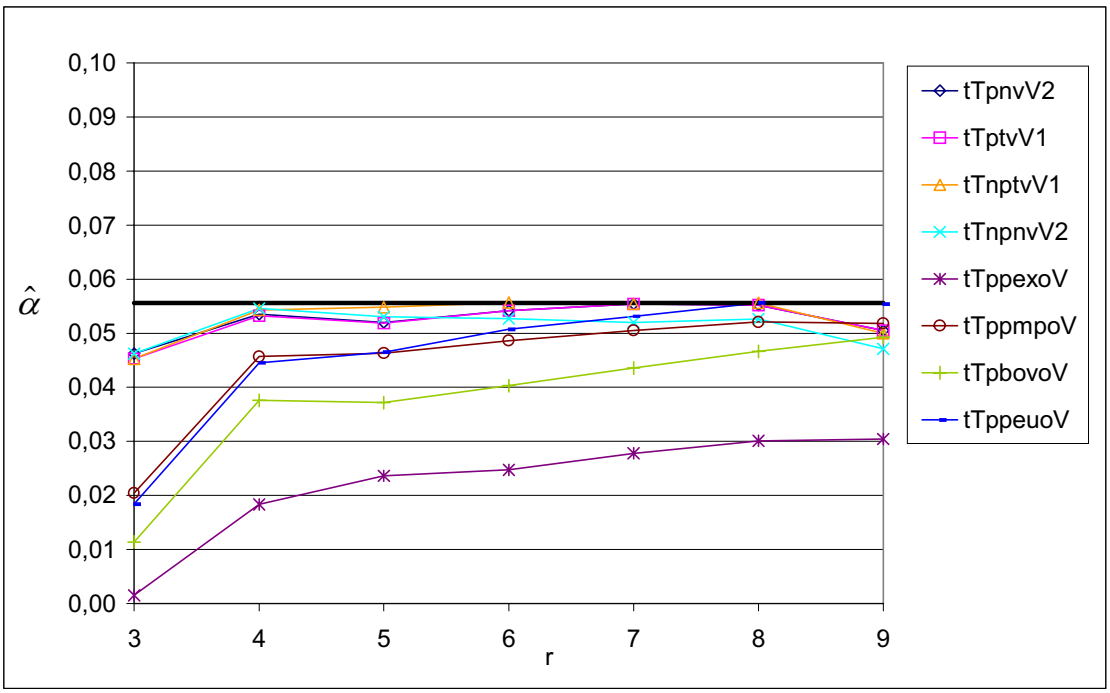


Abbildung 6.11: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

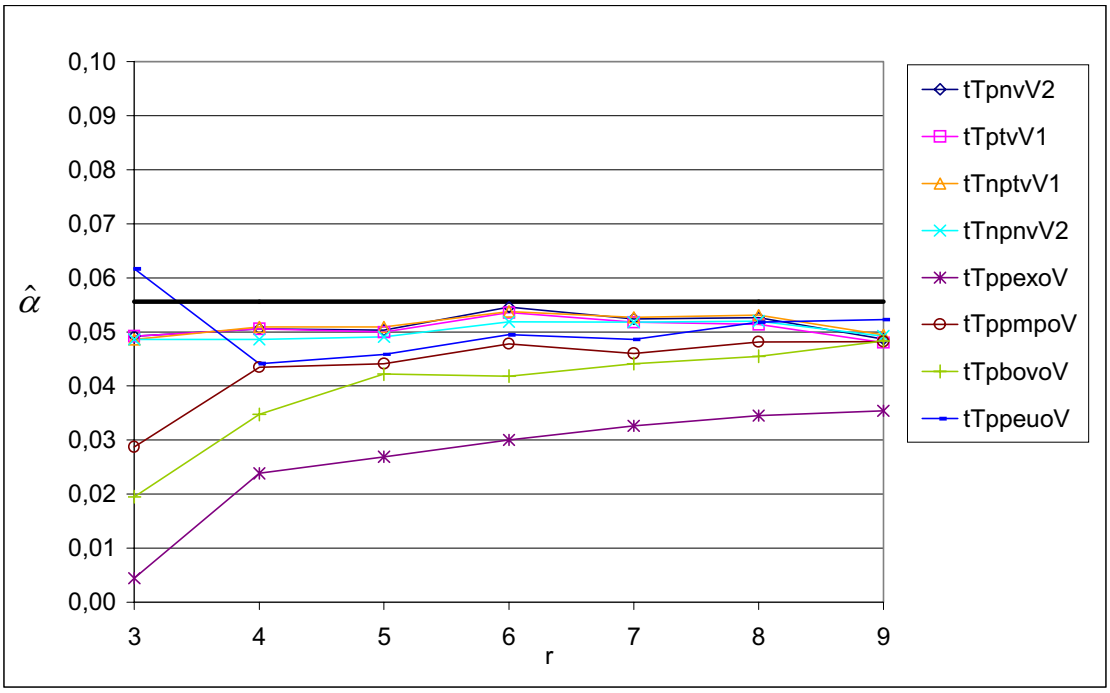


Abbildung 6.12: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp=RSV)

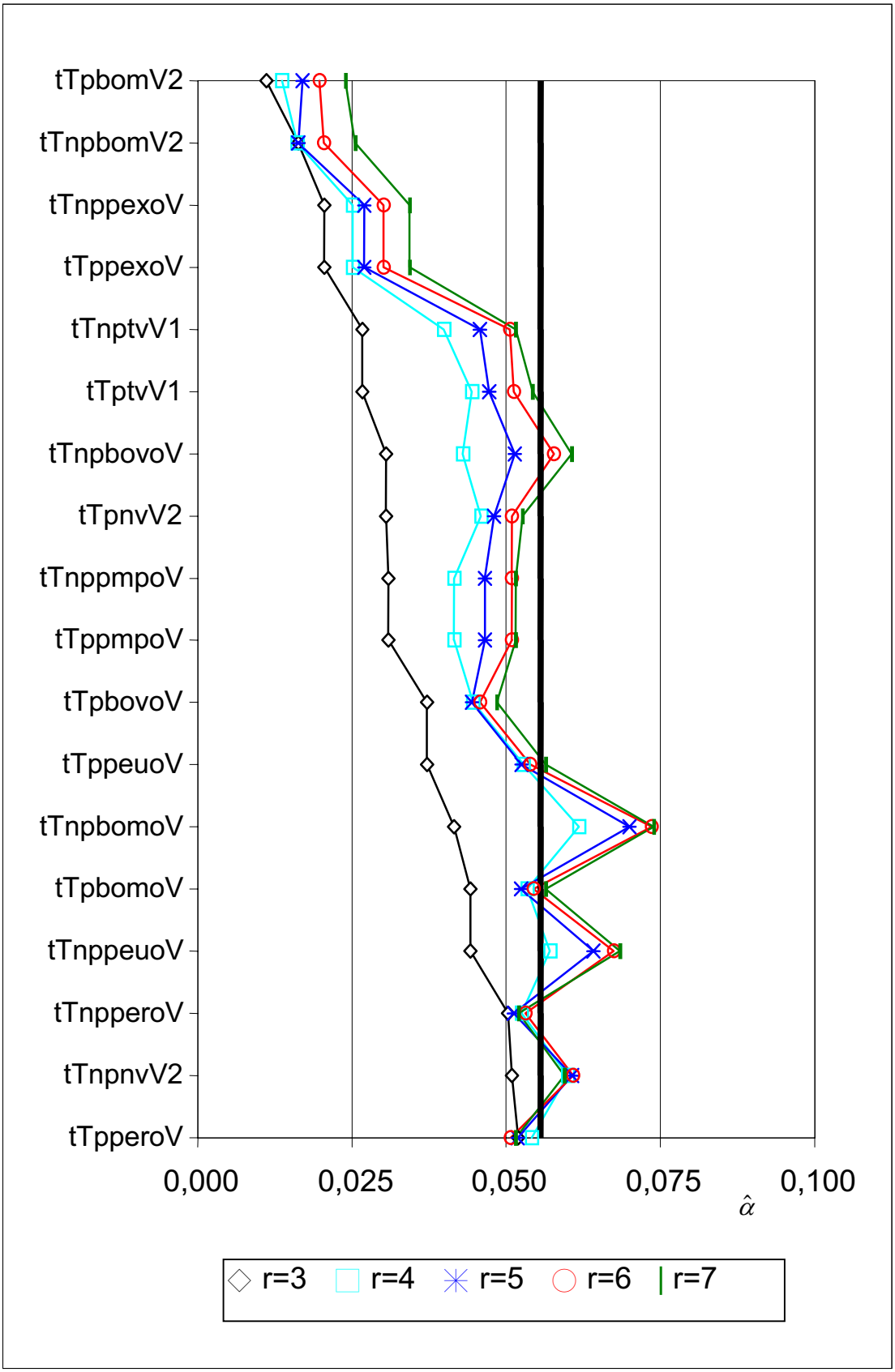


Abbildung 6.13: Güte ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\theta = 0$, $\omega = 0$, Verteilungstyp = GV)

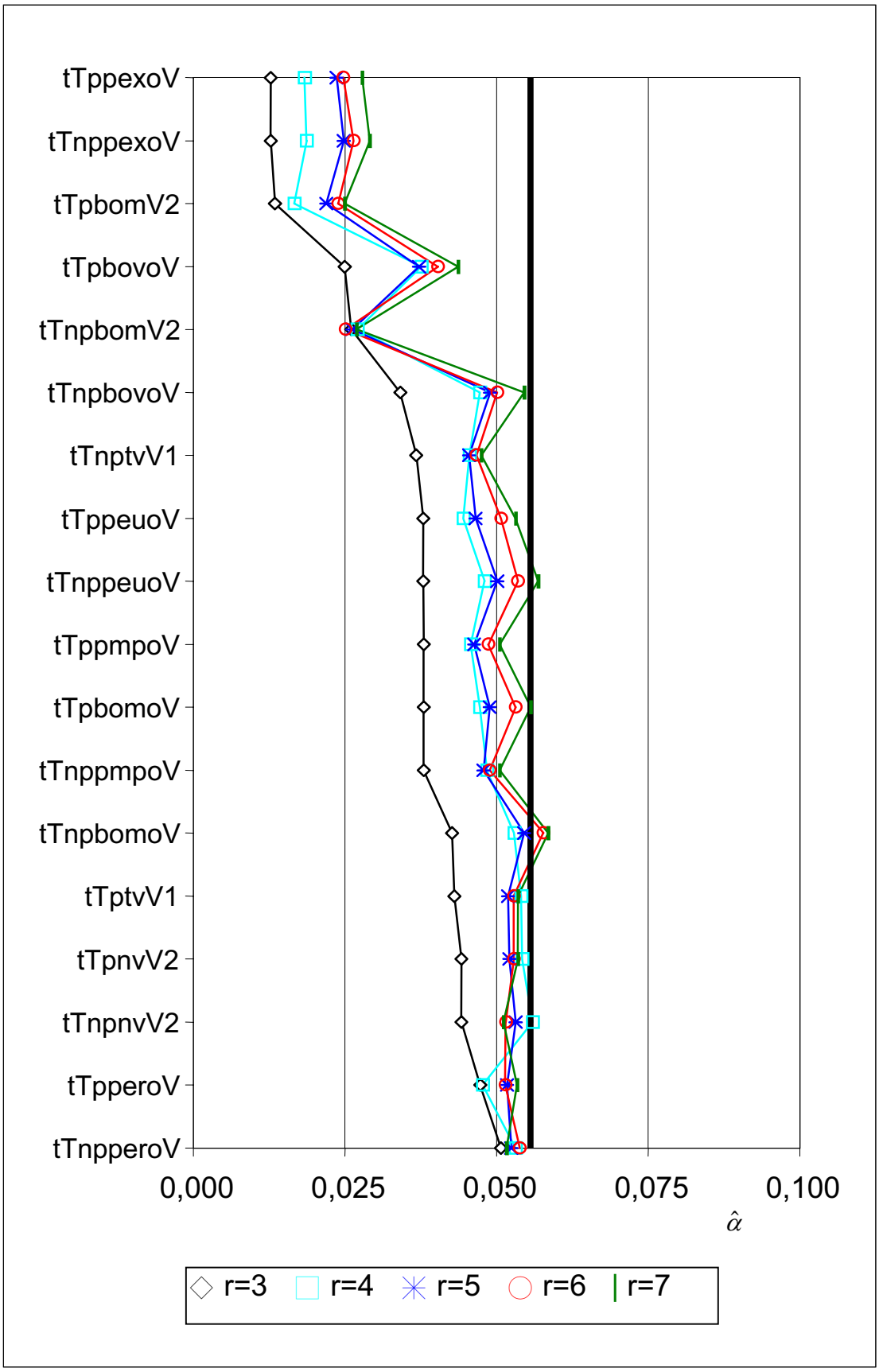


Abbildung 6.14: Güteverhalten ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0$, $\theta = 0$, Verteilungstyp = GV)

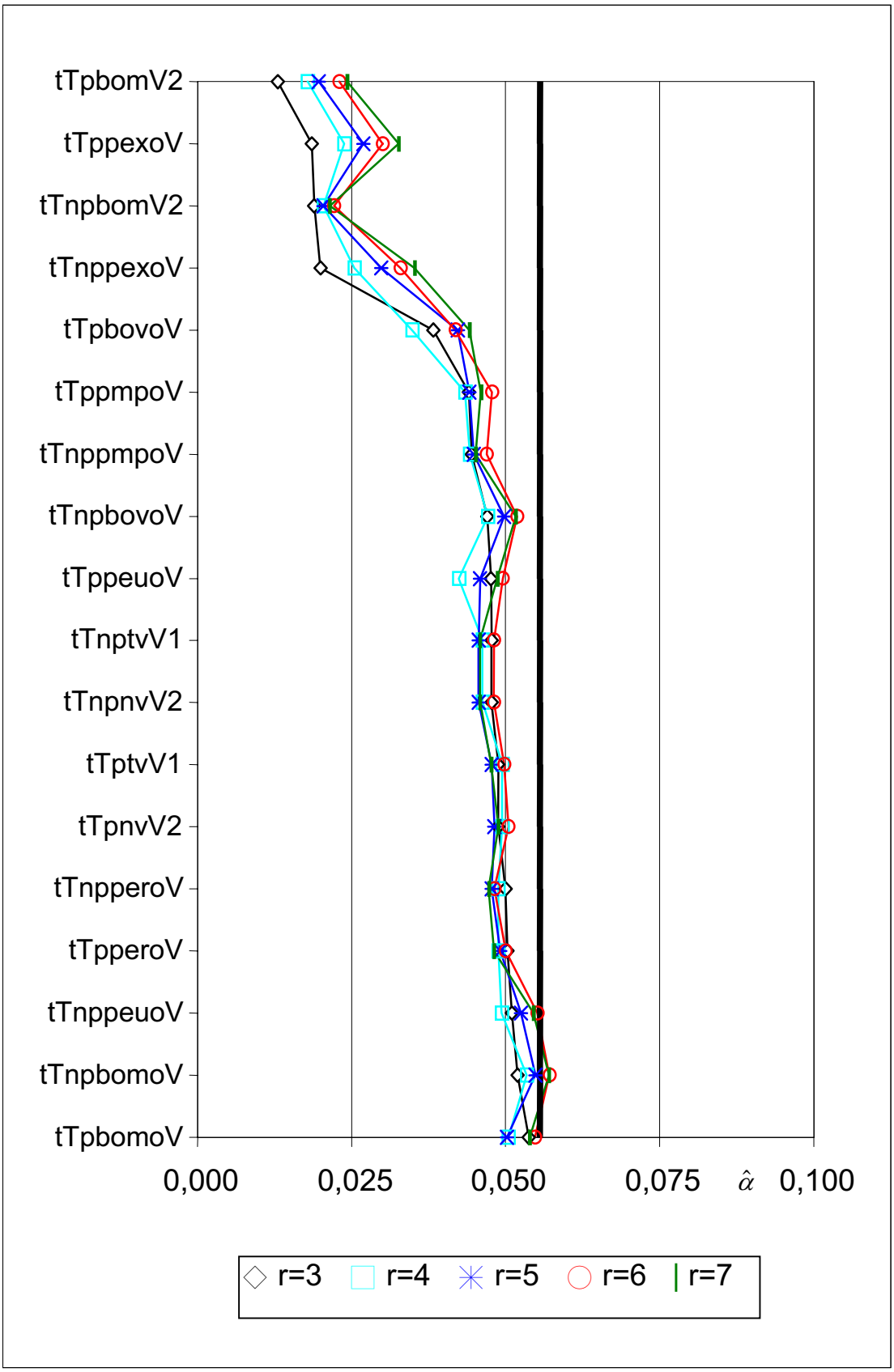


Abbildung 6.15: Güteverhalten ($\hat{\alpha}$) von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0$, $\theta = 0$,Verteilungstyp = GV)

6.3.2.2 Güte unter der Alternativhypothese

In Abbildung 6.16 bis 6.25 ist beispielhaft das Güteverhalten aller bzw. ausgewählter Tests dargestellt. Wie in allen anderen Simulationen ist auch hier deutlich zu erkennen, wie negativ sich die Konservativität der exakten bedingten Permutationstests auf die Güte unter der Alternative auswirkt. Ihre Güte ist grundsätzlich deutlich schlechter als die der anderen Tests. Analog gilt dies für die betrachteten Bootstraptests mit Varianzschätzer. Bei den exakten bedingten Permutationstests treten in manchen Situationen bei $n = 4$ Güteverluste gegenüber dem Fall $n = 3$ auf (alle anderen Parameter fest). Das heißt, hier haben die Tests beim kleineren Stichprobenumfang eine höhere Güte (siehe Abbildung 6.25). Eine Erklärung für dieses Phänomen gibt es bisher nicht. Ähnliche extrem schlechte Eigenschaften beschreibt Duchateau [176] für exakte bedingte Permutationstests im Fall von 2×2-Tafeln. Exakte bedingte Permutationstests sollten daher bei den hier betrachteten Restriktionen nicht genutzt werden. Bei dem unbedingten Permutationstest *tTppeuoV* und beim „exakten" Bootstraptest *tTpbovoV* treten im Fall $r = 4$ oft Anomalien in bezug auf *r* auf. Zum Beispiel haben diese Tests bei $n = 3$ und $r = 4$ teilweise eine höhere Güte als bei $n = 3$ und $r = 5,6,(7)$ (alle anderen Parameter fest; siehe Abbildung 6.21). Bei den anderen Tests tritt dies kaum, bzw. nicht auf.

Die Güte aller anderen Tests liegt beim Niveau $\alpha = 0{,}05$ relativ dicht beieinander. Bei $n = 3$ zeichnen sich im allgemeinen die randomisierten Tests als die Tests mit der höchsten Güte aus (siehe Abbildung 6.22). Aber schon ab $n = 4$ erweisen sich die Mid-p-Tests als die besseren (siehe Abbildung 6.23, 6.24). Die Güte der Mid-p-Tests ist im Vergleich zu den „exakten" Bootstraptests oder den unbedingten Permutationstests recht stabil. Nur bei $n = 3$ und $r < 5$ (vor allem bei $\alpha = 0{,}01$) besitzt der exakte Bootstraptest die etwas bessere Güte. Die Normalverteilungsapproximation erweist sich insgesamt gesehen der t-Verteilungs-approximation überlegen (sowohl bei den Rängen als auch bei den äquidistanten Scores). Allerdings sind auch sie den Mid-p-Tests unterlegen. Die Abhängigkeit der Güte vom Stichprobenumfang ist in Abbildung 6.25 beispielhaft dargestellt für den Fall, daß F_0 eine linksschiefe Verteilung ist. Diese Abbildung spiegelt auch das allgemeine Verhalten der Bootstraptests ohne Varianzschätzer (basierend auf Rängen) wider. Trotz ihrer doch zum Teil beträchtlichen Niveauüberschreitungen besitzen sie unter der Alternative keine höhere Güte als die anderen Tests. Die bekanntesten Zweistichprobentests, der t-Test (*tTptvV1)* und der Wilcoxon-Mann-Whitney-Test (*tTnpnvV2*), schneiden insgesamt betrachtet schlechter als die Mid-p-Tests ab. Bei $\alpha = 0{,}05$ sind die Güteunterschiede jedoch nicht sehr groß.

Bemerkung 6.1: Da eine Überschreitung um ca. 10 Prozent auch den anderen Tests erlaubt wird, wurden exakte Permutationstests zum Niveau $\tilde{\alpha} = \alpha + 0{,}1\alpha$ untersucht. Dies führt jedoch zu keiner wesentlichen Verbesserung.

Aufgrund der Simulationsergebnisse werden unter den betrachteten Restriktionen folgende Empfehlungen gegeben:

1. Für $n > 3$ wird generell ein Mid-p-Test empfohlen. Die Nutzung von Rängen wird dabei besonders bei größeren Skalen $r > 5$ empfohlen. Ähnlich, wie im stetigen Fall, sind die Ränge bei schiefen Verteilungen häufig die bessere Wahl.
2. Der Stichprobenumfang $n = 3$ sollte vermieden werden. Kein Test konnte bei diesem Stichprobenumfang überzeugen. Wenn überhaupt bei $n = 3$ getestet werden soll, dann sollte ein randomisierter Permutationstest genutzt werden.
3. Für das Signifikanzniveau $\alpha = 0{,}01$ sollte der Stichprobenumfang mindestens fünf betragen. Dann wird die Nutzung eines Mid-p-Tests oder eines unbedingten Permutationstests empfohlen (siehe Abbildung 6.16 bis 6.18).

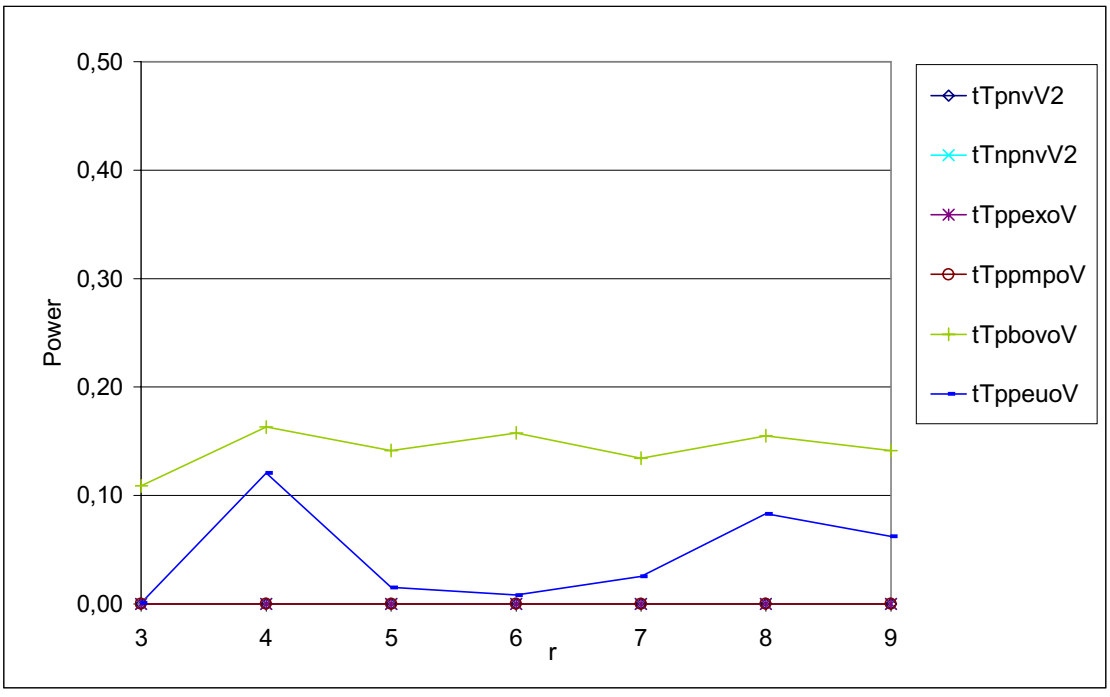


Abbildung 6.16: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 3$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

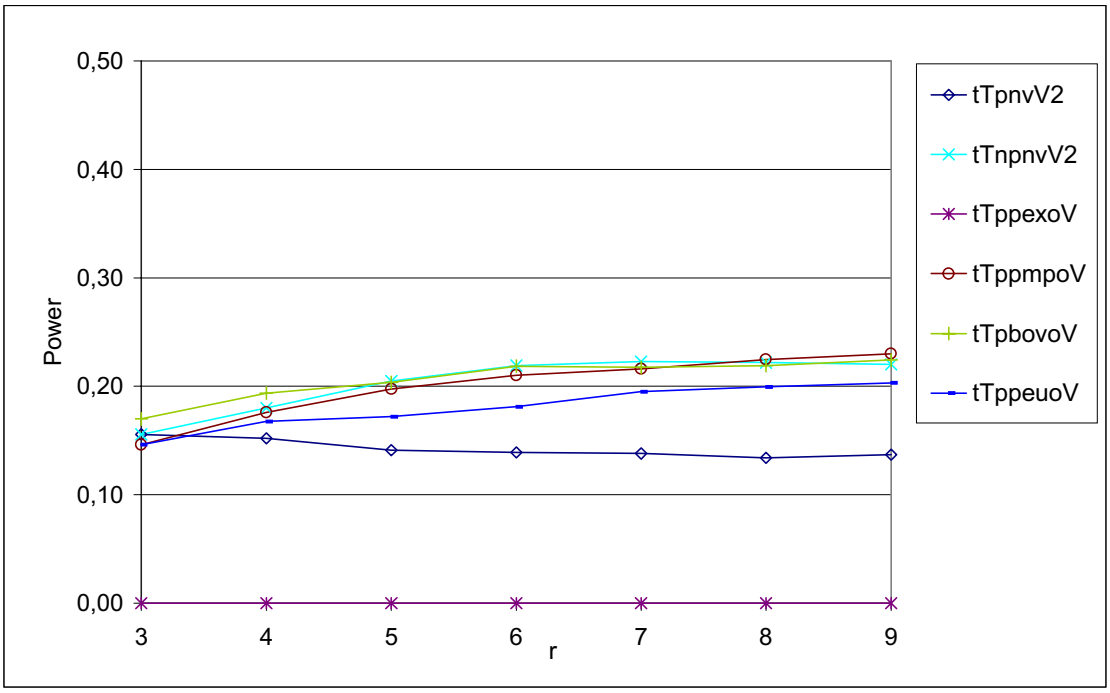


Abbildung 6.17: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 4$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

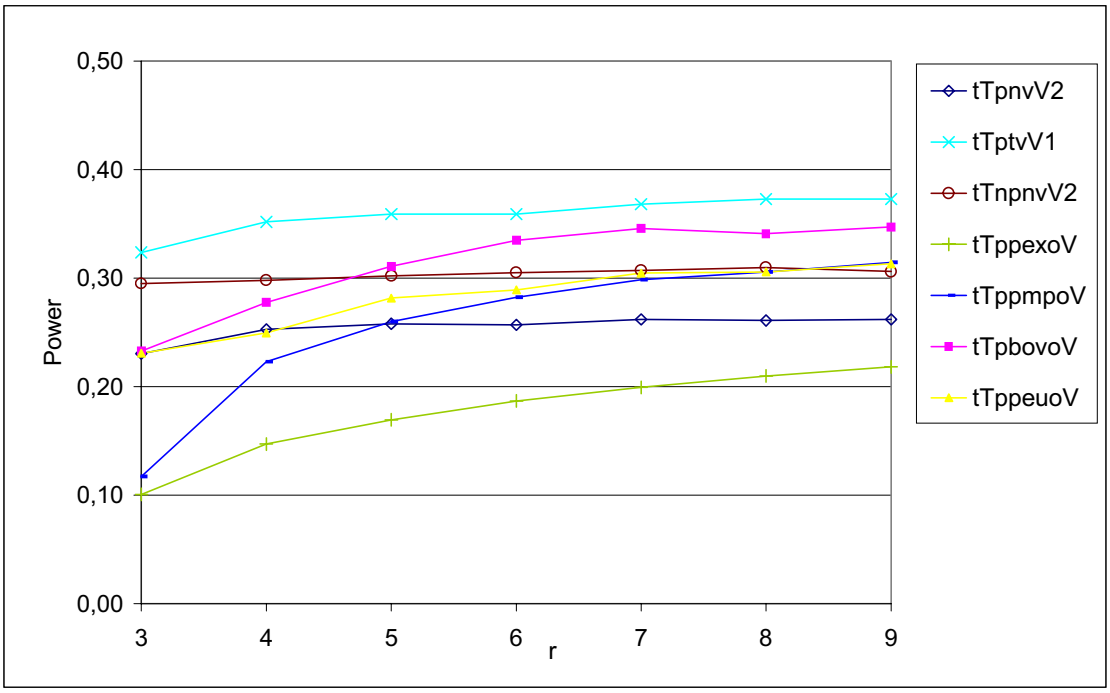


Abbildung 6.18: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}01$, $n = 5$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

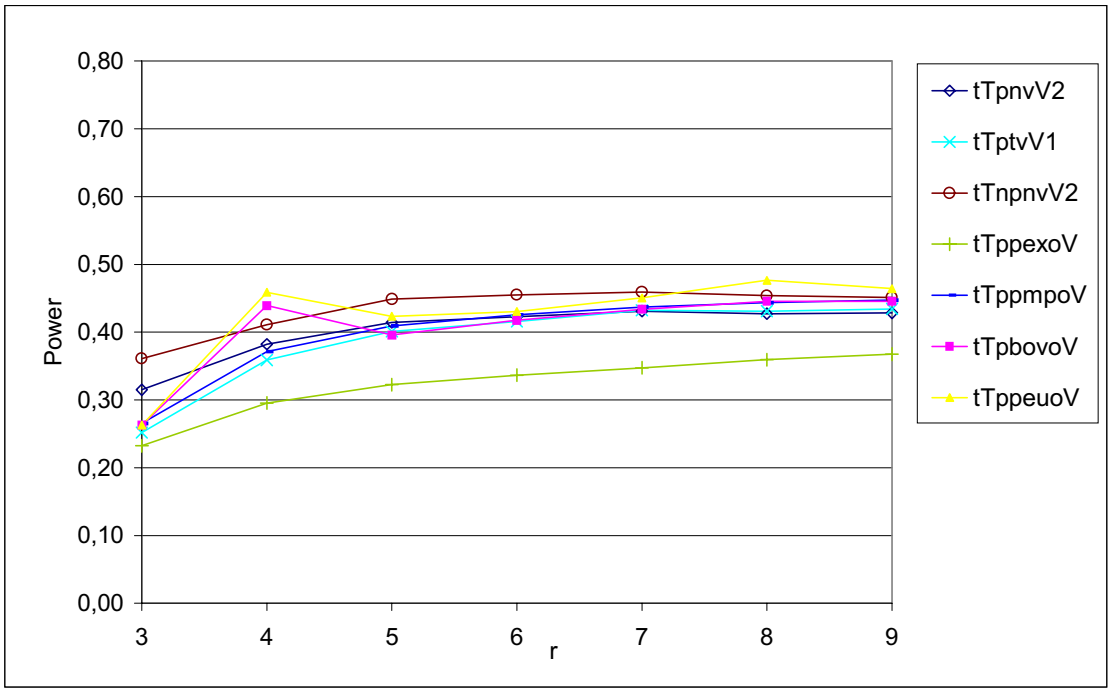


Abbildung 6.19: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

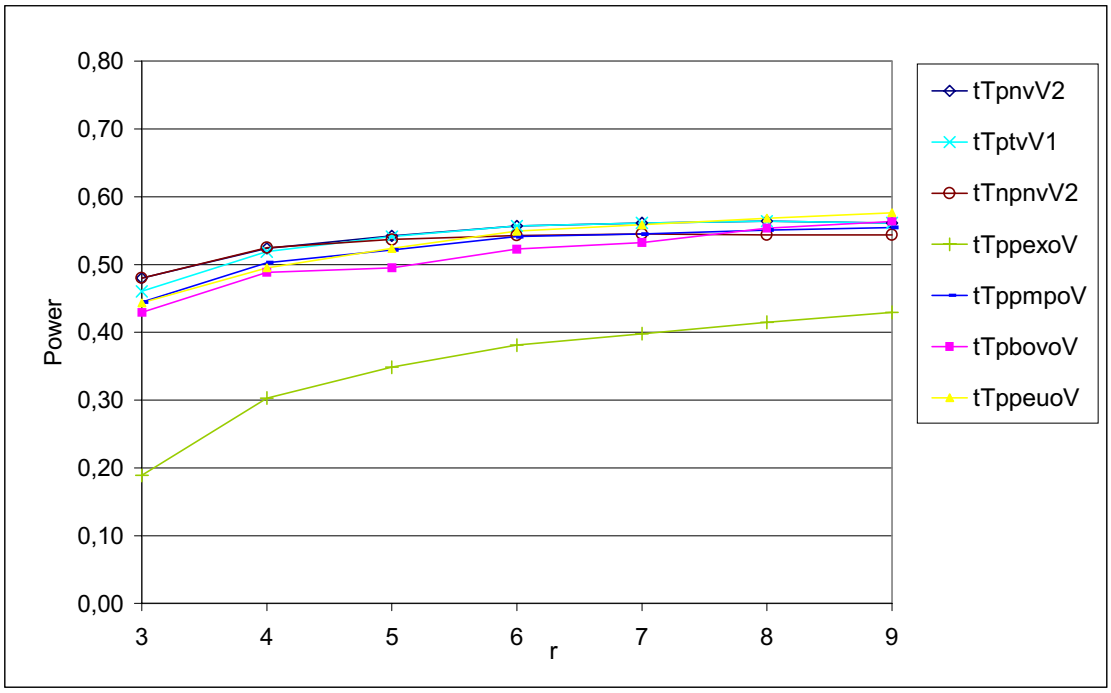


Abbildung 6.20: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

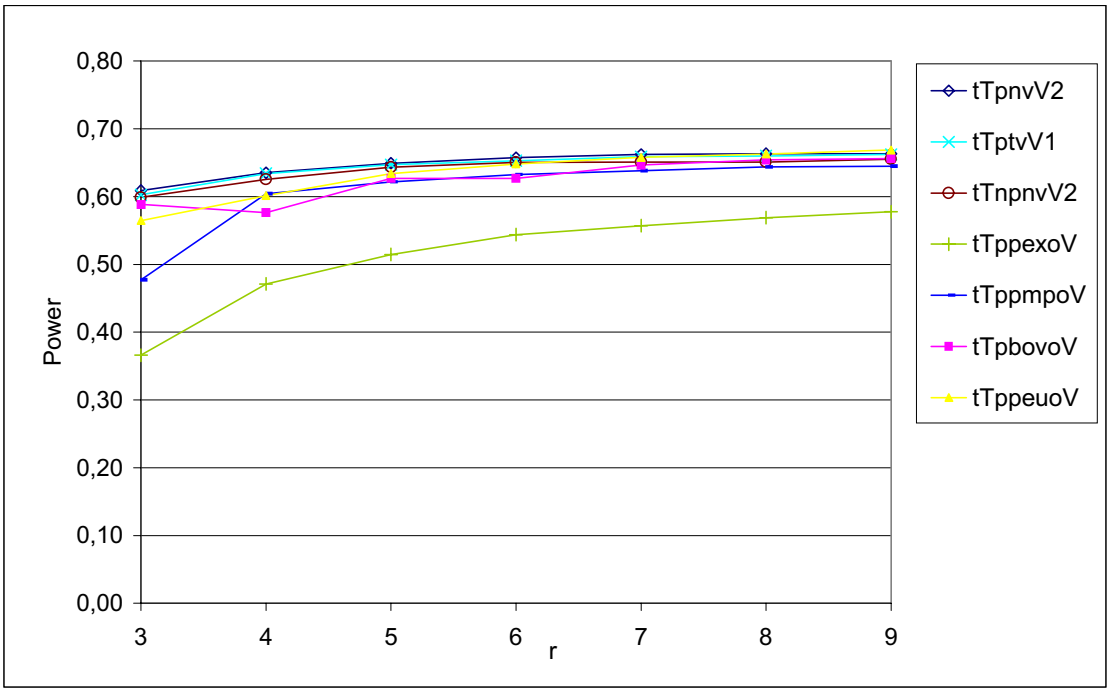


Abbildung 6.21: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = RSV)

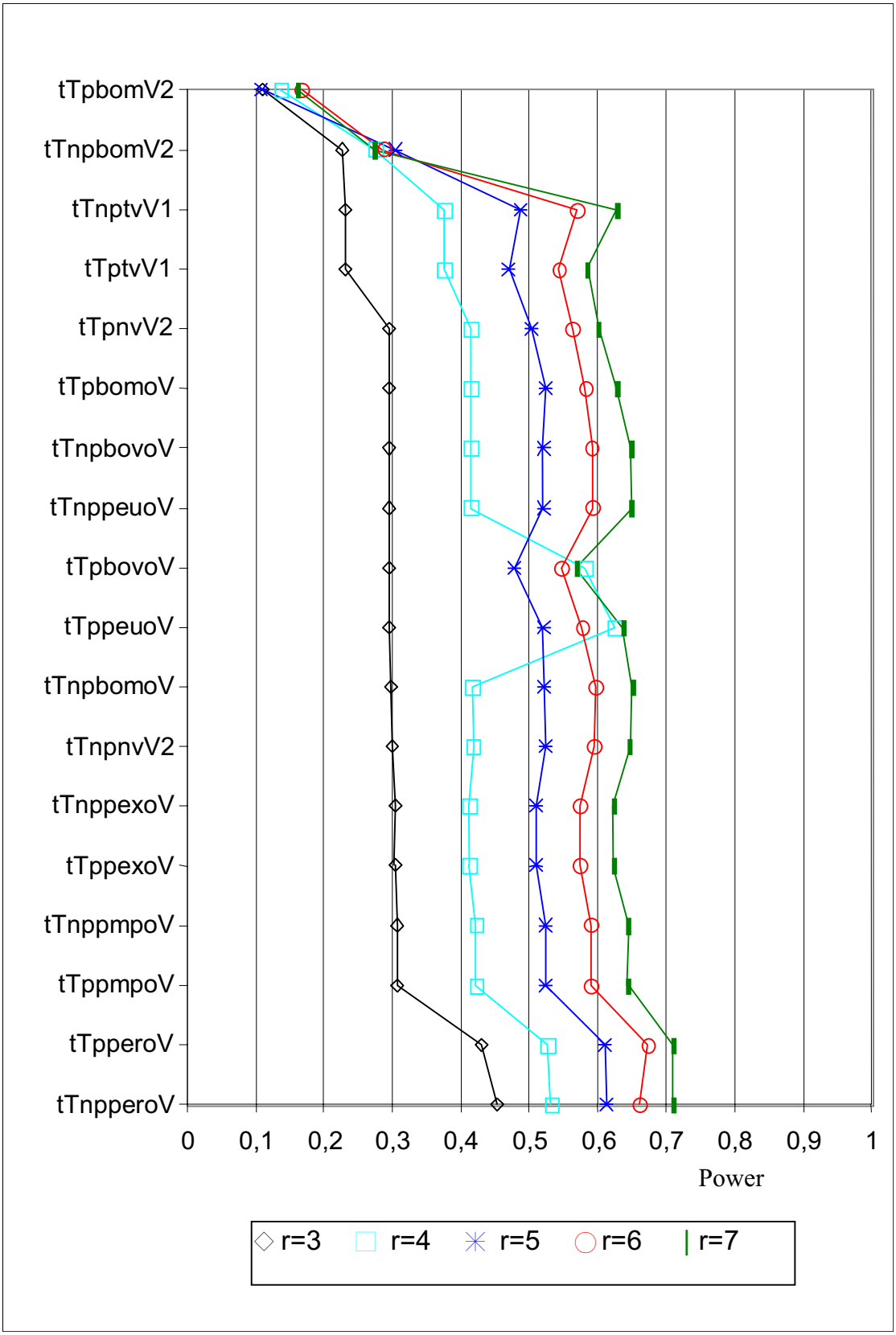


Abbildung 6.22: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = GV)

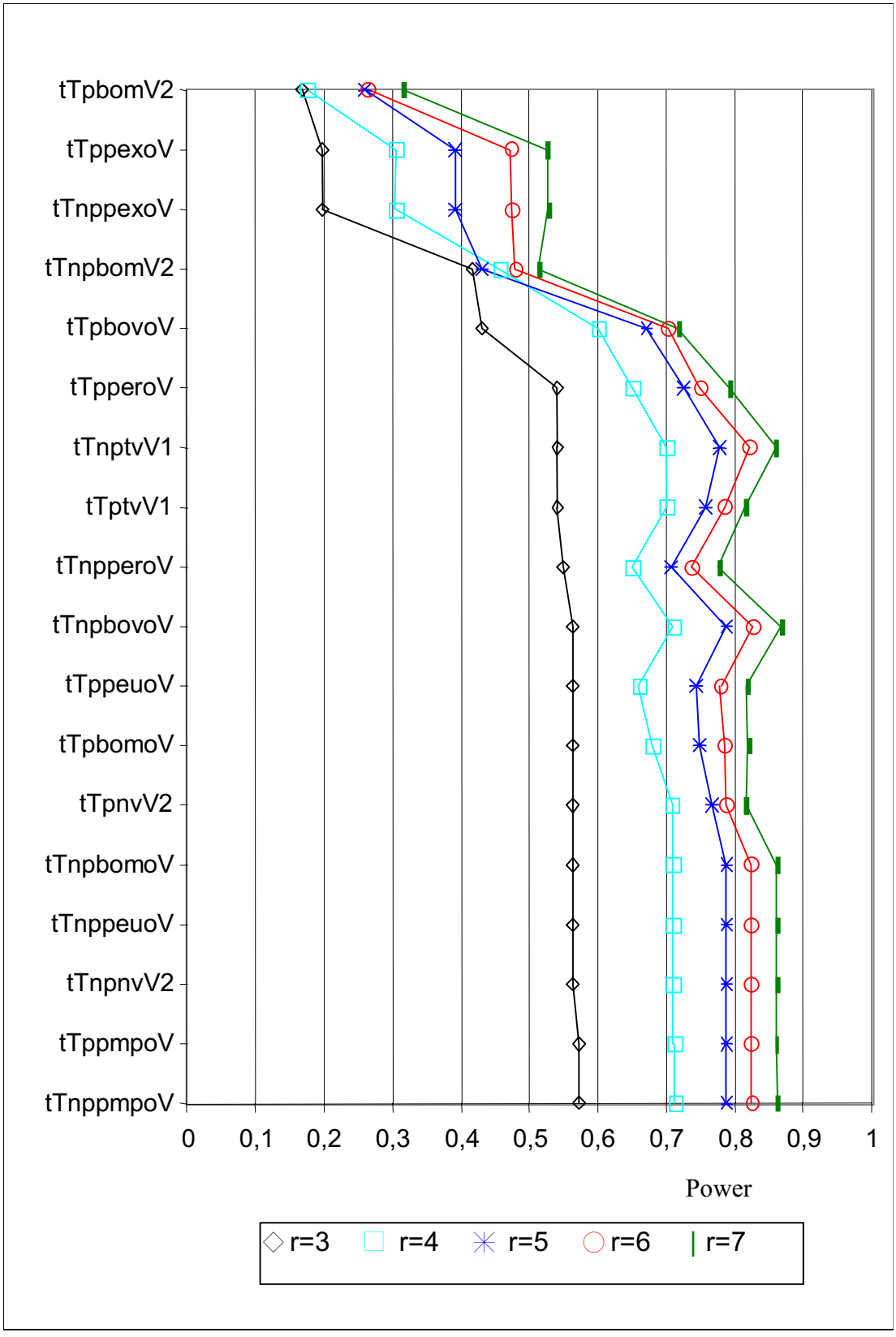


Abbildung 6.23: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = GV)

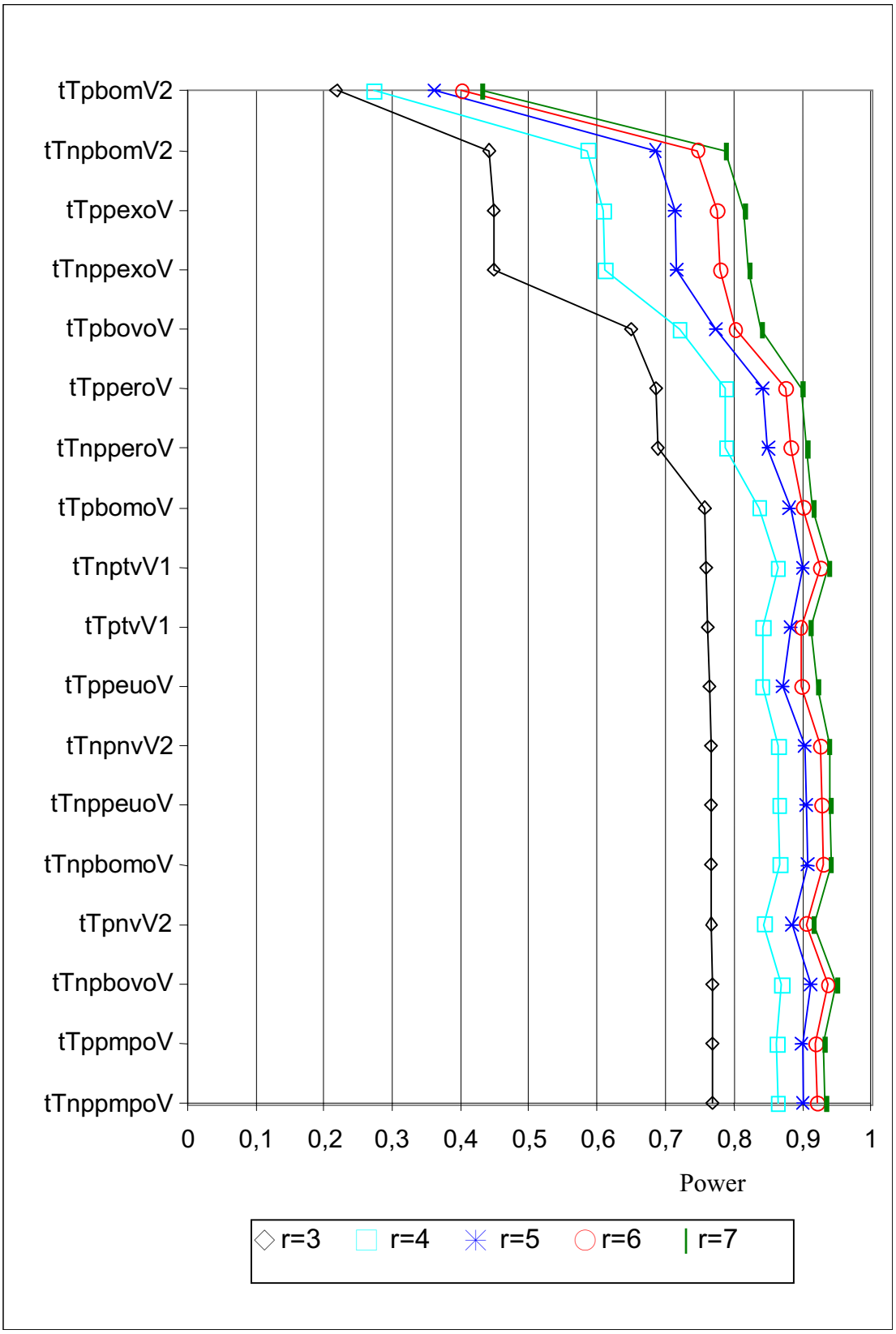


Abbildung 6.24: Power von Zweistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = GV)

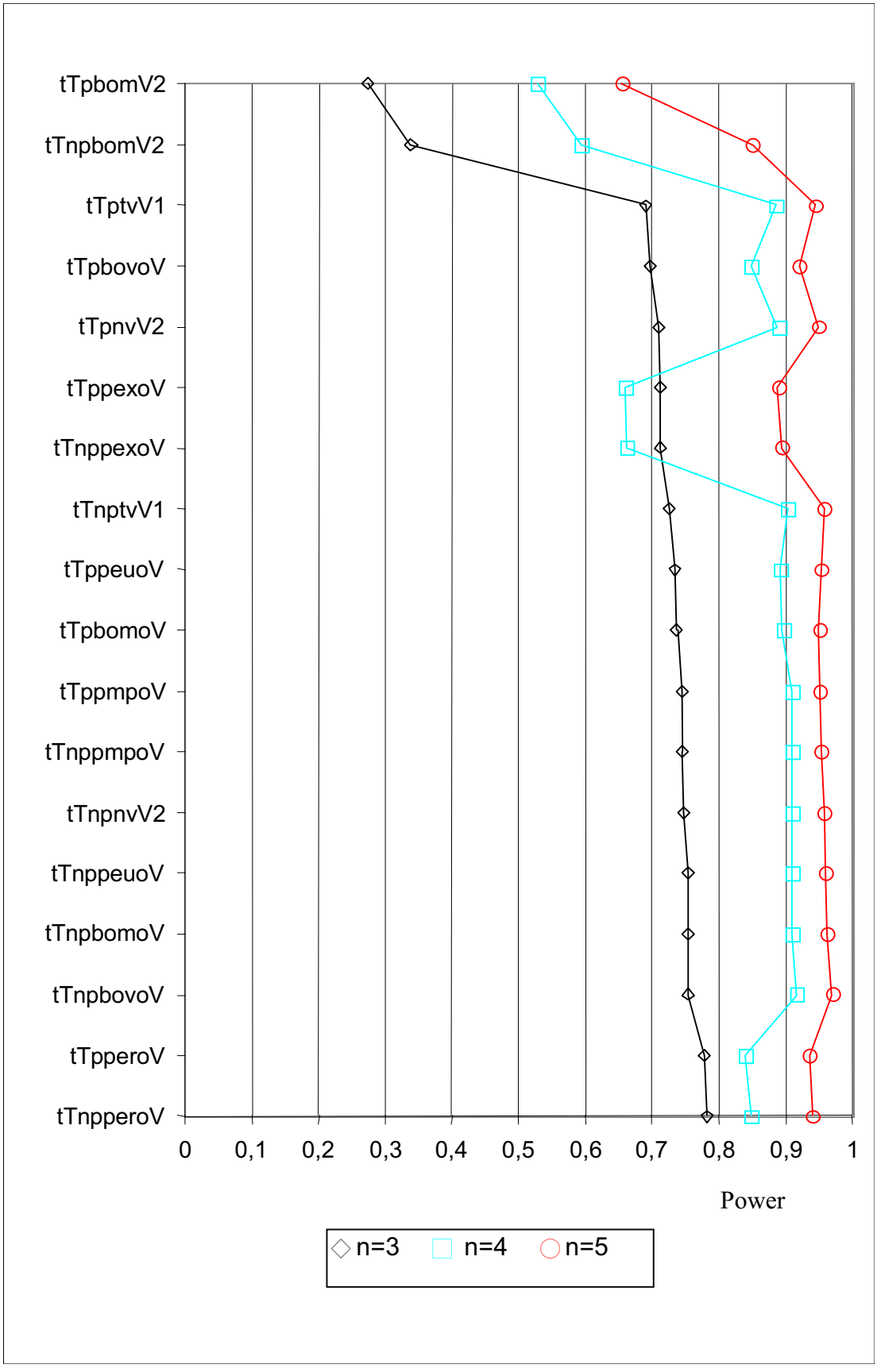


Abbildung 6.25: Power von Zweistichprobentests in Abhängigkeit vom Stichprobenumfang (n) $\alpha = 0{,}05$, $r = 7$, $\theta = 5$, $\omega = 0{,}5$, Verteilungstyp = LSV

6.3.3 Drei- und Vierstichprobentests

„Exakte“ Bootstraptests und unbedingte Permutationstests wurden unter Nutzung einfacher Kontraste untersucht, da für diese Statistiken der Rechenaufwand am geringsten ist. Wie im Zweistichprobenfall erweisen sich diese unbedingten Permutationstests zum Teil als liberal (siehe Tabelle 6.5). Die „exakten“ Bootstraptests neigen vor allem bei größeren Skalen zu Liberalität. Insgesamt ist der Rechenaufwand für die einfachen Kontraste innerhalb von Simulationen zu groß. Für die „interessanteren“ Statistiken (z. B. multiple Kontraste) wurden daher keine Simulationen durchgeführt. Sowohl die „exakten“ Bootstraptests als auch die unbedingten Permutationstests werden daher aufgrund nicht genügender Güteabschätzungen weder als geeignet noch als ungeeignet eingestuft. Ähnlich sieht es beim Doublebootstraptest aus. Hier wurden zwar einfache und multiple Kontraste untersucht, aufgrund der hohen Berechnungszeiten wurden jedoch nur wenige Punkte simuliert. Dabei zeigten sich stets geringe Gütevorteile für den Double-Bootstraptest (im Vergleich zu asymptotischen Tests und „einfachen“ Bootstraptests). Wesentliche Niveauüberschreitungen treten zwar ebenfalls auf, sie sind jedoch seltener als beim „einfachen“ Bootstraptest. In Tabelle 6.6 sind Beispiele für den multiplen Kontrast TK auf der Basis von Rängen und der Nutzung des Double-Bootstraps bzw. des „einfachen“ Bootstraps für konkave Erwartungswertprofile dargestellt (Bestimmung der Ränge nach jedem Ziehen der Resamplingstichproben). Während sich für $r = 3$, $n = 3,4$ deutliche Unterschiede zeigen, sind bei $r = 4,\ 5$, $n = 4,5$ nur noch relativ kleine Unterschiede zu erkennen. Für $\alpha = 0{,}05$, $n = 4,5$ und $r > 2$ zeigten sich auch in den anderen Simulationen für den Double-Bootstrap kaum Powergewinne, die größer als 0,03 waren. Eine weitere Untersuchung der Double-Bootstraptests ist angezeigt, da vor allem das Güteverhalten unter der Nullhypothese besser ist als bei „einfachen“ Bootstraptests. Diese Untersuchungen müssen aber auf einen Zeitpunkt verschoben werden, an dem leistungsfähigere Computer zur Verfügung stehen.

Im folgenden werden Ergebnisse für die in Tabelle 6.7 beschriebenen Tests vorgestellt. Für die Parameter *r, n,* α, ς, ω, θ wurden Werte aus den in Abschnitt 6.2 beschrieben Mengen gewählt. Neben den oben beschriebenen vier Verteilungstypen wurden für $k = 2,3$ die drei beschriebenen Dosis-Wirkungs-Profile (konvex, konkav und linear) simuliert. Die Wahl der beschriebenen Parameter führt zu

$$7 \times 3 \times 2 \times 5 \times 4 \times 4 \times 3 \times 2 + 7 \times 3 \times 2 \times 4 \times 3 \times 2 = 21.168$$

$$21168 - 3528 = 17.640$$

	$r = 3$		$r = 4$		$r = 5$		$r = 6$	
	H_0	H_A	H_0	H_A	H_0	H_A	H_0	H_A
RHKnppexoV	0,021	0,256	0,027	0,343	0,029	0,387	0,030	0,420
RHKnppmpoV	0,032	0,494	0,046	0,537	0,045	0,549	0,048	0,563
RHKnpperoV	0,049	0,444	0,049	0,486	0,046	0,509	0,049	0,528
RHKnppeuoV	0,048	0,581	0,060	0,581	0,059	0,582	0,063	0,590
RHKnpbovoV	0,035	0,487	0,052	0,547	0,052	0,580	0,054	0,604

Tabelle 6.5: Vergleich „exakter“ Bootstraptest versus Permutationstests ($\alpha = 0{,}05$, $k = 2$, $n = 3$, $\theta = 0$ bzw. $\theta = 5$, $\omega = 0$, konkave Profile, Verteilungstyp = GV)

			$n = 3$		$n = 4$		$n = 5$	
r	α	Test	H_0	H_A	H_0	H_A	H_0	H_A
3	0,01	TKnpbomoV	0,009	0,286	0,010	0,560	0,012	0,741
		TKnpdbovoV	0,011	0,415	0,009	0,633	0,010	0,778
	0,05	TKnpbomoV	0,055	0,710	0,049	0,855	0,054	0,930
		TKnpdbovoV	0,056	0,750	0,049	0,881	0,050	0,943
4	0,01	TKnpbomoV	0,012	0,406	0,011	0,692	0,012	0,845
		TKnpdbovoV	0,012	0,520	0,010	0,733	0,095	0,861
	0,05	TKnpbomoV	0,055	0,810	0,050	0,914	0,052	0,962
		TKnpdbovoV	0,055	0,837	0,048	0,931	0,049	0,969
5	0,01	TKnpbomoV	0,012	0,491	0,011	0,762	0,011	0,893
		TknpdbovoV	0,011	0,583	0,010	0,788	0,010	0,898
	0,05	TknpbomoV	0,054	0,864	0,051	0,940	0,052	0,976
		TknpdbovoV	0,051	0,879	0,049	0,951	0,049	0,980

Tabelle 6.6: Vergleich Bootstraptest (TknpbomoV) versus Double-Bootstraptest (TknpdbovoV) auf der Basis von Rängen ($k = 3$, konkave Profile)

(Nullhypothesen sind bei den unterschiedlichen Profilen gleich) verschiedenen Parameterkonfigurationen. In diesem Umfang wurden nicht alle der in Tabelle 6.7 dargestellten Tests untersucht. Vor allem bei den Bootstraptests und den exakten Permutationstests wurden nur ausgewählte Parameterkonfigurationen untersucht.

Statistik	**Verteilung**	**Varianz-schätzer**	**Bezeichner**
$\overline{E}_{01}^2$ äquidistante Scores	Chi-Bar-Verteilung aus Satz 4.1	*V1*	*BARpchbV2*
	Bootstrapverteilung; Resamplingraum RS	*oV*	*BARpbomoV*
	Bootstrapverteilung; Resamplingraum RS_Z	*V2*	*BARpbomV2*
	exakte bedingte Permutationsverteilung	*oV*	*BARppexoV*
	Mid-p-Test	*oV*	*BARppmpoV*
	randomisierter Permutationstest	*oV*	*BARpperoV*
$\overline{\chi}_{01}^2(R_{II})$ Ränge	Chi-Bar-Verteilung Satz 5.4	*V1*	*CHAnpchbV2*
	Bootstrapverteilung; Resamplingraum $RS(R)$	*oV*	*CHAnpbomoV*
	Bootstrapverteilung; Resamplingraum $RS_Z(R)$	*V2*	*CHAnpbomV2*
	exakte bedingte Permutationsverteilung	*oV*	*CHAnppexoV*
	Mid-p-Test	*oV*	*CHAnppmpoV*
	randomisierter Permutationstest	*oV*	*CHAnpperoV*
$T_{ISO}^{\max}$ äquidistante Scores[1]	multivariate t-Verteilung aus Satz 4.4	*V1*	*ISOptvV1*
	multivariate Normalverteilung aus Satz 4.5	*V2*	*ISOpnvV2*
	Bootstrapverteilung; Resamplingraum RS	*oV*	*ISOpbomoV*
	Bootstrapverteilung; Resamplingraum RS_Z	*V2*	*ISOpbomV2*
	exakte bedingte Permutationsverteilung	*oV*	*ISOppexoV*
	Mid-p-Test	*oV*	*ISOppmpoV*
	randomisierter Permutationstest	*oV*	*ISOpperoV*
$T_{ISO}^{\max}(R)$ Ränge[1]	multivariate t-Verteilung aus Bemerkung 5.2	*V1*	*ISOnptvV1*
	multivariate Normalverteilung aus Satz 5.6	*V2*	*ISOnpnvV2*
	Bootstrap-Verteilung; Resamplingraum $RS(R)$	*oV*	*ISOnpbomoV*
	Bootstrap-Verteilung; Resamplingraum $RS_Z(R)$	*V2*	*ISOnpbomV2*
	exakte bedingte Permutations-Verteilung	*oV*	*ISOnppexoV*
	Mid-p-Test	*oV*	*ISOnppmpoV*
	randomisierter Permutationstest	*oV*	*ISOnpperoV*

Tabelle 6.7: Bezeichner für die im Abschnitt 6.3.2 beschriebenen k-Stichprobentests

[1] Bis auf die exakten Permutationstests, Mid-p-Tests und randomisierten Tests wurden die analogen Tests auf Basis der Sugiura-Kontrastmatrix ebenfalls untersucht.

Die Varianzschätzer besitzen im balancierten Fall folgende Gestalt:

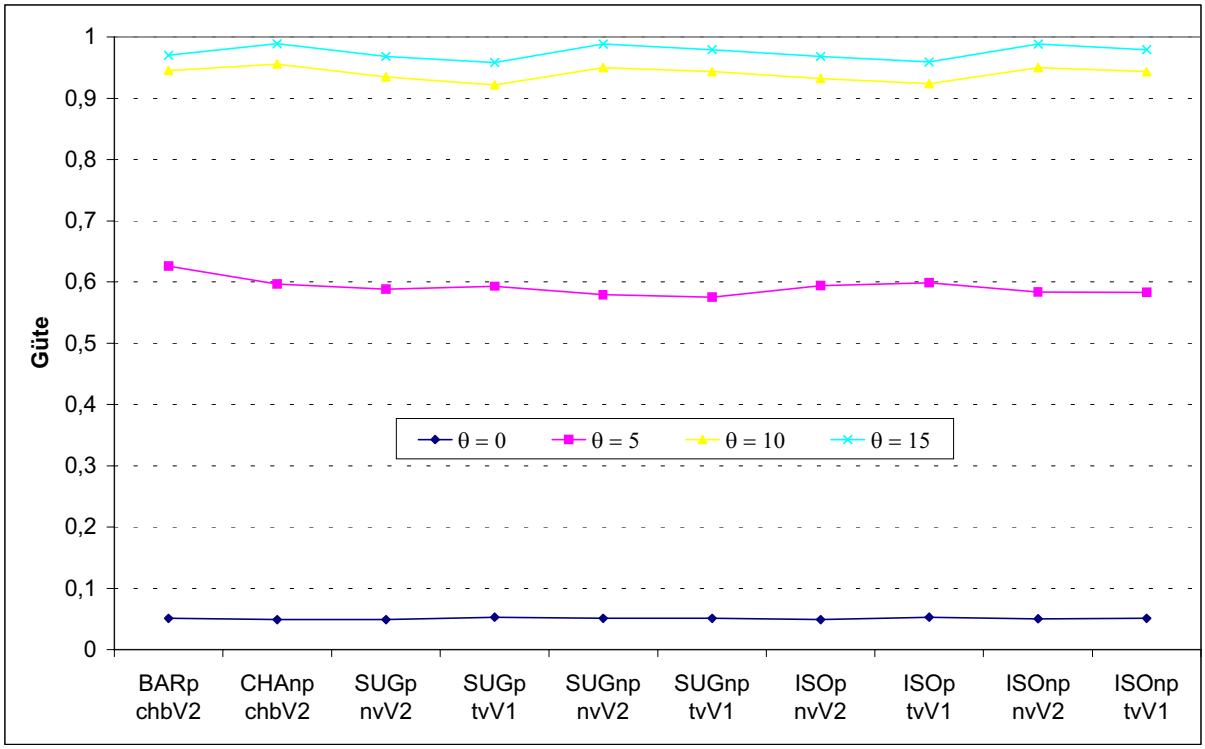
$$V_1 = \sum_{i=0}^{k} c_{li}^2 \frac{1}{n(k+1)(n-1)} \sum_{i=0}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{i.})^2 \text{ , } V_2 = \sum_{i=0}^{k} c_{li}^2 \frac{1}{n(N-1)} \sum_{i=0}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{..})^2 \text{ bzw.}$$

$$V_1 = \sum_{i=0}^{k} c_{li}^2 \frac{1}{n(k+1)(n-1)} \sum_{i=0}^{k} \sum_{j=1}^{n} (R_{ij} - \overline{R}_{i.})^2 \text{ , } V_2 = \sum_{i=0}^{k} c_{li}^2 \frac{1}{n(N-1)} \sum_{i=0}^{k} \sum_{j=1}^{n} (R_{ij} - \overline{R}_{..})^2 \text{ .}$$

In diesem Abschnitt beruhen die multiplen Kontraste auf $\boldsymbol{C}_{SUG}$ bzw. $\boldsymbol{C}_{ISO}$, da diese für beliebiges *k* definiert werden können. Die Güte des auf $\boldsymbol{C}_{TK}$ basierenden multiplen Kontrastes ist für $k = 3$ ähnlich der Güte des auf $\boldsymbol{C}_{SUG}$ beruhenden multiplen Kontrasttests. Für $k > 3$ ist jedoch zu erwarten, daß die drei Kontraste des Trivariaten Kontrastes die Alternative nicht gut genug überdecken und somit für einige bestimmte Alternativen der Trivariate Kontrast eine deutlich geringere Güte besitzt als z. B. ein multipler Kontrast, der auf $\boldsymbol{C}_{SUG}$ bzw. $\boldsymbol{C}_{ISO}$ beruht. In den Abbildungen 6.26 a) und b) bzw. 6.27 a) und b) ist beispielhaft das Verhalten der Tests, deren Verteilungen auf einer infiniten Verteilung (multivariate Normalverteilung bzw. multivariate t-Verteilung) basieren, für konkave bzw. konvexe Dosis-Wirkungs-Profile dargestellt. Es ist zu erkennen, daß die dargestellten Tests relativ robust gegen die Form der speziellen Alternative sind. Der Unterschied zwischen den auf $\boldsymbol{C}_{SUG}$ bzw. $\boldsymbol{C}_{ISO}$ basierenden multiplen Kontrasten ist bei gleicher Verteilungsaussage (z. B. jeweils multivariate t-Verteilung) und der Nutzung gleicher Scores (z. B. jeweils äquidistante Scores) für die untersuchten Profile gering. Theoretisch ist z. B. der *SUGptvV1*-Test dem *ISOptvV1*-Test bei den oben definierten streng konvexen und streng konkaven Profilen überlegen (die Zeilen von $\boldsymbol{C}_{SUG}$ stellen eine Teilmenge der Zeilen von $\boldsymbol{C}_{ISO}$ dar; das Maximum über eine kleinere Menge ist stets kleiner als das Maximum über eine größere Menge; für die simulierten konvexen bzw. konkaven Profile nehmen beide Test in der Regel denselben Wert (*t*) an; somit folgt für die p-Werte $P(T_{ISO}^{\max} \geq t) \geq P(T_{SUG}^{\max} \geq t)$). Die in den Simulationen geschätzten Güten dieser beiden Tests unterscheiden sich für die beschriebenen Profile jedoch nur unbedeutend. Der Gütegewinn des Isotonen Kontrastes bei Erwartungswertprofilen, die mit den Sugiura-Kontrasten nicht so gut korrelieren, ist in der Regel bedeutender. Da alle in Tabelle 6.7 beschriebenen Tests relativ robust gegen die Form der speziellen Alternative sind, wird im weiteren nicht mehr auf die Abhängigkeit vom gewählten Profil eingegangen.

Zwischen dem Isotonen Kontrasttest *ISOnppexoV* und dem *CHAnppexoV* –Test besteht kein praktisch relevanter Güteunterschied. Ähnlich sieht es bei den Vergleichen *ISOppexoV* vs. *BARppexoV, ISOnppmpoV* vs. *CHAnppmpoV, ISOppmpoV* vs. *BARppmpoV, ISOnpperoV* vs. *CHAnpperoV, ISOpperoV* vs. *BARpperoV, ISOnpbomoV* vs. *CHAnpbomoV, ISOpbomoV* vs. *BARpbomoV, ISOnpbomV2* vs. *CHAnpbomV2* und *ISOpbomV2* vs. *BARpbomV2* aus. Das heißt, bei den Bootstraptests und bei den Permutationstests ist es unwesentlich, ob als Statistik der Isotone Kontrast verbunden mit Rängen (äquidistanten Scores) oder die Chacko-Statistik (Bartholomew-Statistik) genutzt wird. Unterschiede gibt es jedoch bei der Verwendung der infiniten Verteilungen und zwischen den einzelnen Verteilungsapproximationen.

a)



b)

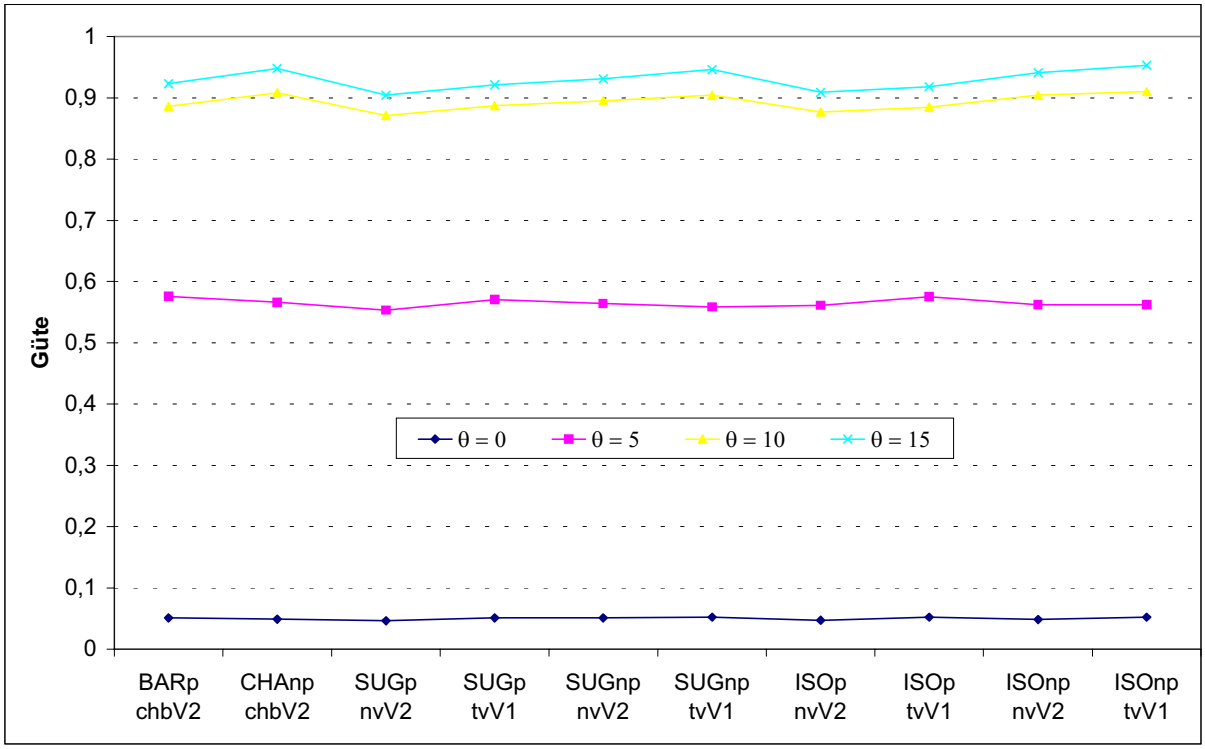
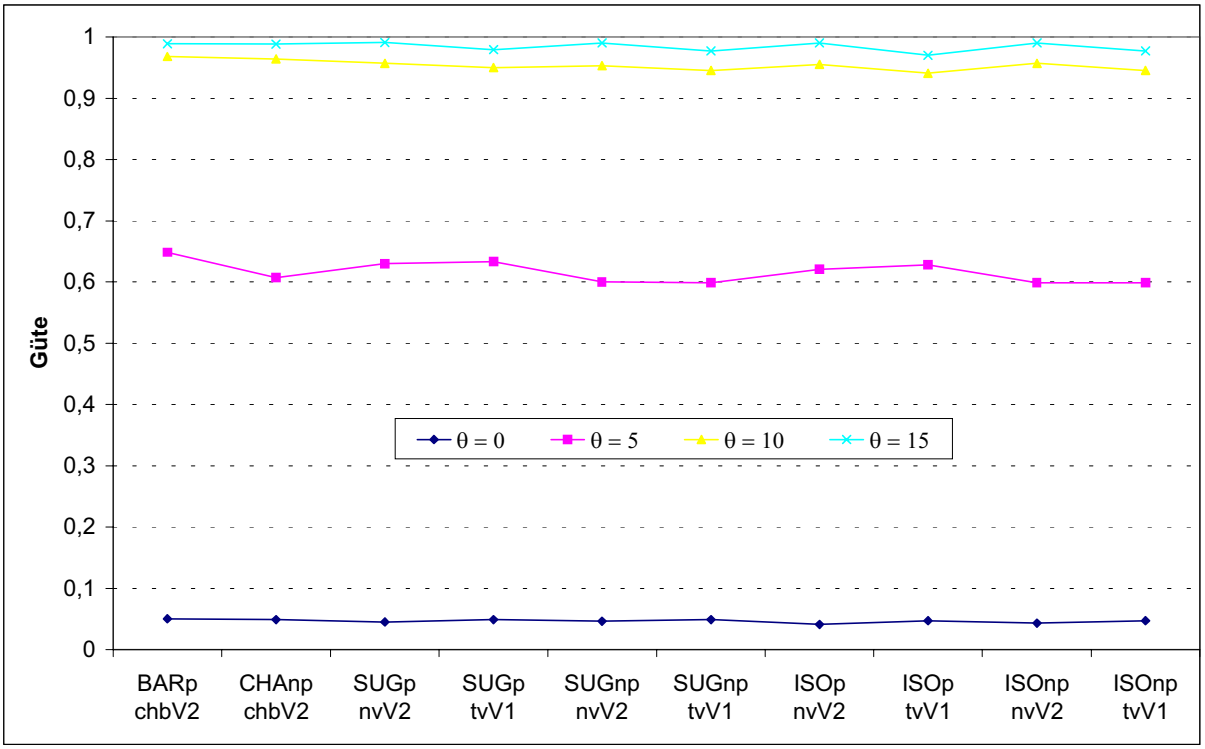


Abbildung 6.26: Güte von Dreistichprobentests (infinite Verteilungen) für a) konkave bzw. b) konvexe Profile ($\alpha = 0{,}05, n = 4, r = 5, \omega = 0{,}5$, Verteilungstyp = RSV)

a)



b)

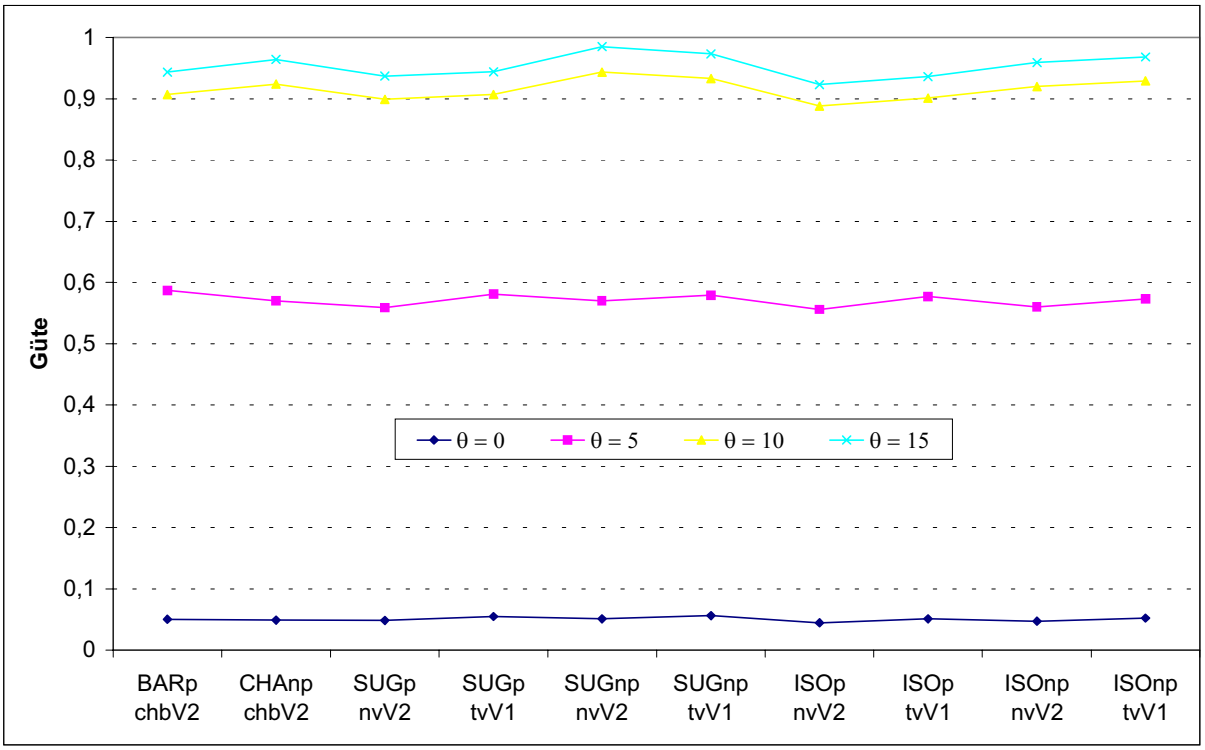


Abbildung 6.27: Güte von Vierstichprobentests (infinite Verteilungen) für a) konkave bzw. b) konvexe Profile ($\alpha = 0{,}05, n = 4, r = 5, \omega = 0{,}5$, Verteilungstyp = RSV)

6.3.3.1 Güte unter der Nullhypothese

BARpchbV2, ISOptvV1, ISOpnvV2

Sowohl der *BARpchbV2*-Test als auch der *ISOptvV1*-Test erweisen sich für $\alpha = 0{,}01$ vor allem bei $k = 2$, aber auch bei $k = 3$ als liberal. Der *BARpchbV2*-Test besitzt dabei noch die bessere Güte. Eine Erhöhung des Stichprobenumfangs von $n = 3$ auf $n = 4$ führt beim *BARpchbV2*-Test zu einer Halbierung der Anzahl der wesentlichen Niveauüberschreitungen. Wird der Stichprobenumfang von $n = 4$ auf $n = 5$ erhöht, verringert sich die Anzahl der Niveauüberschreitungen kaum. Gegenüber $k = 3$ treten im Fall von $k = 2$ bei diesem Niveau fast dreimal soviele Niveauverletzungen beim *BARpchbV2*-Test auf. Die Anzahl der Kategorien (*r*) hat dabei kaum einen Einfluß auf die Güte. Der *ISOpnvV2*-Test ist bei $\alpha = 0{,}01$ extrem konservativ.

Für $\alpha = 0{,}05$ erweist sich der *ISOptvV1*-Test nur für $n = 3$ als zu liberal. Sowohl beim *BARpchbV2*-Test als auch beim *ISOpnvV2*-Test treten selbst bei $n = 3$ fast keine wesentlichen Niveau-verletzungen auf. Die wenigen Niveauüberschreitungen teilen sich dabei zu gleichen Teilen auf $k = 2$ und $k = 3$ auf. Auch hier hat die Anzahl der Kategorien keinen wesentlichen Einfluß.

CHAnpchbV2, ISOnptvV1, ISOnpnvV2

Die t-Verteilungsapproximation erweist sich für $\alpha = 0{,}01$ auch hier als liberal (wenn auch weniger als bei äquidistanten Scores). Der *CHAnpchbV2*-Test und der *ISOnpnvV2*-Test sind bei diesem Niveau konservativ, Niveauüberschreitungen treten nicht auf. Für $\alpha = 0{,}05$ und $n < 5$ treten beim *ISOnptvV1*-Test noch zu viele Niveauverletzungen auf. Der *CHAnpchbV2*-Test und der *ISOnpnvV2*-Test schöpfen das Signifikanzniveau $\alpha = 0{,}05$ gut aus und überschreiten es so gut wie nie.

BARppexoV, BARppmpoV, BARpperoV, CHAnpexoV, CHAnppmpoV, CHAnpperoV

Zwar sind die exakten bedingten Permutationstests nicht so konservativ wie im Zweistichprobenfall, jedoch erweist sich auch im Drei- und Vierstichprobenfall sowohl eine Randomisierung, als auch eine Mid-p-Modifizierung als zweckmäßig. Sowohl das Niveau $\alpha = 0{,}01$ als auch das Niveau $\alpha = 0{,}05$ schöpfen die Mid-p-Tests und die randomisierten Tests am besten von allen Tests aus. Wesentliche Niveauverletzungen treten bei den Mid-p-Tests nur in sehr seltenen Situationen auf. Die Nutzung von äquidistanten Scores oder Rängen beeinflußt das Güteverhalten unter der Nullhypothese nicht.

BARpbomoV, BARpbomV2, CHAnpbomoV, CHAnpbomV2

Für $\alpha = 0{,}01$ halten die Bootstraptests ohne Varianzschätzer sowohl für drei als auch für vier Stichproben das Niveau im wesentlichen ein. Hier treten für $n = 3{,}4$ kaum Niveauverletzungen auf. Mit wachsendem Stichprobenumfang steigt die Anzahl der Niveauverletzungen ($k = 2,\ n = 5$) jedoch an. Die Ursache für diese Anomalie ist die finite Bootstrapverteilung (analog zu Permutationstests, die bei extrem diskreten Verteilungen kaum eine Chance haben, die Nullhypothese abzulehnen). Bootstraptests mit Varianzschätzer neigen für $\alpha = 0{,}01$ eher zu Niveauüberschreitungen. Für $k = 2$ und $\alpha = 0{,}05$ müssen die Tests ohne Varianzschätzer als ungeeignet eingestuft werden, da zu oft deutliche Niveauüberschreitungen auftreten. Gerade für große Werte für *r* ist dies auffällig. Die Nutzung des Varianzschätzers führt zu akzeptablen Fehlern 1. Art. Für $\alpha = 0{,}05,\ k = 3$ und $n > 4$ verbessert sich das Verhalten der vier Tests deutlich. Wesentliche Niveauüberschreitungen treten nicht häufiger als bei den asymptotischen Tests auf. Für $\alpha = 0{,}05,\ n = 3$ sollten die Tests nicht benutzt werden.

Aufgrund der Ergebnisse aller Simulationen kann zusammenfassend gesagt werden:

1. Infinite Verteilungen können bei einem Signifikanzniveau von $\alpha = 0{,}01$ nicht genutzt werden. Vor allem die t-Verteilungsapproximation erweist sich als schlecht.
2. Für $\alpha = 0{,}05$ neigen die Bootstraptests ohne Varianzschätzer am häufigsten zu wesentlichen Niveauüberschreitungen.
3. Die exakten Permutationstests erweisen sich auch für $k = 2{,}3$ als zu konservativ.
4. Die randomisierten Permutationstests sowie die Mid-p-Tests schöpfen das Niveau am besten aus. Wesentliche Niveauüberschreitungen treten praktisch nicht auf.
5. Erst ab $n = 4$ und $\alpha = 0{,}05$ zeigen die meisten Tests ein stabiles Güteverhalten.

In den Abbildungen 6.28 bis 6.33 ist das Güteverhalten beispielhaft für eine rechtsschiefe Verteilung dargestellt.

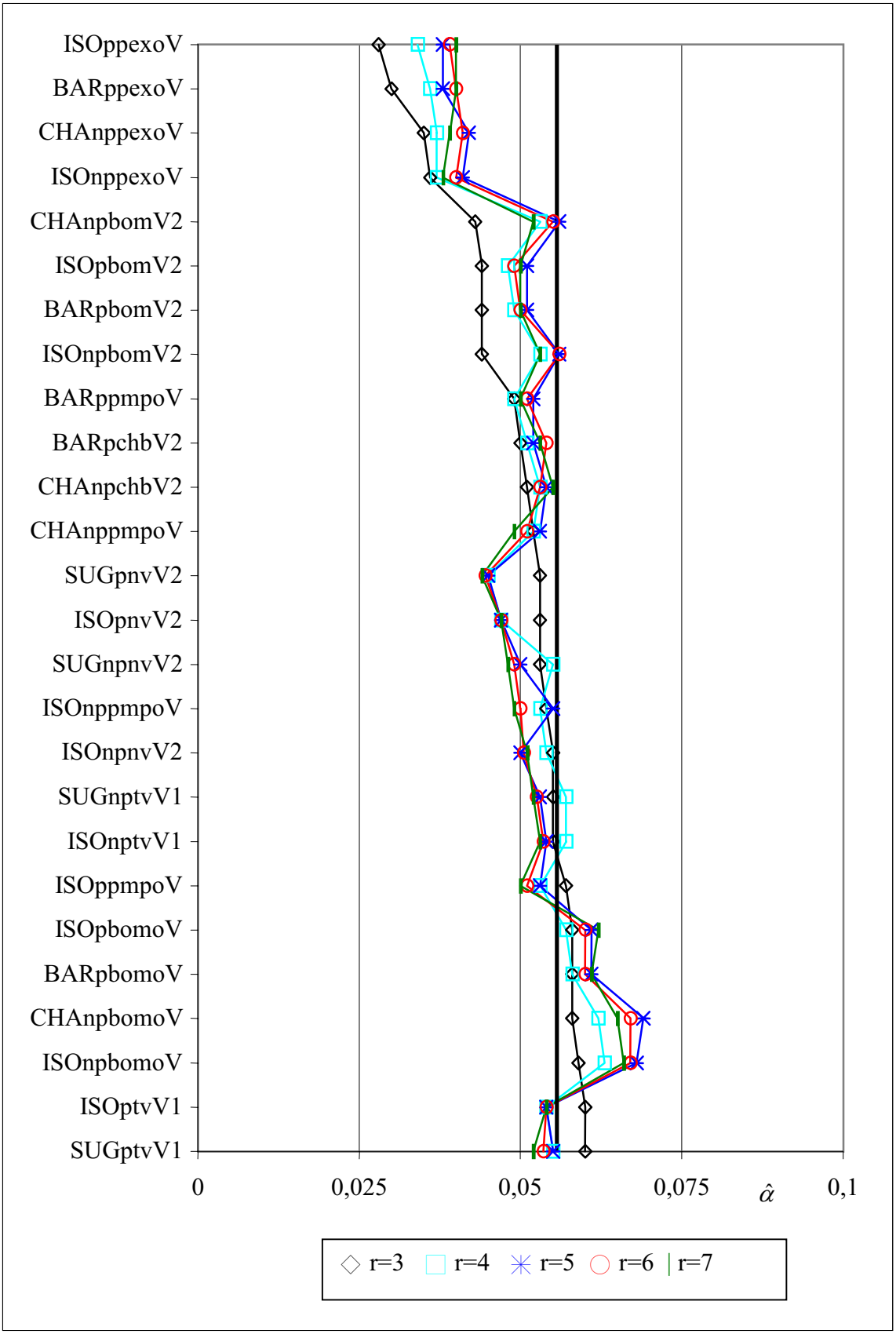


Abbildung 6.28: Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

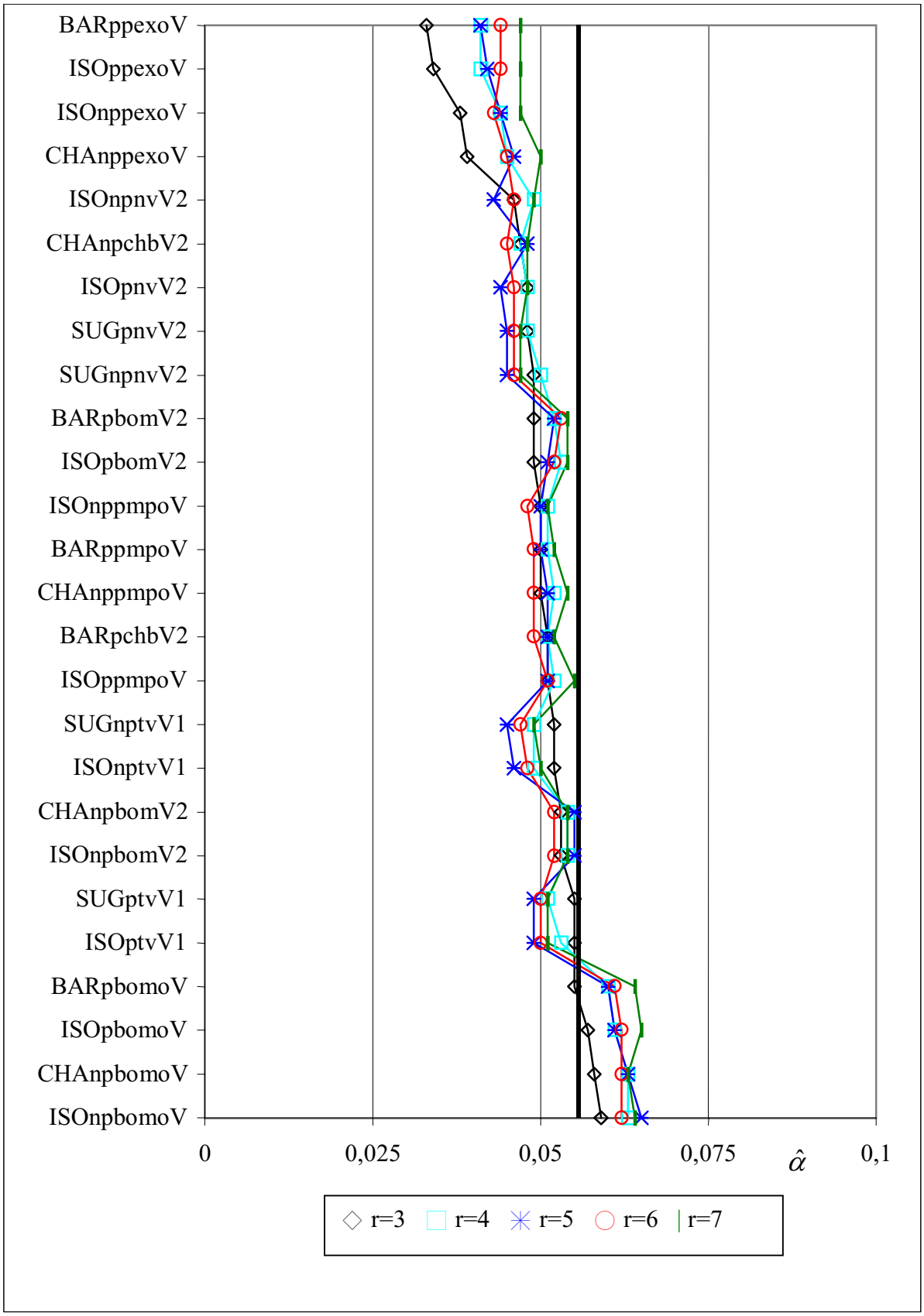


Abbildung 6.29: Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

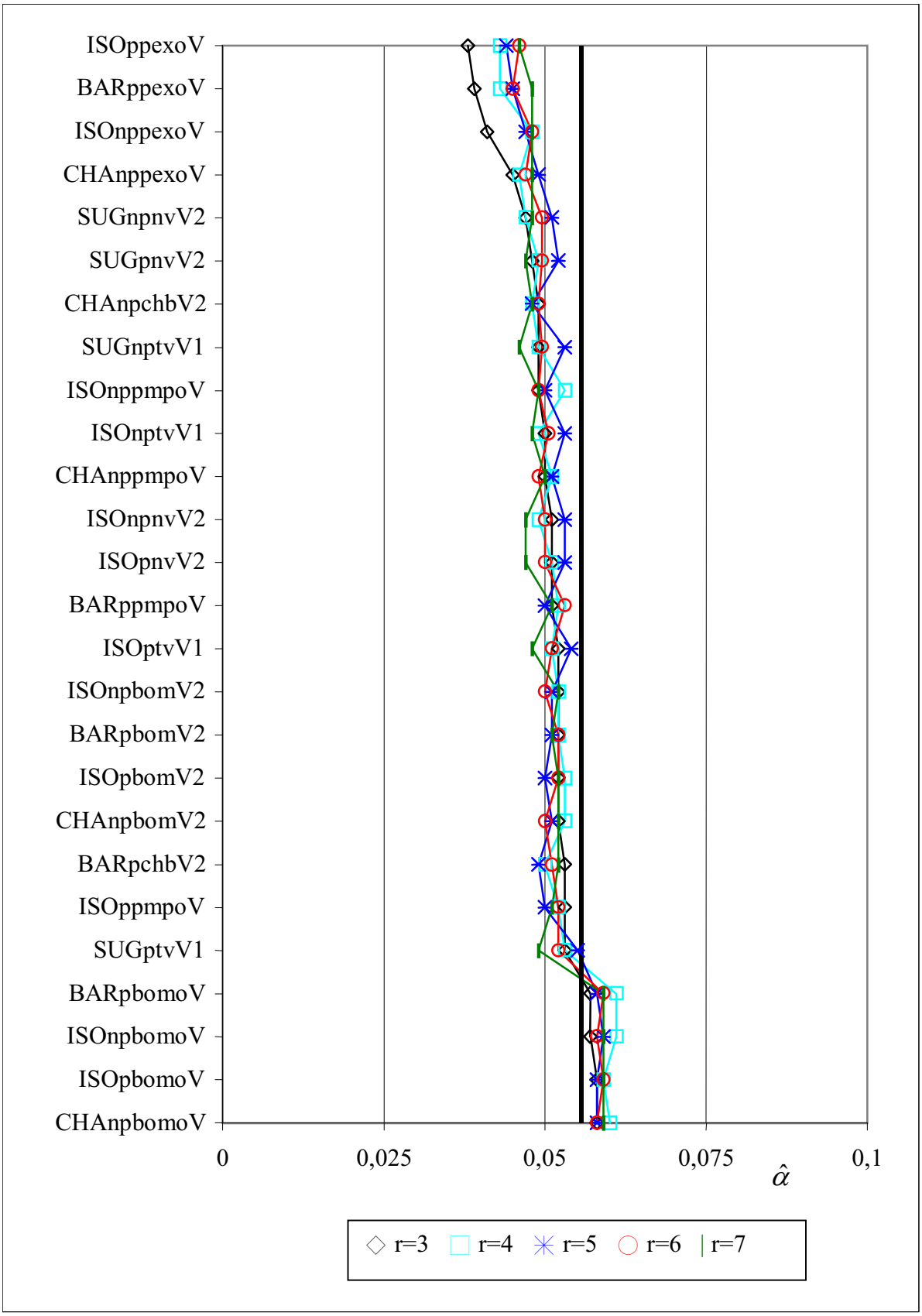


Abbildung 6.30: Güte von Dreistichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

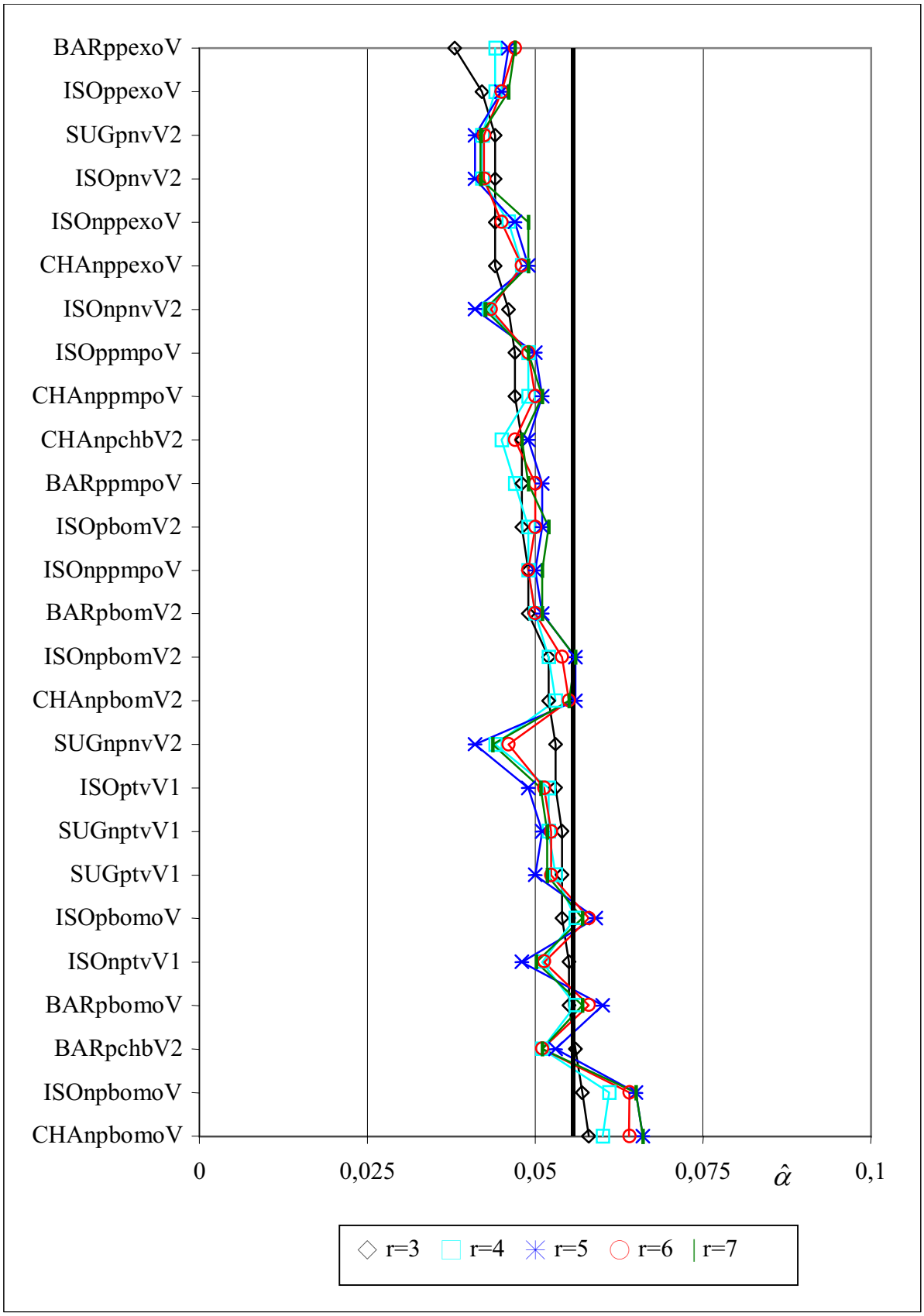


Abbildung 6.31: Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

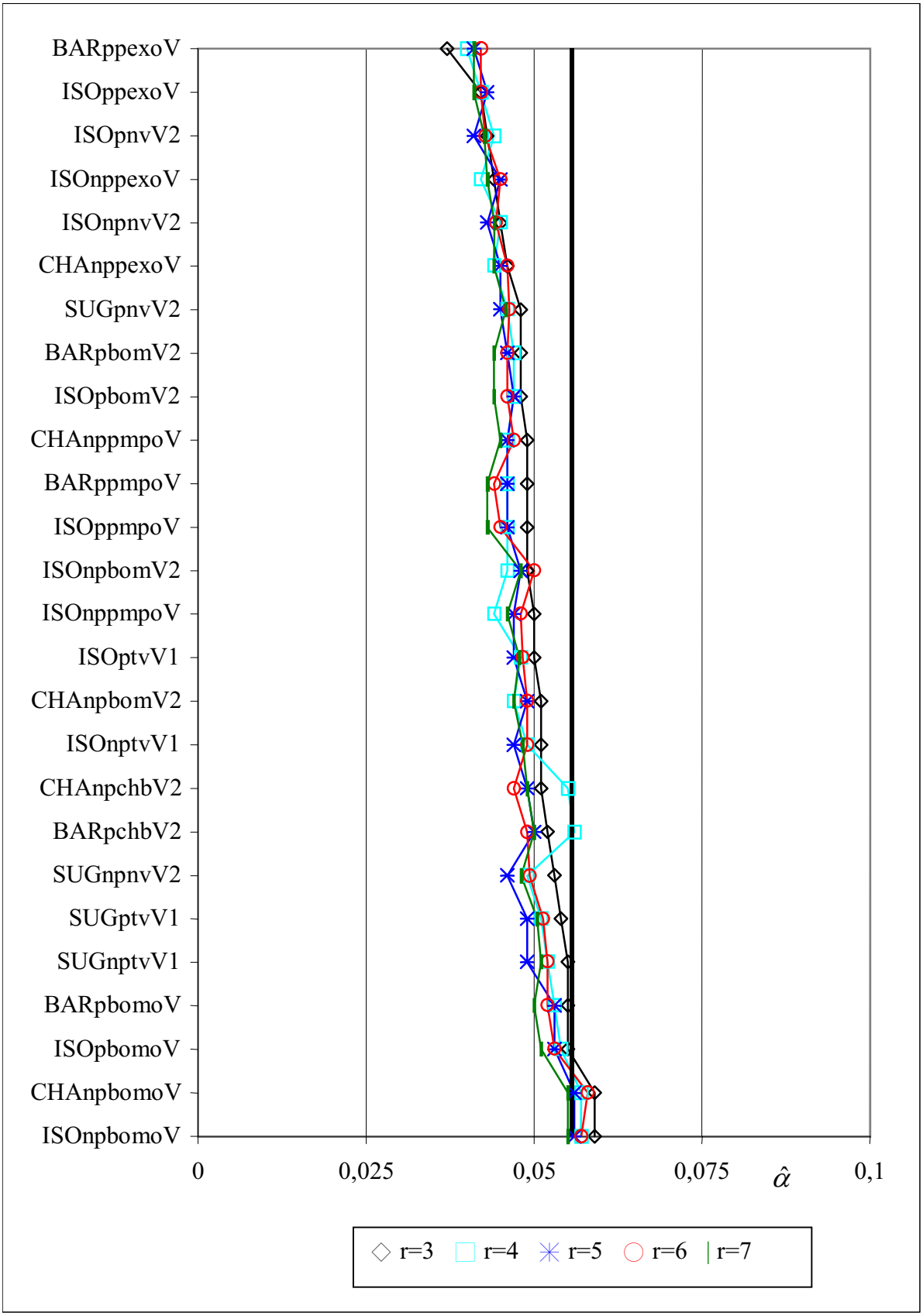


Abbildung 6.32: Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

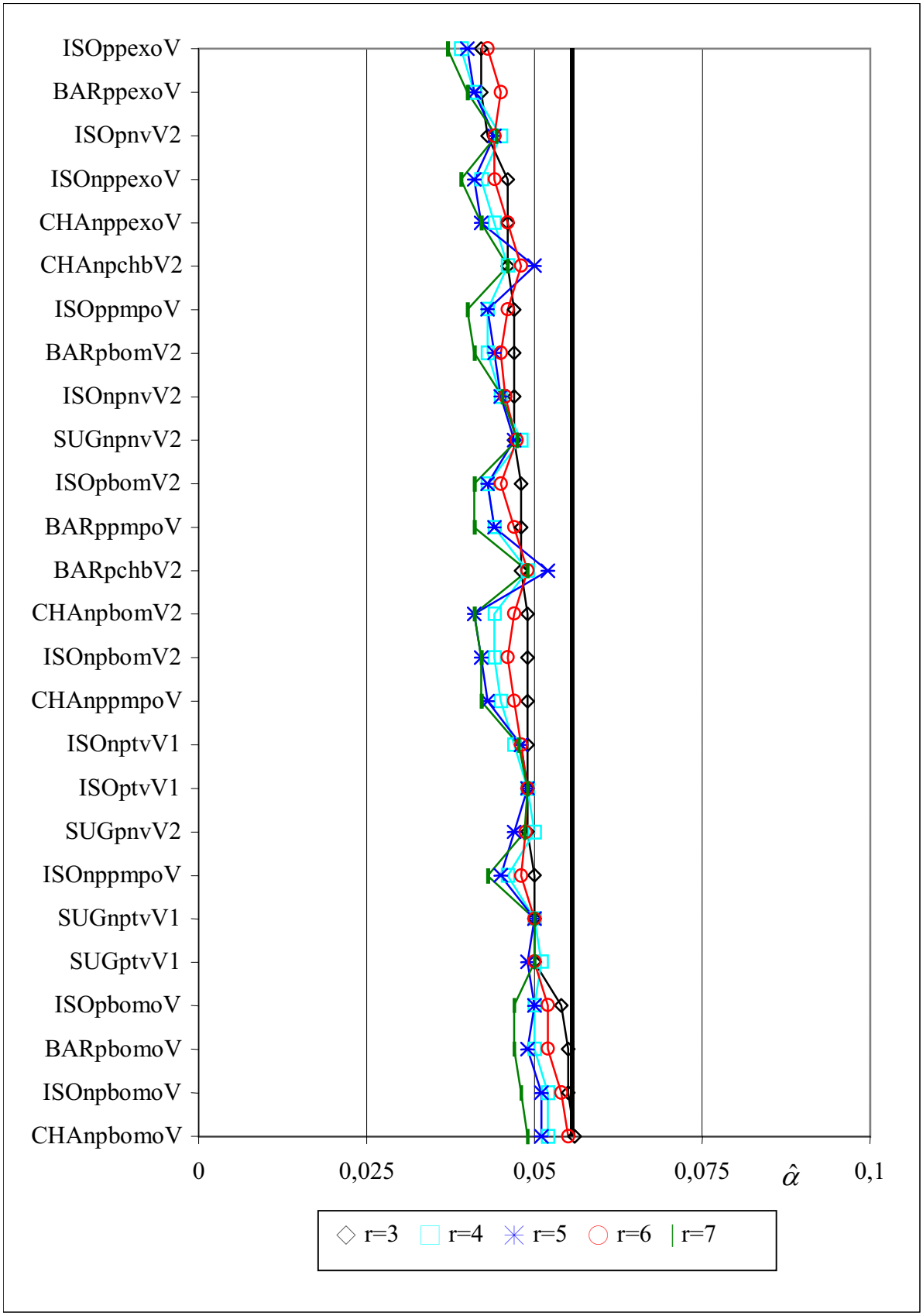


Abbildung 6.33: Güte von Vierstichprobentests in Abhängigkeit von der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 0$, Verteilungstyp = RSV)

6.3.3.2 Güte unter der Alternativhypothese

BAR pchBV2, ISOptvV1, ISOpnvV2

Aufgrund ihres schlechten Güteverhaltens bei $\alpha = 0{,}01$ wird das Verhalten dieser Tests nur für $\alpha = 0{,}05$ diskutiert. Für dieses Niveau zeigen sich klare Vorteile für den *BARpchbV2*-Test und für den *ISOpnvV2*-Test. Die Nutzung der t-Verteilungsapproximation (des Varianzschätzers V1) führt selbst für $n = 5$ noch zu großen Gütenachteilen (0,5-0,8). Erst mit zunehmendem Stichprobenumfang und großem r stellt die t-Verteilung eine Alternative dar. Der Unterschied zwischen dem *BARpchbV2*-Test und dem *ISOpnvV2*-Test ist in der Regel gering (meist kleiner als 0.05). Für kleine r ($r < 6$) ist meist der *BARpchbV2*-Test die bessere Wahl. Besonders für $k = 3$ ist dies deutlich zu erkennen. Für große r ($r \geq 6$) und wachsenden Stichprobenumfang liegen die beiden Tests dichter beieinander. Aufgrund leichter Gütevorteile sollte der *BARpchbV2*-Test bevorzugt werden.

CHAnpchbV2, ISOnptvV1, ISOnpnvV2

Die starke Konservativität des *CHAnpchbV2*-Tests und des *ISOnpnvV2*-Tests für $\alpha = 0{,}01$ führt zu einer vergleichsweise niedrigen Güte unter der Alternative. Der *ISOnptvV1*-Test ist für $\alpha = 0{,}01$ zu liberal. Auch diese drei Tests können daher nur für $\alpha = 0{,}05$ empfohlen werden. Für dieses Niveau sind der *CHAnpchbV2*-Test und der *ISOnpnvV2*-Test dem *ISOnptvV1*-Test überlegen. Die Güteunterschiede zwischen dem *CHAnpchbV2*-Test und dem *ISOnpnvV2*-Test sind marginal (meist kleiner als $0{,}01$). Für $k = 3$ ist dabei der *CHAnpchbV2*-Test fast immer die bessere Wahl. Auch wenn der *CHAnpchbV2*-Test oft der Test mit der minimal höheren Güte ist, sollten diese beiden Tests als gleichwertig angesehen werden.

BARpchbV2, CHAnpchbV2

Der *CHAnpchbV2*-Test ist dem *BARpchbV2*-Tests für $\alpha = 0{,}05$ überlegen. Während die Gütevorteile des *BARpchbV2*-Tests durch $0{,}1$ beschränkt sind, liegt die Güte des *CHAnpchbV2*-Tests teilweise um 0,4 höher. Besonders deutlich ist dies bei $k = 2$. Auch mit steigendem r wird die Überlegenheit des *CHAnpchbV2*-Tests größer. Die Gütekurve des *CHAnpchbV2*-Tests steigt jedoch oftmals etwas langsamer an, als die des *BARpchbV2*-Tests. Die Vorteile des *BARpchbV2*-Tests liegen daher oft im unteren Gütebereich (siehe Tabelle 6.9, 6.10). Der Rangtest stellt gerade bei großen Abweichungen von der Nullhypothese und schiefen Verteilungen die bessere Wahl dar.

BARppexoV, BARppmpoV, BARpperoV, CHAnpexoV, CHAnppmpoV, CHAnpperoV

Auch wenn die Unterschiede zwischen den exakten Permutationstests, den Mid-p-Tests und den randomisierten Tests im allgemeinen nicht mehr so groß sind wie im Zweistichprobenfall, so wird dennoch von den exakten bedingten Permutationstests abgeraten. Sie besitzen stets eine zum Teil deutlich schlechtere Güte als die Mid-p-Tests bzw. die randomisierten Tests. Außerdem treten Situationen auf, bei denen trotz wachsenden Stichprobenumfangs die Güte unter der Alternative fällt (siehe Tabelle 6.8). Ebenso kann eine Erhöhung von $\theta = 10$ auf $\theta = 15$, d. h. eine Verschiebung der Verteilungen in Richtung der Alternative, anders als bei den anderen Tests mit einem Güteverlust verbunden sein (siehe Tabelle 6.9, $r < 7$).

Die Unterschiede zwischen den Mid-p-Tests und den analogen randomisierten Tests sind meist gering. Allerdings existieren auch Konfigurationen, bei denen deutliche Unterschiede auftreten (siehe Tabelle 6.8, 6.9). Dies liegt vor allem an Fällen, wo der beobachtete Wert (s_0) der größte mögliche Wert der Statistik für den aktuellen Datensatz ist und

$$0 < P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) = P(S(\boldsymbol{y}) = s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) \leq 2\alpha$$

bzw.

$$2\alpha < P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) = P(S(\boldsymbol{y}) = s_0 | H_0, \boldsymbol{n}, \boldsymbol{t})$$

gilt. Ersteres kommt für $k = 2, n = 3$ relativ häufig vor. Für diese Datensätze führt der Mid-p-Test im Gegensatz zum randomisierten Test immer zur Ablehnung. Die zweite Ungleichung führt zu der Überlegenheit der randomisierten Tests; der exakte Permutationstest und der Mid-p-Test stimmen dann überein (siehe Tabelle 6.8). Allerdings tritt diese Konstellation wesentlich seltener auf. In den Situationen, in denen der beobachtete Wert nicht der maximal mögliche Wert ist,

$$0 < P(S(\boldsymbol{y}) > s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}) < \alpha \leq P(S(\boldsymbol{y}) \geq s_0 | H_0, \boldsymbol{n}, \boldsymbol{t}),$$

sind die randomisierten Tests und die Mid-p-Tests relativ gleichwertig. Da sowohl für $k = 2$ als auch für $k = 3$ bei den Mid-p-Tests eigentlich fast keine wesentlichen Niveauüberschreitungen auftreten, werden die Mid-p-Tests und die randomisierten Tests empfohlen. Ränge und äquidistante Scores erweisen sich unter den betrachteten Restriktionen als relativ gleichwertig. Gerade für kleine Skalen (kleines *r*) ist dies aufgrund der linearen Transformation $f(a_s) = (a_s - a_1)/(a_r - a_1)$ (d. h. $f(a_1) = 0, f(a_r) = 1$; analog können die Ränge transformiert werden) auch zu erwarten.

Die Gütefunktion eines Tests, der auf äquidistanten Scores basiert, steigt jedoch auch bei $k = 2,3$ häufig etwas steiler an. Das heißt, kleine Abweichungen von der Nullhypothese werden eher mit äquidistanten Scores erkannt. Während die Gütevorteile bei Nutzung äquidistanter Scores meist klein sind ($< 0{,}1$), sind die Gütevorteile bei Nutzung von Rängen in wenigen Fällen deutlich (aber nicht so groß wie bei den Tests mit infiniten Verteilungen). Aufgrund aller Simulationsergebnisse wird zur Nutzung von Rängen geraten. Gerade für große Skalen ($r \geq 6$ oder z. B. für geschätzte ganzzahlige Prozentwerte) sind Rangstatistiken die bessere Wahl. Da der Mid-p-Test ohne ein zusätzliches Zufallsexperiment auskommt, welches in der Praxis auf Ablehnung stößt, wird zu einem Mid-p-Test geraten.

BARpbomoV, BARpbomV2, CHAnpbomoV, CHAnpbomV2

Die konservative Powerschätzung führt bei den Tests mit Varianzschätzer zu einer mangelhaften Güte unter der Alternative. Oftmals ist der Unterschied zwischen *BARpbomoV* und *BARpbomV2* für $\theta = 5$ gering ($< 0{,}01$) und wächst auf bis zu $0{,}2 - 0{,}4$ für $\theta = 10{,}15$ (alle anderen Parameter jeweils fest). Dies ist aufgrund der zunehmenden Anzahl von Bindungen und der damit verbundenen höheren Wahrscheinlichkeit nichtpositiver Varianzschätzer erklärbar. Erst für $\alpha = 0{,}05$, $k = 3$ und $n > 3$ verbessert sich das Verhalten der Tests deutlich; die Varianzschätzer limitieren die Güte nicht mehr so stark. Insgesamt ist ihre Güte jedoch zu niedrig. Sie können unter den hier betrachteten Restriktionen nicht empfohlen werden. Trotz der zum Teil wesentlichen Niveauüberschreitungen der Bootstraptests ohne Varianzschätzer für $\alpha = 0{,}01,\ 0{,}05$ besitzen sie keine höhere Güte als z. B. die Mid-p-Tests. Ob äquidistante Scores oder Ränge benutzt werden, hat auf diese Aussage keinen Einfluß. Der Vergleich des *BARpbomoV*-Tests (*BARpbomV2*-Tests) und des *CHAnpbomoV*-Tests (*CHAnpbomV2*-Tests) ist relativ analog zum Verhalten der analogen Tests mit infiniten Verteilungen. Keiner der Tests ist stets besser. Die deutlicheren Gütevorteile besitzen auch hier die Rangtests, während die Tests mit äquidistanten Scores für kleine Abweichungen von der Nullhypothese die bessere Wahl sind.

Vergleich *BARppmpoV* vs. *BARpchbV2* bzw. *CHAnppmpoV* vs. *CHAnpchbV2*

Dieser Vergleich ist in den Tabellen 6.9 und 6.10 beispielhaft für $\alpha = 0{,}05$ dargestellt. Gerade für $k = 2$ und $n = 3{,}4$ erweisen sich die Mid-p-Tests als die mächtigeren Tests. Während für kleine Skalen und nicht zu großen Abweichungen von der Nullhypothese der *BARpchbV2*-Test und der *CHAnpchbV2*-Test noch relativ gleichmächtig zu den analogen Mid-p-Tests sind, erweisen sich die Mid-p-Tests für große r und große θ ($\theta = 10{,}15$) überlegen. Die

Unterschiede nehmen jedoch bei $k = 3$ und $n = 4,5$ ab. Da sich für $\alpha = 0,01$ die infiniten Verteilungsapproximationen als schlecht erweisen und aufgrund des besseren Güteverhaltens unter der Nullhypothese, sind die Mid-p-Tests die bessere Wahl.

Vergleich *BARpbomoV* vs. *BARppmpoV* bzw. *CHAnpbomoV* vs. *CHAnppmpoV*

Die Bootstraptests überschreiten zu oft das vorgegebene Signifikanzniveau. Zudem besitzen sie unter der Alternative keine höhere Güte als die Mid-p-Tests. Anders als im Zweistichprobenfall übertreffen die Bootstraptests die Mid-p-Tests nicht einmal für das Niveau $\alpha = 0,01$ (siehe Tabelle 6.11). Insgesamt sind die Mid-p-Tests den Bootstraptests vorzuziehen.

	r = 3			**r = 4**		
	$n = 3$	$n = 4$	$n = 5$	$n = 3$	$n = 4$	$n = 5$
unbedingte Güte[1]	0,570	0,754	0,590	0,709	0,866	0,758
CHAnppexoV	0,574	0,752	0,598	0,714	0,866	0,757
CHAnppmpoV	0,575	0,753	0,869	0,714	0,866	0,942
CHAnpperoV	0,734	0,832	0,894	0,806	0,902	0,947
CHAnpbomoV	0,574	0,752	0,868	0,715	0,866	0,941
CHAnpbomV2	0,337	0,618	0,802	0,497	0,75	0,894
CHAnpchbV2	0,565	0,761	0,864	0,71	0,871	0,941

Tabelle 6.8: Chacko-Test in Abhängigkeit von r und n ($\alpha = 0,05, k = 3, \theta = 15$, $\omega = 0,5$, konkave Profile, Verteilungstyp = LSV)

[1] berechnete exakte unbedingte Güte des *CHAnppexoV*-Tests

r	θ	*BAR-ppexoV*	*BAR-ppmpoV*	*BAR-pperoV*	*BAR-pchbV2*	*CHA-nppexoV*	*CHA-nppmpoV*	*CHA-npperoV*	*CHA-npchbV2*
3	0	0,030	0,049	0,052	0,05	0,035	0,052	0,052	0,051
	5	0,358	0,502	0,472	0,474	0,370	0,509	0,474	0,463
	10	0,453	0,828	0,707	0,815	0,454	0,830	0,708	0,812
	15	0,452	0,864	0,731	0,854	0,452	0,864	0,731	0,853
4	0	0,036	0,049	0,050	0,051	0,037	0,052	0,051	0,053
	5	0,425	0,509	0,497	0,48	0,425	0,505	0,486	0,468
	10	0,580	0,878	0,773	0,802	0,585	0,882	0,776	0,854
	15	0,577	0,924	0,801	0,842	0,577	0,925	0,802	0,916
5	0	0,038	0,052	0,053	0,052	0,042	0,053	0,053	0,054
	5	0,457	0,520	0,513	0,493	0,451	0,504	0,494	0,475
	10	0,661	0,902	0,814	0,84	0,668	0,907	0,819	0,876
	15	0,658	0,952	0,844	0,88	0,659	0,953	0,844	0,943
6	0	0,040	0,051	0,053	0,054	0,041	0,051	0,051	0,053
	5	0,476	0,522	0,518	0,491	0,453	0,496	0,490	0,469
	10	0,714	0,909	0,837	0,863	0,720	0,915	0,842	0,887
	15	0,713	0,966	0,870	0,917	0,713	0,966	0,871	0,962
7	0	0,040	0,050	0,051	0,053	0,039	0,049	0,048	0,055
	5	0,485	0,521	0,519	0,504	0,455	0,491	0,486	0,477
	10	0,751	0,914	0,854	0,869	0,758	0,919	0,859	0,894
	15	0,752	0,973	0,888	0,914	0,753	0,973	0,889	0,966
8	0	0,041	0,049	0,050	0,052	0,040	0,048	0,049	0,054
	5	0,491	0,523	0,520	0,506	0,456	0,490	0,487	0,481
	10	0,783	0,919	0,870	0,876	0,789	0,923	0,874	0,898
	15	0,785	0,978	0,903	0,92	0,786	0,978	0,904	0,971
9	0	0,042	0,050	0,050	0,054	0,042	0,051	0,051	0,052
	5	0,498	0,528	0,526	0,502	0,457	0,486	0,484	0,475
	10	0,806	0,922	0,880	0,88	0,810	0,924	0,882	0,896
	15	0,809	0,982	0,915	0,928	0,809	0,983	0,915	0,979

Tabelle 6.9: Vergleich finiter und infiniter Verteilungen anhand des Bartholomew-Tests und des Chacko-Tests ($\alpha = 0{,}05$, $k = 2$, $n = 3$, $\omega = 0{,}5$,konkave Profile, Verteilungstyp = RSV)

r	θ	*BAR-ppexoV*	*BAR-ppmpoV*	*BAR-pperoV*	*BAR-pchbV2*	*CHA-nppexoV*	*CHA-nppmpoV*	*CHA-npperoV*	*CHA-npchbV2*
3	0	0,038	0,048	0,049	0,056	0,044	0,047	0,049	0,048
	5	0,472	0,515	0,520	0,53	0,496	0,524	0,527	0,457
	10	0,817	0,825	0,854	0,825	0,822	0,829	0,858	0,801
	15	0,862	0,863	0,899	0,856	0,862	0,863	0,899	0,855
4	0	0,044	0,047	0,049	0,051	0,048	0,049	0,050	0,045
	5	0,509	0,537	0,540	0,526	0,503	0,526	0,528	0,481
	10	0,871	0,881	0,895	0,878	0,874	0,883	0,898	0,856
	15	0,921	0,921	0,941	0,922	0,920	0,921	0,941	0,921
5	0	0,046	0,051	0,051	0,053	0,049	0,051	0,051	0,049
	5	0,528	0,552	0,552	0,535	0,505	0,522	0,523	0,489
	10	0,897	0,909	0,915	0,906	0,899	0,908	0,917	0,878
	15	0,951	0,952	0,963	0,948	0,950	0,951	0,962	0,947
6	0	0,047	0,050	0,050	0,051	0,048	0,050	0,050	0,047
	5	0,540	0,557	0,558	0,534	0,505	0,516	0,518	0,483
	10	0,915	0,924	0,928	0,907	0,911	0,918	0,925	0,892
	15	0,965	0,966	0,973	0,94	0,965	0,965	0,973	0,956
7	0	0,047	0,049	0,049	0,051	0,049	0,051	0,051	0,048
	5	0,550	0,563	0,564	0,538	0,507	0,516	0,516	0,485
	10	0,922	0,931	0,932	0,911	0,917	0,924	0,930	0,892
	15	0,973	0,974	0,979	0,949	0,972	0,973	0,979	0,966
8	0	0,047	0,050	0,050	0,049	0,049	0,050	0,050	0,046
	5	0,553	0,564	0,563	0,539	0,499	0,507	0,508	0,483
	10	0,928	0,936	0,938	0,919	0,922	0,929	0,934	0,899
	15	0,979	0,980	0,983	0,959	0,978	0,979	0,983	0,973
9	0	0,047	0,049	0,049	0,051	0,048	0,049	0,049	0,047
	5	0,554	0,564	0,564	0,542	0,496	0,503	0,504	0,479
	10	0,933	0,939	0,940	0,922	0,925	0,930	0,935	0,909
	15	0,983	0,983	0,985	0,964	0,982	0,982	0,985	0,977

Tabelle 6.10: Vergleich finiter und infiniter Verteilungen anhand des Bartholomew-Tests und des Chacko-Tests ($\alpha = 0{,}05$, $k = 3$, $n = 3$, $\omega = 0{,}5$, konkave Profile, Verteilungstyp = RSV)

r	θ	BAR-pbomoV	BAR-pbomV2	BAR-ppmpoV	BAR-ppexoV	CHA-npbomoV	CHA-npbomV2	CHA-nppmpoV	CHA-nppexoV
3	0	0,009	0,008	0,011	0,002	0,012	0,011	0,011	0,002
	15	0,449	0,002	0,449	0	0,45	0,002	0,449	0
4	0	0,008	0,009	0,01	0,003	0,01	0,009	0,01	0,004
	15	0,435	0,005	0,572	0,002	0,574	0,006	0,572	0,002
5	0	0,008	0,011	0,011	0,005	0,011	0,012	0,011	0,005
	15	0,544	0,009	0,652	0,003	0,653	0,01	0,652	0,003
6	0	0,009	0,01	0,011	0,005	0,011	0,011	0,011	0,006
	15	0,535	0,012	0,704	0,005	0,706	0,014	0,704	0,005
7	0	0,01	0,012	0,012	0,005	0,01	0,013	0,011	0,006
	15	0,522	0,017	0,742	0,008	0,745	0,019	0,742	0,008
8	0	0,007	0,01	0,01	0,006	0,01	0,011	0,01	0,007
	15	0,549	0,021	0,774	0,009	0,776	0,024	0,774	0,009
9	0	0,009	0,011	0,011	0,006	0,009	0,012	0,01	0,007
	15	0,571	0,025	0,797	0,011	0,799	0,028	0,797	0,011

Tabelle 6.11: Vergleich finiter Verteilungen anhand der Bartholomew-Statistik und der Chacko- Statistik ($\alpha = 0{,}01, k = 2, n = 3, \omega = 0{,}5$, konkave Profile, Verteilungstyp = RSV)

Unter den betrachteten Restriktionen werden aufgrund der Simulationsergebnisse folgende Schlußfolgerungen gezogen und Empfehlungen gegeben:

1. Es gibt keinen Test der durchweg überzeugt. Vor allem bei $n = 3$ und $\alpha = 0{,}01$ fällt es schwer, einen Test zu empfehlen.
2. Werden die Versuchsbedingungen relativ optimal gehalten (bei Simulationen treten ja keine Confounder auf), zeigen die Simulationen, daß selbst mit den sehr kleinen Fallzahlen gute Testergebnisse erreichbar sind.
3. Die Mid-p-Tests *CHAnppmpoV* und *BARppmpoV* erweisen sich unter den betrachteten Restriktionen ($\alpha = 0{,}01,\ 0{,}05,\ 0{,}10,\ k = 2,3, n = 3,4,5, r = 3,\ldots,9$) insgesamt als die beste Wahl. Auch wenn bei den Mid-p-Tests theoretisch Niveauüberschreitungen auftreten können, sind solche in der Praxis sehr selten. Die Wahl der Scores spielt bei diesen Tests keine wesentliche Rolle, da sich die Güten der Tests, die auf Rängen bzw. auf äquidistanten Scores beruhen, kaum unterscheiden. Wenn möglich, sollte eine nicht zu

grobe Skala benutzt werden. Bedenken hinsichtlich schwachbesetzter Zellen (leere Zellen) sind bei diesen Tests überflüssig.

4. Die Gütefunktion des *ISOnppmpoV*-Tests (*ISOppmpoV*-Tests) stimmt gut mit der Gütefunktion des *CHAnppmpoV*-Tests (*BARppmpoV*-Tests) überein. Alle vier Tests können daher empfohlen werden. Aufgrund aller betrachteten Konstellationen wird jedoch zu dem Rangtest *CHAnppmpoV* geraten.
5. Die Bootstraptests ohne Varianzschätzer besitzen unter der Alternative zwar eine gleich hohe Güte wie die Mid-p-Tests, neigen jedoch häufiger zu Niveauverletzungen. Bootstraptests mit Varianzschätzer sind im wesentlichen erst für $n > 4$ zu empfehlen und sollten gerade für kleine Skalen gemieden werden. Auch bei den Bootstraptests ist der Unterschied zwischen den *ISOnpbomoV*-Tests (*ISOpbomoV*-Tests) und den *CHAnpbomoV*-Tests (*BARpbomoV*-Tests) gering.
6. Die Tests, die auf multivariaten Verteilungen beruhen, sind nur für $\alpha = 0{,}05, n > 4$ zu empfehlen. Bei diesem Niveau entsprechen ihre Güten in vielen Fällen denen der Mid-p-Tests. Vereinzelt besitzen sie sogar eine höhere Güte.

Die deutlichsten Effekte hängen mit den Stichprobenumfängen und mit der Mächtigkeit der Skala zusammen. Außerdem fallen die exakten Permutationstests, die Bootstraptests mit Varianzschätzer und die t-Verteilungsapproximation auf. Um diese Effekte nochmal zu verdeutlichen, wurden die Parameter in den folgenden Abbildungen entsprechend gewählt. Die Grafiken basieren stets auf dem Niveau 0,05 und beschränken sich auf die Ergebisse für $r = 3,4,5,6,7$. Die Tests sind in den Abbildungen stets nach dem Güteverhalten bei $r = 3$ geordnet.

In den Abbildungen 6.39-6.44 ist beispielhaft das Verhalten der Tests, die auf multivariaten infiniten Verteilungen basieren, dargestellt. Diese Abbildungen verdeutlichen die zum Teil geringen Unterschiede, aber auch die Vorteile der äquidistanten Scores bei kleinen Abweichungen von der Nullhypothese.

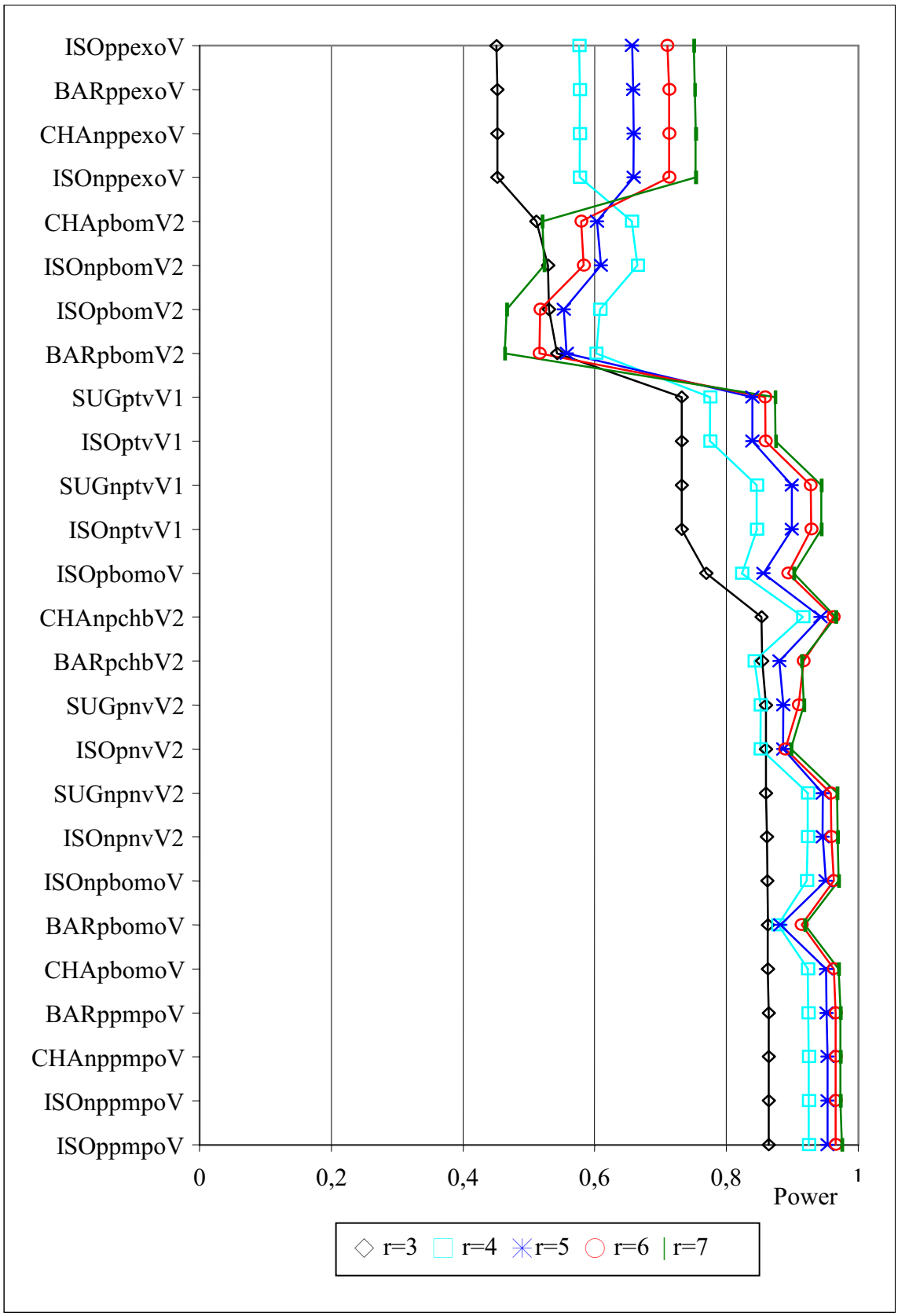


Abbildung 6.34: Power von Dreistichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

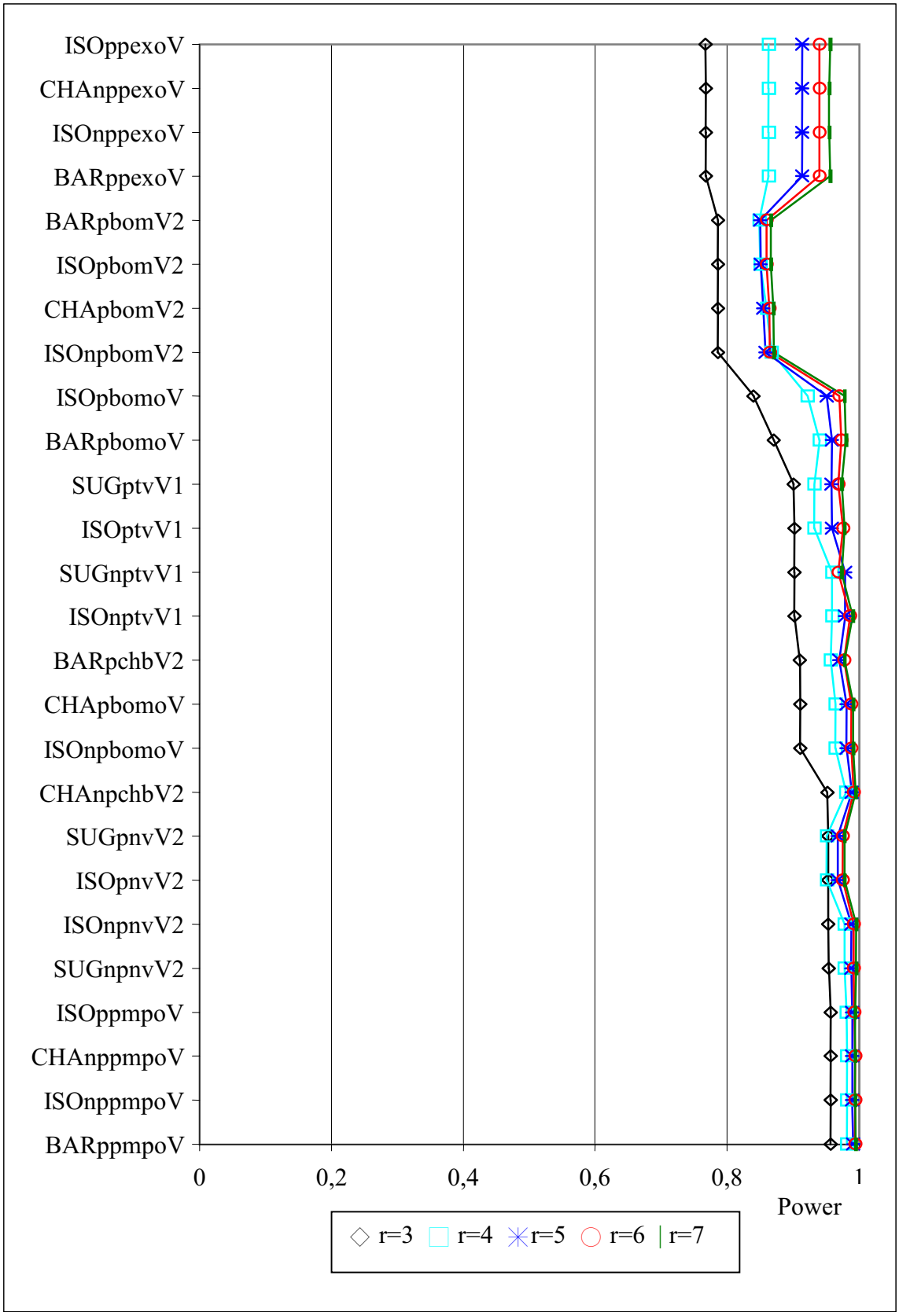


Abbildung 6.35: Power von Dreistichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

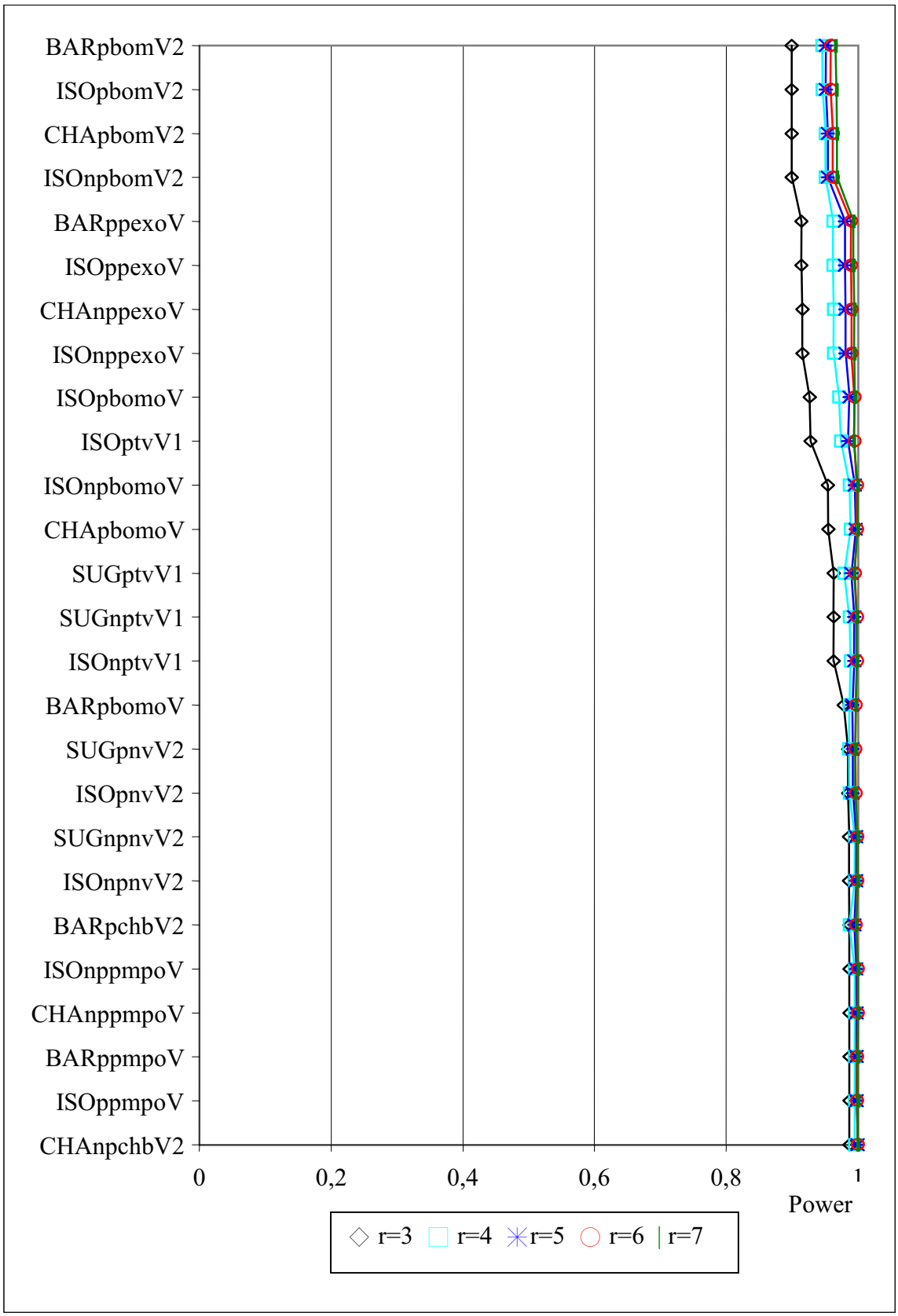


Abbildung 6.36: Power von Dreistichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

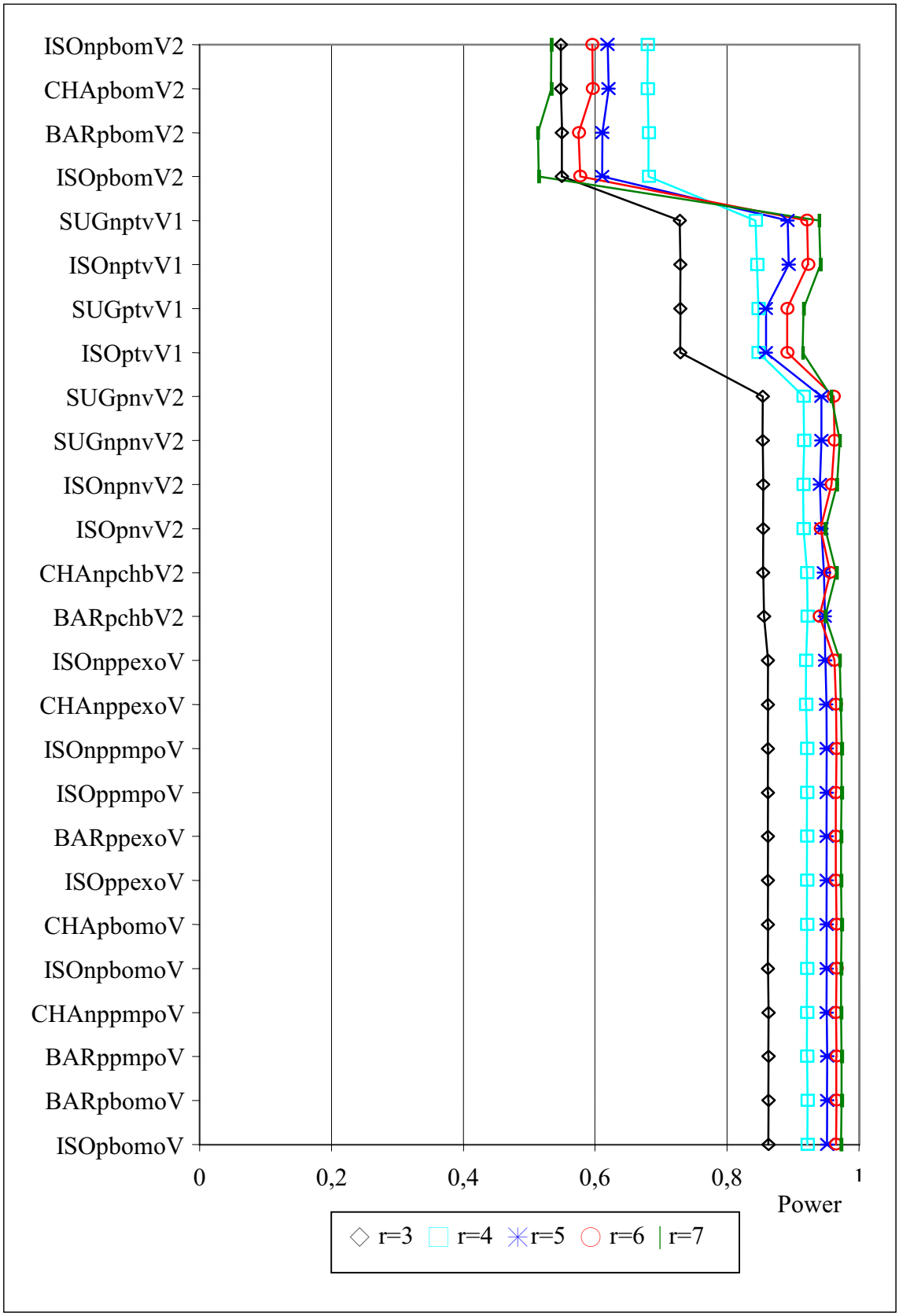


Abbildung 6.37: Power von Vierstichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 3$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

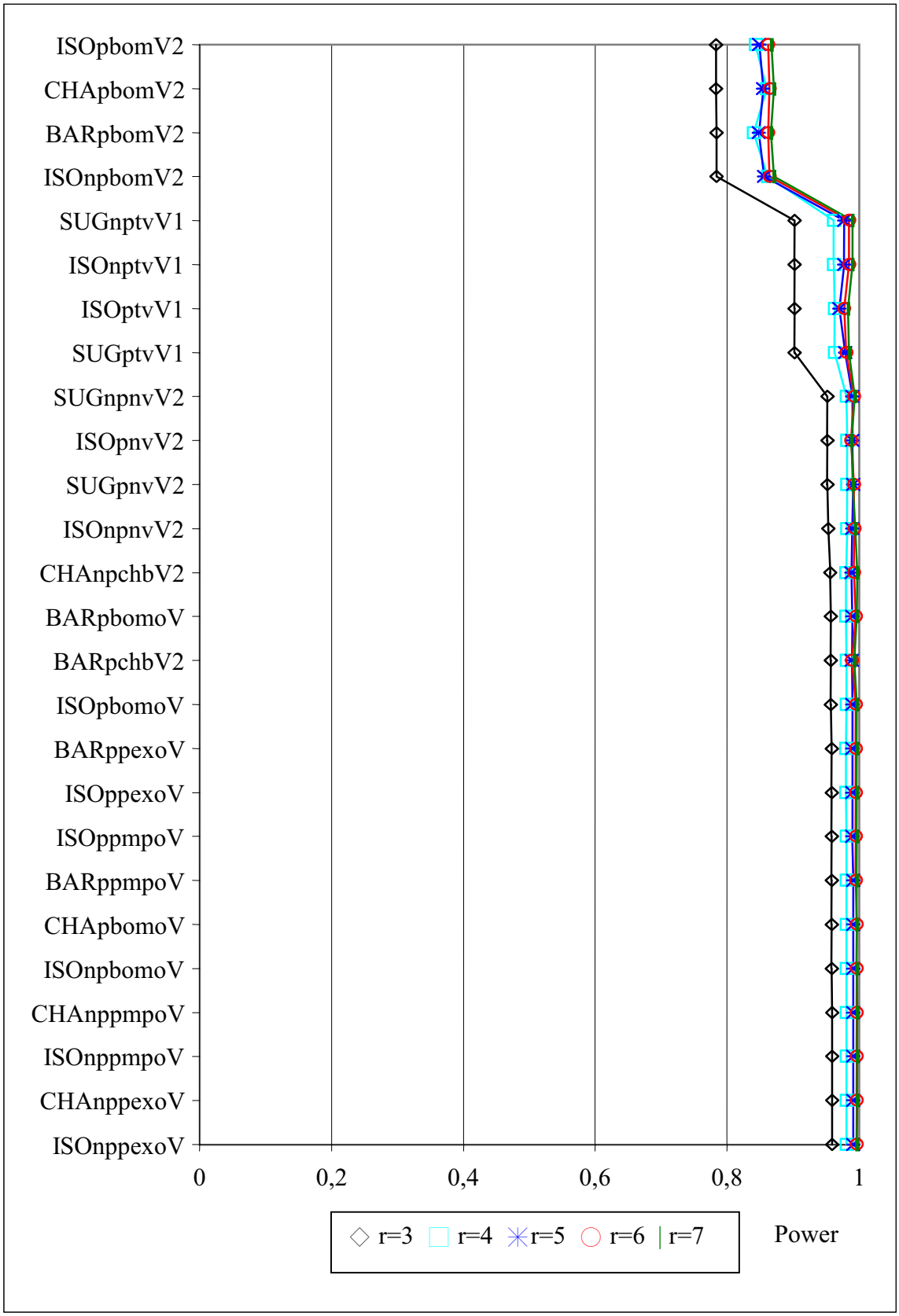


Abbildung 6.38: Power von Vierstichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 4$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

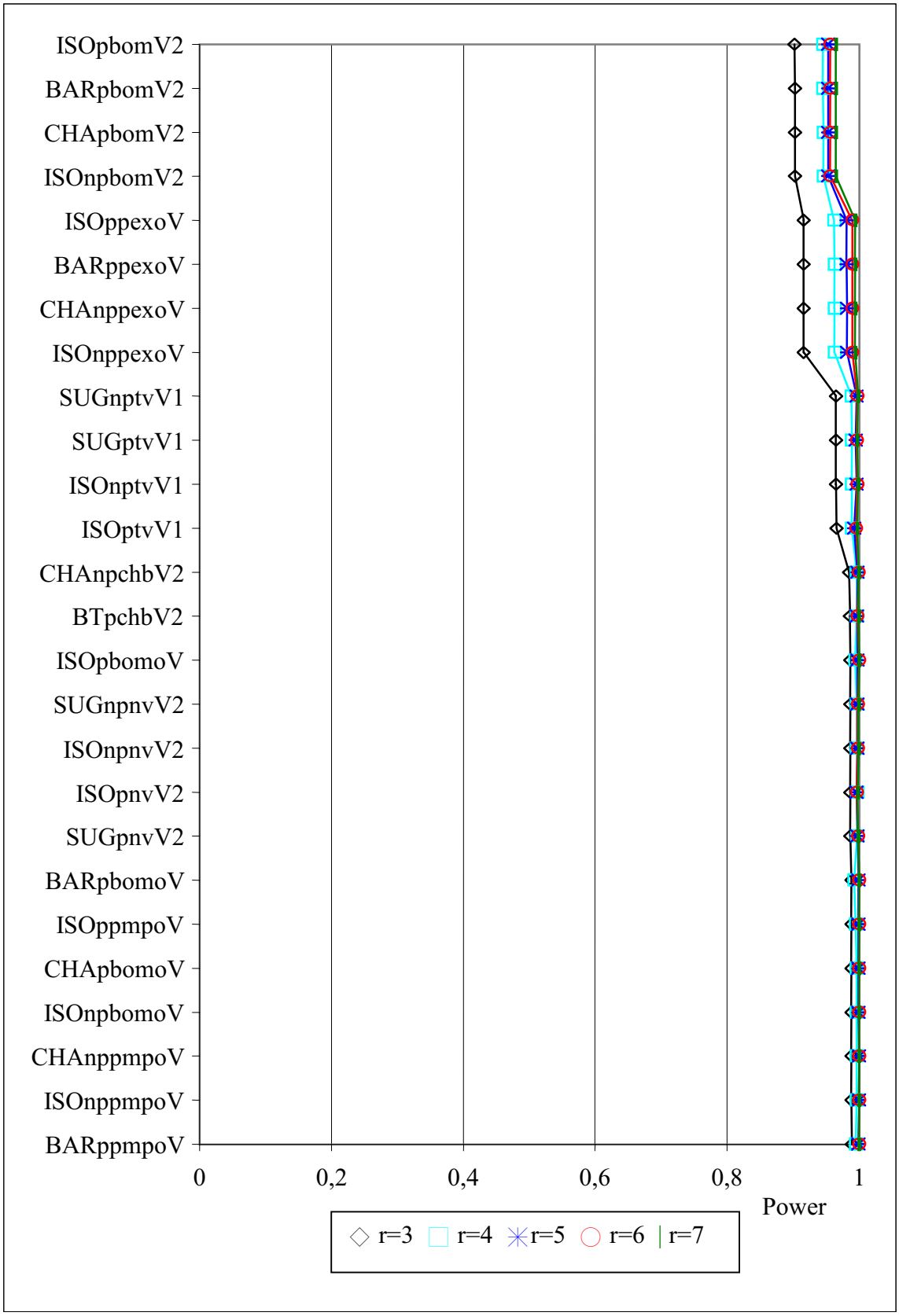


Abbildung 6.39: Power von Vierstichprobentests in Abhängigkeit von der Verteilung und der Anzahl der Skalenpunkte (r) ($\alpha = 0{,}05$, $n = 5$, $\omega = 0{,}5$, $\theta = 15$, Verteilungstyp = RSV)

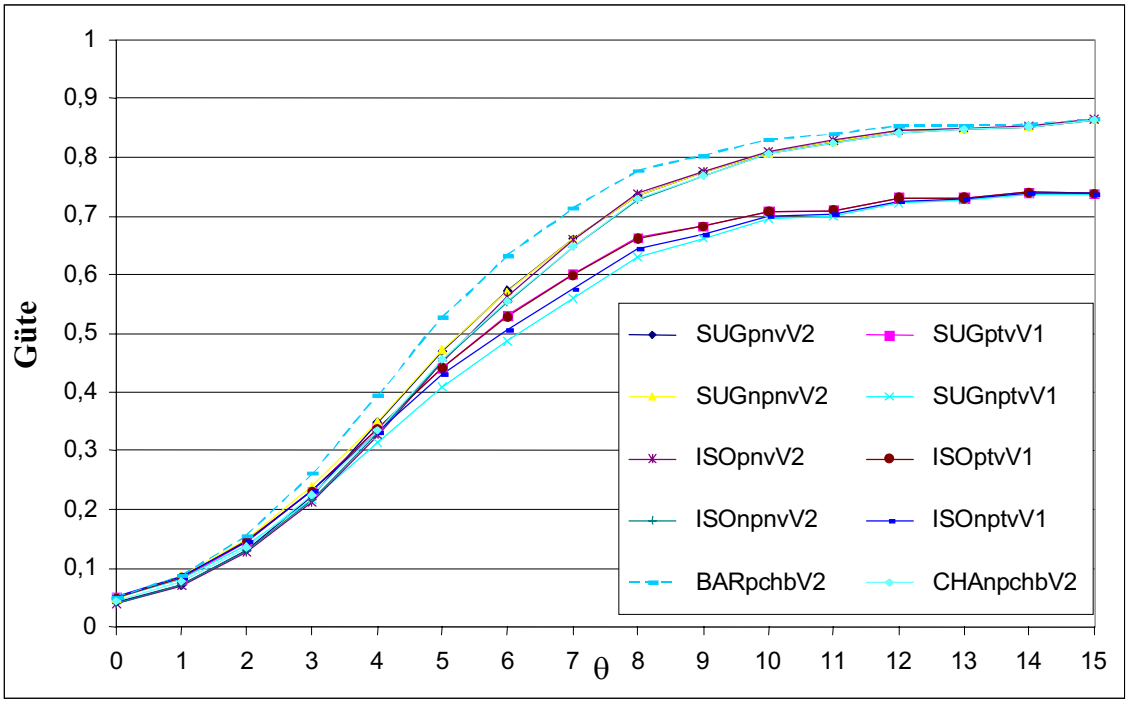


Abbildung 6.40: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 3$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV)

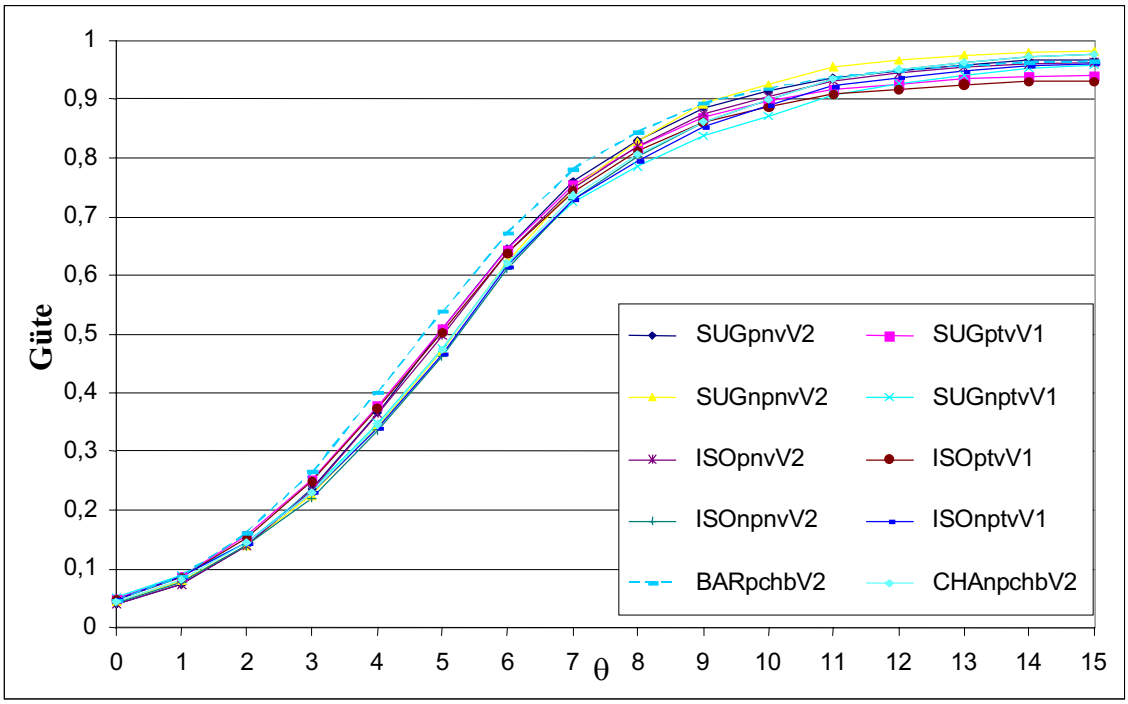


Abbildung 6.41: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 3$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV)

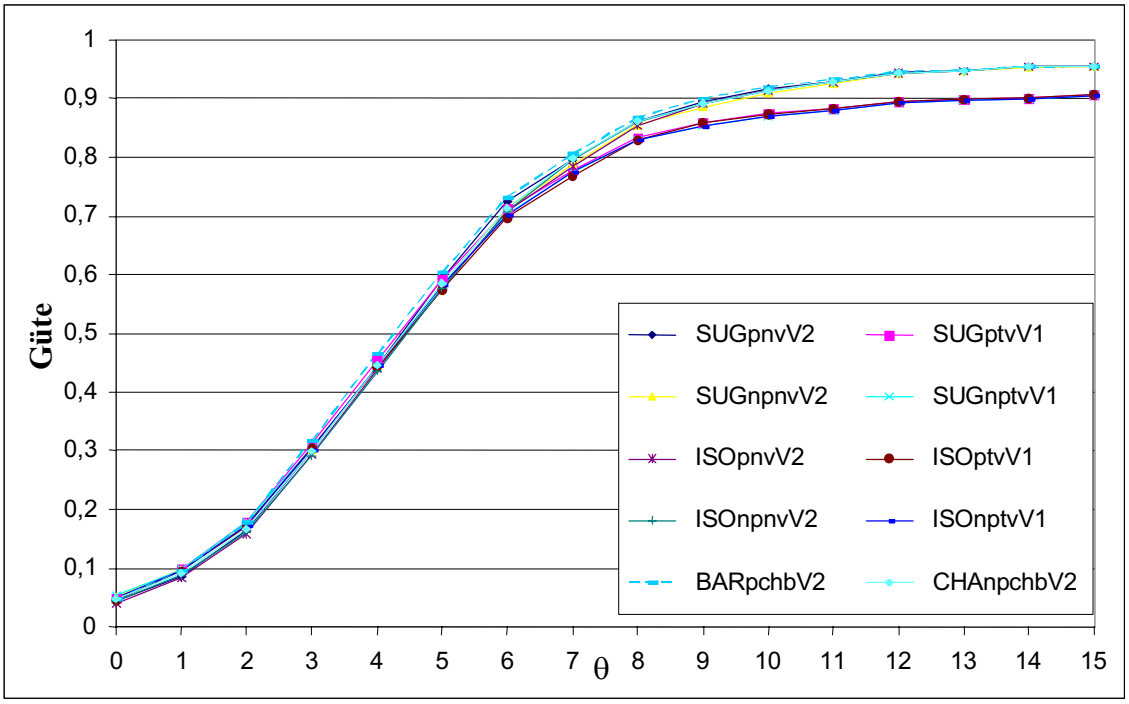


Abbildung 6.42: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 4$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV)

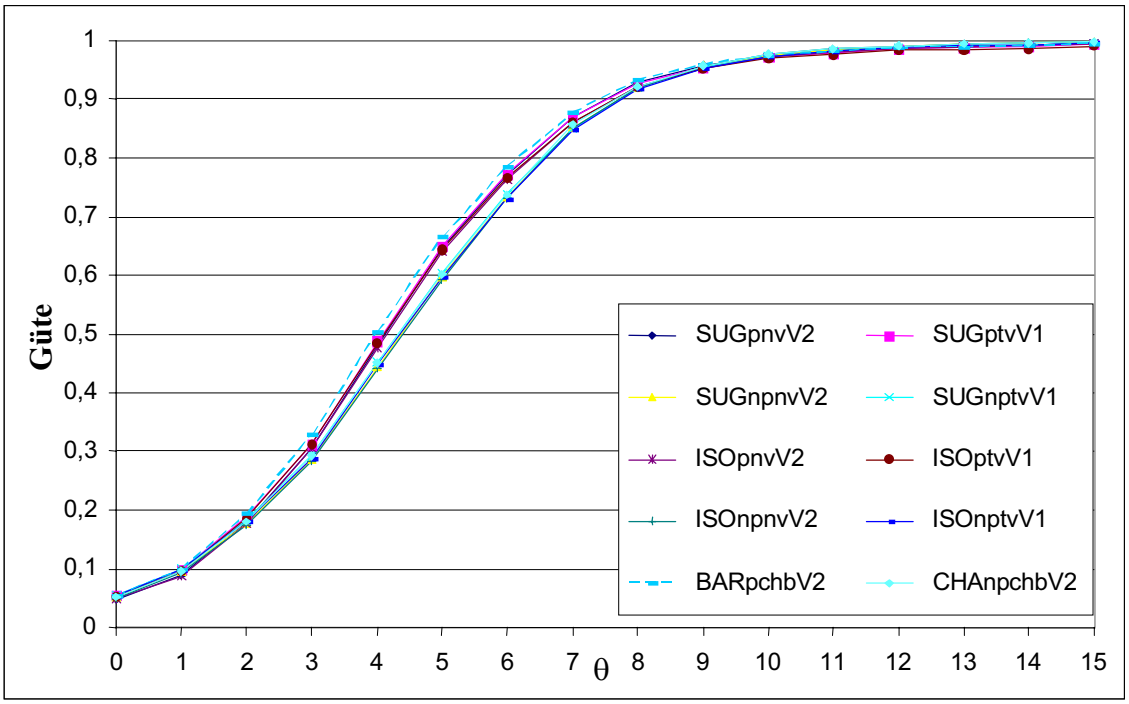


Abbildung 6.43: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 4$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV)

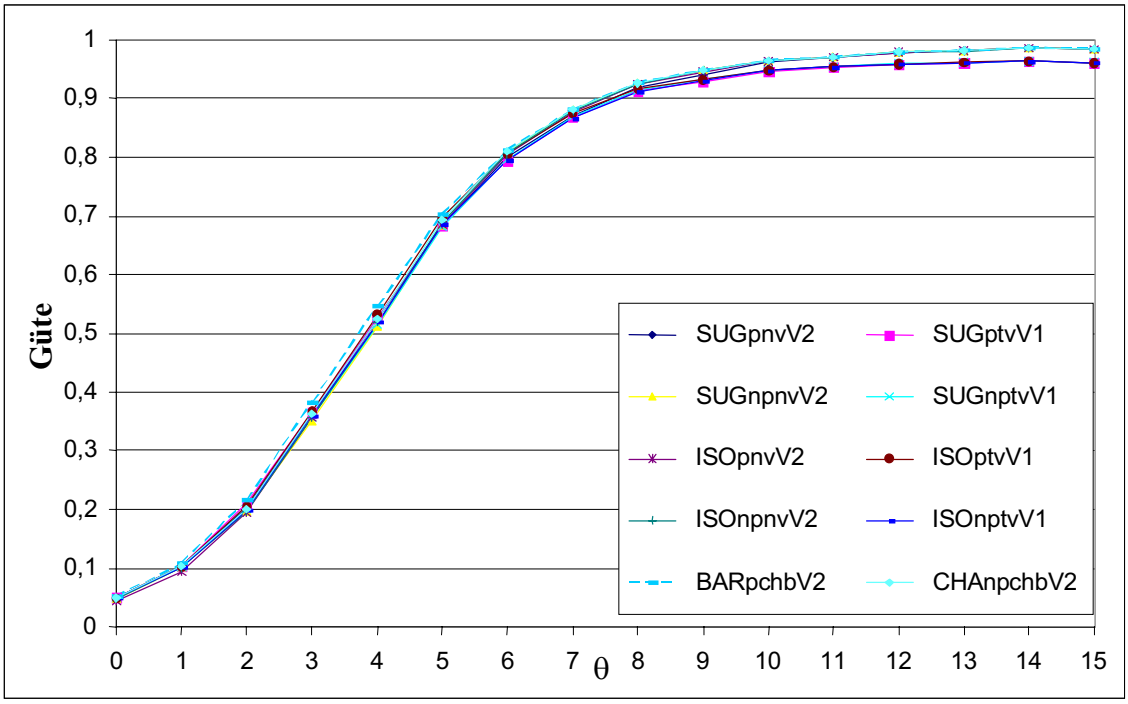


Abbildung 6.44: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 5$, $r = 3$, $\omega = 0{,}5$, Verteilungstyp = RSV)

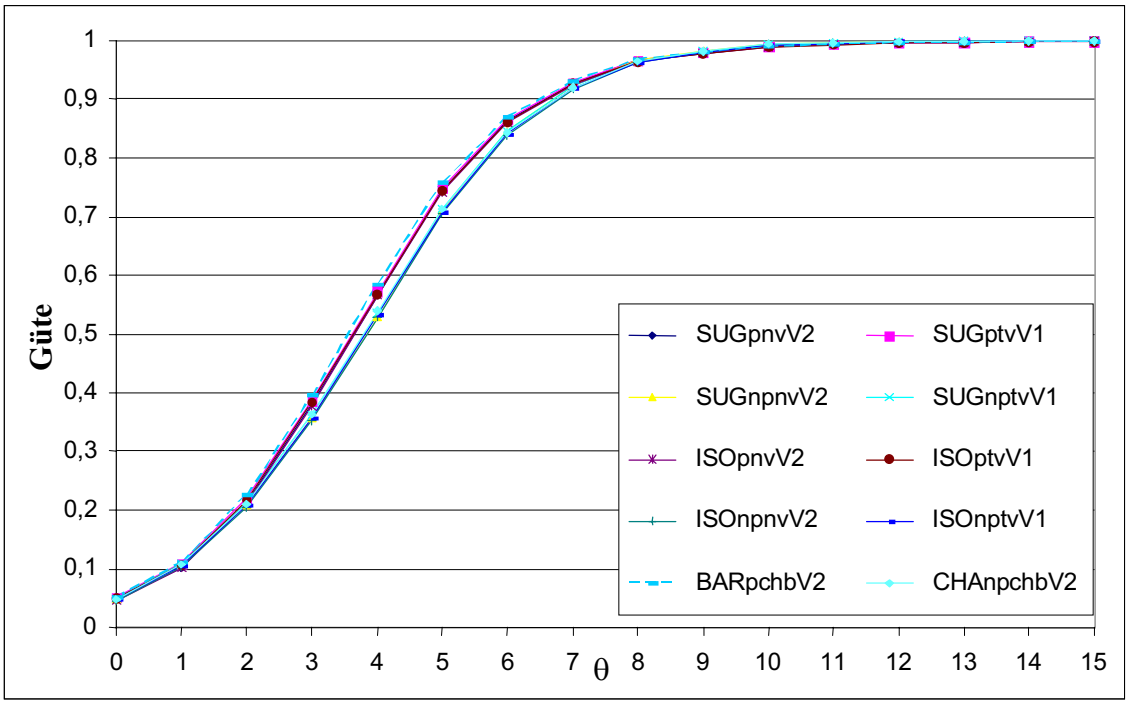


Abbildung 6.45: Güte von Vierstichprobentests auf Basis von infiniten Verteilungen in Abhängigkeit von (θ) ($\alpha = 0{,}05$, $n = 5$, $r = 9$, $\omega = 0{,}5$, Verteilungstyp = RSV)

7 Multiples Testen

Bei der Bestimmung der *MED*, der *MÄD*, des *HEDS* und der *MED&MÄD* (Problem B bis E, siehe Abschnitt 1.1) geht es darum, aus mehreren Dosen die Dosis herauszufinden, die ein bestimmtes Kriterium erfüllt. Die *MED* soll z. B. die kleinste der untersuchten Dosen sein, die sich hinsichtlich ihrer Wirkung noch von der Kontrolle unterscheidet. Zur Lösung der Probleme B bis E werden mehrere Tests durchgeführt, wobei jeder der einzelnen Tests mit den zugehörigen Fehlerwahrscheinlichkeiten verbunden ist. Das Risiko sich für eine falsche Dosis zu entscheiden, hängt somit von mehreren Einzelrisiken (Fehlern) ab. Wie bei allen multiplen Testproblemen stellt sich die Frage, welche Fehlerwahrscheinlichkeiten zu beschränken sind. Vorrangig werden im allgemeinen, wie beim einfachen Testen, nur die Wahrscheinlichkeiten für die Fehler 1. Art kontrolliert. Diese können durch Vorgabe eines lokalen, globalen oder multiplen Niveaus und der Wahl eines geeigneten multiplen Testverfahrens relativ gut beschränkt werden. Das lokale Niveau, welches definiert ist als Wahrscheinlichkeit, irgendeine der Nullhypothesen fälschlicherweise abzulehnen, kann einfach durch die Anwendung von mehreren α-Tests gewährleistet werden. Bei dieser pragmatischen Vorgehensweise, die von einigen Autoren empfohlen wird, werden also mehrfache Fehler ignoriert. Für das strengere multiple Niveau, die Wahrscheinlichkeit dafür, mindestens eine der Nullhypothesen unberechtigt abzulehnen, unabhängig davon, wie viele und welche Nullhypothesen wirklich wahr sind, gibt es mehrere Verfahren. Sie unterteilen sich in Ein- und Mehrschrittverfahren. Ein einfaches Verfahren zur Kontrolle des multiplen Niveaus basiert z. B. auf der Bonferonni-Ungleichung:

$$P\left(\bigcup_{i=1}^{m} A_i\right) \le \sum_{i=1}^{m} P(A_i) \qquad A_i = \text{Ereignis, die Nullhypothese } i \text{ abzulehnen.}$$

Wird jede der m zu prüfenden Einzelhypothesen (sie werden auch als Elementarhypothesen bezeichnet) zum Niveau α / m bzw. mittels der adjustierten p-Werte

$$p_l^B = \min(1, m\, p_l)\,,$$

geprüft, beträgt die obere Grenze für die Wahrscheinlichkeit eines Fehlers 1. Art gerade $m\alpha / m = \alpha$. Wenn zwischen den einzelnen Hypothesen ein starker Zusammenhang besteht, d. h., sie voneinander abhängig sind, ist dies Verfahren jedoch ineffizient, da bestehende

Abhängigkeiten nicht ausgenutzt werden. Besonders bei kleinen Stichprobenumfängen ist die Verringerung des Signifikanzniveaus kritisch. Schon wenn drei Elementarhypothesen geprüft werden sollen, z. B. Bestimmung der *MED* für einen Versuch mit drei Dosen und einer Kontrolle, würde das lokale Signifikanzniveau $0{,}01\overline{6}$ betragen. Die Wahl eines geeigneten Tests ist dann wie bei $\alpha = 0{,}01$ vor allem bei Zweistichprobentests problematisch.

Eine Verbesserung des Bonferronni-Verfahrens beruht auf einem schrittweisen Vorgehen. Dazu werden zunächst die *m* p-Werte aufsteigend sortiert:

$$p_1,...,p_m \ (p_i = p_{H_{0i}}) \rightarrow \ p_{l_1},...,p_{l_m} \ (p_{l_q} = p_{H_{0l_q}}) \quad \text{mit} \ p_{l_q} \le p_{l_r} \ \forall l_q \le l_r \, .$$

Die Elementarhypothesen werden dann schrittweise, z. B. nach Holm [180], mittels folgender Prozedur geprüft:

Schritt 1:

Wenn $p_{l_1} \le \alpha / m$ gilt, wird H_{0l_1} abgelehnt und mit Schritt 2 fortgefahren. Gilt jedoch $p_{l_1} > \alpha / m$, dann kann keine der Nullhypothesen abgelehnt werden, und die Prozedur ist beendet.

Schritt 2:

Gilt $p_{l_2} \le \alpha /(m-1)$, dann wird auch H_{0l_2} abgelehnt und mit Schritt 3 fortgefahren. Gilt jedoch $p_{l_2} > \alpha /(m-1)$, dann kann keine weitere Nullhypothese abgelehnt werden, und die Prozedur endet mit diesem Schritt.

...

Schritt i:

Gilt $p_{l_i} \le \alpha /(m-i+1)$, dann wird auch $H_{0l_{m-i+1}}$ abgelehnt und mit Schritt $i+1$ fortgefahren. Gilt jedoch $p_{l_i} > \alpha /(m-i+1)$, dann kann keine weitere Nullhypothese abgelehnt werden, und die Prozedur endet mit diesem Schritt.

...

Schritt m:

Gilt $p_{l_m} \le \alpha$, dann wird auch H_{0l_m} abgelehnt . Gilt $p_{l_m} > \alpha$, dann kann H_{0l_m} nicht abgelehnt werden, und die Prozedur endet mit diesem Schritt.

Bei einem Versuch mit drei Dosen und einer Kontrolle würde der erste Test immer noch zum Niveau $\alpha = 0{,}01\overline{6}$ durchgeführt werden. Der zweite bzw. dritte Test würde hingegen zum

Niveau $\alpha = 0{,}025$ bzw. $\alpha = 0{,}05$ prüfen. Nachteilig ist außerdem, daß bei diskreten Daten gleichgroße p-Werte auftreten können (vor allem bei Rangtests). Die Sortierreihenfolge ist dann nicht mehr eindeutig. Es müßte also zusätzlich a priori eine Ordnung vorgegeben werden.

Maurer et al. [181] beschreiben eine Prozedur, die gänzlich auf a priori geordneten Hypothesen beruht. Sie zeigen, daß wenn in der festgelegten Reihenfolge sequentiell getestet wird, auf jegliche Adjustierung verzichtet werden kann und trotzdem ein multiples Niveau eingehalten wird. Da sich für die Bestimmung der *MED* die Testreihenfolge ganz natürlich aus der Definition der *MED* und der totalen Ordnung der Versuchsglieder ergibt, kann diese Prozedur zur Bestimmung der *MED* genutzt werden. Wird mit $p_i = p_{H_{0i}^F}$ der p-Wert bezeichnet, der zum Prüfen der Hypothese H_{0i}^F (siehe Kapitel 1.1) berechnet wurde, so wird folgende Prozedur empfohlen.

Bestimmung der *MED*:

Schritt 1:

Wenn $p_k \leq \alpha$ gilt, wird H_{0k}^F abgelehnt und mit Schritt 2 fortgefahren. Gilt jedoch $p_k > \alpha$, dann existiert keine effektive Dosis, und die Prozedur ist beendet.

Schritt 2:

Gilt $p_{k-1} \leq \alpha$, dann wird auch H_{0k-1}^F abgelehnt und mit Schritt 3 fortgefahren. Gilt jedoch $p_{k-1} > \alpha$, dann kann keine weitere Nullhypothese abgelehnt werden. Als *MED* wird dann die Dosis D_k bestimmt, und die Prozedur wird mit diesem Schritt beendet.

...

Schritt i:

Gilt $p_{k-i+1} \leq \alpha$, dann wird auch H_{0k-i+1}^F abgelehnt, und mit Schritt $i+1$ fortgefahren. Gilt jedoch $p_{k-i+1} > \alpha$, dann kann keine weitere Nullhypothese abgelehnt werden, und die Prozedur endet mit diesem Schritt. Als *MED* wird dann die Dosis D_{i+1} definiert

...

Schritt k:

Gilt $p_1 \leq \alpha$, dann wird auch H_{01}^F abgelehnt und D_1 als *MED* definiert. Gilt $p_1 > \alpha$, dann kann H_{01}^F nicht abgelehnt werden, und D_2 wird als *MED* definiert. Die Prozedur endet mit diesem Schritt.

Zu dieser Prozedur würde auch die Nutzung des Abschlußtestprinzips [182] führen, wenn die totale Ordnung der Versuchsglieder vorausgesetzt wird [183]. Sind Wirkungsabfälle bei hohen Dosen ausgeschlossen, d. h., eine höhere Dosis hat stets eine mindestens gleichgroße Wirkung wie eine kleinere Dosis, so können neben den Zweistichprobentests im Schritt i auch Mehrstichprobentests genutzt werden. Dabei gibt es zwei Varianten:

a) die Daten der Kontrolle und der Dosen D_1 bis D_i bilden die Basis des Tests im Schritt i,
b) die Daten der Kontrolle und der Dosen D_1 bis D_k bilden die Basis des Tests im Schritt i.

Aufgrund des höheren Stichprobenumfangs führt die Variante a) zu einer Bevorteilung der hohen Dosen. Selbst eine kleine Dosis, mit derselben Wirkung wie eine höhere Dosis, hat eine geringere Chance als *MED* erkannt zu werden. Trotzdem wird diese Variante empfohlen. Die Variante b) macht, wenn überhaupt, bei der Benutzung einer multivariaten t-Verteilung einen Sinn. Im Schritt i könnte, z. B. als Kontrastmatrix, eine Matrix mit $2^i - 1$ Zeilen und $k+1$ Spalten genutzt werden, wobei die ersten $i+1$ Spalten der isotonen Kontrastmatrix für $i+1$ Gruppen entsprechen und die letzten $k-i$ Spalten mit Nullen aufgefüllt werden. Der so erhöhte Freiheitsgrad wirkt sich zwar positiv auf die Güte aus, da aber in der Regel nicht von homogenen Varianzen ausgegangen werden kann, erscheint es andererseits nicht sinnvoll, im Schritt i noch Informationen der Dosis D_{i+1}, die ja schon als signifikant verschieden von der Kontrolle erkannt wurde, in den Varianzschätzer mit einfließen zu lassen (gilt auch bei Nutzung einer multivariaten Normalverteilung). Bei der Nutzung der Variante b) in Verbindung mit Bootstrap- oder Permutationsverteilungen haben die Daten der Dosen D_{i+1} bis D_k nicht nur einen Einfluß auf den Varianzschätzer und den Freiheitsgrad, sondern auf die gesamte Verteilung der Statistik. Zudem bedeutet eine Ablehnung im Schritt 1 ja, daß sich die Verteilung der Dosis D_k von den Verteilungen anderer Versuchsglieder unterscheidet. D.h., im Schritt 2 würde bewußt mißachtet, daß die genutzte Verteilung nicht invariant gegenüber Permutationen der Daten ist. Der Einfluß kann zwar auch positiv sein, läßt sich im allgemeinen aber nicht klassifizieren.

Statt in jedem Schritt die Ränge neu zu bestimmen, was hier empfohlen wird, könnten die Ränge auch einmalig über alle Daten gebildet werden und anschließend in jedem Schritt beibehalten werden. Williams [184] zeigt jedoch, daß Shirleys-Prozedure [185] für den Vergleich mehrerer Dosen mit einer Kontrolle, die auf einmaliger Rangbildung beruht, verbessert werden kann, wenn in jedem Schritt (Vergleich Dosis i mit der Kontrolle) die Ränge neu berechnet werden.

Eine Prozedur zur Bestimmung der *MÄD* kann analog zur Bestimmung der *MED* beschrieben werden, da auch bei diesem Problem eine Ordnung der Hypothesen naheliegend ist. Kritisch ist jedoch der eigentliche Äquivalenznachweis für jede Dosis. Sind die Hypothesen im Normalverteilungsmodell noch relativ gut mittels der Erwartungswerte formulierbar und mit Hilfe des t-Tests prüfbar, so ist dies für geordnete kategoriale Daten nicht so klar. Dies liegt vor allem daran, daß der Abstand zwischen den Kategorien nicht meßbar ist und unterschiedlich groß sein kann. Wird die auf Seite 1 beschriebene Bonitурskala benutzt und wird im Durchschnitt die Wirkung eines neuen Herbizids mit 2 und die eines Standards mit 1, 2 oder gar 3 bewertet, so kann in diesem Fall sicher von äquivalenten Mitteln gesprochen werden. Am oberen Ende der Skala erscheinen ähnliche Aussagen unsinnig. Werden die Daten jedoch, wie z. B. bei den in Abschnitt 1.1 beschriebenen Fungizid- und Herbizidversuchen, auf einer annähernd stetigen Skala erhoben, oder handelt es sich um Zähldaten (ohne Kategorisierung), so besitzen Äquivalenzaussagen einen Sinn. Für alle anderen geordneten kategorialen Daten bieten sich eher Aussagen an, wie sie z. B. von Wellek und Hampel [186] vorgeschlagen werden. Da für diese Verfahren die Vorgabe der Äquivalenzschranken nicht so offensichtlich ist, wird auf sie nicht weiter eingegangen.

Wurde die Wirkung des Standards schon in anderen Experimenten nachgewiesen, so ist im allgemeinen ein Vergleich des Standards mit der Kontrollgruppe (Sensitivitätstest) zur Bewertung des Experiments sinnvoll [15]. Nur wenn beim Prüfen der Hypothesen

$$H^F_{00S}: F_0 \geq F_S \qquad \text{versus} \qquad H^F_{A0S}: F_0 < F_S$$

ein signifikanter Unterschied nachgewiesen werden kann, sollte mit dem Prüfen der zu untersuchenden Äquivalenzhypothesen fortgefahren werden [15]. Da auch bei diesem bedingten Vorgehen eine Und-Verknüpfung genutzt wird, ist auch hier keine Adjustierung des Signifikanzniveaus notwendig. Bei kleinen Fallzahlen oder Standardprodukten mit nicht so großer Effektivität ist dieses Vorgehen aufgrund der fehlenden Kontrolle eines Fehlers 2. Art kritisch bzw. zu streng. Dieser Tests sollte daher entweder auf Basis eines schwächeren Niveaus (z. B.

$\alpha = 0{,}10$) durchgeführt werden oder nur bei der Interpretation der Ergebnisse diskutiert werden.

Für das Testen der Äquivalenzhypothesen werden die Daten des Standards zunächst um ε verschoben ($Z_{Sj} = X_{Sj} - \varepsilon$, $j = 1,...,n_S$, $Z_{ij} = X_{ij}$, $j = 1,...,n_i$). Auf Basis der transformierten Daten bzw. der gebildeten Rangdaten ($R_{ij} = R_{ij}(Z_{ij})$ bzw. $R_{Sj} = R_{Sj}(Z_{Sj})$) und der berechneten p-Werte $p_i = p_{H^F_{0Si}(\varepsilon)}$ (z. B. mit Mid-p-Tests; Hypothesen siehe Abschnitt 1.1) wird die folgende Prozedur empfohlen.

Bestimmung der *MÄD*:

Schritt 1:

Wenn $p_k \le \alpha$ gilt, wird $H^F_{0,Sk}(\varepsilon)$ abgelehnt und mit Schritt 2 fortgefahren. Gilt jedoch $p_k > \alpha$, dann existierte keine äquivalente Dosis, und die Prozedur ist beendet.

Schritt 2:

Gilt $p_{k-1} \le \alpha$, dann wird auch $H^F_{0,S(k-1)}(\varepsilon)$ abgelehnt und mit Schritt 3 fortgefahren. Gilt jedoch $p_{k-1} > \alpha$, dann kann keine weitere Nullhypothese abgelehnt werden. Als *MÄD* wird dann die Dosis D_k bestimmt, und die Prozedur wird mit diesem Schritt beendet.

...

Schritt i:

Gilt $p_{k-i+1} \le \alpha$, dann wird auch $H^F_{0,S(k-i+1)}(\varepsilon)$ abgelehnt und mit Schritt $i+1$ fortgefahren. Gilt jedoch $p_{k-i+1} > \alpha$, dann kann keine weitere Nullhypothese abgelehnt werden, und die Prozedur endet mit diesem Schritt. Als *MÄD* wird die Dosis D_{i+1} definiert

...

Schritt k:

Gilt $p_1 \le \alpha$, dann wird auch $H^F_{0,S1}(\varepsilon)$ abgelehnt und D_1 als *MÄD* definiert. Gilt $p_1 > \alpha$, dann kann $H^F_{0,S1}(\varepsilon)$ nicht abgelehnt werden, und D_2 wird als *MED* definiert. Die Prozedur endet mit diesem Schritt.

Auch diese Prozedur kann mit dem Abschlußtestprinzip hergeleitet werden, wenn von der totalen Ordnung der Versuchsglieder ausgegangen werden kann [15].

Die Theorie zum Testen der Hypothesen

$$H^{\mu}_{0,Si}(\delta): \mu_S - \delta\,\mu_S \ge \mu_i \quad \text{versus} \qquad H^{\mu}_{A,Si}(\delta): \mu_S - \delta\,\mu_S < \mu_i$$

basiert im wesentlichen auf normalverteilten Daten [187]. Efron und Tibshirani [55] zeigen jedoch, daß nach Transformation der Daten

$$Z_{Sj} = (1-\delta)\, X_{Sj} \;\; j = 1,...,n_S \,,\; Z_{ij} = X_{ij} \,,\; j = 1,...,n_i$$

Bootstraptests zu obigen Hypothesen durchgeführt werden können. Gerade bei Prozentwerten ist es aber relativ einfach, eine Äquivalenzschranke (z. B. $\varepsilon = 5\%$) vorzugeben. Der Äquivalenznachweis sollte daher bevorzugt auf einer Verschiebung der Daten des Standards um ε beruhen, da durch die Verschiebung nur die Lage und nicht auch noch die Variabilität der Daten verändert wird.

Die Bestimmung der *MED&MÄD* kann erfolgen, indem die empfohlenen Prozeduren zur Bestimmung der *MED* bzw. *MÄD* durch eine Und-Beziehung verknüpft werden [15]. D. h., als *MED&MÄD* wird die Dosis definiert, für die letztmalig sowohl bei der Bestimmung der *MED* im Schritt i als auch bei der Bestimmung der *MÄD* im Schritt i eine Ablehnung der Nullhypothese erfolgt.

Für die Bestimmung der *MED*, *MÄD* und der *MED&MÄD* ergibt sich die Testreihenfolge offensichtlich aus den theoretischen Definitionen. Für die Bestimmung des *HEDS* kann das Verfahren der a priori geordneten Hypothesen nicht genutzt werden. Würde nämlich im ersten Schritt D_1 mit D_2 verglichen ($H^F_{0,12}$), so könnte ein möglicher effizienter Dosisschritt bei den höheren Dosen übersehen werden, da nur bei Ablehnung der Nullhypothese zum nächsten Test übergegangen werden darf. Ähnlich sieht es aus, wenn beim Vergleich der Dosen D_{k-1} und D_k begonnen wird und kein Unterschied aufgedeckt werden kann. Beide Verfahren würden somit nicht im Sinne der Fragestellung stehen. Für die Bestimmung des *HEDS* können daher nur die Holm-Prozedur oder weitere Verbesserungen dieser [188; 189] in Verbindung mit Mid-p-Zweistichprobentests empfohlen werden. Da hier jedoch stets α-Adjustierungen für die Einzeltests (zumindest im ersten Schritt) notwendig sind und die Unterschiede im allgemeinen nicht so deutlich sind wie beim Vergleich mit der Kontrolle, sollten für diese Fragestellung höhere Stichprobenumfänge angestrebt werden.

Mit den vorgestellten Prozeduren werden die Fehler 1. Art bei Anwendung geeigneter Tests (siehe Kapitel 6) im wesentlichen kontrolliert. Aufgrund der betrachteten kleinen Stichprobenumfänge können jedoch kaum Aussagen zu den analog definierbaren multiplen Fehlern 2. Art getroffen werden. Da alle Testergebnisse von den verwendeten Stichprobenumfängen abhängen, sind die bestimmten Werte (Schätzungen) für die *MED, MÄD, MED&MÄD* und für den *HEDS* auch stark vom Stichprobenumfang abhängig. Leider kann nicht, wie sonst für

Schätzer üblich, ein sinnvoller Varianzschätzer angegeben werden. Um diese Problematik besser zu behandeln wären andere Mittel der endlichen Entscheidungstheorie, wie z. B. Verlustfunktionen, notwendig [190]. Darauf soll in der vorliegenden Arbeit aber nicht weiter eingegangen werden.

8 Anwendungsbeispiel

Am Beispiel eines Fungizidversuchs wird im folgenden zusammenfassend dargelegt, wie die in Kapitel 1 genannten Fragestellungen unter Anwendung der in Kapitel 7 beschriebenen Prozeduren gelöst werden können.

Aus einem komplexen Versuch der BASF AG aus dem Jahr 1995 wurden die Daten für die Kontrolle, eines Präparates, das in vier aufsteigenden Dosen untersucht wurde, und für einen Standard entnommen. Bei diesem Versuch wurde als Kulturpflanze Wintergerste angebaut. Die in Tabelle 7.1 widergegebenen Ergebnisse stellen die Bonituren für den Befall der Blätter mit Rhynchosporium secalis zu drei verschieden Zeitpunkten (Entwicklungsstadien) dar. Die Fungizide wurden durch Spritzen aufgetragen.

Stadium der Kultur-pflanze	**Wieder-holung**	**D_0 (unbe-handelt)**	**D_1 (0.25 L/HA[1])**	**D_2 (0.5 L/HA)**	**D_3 (0.75 L/HA)**	**D_4 (1.0 L/HA)**	**D_S (1.0 L/HA)**
Blatt-häutchen	1	8	7	7	6	6	7
	2	7	7	7	6	6	7
	3	6	6	5	7	6	7
	4	7	6	7	7	7	7
Blüh-beginn	1	20	8	5	6	2	6
	2	5	7	5	4	5	6
	3	7	6	5	4	3	4
	4	7	7	7	6	3	5
Blühende	1	25	8	8	5	3	15
	2	25	18	10	10	5	12
	3	25	18	10	6	5	15
	4	22	15	8	8	3	18

Tabelle 8.1: Beispieldaten aus einem Fungizidversuch der BASF AG 1995 ([1] Aufwandmengen in Liter je Hektar)

Aufgrund der Empfehlungen des 6. Kapitels werden als Trendtests die Mid-p-Tests auf Basis der isotonen Schätzer (*BARppmpoV*, *CHAnppmpoV*) zur Bestimmung der *MED* benutzt. Als Zweistichprobentests wurden die Mid-p-Tests *tTppmpoV* und *tTnppmpo*V gewählt. Zum Vergleich wurden auch noch die p-Werte der analogen exakten Permutationstests in die Tabellen aufgenommen. Als multiples Niveau wurde $\alpha = 0{,}05$ gewählt.

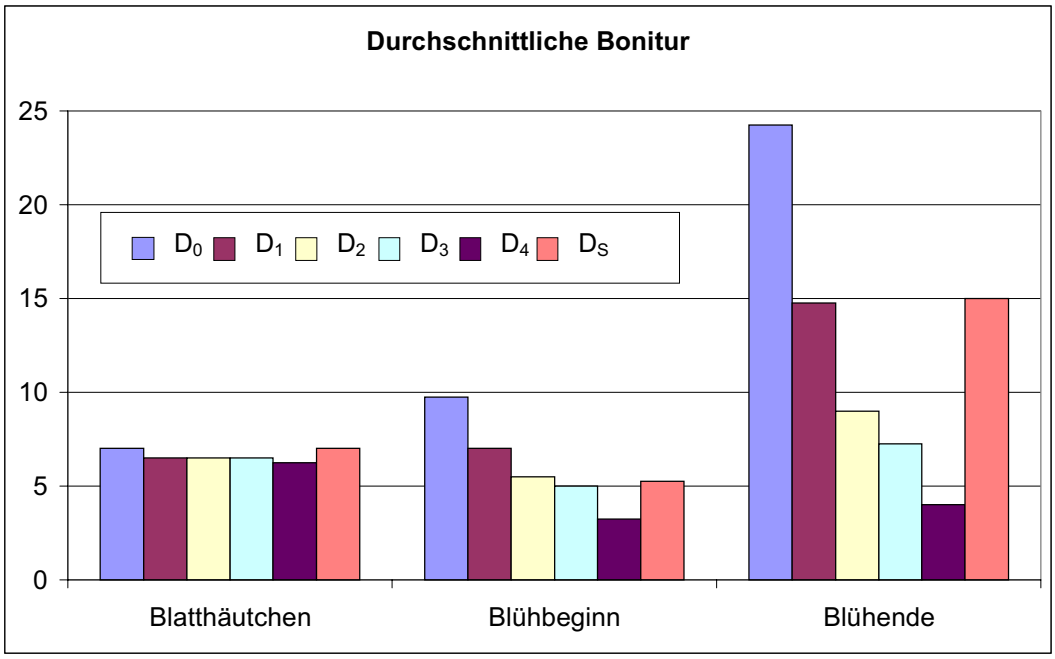


Abbildung 8.1: Mittelwerte der Versuchsglieder zu den drei verschiedenen Zeitpunkten

Stadium der Kulturpflanze	MED	BARppexoV	BARppmpoV	CHAnppexoV	CHAnppmpoV
Blatt-häutchen	H_{04}^{F}	0,2129	0,1896	0,1991	0,1984
	H_{03}^{F}	0,3826	0,2992	0,2971	0,2908
	H_{02}^{F}	0,3545	0,2909	0,3555	0,3232
	H_{01}^{F}	0,3286	0,1928	0,3286	0,2000

Tabelle 8.2: p-Werte für die *MED*-Bestimmung am 1. Zeitpunkt

Zum 1. Zeitpunkt kann weder für die Dosen noch für das Standardprodukt eine Wirkung nachgewiesen werden. Aufgrund des relativ geringen Befalls der Kontrollparzelle ist dies auch nicht überraschend.

Stadium der Kulturpflanze	HEDS	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blatt-häutchen	H_{012}^{F}	0,6429	0,5000	0,6429	0,5714
	H_{023}^{F}	0,6429	0,5000	0,5000	0,4286
	H_{034}^{F}	0,5000	0,2857	0,5000	0,2857

Tabelle 8.3: p-Werte für die *HEDS*-Bestimmung am 1. Zeitpunkt

Stadium der Kulturpflanze	MÄD $\varepsilon = 2$	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blatt-häutchen	H_{00S}^{F}	0,7143	0,5000	0,7143	0,5000
	$H_{0S4}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S3}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S2}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S1}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071

Tabelle 8.4: p-Werte für die *MÄD*-Bestimmung am 1. Zeitpunkt

Wie Abbildung 7.1 schon vermuten läßt, existiert auch kein effektiver Dosisschritt. Die geringe Variabilität in den Daten und das Verschieben um $\varepsilon = 2$ führt dazu, daß sich zu diesem Zeitpunkt alle Dosen zum Standard äquivalent erweisen. Der nicht signifikante Sensitivitätstest und das Fehlen effektiver Dosen legt jedoch nahe, diese Ergebnisse nicht zu hoch zu bewerten. Eine *MED&MÄD* existiert nicht.

Stadium der Kulturpflanze	MED	BARppexoV	BARppmpoV	CHAnppexoV	CHAnppmpoV
Blühbeginn	H_{04}^{F}	0,0026	0,0026	0,0002	0,0002
	H_{03}^{F}	0,0313	0,0307	0,0134	0,0133
	H_{02}^{F}	0,1163	0,1142	0,0693	0,0592
	H_{01}^{F}	0,4429	0,3929	1,0000	0,7000

Tabelle 8.5: p-Werte für die *MED*-Bestimmung am 2. Zeitpunkt

Stadium der Kulturpflanze	HEDS	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blüh-beginn	H_{012}^{F}	0,0571	0,0357	0,0571	0,0357
	H_{023}^{F}	0,3857	0,2714	0,3857	0,2929
	H_{034}^{F}	0,0714	0,0500	0,0429	0,0357

Tabelle 8.6: p-Werte für die *HEDS*-Bestimmung am 2. Zeitpunkt

Stadium der Kulturpflanze	MÄD $\varepsilon = 2$	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blüh-beginn	H_{00S}^{F}	0,0857	0,0571	0,0857	0,0571
	$H_{0S4}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S3}^{F}(\varepsilon)$	0,0429	0,0214	0,0429	0,0214
	$H_{0S2}^{F}(\varepsilon)$	0,0429	0,0286	0,0429	0,0286
	$H_{0S1}^{F}(\varepsilon)$	0,5000	0,3500	0,4571	0,3286

Tabelle 8.7: p-Werte für die *MÄD*-Bestimmung am 2. Zeitpunkt

Zum Zeitpunkt des Blühbeginns ist bei den hohen Dosen schon eine Wirkung erkennbar. Sowohl die Dosis D_3 als auch die Dosis D_4 erweisen sich als effektiv. Nur bei einem lokalen Signifikanzniveau von $\alpha = 0{,}05$ würden sich die Dosisschritte von D_1 nach D_2 und von D_3 nach D_4 als effektiv erweisen, nicht jedoch wenn ein multiples Niveau von $\alpha = 0{,}05$ kontrolliert werden soll. Die zu geringe Konzentration der Dosis D_1 führt dazu, daß sich nur noch die Dosen D_2 bis D_4 zum Standard äquivalent erweisen. Der Sensitivitätstest führt bei $\alpha = 0{,}05$ zwar nicht zur Signifikanz, es deutet sich jedoch eine Wirksamkeit des Standards an. Äquivalenzaussagen erscheinen für diesen Zeitpunkt daher schon als sinnvoll. Die *MÄD* ist somit die Dosis D_2 und die Dosis D_3 ist *MED* und *MED&MÄD*.

Im Stadium Blühende kann eine Wirkung aller Dosen nachgewiesen werden. Die Dosis D_1 ist somit die *MED*. Aber auch in diesem Stadium kann auf Basis eines multiplen Niveaus kein Dosisschritt als effektiv bezeichnet werden. Die Wirksamkeit des Standards ist zu diesem Zeitpunkt mittels des Sensitivitätstests deutlich nachweisbar. Die Wirkung der Dosen D_2 bis

Stadium der Kulturpflanze	MED	BARppexoV	BARppmpoV	CHAnppexoV	CHAnppmpoV
Blüh-ende	H_{04}^{F}	0,0000	0,0000	0,0000	0,0000
	H_{03}^{F}	0,0000	0,0000	0,0000	0,0000
	H_{02}^{F}	0,0001	0,0001	0,0003	0,0002
	H_{01}^{F}	0,0143	0,0071	0,0143	0,0071

Tabelle 8.8: p-Werte für die *MED*-Bestimmung am 3. Zeitpunkt

Stadium der Kulturpflanze	HEDS	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blüh-ende	H_{012}^{F}	0,0714	0,0500	0,0857	0,0643
	H_{023}^{F}	0,1857	0,1143	0,2000	0,1286
	H_{034}^{F}	0,0429	0,0214	0,0429	0,0214

Tabelle 8.9: p-Werte für die *HEDS*-Bestimmung am 3. Zeitpunkt

Stadium der Kulturpflanze	MÄD $\varepsilon = 2$	tTppexoV	tTppmpoV	tTnppexoV	tTnppmpoV
Blüh-ende	H_{00S}^{F}	0,0143	0,0071	0,0143	0,0071
	$H_{0S4}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S3}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S2}^{F}(\varepsilon)$	0,0143	0,0071	0,0143	0,0071
	$H_{0S1}^{F}(\varepsilon)$	0,2714	0,2500	0,4000	0,3929

Tabelle 8.10: p-Werte für die *MÄD*-Bestimmung am 3. Zeitpunkt

D_4 ist in diesem Stadium nicht nur äquivalent zur Wirkung des Standards ($\varepsilon = 2$), sondern ist sogar höher ($\varepsilon = 0$). Die *MÄD* und *MED&MÄD* ist somit die Dosis D_2.

Das Beispiel zeigt, daß die Tests nicht nur in Simulationen überzeugen, sondern auch praktisch bei kleinen Stichproben die Effektivität eines Fungizids und die Äquivalenz zu einem Standardprodukt statistisch belegen können. Die Bestimmung der effektiven Dosisschritte bei

Kontrolle eines multiplen Niveaus scheitert jedoch selbst bei zum Teil klaren Effekten an der notwendigen Niveauadjustierung. An diesem Beispiel wird einerseits nochmal deutlich, wie groß der Unterschied zwischen den p-Werten eines exakten Permutationstests und eines Mid-p-Tests sein kann. Andererseits fallen die vielen gleichgroßen p-Werte auf. Bei kleinen Stichproben gilt somit um so mehr, daß signifikante Ergebnisse nicht nur durch Angabe von p-Werten dargestellt werden sollten. Konfidenzintervalle sind sicher eine bessere Lösung. Leider gibt es nur wenige Arbeiten, die sich mit Konfidenzintervallen für diskrete Daten beschäftigen. Röhmel [193] definiert Konfidenzintervalle auf Basis von Zweistichproben-Permutationstests und diskutiert deren Probleme bei Anwendung dieser Intervalle. Konfidenz-intervalle auf Basis einer (multivariaten) Normalverteilung oder einer t-Verteilung sind zwar Alternativen, aufgrund der Simulationsergebnisse jedoch erst ab $n = 5$ empfehlenswert. Aus dem gleichen Grund wurden auch keine multiplen Testprozeduren diskutiert, die auf multivariaten Normalverteilungs- oder t-Verteilungsapproximationen basieren und z. B. zur Bestimmung des *HEDS* eingesetzt werden könnten.

9 Schlußbetrachtung

In der vorliegenden Arbeit wurden einseitige und gerichtete Tests für zwei oder mehr unabhängige Stichproben untersucht. Das Ziel bestand darin, Testempfehlungen für die Auswertung von jenen Versuchen geben zu können, bei denen geordnete kategoriale Daten erhoben werden und nur sehr kleine Stichproben zur Verfügung stehen. Da nur einseitige Hypothesen bzw. Trendhypothesen von Interesse waren, wurden einseitige Zweistichprobentests und Trendtests für mehr als zwei Stichproben vorgestellt. Das Verhalten der Trendtests wurde für drei und vier Stichproben untersucht.

Die Kombination von geordneten kategorialen Daten und sehr kleinen Fallzahlen führt nicht nur zu der Frage nach einer geeigneten Statistik und einer geeigneten Verteilungsapproximation, sondern auch zur Fragestellung, ob den Daten Scores zugeordnet werden sollen. Werden den Daten keine Scores zugeordnet, führt dies in der Regel zu Statistiken, deren Verteilungen vor allem bei kleinen Fallzahlen nur schwer handhabbar sind. Ein Beispiel hierfür sind die in Abschnitt 4.3 vorgestellten Likelihood-Quotienten-Tests. Werden den Daten Scores zugeordnet, so hängen alle Ergebnisse von der Wahl der Scores ab. Die adaptiven Ansätze, die untersucht wurden, um die Ergebnisse unabhängiger von der Wahl der Scores zu machen, erwiesen sich bei den kleinen Fallzahlen als nicht praktikabel. Als Folge werden auch in der vorliegenden Arbeit die äquidistanten Scores und die Durchschnittsränge empfohlen. Die Simulationen haben gezeigt, daß die Verwendung von Rängen bei den betrachteten Fallzahlen zu leichten Gütevorteilen führt.

Die Zuordnung von Scores führt in vielen Fällen zu einfachen Statistiken, die auf Mittelwerten beruhen. Es wurde gezeigt, daß verschiedene Ansätze zu recht ähnlichen oder gleichen Statistiken führen. Dies erklärt zum Teil, daß Verfahren genutzt und empfohlen werden können, die eigentlich für die Auswertung stetiger Daten entwickelt wurden. Beispiele hierfür sind die multiplen Kontraste und der Bartholomew-Test.

Im Zähler der untersuchten Teststatistiken stehen sowohl bei den Zwei- als auch bei den Mehrstichprobentests Linearkombinationen oder quadratische Funktionen der Gruppenmittelwerte. Sie können gut interpretiert werden und lassen sich einfach berechnen. Gerade aufgrund möglicher Bindungen in den Daten ist die Wahl eines geeigneten Nenners (meist ein Varianzschätzer für den Zählerterm) problematisch. Ansätze, bei denen von Varianzheterogenität ausgegangen wird, führen zu Varianzschätzern, die für die hier betrachteten Versuche ungeeignet sind. Dies äußert sich z. B. in einer hohen Wahrscheinlichkeit für einen Varianzschätzer mit dem Wert Null (Bootstraptests) oder in einem eher konservativen

Güteverhalten (Welch's Zweistichprobentest). Eine grundsätzliche Empfehlung bei den betrachteten Stichprobenumfängen lautet daher, auf einen Varianzschätzer zu verzichten oder einen Varianzschätzer zu benutzten, der auf der Gesamtvariabilität in den zu untersuchenden Stichproben beruht. Dies führt bei den Zweistichprobentests zu einer Empfehlung der Statistik des Wilcoxon-Mann-Whitney-Tests. Da die vorgestellten Mehrstichprobentests auf Basis der isotonen Kontrastmatrix relativ gleichwertig zu den Verfahren sind, die auf den isotonen Schätzer beruhen, kann die Frage nach einer geeigneten Statistik mit der Empfehlung für die Chacko-Statistik auch im Mehrstichprobenfall hinreichend gut beantwortet werden. Diese Statistik erwies sich selbst bei kleinen Skalen als sensitiv für die betrachteten Hypothesen. Im Gegensatz zu Chuang-Stein und Agresti [13] wird diese Statistik oder die Bartholomew-Statistik daher auch für Skalen empfohlen, die auf weniger als fünf Punkte konzentriert sind.

Im allgemeinen ist die Wahl der Verteilung das Problem, welches am meisten Gespür benötigt. Neben den oft genutzten Approximationen durch eine Normalverteilung oder eine t-Verteilung wurden Permutationsverteilungen und Bootstrapverteilungen untersucht. Für den Zweistichprobenfall wurden „exakte" Bootstraptests und unbedingte Permutationstests als Alternative zu den üblichen Monte-Carlo-Bootstraptests und den bedingten Permutationstests umfangreich untersucht. Sie konnten sich im Vergleich mit den bekannten Mid-p-Tests jedoch nicht durchsetzen. Nur bei wenigen Parameterkonfigurationen zeigen sie sich den Mid-p-Tests überlegen. Für den Vergleich zweier Stichproben wird daher ein Mid-p-Test empfohlen. Sie sind den exakten bedingten Permutationstests und den parametrischen bzw. asymptotischen Tests überlegen. Die absolute Kontrolle des Fehlers 1. Art durch exakte bedingte Permutationstests sollte nicht überbewertet werden. Bei den betrachteten Fallzahlen sind diese Tests eher das Gegenteil eines „goldenen Standards". Statistikprogramme, die mit exakten Tests für kleine Fallzahlen werben, sollten die Anwender besonders im Zweistichprobenfall stärker für das Problem der Konservativität und der extrem schlechten Güte unter der Alternative sensibilisieren. Die Informationen für die Berechnung eines Mid-p-Wertes sollten auf jeden Fall stets bereitgestellt werden. Auch für den Mehrstichprobenfall wird ein Mid-p-Test empfohlen. Jedoch ist die Überlegenheit gegenüber den anderen geeigneten Tests nicht so gravierend. Anders als im Zweistichprobenfall ist die Berechnung eines Mid-p-Wertes selbst mit den heutigen Rechnern für den Fall, daß mehr als zwei Stichproben untersucht werden, sehr zeitaufwendig. Das für die Simulationen erstellte Programm zur Berechnung der Mid-p-Tests ist zwar nicht nur für drei oder vier Stichproben geeignet, benötigt aber schon für das Testen der Hypothesen H_{04}^{F} des in Kapitel 8 beschriebenen Beispiels über eine Minute (jeweils für einen Zeitpunkt; auf einem Pentium II

mit 300 MHz). Die Suche nach schnelleren Algorithmen ist daher eine Aufgabe, die sich an diese Arbeit anschließen sollte. Neue Ideen zur Berechnung exakter Permutationstests beschreiben z. B. Kang und Klotz [192].

Die Simulationen haben gezeigt, daß der Güteunterschied zwischen drei, vier oder fünf Wiederholungen gravierend ist. Bei der Versuchsplanung sollte daher genau abgewogen werden, wieviele Versuchsglieder in einem Versuch eingebunden werden. Zum Beispiel stellt sich bei Fungizidversuchen mit einem Präparat, das bereits im Gewächshaus und auf Kleinparzellen untersucht wurde, die Frage, ob bei eventuell nachfolgenden Versuchen auf Großparzellen der Vergleich mit der Kontrolle noch notwendig ist oder ob die sonst freien Parzellen besser mit dem neuen Präparat behandelt werden sollten. Diese Frage stellt sich vor allem dann, wenn ein effizienter Standard mituntersucht wird.

Die Simulationen verdeutlichen den enormen Güteverlust, der bei kleineren Skalen in Kauf genommen wird. Daher sollte eine nicht zu grobe ordinale Skala gewählt werden. Anders als viele asymptotische Tests besitzen Mid-p-Tests auch dann eine hohe Güte, wenn einige Zellen der Kontingenztafel nicht oder nur schwach besetzt sind. Eine feinere Skala kann somit genutzt werden, um die kleinen Fallzahlen etwas zu kompensieren.

Vorsicht ist bei der Wahl des Signifikanzniveaus geboten, da für kleine Signifikanzniveaus besonders im Zweistichprobenfall keine angemessene Kontrolle des Fehlers 2. Art möglich ist. Gerade bei multiplen Fragestellungen sollte daher überlegt werden, wie streng die Fehler 1. Art kontrolliert werden sollen.

In den Simulationen wurden nur balancierte Anlagen untersucht. Grundsätzlich eignen sich zwar alle vorgestellten Tests auch für unbalancierte Anlagen; ob eine Übertragung der Ergebnisse auch auf unbalancierte Anlagen möglich ist, sollte jedoch erst geprüft werden.

Um alle Tests miteinander vergleichen zu können, wurden die Parameter so gewählt, daß oft deutliche Unterschiede simuliert wurden. Die in den Tabellen oder Abbildungen wiedergegeben hohen Powerwerte sollten daher nicht zu optimistisch aufgenommen werden. Liegen nur sehr geringe Unterschiede zwischen den Versuchsgliedern vor, so ist selbst mit den Mid-p-Tests keine angemessene Kontrolle der Fehler 2. Art möglich.

Die Wahl eines geeigneten Designs spielt gerade bei Experimenten mit kleinen Stichproben eine bedeutende Rolle. An randomisierten Blockanlagen sollte z. B. nicht aus Tradition („es wurde schon immer so gemacht“) oder aus bearbeitungstechnischen Gründen festgehalten werden. Gerade bei kleinen Fallzahlen haben Blockanlagen (wenn eigentlich doch kein Blockeffekt existiert) deutliche Nachteile gegenüber vollständig randomisierten Anlagen. Da aufgrund von Simulationen, auf die hier nicht eingegangen wurde, auch für randomisierte

Blockanlagen mit drei bis fünf Blöcken Mid-p-Tests empfohlen werden, kann sich dieser Nachteil schon an der noch diskreteren Permutationsverteilung klargemacht werden. Stehen z. B. acht Parzellen zur Verfügung um zwei Mittel zu vergleichen, so existieren im Fall einer vollständig randomisierten Anlage bei balancierter Aufteilung und Invarianz der Statistik gegenüber Permutationen innerhalb einer Stichprobe $8!/4!4! = 70$ verschiedene Permutationen. Bei Blockanlagen mit vier Blöcken sind es hingegen nur $2^4 = 16$ Permutationen.

Bei Fungizid- und Herbizidversuchen wird oft nicht nur ein, sondern mehrere Merkmale bonitiert (z. B. mehrere Unkräuter, mehrere Krankheiten). Oder es wird, wie im Beispiel des achten Kapitels, die Wirkung zu mehreren Zeitpunkten bonitiert. Ein Wirkungsnachweis oder auch ein Äquivalenznachweis ist dann für mehrere Merkmale bzw. für mehrere Zeitpunkte zu erbringen. Zur Lösung dieser Probleme können multivariate Auswertungsverfahren genutzt werden, wie sie Bregenzer [151] beschreibt.

Entsprechend den Empfehlungen werden C-Programme zur Verfügung gestellt, die zur Durchführung der Mid-p-Tests bei kleinen Stichproben geeignet sind. Prinzipell sind sie auch für größere Fallzahlen einsetzbar, jedoch kann die Berechnungszeit dann sehr schnell steigen. Die C-Programme können sowohl separat als auch in Verbindung mit anderen Programmen genutzt werden. Als Beispiel sind Routinen im Anhang beschrieben, mit deren Hilfe die Tests in SAS genutzt werden können. Desweiteren sind im Anhang SAS-Programme aufgelistet, die zur Berechnung der p-Werte des Isotonen Kontrastes, des Bartholomew-Tests und des Chacko-Tests bei unterschiedlichen Verteilungsapproximationen geeignet sind.

Literaturverzeichnis

[1] **Schumacher**, E., Bleiholder, H., Thöni, H. (1995): Methodische Untersuchungen zur biometrischen Analyse von Boniturwerten aus Freilandversuchen mit Herbiziden. 9 th European Weed research society symposium, Budapest 1995, Proceedings 1, 283-290

[2] **Little**, T.M. (1985): Analysis of percentage and rating scale data. HortScience 20(4), 642-644

[3] **European and Mediterranean Plant protection organization (1986):** Guideline for the biological evaluation of herbicides. Bulletin OEPP EPPO Bulletin 16 (93), 151-167

[4] **Bleiholder**, H. (1989): Methods for the Layout and Evaluation of field trials. BASF, Limburgerhof

[5] **Holtbrügge**, W.(1987): Regressionsmodelle mit ordinalen Zielgrößen und ihre Anwendung in klinischen Studien. Dissertation, Fachbereich Statistik der Universität Dortmund

[6] **Thöni**, H. (1982): Ein statistisches Modell für Bonituren. EDV in Medizin und Biologie 2, 51-56

[7] **Burkhardt**, R., Kienle, G. (1978): Controlled clinical Trials and medical Ethics. Lancet, Dec. 23 /30, 1356-1359

[8] **Fink**, H. (1976): Zur Frage der Zahl der Probanden oder Patienten in klinisch-pharmakologischen Studien. International Journal of Clinical Pharmacology 14, 66-74

[9] **Woggon**, B. (1977): Planung von Psychopharmakaprüfungen. Pharmakopsychiatrie 10,140-146

[10] **Hothorn**, L.A, Hayashi, M., Seidel, D. (2000): Dose-response relationships in mutagenicity assay's including an appropriate positive control group: a multiple testing approach. Enviromental and Ecological Statistics 7, 25-40

[11] **Grömping**, U. (1996): Tests for a monotone dose-response relation in models with ordered categorical dose with emphasis on likelihood ratio tests for linear inequalities on normal means. Dissertation, Fachbereich Statistik der Universität Dortmund

[12] **Ruberg**, S.J.(1989): Contrasts for identifying the minimum effective dose. Journal of the American Statistical Association 84, 816-822

[13] **Chuang-Stein**, C., Agresti, A. (1997) Tutorial in Biostatistics A review of tests for detecting a monotone dose-response relationship with ordinal response data. Statistics in Medicine 16, 2599-2618

[14] **Tamhane**, A.C., Hochberg, Y., Dunnett,C.W .(1996): Multiple test procedures for dose finding. Biometrics 52, 21-37

[15] **Bauer**, P., Röhmel, J., Maurer, W., Hothorn, L.(1998): Testing strategies in multidose trials including active control.Statistics in Medicine 17, 2133-2146

[16] **Agresti**, A. (1984): Analysis of ordinal categorical data. Wiley, New York

[17] **Agresti**, A.(1996): An introduction to categorical data analysis. Wiley, New York

[18] **Stokes**, M.E., Davis, C.S., Koch, G.G.(1995):Categorical data analysis using the SAS system. SAS Institut Inc., Cary, North Carolina

[19] **Mehta**, C.R., Patel, N. (1995): Statxact 3 for windows: Statistical software for exact nonparametric inference, User manual. Cytel software corporation, Cambridge

[20] **Agresti**, A. (1999): Modelling ordered categorical data: Recent advances and future challenges. Statistics in Medicine 18, 2191-2207

[21] **Goodman**, L.A., Kruskal, W.H. (1954): Measures of association for cross classifications. Journal of the American Statistical Association 49, 732-764

[22] **Jonckheere**, A.R. (1954): A distribution-free k-sample test against ordered alternativs. Biometrika 41, 133-154

[23] **Terpstra**, T.J. (1954): A nonparametric test for the problem of k samples. Indagationes Mathematticae 16, 505-512
[24] **Hartung**, J., Elpelt, B., Klösener, K.-H. (1993): Statistik Lehr- und Handbuch der angewandten Statistik. R.Oldenbourg Verlag, München
[25] **Grizzle**, J.E., Starmer,C.F., Koch, G.G. (1969): Analysis of categorical data by linear models. Biometrics 25, 489-504
[26] **Bartholomew**,D.J. (1961): A test of homogeneity of means under restricted alternatives. Journal of the Royal Statistical Society (Series B) 23, 239-281
[27] **Chacko**, V.J. (1963): Testing homogeneity against ordered alternatives. Sankhya (B) 28, 185-190
[28] **McCullagh**, P. (1980): Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society (Series B) 42, 109-142
[29] **Haberman**, S.J. (1980): Contribution to the discussion of a paper by P. McCullagh. Journal of the Royal Statistical Society (Series B) 42, 136-137
[30] **Hamerle**, A., Kemeny, P., Tutz, G. (1984): Kategoriale Regression. In: Fahrmeir,L., Hamerle,A. (Hrsg.): Multivariate statistische Verfahren. W de Gruyter, Berlin
[31] **Krauth**,J. (1988): Distribution – free statistics: An application-oriented approach. Elsevier, Amsterdam
[32] **Robertson**, T., Wright, F.T. (1981): Likelihood ratio tests for and against stochastic ordering between multinomial populations. The Annals of Statistics 9, 1248-1257
[33] **Patefield**, W.M. (1982): Exact tests for trends in ordered contingency tables. Journal of the Royal Statistical Society (Series C) 31, 32-43
[34] **Agresti**, A., Coull, B.A. (1998): Order-restricted inference for monotone trend alternatives in contingency tables. Computational Statistics & Data Analysis 28, 139-155
[35] **Cochran**, W.G. (1954): Some methods of strenthening the common χ^2 tests. Biometrics 10, 417-451
[36] **Agresti**, A., Chuang, C., Kezouh, A. (1987): Order-restricted score parameters in association models for contingency tables. Journal of the American Statistical Association 82, 619-623
[37] **Agresti**, A.(1990): Categorical data analysis. Wiley, New York
[38] **Kimeldorf**, G., Sampson, A.R. (1992): Min and Max Scorings for Two-Sample Ordinal Data. Journal of the American Statistical Association 87,241-247
[39] **Lin**, S., Tang, D.-I. (1995): A hierarchical active constraints search algorithm for optimal scaling of ordered categorical responses. Journal of Statistical Computation and Simulation 53, 197-209
[40] **Graubard**, B.I., Korn, E.L. (1987): Choice of column scores for testing independence in ordered 2xK contingency tables. Biometrics 43, 471-476
[41] **Fleiss**, J.L. (1986): The design and analysis of clinical experiments. Wiley, Ney York
[42] **Stiger**, T.R., Kosinski, A.S., Barnhart, H.X., Kleinbaum, D.G. (1998): Anova for repeated ordinal data with small sample size? A comparision of anova, manaova WLS and GEE methods by simulation. Communications in Statistics- Simulation and Computation 24, 357-375
[43] **Podgor**, M.J., Gastwirth, J.L., Mehta, C.R. (1996): Efficiency robust tests of independence in contingency tables with ordered classifications. Statistics in Medicine 15 (19), 2095-2105
[44] **Gastwirth**, J.L. (1985): The use of maximin efficiency robust tests in combining contingency tables and survival analysis. Journal of the American Statistical Association 80, 380-384
[45] **Hogg,** R.V., Fisher, D.M., Randles, R.H. (1975): A two-sample adaptive distribution free test. Journal of the American Statistical Association 70, 656-661

[46] **Armitage**, R. (1955): Tests for linear trends in proportions and frequencies. Biometrics 11, 375-386

[47] **Chakraborti**, S., Schaafsma, W. (1996): On the choice of scores in Contingency tables. Vortrag gehalten während des Joint Statistical Meetings 1996

[48] **Akritas**, M.G., Brunner, E. (1996): Rank tests for patterned alternatives in factorial designs with interaction. In: Brunner, E. , Denker, M. (Hrsg.):Research developments in probability and statistics Festschrift in honar of Madan L. Puri on the occasion of his 65 th birthday. VSP-International Science Publishers, Utrecht

[49] **Heeren**T., D'Agostino, R.B. (1987): Robustness of the two independent samples t-test when apllied to ordinal scaled data. Statistics in Medicine 6, 79-90

[50] **Snapinn**, S.M., Small, R.D. (1986): Tests of significance using regression models for ordered categorical data. Biometrics 42, 583-592

[51] **Oelerich**, A., Munzel, U. (1998): Testverfahren für ordinale Daten. Vortrag gehalten während des Biometrischen Kolloquiums Mainz 1998

[52] **Reinhard**, I., Krumm, B. (1998): Die Analyse von Summenscores mit Hilfe linearer Modelle für kategoriale Daten. Vortrag gehalten während des Biometrischen Kolloquiums Mainz 1998

[53] **Good**, P. (1994): Permutation tests - A practical guide to resampling methods for testing hypotheses. Springer , New York

[54] **D'Agostino**, R.B., Chase, W., Belanger, A. (1988): The appropriateness of some common procedures for testing two independent binomial populations. American Statistician 42, 198-202

[55] **Efron**, B., Tibshirani, R.J. (1993): An Introduction to the Bootstrap. Chapman and Hall, New York

[56] **Edgington**, E. S. (1995) : Randomization Tests. 3rd Edition, Marcel Dekker, Inc., New York

[57] **Agresti**, A. (1992): A survey of exact inference for contingency tables. Statistical Science 7 (1), 131-177

[58] **Emerson**, J. D., Moses, L.E. (1985): A note on the Wilcoxon-Mann-Whitney test for 2 x k ordered tables. Biometrics 41, 303-309

[59] **Field**, C.A., Ronchetti, E. (1990): Small sample Asymptotics. IMS monograph series 13, Hayward, Institut of Mathematical Statistics, California

[60] **Booth**, J.G., Hall, P., Wood, A.T.A (1992): Edgeworth and saddlepoint approximations for sums of discrete non-lattice random variables. Technical report, Centre for Mathematics & Application, Australian National University, Canberra

[61] **Agresti**, A., Lang, J.B., Mehta, C. (1993): Some empirical comparisons of exact, modified exact and higher order asymptotic tests of independence for ordered categorical variables. Communication in Statistics-Simulation 22(1), 1-18

[62] **Brazzale**, A.R. (1998): Approximate conditional inference in logistic and loglinear models. Vortrag gehalten während des Joint Statistical Meetings Dallas 1998

[63] **Evans**, M., Gilula, Z., Guttman, I., Swartz, T. (1997): Bayesian analysis of stochastically ordered distributions of categorical variables. Journal of the American Statistical Association 92, 208-214

[64] **Adams**, N. M., Kirby, S.P.J., Harris, P., Clegg, D.B. (1995): A review of parallel processing for statistical computation. Statistics and Computing 6, 37-49

[65] **Kaufman**,L., Hopke, P.K., Rousseeuw, P.J. (1988): Using a parallel computer system for statistical resampling methods. Computational Statistics Quarterly 2, 129-141

[66] **Berry**, J.J. (1995): A simulation-based approach to some nonparametric statistics problems. Observations 5, 19-26

[67] **Cochran**, W.G. (1952): The χ^2-test of goodness of fit. Annals of of Mathematical Statistics 23, 315-345

[68] **Gibbons**, J.D. (1985): Permutation tests. In: Kotz, S., Johnson, N.L. (Hsrg.) (1985): Encyclopedia of statistical scienes Volume 6, Wiley, New York

[69] **Edgington**, E.S. (1985): Randomization tests. In: Kotz, S., Johnson, N.L. (Hsrg.) (1985): Encyclopedia of statistical scienes Volume 7, Wiley, New York

[70] **Kempthorne**, O. (1985): Randomization-II. In: Kotz, S., Johnson, N.L. (Hsrg.) (1985): Encyclopedia of statistical scienes Volume 7, Wiley, New York

[71] **Manly**, F.J. (1991): Randomization and Monte Carlo methods in biology. Chapman and Hall, London

[72] **Lehmann**, E.L. (1975): Nonparametrics.Statistical Methods based on ranks. McGraw- Hill, New-York

[73] **Petrondas**, D.A., Gabriel, K.R. (1983): Multiple comparisons by rerandomization tests. Journal of the American Statistical Association 78, 949-957

[74] **Verbeck**, A., Kroonenberg, P.M. (1985): A survey of algorithms for exact distributions of test statistics in *r x c* contingency tables with fixed margins. Computational Statistics & Data Analysis 3, 159-185

[75] **Metha**, C. R., Patel, N. R. (1983): A network algorithm for performing Fisher's Exact test in r x c contingency table. Journal of the American Statistical Association 78 (382),427-434

[76] **Valz**, P.D., Thompson, M.E. (1994): Exact inference for Kendall's S und Spearman's Rho with extensions to Fisher's exact test in *r x c* contingency tables. Journal of Computational and Graphical Statistics 3(4), 459-472

[77] **Hilton**, J., Mehta, C.R. (1993): Power and sample size calculation for exact condi tional tests with ordered categorical data. Biometrics 49, 609-616

[78] **Nijenhuis**,A., Wilf, H.S. (1978): Combinatorial algorithms for Computers and calculators. 2nd Edition, Academic Press, New York

[79] **Mehta**, C.R., Patel, N., Senchaudhuri, P. (1998):Exact power and sample-size computations for the Cochran-Armitage trend test. Biometrics 54 (4), 1371-1379

[80] **Cochran**, W.G. (1954): Some methods for strengthening the common χ^2-test. Biometrics 10, 417-451

[81] **Ludbrock**, J., Dudley,H. (1998): Why permutation tests are superior to t and F tests in biomedical research. The American Statistician 52 (2), 127-132

[82] **Neuhäuser**, M. (1996): Trendtests bei a priori unbekannten Erwartungswertprofil. Dissertation, Fachbereich Statistik der Universität Dortmund

[83] **Mehta**, C.R., Hilton, J.F. (1993): Exact power of conditional and unconditional tests: going beyond the 2x2 contingency table. The American Statistician 47 (2), 91-98

[84] **Koch**, H.-F. (1996): Teststatistiken für die ‚many-to-one' Versuchsanlage im Falle dichotomer Ereignisse. Dissertation, Fachbereich Gartenbau der Universität Hannover

[85] **Haber**, M. (1987): A comparison of some conditional and unconditional exact tests for 2x2 contingency tables. Communications in Statistics-Simulation and Computation 16, 999-1013

[86] **Schrage**, C. (1982): k-Stichprobenpermutationstests bei diskreter Verteilungsannahme. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät der Westfälischen Wilhelms-Universität Münster

[87] **Lancaster**, H.O. (1961): Significance tests in discrete distributions. Journal of the American Statistical Association 56, 223-234

[88] **Kim**, D., Agresti, A. (1995): Improved exact inference about conditional association in three-way contigency tables. Journal of the American Statistical Association 90 (430), 632-639

[89] **Barnard**, G.A. (1990): Must clinical trials be large? The interpretation of p-value and the combination of test results. Statistics in Medicine 9, 601-614

[90] **Cohen**, A., Sackrowitz, H. B. (1992): An evaluation of some tests of trend in

contigency tables. Journal of the American Statistical Association 87 (418), 470-475
[91] **Berger**, V., Sackrowitz, H. (1997): Improving tests for superior treatment in contingency tables. Journal of the American Statistical Association 92 (438), 700-705
[92] **Berger**, V., Permutt, T., Ivanova, A. (1998): Convex hull test for ordered categorical data. Biometrics 54, 1541-1550
[93] **Behnen**, K., Neuhaus, G. (1989): Ranktests with estimated scores and their apllications. Teubner, Stuttgart
[94] **Dwass**, M. (1957): Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181-187
[95] **Jöckel**, K.-H. (1982): Eigenschaften und effektive Anwendung von Monte-Carlo-Tests. Dissertation, Abteilung Statistik der Universität Dortmund
[96] **Efron**, B. (1979): Bootstrapmethods: Another look at the jackknife. Annals of Statistics 7, 1-26
[97] **Beran**, R. (1988): Prepivoting test statistics: A bootstrap view of asymptotic refinements. Journal of the American Statistical Association 83 (403), 687-697
[98] **Hall**, P., Wilson, S.R. (1991): Two guidlines for bootstrap hypothesis testing. Biometrics 47, 757-762
[99] **Davison**, A.C., Hinkley, D.V. (1997): Bootstrap methods and their application. Cambridge University Press
[100] **Shao**, J., Tu, D. (1995): The Jackknife and Bootstrap. Springer, New York
[101] **Fisher**, N.I., Hall, P. (1990): On Bootstrap hypothesis testing. Australian Journal of Statistics 32 (2), 177-190
[102] **Romano**, J.P. (1989): Bootstrap and randomization tests of some nonparametric hypothesis. The Annals of Statistics 17 (1), 141-159
[103] **Fisher**, N.I., Hall, P. (1991): Bootstrap algorithms for small samples. Journal of Statistical Planning and Inference 27, 157-169
[104] **Bishop**, Y.M.M., Fienberg, S.E., Holland, P.W. (1975): Discrete multivariate analysis: Theorie and Practice. The MIT Press, Cambrigde
[105] **Serfling**, R.J. (1980): Approximation theorems of mathematical statistics. Wiley, New York
[106] **Young**, G.A. (1994): Bootstrap: More than a stab in the dark? Statistical Science 9(3), 382-415
[107] **Efron**, B. (1982): The Jackknife, the bootstrap and other resampling plans. Society for industrial and applied mathematics. Philiadelphia
[108] **Shi**, S.G.(1992): Accurate and efficient double-bootstrap confidence limit method. Computational Statistics & Data Analysis 13, 21-32
[109] **Box**, G.E.P. (1953): Non-normality and tests on variance. Biometrika 40, 318-335
[110] **Quebe-Fehling**, E. (1990): Mehrfaktorielle Varianzanalyse mit ordinal-kategorialen Zielgrössen-Aspekte bei der Analyse landwirtschaftlicher Sortenversuche mit Bonituren. Dissertation, Fakultät für Mathematik der Ruhr-Universität Bochum
[111] **Thöni**, H. (1985): Auswertung von Bonituren: ein empirischer Methodenvergleich. EDV in Medizin und Biologie 16 (3), 108-114
[112] **Thöni**, H. (1992): Auswertung von Bonituren: Ein empirischer Methodenvergleich. Biometrie und Informatik in Medizin und Biologie 23(3), 144-156
[113] **Wright**, The one-way analysis of variance with ordered alternatives: a modifikation of Bartholomew's $\overline{E}^2$ test. The Canadian Journal of Statistics 16(1), 75-85
[114] **Robertson**, T., Wright, F.T., Dykstra, R.L. (1988): Order restricted statistical inference .Wiley, New York
[115] **Bretz**, F. (1999): Powerful modification of Williams'trend test. Dissertation, Fachbereich Gartenbau der Universität Hannover
[116] **Roth**, A. J. (1983): Robust trend tests derived and simulated: analogs of the Welch

and Brown and Brown-Forsythe Tests. Journal of the American Statistical Association 78 (384), 972-980

[117] **Grimes**, B., Federer, W.T. (1984): Comparison of means from populations with unequal variances. In: Rao, P.S.R.S., Sedransk, J. (Hrsg.):W. G. Cochran's impact on statistics. Wiley, New York

[118] **Bohrer**, R., Chow, W. (1972): Algorithm AS 122: Weights for One-sided multivariate inference. Applied Statistics 27, 100-104

[119] **Tang**, D.-I., Gnecco, C., Geller, N.L. (1989): An approximate likelihood ratio test for a normal mean vektor with nonnegativ components with application to clinical trials. Biometrika 76, 577-583

[120] **Akkerboom**, J. C. (1990): Testing problems with linear or angular inequality constraints. Springer-Verlag, New York

[121] **Shi**, Yingqi, Meng, Cliff (1991): Bootstrap trend test procedures based on isotonic regression. American Statistical Association Proceedings Biopharmacetical Section, 256-262

[122] **Welch**, B.L. (1951): On the comparison of serveral mean values: an alternative approach. Biometrika 38, 330-336

[123] **Brown**, M.B., Forsythe, A.B. (1974): The small sample behavior of some statistics which test the equality of serveral means. Technometrics 16, 385-389

[124] **Meng**, Cliff Y.K., Davis, S.B., Roth, A. J. (1993): Robust contrast based trend tests. American Statistical Association Proceedings Biopharmacetical Section, 127- 132

[125] **Williams**, D.A.(1971): A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27, 103-117

[126] **Marcus**, R. (1976): The power of some tests of the equality of normal means against an ordered alternative. Biometrika 63, 177-183

[127] **Fligner**, M.A., Wolfe, D.A. (182): Distribution-free tests for comparing several treatments with a control. Statistica 36, 119-127

[128] **Hothorn**, L.A., Lehmacher, W (1991): A simple testing procedure ‚control versus k treatments' for one sided ordered alternatives, with aplication in toxicology. Biometrical Journal 33, 179-189

[129] **Hogg**, R.V. (1965): On models and hypotheses with restricted alternatives. American Statistical Association Journal Dezember, 1153-1162

[130] **Abelson**, R.P., Tukey, J.W. (1963): Efficient utilization of non-numerical information in quantitative analysis. General theory and the case of the simple order. Annals of Mathematical Statistics 34, 1347-1369

[131] **Schaafsma**, W., Smid, L.J. (1966): Most stringent somewhere most powerful test against alternatives restricted by a number of linear inequalities. Annals of Mathematical Statistics 37 (5), 1161-1172

[132] **Cohen**, A., Sackrowitz, H.B. (1992): Improved tests for comparing treatments against a control and other one-sited problems. Journal of the American Statistical Association 87 (420), 1137-1144

[133] **Cohen**, A., Sackrowitz, H.B. (1993): Inadmissibility of studentized tests for normal order resticted models. The Annals of Statistics 21 (2), 746-752

[134] **Bechhofer**, R.E., Dunnett, C.W. (1955): Multiple comparisons for orthogonal contrasts: example and table. Technometrics 24, 213-222

[135] **McDermott**, M.P., Mudholkar, G.S. (1993): A simple approach to testing homogeneity of order–constrained means. Journal of the American Statistical Association 88 (424), 1371-1379

[136] **Sugiura**, N. (1994): Approximating bayes critical region for testing simple and tree ordered normal means. Journal of Statistical Research 28 (1&2), 1-20

[137] **Hirotsu**, C., Kuriki, S., Hayter, A. (1992): Multiple comparison procedures based on

the maximal component of the cumulative chi-squared statistic. Biometrika 79 (2), 381-392

[138] **Genz**,A.,Bretz,F. (1998): Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation and Simulation 63, 361-378

[139] **Hochberg**, Y., Tamhane, A.C. (1987): Multiple comparison procedures. Wiley, New York

[140] **Zhou**, J.L., Tits, A.L. (1994): User's guide for FSQP, Version 3.4: A Fortran code solving constrined nonlinear (minimax) optimization problems, generating iterats satisfying all inequality and linear constraints, technical Report TR-92-107r4, University Maryland, System Research Center

[141] **Wang**, Y. (1996): A likelhood ratio test against stochastic ordering in several population. Journal of the American Statistical Association 91 (436), 1676-1682

[142] **Mohberg**, N.R., Ghosh, M., Grizzle, J.E. (1978): linear models analysis of small samples of categorized ordinal response data. Journal of Statistical Computation & Simulation 7, 237-252

[143] **Berkson**, J. (1955): Maximum-Likelihood and minimum estimation of the logistic function. Journal of the American Statistical Association 50, 130-162

[144] **Hirotsu**, C. (1982): Use of cumulativ efficient scores for testing ordered alternatives in diskrete models. Biometrika 69, 567-577

[145] **Oelerich**, A. (1998): Teststatistiken zur Analyse ordinaler Daten bei kleinen Stichprobenumfängen. Diplomarbeit, Institut für Mathematische-Stochastik der Georg-August-Universität Göttingen

[146] **Brunner**,E., Puri, M.L. (1996): Nonparametric methods in design and analysis of experiments, In: Ghosh, S., Rao, C.R. (Hrsg.): Handbook of statistics 13, Elsevier Science

[147] **Bross**, I.D.J. (1958): How to use ridit analysis. Biometrics 14, 18-38

[148] **Shorack**, G.R. (1967): Testing against ordered alternatives in model 1 analysis of variance; normal theory and nonparametric. Annals of Mathematical Statistics 38, 1740-1753

[149] **Billingsley**, P. (1971): Weak convergence of measures: Application in probability. Society for Industrial and Applied Mathematics, Philadelphia

[150] **Munzel**, U. (1996): Multivariate nichtparametrische Verfahren für feste Faktoren in mehrfaktoriellen Versuchsanlagen. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät der Georg-August-Universität zu Göttingen

[151] **Bregenzer**, T. (1998): Direktionale Tests zur Auswertung klinischer Studien mit multiplen Endpunkten bei unvollständigen Daten. Dissertation, Medizinische Fakultät der Universität zu Köln

[152] **Shiraishi**, T. (1982): Testing homogeneity against trend based on rank in one –way layout. Communication in Statistics (A) 11, 1255-1268

[153] **Conover**, W.J., Iman, R.I. (1976): On some alternative procedures using ranks for the analysis of experimental designs. Comunications in Statistics 5, 1349-1368

[154] **Steland**, A. (1996): Bootstraping linearer Rangstatistiken mit Anwendungen in semiparametrischen Modellen. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät der Georg-August-Universität Göttingen

[155] **Brunner**, E., Puri, M.L. (1996): A class of rank-score tests in factorial designs. Preprint.

[156] **Hettmansperger**, T.P., Norton, R.M. (1987): Tests for patterned alternatives in k-sample problems. Journal of the American Statistical Association 82 (397), 292-299

[157] **Puri**, M.L., Sen, P.K. (1971): Nonparametric methods in multivariate analysis. Wiley, New York

[158] **Conover**, W.J. (1973): Rank tests for one sample, two samples and k samples without the assumption of a continuous distribution function. The Annals of Statistics 1 (6), 1105-1125
[159] **Hogg**, R.V., Fisher, D.M., Randles, R.H. (1975): A two sample adaptive distribution-free test. Journal of the American Statistical Association 70, 656-661
[160] **Hill**, N.J., Padmanabhan, A.R., Puri, M.L. (1988): Adaptive nonparametric procedures and application. Applied Statistics 37(2), 205-218
[161] **Beier**, F., Büning, H. (1997): An adaptive test against ordered alternatives. Computational Statistics & Data Analysis 25, 441-452
[162] **Gastwirth**, J.L. (1965): Percentile modifications of two sample rank tests. Journal of the American Statistical Association 60, 1127-1141
[163] **Büning**, H. (1996): Adaptive tests for the c-sample location problem - the case of twosided alternatives. Communication in Statistics- Theory and Methods 25, 1569-1582
[164] **Wilcoxon**, F. (1945). Individual comparisons by ranking methods. Biometrics 1, 80-83
[165] **Seidel**, D., Neuhäuser, M., Hothorn, L.A., Urfer, W. (1998): Adaptive trend tests with application to mutagenicity data. Vortrag gehalten während des Biometrischen Kolloquiums Mainz 1998
[166] **Neuhäuser**, M., Seidel, D., Hothorn, L.A., Urfer, W.(2000): Robust trend tests with application to toxicology. Enviromental and Ecological Statistics 7, 43-56
[167] **Donegani**, M. (1992): A Bootstrap adaptiv test for two-way analysis of variance. Biometrical Journal 34 (2), 141-146
[168] **Hothorn**, L.A. (1987): k-Stichprobentests und –vergleichsprozeduren in Dosis-Wirkungs-Abhängigkeiten toxikologischer Untersuchungen – Eine biometrische Analyse. Dissertation B, Fakultät für Naturwissenschaften der Martin-Luther-Universität Halle-Wittenberg
[169] **Selwyn**, M.R. (1995): The use of trend tests to determine a no-observable-effect level in animal safety studies. Journal of the AmericanCollege of Toxicology 14, 158-168
[170] **Beer**, E., Grigo, E., Kaspers, H., Frahm, J., Martin, J., Mielke, H., Montag, H., Prillwitz, H.G., Radtke, W., Schreiber, B. (1988): Richtlinien für die amtliche Prüfung von Pflanzenschutzmitteln Teil II. Biologische Bundesanstalt für Land- und Forstwirtschaft. Braunschweig
[171] **Cohen**, A., Sackrowitz, H.B., Sckrowitz, M. (1998): Testing whether treatment is „better“ than control with ordered categorical data: An evaluation of new methology. Rutgers Technical Report #98-004, Rutgers University
[172] **Boyet**, J.M. (1979): Random r x c tables with given row and column totals (Algorithm AS 144). Applied Statistics 28, 329-332
[173] **Patefield**, W.M.(1981): Algoritm AS 159. An efficient method of generating random r x c tables with given row and column totals. Applied Statatictics 30, 91-97
[174] **Fleishman**, A.I. (1978): A method for simulating non-normal distributions. Psychometrika 43(4), 521-532
[175] **Hilton**, J.F. (1996): The appropriateness of the Wilcoxon test in ordinal data. Statistics in Medicine 15, 631-645
[176] **Lesaffre**, E., Scheys, I., Fröhlich, J., Bluhmki, E. (1993): Calculatione of power and sample size with bounded outcome scores. Statistics in Medicine 12, 1063-1078
[177] **Mann**, H.B., Whitney, D.R. (1947): On a test of whether one of two random variables is stochastical larger than the other. Annals of of Mathematical Statistics 18, 50-60
[178] **Duchateau**, L. (1999): Small Vaccination Experiments with binary outcome: The paradox of increasing power with decreasing sample size and/or increasing imbalance. Biometrical Journal 41 (5), 583-600

[179] **Gross**, S.T. (1981): On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. Journal of the American Statistical Association 76(376), 935-941

[180] **Holm, S.** (1979): A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65-70

[181] **Maurer, W.**, Hothorn, l., Lehmacher, W. (1995): Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypothesis. In: Vollmar, J. (Hrsg.): Biometrie in der chemisch-pharmazeutischen Industrie, Band 6, Fischer Verlag, Stuttgart, 3-21

[182] **Marcus, R.,** Peritz, E., Gabriel, K.B. (1976): On closed testing procedures with special references to ordered analysis of variance. Biometrika 63, 655-660

[183] **Hothorn, L.A.** (1997): Modifications of the closure principle for analyzing toxicological studies. Drug Information Journal 31, 403-412

[184] **Williams, D.A.** (1986): A note on Shirley's non-parametric test for comparing several dose levels with a zero-dose control. Biometrics 42, 183-186

[185] **Shirley, E.A.** (1977): A nonparametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics 33, 386-389

[186] **Wellek, S.**, Hampel, B. (1999): A distribution-free two sample equivalence test allowing tied observations. Biometrical Journal 41(2), 171-186

[187] **Hauschke, D.**, Kieser, M., Diletti, E. Burke, M. (1999): Sample size determination for proving equivalencebased on the ratio or two means for normally distributed data. Statistics in Medicine 18, 93-105

[188] **Bauer, P.**, Budde, M. (1994): Multiple testing for detecting efficient dose steps. Biometric Journal 36 (1), 3-15

[189] **Weichert, M.** (2000): Robuste Mittelwertvergleiche mit gartenbaulichen Anwendungen. Dissertation, Fachbereich Gartenbau der Universität Hannover

[190] **Yoshimura, I.**, Wakana, A., Hamada, C. (1997): A performance comparison of maximum contrast methods to detect dose dependency. Drug Information Journal 31, 423-432

[191] **Röhmel, J.** (1996): Precision intervals for estimates of the difference in success rates for binary random variables based on the permutation principle. Biometrical Journal 38 (8), 977-993

[192] **Kang, S.-H.**, Klotz, J.H. (1999): Updating exact p-values for the conditional likelihood ratio test of independence. Journal of Statistical Computation and Simulation 63, 201-215

Anhang A Exakte Mid-p-Tests

Die beiden folgenden SAS-Macros dienen dazu, C-Programme zur Berechnung von Mid-p-Werten aufzurufen. Innerhalb der Macros werden die Daten in eine Textdatei geschrieben, die einem C-Programm als Parameter übergeben werden. Das C-Programm schreibt die berechneten p-Werte in eine Textdatei, die dann von SAS eingelesen und ausgegeben werden. Ein ähnliches Excel-Add-In-Programm und das C-Programm befindet sich auf der beiliegenden CD. Die Programme sind auch auf der Homepage des Lehrgebiets Bioinformatik am Fachbereich Gartenbau der Universität Hannover zu finden (http://bioinf.uni-hannover.de)

```
/*************************************************************************/
/* Macro zur Durchführung eines exakten Mid-p-Tests für zwei unabhängige */ /* Stichproben                                                           */
/*1.exakter t-Test (raenge = 0)                                          */
/*2.exakter Wilcoxon-Mann-Whitney-Test (raenge=1)                        */
/*************************************************************************/
%MACRO midp2sam(
 daten=_LAST_,     /*Datensatz */
 var=,             /*Zielvariable */
 class=,           /*Gruppierungsvariable */
 c_prog=           /*Name des C-Programmes*/
 tempdat=,         /*Textdatei, in der die Daten gespeichert werden. */
 temp_p=,          /*Textdatei, in die das C-Programm die P-Werte schreibt.*/
 raenge=1,         /* Rangtest: raenge=1, Test mit Originaldaten: raenge=0 */
 );
DATA test0;
 SET &daten;
PROC SORT DATA=test0;
 BY &class;
 RUN;
PROC IML;
 USE test0;
 READ all var {&class}  into gr;
 READ all var {&var}  into bonitur;
 CLOSE test0;
 newclass=j(nrow(gr),1,1);
 DO i=2 TO nrow(gr);
  IF gr[i]=gr[i-1] THEN newclass[i]=newclass[i-1];
```

```
                    ELSE newclass[i]=newclass[i-1]+1;
 END;
 CREATE tempdata var{gr newclass bonitur}; APPEND;CLOSE tempdata;
QUIT;
%IF &raenge=1 %then
 %DO;
  PROC RANK DATA=tempdata OUT=tempdata TIES=mean;
   VAR &var;
   RANKS &var;
  DATA tempdata;
   SET tempdata;
   &var=2*&var;
 %END;
DATA tempdata;
 File "&tempdat";
 SET tempdata;
 PUT newclass &var;
RUN;
%PUT &c_prog &tempdat &temp_p;
X "&c_prog &tempdat &temp_p";
RUN;
DATA tempdata;
 SET tempdata;
 IF &raenge=1 THEN &var=&var/2;
RUN;
PROC MEANS DATA=tempdata ;
 VAR &var;
 CLASS &class;
RUN;
DATA ergs;
 INFILE "&temp_p" dlm=";";
 INPUT x1 x2 x3 x4;
 LABEL x1="P(T=t_0)"; LABEL x2="P(T>t_0)";
 LABEL x3="P(T>=t_0)"; LABEL x4="P(T>t_0)+0.5 P(T=t_0))";
RUN;
%IF &raenge=0 %THEN %LET
                    titel="Exakter Permutationstest mit Originaldaten";
%IF &raenge=1 %THEN %LET titel= "Exakter Permutationstest mit Rängen";
PROC PRINT DATA=ergs LABEL NOOBS; TITLE &titel;
 TITLE2 "------------------------------------------------------";
 TITLE3 "T=Differenz der Mittelwerte der Daten bzw. Ränge (Gruppe 2 -
Gruppe 1)";
 TITLE4 "------------------------------------------------------------";
RUN;%MEND;
/* Beispiel */
DATA test;
 INPUT gr bonitur;
 CARDS @@;
0 3 0 2 0 2 0 3 1 5 1 6 1 5 1 5
```

```
;
/************************ Aufruf des Macros ************************/
%midp2sam(daten=test, var=bonitur, class=gr, c_prog=C:\Temp\mid_p_testfor-
twosamples.exe, tempdat=c:\temp\dat1.txt,temp_p=C:\temp\dat2.txt,raenge=0);
/************************ Output *********************************/
Exakter Permutationstest mit Originaldaten
-----------------------------------------------------------
T=Differenz der Mittelwerte der Daten bzw. Ränge (Gruppe 2 - Gruppe 1)
t_0=Beobachteter Wert der Teststatistik
-------------------------------------------------------------------------------
Analysis Variable : BONITUR
          GR  N Obs       Minimum          Mean       Maximum       Std Dev
-------------------------------------------------------------------------------
           0      4     2.0000000     2.5000000     3.0000000     0.5773503
           1      4     5.0000000     5.2500000     6.0000000     0.5000000
-------------------------------------------------------------------------------
Exakter Permutationstest mit Originaldaten
-----------------------------------------------------------
T=Differenz der Mittelwerte der Daten bzw. Ränge (Gruppe 2 - Gruppe 1)
t_0=Beobachteter Wert der Teststatistik
-------------------------------------------------------------------------------
                                    P(T>t_0)+0.5
P(T=t_0)    P(T>t_0)    P(T>=t_0)      P(T=t_0))
0.014286         0        0.014286       .007143

/*************************************************************************/
/*Macro zur Durchführung eines exakten Mid-p-Tests für mehr als zwei     */
/*unabhängige Stichproben                                                */
/*1.exakter Bartholomew-Test (raenge = 0)                                */
/*2.exakter Chacko-Test (raenge=1)                                       */
/*************************************************************************/
%MACRO midpksam(
daten=_LAST_, /* Datensatz */
var=,         /* Zielvariable */
class=,       /* Gruppierungsvariable */
c_prog=       /* Name des C-Programmes*/
anstieg=1,    /* 1=Test auf Anstieg, 0=Test auf Abfall*/
tempdat=,     /* Textdatei, in der die Daten gespeichert werden. */
temp_p=,      /* Textdatei, in die das C-Programm die P-Werte schreibt.*/
raenge=1,     /* Rangtest: raenge=1, Test mit Originaldaten: raenge=0 */
);
data test0;
 set &daten;
title ;title2;title3;title4;
proc sort data=test0;
 by &class;
run;
```

```
proc iml;
use test0;
read all var {&class}  into gr;
read all var {&var}  into bonitur;
close test0;
newclass=j(nrow(gr),1,1);
do i=2 to nrow(gr);
 if gr[i]=gr[i-1] then newclass[i]=newclass[i-1];
                  else newclass[i]=newclass[i-1]+1;
end;
if &anstieg ^=1 then bonitur=(-1)*bonitur;
create tempdata var{gr newclass bonitur};append;close tempdata;
quit;
%if &raenge=1 %then
 %do;
   proc rank data=tempdata out=tempdata ties=mean;
    var &var;
    ranks &var;
   run;
   data tempdata;
    set tempdata;
    &var=2*&var;
 %end;
data tempdata;
 File "&tempdat";
 set tempdata;
 put newclass &var;
run;
%put &c_prog &tempdat &temp_p;
x "&c_prog &tempdat &temp_p";
run;
data tempdata;
 set tempdata;
 if &raenge=1 then &var=&var/2;
 else if &anstieg = 0 then &var=&var * (-1);
 output;
run;
proc means data=tempdata min mean max std ;
 var &var;
 class gr;
run;
data ergs;
infile "&temp_p" dlm=";";
input x1 x2 x3 x4;
label x1="P(T=t_0)";label x2="P(T>t_0)";
label x3="P(T>=t_0)";label x4="P(T>t_0)+0.5 P(T=t_0))";
run;
%if &raenge=0 %then %let titel= "Exakter Permutationstest mit Originaldaten
- Bartholomew-Test";
```

```
%if &raenge=1 %then %let titel= "Exakter Permutationstest mit Rängen -
Chacko-Test";
proc print data=ergs label noobs;
title &titel;
title2 "------------------------------------------------------";
title3 "T=Summe der Quadrate der isotonen Schätzer  ";
title4 "------------------------------------------------------";
run;
title ;title2;title3;title4;
%mend;



/********************** Beispiel ********************************/
DATA test;
 INPUT gr bonitur @@;
 CARDS;
 0 9 0 2 0 2 0 3 2 5 2 6 2 5 2 5 3 8 3 9 3 9 3 9 4 8 4 9 4 9 4 9
;
/************************* Aufruf des Macros ***********************/
%midpksam(daten=test, var=bonitur, class=gr, anstieg=1, c_prog=C:\temp\-
mid_p_LRT_for_k_samples.exe, tempdat=c:\temp\dat1.txt,
temp_p=C:\temp\dat2.txt,raenge=0,

/************************* Output **********************************/
```

Exakter Permutationstest mit Originaldaten

```
Analysis Variable : BONITUR
          GR  N Obs       Minimum          Mean       Maximum       Std Dev
-------------------------------------------------------------------------------
           0      4     2.0000000     4.0000000     9.0000000     3.3665016
           2      4     5.0000000     5.2500000     6.0000000     0.5000000
           3      4     8.0000000     8.7500000     9.0000000     0.5000000
           4      4     8.0000000     8.7500000     9.0000000     0.5000000
-------------------------------------------------------------------------------
Exakter Permutationstest mit Originaldaten - Bartholomew-Test
-----------------------------------------------------------
T=Summe der Quadrate der isotonen Schätzer
t_0=Beobachteter Wert der Teststatistik
-----------------------------------------------------------
                                      P(T>t_0)+0.5
P(T=t_0)    P(T>t_0)    P(T>=t_0)       P(T=t_0))
 .000075     .001931     .002007         .001969
```

Anhang B SAS/IML-Routinen

Im folgenden sind IML-Routinen aufgelistet, mittels derer p-Werte für multiple Kontraste (Scorestatistiken) und für Tests, die auf isotonen Schätzern beruhen, berechnet werden können. Sowohl Monte-Carlo-Bootstraptests als auch approximative Permutationstests und Tests in Verbindung mit multivariaten Verteilungen, gekoppelt mit verschieden Varianzschätzern, können berechnet werden. Wichtig sind die beiden Funktionen *Multcont* und *LRT*. Mittels der im Funktionskopf dieser beiden Routinen beschriebenen Parameter sind die unterschiedlichen Testvarianten selektierbar. Alle anderen Funktionen sind nur Hilfsroutinen.

```
PROC IML;
/************************************************
Permutieren eines Vektors
************************************************/
START permut(feld, n_ges, st_wert);
seed=st_wert; nm=n_ges-1; z=0;
DO i=1 TO nm;
 j=n_ges-i+1; CALL ranuni(seed, z); k=int(j*z)+1;
 IF(k>=j) THEN k=j;
 z=feld[j]; feld[j]=feld[k]; feld[k]=z;
END;
st_wert=seed;
FINISH;
/************************************************
Bootstrappen eines Vektors in gepoolter Stichprobe
************************************************/
START bootf(feld, n_ges, st_wert);
seed=st_wert;
IF ncol(feld)>1 THEN feld1=t(feld); ELSE feld1=feld;
u=J(n_ges,1,0); CALL ranuni(seed,u);
i=int(n_ges*u+j(n_ges,1,1)); feld1=feld1(|I,|);
IF ncol(feld)>1 THEN feld=t(feld1); ELSE feld=feld1;
st_wert=seed;
FINISH;
/*****************************************************
Bootstrappen eines Vektors in den einzelnen Stichproben
*****************************************************/
START boots(feld, c, n, st_wert);
seed=st_wert; top=0;
DO i=1 TO c; dat=feld[top+1:top+n[i]]; RUN bootf(dat,n[i],seed);
 IF i=1 THEN neu=dat; ELSE neu=neu//dat;
 top=top+n[i];
END;
feld=neu; st_wert=seed;
FINISH;
```

```
/*****************************************************
Berechnung der gepoolten Varianz
*****************************************************/
START varipool(daten, grmittel, grumfang, granz, ngesamt);
varianz=1/(ngesamt-granz)*(sum(daten##2)-T(grumfang)*grmittel##2);
RETURN(max(0,varianz));
FINISH;
/*****************************************************
Berechnung der gesamten Varianz
*****************************************************/
START variges(daten, gesmit, granz, ngesamt);
varianz=1/(ngesamt-1)*(sum(daten##2)-ngesamt*gesmit##2);
RETURN(max(0,varianz));
FINISH;
/*****************************************************
Berechnung der Korrelationsmatrix für einen Kontrasttest
*****************************************************/
START korrel(k, n, gew_mat, korr2);
inv_ni=I(k);
DO i=1 TO k; inv_ni[i,i]=1/sqrt(n[i]); END;
korr=gew_mat* inv_ni; korr=korr*T(korr); inv_ni=I(nrow(gew_mat));
DO i=1 TO nrow(gew_mat); inv_ni[i,i]=1/sqrt(korr[i,i]);END;
korr2=inv_ni*korr*inv_ni;
FINISH;
/* *****************************************************
Berechnung der Isotonen Schätzer
********************************************** */
START ordnen(mittel, gewicht, c, geordnet);
i=0; j=0; min=1; s=0; bis_j=0; bis_jp1=0; gew_bisj=0; g_bisjp1=0;
geordnet[c]=mittel[c];
DO i=1 TO c-1;
 DO;
  bis_j=mittel[i]; gew_bisj=gewicht[i]; geordnet[i]=mittel[i]; min=i;
  DO j=i+1 TO c;
   DO;
    g_bisjp1=gew_bisj+gewicht[j];
    bis_jp1=(gew_bisj*bis_j+gewicht[j]*mittel[j])/g_bisjp1;
    IF (geordnet[i]>=bis_jp1) THEN DO;min=j;geordnet[i]=bis_jp1;END;
    gew_bisj=g_bisjp1; bis_j=bis_jp1;
   END;
  END;
  DO s=i TO min; geordnet[s]=geordnet[i]; END; i=min;
 END;
END;
FINISH;
/*****************************************************
Bestimmt bei iterativer Anwendung alle möglichen Zerlegungen einer
natürlichen Zahl in c Summanden (Funktion basiert auf einem bei
Nijenhuis,A., Wilf, H.S. [78] beschriebenen Algorithmus)
*****************************************************/
```

```
START NEXCOM2(n_ges, c, feld, mtc, t, h);
IF (mtc=0) THEN Do;
  feld[1]=n_ges;  IF(c>1) THEN DO i=2 TO c; feld[i]=0;end;
  t=n_ges;h=0;
 END;
 ELSE GOTO marke2;
marke1: IF (feld[c]=n_ges) THEN mtc=2; ELSE mtc=1;
return(1);
marke2:
 IF (t>1) THEN h=0;
 h=h+1; t=feld[h]; feld[h]=0; feld[1]=t-1; feld[h+1]=feld[h+1]+1;
GOTO marke1;
FINISH;
/*********************************************************
Erzeugt alle möglichen Zerlegungen einer natürlichen Zahl in k Summanden
*********************************************************/
START erzall(n,k,mengen);
feld1=j(1,k,0); first=0; mtc1=0; t=0; h=0;
DO WHILE (mtc1<2);
 i=nexcom2(n, k, feld1, mtc1,t,h);
 IF all(feld1) THEN IF first=0 THEN DO;mengen=feld1;first=1; END;
                               ELSE mengen=mengen//feld1;
END;
return(nrow(mengen));
FINISH;
/*************************************
Berechnet die Stirlingschen Zahlen 2. Art (Für balancierten LRT) (Funktion
basiert auf einen bei Nijenhuis,A., Wilf, H.S. [78] beschriebenem Algo-
rithmus)
*************************************/
START stirling( k, a );
 gamma=1; n1=k; eps=1; w=-k*eps;
 DO m=1 TO n1;
  val=0; z=w;
  DO i=m TO k;
   z=z+eps; val=a[k+m-i]+z*val; IF (gamma=1) THEN  a[k+m-i]=val;
  END;
  w=w+1;
 END;
FINISH;
R=0;
/*************************************************
Orthantwahrscheinlichkeiten für multivariate Normalverteilung
(P(X1<0, X2<0, X3<0)) Funktion stammt von F. Bretz [115].
*************************************************/
START mvnorth(r,b,eps);
q=nrow(r); c=t(root(r))+1E-12; f=j(1,q,0); y=f;
f[1]=probnorm(b[1]/c[1,1]); e=f; n=10; vec=0:q-2;
p_vector={157 313 619 1249 2503 5003 10007 20011 40009};
mat={ 1   1   1   1   1    1    1     1     1,
```

```
       46 119 239 512 672 1850 3822  6103 15152,
       46  93 178 136 652 1476 2325  2894  8789,
       17  51  73 197 792  792 1206  8455  9023,
       18  51 104 165 792  380 1927  3629  5632,
       18  80 102 175 253  162 2286  1752  1542,
       11  70 161 303 306  363 3920  1920  2638,
       11  70 161 155 153  137  378   652 11200,
       11  93 106  18 288  520 2240   146   472,
       36  62 106 184 288  186  752   704   144,
       36  15  80  27  29   33 1024   704   144,
       36  15  80 160 288  240 1024   704   144,
        4  44 128  24 144  240 1024   704  1344,
       36  15 128  24  16  240 1024    88  5632,
       36   9 128  24  44  160  144    88  5632,
       36   9  48  24  44  176  144    44  5632,
       36  20  48  24  16  176  768    88   144,
       46  20  48  24  14  112  768    88   144,
       36  20  32  28  64  112  768    80    22};
DO UNTIL (n>50 | error<eps);
 index=1;
 DO UNTIL(index=10 | error<eps);
  p=p_vector[index]; h=mat[q-1,index]; z=mod(j(1,q-1,h)##vec,p);
  intval=0; varsum=0;
  DO l=1 TO n;
   latsum=0; rr=ranuni(j(1,q-1,141071));
   DO j=1 TO p;
    w=abs(2*mod(rr+j#z/p,1)-1);
    DO i=2 TO q;
     y[i-1]=probit(w[i-1]#e[i-1]+1E-12);
     e[i]=probnorm((b[i]-sum(c[i,1:i-1]*y[1:i-1]))/c[i,i]);
    END;
    f=e[#]; latsum=latsum+(f-latsum)/j;
   END;
   varsum=varsum+(l-1)*(latsum-intval)**2/l;
   intval=intval+(latsum-intval)/l;
  END;
  error=3*sqrt(varsum/(n*(n-1))); index=index+1;
 END;
 n=n+2;
END;
prob=intval; return(prob);
FINISH;
/*********************************************************
Funktion, die bei Berechnung der Orthantwahrscheinlichkeiten
auftritt und integriert werden muss.
*********************************************************/
START outer3(v) global (R);
return(1/sqrt(1-v*v)*arsin((R[1,2]*(-1)*sqrt(1-v*v))/sqrt(1-v*v-
R[2,3]*R[2,3])));
FINISH;
```

```
/************************************************
Exakte Berechnung der Orthantwahrscheinlichkeiten für
einige Spezialfälle (k<6) und sonst mittels mvnorth.
************************************************/
START orthant(k,gew,eps) global (R);
pi=3.1415926535898; IF k-1=1 THEN return(0.5);
R=j(k-1,k-1,0);
DO i=1 to k-1;
 R[i,i]=1;
 IF i<k-1 THEN DO;
   R[i,i+1]=-sqrt(gew[i]*gew[i+2]/( (gew[i]+gew[i+1])*
   (gew[i+1]+gew[i+2]))); R[i+1,i]=R[i,i+1];
  END;
END;
IF k-1=2 THEN return(0.25+0.5*arsin(R[1,2])/pi);
R=j(k-1,k-1,0);
DO i=1 TO k-1;
 R[i,i]=1;
 IF i<k-1 THEN DO; R[i,i+1]=-sqrt(gew[i]*gew[i+2]/
                      ( (gew[i]+gew[i+1])* (gew[i+1]+gew[i+2])));
                R[i+1,i]=R[i,i+1];
               END;
END;
IF k-1=3 THEN return(1/8+1/(4*pi)*(arsin(R[1,2])+arsin(R[2,3])));
IF k-1=4 THEN DO;
 q=0;a=0||-R[3,4]; CALL quad(q, "OUTER3", A) eps=1E-06;;
 return(1/16+1/(8*pi)*(arsin(R[1,2])+arsin(R[2,3])
      +arsin(R[3,4]))+1/(4*pi*pi)*q);
END;
IF k-1>4 THEN DO; b=j(1,nrow(R),0); p=mvnorth(R,b,eps);return(p); END;
FINISH;
/************************************************
Exakte Berechnung der Levelwahrscheinlichkeiten für
einige Spezialfälle (k<6) im unbalancierten Fall
************************************************/
START plkw_het(k,gew, p_lkw,eps);
pi=3.1415926535898;
p_lkw=j(k,1,0);
IF k=2 THEN DO; p_lkw[2]=1/2;return(1);END;
IF k=3 THEN DO; p_lkw[2]=0.5;
 r=-sqrt(gew[1]*gew[3]/( (gew[1]+gew[2])* (gew[2]+gew[3])));
 p_lkw[3]=0.25+1/(2*pi)*arsin(r);
 return(1);
END;
IF k=4 THEN DO;
 p_lkw[4]=orthant(k,gew,eps); p_lkw[2]=0.5-p_lkw[4];
 gew1=gew[1]//gew[2]//(gew[3]+gew[4]);
 gew2=(gew[1]+gew[2])//gew[3]//gew[4];
 gew3=gew[1]//(gew[2]+gew[3])//gew[4];
p_lkw[3]=0.5*(orthant(3,gew1,eps)+orthant(3,gew2,eps)+orthant(3,gew3,eps));
```

```
 return(1);
END;
IF k=5 THEN DO;
 p_lkw[5]=orthant(k,gew,eps);
 gew1=gew[1]//gew[2]//gew[3]//(gew[4]+gew[5]);
 gew2=gew[1]//gew[2]//(gew[3]+gew[4])//gew[5];
 gew3=gew[1]//(gew[2]+gew[3])//gew[4]//gew[5];
 gew4=(gew[1]+gew[2])//gew[3]//gew[4]//gew[5];
 p_lkw[4]=0.5*(orthant(4,gew1,eps)+orthant(4,gew2,eps)+orthant(4,gew3,eps)
              +orthant(4,gew4,eps));
 p_lkw[2]=0.5-p_lkw[4];
 gew1=gew[1]//(gew[2]+gew[3])//(gew[4]+gew[5]);
 gew2=(gew[1]+gew[2])//(gew[3]+gew[4])//gew[5];
 gew3=(gew[1]+gew[2])//gew[3]//(gew[4]+gew[5]);
p_lkw[3]=0.25*(orthant(3,gew1,eps)+orthant(3,gew2,eps)+orthant(3,gew3,eps))
;
 gew1=gew[1]//gew[2]//(gew[3]+gew[4]+gew[5]);
 gew2=gew[1]//(gew[2]+gew[3]+gew[4])//gew[5];
 gew3=(gew[1]+gew[2]+gew[3])//gew[4]//gew[5];
 p_lkw[3]=p_lkw[3]+ orthant(3,gew1,eps)*(0.5-orthant(3,gew[3:5],eps))
                  + orthant(3,gew2,eps)*(0.5-orthant(3,gew[2:4],eps))
                  + orthant(3,gew3,eps)*(0.5-orthant(3,gew[1:3],eps));
END;
return(1);
FINISH;
/************************************************
Exakte Berechnung der Levelwahrscheinlichkeiten für den unbalancierten Fall
************************************************/
START plkwunba(k,gewicht,p_lkw,eps);
p_1_k=j(k-1,k-1,1); p_1_k[2,]=j(1,k-1,0.5); p_lkw=j(k,1,0);
DO t=3 TO k;
 DO i=2 TO t;
  mengen=0; anz=erzall(t,i, mengen); nrowmen=nrow(mengen);
  IF t<k THEN index=j(nrowmen,t,1)||j(nrowmen,k-t,0);
         ELSE index=j(nrowmen,k,1);
  IF i>1 THEN DO;
    DO s= 1 TO nrowmen;
     gew=j(1,i,0); top=0; prod=1;
     DO u=1 TO i;
      gew[u]=sum(gewicht[top+1:top+mengen[s,u]]);
      IF mengen[s,u]=2 THEN prod=prod*0.5;
                       ELSE IF mengen[s,u]>2 THEN
                       DO; prod=prod*p_1_k[mengen[s,u],top+1];END;
      top=top+mengen[s,u];
     END;
     IF i=2 THEN z=0.5; ELSE z=orthant(i,gew,eps);
     IF s=1 THEN y=z; ELSE y=y//z;
     IF t<k THEN p_1_k[t,1]=p_1_k[t,1]-z*prod;
            ELSE DO; p_lkw[i]=p_lkw[i]+z*prod; END;
    END;
```

```
   END;
  gewi=gewicht[,1:k];
  DO j=2 TO k-t+1;
   index=j(nrowmen,1,0)||index[,1:k-1];
   gewi=gewi[,2:ncol(gewi)]; mengen1=mengen||index;
   IF i>1 THEN DO;
     DO s= 1 TO nrowmen;
      gew=j(1,i,0); top=0;prod=1;
      DO u=1 TO i;
       gew[u]=sum(gewi[top+1:top+mengen[s,u]]);
       IF mengen[s,u]=2 THEN prod=prod*0.5;
        ELSE IF mengen[s,u]>2 THEN prod=prod*p_1_k[mengen[s,u],top+j-1+1];
       top=top+mengen[s,u];
      END;
      IF i=2 THEN z=0.5; ELSE z=orthant(i,gew,eps);
      IF s=1 THEN y=z; ELSE y=y//z;
      p_1_k[t,j]=p_1_k[t,j]-z*prod;
     END;
    END;
   END;
 END;
END;
FINISH;
/*******************************************************
LRT-Varianten
form=1 LRT von Bartholomew
     2 MLRT von Wright
     3 LRT-Analogon von Chacko
*******************************************************/
START barthol(lrtorg1, n, c, n_ges, eps,form);
balanc=ncol(unique(n));
IF balanc=1 THEN typ=1;
 ELSE IF k<6 THEN typ=2; ELSE typ=3;
 IF typ=1 THEN DO;
   aa=j(c+1,1,0); aa[c+1]=1; RUN stirling(c+1, aa );
   p_lkw=abs(aa/gamma(c+1)); p_lkw=p_lkw[2:c+1];
  END;
 ELSE
 DO;
  IF typ=2 THEN DO; p_lkw=0; i=plkw_het(c,n, p_lkw,eps);  END;
           ELSE DO; p_lkw=0; RUN plkwunba(c,n,p_lkw,eps); END;
 END;
p=0;
IF form=1 THEN DO;
  DO i=2 TO c; p=p+p_lkw[i]*(1-probf(lrtorg1/(i-1),(i-1),(n_ges-i)));
  END;
 END;
IF form=2 THEN DO;
  help=min(1, lrtorg1);
  DO i=2 TO c; p=p+p_lkw[i]*(1-probbeta(help,(i-1)/2, (n_ges-i)/2)); END;
```

```
 END;
IF form=3 THEN DO; Do i=2 TO c; p=p+p_lkw[i]*(1-probchi(lrtorg1,i-1)); END;
               END;
p_lkw[1]=1-sum(p_lkw[2:c]); IF lrtorg1=0 THEN p=p_lkw[1];
return(p);
FINISH;
/************************************************
Berechnung der multivariaten t-Verteilung
Funktion stammt von F. Bretz [115].
************************************************/
START nu3(ss) global(z3,R3,nu3,dim3);
 return(solow(ss#z3,R3,dim3)#ss##(nu3-1)#exp(-
nu3#ss#ss/2)#(nu3/2)##(nu3/2)#2/gamma(nu3/2));
FINISH;
dim3=1; z3=1; R3=1; nu3=1;
START multi_t( z,R,fg,dim) global (z3,R3,nu3,dim3) ;
dim3=dim; z3=z; R3=R; nu3=fg; qq=0; aa=1E-05||.p;
CALL quad(qq,"nu3",aa) eps=1e-5;
return(1-qq);
FINISH;
seed=1613725185;
/************************************************
Berechnung der multivariaten Normalverteilung
Funktion stammt von F. Bretz [115].
************************************************/
START solow(z,R,dim);
D=probbnrm(z,z,round(R,1E-09))-probnorm(z)**2;
sumi=1;
DO i=1 TO dim-1;
 b=inv(D[i+1:dim,i+1:dim]+I(dim-i)*1E-09)*D[i,i+1:dim]`;
 sumj=sum(b[1:dim-i]#(1-probnorm(z))); sumi=sumi#(probnorm(z)+sumj);
END;
return(sumi#probnorm(z));
FINISH;
/************************************************
Berechnung des Maxmin-Kontrastes
************************************************/
START maxminko(k, a);
a=j(1,k,0);
DO i=1 TO k; a[i]=sqrt((i-1)*(1-((i-1)/k)))-sqrt(i*(1-(i/k)));END;
FINISH;
/************************************************
Berechnung des Isotonen Kontrastes
************************************************/
START erzall(n,k,mengen);
feld1=j(1,k,0); first=0; mtc1=0; t=0; h=0;
DO WHILE(mtc1<2);
i=nexcom2(n, k, feld1, mtc1,t,h);
IF all(feld1) THEN IF first=0 THEN DO;mengen=feld1;first=1; END;
                              ELSE mengen=mengen//feld1;
```

```
END;
return(nrow(mengen));
FINISH;
START ergewmat(k,typ,contmat);
kon=0; RUN maxminko(k, kon); norm=sqrt(sum(kon#kon)); contmat=kon/norm;
DO i=2 TO k-1;
 mengen=0; x=erzall(k,i, mengen); col=ncol(mengen);
 DO j=1 TO x;
  top=0;
  DO s=1 TO col;
   mittel=sum(kon[top+1:top+mengen[j,s]]);
   IF s=1 THEN c1=j(1,mengen[j,s],mittel/mengen[j,s]);
          ELSE c1=c1||j(1,mengen[j,s],mittel/mengen[j,s]);
   top=top+mengen[j,s];
  END;
  norm=sqrt(sum(c1#c1)); c1=1/norm*c1; contmat=contmat//c1;
 END;
END;
FINISH;
/************************************************
Berechnung des Sugiura-Kontrastes
************************************************/
START sugiura(k,gewmat);
gewmat=j(k-1,k,0);
DO i=2 TO k;
 DO j=i TO k; gewmat[i-1,j]=1-(k-i+1)/k; END;
 DO j=1 TO i-1; gewmat[i-1,j]=-(k-i+1)/k; END;
END;
DO i=1 TO k-1; gewmat[i,]=sqrt(k/(k-i)*(i))* gewmat[i,]; END;
FINISH;
/*************************************************************************/
/*             LRT nach Bartholomew, Wright bzw. Chacko                  */
START lrt
(daten,   /*=Originaldaten    (Spaltenvektor)                              */
k,        /*=Gruppenanzahl    (Skalar)                                     */
n,        /*=Stichprobenvektor(Spaltenvektor)                              */
datform,  /*=1 Originaldaten verwenden                                     */
          /*=2 Rangtransformierte Daten verwenden                         */
verteil,  /*=1 CHI-Barverteilung (bekannte Varianz)                        */
          /*=2 CHI-Barverteilung (unbekannte Varianz)                     */
          /*=3 approximative permutative Verteilung                       */
        /*=4 Bootstrap in gepoolter Stichprobe der nichtzentrierten Daten*/
          /*=5 Bootstrap in gepoolter Stichprobe der zentrierten Daten    */
          /*=6 Bootstrap in einzelnen zentrierten Stichproben             */
varesti,  /*=1 1/(n_ges-k)                                                */
       /* sum(i=1 to k)sum(j=1to n[i]) (daten(i,j)-mean(i-te sample))##2 */
          /*=2 1/(n_ges-1)                                                */
      /*  sum(i=1 to k)sum(j=1to n[i]) (daten(i,j)-mean(alle sample))##2 */
          /*=3 kein Varianzschätzer                                      */
samplanz, /*= Anzahl der Resample-Stichproben                            */
```

```
start_w,  /*= Startwert für den Zufallsgenerator                          */
eps,      /*= Genauigkeitsschranke für das Integrationsverfahren         */
rgiman    /*= Wright-Test angewandt auf Ränge                            */
);
nges=sum(n); geord=j(k,1,0); lrt_gew=n;
IF verteil=5 | 6 THEN
 DO;
  top=0;
  DO;
   origmean=j(k,1,0);
   DO i=1 TO k;
    origmean[i]=1/n[i]*sum(daten[top+1:top+n[i]]); top=top+n[i];
    IF i=1 THEN origvek=j(n[1],1,origmean[1]);
           ELSE origvek=origvek//j(n[i],1,origmean[i]);
   END;
  END;
 END;
IF datform=2 THEN rangvek=ranktie(daten); ELSE rangvek=daten;
top=0; meanvek=j(k,1,0);
DO i=1 TO k;
 meanvek[i]=1/n[i]*sum(rangvek[top+1:top+n[i]]); top=top+n[i];
END;
gesmean=sum(rangvek)/nges;
IF varesti=1 THEN varianz=varipool(rangvek, meanvek, n, k, nges);
  ELSE IF varesti=2 THEN
                  DO;
                   varianz=variges(rangvek, gesmean, k, nges);
                   IF datform=1 & verteil=1 THEN varianz=varianz*(nges-1);
                  END;
                  ELSE varianz=1;
RUN ordnen(meanvek, lrt_gew, k, geord);
stat=t(n)*((geord-j(k,1,gesmean))##2);
IF varianz>0 THEN stat=stat/varianz;  ELSE return(1);
IF datform=1 & verteil=1 & varesti=1
 THEN DO; p=barthol(stat, n, k, nges, eps,1); return(p); END;
IF datform=1 & verteil=1 & varesti=2 THEN
  DO;   p=barthol(stat, n, k, nges, eps,2); return(p);  END;
IF datform=2 & verteil=2 & varesti=1  THEN
   DO; IF rgiman=1 THEN  p=barthol(stat, n, k, nges, eps,1);
                  ELSE  p=barthol(stat, n, k, nges, eps,3);
       return(p);
   END;
IF datform=2 & verteil=2 & varesti=2  THEN
   DO; p=barthol(stat, n, k, nges, eps,3); return(p); END;
IF datform=1 & verteil=2 & varesti=2  THEN
   DO; p=barthol(stat, n, k, nges, eps,3); return(p); END;
IF verteil=5 | verteil=6 THEN zentfeld=daten-origvek;
zaehler=1;
DO b=1 TO samplanz-1;
 IF verteil=3 THEN DO; bootst=rangvek; RUN permut(bootst, nges, start_w);
```

```
                  END;
 IF verteil=4 THEN DO;bootst=daten; RUN bootf(bootst, nges, start_w); END;
 IF verteil=5 THEN DO; bootst=zentfeld; RUN bootf(bootst,nges,start_w);END;
 IF verteil=6 THEN DO; bootst=zentfeld;RUN boots(bootst,k,n,start_w);END;
 IF datform=2 THEN rangvek=ranktie(bootst); ELSE rangvek=bootst;
 top=0;
 DO i=1 TO k;
  meanvek[i]=1/n[i]*sum(rangvek[top+1:top+n[i]]);top=top+n[i];
 END;
 gesmean=sum(rangvek)/nges;
 IF varesti=1 THEN varianz=varipool(rangvek, meanvek, n, k, nges);
              ELSE IF varesti=2
               THEN DO; varianz=variges(rangvek, gesmean, k, nges);
                     IF datform=1 THEN varianz=(nges-1)*varianz;
                    END;
              ELSE varianz=1;
 lrt_gew=n; RUN ordnen(meanvek, lrt_gew, k, geord);
 resstat=t(n)*((geord-j(k,1,gesmean))##2);
 IF varianz>0 THEN DO;
  resstat=resstat/(varianz); zaehler=zaehler+(resstat>=stat);
  END;
 ELSE zaehler=zaehler+1;
END;
return(zaehler/samplanz);
FINISH;
/*************************************************************************/
/*                       Multipler Kontrast:                             */
START Multcont(
daten,    /*=Originaldaten    (Spaltenvektor)                             */
k,        /*=Gruppenanzahl    (Skalar)                                    */
n,        /*=Stichprobenvektor(Spaltenvektor)                             */
contmat,  /*=Kontrastmatrix   (Dimension: Anzahl der Kontraste x k)       */
datform,  /*=1 Originaldaten verwenden                                    */
          /*=2 Rangtransformierte Daten verwenden                         */
verteil,  /*=1 multivariate Normalverteilung                              */
          /*=2 multivariate t-Verteilung                                  */
          /*  (Freiheitsgrad=n_ges-c   bei Originaldaten                  */
          /*                =n_ges-c-1 bei Rangdaten                      */
          /*=3 approximative permutative Verteilung                       */
       /*=4 Bootstrap in gepoolter Stichprobe der nichtzentrierten Daten */
          /*=5 Bootstrap in gepoolter Stichprobe der zentrierten Daten    */
          /*=6 Bootstrap in einzelnen zentrierten Stichproben             */
varesti,  /*=1 1/(n_ges-k)                                                */
       /*  sum(i=1 to k)sum(j=1to n[i]) (daten(i,j)-mean(i-tesample))##2 */
          /*=2 1/(n_ges-1)                                                */
      /*  sum(i=1 to k)sum(j=1to n[i]) (daten(i,j)-mean(alle sample))##2 */
          /*=3 kein Varianzschätzer                                      */
samplanz, /*= Anzahl der Resample-Stichproben                            */
start_w   /*= Startwert für den Zufallsgenerator                        */
);
```

```
anzcont=nrow(contmat); nges=sum(n);
IF verteil=5 | 6 THEN
 DO;
 top=0;
  DO;
   origmean=j(k,1,0);
   DO i=1 TO k;
    origmean[i]=1/n[i]*sum(daten[top+1:top+n[i]]); top=top+n[i];
    IF i=1 THEN origvek=j(n[1],1,origmean[1]);
           ELSE origvek=origvek//j(n[i],1,origmean[i]);
   END;
  END;
 END;
IF datform=2 THEN rangvek=ranktie(daten); ELSE rangvek=daten;
top=0; meanvek=j(k,1,0);
DO i=1 TO k; meanvek[i]=1/n[i]*sum(rangvek[top+1:top+n[i]]); top=top+n[i];
END;
norm=j(anzcont,1,0);
DO i=1 TO anzcont; norm[i]=sum(contmat[i,]#contmat[i,] /t(n)); END;
gesmean=sum(rangvek)/nges;
IF varesti=1 THEN varianz=varipool(rangvek, meanvek, n, k, nges);
 ELSE IF varesti=2 THEN varianz=variges(rangvek, gesmean, k, nges);
 ELSE IF varesti=4 THEN varianz=(nges-1)*variges(rangvek, gesmean, k,
                                nges)/nges;
   ELSE varianz=1;
stat=contmat*meanvek;
IF varianz>0 THEN stat=stat/sqrt(varianz*norm); ELSE return(1);
stat=max(stat);
IF verteil=1 THEN
 DO;
  kormat=j(anzcont,anzcont,0); RUN korrel(k, n, contmat, kormat);
  IF nrow(contmat) >1 THEN p=1-solow(stat,kormat,anzcont);
                      ELSE p=1-probnorm(stat);
    return(p);
 END;
IF verteil=2 THEN
 DO;
  kormat=j(anzcont,anzcont,0); RUN korrel(k, n, contmat, kormat);
  IF datform=1 THEN fg=nges-k; ELSE fg=nges-k-1;
  IF nrow(contmat)>1 THEN p=multi_t( stat,kormat,fg,anzcont);
                     ELSE p=1-probt(stat,fg);
   return(p);
 END;
IF verteil=5 | verteil=6 THEN zentfeld=daten-origvek;
zaehler=1;
DO b=1 TO samplanz-1;
 IF verteil=3 THEN DO;bootst=rangvek;RUN permut(bootst, nges, start_w);END;
 IF verteil=4 THEN DO;bootst=daten; RUN bootf(bootst, nges, start_w); END;
 IF verteil=5 THEN DO;bootst=zentfeld; RUN bootf(bootst,nges, start_w);END;
 IF verteil=6 THEN DO;bootst=zentfeld; RUN boots(bootst,k,n,start_w); END;
```

```
 IF datform=2 THEN rangvek=ranktie(bootst); ELSE rangvek=bootst;
 top=0;
 DO i=1 TO k;meanvek[i]=1/n[i]*sum(rangvek[top+1:top+n[i]]); top=top+n[i];
 END;
 gesmean=sum(rangvek)/nges;
 IF varesti=1 THEN varianz=varipool(rangvek, meanvek, n, k, nges);
       ELSE IF varesti=2 THEN varianz=variges(rangvek, gesmean, k, nges);
                         ELSE varianz=1;
 resstat=contmat*meanvek;
 IF varianz>0 THEN DO;
    resstat=resstat/sqrt(varianz*norm); resstat=max(resstat);
    zaehler=zaehler+(resstat>=stat);
   END;
 ELSE zaehler=zaehler+1;
END;
return(zaehler/samplanz);
FINISH;
/*************************Beispiel ***************************/
datenorg= {20, 8, 5, 6, 2, 5, 7, 5,4, 5, 7, 6, 5, 4, 3, 7, 7, 7, 6, 3};
n_ges=nrow(datenorg);n={4,4,4,4,4};c=5; eps=0.0001; mat1=1;
datenorg=(-1)*datenorg;
/**************** Aufruf der IML-Funktionen ******************/
run ergewmat(c,1,mat1);
p_MCT=Multcont(datenorg, c, n, mat1, 1, 2, 1, 10000, 1404711);
p_LRT=LRT(datenorg, c, n, 1, 2, 1, 10000, 1404711,0.001,1);
print p_MCT p_LRT;
/* Output */
/*P_MCT     P_LRT
0.0639308 0.0505579 */
QUIT;
```

Danksagung

Bei Herrn Prof. Dr. L. A. Hothorn möchte ich mich für die kritische Begleitung meiner Dissertation, für die hilfreichen fachlichen Gespräche und wissenschaftlichen Anregungen, die ständige Diskussionsbereitschaft sowie für den gewährten Spielraum während der Erstellung der Arbeit ganz besonders bedanken.

Herrn Dr. H. Bleiholder danke ich für die Überlassung des Themas, für die vielen wichtigen praktischen Hinweise, der Bereitstellung zahlreicher Beispieldaten sowie das entgegengebrachten Vertrauen.

Bedanken möchte ich mich auch bei Herrn Dr. F. Bretz und Herrn Dr. M. Weichert für die wertvollen Ratschläge während unserer gemeinsamen Doktorandenzeit.

Frau G. Rettig und Frau I. Canborgil danke ich für das Korrekturlesen der Arbeit.

Abschließend danke ich meiner Frau, Dr. A. Seidel, für die Geduld und das aufgebrachte Verständnis in den letzten Jahren sowie für das Korrekturlesen der Arbeit.

Lebenslauf

Persönliche Daten

Name	Dirk Seidel
Geburtsdatum	08. 04. 1967
Geburtsort	Rostock
Familienstand	verheiratet mit Dr. med. Anja Seidel, zwei Kinder

Schulausbildung

9/1973 - 8/1983	Polytechnische Oberschule in Rostock
9/1989 - 8/1990	Vorkurs zur Erlangung der Hochschulreife an der Technischen Hochschule Wismar, Abschluß: Fachabitur

Berufsausbildung

9/1983 - 7/1985	Lehre zum Zerspanungsfacharbeiter im Dieselmotorenwerk Rostock

Hochschulausbildung

9/1990 - 12/95	Mathematikstudium (Diplom) an der Universität Rostock Spezialisierung: Wahrscheinlichkeitsrechnung und mathematische Statistik; Nebenfachausbildung: Informatik

Wehrdienst

5/1986 - 2/1989	Unteroffizier

Berufstätigkeit:

7/1985 - 4/1986, 3/1989 - 8/1989	Zerspanungsfacharbeiter im Dieselmotorenwerk Rostock
1/1996- 1/1999	Wissenschaftlicher Assistent am Lehrgebiet Bioinformatik des Fachbereichs Gartenbau der Universität Hannover
2/99 -	Statistiker bei der Bau-Berufsgenossenschaft Hannover

Hannover, 6. Juni 2000