Das nichtparametrische Behrens-Fisher-Problem: ein studentisierter Permutationstest und robuste Konfidenzintervalle für den Shift-Effekt

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

vorgelegt von
Karin Neubert
aus Karl-Marx-Stadt

Göttingen, Juni 2006

D7
Referent: Prof. Dr. Edgar Brunner
Koreferent: Prof. Dr. Manfred Denker
Tag der mündlichen Prüfung: 7. Juli 2006

Danksagung

Mein besonderer Dank gilt Herrn Prof. Dr. Edgar Brunner, der die Idee zum Inhalt dieser Arbeit hatte und mich während der Weiterentwicklung dieser Idee mit vielen Hinweisen und Ratschlägen unterstützt und engagiert betreut hat. Herr Prof. Dr. Manfred Denker half mir die zugrunde liegenden Prinzipien mathematisch präzise zu betrachten und zu formulieren. Ohne die hervorragenden Arbeitsmöglichkeiten, insbesondere die technische Ausstattung, der Abteilung Medizinische Statistik wären beispielsweise die Simulationsstudien meiner Arbeit nicht in der Form möglich gewesen.

Ich möchte mich sehr herzlich bei Carola Werner bedanken. Unsere vielen gemeinsamen Gespräche über Themen unserer Dissertationen, die Erfahrungen bei der statistischen Beratung und all die anderen Aspekte des Promotionsstudenten-Daseins haben mir sehr geholfen.

Allen anderen Kollegen aus den Abteilungen Medizinische Statistik und Genetische Epidemiologie möchte ich danken für die herzliche Aufnahme und die freundliche und offene Atmosphäre. Ich war gern Mitglied im Promotionsstudiengang „Angewandte Statistik und Empirische Methoden“, der mir immer wieder den Blick über den eigenen Tellerrand hinaus öffnete.

Außerdem möchte ich Carola, Moritz, Leif, Stephie, Sven und Karthi danken, die mich herzlich aufgenommen haben und trotz vieler Wochenenden in Abwesenheit dafür gesorgt haben, dass Göttingen mein Zuhause war. Den Korrekturlesern Carola, Leif und Thomas danke ich für all ihre Tipps und Hinweise. Meinen Eltern, Großeltern und beiden Schwestern danke ich für ihre stetige Unterstützung und Begleitung, insbesondere in den letzten schwierigen Monaten.

Karin Neubert
Göttingen, den 8. Juni 2006

Inhaltsverzeichnis

1 Einleitung **1**

2 Grundlagen **5**
2.1 Verteilungen, Ränge, Modell 5
2.2 Relativer Effekt 7
2.3 Permutationstests 8
2.3.1 Definition des Permutationstests 9
2.3.2 Die Invarianzeigenschaft 12
2.3.3 Bedingte Monte-Carlo-Simulationen 12

3 Ein studentisierter Permutationstest für das Behrens-Fisher-Problem **15**
3.1 Modell, Hypothese und Teststatistik 16
3.2 Methoden für kleine Stichprobenumfänge 18
3.2.1 t-Approximation 19
3.2.2 Likelihood-Ratio-Test 19
3.2.3 Bootstraptest 21
3.2.4 Eigenschaften der existierenden Methoden 21
3.2.5 Studentisierter Permutationstest 21
3.3 Studentisiertes Permutieren 23
3.4 Nachweis der Bedingungen für den Grenzwertsatz 25
3.5 Simulationsstudie 32
3.5.1 Zwei normalverteilte Stichproben 35
3.5.2 Zwei bimodal-verteilte Stichproben 38
3.5.3 Eine normalverteilte vs. eine χ^2_3-verteilte Stichprobe 39
3.6 Anwendungen 40
3.6.1 Ferritin-Studie 40
3.6.2 Schulter-Schmerz-Studie 42
3.7 Zusammenfassung 43

4 Robuste Konfidenzintervalle für den Shift-Effekt **45**
4.1 Modell 47
4.2 Konfidenzintervalle nach Hodges-Lehmann 48

4.3 Konfidenzintervalle nach Bauer 51
4.4 Konfidenzintervalle nach Bauer für heteroskedastische Gruppen . . . 52
4.5 Datenbeispiel zur Berechnung der Konfidenzintervalle 53
4.6 Anwendung im 2×2-Split-Plot-Design 55
4.7 Simulationsstudie 58
4.7.1 Normalverteilte Stichproben 61
4.7.2 Bimodal-verteilte Stichproben 64
4.7.3 Log-normalverteilte Stichproben 65
4.8 Anwendungen 66
4.8.1 Rückhärtung des Dentins 66
4.8.2 Post-Operatives Ödem 67
4.9 Zusammenfassung 69

A Anhang **71**
A.1 Asymptotischer Äquivalenzsatz 71
A.2 Asymptotische Normalität der Permutationsverteilung 71
A.3 Varianzformel von Hájek 72
A.3.1 Varianzformel von Hájek im Beweis Satz 3.2 73

B SAS-Makro für den Permutationstest **77**

C SAS-Makro für robuste Konfidenzintervalle für den Shift-Effekt **81**

Literaturverzeichnis **87**

Abbildungsverzeichnis

2.1 Dichte einer symmetrischen Verteilung 8

3.1 Dichten bimodaler Verteilungen 33
3.2 Niveausimulationen schematisch 35
3.3 Normalverteilungen 1 36
3.4 Normalverteilungen 2 37
3.5 Normalverteilungen, Güte bei verschiedenen Varianzen 38
3.6 Bimodale Verteilungen 39
3.7 Normalverteilung vs. χ^2_3-Verteilung 40
3.8 Ferritin-Daten 41

4.1 Coverage-Simulationen schematisch 60
4.2 Normalverteilung 62
4.3 Bimodale Verteilungen 64
4.4 Lognormale Verteilungen 65
4.5 Rückhärtung des Dentins 66
4.6 Temperaturdifferenzen zu Baseline 68
4.7 paarweise Temperaturdifferenzen 69

B.1 Makro Output PERM_BF 79

C.1 Makro Output PERM_KI 85

Tabellenverzeichnis

3.1 Ränge und interne Ränge im Fall kleinster Varianz 18
3.2 Daten der Ferritin-Studie. 41
3.3 p-Werte für die Ferritin-Studie. 41
3.4 Daten der Schulter-Schmerz-Studie. 42
3.5 Adjustierte p-Werte für die Schulter-Schmerz-Studie 43

4.1 Datenbeispiel: paarweise Differenzen und Werte von T_N 54
4.2 Datenbeispiel: Quantile, Indizes und Grenzen der Konfidenzintervalle 55
4.3 Split-Plot-Design 55
4.4 Eigenschaften der Konfidenzintervalle bei Normalverteilung 63
4.5 Dentinrückhärtung: Mittelwerte und Varianzen 67
4.6 Dentinrückhärtung: Konfidenzintervalle 67
4.7 Temperaturdifferenzen: Mittelwerte und Varianzen 68
4.8 Temperaturdifferenzen: Konfidenzintervalle 69

C.1 Split-Plot-Design 82

1 Einleitung

Ein wesentliches Merkmal eines wissenschaftlichen Experiments ist dessen wiederholte Durchführung, die ausschließen soll, dass die gewonnenen Ergebnisse nur zufallsbedingt sind. Allerdings sind in vielen Anwendungsgebieten, wie beispielsweise der Medizin oder Biologie, oft nur wenige Versuchseinheiten vorhanden an denen unabhängige Messwiederholungen vorgenommen werden können. Eine Ursache für diese sehr beschränkte Anzahl an Versuchseinheiten ist die eingeschränkte Verfügbarkeit, z.B. wenn es bei seltenen Krankheiten nur wenige erkrankte Patienten gibt. Oft gibt es außerdem ethische Einwände, zusätzliche Patienten oder Probanden den für den Versuch notwendigen Behandlungsmaßnahmen auszusetzen. Um in einer solchen Situation verlässliche Aussagen über das Vorliegen einer Wirkung der betrachteten Behandlung machen zu können, müssen statistische Methoden speziell für diese Situation geeignet sein.

Die beiden klassischen Herangehensweisen, die *parametrische* und die *nichtparametrische* Statistik, stoßen hier an ihre Grenzen. Parametrische Testverfahren beruhen auf der Annahme, dass den gewonnen Daten eine bestimmte Wahrscheinlichkeitsverteilung zugrunde liegt, die eindeutig durch einen endlich-dimensionalen Parameter bestimmt ist. Die Entscheidung für eine bestimmte parametrische Modellklasse ist eine grundlegende Schwierigkeit dieser Methoden. So verlieren Parameter ihre Bedeutung und schließende statistische Verfahren ihre Gültigkeit, wenn eine ungeeignete Modellklasse ausgewählt wird. Insbesondere bei wenigen unabhängigen Versuchswiederholungen ist es oft schwierig, die Wahl einer bestimmten Verteilungsklasse zu rechtfertigen. Nichtparametrische Verfahren benötigen keine solche Verteilungsannahme oder lassen unendlich-dimensionale Parameter zu. Bei vielen nichtparametrischen Methoden ist die Anwendung dann aber mit der korrekten Wahl bestimmter Kenngrößen verbunden, wie beispielsweise bei der Verwendung von Kernschätzern, oder sie hängen von der Gültigkeit asymptotischer Verteilungsaussagen ab, wie viele Rangmethoden. Beides ist bei geringen Stichprobenumfängen problematisch. Einen Lösungsansatz stellen hier Permutationsmethoden dar. Dabei wird die Verteilung, die für die Anwendung von Testverfahren oder die Berechnung von Konfidenzintervallen gebraucht wird, aus den gewonnenen Daten bestimmt. Die Verallgemeinerbarkeit der durch solche Verfahren gewonnenen Ergebnisse über die betrachtete Stichprobe hinaus, hängt allerdings vom Grad der Repräsentativität der Stichprobe für die Allgemeinheit ab. In diesem Sinne sind Permutationstests beding-

te Tests, bedingt auf die beobachteten Daten. Die Repräsentativität der Daten kann durch geeignete Verfahren wie beispielsweise durch Randomisierungsverfahren in klinischen Studien erhöht werden. Permutationsmethoden bieten sich insbesondere an, wenn nur wenige unabhängige Messwiederholungen zur Verfügung stehen, da man dann gegebenenfalls in der Lage ist, alle möglichen Permutationen zu bestimmen. Dadurch erhält man vollständige Informationen über die Verteilung und die durchgeführten Tests halten das Niveau exakt ein. Allerdings sind Permutationsverfahren nur exakt, wenn die Zufallsvariablen unter der Hypothese austauschbar sind.

Das Ziel vieler biometrischer Untersuchungen ist der Vergleich von zwei Stichproben, beispielsweise wenn die Wirkung eines Medikaments mit Placebo oder eine neue Behandlungsmethode mit einer etablierte Methode verglichen werden sollen. Aber auch bei komplexeren Designs der Experimente werden am Ende der Durchführung einer hierarchisch gegliederten statistischen Analyse, z.B. einer mehrfaktoriellen Varianzanalyse, oft Zwei-Stichproben-Vergleiche durchgeführt. Wir bezeichnen die n_1 Beobachtungen der einen Gruppe mit $X_{11}, \ldots, X_{1n_1}$ und die n_2 Beobachtungen der zweiten Gruppe mit $X_{21}, \ldots, X_{2n_2}$. Die Zufallsvariablen X_{ik} seine alle unabhängig und innerhalb einer Gruppe identisch mit Verteilungsfunktion F_i, $i = 1{,}2$ verteilt.

Eine spezielle Rolle unter den Zwei-Stichproben-Problemen nimmt das *Behrens-Fisher-Problem* ein. Dabei soll die Lage der zwei Stichproben verglichen werden, wenn beide Gruppen eventuell unterschiedliche Streuungen haben. Beim klassischen parametrischen Ansatz betrachten wir die Hypothese gleicher Erwartungswerte μ_i bei möglicherweise verschiedenen Varianzen, das heißt $H_0 : \ \mu_1 = \mu_2$. Für normalverteilte Daten wurden verschiedene Tests entwickelt und in der Literatur diskutiert (z.B. Smith, 1936; Welch, 1937; Satterthwaite, 1946; Cochran, 1964; Moser & Stevens, 1992).

Kann die Normalverteilung der Daten nicht voraussetzt werden, verwendet man den Wilcoxon-Mann-Whitney-Test (WMW-Test), um zwei unabhängige Stichproben zu vergleichen (Wilcoxon, 1945; Mann & Whitney, 1947). Dabei wird die Hypothese $F_1 = F_2$ getestet. Sind die Streuungen der beiden Verteilungen allerdings verschieden, so hält der WMW-Test das Niveau nicht mehr ein (Pratt, 1964). Für semiparametrische Modelle wurden Modifikationen des WMW-Tests vorgeschlagen, die heteroskedastische Verteilungen zulassen (Fligner & Policello, 1981) bzw. auch für nicht-symmetrische Verteilungen geeignet sind (Babu & Padmanabhan, 2002). Fligner & Policello (1981) bemerken in ihrem Paper außerdem, dass ihr Verfahren auch zum Testen der Hypothese $\int F_1 dF_2 = \frac{1}{2}$ geeignet ist. Die dieser Hypothese zugrunde liegende Größe $p = \int F_1 dF_2$ wurde von Mann & Whitney (1947) als *relativer Effekt* eingeführt. Der relative Effekt kann als Wahrscheinlichkeit interpretiert werden, mit der die Beobachtungen der einen Stichprobe tendenziell größere (kleinere) Werte annehmen als die Beobachtungen der anderen Stichprobe. Für symmetrische Verteilungen ist der relative Effekt invariant unter reinen Skalenalternativen, so dass dann die Hypothese des parametrischen Behrens-Fisher-Problems $H_0 : \ \mu_1 = \mu_2$ äquivalent ist zu $H_0 : \ p = \frac{1}{2}$. Ein Testverfahren für die Hypothese $H_0 : \ p = \frac{1}{2}$ ist der Rangtest von

Brunner & Munzel (2000) und die dazugehörige t-Approximation für kleine Stichproben. Dieser Rangtest ist auf viele Modelle anwendbar, da beliebige Verteilungen der Daten zugelassen werden (nur die trivialen Ein-Punkt-Verteilungen sind ausgeschlossen). Beispielsweise kann dieses Verfahren auch für die Analyse von ordinalen Daten oder Scores verwendet werden. Weitere Testverfahren, die für die Hypothese $H_0: \ p = \frac{1}{2}$ vorgestellt wurden, sind ein Likelihood-Ratio-Test von Troendle (2002) sowie Bootstrap-Prozeduren (Chen & Kianifard, 2000; Reiczigel *et al.* , 2005).

Wir schlagen für diese Hypothese einen neuen studentisierten Permutationstest vor, das heißt einen Permutationstest der auf einer Teststatistik beruht, die durch einen geeigneten Varianzschätzer dividiert wird. Per Definition sind Permutationstests in einer Behrens-Fisher-Situation zunächst einmal nicht gültig, da die Beobachtungen unter der Hypothese nicht austauschbar sind. Verwendet man nun eine geeignet studentisierte Teststatistik, so erhält man durch das Dividieren mit dem Varianzschätzer asymptotisch die richtige Varianz und kann unter gewissen Bedingungen zeigen, dass die Permutationsverteilung der Teststatistik gegen eine Normalverteilung konvergiert. Diese Aussage wird in Janssen (1997) allgemein für lineare Teststatistiken bewiesen und speziell für die Welch-Statistik im parametrischen Behrens-Fisher-Problem nachgewiesen. Weitere asymptotisch gültige Permutationstests werden in Pesarin (2001) beschrieben.

Häufig werden Permutationstests für Teststatistiken durchgeführt, die direkt von den Daten X_{ik} abhängen (vgl. Janssen, 1997). Durch die Verwendung einer Teststatistik, die ausschließlich auf den Rängen der Daten definiert ist, erhält man eine zusätzliche Robustheit des Verfahrens gegenüber Ausreißern. Dies ist besonders bei kleinen Stichprobenumfängen wichtig. Wie in Neubert & Brunner (2006) beschrieben schlagen wir hier vor, die lineare Rangstatistik von Brunner & Munzel (2000) für den Permutationstest zu verwenden. Die durchgeführte Simulationsstudie (Abschnitt 3.5) bestätigt gute Eigenschaften des Testverfahrens bei Anwendung auf kleine Stichprobenumfänge. Mithilfe des Zentralen Grenzwertsatzes für studentisierte Permutationstests von Janssen (1997) können wir außerdem die asymptotische Normalität dieser Teststatistik nachweisen (vgl. Abschnitt 3.3).

Neben einem Test auf Lageunterschiede zweier heteroskedastischer Stichproben stellen wir ein Konfidenzintervall für einen zwischen diesen Gruppen auftretenden *Verschiebungseffekt (Shift-Effekt)* vor. Um einen Shift-Effekt untersuchen zu können, wird in der Literatur häufig ein (reines) Lokationsmodell für die Verteilungsfunktionen F_i vorausgesetzt. Das heißt wir fordern, dass es eine Verteilungsfunktion F gibt, so dass

$$F_1(x) = F(x - \mu_1), \quad \text{und} \quad F_2(x) = F(x - \mu_2), \quad x \in \mathbb{R},$$

wobei μ_i der Erwartungswert bezüglich F_i sei. Der Shift-Effekt ist dann als $\theta = \mu_2 - \mu_1$ definiert. Außerdem wird vorausgesetzt, dass die Verteilungsfunktionen F_i stetig sind. Im reinen Lokationsmodell ist die parametrische Hypothese $H_0: \ \theta = 0$

äquivalent zur nichtparametrischen Hypothese $H_0 : \ F_1 = F_2$. Unter diesen Annahmen haben Lehmann (1963) und Bauer (1972) Konfidenzintervalle für den Shift-Effekt vorgestellt. Allerdings ist die Methode zur Konstruktion des Konfidenzintervalls von Bauer (1972) allgemein für lineare Rangstatistiken formuliert. Dies werden wir nutzen, um auch für heteroskedastische Verteilungen der Gruppen ein Konfidenzintervall für den Shift-Effekt herzuleiten. Statt des reinen Lokationsmodells lassen wir dann ein Lokations-Skalen-Modell zu, das heißt

$$F_1(x) = F\left(\frac{x-\mu_1}{\sigma_1}\right), \quad \text{und} \quad F_2(x) = F\left(\frac{x-\mu_2}{\sigma_2}\right), \quad x \in \mathbb{R},$$

wenn σ_i^2 die Varianz von F_i ist. Dieses Modell entspricht dann wiederum einer Behrens-Fisher-Situation. Entsprechend betrachten wir statt der nichtparametrischen Hypothese $H_0 : \ F_1 = F_2$ des reinen Lokationsmodells, die Hypothese des nichtparametrischen Behrens-Fisher-Problems $H_0 : \ p = \frac{1}{2}$. Um die Äquivalenz der Hypothese $H_0 : \ p = \frac{1}{2}$ mit der parametrischen Hypothese $H_0 : \ \theta = 0$ zu gewährleisten, müssen wir nun fordern, dass die Verteilungsfunktionen F_i stetig und symmetrisch sowie an der Stelle des Erwartungswertes μ_i invertierbar sind.

Da wir nun die Hypothese $H_0 : \ p = \frac{1}{2}$ betrachten, verwenden wir wieder die von Brunner & Munzel (2000) vorgestellte lineare Rangstatistik und konstruieren damit Konfidenzintervalle nach der Methode von Bauer (vgl. Abschnitt 4.4). Für die Berechnung dieser Konfidenzintervalle benötigen wir die Quantile der Verteilung der Teststatistik. Wir schlagen unter anderem vor, die Quantile der Permutationsverteilung der Statistik zu verwenden. Diese Konfidenzintervalle sowie die von Lehmann (1963) vorgestellten Intervalle vergleichen wir in einer Simulationsstudie (siehe Abschnitt 4.7). Die Simulationen zeigen, dass das Konfidenzintervall nach Bauer mit Permutationsverteilungsquantilen die besten Eigenschaften aufweist.

Die Arbeit ist wie folgt aufgebaut: Zunächst werden wir einige grundlegende Begriffe zu Rangstatistiken und Permutationstests definieren und erklären (Kapitel 2). Danach wird in Kapitel 3 der studentisierte Permutationstest vorgestellt und gezeigt, dass seine asymptotische Permutationsverteilung eine Normalverteilung ist. Das Verhalten des Permutationstests bei kleinen Stichprobenumfängen wird anhand einer Simulationsstudie mit anderen Verfahren verglichen. Alle Verfahren werden dann auf Beispieldatensätze aus der Medizin angewendet. In Kapitel 4 wird zunächst die Konstruktion von Konfidenzintervallen für den Shift-Effekt nach Hodges-Lehmann und nach Bauer vorgestellt. Die Methode nach Bauer wird auf die Anwendung bei vorliegender Heteroskedastizität erweitert und es werden drei verschiedene Intervalle für diese Situation vorgestellt. Alle vorgestellten Intervalle werden mithilfe einer Simulationsstudie verglichen und auf Daten medizinischer Studien angewendet. Die verwendeten Sätze sind im Anhang A zitiert. Im Rahmen dieser Arbeit wurden außerdem SAS-IML-Makros erstellt. Das Makro zur Berechnung des Permutationstests wird im Anhang B beschrieben und das Makro zur Berechnung der Konfidenzintervalle für den Shift-Effekt im Anhang C.

2 Grundlagen

In diesem Kapitel werden zunächst einige im Folgenden häufig verwendet Begriffe eingeführt und deren Notation festgelegt. Außerdem werden Permutationstests definiert und deren praktische Durchführung mit bedingten Monte-Carlo-Simulationen beschrieben.

2.1 Verteilungen, Ränge, Modell

Wir betrachten einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}, \mathrm{P})$ und die Zufallsvariable X,

$$X : \Omega \mapsto \mathcal{X},$$

wobei wir im Folgenden stets $\mathcal{X} = \mathbb{R}$ betrachten. $\mathcal{B}$ sei die Borelsche σ-Algebra auf Ω und X sei $\mathcal{B}$-messbar. Weiterhin sei P eine Wahrscheinlichkeitsverteilung auf $\mathcal{B}$. Ist $\mathcal{P}$ eine auf $\mathcal{X}$ definierte Familie von Wahrscheinlichkeitsmaßen, so können wir das betrachtete Experiment als $(X, \mathcal{X}, \mathcal{B}, \mathrm{P} \in \mathcal{P})$ zusammenfassen. Die *(Wahrscheinlichkeits-)Verteilung von* X auf $\mathcal{X}$ ist dann P_X, wobei für $A \subset \Omega$:

$$\mathrm{P}_X(A) := \mathrm{P}(X \in A).$$

Die *Verteilungsfunktion der Zufallsvariablen* X werden wir mit F bezeichnen. Wir verwenden dabei stets die *normalisierte Verteilungsfunktion*, die an der Stelle $x \in \mathbb{R}$ als

$$F(x) := \frac{1}{2}(F^+(x) + F^-(x))$$

definiert ist (Ruymgaart, 1980). Dabei ist $F^+(x) = P(X \leq x) = P_X(]-\infty, x])$ die klassische rechtsseitig stetige Verteilungsfunktion und $F^-(x) = P(X < x)$ die linksstetige Version. Diese Definition gewährleistet, dass auch bei nicht-stetigen Verteilungen und daraus resultierendem Auftreten von identischen Beobachtungen (Bindungen) jeder Beobachtung eindeutig ein Wert zugeordnet werden kann. Dabei machen wir keine weiteren Annahmen über die Form von F, einzig Ein-Punkt-Verteilungen schließen wir aus.

Sei $\mathbf{X} = (X_1, \ldots, X_N)'$ ein Vektor von N unabhängigen Zufallsvariablen, die identisch nach F verteilt sind. Wir schreiben dafür abkürzend $X_k \overset{\text{u.i.v.}}{\sim} F$, $k =$

$1, \ldots, N$, wobei u.i.v. für „unabhängig identisch verteilt“ steht. Im Folgenden werden wir Vektoren und Matrizen immer durch Fettdruck kennzeichnen. Als Schätzer für die Verteilungsfunktion F verwenden wir die *empirische Verteilungsfunktion*

$$\widehat{F}(x) := \frac{1}{N} \sum_{k=1}^{N} c(x - X_k),$$

die mithilfe der normalisierten Version der Zählfunktion c definiert wird, wobei

$$c(x) := \begin{cases} 1 & x > 0 \\ \frac{1}{2} & x = 0 \\ 0 & x < 0. \end{cases}$$

Diese Definitionen gewährleisten, dass $\widehat{F}$ ein konsistenter und erwartungstreuer Schätzer für F ist (Brunner & Munzel, 2002, S. 32). Außerdem können wir nun den *Mittelrang* von X_k bezüglich aller Zufallsvariablen $X_1, \ldots, X_N$ definieren:

$$R_k := N\widehat{F}(X_k) + \frac{1}{2} = \sum_{l=1}^{N} c(X_k - X_l) + \frac{1}{2}.$$

Die Verwendung der Mittelränge ergibt sich aus der Verwendung der normalisierten Verteilungs- und Zählfunktion. Wenn Bindungen auftreten kann so der Rang jeder Beobachtung eindeutig bestimmt werden.

Im Folgenden werden wir ein Zwei-Stichproben-Problem betrachten. Wir wollen also die Beobachtungen zweier Gruppen miteinander vergleichen. Seien

$$\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})'$$

die n_1 Beobachtungen der ersten Gruppe und

$$\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})'$$

die n_2 Beobachtungen der zweiten Gruppe. Wir schreiben dann für den Vektor aller Daten $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ und es sei $N = n_1 + n_2$. Es sei P_1 die Verteilung der Zufallsvariablen in der ersten Gruppe und P_2 die Verteilung in der zweiten Gruppe. Entsprechend wird die Verteilungsfunktion in der ersten Gruppe mit F_1 und die Verteilungsfunktion in der zweiten Gruppe mit F_2 bezeichnet. Die Beobachtungen seien unabhängig und innerhalb einer Gruppe identisch verteilt mit Verteilungfunktion F_i:

$$X_{ik} \overset{\text{u.i.v.}}{\sim} F_i.$$

Der Mittelrang R_{ik} einer Beobachtung X_{ik} (im Folgenden meist nur mit „Rang“ bezeichnet) ist der Rang bezüglich aller N Beobachtungen beider Gruppen. Für

die Berechnung des *internen Rangs* von X_{ik} werden nur die n_i Beobachtungen der Gruppe i verwendet und er ist definiert als:

$$R_{ik}^{(i)} := \sum_{l=1}^{n_i} c(X_{ik} - X_{il}) + \frac{1}{2}.$$

Mithilfe der internen Ränge und der Mittelränge können wir nun die *normierten Platzierungen* von X_{ik} bezüglich der n_i Zufallsvariablen $X_{i1}, \ldots, X_{in_i}$ der eigenen Gruppe ($\widehat{F}_i(X_{ik})$) und bezüglich der n_j Zufallsvariablen $X_{j1}, \ldots, X_{jn_j}$, $j \neq i$, $i,j \in \{1,2\}$ der anderen Gruppe ($\widehat{F}_j(X_{ik})$) bestimmen:

$$\begin{aligned} \widehat{F}_i(X_{ik}) &:= \frac{1}{n_i}\left(R_{ik}^{(i)} - \frac{1}{2}\right) \\ \widehat{F}_j(X_{ik}) &:= \frac{1}{n_j}\left(R_{ik} - R_{ik}^{(i)}\right) \quad i \neq j. \end{aligned} \tag{2.1}$$

2.2 Relativer Effekt

Wir werden die Verteilungsfunktionen F_i $(i = 1, 2)$ nun benutzen, um einen Unterschied zwischen den zwei betrachteten Gruppen zu beschreiben. Wir verwenden dazu den *relativen Effekt* p, für den wir die folgende, auch für nicht-stetige Verteilungen geeignete, Definition verwenden (Mann & Whitney, 1947):

$$p := \mathrm{P}(X_{11} < X_{21}) + \frac{1}{2}P(X_{11} = X_{21}) = \int F_1 dF_2.$$

Ist $p < \frac{1}{2}$ $(p > \frac{1}{2})$ so sagt man, dass X_{11} zu größeren (kleineren) Werten tendiert als X_{21}. Gilt $p = \frac{1}{2}$, dann spricht man davon, dass „X_{11} und X_{21} (stochastisch) tendenziell gleich sind“. Dies motiviert die Formulierung der Hypothese, dass es keine Gruppenunterschiede gibt, als $H_0 : \ p = \frac{1}{2}$. Sind X_{11} und X_{21} identisch verteilt, so folgt, dass $p = \frac{1}{2}$ ist (Brunner & Munzel, 2002, S. 19) und die Hypothese H_0^F : $F_1 = F_2$ somit ein Spezialfall von $H_0 : \ p = \frac{1}{2}$ ist.

Für stetige und symmetrische Verteilungen enthält $H_0 : \ p = \frac{1}{2}$ außerdem die verallgemeinerte parametrische Behrens-Fisher-Hypothese: $H_0 : \ \mu_1 = \mu_2$, wenn μ_i der Erwartungswert in Gruppe i ist. Außer der Stetigkeit und Symmetrie müssen wir fordern, dass die Verteilungsfunktion F_i an der Stelle μ_i invertierbar ist bzw. ihre Dichte f_i dort strikt größer als Null ist. Unter $H_0 : \ p = \frac{1}{2}$ gilt zunächst für stetige Verteilungen $p = \mathrm{P}(X_{11} < X_{21}) = \mathrm{P}(X_{11} > X_{21}) = \frac{1}{2}$. Aufgrund der Symmetrie gilt für X_{11} (vgl. Abbildung 2.1):

$$\begin{aligned} \mathrm{P}(X_{11} < X_{21}) &= 1 - \mathrm{P}(X_{11} < X_{21} - 2(X_{21} - \mu_1)) \\ &= 1 - \mathrm{P}(X_{11} < 2\mu_1 - X_{21}) \\ &= 1 - \mathrm{P}(X_{11} + X_{21} < 2\mu_1) \end{aligned}$$

und analog für X_{21}

$$\mathrm{P}(X_{21} < X_{11}) = 1 - \mathrm{P}(X_{11} + X_{21} < 2\mu_2).$$

Daraus folgt

$$\frac{1}{2} = \mathrm{P}\left(\mu_1 > \frac{X_{11} + X_{21}}{2}\right) = \mathrm{P}\left(\mu_2 > \frac{X_{11} + X_{21}}{2}\right).$$

Durch die Forderung der Invertierbarkeit der Verteilungsfunktionen F_i an der Stelle μ_i gilt damit für stetige Verteilungen $\mu_1 = \mu_2$.

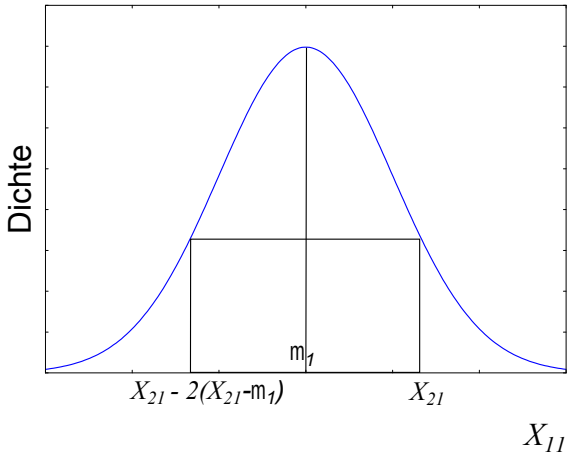


Abbildung 2.1: Dichte der symmetrischen Verteilung von X_{11}

Einen konsistenten und erwartungstreuen Schätzer für p erhält man, wenn die Verteilungsfunktionen F_1 und F_2 durch die empirischen Pendants ersetzt werden (Brunner & Munzel, 2002):

$$\widehat{p} := \int \widehat{F}_1 d\widehat{F}_2 = \frac{1}{n_1}\left(\overline{R}_{2\cdot} - \frac{n_2 + 1}{2}\right) = \frac{1}{N}(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}) + \frac{1}{2}, \tag{2.2}$$

$$\text{wobei} \quad \overline{R}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik}.$$

2.3 Permutationstests

In diesem Abschnitt werden wir zunächst einige Begriffe aus der Theorie der Permutationstests vorstellen, gefolgt von der formalen Definition der Permutationsver-

teilung und eines Permutationstests. Dann wird eine Invarianzeigenschaft der Permutationsverteilung und der Teststatistik definiert. Abschließend stellen wir einen Algorithmus zur Berechnung von Monte-Carlo-Simulationen der Permutationsverteilung vor und zeigen die Konvergenz der so gewonnenen Approximation der Permutationsverteilung gegen die tatsächliche Permutationsverteilung.

2.3.1 Definition des Permutationstests

Permutationstests sind zunächst einmal nichtparametrische Tests, da sie keine Annahme über die den Daten zugrunde liegenden Verteilungen benötigen. Sie gehören wie die Bootstraptests zu den *resampling Verfahren*, bei denen die zur Durchführung von Tests oder zur Berechnung von Konfidenzintervallen benötigte Verteilung der betrachteten Teststatistik aus der gezogenen Stichprobe ermittelt wird. Permutationstests sind entsprechend auf die beobachteten Daten bedingte Tests. Dabei wird aus den Daten mehrmals eine neue Stichprobe gezogen. Im Gegensatz zu Bootstraptests geschieht das Ziehen bei Permutationstests allerdings ohne Zurücklegen. Im Zwei-Stichproben-Fall wird also jeder Beobachtung nur die Gruppe, zu der sie gehört, neu zugeordnet. Wenn die Beobachtungen mit den höchsten Werten alle einer Gruppe zugeordnet werden, können sich dadurch vorhandene Gruppenunterschiede verstärken. Andererseits können die Gruppenunterschiede auch abgeschwächt werden, wenn die Beobachtungen mit hohen Werten gleichmäßig auf beide Gruppen verteilt werden. Anhand des Anteils von Permutationen, die zu noch größeren Gruppenunterschieden führen als zu dem tatsächlich beobachteten Gruppenunterschied, lässt sich dann ablesen wie „extrem“ dieser im Verhältnis ist. Auf diesem Prinzip basiert die Testentscheidung eines Permutationstests.

Zunächst definieren wir allgemein für eine Gruppe von Transformationen einige Begriffe, die zur formalen Definition eines Permutationstests notwendig sind. Sei $\mathrm{P}^{(N)}$ die gemeinsame Verteilung des Vektors $\mathbf{X}$. Dann betrachten wir zunächst das allgemeine Testproblem:

$$H_0^P:\ \mathrm{P}^{(N)} \in \mathcal{P}_0 \quad \text{gegen} \quad H_1^P:\ \mathrm{P}^{(N)} \in \mathcal{P}_1, \quad \text{wobei} \quad \mathcal{P}_1 = \mathcal{P} \backslash \mathcal{P}_0.$$

Sei *G eine Gruppe von Transformationen*, die auf $\mathcal{X}^N = \mathbb{R}^N$ operiert. G induziert eine Äquivalenzrelation auf $\mathcal{X}^N$ mit

$$\mathbf{x}^* \sim \mathbf{x} \quad \Leftrightarrow \quad \exists\, g \in G:\ \mathbf{x}^* = g(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathcal{X}^N.$$

Damit erhalten wir eine Partition $\mathcal{X}^N/_{\sim}$ auf $\mathcal{X}^N$ und zerlegen $\mathcal{X}^N$ somit in Äquivalenzklassen $O(\mathbf{x})$, die wir als *Orbits* bezeichnen:

$$O(\mathbf{x}) := \{\mathbf{x}^*,\, \mathbf{x}^* = g(\mathbf{x}),\, g \in G\}, \quad \#(O(\mathbf{x})) = M(\mathbf{x}).$$

Konkret werden wir hier die Permutationsgruppe von N Elementen, das heißt

$$G = \mathcal{S}_N = \{\pi, \pi : \{1,\ldots,N\} \to \{1,\ldots,N\}\}, \tag{2.3}$$
$$\text{wobei} \quad \pi(ik) = \begin{cases} (1\pi(k)) & \pi(k) \leq n_1, \\ (2\pi(k)) & \pi(k) > n_1. \end{cases}$$

betrachten, so dass $O(\mathbf{x})$ alle Permutationen des Vektors $\mathbf{x}$ enthält und damit $M(\mathbf{x}) \leq N!$. Anhand des Orbits lässt sich noch einmal verdeutlichen, dass Permutationstests bedingte Tests sind. Der Zufallsmechanismus teilt sich in 2 Schritte:

- die Auswahl des Orbits $O(\mathbf{x})$ und
- die Auswahl eines Elements des Orbits.

Die *bezüglich des Orbits* $O(\mathbf{x})$ *bedingte Verteilung* definieren wir als

$$\mathrm{P}^{(N)}|_{O(\mathbf{x})}(A) := \mathrm{P}^{(N)}(\mathbf{x}^* \in A | \mathbf{x}^* \in O(\mathbf{x})) \quad \forall A \subseteq \mathcal{X}^N \quad \text{messbar.}$$

Wir bezeichnen $\mathrm{P}^* = \mathrm{P}^{(N)}|_{O(\mathbf{x})}$ auch als *Permutationsverteilung*.

Wir wollen nun formal den zugehörigen zweiseitigen Permutationstest bezüglich des Orbits $O(\mathbf{x})$ definieren. Wir betrachten das Testproblem

$$H_0^\theta : \ \theta = 0 \quad H_1^\theta : \ \theta \neq 0$$

mit einem Parameter $\theta \in \Theta$, der eine Eigenschaft der Wahrscheinlichkeitsverteilung der Daten beschreibt. Sei

$$T : \mathcal{X}^N \to \mathbb{R}$$

eine geeignete reellwertige Teststatistik für dieses Testproblem und $\mathcal{T}(\mathbf{x})$ der *Permutationssupport der Statistik* T unter $O(\mathbf{x})$, das heißt die Menge aller Werte, die T unter $O(\mathbf{x})$ annimmt:

$$\mathcal{T}(\mathbf{x}) := \{T^*, T^* = T(\mathbf{x}^*), \mathbf{x}^* \in O(\mathbf{x})\}.$$

Dann können die Elemente des Permutationssupports $\mathcal{T}(\mathbf{x})$ aufsteigend sortiert werden

$$T^*_{(1)} \leq \cdots \leq T^*_{(M(\mathbf{x}))},$$

so dass für ein Niveau $\alpha \in (0,1)$ die kritischen Werte T_1^α, T_2^α der Teststatistik für den Permutationstest wie folgt definiert sind:

$$T_1^\alpha := T^*_{(M_1)}, \quad \text{wobei} \quad M_1 = \lceil (\alpha/2) \cdot M(\mathbf{x}) \rceil,$$
$$T_2^\alpha := T^*_{(M_2)}, \quad \text{wobei} \quad M_2 = \lfloor (1 - \alpha/2) \cdot M(\mathbf{x}) \rfloor.$$

Dabei steht $\lfloor x \rfloor$ für die größte ganze Zahl, die kleiner oder gleich x ist und $\lceil x \rceil$ für die kleinste ganze Zahl, die größer oder gleich x ist. Mithilfe der Permutationsverteilung P^* können wir die Permutationsverteilung der Teststatistik T definieren:

$$F_T^*(t|O(\mathbf{x})) := \mathrm{P}^*(T \leq t) = \mathrm{P}(T^* \leq t|O(\mathbf{x})) = \frac{\#\{T^* \leq t\}}{M(\mathbf{x})}, \quad t \in \mathbb{R}.$$

$F_T^*(t|O(\mathbf{x}))$ gibt also den Anteil der Permutationen an, für die der Wert der Teststatistik kleiner oder gleich t ist. F_T^* ist eine Treppenfunktion mit Sprüngen an den Stellen des *Permutationssupports der Verteilung* F_T^*:

$$\mathcal{S}_F := \left\{ \frac{h}{M(\mathbf{x})}, \;\; h = 1, \ldots, M(\mathbf{x}) \right\}.$$

$\mathcal{S}_F$ entspricht also der Menge der möglichen Werte, die die Permutationsverteilung der Teststatistik annehmen kann.

Definition 2.1
Der zweiseitige Permutationstest Φ *ist definiert als:*

$$\Phi(\mathbf{x}) := \begin{cases} 1 & T(\mathbf{x}) \leq T_1^\alpha \text{ oder } T_2^\alpha \leq T(\mathbf{x}) \\ 0 & T_1^\alpha < T(\mathbf{x}) < T_2^\alpha, \end{cases}$$

so dass unter H_0^θ *gilt:*

$$\int_{O(\mathbf{x})} \Phi(\mathbf{x}) dP^* = E(\Phi(\mathbf{x})|O(\mathbf{x})) = \alpha' \geq \alpha, \quad \alpha' \in \mathcal{S}_F$$

Für stetige Verteilungen erfüllt der Test Φ für fast alle $\mathbf{x} \in \mathcal{X}^N$ die *Ähnlichkeitseigenschaft*, das heißt, dass das tatsächlich erreichbare Niveau α' nicht von den Daten $\mathbf{x}$ abhängt. Wenn mit einer positiven Wahrscheinlichkeit Bindungen auftreten, ist das tatsächlich erreichbare Niveau von den beobachteten Daten abhängig $\alpha' = \alpha'(\mathbf{x})$ und die Ähnlichkeitseigenschaft somit nicht erfüllt. Sie ist aber (außer für Ein-Punkt-Verteilungen) dennoch asymptotisch erfüllt (Pesarin, 2001, S. 48). Außerdem hängt α' über die Definition der kritischen Werte T_1^α, T_2^α natürlich vom Niveau α ab.

Damit können wir den *zweiseitigen p-Wert* für diesen Test als Anteil der Permutationen definieren, für die die Teststatistik einen größeren Wert annimmt als für die Originaldaten:

$$\lambda := \lambda(T(\mathbf{x})) = \mathrm{P}^{(N)}\left(|T^*| \geq |T(\mathbf{x})| \;\middle|\; O(\mathbf{x}) \right) = \frac{\#\{T^*, |T^*| \geq |T(\mathbf{x})|\}}{M(\mathbf{x})}.$$

$\lambda(T(\mathbf{x}))$ ist somit nicht-steigend in $T(\mathbf{x})$. Es gilt:

$$\lambda > \alpha' \quad \Leftrightarrow \quad T_1^\alpha < T(\mathbf{x}) < T_2^\alpha.$$

2.3.2 Die Invarianzeigenschaft

Um einen gültigen Permutationstest durchführen zu können, müssen die Zufallsvariablen unter der betrachteten Hypothese zwischen den beiden Gruppen austauschbar sein. Das bedeutet, dass die Gruppen unter H_0 die gleiche Verteilung haben müssen. Diese Eigenschaft soll hier formal und allgemein für die Gruppe G von Transformationen definiert werden, die auf $\mathcal{X}^N$ eine Äquivalenzrelation definiert. Die Austauschbarkeit der Beobachtungen unter der Hypothese kann äquivalent durch die *Invarianzeigenschaft* der Permutationsverteilung $\mathrm{P}^* = \mathrm{P}^{(N)}|_{O(\mathbf{x})}$ ausgedrückt werden (vgl. Pesarin, 2001).

Definition 2.2 (Invarianzeigenschaft der Permutationsverteilung)
Die Permutationsverteilung $P^* = P^{(N)}|_{O(\mathbf{x})}$ besitzt die Invarianzeigenschaft, wenn sie für alle Punkte aus einem Orbit $O(\mathbf{x})$ unabhängig von der Populationsverteilung $P^{(N)}$ ist, wobei $P^{(N)} \in \mathcal{P}_0$ sei.

Folgenden Aussagen sind äquivalente Formulierungen der Invarianzeigenschaft:

- $\forall\, g \in G:\ g(\mathbf{x})$ und $\mathbf{x}$ haben die gleiche Wahrscheinlichkeit bezüglich P^* unter H_0^P.
- P^* ist die Gleichverteilung auf $O(\mathbf{x})$ unter H_0^P.

Aus der Definition der Invarianzeigenschaft für P^* auf $O(\mathbf{x})$ leitet sich die Definition der Invarianzeigenschaft der Teststatistik T ab:

Definition 2.3 (Invarianzeigenschaft der Teststatistik)
Die Teststatistik T besitzt die Invarianzeigenschaft, wenn die von P^* induzierte Permutationsverteilungsfunktion der Teststatistik F_T^* unabhängig von $P^{(N)}$ ist, für $P^{(N)} \in \mathcal{P}_0$.

Erfüllt die Permutationsverteilung P^* die Invarianzeigenschaft, so folgt aus der Definition der Permutationsverteilung der Teststatistik F_T^* sofort, dass auch die Teststatistik T die Invarianzeigenschaft erfüllt:

$$F_T^*(t|O(\mathbf{x})) = \mathrm{P}^*(T \leq t).$$

2.3.3 Bedingte Monte-Carlo-Simulationen

Schon bei verhältnismäßig kleinen Stichprobenumfängen von beispielsweise $n_i = 7$ pro Gruppe ist die Anzahl aller möglichen Permutationen $14! \approx 87{,}2 \cdot 10^9$ sehr hoch und führt zu langen Rechenzeiten bei der praktischen Durchführung. Man kann aber statt aller möglichen Permutationen auch nur eine gewisse Anzahl von zufällig ausgewählten Permutationen betrachten und die Permutationsverteilung dadurch

approximieren (Eden & Yates, 1933; Dwass, 1957). Diese Monte-Carlo-Simulationen werden dabei auf dem Orbit $O(\mathbf{x})$ ausgeführt und deshalb als „bedingt“ bezeichnet.

Ist unter H_0 die Invarianzeigenschaft bezüglich der Daten $\mathbf{x}$ erfüllt und sind damit alle Elemente des Orbits $O(\mathbf{x})$ gleichwahrscheinlich, können wir bedingte Monte-Carlo-Simulationen durchführen, um damit die Permutationsverteilungsfunktion der Teststatistik $F_T^*(t|O(\mathbf{x}))$, $t \in \mathbb{R}$ und den p-Wert λ zu schätzen. Dabei geht man wie folgt vor:

1. Berechne $T(\mathbf{x})$.
2. Bestimme eine Permutation der Daten $\mathbf{x}^* \in O(\mathbf{x})$ und berechne $T^* = T(\mathbf{x}^*)$.
3. Führe B unabhängige Wiederholungen von Schritt 2 durch, so dass die Menge $\{T_i^*,\, i = 1, \ldots, B\}$ eine zufällige Stichprobe der Permutationsverteilung von T ist.
4. Berechne die geschätzte Permutationsverteilungsfunktion $\widehat{F}_T^*(t)$, $t \in \mathbb{R}$ und den Schätzer für den p-Wert $\widehat{\lambda}$:
$$\widehat{F}_T^*(t) = \frac{1}{B}\sum_{i=1}^{B} \mathbf{I}_{\{T_i^* \leq t\}}, \quad \widehat{\lambda} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{I}_{\{|T_i^*| \geq |T(\mathbf{x})|\}}.$$
5. Sei α ein festes Signifikanzniveau. Ist $\widehat{\lambda} \leq \alpha$, dann lehne H_0 ab.

Die geschätzte Permutationsverteilungsfunktion $\widehat{F}_T^*(t)$ ist bedingt auf den Orbit $O(\mathbf{x})$ ein unverzerrter und konsistenter Schätzer für die wahre Permutationsverteilungsfunktion $F_T^*(t|O(\mathbf{x}))$ und ebenso ist der Schätzer $\widehat{\lambda}$ bedingt auf $O(\mathbf{x})$ ein unverzerrter und konsistenter Schätzer für den wahren p-Wert λ. Außerdem sichert das folgende Lemma die Konvergenz der geschätzten Permutationsverteilungsfunktion gegen die wahre Permutationsverteilungsfunktion.

Lemma 2.4
Geht die Anzahl der betrachteten zufälligen bedingten Monte-Carlo-Simulationsdurchgänge B gegen ∞, so folgt, dass $\widehat{F}_T^*(t)$ fast sicher bezüglich der Supremumsnorm gegen die wahre Permutationsverteilungsfunktion $F_T^*(t|O(\mathbf{x}))$ konvergiert.

Dabei konvergiert eine Folge von beschränkten Funktionen $(f_n)_{n\in\mathbb{N}}$ mit $f_n : \mathcal{X} \to \mathbb{R}$ in Supremumsnorm gegen eine Funktion $f : \mathcal{X} \to \mathbb{R}$, wenn gilt
$$\lim_{n\to\infty} \|f_n - f\|_\infty = 0, \quad \text{wobei} \quad \|f\|_\infty := \sup\left\{|f(x)| \;\middle|\; x \in \mathcal{X}\right\}.$$

Beweis. Die Aussage des Satzes erhält man durch folgende Argumentation aus dem Satz von Glivenko-Cantelli (Glivenko, 1933; Cantelli, 1933). Sei $\pi \in \mathcal{S}_N$ eine Permutation und $\pi \sim \mathrm{P}^*$, wobei $\mathrm{P}^* = \mathrm{P}^{(N)}|_{O(\mathbf{x})}$ die Gleichverteilung auf allen Permutationen von $\mathbf{x}$ ist. Die Verteilungsfunktion der Teststatistik T nach Anwendung von

π auf $\mathbf{x}$ ist die Permutationsverteilungsfunktion F_T^*. Seien $\pi_1, \ldots, \pi_B \in \mathcal{S}_N$ die B zufällig ausgewählten Permutationen der Monte-Carlo-Simulation. Diese sind damit also unabhängig und identisch nach P^* verteilt. Weiterhin seien $T_i^* = T(\mathbf{x}_{\pi_i})$ die Werte, die die Statistik T an den durch Anwendung von π_i permutierten Beobachtungen $\mathbf{x}_{\pi_i}$ annimmt. Dann sind auch die T_i^* nach F_T^* verteilt. Mit dem Satz von Glivenko-Cantelli folgt dann, dass die empirische Permutationsverteilungsfunktion der T_i^*, also $\widehat{F}_T^*$, in Supremumsnorm mit Rate $c/\sqrt{B}$ gegen die Permutationsverteilungsfunktion F_T^* konvergiert. Angaben für Raten $c/\sqrt{B} \in \mathbb{R}$ können der klassischen Literatur entnommen werden (z.B. Devroye & Lugosi, 2001). □

3 Ein studentisierter Permutationstest für das Behrens-Fisher-Problem

In diesem Kapitel wollen wir einen Permutationstest für das nichtparametrische Zwei-Stichproben Behrens-Fisher-Problem vorstellen und analysieren. Die Definition des Behrens-Fisher-Problems impliziert allerdings, dass die Verteilungen der beobachteten Variablen auch unter der Hypothese nicht gleich sind. Entsprechend sind die Variablen unter H_0 auch nicht austauschbar, so dass die Invarianzeigenschaft (vgl. Abschnitt 2.3.2) nicht erfüllt ist und die in Abschnitt 2.3.1 definierten Permutationstests nicht anwendbar sind.

Aus diesem Grund betrachten wir hier für die nichtparametrische Hypothese $H_0 : \ p = \frac{1}{2}$ eine studentisierte Teststatistik T_N. Bei einem Testproblem, für das die Invarianzeigenschaft erfüllt ist, besteht kein Unterschied zwischen dem Test mit der studentisierten Teststatistik und dem Test mit der gleichen aber nicht studentisierten Teststatistik. Diese beiden Tests sind *permutationsäquivalent*, das heißt, dass ihre Permutationsverteilungen übereinstimmen und sie somit immer zur gleichen Testentscheidung kommen (Pesarin, 2001, S. 43). Betrachtet man einen Test für das Behrens-Fisher-Problem, besteht dagegen allerdings ein Unterschied zwischen Verwendung der studentisierten und nicht studentisierten Teststatistik.

So ermöglicht uns das Betrachten der studentisierten Teststatistik, den Zentralen Grenzwertsatz von Janssen (1997) anzuwenden. Er besagt, dass die Invarianzeigenschaft für den studentisierten Permutationstest asymptotisch erfüllt ist. Der hier vorgestellte studentisierte Permutationstest ist also für das Behrens-Fisher-Problem zumindest asymptotisch ein gültiger Permutationstest. Für kleine Stichprobenumfänge untersuchen wir seine Eigenschaften mithilfe einer Simulationsstudie.

Im nächsten Abschnitt (3.1) werden wir die Teststatistik vorstellen. Danach in Abschnitt 3.2 beschreiben wir drei Methoden, die für das nichtparametrische Behrens-Fisher-Problem vorgeschlagen wurden und insbesondere für kleine Stichprobenumfänge geeignet sein sollen. In Abschnitt 3.3 wird der Zentrale Grenzwertsatz von Janssen (1997) auf die studentisierte Rangstatistik übertragen. Der Beweis des Grenzwertsatzes für die hier verwendete Rangstatistik wird in Abschnitt 3.4 ge-

führt. Die Ergebnisse unserer Simulationsstudie sind in Abschnitt 3.5 beschrieben und zwei Anwendungsbeispiele in Abschnitt 3.6.

3.1 Modell, Hypothese und Teststatistik

Wir betrachten zwei Stichproben von unabhängigen Zufallsvariablen $X_{11},\ldots,X_{1n_1}$ und $X_{21},\ldots,X_{2n_2}$, wobei $X_{ik} \overset{\text{u.i.v.}}{\sim} F_i$, $k = 1,\ldots,n_i$ innerhalb einer Gruppe $i = 1{,}2$. Sei $N = n_1 + n_2$ die Anzahl aller Beoachtungen.

Als Vergleichsmaß betrachten wir den relativen Effekt p:

$$p = P(X_{11} < X_{21}) + \frac{1}{2}P(X_{11} = X_{21}) = \int F_1 dF_2.$$

Entsprechend lautet die Hypothese „kein Behandlungseffekt vorhanden“ (vgl. Abschnitt 2.2):

$$H_0 : \; p = \frac{1}{2}.$$

Der relative Effekt ist invariant unter reinen Skalenalternativen. Deshalb ist in dieser Hypothese die parametrische Hypothese der Gleichheit der Erwartungswerte bei eventuell ungleichen Varianzen als Spezialfall enthalten, wenn die zugrunde liegenden Verteilungsfunktionen stetig und symmetrisch sowie an der Stelle des Erwartungswertes invertierbar sind (vgl. Abschnitt 2.2).

Ein unverzerrter und konsistenter Schätzer für den relativen Effekt p ist (vgl. Abschnitt 2.2)

$$\widehat{p} = \int \widehat{F}_1 d\widehat{F}_2 = \frac{1}{N}(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}) + \frac{1}{2}.$$

Dabei sind $\widehat{F}_i$ die normalisierten Versionen der empirischen Verteilungsfunktionen und R_{ik} die Mittelränge. Zum Testen der Hypothese $p = \frac{1}{2}$ werden wir die Teststatistik $K_N = \sqrt{N}(\widehat{p} - \frac{1}{2})$ verwenden. Um ihre asymptotische Verteilung zu bestimmen, betrachten wir eine *asymptotisch äquivalente* Statistik von unabhängigen Zufallsvariablen. Diese Statistik erhält man aus dem *Asymptotischen Äquivalenzsatz* (Brunner & Munzel (2002), vgl. Anhang Seite 71). Gilt $\frac{N}{n_i} \leq N_0 < \infty$, $i = 1{,}2$ für $N \to \infty$, so erhalten wir

$$\sqrt{N}\,(\widehat{p} - p) \doteqdot \sqrt{N}(\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} + 1 - 2p), \quad \overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik},$$

wobei die nicht-beobachtbaren Zufallsvariablen $Y_{1k} = F_2(X_{1k})$ und $Y_{2k} = F_1(X_{2k})$ die so genannten *Asymptotischen Rang-Transformationen* (ART) sind. Das Symbol $\doteqdot$ steht für *asymptotische Äquivalenz*, das heißt die Differenz der zwei Folgen auf der linken und rechten Seite von $\doteqdot$ konvergiert in Wahrscheinlichkeit gegen 0.

Per Definition sind die ARTs Y_{ik}, $i = 1{,}2$, $k = 1,\ldots,n_i$ gleichmäßig beschränkte, unabhängige und innerhalb einer Gruppe $i = 1{,}2$ identisch verteilte Zufallsvariablen mit Varianzen

$$\sigma_1^2 = Var(F_2(X_{11})) \quad \text{und} \quad \sigma_2^2 = Var(F_1(X_{21})).$$

Der Beweis des Asymptotischen Äquivalenzsatzes beruht auf der Unabhängigkeit der ARTs und deren Mittelwerte $\overline{Y}_{i.}$. Sind $\sigma_1^2, \sigma_2^2 > 0$, dann folgt bei Anwendung des Zentralen Grenzwertsatzes unter H_0, dass

$$\frac{K_N}{\sigma_N} = \frac{\sqrt{N}}{\sigma_N}\left(\widehat{p} - \frac{1}{2}\right) \tag{3.1}$$

asymptotisch standardnormalverteilt ist, wobei die unbekannte Varianz

$$\sigma_N^2 = \frac{N}{n_1 n_2}\left(n_1\sigma_2^2 + n_2\sigma_1^2\right)$$

konsistent aus den Daten geschätzt werden muss. Der von Brunner & Munzel (2002) angegebene Schätzer V_N^2 für σ_N^2 erfüllt diese Eigenschaft, wobei:

$$V_N^2 = N\left(\frac{1}{n_2}\widehat{\sigma}_1^2 + \frac{1}{n_1}\widehat{\sigma}_2^2\right),$$

mit

$$\widehat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(R_{ik} - R_{ik}^{(i)} - \overline{R}_{i.} + \frac{n_i + 1}{2}\right)^2.$$

Dabei ist R_{ik} der Gesamtrang von X_{ik} bezüglich aller N Beobachtungen und $R_{ik}^{(i)}$ der interne Rang von X_{ik} bezüglich der n_i Beobachtungen $X_{i1},\ldots,X_{in_i}$ innerhalb von Gruppe i, $i = 1{,}2$. Der Varianzschätzer V_N^2 hängt also nur über die Ränge R_{ik}, $R_{ik}^{(i)}$ von den Daten ab.

Die Differenz der Ränge $R_{ik} - R_{ik}^{(i)}$ entspricht gerade der normierten Platzierung $\widehat{F}_j(X_{ik})$ (vgl. 2.1). Stellen wir diese mithilfe der Zählfunktion dar

$$\widehat{F}_j(X_{ik}) = \frac{1}{n_j}\sum_{l=1}^{n_j} c(X_{ik} - X_{jl}),$$

so wird deutlich, dass die Rangdifferenz die Lage der Beobachtung X_{ik} aus Gruppe i bezüglich der Beobachtungen aus Gruppe j beschreibt. Aus diesem Grund wird der Varianzschätzer V_N^2 Null, wenn die Beobachtungen beider Gruppen komplett getrennte Wertebereiche haben, weil dann

$$\widehat{F}_j(X_{ik}) - \frac{1}{n_i}\sum_{k=1}^{n_i}\widehat{F}_j(X_{ik}) = 0$$

für $k = 1, \ldots, n_i$ und $i = 1,2$. In diesem Fall wird der Varianzschätzer V_N^2 durch eine untere Grenze ersetzt, die gerade größer als Null ist. Diese untere Grenze wird zum Beispiel angenommen, wenn alle n_1 kleinsten Beobachtungen in Gruppe 1 auftreten und die größte Beobachtung aus Gruppe 1 den gleichen Wert annimmt wie die kleinste Beobachtung aus Gruppe 2. Dies sei außerdem die einzige Bindung, die auftritt. Nehme ohne Einschränkung X_{1n_1} den größten Wert in Gruppe 1 an und X_{21} den kleinsten Wert in Gruppe 2. Dann erhält man für die Gesamtränge und die internen Ränge die in Tabelle 3.1 angegebenen Werte.

Tabelle 3.1: Ränge und interne Ränge im Fall kleinster Varianz

R_{ik}	$1,$	$\ldots,$	$n_1 - 1,$	$n_1 + \frac{1}{2}$	$n_1 + \frac{1}{2},$	$n_1 + 2,$	$\ldots,$ $n_1 + n_2$
$R_{ik}^{(i)}$	$1,$	$\ldots,$	$n_1 - 1,$	n_1	$1,$	$2,$	$\ldots,$ n_2
$R_{ik} - R_{ik}^{(i)}$	$0,$	$\ldots,$	$0,$	$\frac{1}{2}$	$n_1 - \frac{1}{2},$	$n_1,$	$\ldots,$ n_1
$\overline{R}_{i\cdot} - \overline{R}_{i\cdot}^{(i)}$	$\frac{1}{2n_1}$				$n_1 - \frac{1}{2n_2}$		

Daraus folgt dann als Wert für die untere Grenze des Varianzschätzers:

$$V_{N,\min}^2 = \frac{N}{n_2}\left(\frac{1}{4n_1^2} + \frac{n_1 - 1}{4n_1^2}\right) + \frac{N}{n_1}\left(\frac{n_2 - 1}{4n_2^2} + \frac{1}{4n_2^2}\right) = \frac{N}{2n_1 n_2}.$$

Da

$$\frac{N}{2n_1 n_2} \to 0 \quad \text{für} \quad N \to \infty$$

spielt diese Ersetzung bei der Betrachtung der asymptotischen Eigenschaften der Teststatistik keine Rolle. Sie kann allerdings zu konservativem Verhalten des Tests führen (vgl. dazu die Simulationsstudie Abschnitt 3.5).

Ersetzen wir σ_N in (3.1) durch V_N, so erhalten wir die Teststatistik

$$T_N = \frac{\overline{R}_{2\cdot} - \overline{R}_{1\cdot}}{V_N} \sqrt{\frac{n_1 n_2}{N}}. \tag{3.2}$$

Satz 3.1 (Asymptotische Verteilung von T_N)
Die Teststatistik T_N ist unter $H_0 : p = \frac{1}{2}$ asymptotisch standardnormalverteilt.

Beweis. Folgt mit obiger Herleitung aus dem Asymptotischen Äquivalenzsatz (siehe Brunner & Munzel (2002), vgl. Anhang A.1). □

3.2 Methoden für kleine Stichprobenumfänge

Die Statistik T_N ist also nur asymptotisch normalverteilt. Simulationsstudien haben gezeigt, dass recht große Stichprobenumfänge nötig sind, um eine zufriedenstellende

Approximation zu erreichen. In vielen medizinischen und biologischen Anwendungen stehen allerdings oft nur wenige Beobachtungen zur Verfügung. Außerdem sind Annahmen über die Stetigkeit der Verteilungen oft nicht angebracht, was bei einigen für diese Situation sonst geeigneten Verfahren (z.B. Pesarin, 2001) eine wichtige Voraussetzung ist.

Im Folgenden wollen wir Verfahren vorstellen, die für die Anwendung auf ein Behrens-Fisher-Design konzipiert wurden und speziell auch für kleine Stichprobenumfänge geeignet sind. In den nächsten Abschnitten beschreiben wir neben dem studentisierten Permutationstest drei weitere Verfahren. Alle vier Verfahren werden in Abschnitt 3.5 in einer Simulationsstudie verglichen.

3.2.1 t-Approximation

Brunner & Munzel (2000) schlagen vor, die Verteilungsfunktion der Teststatistik T_N mit einer $t_{\widehat{f}}-$Verteilung zu approximieren. Der Freiheitsgrad $\widehat{f}$ wird dabei wie im parametrischen Fall mithilfe der Satterthwaite-Smith-Welch Approximation bestimmt und durch

$$\widehat{f} = \frac{\left(\sum_{i=1}^{2} \widehat{\sigma}_i^2/(N-n_i)\right)^2}{\sum_{i=1}^{2}(\widehat{\sigma}_i^2/(N-n_i))^2/(n_i-1)} \tag{3.3}$$

geschätzt. Dabei konvergiert $\widehat{f} \to \infty$ wenn $n_i \to \infty$ geht. Das heißt, dass die $t_{\widehat{f}}-$Verteilung gegen eine Standardnormalverteilung konvergiert und die Approximation somit asymptotisch korrekt ist.

3.2.2 Likelihood-Ratio-Test

Troendle (2002) gibt für diese Situation einen Likelihood-Ratio-Test an. Er stellt eine rekursive Methode vor, die das Bestimmen der n_1+n_2+3 Parameter des Maximierungsproblems auf Dimension 1 reduziert und somit eine numerische Berechnung überhaupt praktikabel macht.

Für die Anwendung der Methode wird vorausgesetzt, dass die Verteilungsfunktionen F_i diskret sind. Sei $n \leq N$ die Anzahl unterschiedlicher Werte, die die Beobachtungen $x_{11}, \ldots, x_{1n_1}$, $x_{21}, \ldots, x_{2n_2}$ annehmen. Dabei können manche der n Werte also mehrfach vorkommen. Sei deshalb m_t die Multiplizität der Werte der ersten Stichprobe $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1})'$ für den t-ten Wert der gemeinsamen geordneten Liste von Werten beider Gruppen, $t = 1, \ldots, n$. Ebenso sei $m'_{t'}$ die Multiplizität der Werte der zweiten Stichprobe $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n_2})'$, $t' = 1, \ldots, n$. Da die Likelihoodfunktion der Daten von den Ableitungen der Verteilungen abhängt, liegt es nahe, die Höhe der Sprünge der Verteilungsfunktionen F_i zu betrachten. Diese Sprunghöhen bezüglich der gemeinsamen geordneten Liste von angenommenen Werten bezeichnen

wir mit $q_{i1}, \ldots, q_{in}$. Das Maximieren der Log-Likelihood

$$L(q_{11}, \ldots, q_{1n}, q_{21}, \ldots, q_{2n}) = \sum_{t=1}^{n} \log q_{1t}^{m_t} + \sum_{t'=1}^{n} \log q_{2t'}^{m'_{t'}}$$

unter den Nebenbedingungen

$$\sum_{t=1}^{n} q_{1t} = 1, \quad \sum_{t'=1}^{n} q_{2t'} = 1$$

führt zu den Schätzern

$$\widehat{q}_{1t} = \frac{m_t}{n_1}, \quad \widehat{q}_{2t'} = \frac{m'_{t'}}{n_2}.$$

Damit ist die Log-Likelihood unter der Alternative bestimmbar. Unter der Hypothese $H_0 : p = \frac{1}{2}$ lautet das Maximierungsproblem:

$$\max \left\{ \sum_{t=1}^{n} \log q_{1t}^{m_t} + \sum_{t'=1}^{n} \log q_{2t'}^{m'_{t'}} \right\}$$

mit den Nebenbedingungen

$$\sum_{t'=1}^{n} q_{2t'} \sum_{t=t'+1}^{n} q_{1t} = \sum_{t'=1}^{n} q_{2t'} \sum_{t=1}^{t'-1} q_{1t}, \quad \sum_{t=1}^{n} q_{1t} = 1, \quad \sum_{t'=1}^{n} q_{2t'} = 1. \tag{3.4}$$

Dabei entspricht die erste Nebenbedingung in (3.4) der zur Hypothese äquivalenten Forderung $\mathrm{P}(X_{11} > X_{21}) = \mathrm{P}(X_{11} < X_{21})$. Wendet man auf dieses Maximierungsproblem Lagrange-Multiplikatoren an, so erhält man $2n + 3$ Gleichungen. Troendle zeigt nun, dass man dieses System durch geschicktes Einsetzen dieser Gleichungen auf das Bestimmen eines unbekannten Parameters reduzieren kann. Dafür muss man eine Gleichung numerisch durch ein eindimensionales Nullstellenverfahren lösen. Dabei kann es zu Lösungen kommen, die die Nebenbedingungen nicht erfüllen. Troendle schreibt, dass sein Fortran-Simulationsprogramm für $n_1 > 10$ immer eine mögliche Lösung gefunden hat.

Sind die Maximum-Likelihood-Schätzer

$$\widetilde{q}_{11}, \ldots, \widetilde{q}_{1n}, \widetilde{q}_{21}, \ldots, \widetilde{q}_{2n}$$

bestimmt, so kann die Likelihood-Ratio-Teststatistik

$$L(\widehat{q}_{11}, \ldots, \widehat{q}_{1n}, \widehat{q}_{21}, \ldots, \widehat{q}_{2n}) - L(\widetilde{q}_{11}, \ldots, \widetilde{q}_{1n}, \widetilde{q}_{21}, \ldots, \widetilde{q}_{2n})$$

berechnet werden. Die Verteilung der Likelihood-Ratio-Statistik wird dann durch ein Permutationsverfahren bestimmt. Dabei werden neue Datenvektoren nicht durch Anwendung von Permutationen auf die Daten sondern durch Simulationen aus der diskreten Verteilung gewonnen, die durch die bedingten Maximum-Likelihood-Schätzer definiert wird. Eine ausführliche Beschreibung des Testverfahrens findet man in Troendle (2002).

3.2.3 Bootstraptest

Der Bootstraptest von Reiczigel *et al.* (2005) basiert auf Welchs Rang-Test und verwendet entsprechend die folgende Teststatistik:

$$T = \frac{\overline{R}_{2.} - \overline{R}_{1.}}{\sqrt{\sum_{i=1}^{2}\sum_{k=1}^{n_i}\frac{1}{n_i}\frac{1}{n_i-1}(R_{ik}-\overline{R}_{..})^2}}, \quad \overline{R}_{..} = \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i} R_{ik}.$$

Um die Bootstrapverteilung der Teststatistik zu ermitteln, werden die Beobachtungen der einer Stichprobe zunächst transformiert, so dass die Nullhypothese $H_0 : p = \frac{1}{2}$ erfüllt ist. Durch Sensitivitätsanalysen begründet, verwenden Reiczigel *et al.* (2005) die folgende Transformation mit dem Hodges-Lehmann-Schätzer für den Shift-Effekt:

$$x'_{2k} = x_{2k} + c, \quad k = 1, \ldots, n_2,$$

wobei

$$c = \text{Median}\{x_{1k} - x_{2l},\, k = 1, \ldots, n_1,\, l = 1, \ldots, n_2\}.$$

Aus den transformierten Daten $x_{11}, \ldots, x_{1n_1}, x'_{21}, \ldots, x'_{2n_2}$ werden dann separat aus den beiden Stichproben mit Zurücklegen jeweils n_1 bzw. n_2 Beobachtungen gezogen.

Die dort vorgestellte Simulationsstudie legt nahe, dass sich dieser Bootstraptest und die t-Approximation von Brunner und Munzel sehr ähnlich verhalten.

3.2.4 Eigenschaften der existierenden Methoden

Simulationsstudien zeigen, dass die oben genannten Verfahren bei der Einhaltung des festgelegten Niveaus Defizite haben, wenn die zugrunde liegenden Verteilungen nicht symmetrisch sind oder wenn die Stichprobenumfänge klein sind. Troendles Likelihood-Ratio-Test wird z.B. bei bimodalen Verteilungen etwas liberal. Außerdem war der Maximum-Likelihood-Schätzer bei Stichprobenumfang sieben bis zu 173 mal bei 10'000 Simulationen nicht berechenbar. Der Bootstraptest von Reiczigel *et al.* (2005) tendiert zu konservativem Verhalten, insbesondere bei kleinen Stichprobenumfängen. Die t-Approximation wiederum wird leicht liberal, wenn man zweiseitige Tests durchführt und weist insbesondere für kleine nominale Niveaus recht große Abweichungen davon auf. Die Verwendung von Adjustierungsmethoden bei Multiplizität ist dadurch problematisch.

3.2.5 Studentisierter Permutationstest

Aufgrund dieser Probleme stellen wir für das nichtparametrische Behrens-Fisher-Problem einen Permutationstest vor. Für das parametrische Modell beschreibt Janssen (1997) einen Permutationstest, der auf einer studentisierten Teststatistik basiert. Diese Idee geht auf Neuhaus (1993) zurück, der sie bei Überlebenszeitanalysen anwendete. Wir wollen dies nun auf die standardisierte Rangstatistik T_N übertragen,

indem wir die Entscheidung, die Hypothese abzulehnen, anhand der Permutationsverteilung der Statistik treffen.

Da wir die Hypothese $H_0 : \ p = \frac{1}{2}$ betrachten, sind die Beobachtungen unter der Hypothese nicht austauschbar, weil sie nicht die gleichen Verteilungen haben müssen. Deswegen ist die Permutationsverteilung P^* nicht die Gleichverteilung.

Janssen (1997) zeigt für lineare studentisierte Statistiken einen Grenzwertsatz, nach dem die asymptotische Permutationsverteilung dieser Teststatistiken die Normalverteilung ist. Damit sind die Permutationsverteilungen solcher Teststatistiken zumindest asymptotisch unabhängig von der Verteilung der Daten und erfüllen asymptotisch die in Abschnitt 2.3.2 geforderte Invarianzeigenschaft.

Wie wir im Abschnitt 3.3 zeigen werden, kann man eine solche Aussage auch für die studentisierte Rangstatistik T_N nachweisen. Die Gültigkeit der dafür nötigen Bedingungen an die Koeffizienten und den Varianzschätzer der Teststatistik werden in Abschnitt 3.4 bewiesen.

Allerdings liegt unser Fokus auf der Anwendbarkeit der Verfahren bei geringen Stichprobenumfängen. Dabei kann gerade die Kombination der zwei nichtparametrischen Methoden (Permutationstest mit Rangstatistik) zu einem für solche Situationen geeigneten Verfahren führen. Die Verwendung eines Permutationstests bei kleinen Stichprobenumfängen liegt nahe, weil keine Verteilungsannahme getroffen werden muss. Rangstatistiken dagegen haben den Vorteil, robust gegen Ausreißer zu sein und sind außerdem nicht nur auf stetige Daten, sondern auch beispielweise auf Scores anwendbar. Deshalb kann man auf gute Eigenschaften des vorgestellten Permutationstests insbesondere bei kleinen Stichprobenumfängen hoffen. Dass dies tatsächlich der Fall ist, zeigt die Simulationsstudie in Abschnitt 3.5.

Für die Berechnung der Permutationsverteilung betrachten wir alle Permutationen von N Elementen $\pi \in \mathcal{S}_N$ (vgl. Gleichung (2.3)). Die Vektoren der permutierten Zufallsvariablen werden mit $\mathbf{X}_\pi$, $\mathbf{R}_\pi$ und $\mathbf{R}_\pi^{(1)}, \mathbf{R}_\pi^{(2)}$ bezeichnet, wobei

$$\begin{aligned}
\mathbf{X}_\pi &= (X_{\pi(11)}, \ldots, X_{\pi(1n_1)}, X_{\pi(21)}, \ldots, X_{\pi(2n_2)}), \\
\mathbf{R}_\pi &= (R_{\pi(11)}, \ldots, R_{\pi(1n_1)}, R_{\pi(21)}, \ldots, R_{\pi(2n_2)}), \\
\mathbf{R}_\pi^{(i)} &= (R^{(i)}_{\pi(i1)}, \ldots, R^{(i)}_{\pi(in_i)}), \quad i = 1{,}2.
\end{aligned}$$

Wenn die Originaldaten $\mathbf{x}$ durch die permutierten Beobachtungen $\mathbf{x}_\pi$ ersetzt werden, bezeichnen wir die damit neu berechnete Varianz mit V_N^{*2}, wobei

$$V_N^{*2} = V_N^2(\mathbf{R}_\pi,\ \mathbf{R}_\pi^{(1)},\ \mathbf{R}_\pi^{(2)}),$$

da der Varianzschätzer ja nur über die Ränge und internen Ränge von den Daten abhängt. Die mit den permutierten Beobachtungen neu berechnete Teststatik sei T_N^*, wobei

$$T_N^* = T_N(\mathbf{R}_\pi,\ \mathbf{R}_\pi^{(1)},\ \mathbf{R}_\pi^{(2)}).$$

Auch T_N ist nur über die Ränge und internen Ränge von den Daten abhängig.

Wird eine Permutation auf den Vektor aller Beobachtungen angewendet, ändert sich die Gruppenzugehörigkeit der Beobachtungen und dadurch auch der Wert der internen Ränge und Rangsummen. Deswegen muss für jede Permutation die Varianz V_N^{*2} neu berechnet werden, um den neuen Wert der Teststatistik T_N^* bestimmen zu können.

Der p-Wert des Permutationstests ist dann als Anteil der Permutationen definiert, für die der Wert der Teststatistik mit den Originalbeobachtungen $T_N(\mathbf{x})$ betragsmäßig kleiner oder gleich dem Wert der Teststatistik mit den permutierten Beobachtungen T_N^* ist. Führen wir $n_{\text{perm}} = \#S$ zufällig ausgewählte Permutationen mit $\pi \in S \subset \mathcal{S}_N$ durch, so definieren wir den p-Wert (vgl. Abschnitte 2.3.1 und 2.3.3) als

$$p = \frac{1}{n_{\text{perm}}} \# \left\{ \pi \in S,\, |T_N(\mathbf{x})| \leq |T_N^*| \right\}.$$

3.3 Studentisiertes Permutieren

In diesem Abschnitt wenden wir den Zentralen Grenzwertsatz von Janssen (1997) für bedingte Permutationsverteilungen linearer studentisierter Teststatistiken an (siehe Anhang, Seite 72), um zu zeigen, dass unser Permutationstest asymptotisch das festgelegte Niveau einhält. Insbesondere zeigen wir, dass die asymptotische Permutationsverteilung von T_N unabhängig von der den Daten zugrunde liegenden Verteilung ist. Damit erfüllt der auf T_N basierte Permutationstest zumindest asymptotisch die in Abschnitt 2.3.2 beschriebene Invarianzeigenschaft.

Da der Grenzwertsatz von Janssen nicht voraussetzt, dass die Zufallsvariablen unabhängig sind, kann er direkt auf die Rangstatistik T_N angewendet werden (vgl. Janssen, 1997). Um der Notation von Janssen zu folgen, formulieren wir die Teststatistik um und definieren dazu die Zufallsvariablen $\widehat{Z}_{ik}$:

$$\widehat{Z}_{ik} := \frac{1}{N}\left(R_{ik} - R_{ik}^{(i)} - \frac{n_j}{2}\right) = \frac{n_j}{N}\left(\widehat{F}_j(X_{ik}) - \frac{1}{2}\right), \quad j \neq i,\, i,j = 1,2. \tag{3.5}$$

Da die empirischen Verteilungsfunktionen $\widehat{F}_i$ nur Werte in $[0,1]$ annehmen, sind die $\widehat{Z}_{ik}$ durch $1/2$ gleichmäßig beschränkt:

$$\mathrm{P}\left(|\widehat{Z}_{ik}| > \frac{1}{2}\right) = 0.$$

Wir setzen nun die $\widehat{Z}_{ik}$ in die Teststatistik T_N ein:

$$T_N = \frac{\overline{R}_{2.} - \overline{R}_{1.}}{\sqrt{N \sum_{i=1}^{2} \sum_{k=1}^{n_i} \frac{1}{N-n_i} \frac{1}{n_i-1} \left(R_{ik} - R_{ik}^{(i)} - \overline{R}_{i.} + \frac{n_i+1}{2} \right)^2}} \sqrt{\frac{n_1 n_2}{N}}$$

$$= \frac{\overline{\widehat{Z}}_{2.} - \overline{\widehat{Z}}_{1.}}{\sqrt{\frac{1}{N} \sum_{i=1}^{2} \sum_{k=1}^{n_i} \frac{N-n_i}{n_i-1} \left(\frac{N}{N-n_i} \right)^2 \left(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{i.} \right)^2}} \sqrt{\frac{n_1 n_2}{N}},$$

da

$$\overline{\widehat{Z}}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} \widehat{Z}_{ik} = \frac{1}{N} \left(\overline{R}_{i.} - \frac{n_i+1}{2} - \frac{n_j}{2} \right) = \frac{1}{N} \left(\overline{R}_{i.} - \frac{N+1}{2} \right).$$

In der Notation von Janssen für lineare Rangstatistiken lautet T_N dann:

$$T_N = \frac{\sum_{i=1}^{2} \sum_{k=1}^{n_i} c_{ik} \widehat{Z}_{ik}}{\sqrt{\sum_{i=1}^{2} \sum_{k=1}^{n_i} \frac{N}{N-n_i} \frac{1}{n_i-1} (\widehat{Z}_{ik} - \overline{\widehat{Z}}_{i.})^2}} \tag{3.6}$$

$$\text{mit} \quad c_{ik} = \begin{cases} -\sqrt{\frac{n_1 n_2}{N}} \frac{1}{n_1} & i = 1 \\ \sqrt{\frac{n_1 n_2}{N}} \frac{1}{n_2} & i = 2. \end{cases}$$

Um den Zentralen Grenzwertsatz von Janssen (1997) anwenden zu können, müssen einige Bedingungen für die Koeffizienten und den Varianzschätzer der Teststatistik erfüllt sein. Unter diesen Bedingungen zeigt Janssen, dass die asymptotische Permutationsverteilung der Teststatistik gegen eine Normalverteilung konvergiert. Für unsere Teststatistik T_N zeigen wir die Konvergenz für eine geeignete Auswahl der Bedingungen, die im folgenden Satz angegeben ist.

Satz 3.2 (Asymptotische Permutationsverteilung)
Es gelte, dass

$$\exists\, \kappa \in (0,1): \ \frac{n_1}{N} \overset{N\to\infty}{\longrightarrow} \kappa \ \Rightarrow \ \frac{n_2}{N} \overset{N\to\infty}{\longrightarrow} 1-\kappa$$

und die Varianzen der ART Y_{ik} seien strikt positiv:

$$\sigma_1^2 = Var(Y_{11}) = Var(F_2(X_{11})) > 0, \quad \sigma_2^2 = Var(Y_{21}) = Var(F_1(X_{21})) > 0.$$

Sind für die Teststatistik T_N dann die Bedingungen (3.7) - (3.10) erfüllt, so ist die asymptotische Permutationsverteilung von T_N eine Normalverteilung mit Erwartungswert 0 und Varianz $\tau^2 = \kappa(1-\kappa)$:

$$\sup_{t\in\mathbb{R}} \left(\left| P^* \left(T_N(\mathbf{R}_\pi, \mathbf{R}_\pi^{(1)}, \mathbf{R}_\pi^{(2)}) \le t \right) - \Phi(t/\tau) \right| \right) \xrightarrow{P} 0,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Die Bedingungen an die Teststatistik T_N sind:

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik}^2 = 1 \quad \forall N \in \mathbb{N} \tag{3.7}$$

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik} = 0 \quad \forall N \in \mathbb{N} \tag{3.8}$$

$$\max_{i,k} |c_{ik}| \to 0 \quad wenn \quad N \to \infty \tag{3.9}$$

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..})^2 1_{[d,\infty)}\left(\left|\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..}\right|\right) \longrightarrow 0 \quad P-f.s. \quad d \to \infty \tag{3.10}$$

$$\liminf_{N\to\infty} \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..})^2 > 0 \quad P-f.s. \tag{3.11}$$

$$\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..})^2 (V_N^{*2})^{-1} \overset{P\otimes P^*}{\longrightarrow} \kappa(1-\kappa) \quad wenn \quad N \to \infty. \tag{3.12}$$

Beweis. Die Aussage gilt nach Satz (3.3) aus Janssen (1997). □

Die asymptotische Permutationsverteilungsfunktion F_T^* von T_N ist also eine Normalverteilung $\mathcal{N}(0,\, \kappa(1-\kappa))$ und somit unabhängig von der den Daten zugrunde liegenden Verteilung. Damit ist die Invarianzeigenschaft für T_N asymptotisch unter $H_0 : \; p = \frac{1}{2}$ erfüllt und der auf T_N basierende Permutationstest ist asymptotisch verteilungsfrei.

Wir müssen nun zeigen, dass die Bedingungen (3.7) bis (3.10) von der Teststatistik T_N erfüllt werden. Dies tun wir im folgenden Abschnitt.

3.4 Nachweis der Bedingungen für den Grenzwertsatz

- Nachweis von (3.7)-(3.9):
  Die Bedingungen (3.7) bis (3.9) folgen sofort aus der Definition der c_{ik} in (3.6). So gilt Bedingung (3.7), da

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik}^2 = n_1 \frac{n_1 n_2}{N}\frac{1}{n_1^2} + n_2 \frac{n_1 n_2}{N}\frac{1}{n_2^2} = \frac{n_2 + n_1}{N} = 1$$

  ist. Bedingung (3.8) ist erfüllt, wegen

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik} = -n_1\sqrt{\frac{n_1 n_2}{N}}\frac{1}{n_1} + n_2\sqrt{\frac{n_1 n_2}{N}}\frac{1}{n_2} = 0$$

und dass Bedingung (3.9) gilt, sieht man durch

$$\max_{i,k}|c_{ik}| = \max_{i,k}\left\{\sqrt{\frac{n_2}{Nn_1}}, \sqrt{\frac{n_1}{Nn_2}}\right\} \overset{N\to\infty}{\longrightarrow} 0, \quad \text{da} \quad 0 < \frac{N}{n_i} \leq N_0 < \infty.$$

- Um Bedingung (3.10) zu zeigen, berechnen wir zunächst den Mittelwert der $\widehat{Z}_{ik}$:

$$\begin{aligned}\overline{\widehat{Z}}_{..} &= \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik} = \frac{1}{N}(n_1\overline{\widehat{Z}}_{1.} + n_2\overline{\widehat{Z}}_{2.}) \\ &= \frac{1}{N}\left[\frac{n_1}{N}\left(\overline{R}_{1.} - \frac{n_1+1}{2} - \frac{n_2}{2}\right) + \frac{n_2}{N}\left(\overline{R}_{2.} - \frac{n_2+1}{2} - \frac{n_1}{2}\right)\right] \\ &= \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}R_{ik} - \frac{(n_1+n_2)(N+1)}{2N}\right) = 0. \end{aligned} \tag{3.13}$$

  Daraus folgt

$$|\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..}| = |\widehat{Z}_{ik}| \in \left[0, \frac{1}{2}\right].$$

  Der obige Ausdruck ist also durch 1/2 beschränkt, weshalb der Limes in (3.10) P−f.s. gleich 0 ist für $d > \frac{1}{2}$.

- Nachweis von (3.11): Da der Mittelwert $\overline{\widehat{Z}}_{..} = 0$ ist, bleibt zu zeigen, dass

$$\liminf_{N\to\infty}\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2 > 0 \quad \text{P} - \text{f.s.}$$

  Um diesen Grenzwert bei Konvergenz in Wahrscheinlichkeit zu bestimmen, betrachten wir zunächst die Zufallsvariablen

$$Z_{ik} = \frac{n_j}{N}\left(F_j(X_{ik}) - \frac{1}{2}\right), \quad i \neq j,\ i,j = 1,2$$

  deren empirische Pendants gerade die

$$\widehat{Z}_{ik} = \frac{n_j}{N}\left(\widehat{F}_j(X_{ik}) - \frac{1}{2}\right) = \frac{1}{N}\left(R_{ik} - R_{ik}^{(i)} - \frac{n_j}{2}\right)$$

  sind. Es gilt mit $\widetilde{p}_i = \int F_j dF_i$ und $\sigma_i^2 = Var(F_j(X_{ik}))$

$$\begin{aligned}\mathrm{E}(Z_{ik}) &= \frac{n_j}{N}\left(\int F_j dF_i - \frac{1}{2}\right) = \frac{n_j}{N}\left(\widetilde{p}_i - \frac{1}{2}\right), \\ Var(Z_{ik}) &= \left(\frac{n_j}{N}\right)^2 Var(F_j(X_{ik})) = \left(\frac{n_j}{N}\right)^2\sigma_i^2,\end{aligned}$$

wobei $i \neq j$, $i, j = 1, 2$. Wir können Z_{ik} damit auffassen als

$$Z_{ik} = \left(\frac{n_j}{N}\right)\left(\widetilde{p}_i - \frac{1}{2} + \sigma_i \xi_{ik}\right),$$

wobei ξ_{ik} unabhängige Zufallsvariablen mit Erwartungswert 0 und Varianz 1 sind, die innerhalb einer Gruppe, das heißt für festes i, die gleiche Verteilung haben. Die Unabhängigkeit der ξ_{ik} folgt, da wir vorausgesetzt haben, dass die X_{ik} und damit auch die Z_{ik} unabhängig sind. Somit erhalten wir:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i} Z_{ik}^2 =& \frac{1}{N}\left[\sum_{k=1}^{n_1}\left(\frac{n_2}{N}\right)^2\left(\widetilde{p}_1 - \frac{1}{2} + \sigma_1\xi_{1k}\right)^2 + \sum_{k=1}^{n_2}\left(\frac{n_1}{N}\right)^2\left(\widetilde{p}_2 - \frac{1}{2} + \sigma_2\xi_{2k}\right)^2\right] \\
=& \frac{1}{N}\left(\frac{n_2}{N}\right)^2\left(n_1\left(\widetilde{p}_1^2 - \widetilde{p}_1 + \frac{1}{4}\right) + \sigma_1^2\sum_{k=1}^{n_1}\xi_{1k}^2 + 2\left(\widetilde{p}_1 - \frac{1}{2}\right)\sigma_1\sum_{k=1}^{n_1}\xi_{1k}\right) \\
&+ \frac{1}{N}\left(\frac{n_1}{N}\right)^2\left(n_2\left(\widetilde{p}_2^2 - \widetilde{p}_2 + \frac{1}{4}\right) + \sigma_2^2\sum_{k=1}^{n_2}\xi_{2k}^2 + 2\left(\widetilde{p}_2 - \frac{1}{2}\right)\sigma_2\sum_{k=1}^{n_2}\xi_{2k}\right) \\
\xrightarrow{\text{P-f.s.}}& \kappa(1-\kappa)\left[(1-\kappa)\left(\widetilde{p}_1^2 - \widetilde{p}_1 + \frac{1}{4} + \sigma_1^2\right) + \kappa\left(\widetilde{p}_2^2 - \widetilde{p}_2 + \frac{1}{4} + \sigma_2^2\right)\right] \\
=:& \ \rho \quad \text{für} \quad N \to \infty. \qquad (3.14)
\end{aligned}$$

Dabei ist die Konstante ρ positiv. Das gilt da $\widetilde{p}_1 = 1 - p$, $\widetilde{p}_2 = p$ und der relative Effekt p im Intervall $[0, 1]$ liegt, woraus $\widetilde{p}_i^2 - \widetilde{p}_i + \frac{1}{4} \in [0, \frac{1}{4}]$ folgt. Außerdem sind nach Voraussetzung die $\sigma_1^2 > 0$ und $\kappa \in (0, 1)$ und damit $\rho > 0$.

Allerdings benötigen wir die Konvergenzaussage für die empirisch bestimmbaren Ersetzungen $\widehat{Z}_{ik}$ der nicht-beobachtbaren Zufallsvariablen Z_{ik}. Für die Verteilungsfunktionen F_i und ihre Schätzer $\widehat{F}_i$ wird in Brunner & Munzel (2002) Konvergenz im quadratischen Mittel gezeigt:

$$\mathrm{E}\left[\widehat{F}_j(X_{ik}) - F_j(X_{ik})\right]^2 \leq \frac{1}{n_j} \quad i \neq j,\ i, j = 1{,}2.$$

Dies folgt da $\mathrm{E}(c(X_{ik} - X_{jl})) = \int F_j dF_i$ ist und aufgrund der Unabhängigkeit der X_{ik}. Daraus folgt, dass $\widehat{F}_j(X_{ik}) - F_j(X_{ik}) \xrightarrow{\mathrm{P}} 0$. Dies überträgt sich auf die $\widehat{Z}_{ik}$:

$$\widehat{Z}_{ik} - Z_{ik} = \frac{n_j}{N}\left(\widehat{F}_j(X_{ik}) - \frac{1}{2}\right) - \frac{n_j}{N}\left(F_j(X_{ik}) - \frac{1}{2}\right) \xrightarrow{\mathrm{P}} 0$$

Wir erhalten also, dass

$$\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2 \xrightarrow{\mathrm{P}} \rho > 0. \qquad (3.15)$$

- Nachweis von (3.12): Um die bedingte Permutationsverteilung von V_N^2 zu bestimmen, betrachten wir die folgende Zerlegung $V_N^2 = W_1 - W_2^2 - W_3^2$, wobei

$$V_N^2 = \sum_{i=1}^{2}\sum_{k=1}^{n_i} \frac{N}{n_j}\frac{1}{n_i - 1}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{i.})^2$$

ist und

$$W_1 = \frac{N}{n_2}\frac{1}{n_1 - 1}\sum_{l=1}^{n_1}\widehat{Z}_{1l}^2 + \frac{N}{n_1}\frac{1}{n_2 - 1}\sum_{k=1}^{n_2}\widehat{Z}_{2k}^2,$$
$$W_2 = \sqrt{\frac{N}{n_2}\frac{n_1}{n_1 - 1}}\overline{\widehat{Z}}_{1.}, \quad W_3 = \sqrt{\frac{N}{n_1}\frac{n_2}{n_2 - 1}}\overline{\widehat{Z}}_{2.}.$$

Wir können also nun die Verteilung für jeden Teil einzeln bestimmen, indem wir die bedingten Erwartungswerte und Varianzen berechnen. Wenn wir die permutierten Beobachtungen verwenden, um die W_i zu berechnen, bezeichnen wir sie mit W_i^*. Mit (3.13) erhalten wir für W_2^* und W_3^*:

$$\mathrm{E}(W_2^*|\mathbf{R}) = \sqrt{\frac{N}{n_2}\frac{n_1}{n_1 - 1}}\,\overline{\widehat{Z}}_{..} = 0, \quad \mathrm{E}(W_3^*|\mathbf{R}) = \sqrt{\frac{N}{n_1}\frac{n_2}{n_2 - 1}}\,\overline{\widehat{Z}}_{..} = 0.$$

Zur Berechnung der Varianzen verwenden wir die Varianzformel aus Hájek & Šidák (1967) (vgl. Anhang Satz A.3). Der Nachweis der Voraussetzungen für die Anwendung der Formel wird im Anhang (ab Seite 73) geführt.

$$W_2 = \sum_{i=1}^{2}\sum_{k=1}^{n_i} m_{2ik}\widehat{Z}_{ik} \quad \text{mit} \quad m_{2ik} = \begin{cases} \sqrt{\frac{N}{n_2}\frac{n_1}{n_1-1}}\frac{1}{n_1} & i = 1 \\ 0 & i = 2 \end{cases}$$
$$Var(W_2^*|\mathbf{R}) = \left(\sum_{i=1}^{2}\sum_{k=1}^{n_i}(m_{2ik} - \overline{m}_{2..})^2\right)\left(\frac{1}{N-1}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..})^2\right).$$

Dabei ist:

$$\overline{m}_{2..} = \frac{1}{N}\sqrt{\frac{N}{n_2}\frac{n_1}{n_1 - 1}}$$
$$\Rightarrow \quad m_{2ik} - \overline{m}_{2..} = \begin{cases} \sqrt{\frac{N}{n_2}\frac{n_1}{n_1-1}}\left(\frac{1}{n_1} - \frac{1}{N}\right) & i = 1 \\ -\frac{1}{N}\sqrt{\frac{N}{n_2}\frac{n_1}{n_1-1}} & i = 2 \end{cases}$$
$$\Rightarrow \quad \sum_{i=1}^{2}\sum_{k=1}^{n_i}(m_{2ik} - \overline{m}_{2..})^2 = \frac{N}{n_2}\frac{n_1}{n_1 - 1}\left(n_1\left(\frac{1}{n_1} - \frac{1}{N}\right)^2 + n_2\frac{1}{N^2}\right)$$
$$= \frac{N}{n_2}\frac{n_1}{n_1 - 1}\left(\frac{1}{n_1} + \frac{n_1}{N^2} - \frac{2n_1}{n_1 N} + \frac{n_2}{N^2}\right)$$
$$= \frac{N}{n_2}\frac{n_1}{n_1 - 1}\left(\frac{1}{n_1} - \frac{1}{N}\right) = \frac{N}{n_2}\frac{n_1}{n_1 - 1}\frac{n_2}{n_1 N} = \frac{1}{n_1 - 1}.$$

Damit und mit (3.13) ist

$$Var(W_2^*|\mathbf{R}) = \frac{1}{n_1 - 1}\frac{N}{N-1}\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2.$$

Da $\widehat{Z}_{ik} \in [-\frac{1}{2}, \frac{1}{2}]$, ist $\widehat{Z}_{ik}^2$ durch $\frac{1}{4}$ gleichmäßig beschränkt. Deshalb erhalten wir:

$$Var(W_2^*|\mathbf{R}) \leq \frac{1}{n_1 - 1}\frac{N}{N-1}\frac{1}{4} \to 0 \quad \text{für} \quad N \to \infty.$$

Daraus folgt $W_2^* \to 0$ in $\mathrm{P} \otimes \mathrm{P}^*$-Wahrscheinlichkeit. Das gleiche Ergebnis erhalten wir analog für W_3^*.

Für die Herleitung der bedingten Verteilung von W_1^* verwenden wir die folgende Darstellung:

$$W_1 = \sum_{i=1}^{2}\sum_{k=1}^{n_i} d_{ik}\widehat{Z}_{ik}^2 \quad \text{mit} \quad d_{ik} = \begin{cases} \frac{N}{n_2}\frac{1}{n_1-1} & i = 1 \\ \frac{N}{n_1}\frac{1}{n_2-1} & i = 2. \end{cases}$$

Der bedingte Erwartungswert von W_1^* ist

$$\mathrm{E}(W_1^*|\mathbf{R}) = \left(\sum_{i=1}^{2}\sum_{k=1}^{n_i} d_{ik}\right)\left(\frac{1}{N}\sum_{j=1}^{2}\sum_{l=1}^{n_i}\widehat{Z}_{jl}^2\right).$$

Dabei gilt

$$N\bar{d}_{\cdot\cdot} = n_1\frac{N}{n_2}\frac{1}{n_1 - 1} + n_2\frac{N}{n_1}\frac{1}{n_2 - 1} \overset{N\to\infty}{\longrightarrow} \frac{1}{1-\kappa} + \frac{1}{\kappa} = \frac{1}{\kappa(1-\kappa)}.$$

Wie beim Nachweis von Bedinung (3.11) gezeigt wurde, konvergiert

$$\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2 \xrightarrow{\mathrm{P}} \rho.$$

Damit erhalten wir für den Grenzwert der bedingten Erwartung von W_1^*:

$$\mathrm{E}(W_1^*|\mathbf{R}) = \sum_{i=1}^{2}\sum_{k=1}^{n_i} d_{ik}\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2 \xrightarrow{\mathrm{P}} \frac{1}{\kappa(1-\kappa)}\rho =: \widetilde{\rho} \quad \text{für} \quad N \to \infty. \tag{3.16}$$

Außerdem berechnen wir wieder die bedingte Varianz mit der Formel nach Hájek & Šidák (1967), siehe Anhang Satz A.3, deren Anwendbarkeit dort ab Seite 75 nachgewiesen wird.

$$Var(W_1^*|\mathbf{R}) = \left(\sum_{i=1}^{2}\sum_{k=1}^{n_i}(d_{ik} - \bar{d}_{\cdot\cdot})^2\right)\left(\frac{1}{N-1}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left(\widehat{Z}_{ik}^2 - \frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}\widehat{Z}_{ik}^2\right)^2\right).$$

Da Z_{ik}^4 durch $C = 1/2^4$ gleichmäßig beschränkt ist und die Quadratsumme der d_{ik} gegen Null konvergiert:

$$\begin{aligned}
\sum_i \sum_k (d_{ik} - \overline{d}_{..})^2 &= n_1 \left(\frac{N}{n_2} \frac{1}{n_1 - 1} - \left(\frac{1}{n_2} \frac{n_1}{n_1 - 1} + \frac{1}{n_1} \frac{n_2}{n_2 - 1} \right) \right)^2 \\
&\quad + n_2 \left(\frac{N}{n_1} \frac{1}{n_2 - 1} - \left(\frac{1}{n_2} \frac{n_1}{n_1 - 1} + \frac{1}{n_1} \frac{n_2}{n_2 - 1} \right) \right)^2 \\
&= n_1 \left(\frac{N - n_1}{n_2} \left(\frac{1}{n_1 - 1} \right) - \frac{1}{n_1} \frac{n_2}{n_2 - 1} \right)^2 \\
&\quad + n_2 \left(\frac{N - n_2}{n_1} \left(\frac{1}{n_2 - 1} \right) - \frac{1}{n_2} \frac{n_1}{n_1 - 1} \right)^2 \\
&= \frac{n_1}{(n_1 - 1)^2} + \frac{1}{n_1} \left(\frac{n_2}{n_2 - 1} \right)^2 - \frac{2n_2}{(n_1 - 1)(n_2 - 1)} \\
&\quad + \frac{n_2}{(n_2 - 1)^2} + \frac{1}{n_2} \left(\frac{n_1}{n_1 - 1} \right)^2 - \frac{2n_1}{(n_1 - 1)(n_2 - 1)} \\
&\overset{N \to \infty}{\longrightarrow} 0,
\end{aligned}$$

erhalten wir für die bedingte Varianz von W_1^*

$$Var(W_1^* | \mathbf{R}) \to 0 \quad \text{für} \quad N \to \infty.$$

Die Konvergenz von W_1^* gegen $\widetilde{\rho} = \rho/(\kappa(1-\kappa))$ in $\mathrm{P} \otimes \mathrm{P}^*$-Wahrscheinlichkeit folgt aus der Konvergenz der bedingten Erwartung und Varianz, denn für $\varepsilon > 0$ gilt aufgrund der Dreiecksungleichung:

$$\begin{aligned}
\mathrm{P}^*\left(\left| W_1^* - \widetilde{\rho} \right| > \varepsilon \right) &= \mathrm{P}^*\left(\left| W_1^* - \mathrm{E}(W_1^* | \mathbf{R}) + \mathrm{E}(W_1^* | \mathbf{R}) - \widetilde{\rho} \right| > \varepsilon \right) \\
&\leq \mathrm{P}^*\left(\left| W_1^* - \mathrm{E}(W_1^* | \mathbf{R}) \right| > \varepsilon \right) + \mathrm{P}^*\left(\left| \mathrm{E}(W_1^* | \mathbf{R}) - \widetilde{\rho} \right| > \varepsilon \right)
\end{aligned} \tag{3.17}$$

Der zweite Term in (3.17) konvergiert gegen Null, wie in (3.16) gezeigt. Für den

ersten Term erhalten wir mithilfe der Tschebyschow-Ungleichung:

$$\begin{aligned}
&\mathrm{P}^*\left(\left|W_1^* - \mathrm{E}(W_1^*|\mathbf{R})\right| > \varepsilon\right) \\
&\leq \frac{1}{\varepsilon^2}\mathrm{E}\left[(W_1^* - \mathrm{E}(W_1^*|\mathbf{R}))^2\middle|\mathbf{R}\right] \\
&= \frac{1}{\varepsilon^2}\mathrm{E}\left[W_1^{*2} + \mathrm{E}(W_1^*|\mathbf{R})^2 - 2W_1^*\mathrm{E}(W_1^*|\mathbf{R})\middle|\mathbf{R}\right] \\
&= \frac{1}{\varepsilon^2}\left(\mathrm{E}\left[W_1^{*2}\middle|\mathbf{R}\right] + \mathrm{E}\left[\mathrm{E}(W_1^*|\mathbf{R})^2\middle|\mathbf{R}\right] - 2\mathrm{E}\left[W_1^*\mathrm{E}(W_1^*|\mathbf{R})\middle|\mathbf{R}\right]\right) \\
&= \frac{1}{\varepsilon^2}\left(\mathrm{E}\left[W_1^{*2}\middle|\mathbf{R}\right] + \mathrm{E}(W_1^*|\mathbf{R})^2 - 2\mathrm{E}\left[W_1^*|\mathbf{R}\right]\mathrm{E}(W_1^*|\mathbf{R})\right) \\
&= \frac{1}{\varepsilon^2}\left(\mathrm{E}\left[W_1^{*2}\middle|\mathbf{R}\right] - \mathrm{E}\left[W_1^*|\mathbf{R}\right]^2\right) \\
&= \frac{1}{\varepsilon^2}\, Var\left[W_1^{*2}\middle|\mathbf{R}\right] \to 0 \quad \text{für} \quad N \to \infty.
\end{aligned}$$

Damit folgt

$$W_1^* \xrightarrow{\mathrm{P}\otimes\mathrm{P}^*} \widetilde{\rho} = \rho/(\kappa(1-\kappa)).$$

Mit diesen Berechnungen für W_2^*, W_3^* und W_1^* erhalten wir insgesamt für V_N^{*2}

$$V_N^{*2} = W_1^* - W_2^{*2} - W_3^{*2} \xrightarrow{\mathrm{P}\otimes\mathrm{P}^*} \frac{1}{\kappa(1-\kappa)}\rho \quad \text{für} \quad N \to \infty.$$

Unter Berücksichtigung des in Gleichung (3.15) berechneten Grenzwertes, ergibt sich letztendlich für den Limes des Quotienten der beiden Folgen aus Bedingung (3.12):

$$\frac{\frac{1}{N}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\widehat{Z}_{ik} - \overline{\widehat{Z}}_{..})^2}{V_N^{*2}} \xrightarrow{\mathrm{P}\otimes\mathrm{P}^*} \kappa(1-\kappa) = \tau^2 \quad \text{für} \quad N \to \infty.$$

Damit sind alle Bedingungen des Satzes 3.2 erfüllt und wir haben bewiesen, dass die asymptotische Permutationsverteilung der Teststatistik T_N eine Normalverteilung ist. Dadurch erfüllt der auf T_N basierende studentisierte Permutationstest asymptotisch die Invarianzeigenschaft.

3.5 Simulationsstudie

Die Eigenschaften des oben hergeleiteten studentisierten Permutationstests bei kleinen Stichprobenumfängen werden nun mithilfe einer Simulationsstudie analysiert. Niveau und Güte werden für zwei normalverteilte und zwei bimodal-verteilte Stichproben sowie für eine normalverteilte gegen eine χ^2_3-verteilte Stichprobe simuliert. Um die Anwendung auf Scores oder ordinale Daten zu simulieren, werden die Beobachtungen diskretisiert, so dass Bindungen auftreten. Der Einfluss von sehr kleinen und deutlich verschiedenen Stichprobenumfängen wird durch die Verwendung der folgenden Stichprobenumfänge untersucht:

$$(n_1,n_2) \in \{(15,15),(15,7),(7,15),(7,7)\}.$$

Um die Güte des Tests zu untersuchen, wird in jeder der drei Verteilungskonfigurationen bei einer der Verteilungen der Erwartungswert μ um 0.5 bzw. 1 verschoben. Die Einhaltung des Niveaus wird durch Verwendung von $\mu = 0$ überprüft.

1. Zwei normalverteilte Stichproben $\mathcal{N}_{n_1}(0,\, \sigma_1^2)$, $\mathcal{N}_{n_2}(\mu,\, \sigma_2^2)$:
   Für die beiden Stichproben werden verschiedene Varianzen verwendet, wobei die größere Varianz sowohl auf die größere Stichprobe als auch auf die kleinere Stichprobe angewendet wurde:

   $$(\sigma_1^2,\sigma_2^2) \in \{(1,1),(1,2),(1,4)\}.$$

   Um das Verhalten der Tests bei extrem unterschiedlichen Varianzen zu untersuchen, wird für $n_1 = n_2 = 15$ Beobachtungen in beiden Gruppen zusätzlich die Konstellation mit $\sigma_1^2 = 1$ in einer Gruppe und $\sigma_2^2 = 25$ in der anderen Gruppe simuliert. Für Datensätze mit solchen Eigenschaften und noch kleinerem Stichprobenumfang stellt sich die Frage, inwieweit ein Test auf Verschiebungseffekte noch sinnvoll interpretiert werden kann. Trotzdem werden wir die beschrieben Varianzstruktur simulieren, da solche Situationen auch in anderen Arbeiten betrachtet wurden (vgl. Reiczigel *et al.* , 2005). Die Simulationsergebnisse werden in Abschnitt 3.5.1 beschrieben.

2. Zwei bimodal-verteilte Stichproben:
   Die bimodalen Verteilungen werden aus jeweils zwei Normalverteilungen zusammengesetzt. Die verwendeten Werte für die Erwartungswerte, die Varianzen und die Anteile, zu der die beiden Normalverteilungen verwendet werden, stellen sicher, dass $p = \frac{1}{2}$ ist, wenn $\mu = 0$ gesetzt wird (vgl. Abbildung 3.1).

   $$F_1 = \frac{7}{10}\,\mathcal{N}_{n_1}(4,1) + \frac{3}{10}\,\mathcal{N}_{n_1}(8,1),$$
   $$F_2 = \frac{3}{10}\,\mathcal{N}_{n_2}(2.07+\mu,2) + \frac{7}{10}\,\mathcal{N}_{n_2}(3(2.07+\mu),2).$$

Für zwei normalverteilte Zufallsvariablen $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ berechnet sich der relative Effekt als

$$p = \mathrm{P}(X_1 \leq X_2) = \mathrm{P}\left(\frac{X_2 - X_1 - (\mu_2 - \mu_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \leq \frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

wenn Φ die Verteilungsfunktion der Standardnormalverteilung ist. Für die zusammengesetzten Normalverteilungen erhalten wir für die gewählten Parameter dann entsprechend:

$$\begin{aligned} p = \mathrm{P}(X_{11} \leq X_{21}) &= \frac{7}{10}\frac{3}{10}\Phi\left(\frac{2.07-4}{\sqrt{1+2}}\right) + \frac{7}{10}\frac{7}{10}\Phi\left(\frac{3\cdot 2.07-4}{\sqrt{1+2}}\right) \\ &\quad + \frac{3}{10}\frac{3}{10}\Phi\left(\frac{2.07-8}{\sqrt{1+2}}\right) + \frac{3}{10}\frac{7}{10}\Phi\left(\frac{3\cdot 2.07-8}{\sqrt{1+2}}\right) \\ &= 0.500003123. \end{aligned}$$

Die Ergebnisse der Simulationen mit diesen bimodalen Verteilungen sind in Abschnitt 3.5.2 dargestellt.

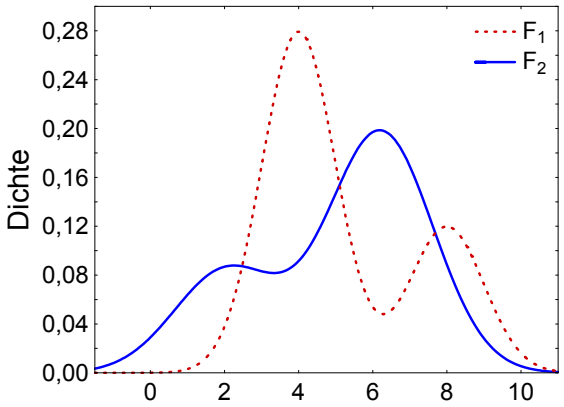


Abbildung 3.1: Dichten zweier bimodaler Verteilungen für die die Hypothese $p = \frac{1}{2}$ erfüllt ist.

3. Eine normalverteilte Stichprobe $F_1 = \mathcal{N}(\mu_0 + \mu, \sigma^2)$ gegen eine χ^2-verteilte Stichprobe $F_2 = \chi_3^2(0)$:
   Numerische Integration ergab, dass bei Verwendung einer Normalverteilung mit Erwartungswert $\mu_0 = 2.5745$ und Varianz $\sigma^2 = 2$ der Wert des relativen Effekts unter der Hypothese mit $p = 0.500013$ recht gut approximiert wird. In Abschnitt 3.5.3 sind die Ergebnisse zu diesen Verteilungen zu finden.

Zum Vergleich werden das Niveau und die Güte außer für den studentisierten Permutationstest auch für die t-Approximation von Brunner & Munzel (2000), den Bootstraptests von Reiczigel *et al.* (2005) sowie den Likelihood-Ratio-Test von Troendle (2002) simuliert.

Da die Anzahl aller möglichen Permutationen (Bootstrap-Stichproben) und damit die Simulationszeiten sehr rasch mit steigendem Stichprobenumfang ansteigen, werden bedingte Monte-Carlo-Simulationen mit jeweils n_{perm} =10'000 zufällige Permutationen (Bootstrap-Stichproben) berechnet. Die Niveau- und Güte-Berechnungen beruhen auf n_{sim} =10'000 Simulationsdurchgängen. Um Aussagen über die Eigenschaften von Adjustierungsverfahren für auftretende Multiplizität treffen zu können, wurden die empirischen Niveaus für 100 Werte von α zwischen 0.1% und 10% berechnet. Insbesondere für Approximationsverfahren wie die t-Approximation ist die Approximation bei kleinen nominalen Niveaus oft ungenau. Aufgrund langer Rechenzeiten konnte der Likelihood-Ratio-Test nur für das 5%-Niveau analysiert werden.

Außerdem werden *Zufallsstreifen* (α_{low}, α_{up}) für das empirische Niveau $\widetilde{\alpha}$ eines simulierten „korrekten“ Tests zum Niveau α berechnet (vgl. Stange, 1970). Die Grenzen werden dabei so gewählt, dass das empirische Niveau $\widetilde{\alpha}$ zu 95% innerhalb der Grenzen liegt:

$$\mathrm{P}(\alpha_{\text{low}} \leq \widetilde{\alpha} \leq \alpha_{\text{up}}) = 0.95,$$

falls des tatsächliche (nominale) Niveau des simulierten Tests α ist. Der Zufallsstreifen ist das Gegenstück eines Konfidenzintervalls, da hier bei bekanntem wahren Wert des Parameters (nominales Niveau α) ein Bereich für den Schätzer (empirisches Niveau $\widetilde{\alpha}$) angegeben wird. Der Schätzer $\widetilde{\alpha}$, das heißt das empirische Niveau, ist binomialverteilt. Über die Approximation der Binomialverteilung durch die F-Verteilung werden die Grenzen (α_{low}, α_{up}) wie folgt berechnet

$$\alpha_{\text{low}} = \frac{\alpha n_{\text{sim}} - (1-\alpha)F_{1-\alpha/2}}{\alpha + (1-\alpha)F_{1-\alpha/2}} \quad \text{und} \quad \alpha_{\text{up}} = \frac{\alpha(n_{\text{sim}}+1)}{\alpha + (1-\alpha)F_{\alpha/2}},$$

wobei $F_{1-\alpha/2}$ und $F_{\alpha/2}$ Quantile der F-Verteilung sind. Konkret ist

$$\begin{aligned} F_{1-\alpha/2} &= F_{1-\alpha/2}(f_1', f_2') \quad \text{mit} \quad f_1' = 2(\alpha n_{\text{sim}} + 1), \;\; f_2' = 2(1-\alpha)n_{\text{sim}}, \\ F_{\alpha/2} &= F_{\alpha/2}(f_1'', f_2'') \quad \text{mit} \quad f_1'' = 2\alpha n_{\text{sim}}, \;\; f_2'' = 2((1-\alpha)n_{\text{sim}} + 1). \end{aligned}$$

Wenn der Varianzschätzer V_N^2 Null wird, das heißt wenn alle Beobachtungen in einer Gruppe kleinere Ränge haben als die Beobachtungen in der anderen Gruppe, wird er durch eine geeignete untere Grenze der Varianz ersetzt, die gerade größer als Null ist. Dadurch wird die Testentscheidung des Permutationstests und der t-Approximation in diesem Fall konservativ. Als Wert für die untere Grenze des Varianzschätzers verwenden wir (vgl. Abschnitt 3.1):

$$V_{N,\min}^2 = \frac{N}{2n_1n_2}.$$

Die Simulationen für den Permutationstest, den Bootstraptest und die t-Approximation wurden mit SAS IML 9.1 durchgeführt. Für den Likelihood-Ratio-Test erhielten wir freundlicherweise von Dr. Troendle sein FORTRAN 77 Simulationsprogramm. Alle Tests wurden als zweiseitige Tests durchgeführt.

In den folgenden Abschnitten werden für jede Verteilungskonfiguration die Simulationsergebnisse vorgestellt. Die Fähigkeit der Tests das nominale Niveau einzuhalten wird durch Graphiken illustriert, auf deren x-Achse das nominale und auf deren y-Achse das empirische Niveau dargestellt ist (vgl. Abbildung 3.2). Die Grenzen des Zufallsstreifens verlaufen über bzw. unterhalb der Diagonale, so dass der Zufallsstreifen gerade der grau schraffierten Fläche in Abbildung 3.2 entspricht. Wir sprechen von einem konservativen Test, wenn sein empirisches Niveau kleiner ist als das nominale Niveau, seine Niveau-Kurve also unterhalb der unteren Grenze des Zufallsstreifens liegt. Das empirische Niveau eines liberalen Tests übersteigt das nominale Niveau, seine Niveau-Kurve verläuft entsprechend oberhalb der Diagonale bzw. der oberen Zufallsstreifengrenze.

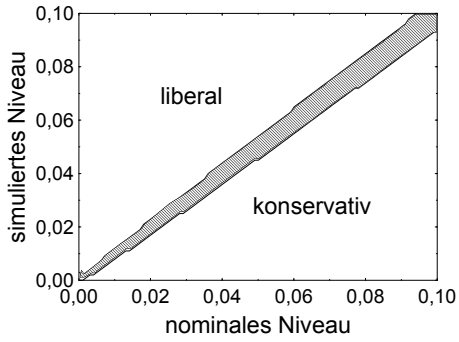
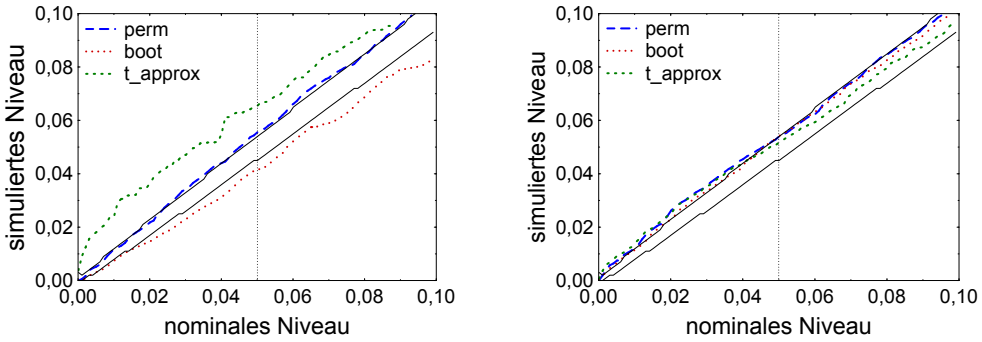


Abbildung 3.2: Niveausimulationen: Schematische Darstellung und Interpretation

3.5.1 Zwei normalverteilte Stichproben

Betrachtet man zwei Normalverteilungen mit gleichen Varianzen (ohne Abbildung), so liegen bei jeweils 15 Beobachtungen pro Gruppe die Kurven der Tests entweder auf der oberen (t-Approximation) oder der unteren (Bootstraptest) Grenze des Zufallsstreifens oder innerhalb dieser Grenzen (Permutationstest), das heißt die Tests erzielen alle vergleichbare Niveaus. Wird der Stichprobenumfang nur in einer Stichprobe auf 7 Beobachtungen reduziert, so wird die t-Approximation etwas liberal, während sich die Kurven der anderen beiden Tests kaum verändern. Bestehen beide Gruppen nur aus 7 Beobachtungen (vgl. Abbildung 3.3(a)), dann verläuft die Niveaukurve des Permutationstests entlang der oberen Zufallsstreifengrenze, die des

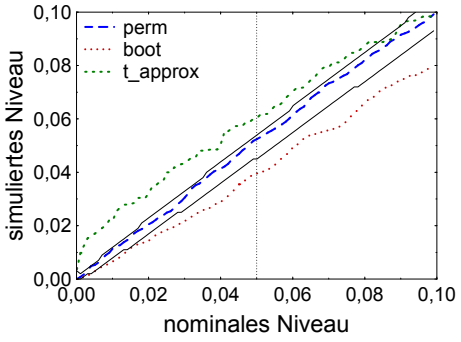
Bootstraptests leicht unterhalb der unteren Grenze und die t-Approximation ist liberal. Das empirische Niveau des Likelihood-Ratio-Tests liegt für alle betrachteten Stichprobenumfänge bei einem nominalem Niveau von 5% innerhalb der Grenzen des Zufallsstreifens.



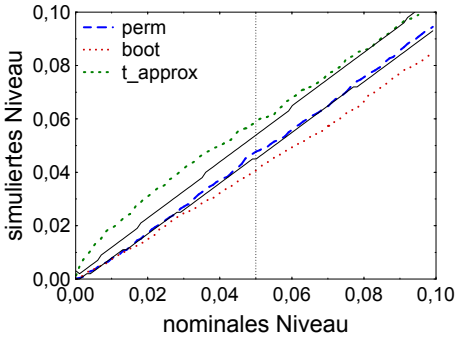
(a) $\sigma_1^2 = \sigma_2^2 = 1$, $n_1 = n_2 = 7$ (b) $\sigma_1^2 = 1$, $\sigma_2^2 = 25$, $n_1 = n_2 = 15$

Abbildung 3.3: Zwei Normalverteilungen. Nominales gegen empirisches Niveau, durchgezogene Linien stellen die Zufallsstreifen dar.

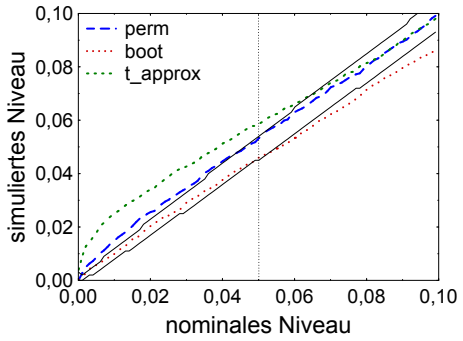
Ist die Varianz in der zweiten Stichprobe 2 und der Stichprobenumfang in beiden Gruppen gleich 7 (vgl. Abbildung 3.4(a)), dann liegt die Kurve des Permutationstests für alle nominalen Niveaus innerhalb des Zufallsstreifens. Der Bootstraptest wird dagegen leicht konservativ. Bei verschiedenen Stichprobenumfängen in den beiden Gruppen verläuft die Kurve des Permutationstests nahe an der unteren (Abbildung 3.4(b)) bzw. oberen (Abbildung 3.4(c)) Grenze des Zufallsstreifens. Tritt die größere Varianz in der größeren Stichprobe auf, so ist der Bootstraptest etwas konservativ (Abbildung 3.4(b)). Im entgegengesetzten Fall liegt seine Kurve für nominale Niveaus bis 5% innerhalb des Zufallsstreifens und für größere Werte knapp unter der unteren Zufallsstreifengrenze (Abbildung 3.4(c)). Die t-Approximation ist in allen drei Situationen leicht liberal für nominale Niveaus bis 6% und erreicht dann empirische Niveaus innerhalb bzw. sehr nahe am Bereich des Zufallsstreifens. In Tabelle 3.4(d) werden das Niveau und die Güte aller vier Tests verglichen. Dabei markieren fettgedruckte Zahlen Niveaus, die innerhalb des Zufallsstreifens liegen. Die angegebenen Werte sind für $n_1 = 15$ und $n_2 = 7$ berechnet, so dass die größere Varianz in der kleineren Stichprobe auftritt. Hier liegen die Niveaus des Permutations- und Bootstraptests innerhalb des Zufallsstreifens und alle vier Tests haben eine vergleichbare Güte.



(a) $n_1 = n_2 = 7$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$



(b) $n_1 = 7$, $n_2 = 15$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$



(c) $n_1 = 15$, $n_2 = 7$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$

Test	$\mu = 0$	$\mu = 0.5$	$\mu = 1$
Φ_{Perm}	**0.053**	0.125	0.343
Φ_{Boot}	**0.046**	0.108	0.305
$\Phi_{t\text{-approx}}$	0.059	0.132	0.349
Φ_{LR}	0.056	0.126	0.327

(d) $n_1 = 15$, $n_2 = 7$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$

Abbildung 3.4: Zwei Normalverteilungen. (a)-(c): Nominales gegen empirisches Niveau, durchgezogene Linien stellen die Zufallsstreifen dar. (d): Simuliertes Niveau und Güte für $\alpha = 5\%$, fettgedruckte Zahlen markieren Werte innerhalb des Zufallsstreifens.

Bei Varianz $\sigma_2^2 = 4$ (ohne Abbildung) verstärken sich die beobachteten Effekte, wobei der Bootstraptest die geringsten Veränderungen zeigt. Um das Verhalten der Tests auch unter extrem unterschiedlichen Varianzen zu untersuchen, haben wir außerdem Varianz $\sigma_2^2 = 25$ in der einen gegenüber Varianz $\sigma_1^2 = 1$ in der anderen Gruppe simuliert. Bei so großen Varianzunterschieden halten wir die Interpretation eines Shifteffekts für schwierig, wenn die Stichprobenumfänge sehr klein sind. Aus diesem Grund betrachten wir nur Stichprobenumfang 15 in beiden Gruppen. Abbildung 3.3(b) zeigt, dass in diesem Fall die Kurven aller drei Tests entlang der oberen Grenze des Zufallsstreifens verlaufen. Für nominale Niveaus über 5% liegt die Kurve des Bootstraptests dann innerhalb des Zufallsstreifens. Auch der Likelihood-Ratio-Test hält das nominale 5% Niveau innerhalb der Zufallsstreifengrenzen ein.

Der Einfluss von verschiedenen Varianzen bei $n_1 = 7$ und $n_2 = 15$ auf die Güte

wird in Abbildung 3.5 dargestellt. Wie zu erwarten nimmt die Güte mit steigender Varianz ab. Außerdem ist die Güte geringer, wenn die größere Varianz in der kleineren Stichprobe auftritt, als im umgekehrten Fall. Hat die größere Stichprobe die größere Varianz ($\sigma_1^2 : \sigma_2^2 \in \{1:4, 1:2\}$), so erreicht der Permutationstest die gleiche Güte wie der Bootstraptest, der als einziger anderer Test auch das Niveau einhält. In der entgegengesetzten Situation ($\sigma_1^2 : \sigma_2^2 \in \{2:1, 4:1\}$) ist die Güte des Permutationstests mit der Güte der liberalen t-Approximation vergleichbar und liegt über der Güte des Bootstraptests.

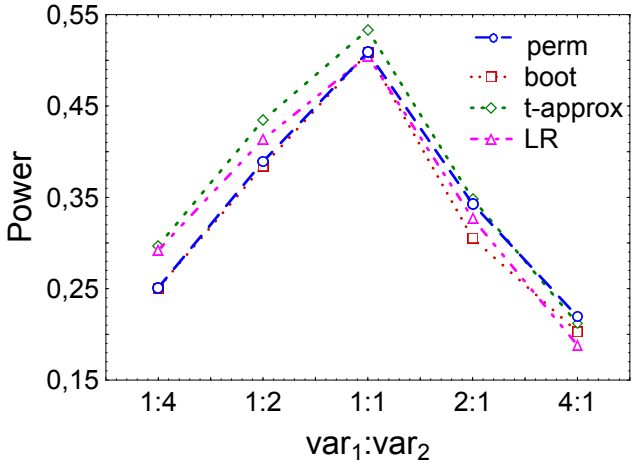
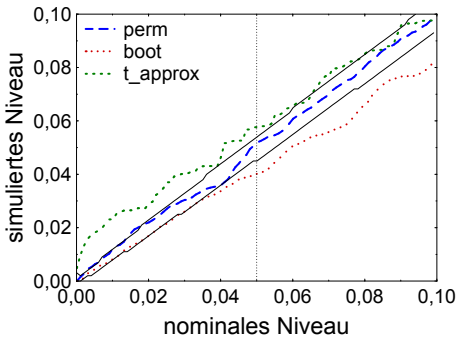


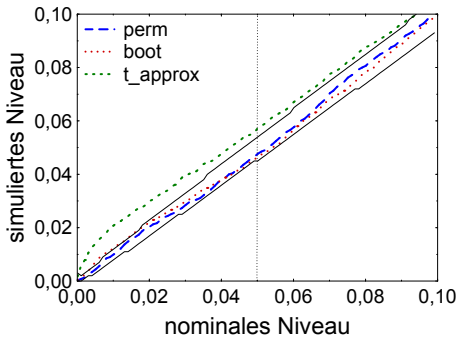
Abbildung 3.5: Güte-Verläufe für verschiedene Varianzen bei zwei Normalverteilungen mit $n_1 = 7$, $n_2 = 15$, $\alpha = 5\%$, $\mu = 1$.

3.5.2 Zwei bimodal-verteilte Stichproben

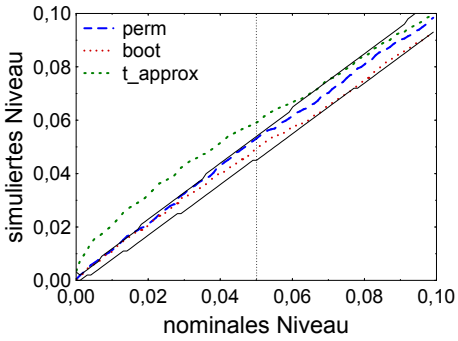
Haben die Verteilungen sehr unterschiedliche Formen, wie die beschriebenen bimodalen Verteilungen, so kann man Abbildung 3.6 entnehmen, dass die t-Approximation und der Permutationstest davon wenig beeinflusst werden und sich sehr ähnlich verhalten, wie bei zugrunde liegenden Normalverteilungen. Die Niveau-Kurve des Bootstraptests sieht bei 7 Beobachtungen pro Gruppe auch sehr ähnlich aus wie im normalverteilten Fall (Abbildung 3.6(a)). Für Stichprobenumfang 7 in der einen und 15 in der anderen Gruppe dagegen verlaufen seine Kurven nun innerhalb des Zufallsstreifens (vgl. Abbildung 3.6(b) + (c)). Die Werte in Tabelle 3.6(d) sind für $n_1 = 15$ und $n_2 = 7$ berechnet. Unter H_0 sind der Likelihood-Ratio-Test und die t-Approximation etwas liberal. Der Permutations- und der Bootstraptest haben sehr ähnliche Gütekurven unter beiden simulierten Alternativen. Der Likelihood-Ratio-Test hat trotz seiner Liberalität die geringste Güte.



(a) $n_1 = n_2 = 7$



(b) $n_1 = 7$, $n_2 = 15$



(c) $n_1 = 15$, $n_2 = 7$

Test	$\mu = 0$	$\mu = 0.5$	$\mu = 1$
Φ_{Perm}	**0.053**	0.129	0.288
Φ_{Boot}	**0.049**	0.122	0.269
$\Phi_{t\text{-approx}}$	0.059	0.134	0.274
Φ_{LR}	0.062	0.111	0.221

(d) $n_1 = 15$, $n_2 = 7$

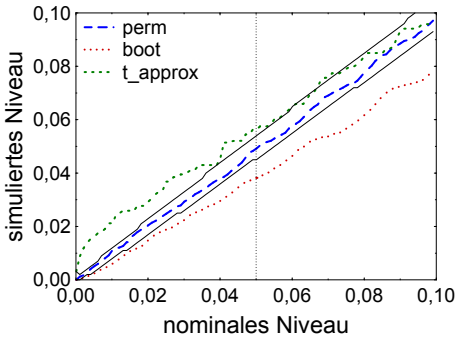
Abbildung 3.6: Zwei bimodale Verteilungen. (a)-(c): Nominales gegen empirisches Niveau, durchgezogene Linien stellen die Zufallsstreifen dar. (d): Simuliertes Niveau und Güte für $\alpha = 5\%$, fettgedruckte Zahlen markieren Werte innerhalb des Zufallsstreifens.

3.5.3 Eine normalverteilte vs. eine χ^2_3-verteilte Stichprobe

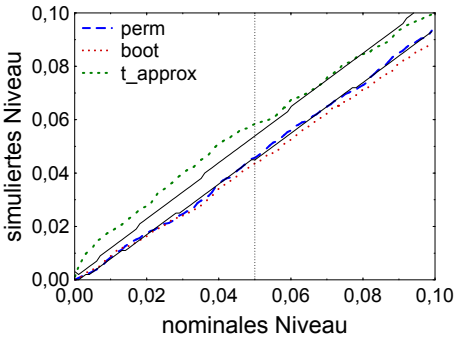
Betrachten wir nun eine symmetrische und eine schiefe Verteilung, z.B. eine normalverteilte gegen eine χ^2_3-verteilte Stichprobe. Die Eigenschaften der t-Approximation und des Permutationstests ändern sich dadurch wiederum kaum im Vergleich zu zwei normalverteilten Stichproben (Abbildung 3.7). Bei 7 Beobachtungen in beiden Gruppen gilt das auch für den Bootstraptest (Abbildung 3.7(a)). Bei 7 Beobachtungen in der normalverteilten und 15 Beobachtungen in der χ^2_3-verteilten Stichprobe verläuft seine Niveau-Kurve entlang der unteren Grenze des Zufallsstreifens. Im umgekehrten Fall liegt sein empirisches Niveau bis zu einem nominalen Niveau von 6% innerhalb des Zufallsstreifens und danach knapp darunter.

Tabelle 3.7(d) entnimmt man, dass für $n_1 = 15$ und $n_2 = 7$ nur die t-Approxi-

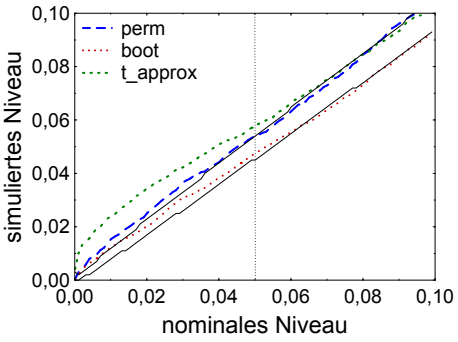
mation etwas liberal ist. Alle vier Tests haben hier vergleichbare Güte.



(a) $n_1 = n_2 = 7$



(b) $n_1 = 7$, $n_2 = 15$



(c) $n_1 = 15$, $n_2 = 7$

Test	$\mu = 0$	$\mu = 0.5$	$\mu = 1$
Φ_{Perm}	**0.054**	0.090	0.175
Φ_{Boot}	**0.047**	0.083	0.163
$\Phi_{t\text{-approx}}$	0.058	0.094	0.175
Φ_{LR}	**0.052**	0.085	0.159

(d) $n_1 = 15$, $n_2 = 7$

Abbildung 3.7: Normalverteilung vs. χ^2_3-Verteilung. (a)-(c): Nominales gegen empirisches Niveau, durchgezogene Linien stellen die Zufallsstreifen dar. (d): Simuliertes Niveau und Güte für $\alpha = 5\%$, fettgedruckte Zahlen markieren Werte innerhalb des Zufallsstreifens.

3.6 Anwendungen

3.6.1 Ferritin-Studie

In dieser Studie an Kindern mit hormonell bedingtem Kleinwuchs wurde der Zusammenhang zwischen einer verminderten Synthese des *insulin-like-growth-factor (IGF-1)* und einer Erhöhung der Ferritin-Werte untersucht. Die Daten wurden uns freundlicherweise von Herrn Prof. Lakomek der Abteilung Kinderheilkunde der Universität Göttingen zur Verfügung gestellt. Die Gruppe der Kinder mit vermindertem IGF-1-Wert besteht aus 12 Kindern, während 7 Kinder einen IGF-1-Wert im

altersbedingten Normalbereich hatten. In Tabelle 3.2 sind für beide Gruppen die Ferritin-Werte in ng/ml angegeben.

Tabelle 3.2: Daten der Ferritin-Studie.

IGF-1	Ferritin [ng/ml]
erniedrigt	1956, 8828, 2051, 3721, 3233, 6606, 2244 5332, 5428, 2603, 2370, 7565
normal	820, 3364, 1497, 1851, 2984, 744, 2044

Anhand des Boxplots (Abbildung 3.8) sieht man, dass die Verteilung schief ist und wir nicht von einer Normalverteilung der Daten ausgehen können.

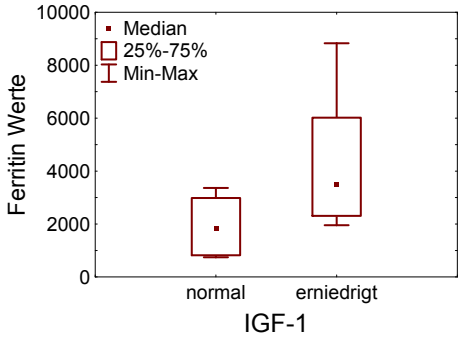


Abbildung 3.8: Boxplot der Ferritinwerte nach IGF-1 Gruppe

Der Rangmittelwert in der Gruppe mit erniedrigtem IGF-1-Wert liegt bei $\overline{R}_{1\cdot} = 12.5$ und in der Gruppe mit normalem IGF-1-Wert bei $\overline{R}_{2\cdot} = 5.7$. Die Varianzschätzer der Ränge haben den Wert $\widehat{\sigma}_1^2 = 1.27$ (erniedrigt) bzw. $\widehat{\sigma}_2^2 = 6.90$ (normal), so dass die Gruppen als heteroskedastisch anzusehen sind. Der Schätzer für den relativen Effekt ist $\widehat{p} = 0.143$. Für den Permutationstest wurden wieder 10'000 Permutationen und für den Bootstraptest 10'000 Bootstrap-Stichproben berechnet. Alle 4 Tests lehnen die Hypothese $H_0 : \ p = \frac{1}{2}$ auf dem 5% Niveau ab (vgl. Tabelle 3.3).

Tabelle 3.3: p-Werte für die Ferritin-Studie.

Φ_{Perm}	Φ_{Boot}	$\Phi_{t\text{-approx}}$	Φ_{LR}
0.0092	0.0001	0.0038	0.010

3.6.2 Schulter-Schmerz-Studie

Hier werden die Daten der in Lumley (1996) beschriebenen Schulter-Schmerz-Studie ausgewertet. In dieser Studie wurde an 6 Zeitpunkten nach der Durchführung einer Operation ein Schmerz-Score bei 25 weiblichen Patienten erhoben. Dieser Score kann Werte zwischen 1 (kaum Schmerzen) und 5 (große Schmerzen) annehmen. Die Patienten wurden zufällig zwei Behandlungsgruppen zugeteilt, wobei 14 Patienten die aktive Behandlung (Y) und 11 Patienten die Kontrollbehandlung (N) erhielten. Die Daten sind in Tabelle 3.4 aufgeführt. Wir wollen hier die Frage beantworten, ob und ab welchem Zeitpunkt sich die Werte der Behandlungs- und Kontrollgruppe unterschieden.

Die p-Werte in Tabelle 3.5 sind nach Bonferroni für die Multiplizität adjustiert, die durch die Durchführung der Tests zu allen 6 Zeitpunkten entsteht. Dadurch werden p-Werte größer als 0.1667 auf 1 gesetzt. Die Varianzschätzer unterscheiden sich in den beiden Gruppen recht stark, wobei die Varianz in der Behandlungsgruppe ($\widehat{\sigma}_1^2$) deutlich geringer ist als in der Kontrollgruppe ($\widehat{\sigma}_2^2$).

Tabelle 3.4: Daten der Schulter-Schmerz-Studie.

Behandlung (Y)							Kontrolle (N)						
	Zeitpunkt							Zeitpunkt					
Pat.	1	2	3	4	5	6	Pat.	1	2	3	4	5	6
1	1	1	1	1	1	1	23	5	2	3	5	5	4
3	3	2	2	2	1	1	24	1	5	3	4	5	3
4	1	1	1	1	1	1	25	4	4	4	4	1	1
5	1	1	1	1	1	1	28	3	4	3	3	3	2
8	2	2	1	1	1	1	30	1	1	1	1	1	1
9	1	1	1	1	1	1	33	1	3	2	2	1	1
10	3	1	1	1	1	1	34	2	2	3	4	2	2
12	2	1	1	1	1	2	35	2	2	1	3	3	2
16	1	1	1	1	1	1	36	1	1	1	1	1	1
18	2	1	1	1	1	1	38	5	5	5	4	3	3
19	4	4	2	4	2	2	40	5	4	4	4	2	2
20	4	4	4	2	1	1							
21	1	1	1	2	1	1							
22	1	1	1	2	1	2							

Alle 4 Tests zeigen, dass die Behandlung (Y) zu signifikant niedrigeren Schmerz-Scores zu den Zeitpunkten 2 und 4 führt. Zu den Zeitpunkten 3 und 5 findet der Permutationstest, der Bootstraptest und die t-Approximation einen signifikanten Behandlungseffekt, nur der Likelihood-Ratio-Test nicht. Zu Zeitpunkt 1 ist noch

kein signifikanter Unterschied festzustellen und zu Zeitpunkt 6 deckt nur der Bootstraptest einen signifikanten Einfluss der Behandlung auf.

Tabelle 3.5: Adjustierte p-Werte für die Schulter-Schmerz-Studie. Signifikante Werte sind fett gedruckt.

Zeitpunkt	relativer Effekt	$\widehat{\sigma}_Y^2$	$\widehat{\sigma}_N^2$	Φ_{Perm}	Φ_{Boot}	$\Phi_{t\text{-approx}}$	Φ_{LR}
1	0.630	5.2	21.6	1	1	1	1
2	0.792	5.7	10.3	**0.0444**	**0.0360**	**0.0216**	**0.0462**
3	0.789	4.2	12.9	**0.0450**	**0.0114**	**0.0348**	0.0522
4	0.828	3.1	12.9	**0.0258**	**<0.0001**	**0.0114**	**0.0282**
5	0.799	0.6	13.8	**0.0246**	**0.0042**	**0.0228**	0.2604
6	0.740	2.9	15.3	0.1224	**0.0324**	0.1308	0.1476

3.7 Zusammenfassung

Der hier vorgestellte studentisierte Permutationstest, basierend auf der Zwei-Stichproben-Rangstatistik von Brunner & Munzel (2000), scheint eine robuste und weithin anwendbare Methode zu sein. Unsere Simulationsstudie zeigt, dass dieser Permutationstest unempfindlich ist bezüglich nicht symmetrischer Verteilungen, z.B. wenn die zugrunde liegenden Verteilungen schief oder bimodal sind. Dabei hält der Permutationstest das Niveau unter der Hypothese recht gut ein und erreicht eine hohe Güte unter der Alternative, selbst wenn die Stichprobenumfänge sehr klein (bis 7 pro Gruppe) oder sehr verschieden sind (z.B. $n_1 = 7$, $n_2 = 15$). Andere Methoden, die für kleine Stichprobenumfänge empfohlen werden, haben vergleichbare Eigenschaften, zeigen aber Tendenzen zu liberalem (t-Approximation) oder konservativem (Bootstraptest) Verhalten, insbesondere bei kleinen Stichprobenumfängen. Diese Effekte könnten durch die Verwendung verschiedener Varianzschätzer entstehen. So ist der für die t-Approximation verwendete Varianzschätzer eine Approximation, die vom parametrischen Fall übernommen wurde. Der Bootstraptest von Reiczigel *et al.* (2005) andererseits basiert auf der Welch-Statistik und verwendet den entsprechenden Varianzschätzer. Besonders problematisch ist die Verwendung von Adjustierungsverfahren bei Multiplizität mit der t-Approximation. Ihre empirischen Niveaus weisen gerade bei kleinen nominalen Niveaus recht große Abweichungen auf. Niveau und Güte des Likelihood-Ratio-Tests sind vergleichbar mit den Werten des Permutationstests. Für bimodale Verteilungen ist der Likelihood-Ratio-Test allerdings etwas liberal und bei Stichprobenumfang sieben in beiden Gruppen war der Maximum-Likelihood-Schätzer in bis zu 173 von den 10'000 Simlationsdurchgängen nicht berechenbar.

Für große Stichprobenumfänge konvergiert die Permutationsverteilung des stu-

dentisierten Permutationstests gegen eine Normalverteilung. Damit erfüllt der Permutationstest asymptotisch die Invarianzeigenschaft und ist für großes N ein gültiger Permutationstest.

Da bezüglich der Verteilungen der Stichproben nur triviale Annahmen nötig sind, ist dieser Permutationstest auch auf ordinale Daten und Scores anwendbar. Zur Durchführung des Tests wurde ein SAS-IML Makro erstellt, dessen Verwendung im Anhang B näher erläutert wird. Das Makro kann auf der Web-Seite der Abteilung Medizinische Statistik der Universität Göttingen heruntergeladen werden (PERM_BF, 2006).

4 Robuste Konfidenzintervalle für den Shift-Effekt in heteroskedastischen Stichproben

Viele der gebräuchlichen statistischen Methoden setzen voraus, dass die zu analysierenden Effekte rein additiv sind. Das entspricht in vielen Situationen auch der Vorstellung des Anwenders der Modelle. Eine besondere Stellung unter den additiven Modellen nimmt das *Lokationsmodell* ein. Im Zwei-Stichprobenfall betrachten wir n_1 Beobachtungen X_{1k}, $k = 1, \ldots, n_1$ in der ersten Gruppe und n_2 Beobachtungen X_{2k}, $k = 1, \ldots, n_2$ in der zweiten Gruppe. Sei

$$X_{1k} \overset{\text{u.i.v.}}{\sim} F_1,\ k = 1, \ldots, n_1, \quad X_{2l} \overset{\text{u.i.v.}}{\sim} F_2,\ l = 1, \ldots, n_2.$$

Gilt dann für eine Verteilungsfunktion F

$$F(x) = F_1(x) \quad \text{und} \quad F(x - \theta) = F_2(x), \quad x, \theta \in \mathbb{R},$$

so liegt ein Lokationsmodell vor. Wenn wie hier vorausgesetzt wird, dass sich die Verteilungsfunktionen F_i ausschließlich bezüglich einer *Verschiebung (Shift)* θ unterscheiden, bezeichnen wir das Modell als *reines Lokationsmodell*. Die parametrische Hypothese für den Shift-Effekt $H_0 : \ \theta = 0$ ist im reinen Lokationsmodell äquivalent zur nichtparametrischen Hypothese $H_0 : \ F_1 = F_2$. Bei stetiger Verteilungsfunktion F ist der Hodges-Lehmann-Schätzer (siehe Abschnitt 4.1) ein robuster und asymptotisch erwartungstreuer Schätzer für den Shifteffekt θ (Hodges & Lehmann, 1963).

Außer der Angabe des Punktschätzers für den vorliegenden Shifteffekt benötigt man zu dessen Bewertung noch weitere Informationen, zum Beispiel über die Varianz und Größe der betrachteten Stichproben. Aus diesem Grund fordern die Zulassungsbehörden für Medizinische Produkte wie beispielsweise die "European Agency for the Evaluation of Medical Products" (EMEA) in ihren Guidelines und PtC (Points to Consider) für klinische Studien die Angabe von Konfidenzintervallen für die geschätzten Effekte: „Estimates of treatment effects should be accompanied by confidence intervals, whenever possible (...)" (EMEA, 1998).

Da wir insbesondere an Methoden interessiert sind, die für kleine Stichprobenumfänge geeignet sind, betrachten wir nur Verfahren, die keine spezielle Verteilung

der Daten voraussetzen. Ein solches Konfidenzintervall für den Shift-Effekt wurde von Lehmann (1963) vorgestellt. Die Konstruktion beruht auf dem Zusammenhang der Verteilungen des Hodges-Lehmann-Schätzers für den Shift-Effekt und der Wilcoxon-Rangsumme

$$R_{2\cdot} = \sum_{k=1}^{n_2} R_{2k}, \quad \text{wobei} \quad R_{2k} = \text{Rang}\{X_{2k} | X_{ik},\, i = 1{,}2,\, k = 1, \ldots, n_i\},$$

der unter bestimmten Annahmen besteht (vgl. Abschnitt 4.2). Die daraus resultierenden Konfidenzintervalle werden wir als *Hodges-Lehmann-Konfidenzintervalle* bezeichnen. Eine andere Konstruktionsmethode für ein robustes Konfidenzintervall für den Shift-Effekt θ wurde von Bauer (1972) beschrieben. Dieses Verfahren beruht auf einer linearen Rangstatistik und wird in Abschnitt 4.3 näher erläutert. Diese beiden Verfahren setzen ein reines Lokationsmodell voraus.

Wir wollen diese restriktive Annahme etwas erweitern und unterschiedliche Varianzen in den beiden Gruppen zulassen. Dieses Modell nennen wir *Lokations-Skalen-Modell*, wobei für eine Verteilungsfunktion F folgendes gelte:

$$F_i(x) = F\left(\frac{x - \mu_i}{\sigma_i}\right), \quad i = 1{,}2,\, x \in \mathbb{R},$$

wenn μ_i der Erwartungswert und σ_i^2 die Varianz in der i-ten Stichprobe ist. Zur Konstruktion eines Konfidenzintervalls für den Shift-Effekt $\theta = \mu_2 - \mu_1$ in diesem Modell erweitern wir die Konstruktionsmethode von Bauer (1972). Dies ist möglich, weil das Verfahren von Bauer allgemein für lineare Rangstatistiken formuliert ist. Durch die zugelassene Heteroskedastizität der Gruppen haben wir im nichtparametrischen Modell wiederum ein Behrens-Fisher-Design. Wir betrachten deshalb die nichtparametrische Hypothese $H_0 : \; p = \frac{1}{2}$, die in der klassischen nichtparametrischen Hypothese für reine Lokationsmodelle $H_0 : \; F_1 = F_2$ enthalten ist. Zur parametrischen Hypothese $H_0 : \; \theta = 0$ ist die Hypothese $H_0 : \; p = \frac{1}{2}$ äquivalent, wenn die Verteilungsfunktionen F_i symmetrisch, stetig und in μ_i invertierbar sind (vgl. Abschnitt 2.2). Unter Verwendung der Teststatistik T_N (vgl. Gleichung 3.2) können wir dann das Konfidenzintervall nach dem Vorschlag von Bauer (1972) berechnen. Für die dafür benötigten Quantile der Teststatistik schlagen wir hier unter anderem vor, die Quantile der Permutationsverteilung von T_N zu benutzen. Dass in einer Behrens-Fisher-Situation ein asymptotisch gültiger Permutationstest durchgeführt werden kann, wurde in Abschnitt 3.3 durch einen Grenzwertsatz für die Permutationsverteilung der Teststatistik T_N bewiesen. Das dieser Test auch für kleine Stichprobenumfänge gute Eigenschaften hat, zeigen die Simulationsergebnisse in Abschnitt 3.5. Dies rechtfertigt die Verwendung der Permutationsquantile zur Berechnung des Konfidenzintervalls. Die Simulationsstudie zur Überdeckungswahrscheinlichkeit aller vorgestellten Konfidenzintervalle zeigt, dass das Bauer-Konfidenzintervall mit Permutationsverteilungsquantilen die besten Ergebnisse erreicht.

Dadurch, dass wir für unser Konfidenzintervall unterschiedliche Varianzen in den beiden Gruppen zulassen können, erweitern sich auch die Anwendungsmöglichkeiten. Eine Anwendung, auf die wir speziell eingehen werden und die in der Praxis häufig auftritt, sind 2×2-Split-Plot-Designs. Ein solches Design liegt beispielsweise vor, wenn die Veränderung einer Variable gegenüber einer Baseline-Messung dieser Variable betrachtet werden soll. Ist X_{ik1} die Baseline-Messung des k-ten Patienten aus Gruppe i und X_{ik2} die Messung des gleichen Patienten nach der Behandlung, so betrachten wir die Differenz gegenüber Baseline $D_{ik} = X_{ik2} - X_{ik1}$. Die Differenzen D_{ik} wollen wir dann auf einen Shift-Effekt zwischen den Gruppen ($i = 1{,}2$) untersuchen. Die Annahmen an die Abhängigkeitsstruktur und die Verteilungsfunktionen werden in Abschnitt 4.6 diskutiert.

Zunächst wird nun im nächsten Abschnitt (4.1) das zugrunde liegende Modell beschrieben. Danach stellen wir für das reine Lokationsmodell die klassischen Konfidenzintervalle nach Lehmann (1963) (Abschnitt 4.2) und die Intervalle nach Bauer (1972) (Abschnitt 4.3) vor. Für möglicherweise heteroskedastische Gruppen schlagen wir in Abschnitt 4.4 drei Intervalle mit verschiedenen Quantilen vor, die auf der Konstruktion nach Bauer beruhen und demonstrieren die Berechnung aller vorgestellten Intervalle in Abschnitt 4.5. Die Anwendung auf das 2×2-Split-Plot-Design diskutieren wir in Abschnitt 4.6. Eine Simulationsstudie (Abschnitt 4.7) und zwei Anwendungen (Abschnitt 4.8) vergleichen die Eigenschaften aller betrachteten Intervalle.

4.1 Modell

Wir betrachten zwei Gruppen mit insgesamt $N = n_1 + n_2$ unabhängig verteilten Zufallsvariablen. Dabei sei

$$X_{1k} \overset{\text{u.i.v.}}{\sim} F_1,\ k = 1, \ldots, n_1, \quad X_{2k} \overset{\text{u.i.v.}}{\sim} F_2,\ k = 1, \ldots, n_2.$$

Gilt für F_1, F_2 ein reines Lokationsmodell, so ist

$$F_1(x) = F(x - \mu_1) \quad \text{und} \quad F_2(x) = F(x - \mu_2), \quad x \in \mathbb{R} \tag{4.1}$$

für eine Verteilungsfunktion F, wobei μ_i der Erwartungswert der Verteilungsfunktion F_i sei. Im Lokations-Skalen-Modell sind außerdem verschieden Varianzen zugelassen, das heißt:

$$F_1(x) = F\left(\frac{x - \mu_1}{\sigma_1}\right) \quad \text{und} \quad F_2(x) = F\left(\frac{x - \mu_2}{\sigma_2}\right), \tag{4.2}$$

wenn σ_i^2 die Varianz der Verteilungsfunktion F_i ist.

Für beide Konstruktionsmethoden (Lehmann, Bauer) müssen wir voraussetzen, dass die Verteilungsfunktionen F_i stetig sind, also Bindungen mit Wahrscheinlichkeit Null auftreten.

Für die Anwendung der Konstruktionsmethode nach Bauer im Lokations-Skalen-Modell müssen wir für die Verteilungsfunktionen F_i außerdem voraussetzen, dass sie symmetrisch und am Erwartungswert μ_i invertierbar sind. Diese Voraussetzungen sind für dieses Modell notwendig, um die Äquivalenz der Hypothesen $H_0: \ \theta = 0$ und $H_0: \ p = \frac{1}{2}$ zu gewährleisten (vgl. Abschnitt 2.2).

Bildet man nun die paarweisen Differenzen zwischen den beiden unabhängigen Gruppen

$$\Delta_{kl} = X_{2k} - X_{1l}, \ k = 1, \ldots, n_2, \ l = 1, \ldots, n_1$$

so sind diese im reinen Lokationsmodell symmetrisch um den Shift-Effekt $\theta = \mu_2 - \mu_1$ verteilt. Dann ist

$$\widehat{\theta} = \text{Median}\{\Delta_{kl}, \ k = 1, \ldots, n_2, \ l = 1, \ldots, n_1\}$$

ein robuster und asymptotisch erwartungstreuer Schätzer für θ. $\widehat{\theta}$ bezeichnen wir als *Hodges-Lehmann-Schätzer für Shift-Effekt*. Im Lokations-Skalen-Modell ist die Verteilung der paarweisen Differenzen nicht symmetrisch, dennoch ist $\widehat{\theta}$ asymptotisch erwartungstreu (Hodges & Lehmann, 1963).

Wir werden nun in den folgenden Abschnitten die Konstruktion der Konfidenzintervalle nach Lehmann (1963) und nach Bauer (1972) skizzieren. Bei beiden Konstruktionsmethoden sind die Intervallgrenzen Elemente der Menge der paarweisen Differenzen $\{\Delta_{kl}, \ k = 1, \ldots, n_2, \ l = 1, \ldots, n_1\}$.

4.2 Konfidenzintervalle nach Hodges-Lehmann

Es gelte das in (4.1) beschriebene reine Lokationsmodell für die Verteilungsfunktionen F_i. Für die Konstruktion des Konfidenzintervalls nach Lehmann (1963) werden zunächst alle $M = n_1 \cdot n_2$ paarweisen Differenzen $\Delta_{kl} = X_{2k} - X_{1l}$ berechnet und aufsteigend sortiert:

$$\Delta_{(1)} \leq \cdots \leq \Delta_{(M)}.$$

Wir betrachten nun die Zufallsvariablen

$$Z_{ik} = \begin{cases} X_{1k} & i = 1 \\ X_{2k} - \theta & i = 2 \end{cases}$$

und bestimmen R_{ik}, den Gesamtrang von Z_{ik}, sowie $R_{ik}^{(i)}$, den internen Rang von Z_{ik} innerhalb der n_i Zufallsvariablen $Z_{i1}, \ldots, Z_{in_i}$ einer Gruppe $i = 1, 2$. Durch die Annahme des reinen Lokationsmodells für F_1 und F_2 gilt also für die Z_{ik}

$$Z_{1k} \overset{\text{u.i.v.}}{\sim} \widetilde{F}_1, \ k = 1, \ldots, n_1 \quad Z_{2l} \overset{\text{u.i.v.}}{\sim} \widetilde{F}_2, \ l = 1, \ldots, n_2$$

wobei

$$\widetilde{F}_1(x) = \widetilde{F}_2(x) = F(x), \quad x \in \mathbb{R}.$$

Die Zufallsvariablen Z_{ik} sind also unabhängig und identisch verteilt. Die Verteilung der geordneten Differenzen $\Delta_{(i)}$ kann nun mithilfe der Verteilung der Wilcoxon-Rangsumme $R_{2\cdot} = \sum_k R_{2k}$ bestimmt werden. Für ein beliebiges $u \in \mathbb{N}$ bedeutet $\Delta_{(u)} \leq \theta$, dass mindestens $M - u$ der paarweisen Differenzen Δ_{kl} kleiner oder gleich θ sind. Das heißt,

$$\mathrm{P}(\Delta_{(u)} \leq \theta) = \mathrm{P}\left(\sum_{k=1}^{n_2}\sum_{l=1}^{n_1} c(X_{2k} - X_{1l} - \theta) \leq M - u\right).$$

Da nach Definition

$$\sum_{l=1}^{n_1} c(X_{2k} - X_{1l} - \theta) = \sum_{l=1}^{n_1} c(Z_{2k} - Z_{1l}) = n_1 \widehat{F}_1(X_{2k}) = R_{2k} - R_{2k}^{(2)}$$

und

$$\sum_{k=1}^{n_2} R_{2k} - R_{2k}^{(2)} = R_{2\cdot} - R_{2\cdot}^{(2)} = R_{2\cdot} - \frac{n_2(n_2+1)}{2},$$

folgt:

$$\mathrm{P}(\Delta_{(u)} \leq \theta) = \mathrm{P}\left(R_{2\cdot} - \frac{n_2(n_2+1)}{2} \leq M - u\right).$$

Entsprechend erhalten wir ebenso:

$$\mathrm{P}(\Delta_{(u)} > \theta) = \mathrm{P}\left(R_{2\cdot} - \frac{n_2(n_2+1)}{2} \geq M - u + 1\right).$$

Wir können nun entweder die asymptotische Verteilung der Wilcoxon-Rangsumme oder ihre exakte Verteilung verwenden, um die Quantile zur Berechnung der Konfidenzintervalle zu bestimmen. Die rechtsseitig-stetige Version der exakten Verteilungsfunktion F_W^+ von $R_{2\cdot}$ erhält man aus der Permutationsverteilung. Sei $h(s, n_2, N)$ die Anzahl der möglichen Aufteilungen der X_{ik} auf die beiden Gruppen, die zur Rangsumme $R_{2\cdot} = s$ führen. Dann gilt unter $H_0: \ F_1 = F_2$

$$\mathrm{P}(R_{2\cdot} = s) = \frac{h(s, n_2, N)}{\binom{N}{n_2}}$$

wobei $h(s, n_2, N)$ rekursiv durch folgende Formel bestimmt werden kann:

$$h(s, m, N) = h(s, m, N-1) + h(s - N, m - 1, N - 1)$$

mit den Startwerten:

$$h(s,m,N) = 0 \quad \text{für } s < 0$$
$$h(s,m,m) = \begin{cases} 1 & s = m(m+1)/2 \\ 0 & \text{sonst,} \end{cases}$$
$$h(s,0,N) = \begin{cases} 1 & s = 0 \\ 0 & \text{sonst.} \end{cases}$$

Daraus kann F_W^+ für $x \in \mathbb{R}$ nach folgender Formel bestimmt werden:

$$F_W^+(x|n_2,N) = \frac{1}{\binom{N}{n_2}} \sum_{s \le x} h(s,n_2,N).$$

Einen Beweis der Rekursionsformel findet man z.B. in Brunner & Munzel (2002). Die Permutationsquantile sind dann für $q \in (0,1)$ definiert als:

$$w_q(n_2,N) = \max\{x = 1,\ldots,N(N+1)/2 \mid F_W^+(x|n_2,N) \le q\}.$$

Damit können wir aus den geordneten paarweisen Differenzen $\Delta_{(i)}$ die Grenzen des Konfidenzintervalls $\mathcal{I}_{\text{HL}}^{\text{exact}} = [\Delta_{(U)}^{\text{ex}}, \Delta_{(O)}^{\text{ex}})$ bestimmen, wobei sich die Indizes U und O wie folgt berechnen:

$$U = M + n_2(n_2+1)/2 - w_{1-\alpha/2}(n_2,N),$$
$$O = M + n_2(n_2+1)/2 + w_{\alpha/2}(n_2,N).$$

Aus der Rekursionsformel für die Verteilungsfunktion F_W^+ lässt sich ein Shift-Algorithmus zur schnellen Berechnung ableiten, der von Streitberg & Röhmel (1986) vorgestellt wurde.

Außerdem kann man die asymptotische Verteilung der Wilcoxon-Statistik verwenden, um die Quantile zu bestimmen. Die asymptotische Verteilung ist unter $H_0: F_1 = F_2$ eine Standardnormalverteilung:

$$\frac{(\overline{R}_{2\cdot} - \overline{R}_{1\cdot})}{\sqrt{\frac{N(N+1)}{12}}} \sqrt{\frac{n_1 n_2}{N}} \overset{N\to\infty}{\longrightarrow} U \sim \mathcal{N}(0,1).$$

Dabei wird verwendet, dass bei stetigen Verteilungen keine Bindungen auftreten und der Varianzschätzer der WMW-Statistik dann den Wert $\widehat{\sigma}_R^2 = \frac{N(N+1)}{12}$ annimmt. Der Beweis der asymptotischen Normalität der Wilcoxon-Statistik kann mithilfe des Asymptotischen Äquivalenzsatzes (vgl. Anhang Satz A.1) geführt werden (Brunner & Munzel, 2002). Die Grenzen des asymptotischen Hodges-Lehmann-Konfidenzintervalls

$$\mathcal{I}_{\text{HL}}^{\text{norm}} = [\Delta_{(U)}^{\text{asy}}, \Delta_{(O)}^{\text{asy}})$$

werden dann mithilfe des $1-\alpha/2$-Quantils der Standardnormalverteilung $u_{1-\alpha/2}$ bestimmt. Die Indizes $U, O \in \{1,\dots,M\}$ sind bestimmt durch die Forderung, dass der Index U derjenige ist, der am nächsten an

$$\frac{n_1 n_2}{2} - u_{1-\alpha/2}\sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

liegt und O derjenige Index, der am nächsten an

$$\frac{n_1 n_2}{2} + u_{1-\alpha/2}\sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

liegt.

4.3 Konfidenzintervalle nach Bauer

In Bauer (1972) wird eine alternative Methode zur Berechnung von Konfidenzintervallen für den Shift-Effekt beschrieben. Die Konstruktion dieser Konfidenzintervalle beruht auf einer beliebigen linearen Rangstatistik T, die geeignet ist, die Hypothese $H_0: F_1 = F_2$ gegen die Alternative eines Shift-Effekts zu testen. Auch hier setzen wir für die Verteilungsfunktionen F_i das reine Lokationsmodell (4.1) voraus. Sei $\mathbf{x} = (x_{11},\dots,x_{1n_1},x_{21},\dots,x_{2n_2})'$ der Vektor einer Realisation des Zufallsvektors $\mathbf{X}$.

Man kann die lineare Rangstatistik $T = T(\theta)$ als eine Funktion von x_{1l} und $x_{2k} - \theta$ auffassen, das heißt letztlich als eine Funktion des Shift-Effekts θ. Das gesuchte Konfidenzintervall ist dann die Menge

$$\{\theta | T(\theta) \notin C_\alpha\},$$

wobei C_α der Ablehnungsbereich des betrachteten Tests mit Statistik T zum Niveau $\alpha \in (0,\,1)$ ist. Als Funktion in θ ist T eine Treppenfunktion mit Sprüngen ausschließlich in den Punkten $\{\Delta_{kl} = x_{2k} - x_{1l}\}$.

Man berechnet zunächst alle $M = n_1 \cdot n_2$ Werte $\{\Delta_{kl} = x_{2k} - x_{1l}\}$ und sortiert sie aufsteigend:

$$\Delta_{(1)} \le \dots \le \Delta_{(M)}.$$

Dann berechnet man die Werte von T an den Stellen x_{1l}, $l = 1,\dots,n_1$, $x_{2k} - \Delta$, $k = 1,\dots,n_2$, wobei Δ gerade rechts von der j-ten Ordnungsstatistik von $(\Delta_{kl})_{k,l}$ liegt. Dann erhält man eine absteigende Folge von Werten der Teststatistik

$$T(\Delta)_{(1)} \ge \dots \ge T(\Delta)_{(M)}.$$

Außerdem bestimmt man die Quantile der Verteilung der Teststatistik T, beispielsweise die Quantile der asymptotischen Verteilung von T. Dabei sei $v_{1-\alpha/2}$ das obere

Quantil und $v_{\alpha/2}$ das untere Quantil zum Niveau α. Die Grenzen des Konfidenzintervalls ergeben sich dann wieder als diejenigen der geordneten paarweisen Differenzen $\Delta_{(i)}$, die Index U bzw. O haben. Dabei bestimmt man $U \in \{1, \ldots, M\}$ durch

$$U = \max\{i | T(\Delta)_{(i)} \leq v_{1-\alpha/2}\}$$

und entsprechend $O \in \{1, \ldots, M\}$ aus

$$O = \max\{i | T(\Delta)_{(i)} \leq v_{\alpha/2}\},$$

so dass das Konfidenzintervall gerade $\mathcal{I}_{\mathrm{B}} = [\Delta_{(U)}, \Delta_{(O)})$ ist.

4.4 Konfidenzintervalle nach Bauer für heteroskedastische Gruppen

Wie anfangs erwähnt wollen wir in den unabhängigen Gruppen unterschiedliche Varianzen zulassen. Möglicherweise gilt also für die Zufallsvariablen X_{1l}, $l = 1, \ldots, n_1$, X_{2k}, $k = 1, \ldots, n_2$, dass $Var(X_{11}) = \sigma_1^2 \neq \sigma_2^2 = Var(X_{21})$ ist. Wir setzen also voraus, dass das Lokations-Skalen-Modell (4.2) für die Verteilungsfunktionen F_i gilt und betrachten die Hypothese $H_0 : \theta = 0$. Durch die Symmetrie- und Stetigkeitsannahmen an die Verteilungsfunktionen F_i, sowie die Forderung, dass F_i in μ_i invertierbar ist, ist diese Hypothese äquivalent zur nichtparametrischen Hypothese $H_0 : p = \frac{1}{2}$.

Die Konstruktionsmethode von Bauer (1972) kann auch für Konfidenzintervalle dieser erweiterten Hypothesen verwendet werden, weil sie nicht für eine feste Teststatistik formuliert wurde. Die Formulierung der Hypothese mit dem relativen Effekt legt nahe, die Statistik T_N (vgl. Gleichung 3.2) zu verwenden. Sie ist eine lineare Rangstatistik und geeignet, die Hypothese $p = \frac{1}{2}$ zu testen, so dass sie alle gestellten Anforderungen erfüllt.

Man berechnet die Werte der Teststatistik an den Stellen $\mathbf{x}_1$ und an den um die Differenz Δ verschobenen Stellen $x_{2k} - \Delta$, $k = 1, \ldots, n_2$, wobei Δ gerade rechts von einer der geordneten paarweisen Differenzen $\Delta_{(1)} \leq \cdots \leq \Delta_{(M)}$ liegt. Für alle paarweisen Differenzen Δ_{kl} erhalten wir so einen zugehörigen Wert der Teststatistik T_N. Außerdem benötigen wir zur Konstruktion der Intervalle noch die Quantile der Verteilung der Statistik. Wie wir in Kapitel 3 gesehen haben, ist T_N unter $H_0 : p = \frac{1}{2}$ asymptotisch standardnormalverteilt. Man könnte also das Quantil der Standardnormalverteilung $u_{1-\alpha/2}$ verwenden. Für kleine Stichprobenumfänge schlagen Brunner & Munzel (2000) vor, statt der Normalverteilung eine t-Verteilung mit geschätztem Freiheitsgrad (vgl. Gleichung 3.3) zu verwenden. Wir können also auch t-Verteilungsquantile benutzen. Außerdem läßt sich wiederum die Permutationsverteilung von T_N bestimmen und daraus die Quantile berechnen. Obwohl wir unter der Hypothese lediglich $p = \frac{1}{2}$ fordern und somit für die beiden Gruppen nicht voraussetzen, dass die Verteilungen identisch sind, können wir für diese Hypothese

einen asymptotisch gültigen Permutationstest durchführen. Wie in Abschnitt 3.3 gezeigt, erfüllt dieser Permutationstest asymptotisch die geforderte Invarianzeigenschaft, weil die Teststatistik T_N eine standardisierte Statistik ist. Dass die Permutationsverteilung - zumindest für das Treffen der Testentscheidung - auch für kleine Stichprobenumfänge geeignet ist, haben die Simulationen in Kapitel 3.5 gezeigt.

Zur Berechnung der Permutationsverteilungsquantile wenden wir die Permutationen $\pi \in \mathcal{S}_N$ auf den Vektor aller N Beobachtungen $\mathbf{x}$ an:

$$\mathbf{x}_\pi = (x_{\pi(11)}, \ldots, x_{\pi(1n_1)}, x_{\pi(21)}, \ldots, x_{\pi(2n_2)}).$$

Für jeden permutierten Vektor $\mathbf{x}_\pi$ berechnen wir $T_N(\mathbf{x}_\pi)$ und sortieren alle auf diese Weise berechneten $N!$ Werte aufsteigend. Da die Verteilungsfunktionen nach Voraussetzung stetig sind, gibt es mit Wahrscheinlichkeit 1 tatsächlich $N!$ verschiedene Permutationen. Die Quantile entsprechen dann dem U und O größten Wert, das heißt

$$\begin{aligned} v_{1-\alpha/2} &= T_N(\mathbf{x}_\pi)_{(O)} \quad \text{mit} \quad O = \lfloor N! \cdot (1-\alpha/2) \rfloor, \\ v_{\alpha/2} &= T_N(\mathbf{x}_\pi)_{(U)} \quad \text{mit} \quad U = \lfloor N! \cdot (\alpha/2) \rfloor. \end{aligned}$$

Wir haben also drei Möglichkeiten für die Wahl der Quantile $v_{1-\alpha/2}$, $v_{\alpha/2}$ von T_N und damit für die Berechnung von Konfidenzintervallen nach Bauer, die auf dieser Statistik basieren:

- über die asymptotische Verteilung von T_N:
  - Normalverteilung: $\mathcal{I}_{\mathrm{B}}^{\text{norm}} = [\Delta_{(U)}^{\text{norm}}, \Delta_{(O)}^{\text{norm}})$,
  - t-Verteilung: $\mathcal{I}_{\mathrm{B}}^{\text{t-approx}} = [\Delta_{(U)}^{\text{t-approx}}, \Delta_{(O)}^{\text{t-approx}})$
- über die Permutationsverteilung von T_N: $\mathcal{I}_{\mathrm{B}}^{\text{perm}} = [\Delta_{(U)}^{\text{perm}}, \Delta_{(O)}^{\text{perm}})$.

4.5 Datenbeispiel zur Berechnung der Konfidenzintervalle

Wir wollen nun anhand eines kleinen Datenbeispiels die Berechnung der Konfidenzintervalle demonstrieren. Gegeben seien die folgenden Beobachtungen:

$$\mathbf{x}_1 = (1.5, 5.1, 1.6, 3.2, 4.3)', \quad \mathbf{x}_2 = (3.1, 1.4, 2.6)'.$$

Wir haben also $n_1 = 5$ Beobachtungen in der ersten Gruppe und $n_2 = 3$ Beobachtungen in der zweiten Gruppe. Die Gesamt- und internen Ränge sind:

$$\begin{aligned} \mathbf{R}_1 &= (2, 8, 3, 6, 7)', \quad & \mathbf{R}_2 &= (5, 1, 4)', \\ \mathbf{R}_1^{(1)} &= (1, 5, 2, 3, 4)', \quad & \mathbf{R}_2^{(2)} &= (3, 1, 2)' \end{aligned}$$

Entsprechend ist der Rangmittelwert in Gruppe 1 $\overline{R}_{1\cdot} = 5.2$ und der Rangmittelwert in Gruppe 2 $\overline{R}_{2\cdot} = 3.33$. Die geschätzte Varianz in Stichprobe 1 beträgt $\widehat{\sigma}_1^2 = 1.2$ und in Stichprobe 2 $\widehat{\sigma}_2^2 = 1.33$, so dass der Varianzschätzer $V_N^2 = 5.33$ ist. Daraus ergibt sich für die Teststatistik ein Wert von $T_N = -1.106797$. Der geschätzte Freiheitsgrad für die t-Approximation ist $\widehat{f} = 5.88$. Die $M = 5 \cdot 3 = 15$ paarweisen Differenzen Δ_{kl} und die für die Berechnung der Bauer-Konfidenzintervalle benötigten Werte der Teststatistik an den Stellen $\mathbf{x}_1, \mathbf{x}_2 - \Delta_{kl} - \delta$, $\delta = 1 \cdot 10^{-7}$ sind in Tabelle 4.1 angegeben.

Tabelle 4.1: Die Indizes der Beobachtungen (l=Gruppe 1, k=Gruppe 2), deren paarweisen Differenzen Δ_{kl} und Werte der Teststatistik T_N in $\mathbf{x}_1$ und gerade rechts von den um Δ_{kl} verschobenen Beobachtungen $\mathbf{x}_2$.

Index	l	k	Δ_{kl}	$T_N(\mathbf{x}_1, \mathbf{x}_2 - \Delta_{kl} - \delta)$
1	2	2	-3.7	4.5961941
2	5	2	-2.9	2.3452079
3	2	3	-2.5	1.7008401
4	2	1	-2	1.1355499
5	4	2	-1.8	0.7372098
6	5	3	-1.7	0.4423259
7	5	1	-1.2	0.1414214
8	4	3	-0.6	−0.141421
9	1	2	-0.2	−0.442326
10	3	2	-0.1	−1.106797
11	4	1	-0.1	−1.106797
12	1	3	1	−1.70084
13	3	3	1.1	−2.345208
14	1	1	1.5	−4.596194
15	3	1	1.6	−10.6066

Aus diesen Werten und den jeweiligen Quantilen der Verteilungen ergeben sich nun die Indizes derjenigen paarweisen Differenzen Δ_{kl}, die die Grenzen der zweiseitigen Konfidenzintervalle sind. Diese Werte sind für ein Niveau von $\alpha = 5\%$ in Tabelle 4.2 angegeben.

Tabelle 4.2: Quantile, Indizes und Grenzen der Konfidenzintervalle

	Quantile		Indizes		Grenzen	
Intervall	$(1-\alpha/2)$	$(\alpha/2)$	unten	oben	unten	oben
$\mathcal{I}_{\mathrm{B}}^{\mathrm{perm}}$	4.60	-4.60	1	14	-3.7	1.5
$\mathcal{I}_{\mathrm{B}}^{\mathrm{norm}}$	1.96	-1.96	2	12	-2.9	1
$\mathcal{I}_{\mathrm{B}}^{\text{t-approx}}$	2.46	-2.46	1	13	-3.7	1.1
$\mathcal{I}_{\mathrm{HL}}^{\mathrm{norm}}$	1.96	1.96	1	14	-3.7	1.5
$\mathcal{I}_{\mathrm{HL}}^{\mathrm{exact}}$	-2.9	1.1	2	13	-2.9	1.1

4.6 Anwendung im 2×2-Split-Plot-Design

In diesem Abschnitt wollen wir darstellen, wie die Annahmen, die zur Berechnung der Bauer-Konfidenzintervalle im Lokations-Skalen-Modell nötig waren, für ein 2×2-Split-Plot-Design lauten. Dabei betrachten wir zwei Faktoren bezüglich derer die Versuchseinheiten eingeteilt werden können. Häufig treten 2×2-Split-Plot-Designs in der Form eines Vergleichs zweier Gruppen auf, wobei pro Versuchseinheit zwei Messungen zu verschiedenen Zeitpunkten vorliegen. Deshalb werden wir später auch von "Zeiteffekt" sprechen, wenn der Effekt des abhängigen Faktors gemeint ist.

Zunächst formulieren wir das Modell des 2×2-Split-Plot-Designs. Wir betrachten N unabhängige Zufallsvektoren (vgl. Tabelle 4.3)

$$\mathbf{X}_{ik} = (X_{ik1}, X_{ik2})', \; i = 1,2, \; k = 1, \ldots, n_i.$$

Tabelle 4.3: Split-Plot-Design

		Zeit	
		1	2
Gruppe	1	X_{111}	X_{112}
		$\vdots$	$\vdots$
		X_{1n_11}	X_{1n_12}
	2	X_{211}	X_{212}
		$\vdots$	$\vdots$
		X_{2n_21}	X_{2n_22}

Für die Beobachtungen gelte das folgende additive Modell

$$X_{ikj} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ikj},$$

wobei

- $\alpha_i,\ i = 1{,}2$ der Gruppeneffekt,
- $\beta_j,\ j = 1{,}2$ der Zeiteffekt und
- $(\alpha\beta)_{ij}\ i = 1{,}2,\ j = 1{,}2$ die Wechselwirkung zwischen Zeit und Gruppe ist.

Weiterhin seien

$$\boldsymbol{\epsilon}_{ik} = (\varepsilon_{ik1},\varepsilon_{ik2})' \quad \text{mit} \quad \mathrm{E}(\boldsymbol{\epsilon}_{ik}) = \mathbf{0}, \quad Var(\boldsymbol{\epsilon}_{ik}) = \mathbf{V}_i = \begin{pmatrix} \tau_{i,1}^2 & \tau_{i,12} \\ \tau_{i,12} & \tau_{i,2}^2 \end{pmatrix}$$

unabhängige Vektoren von Fehlern, die innerhalb einer Gruppe ($i = 1{,}2$) identisch verteilt sind. Dabei lassen wir ausdrücklich verschiedene (positiv definite) Kovarianzmatrizen $\mathbf{V}_i$ in den beiden Gruppen und unterschiedliche Varianzen $\tau_{i,j}^2$ und Kovarianzen $\tau_{i,12}$ an den beiden Zeitpunkten zu.

Die Zufallsvariablen $X_{ikj},\ k = 1,\ldots,n_i$ sind für eine feste Gruppe i und zu einem festen Zeitpunkt j identisch verteilt. Die zugehörigen marginalen Verteilungsfunktionen bezeichnen wir mit G_{ij}. Wir setzen voraus, dass sich die Verteilungsfunktionen G_{ij} nur durch Shift-Effekte und Skalenalternativen unterscheiden, das heißt, dass das folgende Lokations-Skalen-Modell vorliegt

$$G_{ij}(x) = G\left(\frac{x - \gamma_{ij}}{\tau_{ij}}\right), \quad \text{wobei} \quad \gamma_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

der Erwartungswert in Gruppe i zum Zeitpunkt j sei.

Da wir den Shift-Effekt zwischen den beiden Gruppen untersuchen wollen, bilden wir zunächst die Differenzen zwischen den zwei Beobachtungen einer Versuchseinheit:

$$\begin{aligned} D_{ik} &= X_{ik2} - X_{ik1} \\ &= \beta_2 - \beta_1 + (\alpha\beta)_{i2} - (\alpha\beta)_{i1} + \varepsilon_{ik2} - \varepsilon_{ik1} \\ &= \mu + \mu_i + \eta_{ik}. \end{aligned}$$

Bezüglich der Differenzen D_{ik} ist also μ der Mittelwert, μ_i der Gruppeneffekt und η_{ik} sind unabhängige Fehlerterme mit Erwartungswert 0 und möglicherweise verschiedenen Varianzen σ_i^2. Die Differenzen D_{ik} sind damit unabhängig und innerhalb einer Gruppe identisch verteilt mit Verteilungsfunktion F_i. Damit entsprechend die Differenzen D_{ik} den Beobachtungen X_{ik} aus Abschnitt 4.4.

Für die Konstruktion des Bauer-Konfidenzintervalls für den Shift-Effekt $\theta = \mu_2 - \mu_1$ zwischen den beiden Gruppen müssen wir wie in Abschnitt 4.4 für die Verteilung der Differenzen D_{ik} ein Lokations-Skalen-Modell voraussetzen, das heißt:

$$F_i(x) = F\left(\frac{x - \mu_i}{\sigma_i}\right), \quad i = 1{,}2$$

Zur Konstruktion des Bauer-Konfidenzintervalls müssen wir zunächst voraussetzen, dass die Verteilungsfunktionen F_i stetig sind. Diese Annahme müssen wir im allgemeinen Split-Plot-Design auch weiterhin an die Verteilungsfunktionen F_i der Differenzen pro Versuchseinheit stellen. Sind die Zufallsvariablen X_{ikj} bezüglich des Faktors β_j, $j = 1{,}2$ unabhängig ($\tau_{i,12} = 0$) und besitzen die Marginalverteilungen G_{ij} eine Dichte g_{ij}, so folgt die Stetigkeit der Verteilungsfunktionen F_i aus der Faltung. Sei g_{i1} die Dichte von $-X_{ik1}$ und g_{i2} die Dichte von X_{ik2}, dann ist die Dichte f_i von F_i gegeben durch:

$$f_i(d) = (g_{i1} * g_{i2})(d) = \int g_{i1}(x) g_{i2}(d - x)dx, \quad d \in \mathbb{R},$$

so dass F_i differenzierbar und damit stetig ist.

Um im Lokations-Skalen-Modell ein Konfidenzintervall für den Shift-Effekt berechnen zu können, müssen wir außerdem annehmen, dass die Verteilungsfunktionen F_i symmetrisch sind. Dies ist zum Beispiel erfüllt, wenn die gleichen Bedingungen wie für die Stetigkeit gegeben sind, das heißt, dass die Zufallsvariablen X_{ikj} bezüglich des Faktors β_j, $j = 1{,}2$ unabhängig sind und die Marginalverteilungen G_{ij} eine Dichte g_{ij} besitzen. Darüber hinaus darf bezüglich des Faktors β_j nur ein reines Lokationsmodell vorliegen, das heißt für eine Verteilungsfunktion G_i mit Dichte g_i gilt:

$$g_{i1}(x) = g_i(-x - \gamma_{i1}), \quad g_{i2}(x) = g_i(x - \gamma_{i2}), \quad x \in \mathbb{R},$$

wobei wiederum g_{i1} die Dichte von $-X_{ik1}$ und g_{i2} die Dichte von X_{ik2} ist. Dabei ist $g_{i1}(x) = g_i(-x - \gamma_{i1})$, weil die Verteilungsfunktion G_{-X} einer Zufallsvariable $-X$ in folgendem Zusammenhang zur Verteilungsfunktion G_X der Zufallsvariable X steht:

$$G_{-X}(x) = \mathrm{P}(-X \leq x) = \mathrm{P}(X \geq -x) = 1 - P(X \leq -x) = 1 - G_X(-x).$$

Für die Dichte g_{-X} von G_{-X} folgt dann:

$$g_{-X}(x) = \frac{d}{dx} G_{-X}(x) = \frac{d}{dx} \left(1 - G_X(-x)\right) = g_X(-x).$$

Wir können nun die Dichte f_i wieder aus der Faltung bestimmen und zeigen, dass F_i um $\gamma_{i2} - \gamma_{i1}$ symmetrisch ist. Für $d \in \mathbb{R}$ ist:

$$\begin{aligned} f_i((\gamma_{i2} - \gamma_{i1}) + d) &= (g_{i1} * g_{i2})((\gamma_{i2} - \gamma_{i1}) + d) \\ &= \int g_{i1}(x) g_{i2}((\gamma_{i2} - \gamma_{i1}) + d - x)dx. \end{aligned}$$

Aufgrund des reinen Lokationsmodells gilt

$$\begin{aligned} &= \int g_i(-x - \gamma_{i1}) g_i((\gamma_{i2} - \gamma_{i1}) + d - x - \gamma_{i2})dx \\ &= \int g_i(-x - \gamma_{i1}) g_i(-\gamma_{i1} + d - x)dx. \end{aligned}$$

Ersetze nun $x \to x + d$

$$
\begin{aligned}
&= \int g_i(-x - d - \gamma_{i1}) g_i(-x - \gamma_{i1}) dx \\
&= \int g_i(\gamma_{i2} - \gamma_{i1} - d - x - \gamma_{i2}) g_{i1}(x) dx \\
&= \int g_{i2}(\gamma_{i2} - \gamma_{i1} - d - x) g_{i1}(x) dx \\
&= (g_{i1} * g_{i2})((\gamma_{i2} - \gamma_{i1}) - d) = f_i((\gamma_{i2} - \gamma_{i1}) - d)
\end{aligned}
$$

und damit ist F_i symmetrisch um $(\gamma_{i2} - \gamma_{i1})$.

Um für das 2×2-Split-Plot-Design mit Lokations-Skalen-Modell bezüglich der Differenzen D_{ik} ein Bauer-Konfidenzintervall für den Shift-Effekt berechnen zu können, müssen die Verteilungen der Differenzen F_i also stetig und symmetrisch sowie im Erwartungswert μ_i invertierbar sein. Dies kann entweder direkt für die Verteilungsfunktionen F_i der Differenzen nachgewiesen werden oder folgt unter den genannten Voraussetzungen an die Verteilungsfunktionen G_{ij}.

4.7 Simulationsstudie

Im nächsten Kapitel wollen wir mit einer Simulationsstudie die Eigenschaften der Konfidenzintervalle nach Bauer mit den verschiedenen Möglichkeiten zur Berechnung der Quantile und den beiden Hodges-Lehmann-Konfidenzintervallen bezüglich ihrer *Überdeckungswahrscheinlichkeiten* vergleichen.

In dieser Simulationsstudie sollen die Eigenschaften der verschiedenen Konfidenzintervalle sowohl bei sehr kleinen oder sehr unterschiedlichen Stichprobenumfängen als auch bei verschiedenen den Beobachtungen zugrunde liegenden Verteilungen analysiert werden. Dazu werden wieder die folgenden Stichprobenumfänge verwendet:

$$(n_1,n_2) \in \{(15,15),(15,7),(7,15),(7,7)\}.$$

Als zugrunde liegende Verteilungen betrachten wir hier die Normalverteilung, bimodale Verteilungen, die jeweils als eine Mischung von zwei Normalverteilungen gebildet werden, und Log-Normalverteilungen. Wir führen die Simulationen für ein 2×2-Design durch, wobei die Beobachtungen der beiden Gruppen jedes Faktors unabhängig seien. Es wird ein Effekt für Faktor β von $\beta_2 - \beta_1 = 2$ und ein Shift-Effekt $\gamma_2 - \gamma_1 = 3$ verwendet. Konkret werden die folgenden Verteilungsfunktionen $G_{ij},\ i = 1{,}2,\ j = 1{,}2$ simuliert:

- Normalverteilte Stichproben:
  Neben dem homoskedastischen Fall mit Varianz 1 zu beiden Zeitpunkten und in beiden Gruppen betrachten wir den Einfluss von verschiedenen Varianzen.

Dabei werden die Beobachtungen bei Faktorstufe 2 des Faktors β mit einem Faktor skaliert, der in den beiden unabhängigen Gruppen variiert:

$$G_{11} = \mathcal{N}_{n_1}(0,1), \quad G_{12} = \mathcal{N}_{n_1}(2,2),$$
$$G_{21} = \mathcal{N}_{n_2}(0,1), \quad G_{22} = \mathcal{N}_{n_2}(5,4),$$

- Bimodal-verteilte Stichproben:
  Es werden die folgenden bimodalen Verteilungsfunktionen erzeugt:

$$G_{11} = \frac{7}{10}\,\mathcal{N}_{n_1}(4,1) + \frac{3}{10}\,\mathcal{N}_{n_1}(8,1),$$
$$G_{12} = \frac{7}{10}\,\mathcal{N}_{n_1}(9,1) + \frac{3}{10}\,\mathcal{N}_{n_1}(13,1),$$
$$G_{21} = \frac{3}{10}\,\mathcal{N}_{n_2}(2,2) + \frac{7}{10}\,\mathcal{N}_{n_2}(6,2),$$
$$G_{22} = \frac{3}{10}\,\mathcal{N}_{n_2}(4,2) + \frac{7}{10}\,\mathcal{N}_{n_2}(8,2).$$

  Da sich die Verteilungen einer Gruppe für die beiden Faktorstufen von Faktor β nur durch einen Shift unterscheiden, ist die Verteilung der Differenzen D_{ik}, $i = 1{,}2$, $k = 1, \ldots, n_i$ auch hier symmetrisch.

- Log-Normalverteilte Stichproben:
  Um reine Lokationsalternativen zu erzeugen, werden für alle 4 Stichproben zunächst standardnormalverteilte Zufallsvariablen erzeugt, die dann mit der Exponentialfunktion transformiert werden. Danach werden die Zufallsvariablen verschoben. Dadurch haben wir bezüglich der Faktorstufen des Faktors β ein reines Lokationsmodell, woraus die geforderte Symmetrie der Verteilung der Differenzen D_{ik} folgt.

Wir untersuchen nun die Eigenschaften der fünf oben beschriebenen Konfidenzintervalle. Obwohl wir auch Szenarien betrachten, die das reine Lokationsmodell nicht erfüllen, so dass die Voraussetzungen zur Berechnung der beiden Konfidenzintervalle nach Hodges-Lehmann nicht erfüllt sind, wollen wir sie zum Vergleich dennoch immer berechnen.

Als Vergleichsmaß wird für jedes Konfidenzintervall die empirische *Überdeckungswahrscheinlichkeit (Coverage)* berechnet, das heißt die Wahrscheinlichkeit, dass der zugrundegelegte Shift-Effekt auch im Konfidenzintervall enthalten ist:

$$\mathrm{P}(\theta \in \mathcal{I}).$$

Zum Vergleich werden die in Kapitel 3.5 beschriebenen Zufallsstreifen für die empirische Überdeckungswahrscheinlichkeit berechnet, allerdings entsprechend für die nominellen Überdeckungswahrscheinlichkeiten $1 - \alpha$, $\alpha \in [0.001, 0.10]$. Es werden

10'000 Simulationen durchgeführt. Zur Darstellung der Simulationsergebnisse verwenden wir Graphiken in denen die nominale gegen die simulierte (das heißt erreichte) Überdeckungswahrscheinlichkeit aufgetragen ist (vgl. Abbildung 4.1). Dabei entsprechen Punkte unterhalb des Zufallsstreifens Werten von liberalen Konfidenzintervallen, denn liberale Konfidenzintervalle sind zu schmal und erreichen deshalb nur eine simulierte Überdeckungswahrscheinlichkeit unterhalb der nominalen Überdeckungswahrscheinlichkeit. Entsprechend liegen die Kurven von konservativen Konfidenzintervallen oberhalb des Zufallsstreifens, da deren Konfidenzintervalle zu breit sind und sie eine zu hohe simulierte Überdeckungswahrscheinlichkeit haben.

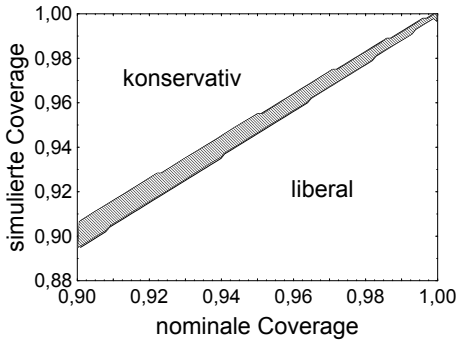


Abbildung 4.1: Coverage-Simulationen: Schematische Darstellung und Interpretation

Zur Berechnung der Permutationsquantile werden bedingte Monte-Carlo-Simulationen verwendet, indem 10'000 zufällige Permutationen auf die Differenzen $D_{ik},\ k = 1,\ldots,n_i,\ i = 1{,}2$ angewendet werden. Bei Bestimmung der Permutationsquantile ersetzen wir den Varianzschätzer wiederum durch $N/(2n_1n_2)$, wenn dieser Null wird, weil die Gruppen komplett getrennte Wertebereiche annehmen (vgl. Kapitel 3.5).

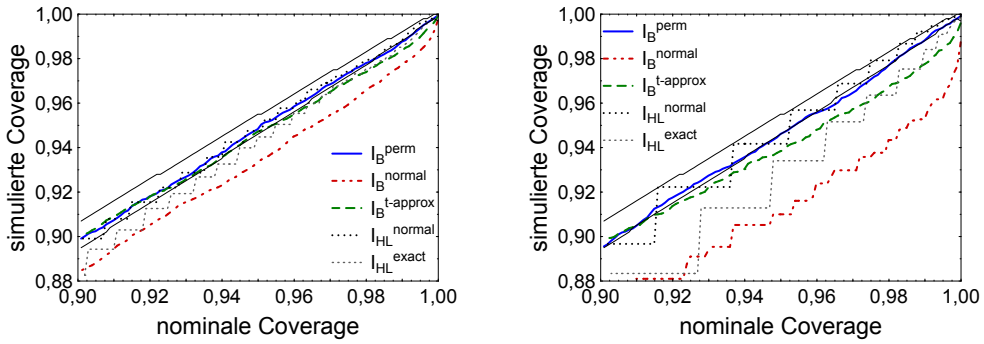
Bei den Hodges-Lehmann-Konfidenzintervallen fällt die mit sinkendem Stichprobenumfang zunehmend ausgeprägte Treppenform der Überdeckungskurven auf. Die Ursache dafür liegt in der Konstruktion der Konfidenzintervalle. Bei der Berechnung der Indizes der Intervallgrenzen gehen außer dem Stichprobenumfang keine weiteren Merkmale der Stichproben ein. Die Quantile werden lediglich mit einem nur vom Stichprobenumfang abhängigen Term multipliziert bzw. um einen solchen Term verschoben. Deshalb nehmen die Konfidenzintervallgrenzen erst dann einen neuen Wert an, wenn sich das Niveau um so viel geändert hat, dass der Index eine neue der paarweisen Differenzen $\Delta_{(i)}$ erreicht. Die Werte der Indixes nehmen für steigende nominale Überdeckungswahrscheinlichkeit also monoton zu (oberer Index) bzw. ab (unterer Index). Dadurch wird die nominale Überdeckungswahrscheinlichkeit der beiden Hodges-Lehmann-Intervalle kurz vor und kurz nach einem Sprung des Indizes insbesondere bei kleinem Stichprobenumfang erst unter- und dann über-

troffen. Bei der Berechnung der Konfidenzintervalle nach der Methode von Bauer (1972) wird dagegen für jede nominale Überdeckungswahrscheinlichkeit der Index der paarweisen Differenzen bestimmt, für den der Wert der Teststatistik am nächsten am Quantil liegt. Dadurch wird für jedes Niveau ein passender Index für die Grenzen des Intervalls ausgewählt und es werden eventuell nicht kontinuierlich fallende bzw. steigende Indizes verwendet. Die beschriebene ausgeprägte Treppenform tritt auch bei den als „exakt“ bezeichneten Hodges-Lehmann-Konfidenzintervallen auf. Diese Intervalle werden mithilfe der Quantile der exakten Permutationsverteilung der Wilcoxon-Rangsumme bestimmt. Da ihre Grenzen dann aber über den Index aus einer Menge geordneter Differenzen ausgewählt werden (vgl. Abschnitt 4.2) entspricht ihre empirische Überdeckungswahrscheinlichkeit dennoch nicht genau der nominalen Überdeckungswahrscheinlichkeit.

4.7.1 Normalverteilte Stichproben

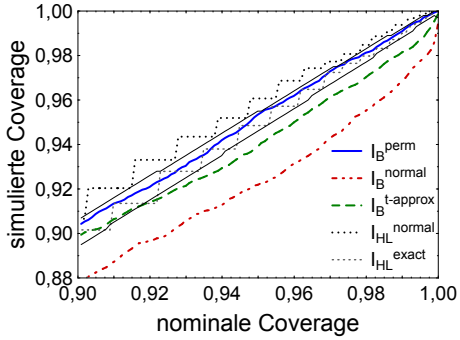
Betrachten wir für normalverteilte Stichproben zunächst den heteroskedastischen Fall. Bei Stichprobenumfang 15 in beiden Gruppen liegen die Überdeckungskurven der Bauer-Konfidenzintervalle mit Permutations- und t-Verteilungsquantilen sowie die Kurve des Hodges-Lehmann-Konfidenzintervalls mit Normalverteilungsquantilen und ab einer nominalen Überdeckungswahrscheinlichkeit von 96% auch die des exakten Hodges-Lehmann-Intervalls sehr dicht beieinander und verlaufen entlang der unteren Zufallsstreifengrenze (Abbildung 4.2(a)). Das Bauer-Intervall mit Normalverteilungsquantilen ist liberal.

Bei nur 7 Beobachtungen in beiden Gruppen werden die Unterschiede zwischen den Intervallen deutlicher (Abbildung 4.2(a)). Von der starken Treppenform abgesehen verläuft die Überdeckungskurve des Intervalls mit Normalverteilungsquantilen im Mittel entlang der unteren Zufallsstreifengrenze, während das exakte Intervall liberal ist. Die Überdeckungskurve des Bauer-Konfidenzintervall mit Permutationsquantilen verläuft sehr dicht entlang der unteren Grenze des Zufallsstreifens, während die Intervalle mit Normal- bzw. t-Verteilungsquantilen sehr bzw. etwas liberal sind.

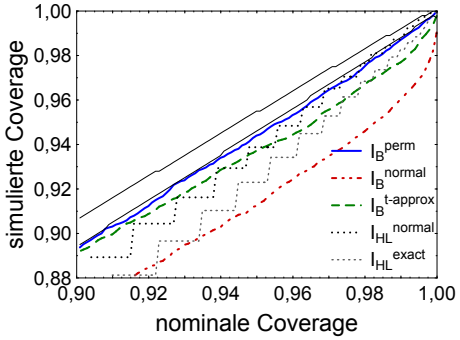


(a) $n_1 = n_2 = 15$

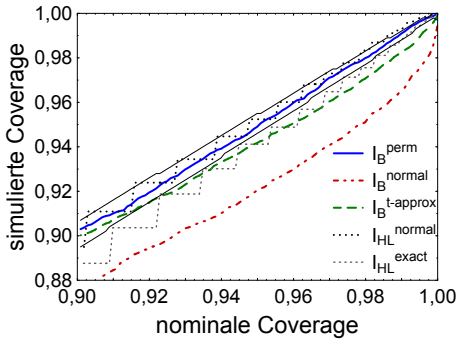
(b) $n_1 = n_2 = 7$



(c) $n_1 = 7$, $n_2 = 15$



(d) $n_1 = 15$, $n_2 = 7$



(e) $n_1 = 7$, $n_2 = 15$ homoskedastisch

Abbildung 4.2: Simulierte gegen nominale Überdeckungswahrscheinlichkeit bei normalverteilten Stichproben, durchgezogene Linien stellen die Zufallsstreifen dar.

Bei 7 Beobachtungen in der Gruppe mit geringer Varianz und 15 Beobachtun-

gen in der anderen Gruppe (Abbildung 4.2(c)) liegt die Überdeckungswahrscheinlichkeit des Bauer-Konfidenzintervalls mit Permutationsquantilen knapp unterhalb der oberen Grenze des Zufallsstreifens. Die Kurven der Hodges-Lehmann-Intervalle sind ebenfalls nach oben verschoben, so dass das Intervall mit Normalverteilungsquantilen nun etwas konservativ wird, während die Kurve des exakten Intervalls recht genau innerhalb des Zufallsstreifens liegt. Tritt die größere Varianz in der kleineren Stichprobe auf (Abbildung 4.2(d)), so verläuft die Überdeckungskurve des Bauer-Intervalls mit Permutationsquantilen wieder entlang der unteren Zufallsstreifengrenze, während beide Hodges-Lehmann-Intervalle hier liberal sind. Die Kurven der Bauer-Konfidenzintervalle mit Normal- und t-Verteilungsquantilen verlaufen bei ungleichen Stichprobenumfängen in beiden Fällen mehr (Normalverteilungsquantil) oder weniger (t-Verteilungsquantil) weit unterhalb der unteren Zufallsstreifengrenze.

Zum Vergleich haben wir außerdem ein reines Lokationsmodell, das heißt homoskedastische Gruppen simuliert. Hier haben die Überdeckungskurven bei gleichen Stichproben einen nahezu identischen Verlauf wie wenn die beschriebene heteroskedastische Varianzstruktur angewendet wird. Für ungleiche Stichprobenumfänge sind bei Vergleich von Abbildungen 4.2(c) und Abbildung 4.2(e) für die Bauer-Intervalle nur sehr geringe Unterschiede zu erkennen. Die Kurven der Hodges-Lehmann Konfidenzintervalle sind bei Simulation gleicher Varianzen etwas nach unten verschoben, so dass das Intervall mit Normalverteilungsquantilen die nominale Überdeckungswahrscheinlichkeit im Rahmen des Zufallsstreifens recht genau trifft, während das exakte Intervall etwas liberal ist. Obwohl hier die Voraussetzungen für die Berechnung der Konfidenzintervalle nach Hodges & Lehmann (1963) erfüllt sind, zeigen sie trotzdem keine klare Überlegenheit zu den Konfidenzintervallen nach Bauer. Eine Zusammenfassung der Eigenschaften aller fünf Konfidenzintervalle findet sich in Tabelle 4.4.

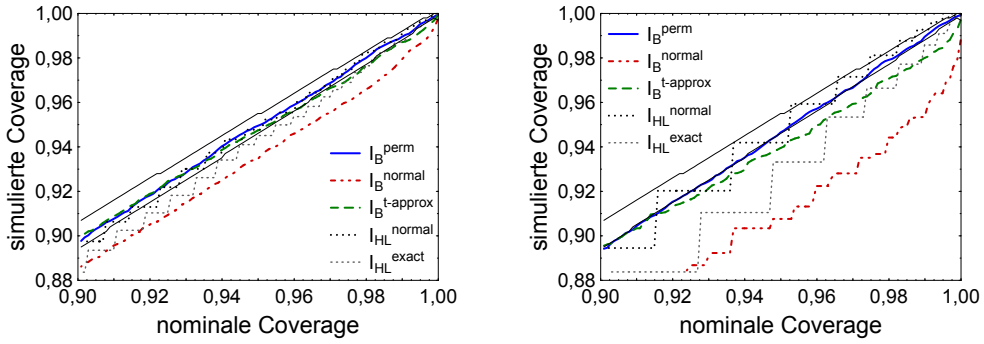
Tabelle 4.4: Eigenschaften der Konfidenzintervalle bei Normalverteilung. ZS = innerhalb des Zufallsstreifens, ZS/lib = teilweise innerhalb des Zufallsstreifens/teilweise liberal, lib=liberal, kons=konservativ.

Varianzstruktur	n_1,	n_2	$\mathcal{I}_{\mathrm{B}}^{\mathrm{perm}}$	$\mathcal{I}_{\mathrm{B}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{B}}^{\text{t-approx}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{exact}}$
homo-skedastisch	15,	15	ZS	lib	ZS	ZS	lib
	7,	7			ZS/lib		
	7,	15			lib		
	15,	7			lib		
hetero-skedastisch	15,	15	ZS	lib	ZS/lib	lib	lib
	7,	7	ZS/lib		lib	ZS/lib	lib
	7,	15	ZS		lib	kons	ZS
	15,	7	ZS/lib		lib	lib	lib

4.7.2 Bimodal-verteilte Stichproben

Sind die zugrunde liegenden Verteilungen bimodal und ist der Stichprobenumfang in beiden Gruppen gleich, so verlaufen die Überdeckungskurven aller fünf Konfidenzintervalle sehr ähnlich wie bei zugrunde liegender Normalverteilung (Abbildungen 4.3(a) und (b)).

Bei Stichprobenumfang 15 in der Gruppe mit der größeren Heteroskedastizität sind die Überdeckungskurven aller Intervalle gegenüber normalverteilten Beobachtungen etwas weiter nach oben verschoben, ohne dass dies die Interpretation ändert (Abbildung 4.3(c)). Im umgekehrten Fall mit dem geringeren Stichprobenumfang bei größerer Varianz sind alle Kurven entsprechend ein klein wenig nach unten verschoben (Abbildung 4.3(d)).



(a) $n_1 = n_2 = 15$ (b) $n_1 = n_2 = 7$

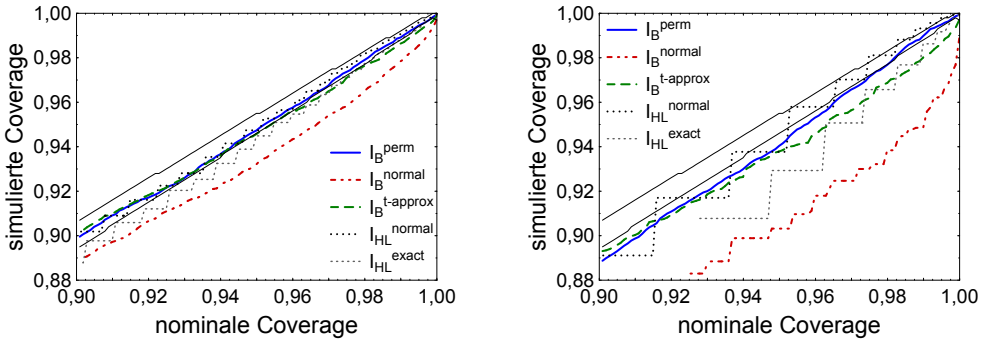
(c) $n_1 = 7, n_2 = 15$ (d) $n_1 = 15, n_2 = 7$

Abbildung 4.3: Simulierte gegen nominale Überdeckungswahrscheinlichkeit bei bimodalverteilten Stichproben, durchgezogene Linien stellen die Zufallsstreifen dar.

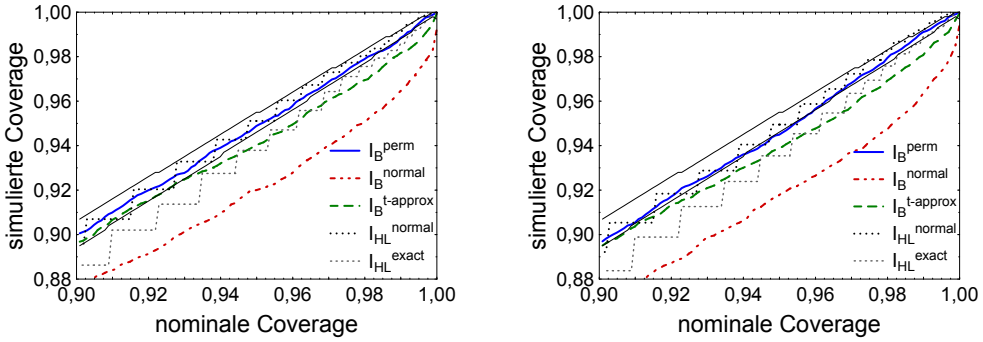
4.7.3 Log-normalverteilte Stichproben

Sind die Stichproben log-normalverteilt und unterscheiden sich wie oben beschrieben nur durch einen Shift-Effekt und einen Effekt für Faktor β, so sind die Überdeckungskurven bei gleichen Stichprobenumfängen gegenüber dem normalverteilten Fall etwas nach unten verschoben, insbesondere bei $n_1 = n_2 = 7$ (vgl. Abbildungen 4.4(a) und (b)).

Ist $n_1 = 7$ und $n_2 = 15$, so verläuft die Kurve des Bauer-Konfidenzintervalls mit Permutationsverteilungsquantilen und die des Hodges-Lehmann-Intervalls mit Normalverteilungsquantilen im Mittel innerhalb des Zufallsstreifens. Im umgekehrten Fall sind beide Überdeckungskurven etwas nach unten verschoben und liegen damit auf der unteren Grenze des Zufallsstreifens. Die anderen drei Intervalle sind in beiden Fällen liberal.



(a) $n_1 = n_2 = 15$ (b) $n_1 = n_2 = 7$



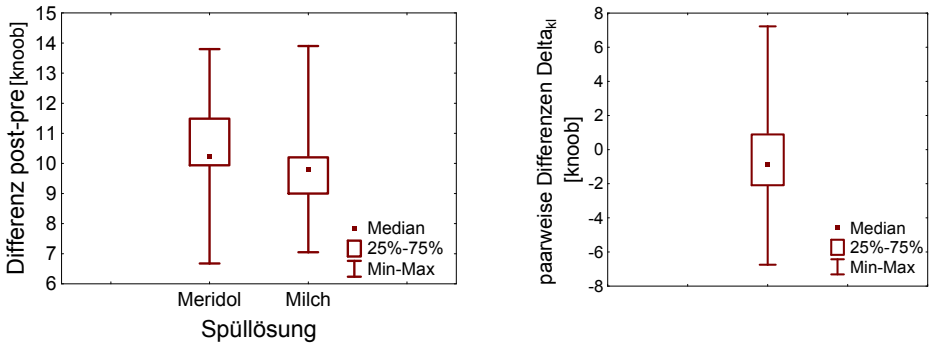
(c) $n_1 = 7$, $n_2 = 15$ (d) $n_1 = 15$, $n_2 = 7$

Abbildung 4.4: Simulierte gegen nominale Überdeckungswahrscheinlichkeit bei log-normalverteilten Stichproben, durchgezogene Linien stellen die Zufallsstreifen dar.

4.8 Anwendungen

4.8.1 Rückhärtung des Dentins

Hier sollen die verschiedenen Konfidenzintervalle auf einen Versuch aus der Zahnmedizin anwendet werden. In der von Wiegand *et al.* (2005) beschriebenen Studie wurde der Einfluss von verschiedenen Spüllösungen auf die Rückhärtung von mit Limonade aufgeweichten Zahnproben untersucht. Dazu wurden die Zahnproben zunächst in Limonade eingelegt und es wurde eine Baselinemessung (pre) durchgeführt. Dann wurden die auf einer Prothese aufgebrachten Proben in den Mund der Probanden eingesetzt. Nach einer Wartezeit von 5 bzw. 30 Minuten spülten die Probanden mit einer von drei Spüllösungen und behielten die Zahnproben danach für weitere 4 Stunden im Mund. Nach dieser Zeit erfolgte die zweite Messung (post).



(a) Differenzen Post - Pre

(b) paarweise Differenzen Δ_{kl}

Abbildung 4.5: Differenzen der Oberflächenhärte

Wir betrachten hier nur Zahnproben, die einer Wartezeit von 5 Minuten unterlagen und vergleichen nur die beiden Spüllösungen Meridol® und Milch, so dass wir einen Stichprobenumfang von 20 Zahnproben pro Gruppe haben. Die Messgröße ist die Oberflächenhärte des Dentins, die in Knoob gemessen wird. In Abbildung 4.5(a) sind die Differenzen der Post- und Pre-Messung dargestellt. Da ihre Verteilung in beiden Gruppen einigermaßen symmetrisch ist, sind die Voraussetzungen für die Berechnung der Konfidenzintervalle für Shifteffekte erfüllt. Die Mittelwerte und Varianzen für die beiden Spüllösungen sind in Tabelle 4.5 aufgeführt.

Tabelle 4.5: Mittelwerte und Varianzen der Dentindifferenzen Post - Pre

Spüllösung	Meridol	Milch
Mittelwert	10.50	10.02
Varianz	2.60	3.43

Abbildung 4.5(b) kann man entnehmen, dass auch die paarweisen Differenzen Milch − Meridol Δ_{kl} symmetrisch um den Median verteilt sind. Der Hodges-Lehmann-Schätzer für den Shifteffekt nimmt den Wert $\widehat{\theta} = -0.9$ an. Die Grenzen der vier untersuchten Konfidenzintervalle sind in Tabelle 4.6 zum Niveau 95% angegeben. Alle Konfidenzintervalle haben ähnliche Grenzen und enthalten alle die Null, das heißt man kann die Hypothese $H_0 : \ \theta = 0$ nicht ablehnen.

Tabelle 4.6: Grenzen der Konfidenzintervalle für einen Shifteffekt der Rückhärtung des Dentins bei Spülen mit Meridol gegenüber Spülen mit Milch.

Grenzen	$\mathcal{I}_{\mathrm{B}}^{\mathrm{perm}}$	$\mathcal{I}_{\mathrm{B}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{B}}^{\mathrm{t\text{-}approx}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{exact}}$
oben	0.125	0.125	0.125	0.125	0.1
unten	-1.6	-1.575	-1.6	-1.6	-1.6

4.8.2 Post-Operatives Ödem

Eine weitere Studie, die wir auf einen Shift-Effekt analysieren wollen, beschäftigt sich mit der Bildung eines Ödems bzw. Rötungen der Haut, die nach einem chirurgischen Eingriff auftreten. Dabei sollte in dieser randomisierten klinischen Studie untersucht werden, ob ein Medikament im Vergleich mit Placebo das Ödem bzw. die Rötungen schneller abklingen lässt. Es wurden jeweils 29 Patienten mit dem Medikament bzw. Placebo behandelt. Die Zielvariable ist die Temperatur (in °C mit 10 multipliziert), die sowohl an den operierten als auch an den nicht-operierten Händen gemessen wurde. Diese Messungen erfolgten am Tag vor der Operation (Baseline) sowie am ersten, dritten und fünften Tag nach der Operation. Wir danken Herrn Freundenstein der Firma Schaper & Brümmer, Salzgitter-Ringelheim, für die freundliche Überlassung dieser Daten.

Wir wollen hier nur die operierten Hände betrachten und dabei für jeden Zeitpunkt die Differenzen zur Baseline-Messung analysieren. Die Differenzen zur Baseline-Messung sind in Abbildung 4.6 dargestellt.

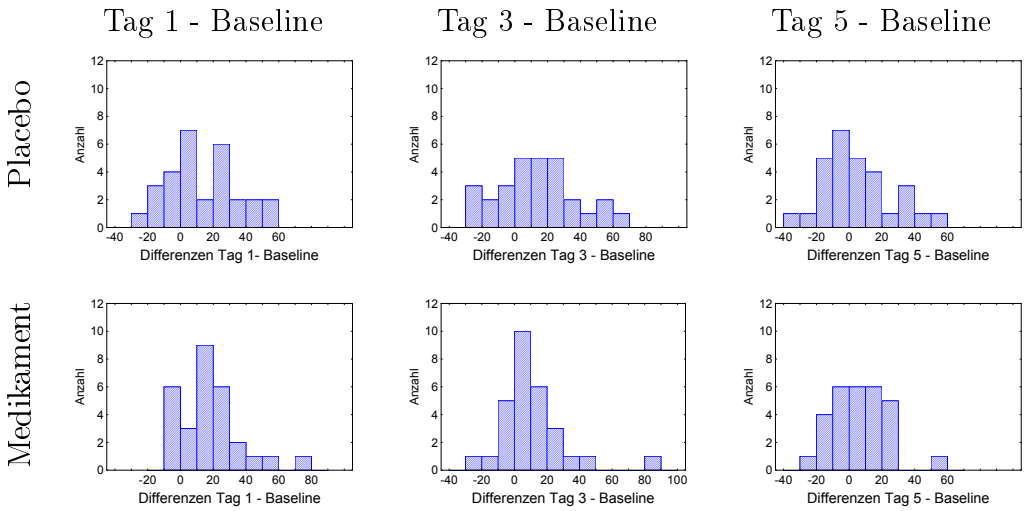


Abbildung 4.6: Histogramme der Differenzen zu Baseline

Die Temperatur steigt durch die Operation also gegenüber Baseline in beiden Behandlungsgruppen eher an und ist ungefähr symmetrisch, wie gefordert. Die Konfidenzintervalle sollen einen möglichen Shift-Effekt zwischen den mit dem Medikament behandelten Patienten und den mit Placebo behandelten Patienten aufdecken.

Die Mittelwerte und Varianzen in den beiden Gruppen sind für die Differenzen aller drei Zeitpunkte in Tabelle 4.7 zusammengefasst. Die paarweisen Differenzen Placebo $-$ Medikament Δ_{kl} sind in Abbildung 4.7 dargestellt und auch sie sind hinreichend symmetrisch um den Median verteilt.

Tabelle 4.7: Mittelwerte und Varianzen der Differenzen zu Baseline

	Behandlung	Tag 1 - Baseline	Tag 3 - Baseline	Tag 5 - Baseline
Mittelwert	Placebo	14.34	14.03	5.79
	Medikament	18.14	10.83	6.93
Varianz	Placebo	445.6	521.6	482.2
	Medikament	366.3	390.1	281.3

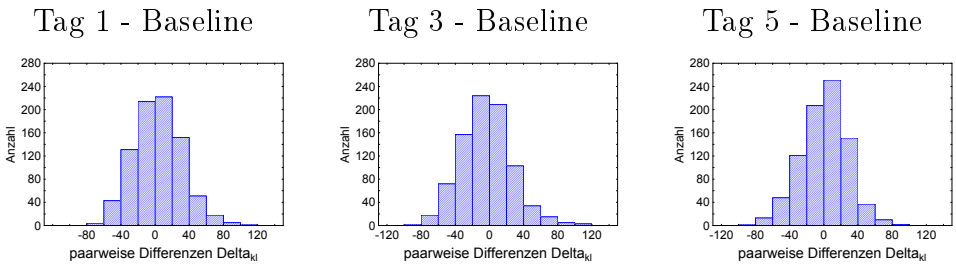


Abbildung 4.7: Paarweise Differenzen Δ_{kl} für die Differenzen zu Baseline

Da wir hier für alle 3 Zeitpunkte Konfidenzintervalle berechnen, müssen wir für die auftretende Multiplizität adjustieren, das heißt bei Korrektur nach Bonferroni das Niveau $\alpha = 0.05/3 = 0.0167$ verwenden. Die Grenzen der adjustierten Konfidenzintervalle für die Differenzen aller drei Tage finden sich in Tabelle 4.8. Für alle drei Zeitpunkte enthalten alle fünf Konfidenzintervalle die Null und lehnen die Hypothesen jeweils auf dem 1.67%-Niveau somit nicht ab.

Tabelle 4.8: Grenzen der Konfidenzintervalle für einen Shifteffekt von Placebo − Medikament für die Differenzen zu Baseline und Hodges-Lehmann-Schätzer für den Shifteffekt

Differenz bzgl.	Grenzen	$\mathcal{I}_{\mathrm{B}}^{\mathrm{perm}}$	$\mathcal{I}_{\mathrm{B}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{B}}^{\mathrm{t\text{-}approx}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{norm}}$	$\mathcal{I}_{\mathrm{HL}}^{\mathrm{exact}}$	Shift-Effekt
Tag 1	oben	9	9	9	10	10	-3
	unten	-17	-17	-17	-17	-17	
Tag 3	oben	17	17	17	18	17	4
	unten	-10	-10	-11	-11	-10	
Tag 5	oben	10	10	10	11	10	-2
	unten	-14	-14	-14	-14	-14	

4.9 Zusammenfassung

Neben den klassischen Konfidenzintervallen für den Shift-Effekt nach Lehmann (1963) mit Permutations- bzw. asymptotischer Verteilung wurden die von Bauer (1972) beschriebenen Konfidenzintervalle betrachtet. Die Konstruktionsmethode nach Bauer ermöglicht es, das reine Lokationsmodell durch das Lokations-Skalen-Modell zu ersetzen, bei dem auch Skalenalternativen zugelassen werden: $F_i(x) = F((x-\gamma_i)/\sigma_i)$, $i = 1{,}2$. Auf die nichtparametrischen Hypothesen übertragen bedeutet dies, dass wir die klassische Hypothese $H_0: \ F_1 = F_2$ durch die Hypothese $H_0: \ p = \frac{1}{2}$ ersetzen. Der heteroskedastische Fall entspricht dann wiederum einer Behrens-Fisher-Situation. Deshalb wurde zur Berechnung der Konfidenzintervalle nach Bauer die

für diesen Fall geeignete Teststatistik T_N (vgl. 3.2) verwendet. Die Quantile von T_N können außer über die asymptotisch angenommene Normalverteilung bzw. über die von Brunner & Munzel (2000) für kleine Stichprobenumfänge empfohlene Approximation mit einer t-Verteilung, auch über die Permutationsverteilung bestimmt werden.

Die Simulationsergebnisse zeigen, dass die Bauer-Konfidenzintervalle mit Permutationsverteilungsquantilen für alle untersuchten Verteilungsfunktionen der Daten sowohl im homoskedastischen Fall als auch bei verschiedenen Varianzen in den Gruppen die besten Überdeckungswahrscheinlichkeiten erreichen. Das gilt insbesondere bei sehr kleinen (7 pro Gruppe) und sehr verschiedenen (7 gegenüber 15 Beobachtungen) Stichprobenumfängen in den beiden Gruppen. Meist liegt die empirische Überdeckungswahrscheinlichkeit innerhalb des Zufallsstreifens und anderenfalls wird das Intervall leicht liberal. Es kommt aber unter allen verglichenen Intervallen der nominalen Überdeckungswahrscheinlichkeit stets am nächsten. Das Hodges-Lehmann-Konfidenzintervall mit Normalverteilungsquantilen erreicht im Mittel zwar eine ähnliche Überdeckungswahrscheinlichkeit, ist aber aufgrund der vor allem bei kleinen Stichprobenumfängen sehr groben Treppenstruktur der Überdeckungskurve ungenau. Das Bauer-Konfidenzintervall mit t-Verteilungsquantilen ist für kleine Stichprobenumfänge leicht liberal, wogegen das Bauer-Intervall mit Normalverteilungsquantilen in allen Fällen liberal ist. Das exakte Hodges-Lehmann-Konfidenzintervall mit Permutationsverteilungsquantilen ist ebenfalls in den meisten Fällen liberal, für kleine Stichprobenumfänge auch im homoskedastischen Fall.

Der Vorteil der Verwendung der Teststatistik T_N bei der Konstruktionsmethode nach Bauer (1972) ist, dass keine Homoskedastizität der beiden Gruppen gefordert werden muss. Allerdings muss vorausgesetzt werden, dass die Verteilungsfunktionen der Gruppen stetig und symmetrisch, sowie am Erwartungswert invertierbar sind. Benutzt man die Permutationsquantile bei der Berechnung der Intervallgrenzen, so ist man weder auf die Annahme einer bestimmten Verteilung der Daten noch auf die Gültigkeit von asymptotischen oder approximativen Verteilungsaussagen angewiesen. Die Verwendung des im Rahmen dieser Arbeit erstellten SAS-IML-Makros zur Berechnung der hier vorgestellten Konfidenzintervalle wird im Anhang C beschrieben. Das Makro kann von der Web-Seite der Abteilung Medizinische Statistik heruntergeladen werden (PERM_KI, 2006).

A Anhang

A.1 Asymptotischer Äquivalenzsatz

SATZ A.1 (Asymptotischer Äquivalenzsatz)
Die Zufallsvariablen $X_{ik} \sim F_i$, $i = 1{,}2$, $k = 1,\ldots,n_i$ seien unabhängig, es sei $\sigma_1^2 = Var(F_2(X_{11}))$ die Varianz in Gruppe 1 und $\sigma_2^2 = Var(F_1(X_{21}))$ die Varianz in Gruppe 2. Ferner bezeichne $\widehat{p} = \frac{1}{N}(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}) + \frac{1}{2}$ den Rangschätzer für den relativen Effekt $p = \int F_1 dF_2$. Falls $\frac{N}{n_i} \leq N_0 < \infty$ für $N \to \infty$, $i = 1{,}2$ und $\sigma_1^2, \sigma_2^2 > 0$ sind, dann gilt:

$$\sqrt{N}(\widehat{p} - p) \doteqdot \sqrt{N}\left(\frac{1}{n_2}\sum_{k=1}^{n_2} F_1(X_{2k}) - \frac{1}{n_1}\sum_{k=1}^{n_1} F_2(X_{1k}) + 1 - 2p\right).$$

Beweis. Der Satz folgt aus der allgemeineren Version des Asymptotischen Äquivalenzsatzes, der in Brunner & Munzel (2002, S. 207) bewiesen wird. □

A.2 Asymptotische Normalität der Permutationsverteilung

Seien $X_{N1}, \ldots, X_{NN}$ beliebige reellwertige Zufallsvariablen und sei

$$T_N = T_N(\mathbf{X}) = \sum_{i=1}^{N} c_{Ni} X_{Ni}$$

eine lineare Statistik, die durch das Schema von Koeffizienten $(c_{Ni})_{i \leq N}$ $\forall N \in \mathbb{N}$ bestimmt ist. Sei

$$\widetilde{T}_N = T_N / V_N^{1/2}$$

eine studentisierte Version von T_N, wobei $V_N = V_N((X_{Ni})_{i \leq N})$ als Varianzschätzer aufgefasst werden kann. Die asymptotische Normalität der Permutationsverteilung von $\widetilde{T}_N$ beruht auf der Gleichverteilung der Permutationen $\pi \in \mathcal{S}_N$ bezüglich der von der Verteilung der Daten P unabhängigen Verteilung P*. Die Permutationsstatistik wird für festes $(X_{N\pi(i)})_{i \leq N}$, $\pi \in \mathcal{S}_N$ mit

$$\pi \mapsto \widetilde{T}_N((X_{N\pi(i)})_{i \leq N})$$

bezeichnet.

SATZ A.2 (Asymptotische Normalität)
Es seien die folgenden Bedingungen erfüllt:

$$\sum_{i=1}^{N} c_{Ni}^2 = 1, \quad \sum_{i=1}^{N} c_{Ni} = 0 \quad \forall N \in \mathbb{N} \tag{A.1}$$

$$\max_{1\le i\le N} |c_{Ni}| \to 0 \quad \textit{wenn} \quad N \to \infty \tag{A.2}$$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (X_{Ni} - \overline{X}_{N.})^2 1_{[d,\infty)} \left(\left| X_{Ni} - \overline{X}_{N.} \right| \right) \longrightarrow 0 \quad P - f.s. \quad d \to \infty \tag{A.3}$$

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (X_{Ni} - \overline{X}_{N.})^2 > 0 \quad P - f.s. \tag{A.4}$$

$$\exists\, \tau > 0 \; \frac{1}{N} \sum_{i=1}^{N} (X_{Ni} - \overline{X}_{N.})^2 (V_N((X_{N\pi(i)})_{i\le N}))^{-1} \overset{P\otimes P^*}{\longrightarrow} \tau^2 \quad \textit{wenn} \quad N \to \infty. \tag{A.5}$$

Dann folgt die asymptotische Normalität:

$$\sup_{t\in\mathbb{R}} \left(\left| P^* \left(\widetilde{T}_N((X_{N\pi(i)})_{i\le N}) \le t \right) - \Phi(t/\tau) \right| \right) \overset{P}{\to} 0,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist.

Beweis. Der Beweis findet sich in Abschnitt 3 in Janssen (1997). □

A.3 Varianzformel von Hájek

Betrachte Zufallsvariablen X_i, $i = 1, \ldots, N$ und deren Ränge R_i. Die Verteilung F der X_i sei unstetig an den Stellen $\{\alpha_h\}$. Seien $A_1, \ldots, A_g$ die Mengen der Indizes auf denen Bindungen der X_i auftreten. Sei τ_k die Anzahl der Elemente in A_k. Die $a_N(i)$ seien Scores, so dass es eine quadratintegrierbare Funktion ϕ gibt, mit

$$\int_0^1 (\phi(u) - \overline{\phi})^2 du > 0, \quad \text{wobei} \quad \overline{\phi} = \int_0^1 \phi(u) du.$$

Es gelte:

$$\lim_{N\to\infty} \int_0^1 [a_N(1 + [uN]) - \phi(u)]^2 du = 0.$$

Bei Bindungen sei

$$a_N(i, \tau) := \frac{1}{\tau_k} \sum_{j\in A_k} a_N(j) \quad \text{für} \quad i \in A_k.$$

Außerdem seien für $\varepsilon > 0$ $I_h = (F(\alpha_h - \varepsilon), F(\alpha_h))$ Intervalle um die Werte von F an den Sprungstellen und

$$\phi^F(u) := \begin{cases} \phi(u) & u \notin \cup_{h=1}^{\infty} I_h \\ \int_{I_h} \phi(v)dv / \int_{I_h} dv & u \in I_h,\ h \geq 1. \end{cases}$$

Sei nun

$$S = \sum_{i=1}^{N} c_i a_N(R_i)$$

die zugehörige lineare Rangstatistik und

$$\overline{a}_\cdot = \frac{1}{N}\sum_{i=1}^{N} a_N(i), \quad \overline{c}_\cdot = \frac{1}{N}\sum_{i=1}^{N} c_i, \quad \sigma_a^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a_N(i) - \overline{a}_\cdot)^2.$$

Satz A.3 (Varianzformel)
Unter der Hypothese, dass die Zufallsvariablen X_i, $i = 1,\ldots,N$ unabhängig sind, und wenn

$$\int_0^1 (\phi^F(u) - \overline{\phi})^2 du > 0$$

sowie

$$\frac{\sum_{i=1}^{N}(c_i - \overline{c}_\cdot)^2}{\max_{1\leq i\leq N}\{(c_i - \overline{c}_\cdot)^2\}} \to \infty \quad \textit{für} \quad N \to \infty \tag{A.6}$$

erfüllt sind, ist die auf die Ränge $R_1,\ldots,R_N$ bedingte Verteilung von S asymptotisch eine Normalverteilung. Die bedingte Varianz berechnet sich nach folgender Formel:

$$Var(S|R_1,\ldots,R_N) = \sigma_a^2 \sum_{i=1}^{N}(c_i - \overline{c}_\cdot)^2. \tag{A.7}$$

Beweis. Die Konvergenzaussage steht auf Seite 198 in Hájek & Ŝidák (1967). Die Varianzformel ist ein Spezialfall des Satzes II 3.1 (a) (Hájek & Ŝidák, 1967, Seite 57), der dort bewiesen wird. □

A.3.1 Varianzformel von Hájek im Beweis Satz 3.2

Beim Beweis der asymptotischen Normalität der Permutationsverteilung wenden wir die Varianzformel auf die Teilsummen W_i an. Dabei haben W_2 und W_3 die Form

$$W_r = \sum_{i=1}^{2}\sum_{k=1}^{n_i} m_{rik}\widehat{Z}_{ik} = \sum_{i=1}^{2}\sum_{k=1}^{n_i} c_{ik} a_N(n_j \widehat{F}_j(X_{ik})) \quad i \neq j,\ i,j = 1,2,\ r = 2,3,$$

und W_1 ist

$$W_1 = \sum_{i=1}^{2}\sum_{k=1}^{n_i} d_{ik}\widehat{Z}_{ik}^2 = \sum_{i=1}^{2}\sum_{k=1}^{n_i} d_{ik} a_N(n_j \widehat{F}_j(X_{ik})) \quad i \neq j,\ i,j = 1,2,$$

wobei $n_j\widehat{F}_j(X_{ik}) = R_{ik} - R_{ik}^{(i)}$. Betrachten wir zunächst W_2 und W_3. Dann sind die Scores $a_N(i)$ gleich:

$$a_N(k) = \frac{1}{N}\left(k - \frac{n_j}{2}\right), \quad j \neq i.$$

Nach der Definition der normalisierten Verteilungsfunktionen stimmt a_N bei Bindungen dann mit der Definition von Hájek & Šidák (1967) überein. Die gesuchte Funktion ϕ definieren wir als

$$\phi_i(u) := \begin{cases} \frac{k}{N} - \frac{1-\kappa}{2} & i = 1 \\ \frac{k}{N} - \frac{\kappa}{2} & i = 2 \end{cases} \quad \text{wobei} \quad k = \max\left\{ l = 1,\ldots,N \middle| \frac{l}{N} \leq u \right\}.$$

Dann gilt für ϕ_i:

$$\overline{\phi}_i = \int_0^1 \phi_i(u)du = \begin{cases} \frac{1}{N}\sum_{k=0}^{N} \frac{k}{N} - \frac{1-\kappa}{2} & i = 1 \\ \frac{1}{N}\sum_{k=0}^{N} \frac{k}{N} - \frac{\kappa}{2} & i = 2 \end{cases} = \begin{cases} \frac{1}{2N} + \frac{\kappa}{2} & i = 1 \\ \frac{1}{2N} + \frac{1-\kappa}{2} & i = 2. \end{cases}$$

Damit folgt:

$$\begin{aligned} &\int_0^1 [\phi_i(u) - \overline{\phi}_i]^2 du \\ &= \frac{1}{N}\sum_{k=0}^{N}\left[\frac{k}{N} - \frac{1}{2} - \frac{1}{2N}\right]^2 \\ &= \frac{1}{N}\sum_{k=0}^{N}\frac{k^2}{N^2} + \left(\frac{1}{2} + \frac{1}{2N}\right)^2 - 2\left(\frac{1}{2} + \frac{1}{2N}\right)\frac{1}{N}\sum_{k=0}^{N}\frac{k}{N} \\ &= \frac{1}{N^3}\frac{N(N+1)(2N+1)}{6} + \left(\frac{1}{2} + \frac{1}{2N}\right)^2 - 2\left(\frac{1}{2} + \frac{1}{2N}\right)\left(\frac{1}{N^2}\frac{N(N+1)}{2}\right) \\ &= \frac{1}{3} + \frac{3N-1}{12N^2} > 0. \end{aligned}$$

Dies folgt aufgrund der Definition von ϕ auch für ϕ^F. Für a_N gilt:

$$\begin{aligned} a_N(1 + [uN]) &= \begin{cases} \frac{1}{N}\left(k + 1 - \frac{n_j}{2}\right) & u \in \left[\frac{k}{N}, \frac{k}{N} + \frac{1}{2N}\right) \\ \frac{1}{N}\left(k + 2 - \frac{n_j}{2}\right) & u \in \left[\frac{k}{N} + \frac{1}{2N}, \frac{k+1}{N}\right) \end{cases} \\ &= \begin{cases} \frac{k}{N} + \frac{1}{N} - \frac{n_j}{2N} & u \in \left[\frac{k}{N}, \frac{k}{N} + \frac{1}{2N}\right) \\ \frac{k}{N} + \frac{2}{N} - \frac{n_j}{2N} & u \in \left[\frac{k}{N} + \frac{1}{2N}, \frac{k+1}{N}\right), \end{cases} \end{aligned}$$

wobei $k \in \{1,\ldots,N\}$. Damit folgt für die Differenz von a_N und ϕ für $i = 1$:

$$a_N(1+[uN]) - \phi_1(u) = \begin{cases} \frac{1}{N} - \frac{n_j}{2N} + \frac{1-\kappa}{2} & u \in \left[\frac{k}{N}, \frac{k}{N} + \frac{1}{2N}\right) \\ \frac{2}{N} - \frac{n_j}{2N} + \frac{1-\kappa}{2} & u \in \left[\frac{k}{N} + \frac{1}{2N}, \frac{k+1}{N}\right) \end{cases}$$

und entsprechend für $i = 2$:

$$a_N(1+[uN]) - \phi_2(u) = \begin{cases} \frac{1}{N} - \frac{n_j}{2N} + \frac{\kappa}{2} & u \in \left[\frac{k}{N}, \frac{k}{N} + \frac{1}{2N}\right) \\ \frac{2}{N} - \frac{n_j}{2N} + \frac{\kappa}{2} & u \in \left[\frac{k}{N} + \frac{1}{2N}, \frac{k+1}{N}\right) \end{cases}.$$

Damit folgt

$$a_N(1+[uN]) - \phi_i(u) \overset{N\to\infty}{\longrightarrow} 0 \quad \text{für} \quad i = 1,2$$

und damit auch

$$\int_0^1 [a_N(1+[uN]) - \phi(u)]^2 du \overset{N\to\infty}{\longrightarrow} 0.$$

Bleibt noch zu zeigen, dass auch Bedingung (A.6) erfüllt ist. Wie im Beweis zu Satz 3.2 gezeigt wird ist

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i}(m_{2ik} - \overline{m}_{2..})^2 = \frac{1}{n_1 - 1}.$$

Für das Maximum gilt:

$$\max\left\{(m_{2ik} - \overline{m}_{2..})^2\right\} = \frac{1}{N}\frac{1}{n_1 - 1}\max\left\{\frac{n_2}{n_1}, \frac{n_1}{n_2}\right\}.$$

Damit folgt:

$$\frac{\sum_{i=1}^{2}\sum_{k=1}^{n_i}(m_{2ik} - \overline{m}_{2..})^2}{\max\left\{(m_{2ik} - \overline{m}_{2..})^2\right\}} = \begin{cases} \frac{Nn_2}{n_1} & n_2 > n_1 \\ \frac{Nn_1}{n_2} & n_2 < n_1 \\ N & n_2 = n_1 \end{cases} \geq N \overset{N\to\infty}{\longrightarrow} \infty.$$

Damit sind für W_2 und W_3 alle Voraussetzungen für die Anwendung der Varianzformel erfüllt.

Für W_1 definieren wir a_N als

$$a_N(i) = \left[\frac{1}{N}\left(k - \frac{n_j}{2}\right)\right]^2$$

und entsprechend ϕ als

$$\phi_i(u) := \begin{cases} \left(\frac{k}{N} - \frac{1-\kappa}{2}\right)^2 & i = 1 \\ \left(\frac{k}{N} - \frac{\kappa}{2}\right)^2 & i = 2 \end{cases} \quad \text{wobei} \quad k = \max\left\{l = 1,\ldots,N \,\middle|\, \frac{l}{N} \leq u\right\}.$$

Die Bedingungen des Satzes A.3 können dann analog zu obigen Berechnungen gezeigt werden. Die Bedingung an die Koeffizienten ist auch erfüllt, denn wie im Beweis zu Satz 3.2 berechnet wird, ist:

$$\begin{aligned}\sum_i \sum_k (d_{ik} - \overline{d}_{..})^2 &= n_1 \left(\frac{1}{n_1 - 1} - \frac{1}{n_1} \frac{n_2}{n_2 - 1} \right)^2 + n_2 \left(\frac{1}{n_2 - 1} - \frac{1}{n_2} \frac{n_1}{n_1 - 1} \right)^2 \\ &=: n_1 C_1^2 + n_2 C_2^2\end{aligned}$$

Das Maximum lautet mithilfe der Konstanten C_1^2, C_2^2:

$$\max \left\{ (d_{ik} - \overline{d}_{..})^2 \right\} = \max\{C_1^2, C_2^2\}.$$

Daraus folgt für den Quotient:

$$\begin{aligned}\frac{\sum_i \sum_k (d_{ik} - \overline{d}_{..})^2}{\max (d_{ik} - \overline{d}_{..})^2} &= \begin{cases} n_1 + n_2 \frac{C_2^2}{C_1^2} & C_1^2 > C_2^2 \\ n_1 \frac{C_1^2}{C_2^2} + n_2 & C_1^2 < C_2^2 \\ N & C_1^2 = C_2^2 \end{cases} \\ &\geq \begin{cases} n_1 & C_1^2 > C_2^2 \\ n_2 & C_1^2 < C_2^2 \\ N & C_1^2 = C_2^2 \end{cases} \overset{N \to \infty}{\longrightarrow} \infty.\end{aligned}$$

Damit sind auch für die Anwendung der Varianzformel bei W_1 alle Voraussetzungen erfüllt.

B SAS-Makro für den Permutationstest

Das SAS-IML-Makro PERM_BF führt in einem Zwei-Stichproben-Behrens-Fisher-Design für die nichtparametrische Hypothese $H_0 : \ p = \frac{1}{2}$ einen zweiseitigen Permutationstest durch. Für die Teststatistik werden die Ränge berechnet, wobei R_{ik} der Mittelrang der Beobachtung X_{ik} unter allen $N = n_1 + n_2$ Beobachtungen und $R_{ik}^{(i)}$ der interne Rang der Beobachtung X_{ik} unter allen n_i Beobachtungen aus Gruppe i ($i = 1,2$, $k = 1,\ldots,n_i$) ist. Der Permutationstest basiert auf der studentisierten Teststatistik T_N

$$T_N = \frac{\overline{R}_{2.} - \overline{R}_{1.}}{V_N} \sqrt{\frac{n_1 n_2}{N}}, \tag{B.1}$$

wobei die Varianz V_N^2 das gewichtete Mittel der empirischen Varianzen in den beiden Stichproben ist:

$$V_N^2 = N \left(\frac{1}{n_2} \widehat{\sigma}_1^2 + \frac{1}{n_1} \widehat{\sigma}_2^2 \right)$$

mit

$$\widehat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(R_{ik} - R_{ik}^{(i)} - \overline{R}_{i.} + \frac{n_i + 1}{2} \right)^2 .$$

Der p-Wert des Tests wird ermittelt, indem zufällig ausgewählte Permutationen $\pi \in \mathcal{S}_N$ auf den Datenvektor $\mathbf{X} = (X_{11},\ldots,X_{1n_1},X_{21},\ldots,X_{2n_2})'$ angewendet werden und die Statistik T_N jeweils neu berechnet wird. Dabei wird der Anteil der Permutationen bestimmt, für die die neu berechnete Statistik T_N^* einen betragsmäßig größeren Wert annimmt als die Teststatistik T_N mit den Originaldaten. Außerdem wird der zweiseitige p-Wert der t-Approximation (Brunner & Munzel, 2000) bestimmt.

Das Makro wird im SAS-Programm-Editor durch den Befehl

```
%INCLUDE 'Pfad\PERM_BF.SAS';
```

eingebunden. Der Datensatz muss als SAS-Datei bereitstehen, also beispielsweise durch folgenden DATA-Step eingelesen werden:

DATA SAS-Name;
INPUT Gruppe Zielvariable;
$1 \quad X_{11}$
$\vdots \quad \vdots$
$1 \quad X_{1n_1}$
$2 \quad X_{21}$
$\vdots \quad \vdots$
$2 \quad X_{2n_2}$
;
RUN;

Aufgerufen wird das Makro im SAS-Programm-Editor mit dem Befehl

%PERM_BF(DATA=SAS-Name, VAR=Zielvariable, GROUP=Gruppe);

Mit zusätzlichen Optionen können folgende Einstellungen gemacht werden:

- die Anzahl der zufälligen Permutationen kann durch Setzen der Variable LOOP festgelegt werden, die standardmäßig den Wert 10'000 hat,
- die graphische Ausgabe kann für den Output auf dem SAS-Output-Fenster oder für den Output als HTML-Datei optimiert werden. Dies geschieht über den Wert der Variable OUTSTYLE, der 'HTML' oder irgendein anderer Wert sein kann. Standard ist die optimierte Ausgabe als HTML-Datei, wobei die Ergebnisse dabei auch im SAS-Output-Fenster angezeigt werden.

Ein Beispielaufruf, in dem die Anzahl der Permutationen auf 100'000 und der Output für das SAS-Output-Fenster optimiert wird, sieht wie folgt aus:

%PERM_BF(DATA=SAS-Name, VAR=Zielvariable, GROUP=Gruppe,
LOOP=100000, OUTSTYLE=' ');

Das Makro gibt die Gesamtanzahl der Beobachtungen N, sowie die Anzahl Beobachtungen in den beiden Gruppen n_1, n_2 aus. Außerdem den Schätzer für den relativen Effekt

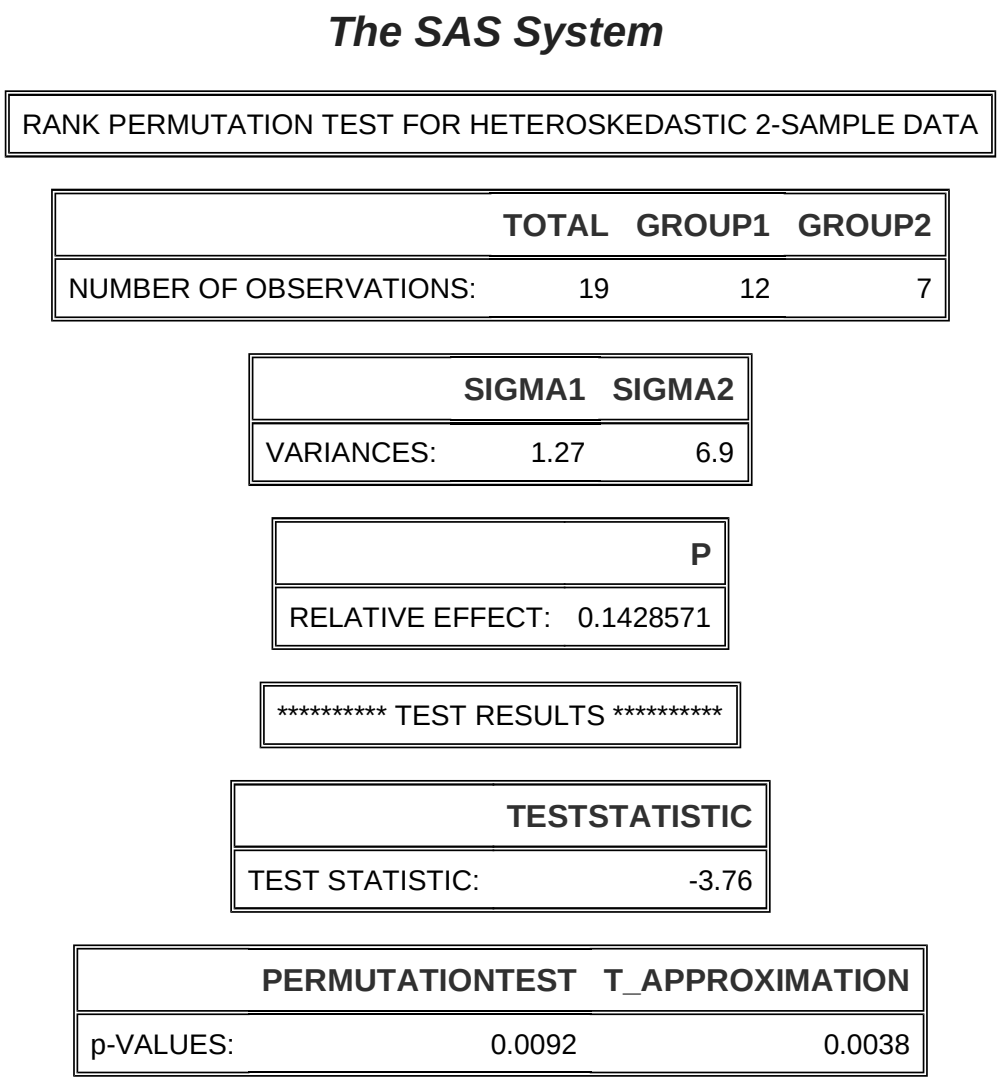
$$\widehat{p} = \frac{1}{N}(\overline{R}_{2\cdot} - \overline{R}_{1\cdot}) + \frac{1}{2}, \quad \text{wobei} \quad \overline{R}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik},$$

sowie die Werte der Varianzschätzer $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$ in beiden Gruppen und der Teststatistik T_N bei Verwendung der Originaldaten. Danach folgen die p-Werte für den Permutationstest und für die t-Approximation. Es wird SAS in Version 9.1 benötigt.

Anhand des Ferritin-Datensatzes aus Kapitel 3.6.1 soll die Verwendung des SAS-Makros demonstriert werden. Die Daten seien in der Datei ferritin.txt gespeichert. Die folgenden Befehle lesen den Datensatz ein und rufen das Makro auf:

```
DATA ferritin;
INFILE 'Pfad\ferrtin.txt';
INPUT treatment score;
RUN;
%PERM_BF(DATA=ferritin, VAR=score, GROUP=treatment);
```

Der HTML-Output ist in Abbildung B.1 dargestellt.



The SAS System

RANK PERMUTATION TEST FOR HETEROSKEDASTIC 2-SAMPLE DATA

	TOTAL	GROUP1	GROUP2
NUMBER OF OBSERVATIONS:	19	12	7

	SIGMA1	SIGMA2
VARIANCES:	1.27	6.9

	P
RELATIVE EFFECT:	0.1428571

********** TEST RESULTS **********

	TESTSTATISTIC
TEST STATISTIC:	-3.76

	PERMUTATIONTEST	T_APPROXIMATION
p-VALUES:	0.0092	0.0038

Abbildung B.1: Output des Makros PERM_BF für den Datensatz *ferritin*.

Die p-Werte beider Tests liegen unterhalb von 5%, so dass die Hypothese abzulehnen ist.

C SAS-Makro für robuste Konfidenzintervalle für den Shift-Effekt

Mit dem SAS-IML-Makro PERM_KI können für ein 2×2-Split-Plot-Design Konfidenzintervalle für einen Shift-Effekt zwischen den beiden unabhängigen Gruppen berechnet werden.

Ein 2×2-Split-Plot-Design liegt vor, wenn es für zwei unabhängige Gruppen pro Versuchseinheit zwei Wiederholungsmessungen gibt, zum Beispiel eine Baseline-Messung und eine Messung nach der Behandlung. Wir betrachten also die unabhängigen Zufallsvektoren $\mathbf{X}_{ik} = (X_{ik1}, X_{ik2})'$, $i = 1{,}2$, $k = 1, \ldots, n_i$ (vgl. Tabelle C.1) und das folgende lineare Modell

$$X_{ikj} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ikj},$$

wobei

- $\alpha_i,\ i = 1{,}2$ der Gruppeneffekt,
- $\beta_j,\ j = 1{,}2$ der Zeiteffekt,
- $(\alpha\beta)_{ij},\ i = 1{,}2,\ j = 1{,}2$ die Wechselwirkung zwischen Zeit und Gruppe und
- $\boldsymbol{\epsilon}_{ik} = (\varepsilon_{ik1}, \varepsilon_{ik2})',\ i = 1{,}2,\ k = 1, \ldots, n_i$ unabhängige Vektoren von Fehlern mit $\mathrm{E}(\boldsymbol{\epsilon}_{ik}) = \mathbf{0},\ Var(\boldsymbol{\epsilon}_{ik}) = \mathbf{V}_i$ sind.

Dabei lassen wir ausdrücklich verschiedene Varianzen $\mathbf{V}_i$ in den beiden Gruppen und unterschiedliche Varianzen τ_{ij}^2 zu den beiden Zeitpunkten zu.

Tabelle C.1: Split-Plot-Design

		Zeit	
		1	2
Gruppe	1	X_{111}	X_{112}
		$\vdots$	$\vdots$
		X_{1n_11}	X_{1n_12}
	2	X_{211}	X_{212}
		$\vdots$	$\vdots$
		X_{2n_21}	X_{2n_22}

Zunächst werden die Differenzen zwischen den Messwiederholungen berechnet:

$$\begin{aligned} D_{ik} &= X_{ik2} - X_{ik1} \\ &= \beta_2 - \beta_1 + (\alpha\beta)_{i2} - (\alpha\beta)_{i1} + \varepsilon_{ik2} - \varepsilon_{ik1} \\ &= \mu + \gamma_i + \eta_{ik}, \end{aligned}$$

wobei

- μ der Mittelwert,
- $\gamma_i,\ i = 1{,}2$ die Gruppeneffekte und
- $\eta_{ik},\ i = 1{,}2,\ k = 1,\ldots,n_i$ die unabhängigen Fehler bezüglich der Differenzen D_{ik} sind, mit Erwartungswert 0 und möglicherweise verschiedenen Varianzen $\sigma_i^2,\ i = 1{,}2$ in den beiden Gruppen.

Damit lautet das erweiterte Lokationsmodell für die Differenzen innerhalb einer Gruppe:

$$D_{ik} \overset{\text{u.i.v.}}{\sim} F_i(x) = F\left(\frac{x - \gamma_i}{\sigma_i}\right) \quad i = 1{,}2.$$

Voraussetzung für die Berechnung der robusten Konfidenzintervalle ist die Stetigkeit und Symmetrie der Verteilungsfunktionen F_i der Differenzen D_{ik2}. Außerdem muss F_i an seinem Erwartungswert μ_i invertierbar sein. Für den Shift-Effekt $\theta = \gamma_2 - \gamma_1$ zwischen den beiden Gruppen soll dann das Konfidenzintervall bestimmt werden. Ein robuster und asymptotisch erwartungstreuer Schätzer für θ ist der Hodges-Lehmann Schätzer $\widehat{\theta}$.

Das Makro berechnet einerseits Konfidenzintervalle nach Lehmann (1963) und andererseits nach der Methode von Bauer (1972). Dabei gibt es jeweils verschiedene Möglichkeiten für die Wahl der Quantile. Es werden insgesamt 5 verschiedene Intervalle bestimmt:

- Bauer-Konfidenzintervalle
  - mit Permutationsverteilungsquantilen aus Monte-Carlo-Simulationen: $\mathcal{I}_{\mathrm{B}}^{\mathrm{perm}}$
  - mit Normalverteilungsquantilen: $\mathcal{I}_{\mathrm{B}}^{\mathrm{norm}}$
  - mit t-Verteilungsquantilen: $\mathcal{I}_{\mathrm{B}}^{\text{t-approx}}$
- Hodges-Lehmann-Konfidenzintervalle
  - mit Normalverteilungsquantilen: $\mathcal{I}_{\mathrm{HL}}^{\mathrm{norm}}$
  - mit exakten Permutationsverteilungsquantilen: $\mathcal{I}_{\mathrm{HL}}^{\mathrm{exact}}$

Die Konfidenzintervalle nach Hodges-Lehmann sind allerdings nur für den Fall von homoskedastischen Gruppen gültig, das heißt wenn ein reines Lokationsmodell der Form $F_i(x) = F(x-\gamma_i), \quad i = 1{,}2$ vorliegt und sonst nur als Vergleich zu betrachten.

Das Makro wird im SAS-Programm-Editor durch den Befehl

%INCLUDE 'Pfad\PERM_KI.SAS';

eingebunden. Der Datensatz muss als SAS-Datei bereitstehen, also beispielsweise durch folgenden DATA-Step eingelesen werden:

DATA SAS-Name;
INPUT Gruppe Zeitpunkt Zielvariable;
$1 \quad 1 \quad X_{111}$
$1 \quad 2 \quad X_{112}$
$\vdots \qquad \vdots$
$1 \quad 1 \quad X_{1n_11}$
$1 \quad 2 \quad X_{1n_12}$
$2 \quad 1 \quad X_{211}$
$2 \quad 2 \quad X_{212}$
$\vdots \qquad \vdots$
$2 \quad 1 \quad X_{2n_21}$
$2 \quad 2 \quad X_{2n_22}$
;
RUN;

Aufgerufen wird das Makro im SAS-Programm-Editor mit dem Befehl

%PERM_KI(DATA = SAS-Name, VAR=Zielvariable, GROUP=Gruppe,
TIME=Zeitpunkt);

Mit zusätzlichen Optionen können folgende Einstellungen gemacht werden

- das Niveau der Intervalle kann durch die Variable ALPHA verändert werden, deren default-Einstellung 0.05 ist,

- die Anzahl der zufälligen Permutationen kann durch Setzen der Variable LOOP festgelegt werden, die standardmäßig den Wert 10'000 hat,
- die graphische Ausgabe kann für den Output auf dem SAS-Output-Fenster oder für den Output als HTML-Datei optimiert werden. Dies geschieht über den Wert der Variable OUTSTYLE, der 'HTML' oder irgendein anderer Wert sein kann. Standard ist die optimierte Ausgabe als HTML-Datei, wobei die Ergebnisse dabei zusätzlich im SAS-Output-Fenster angezeigt werden.

Ein Beispielaufruf,in dem das Niveau auf $\alpha = 0.01$, die Anzahl Permutationen auf 100'000 und der Output für das SAS-Output-Fenster optimiert wird, sieht wie folgt aus:

```
%PERM_BF(DATA=SAS-Name, VAR=Zielvariable, GROUP=Gruppe,
         TIME = Zeitpunkt, ALPHA = 0.01, LOOP=100000,
         OUTSTYLE=' ');
```

Das Makro gibt die Gesamtanzahl der Beobachtungen N, sowie die Anzahl der Beobachtungen in den beiden Gruppen n_1, n_2 aus. Außerdem den Hodges-Lehmann-Schätzer für den Shifteffekt

$$\widehat{\theta} = \text{Median}\{\Delta_{kl} = D_{2k} - D_{1l}, k = 1, \ldots, n_2,\ l = 1, \ldots, n_1\}$$

sowie den Wert der Varianzschätzer $\widehat{\sigma}_1^2,\ \widehat{\sigma}_2^2$ in beiden Gruppen, wobei

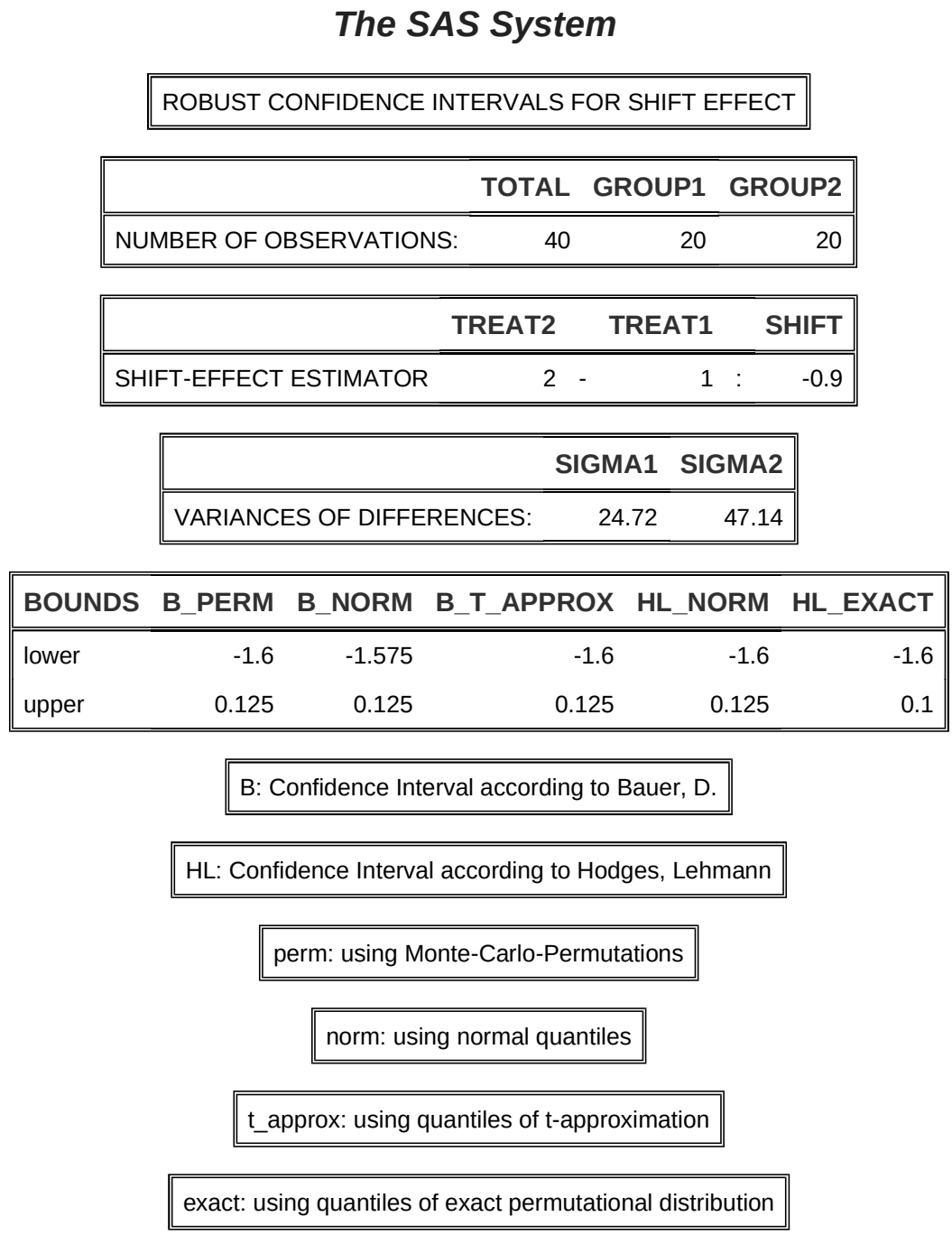
$$\widehat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(R_{ik} - R_{ik}^{(i)} - \overline{R}_{i.} + \frac{n_i + 1}{2} \right)^2,$$

wenn R_{ik} der Mittelrang der Beobachtung D_{ik} unter allen N Beobachtungen und $R_{ik}^{(i)}$ der interne Rang der Beobachtung D_{ik} unter allen n_i Beobachtungen aus Gruppe i ($i = 1{,}2,\ k = 1, \ldots, n_i$) ist. Danach folgen die Konfidenzintervallgrenzen für alle 5 genannten Konfidenzintervalle für θ. Es wird SAS in Version 9.1 benötigt.

Anhand der Daten zur Rückhärtung des Dentins (Wiegand *et al.* , 2005) soll die Verwendung des SAS-Makros demonstriert werden. Die Daten seien in der Datei dentin.txt gespeichert. Die folgenden Befehle lesen den Datensatz ein und rufen das Makro auf:

```
DATA dentin;
INFILE 'Pfad\dentin.txt';
INPUT rinsing time dentin;
RUN;
%PERM_KI(DATA=dentin, VAR=dentin, GROUP=rinsing, TIME=time);
```

Der HTML-Output ist in Abbildung C.1 dargestellt.



The SAS System

ROBUST CONFIDENCE INTERVALS FOR SHIFT EFFECT

	TOTAL	GROUP1	GROUP2
NUMBER OF OBSERVATIONS:	40	20	20

	TREAT2		TREAT1		SHIFT
SHIFT-EFFECT ESTIMATOR	2	-	1	:	-0.9

	SIGMA1	SIGMA2
VARIANCES OF DIFFERENCES:	24.72	47.14

BOUNDS	B_PERM	B_NORM	B_T_APPROX	HL_NORM	HL_EXACT
lower	-1.6	-1.575	-1.6	-1.6	-1.6
upper	0.125	0.125	0.125	0.125	0.1

B: Confidence Interval according to Bauer, D.

HL: Confidence Interval according to Hodges, Lehmann

perm: using Monte-Carlo-Permutations

norm: using normal quantiles

t_approx: using quantiles of t-approximation

exact: using quantiles of exact permutational distribution

Abbildung C.1: Output des Makros PERM_KI für den Datensatz *dentin*.

Alle 5 Konfidenzintervalle schließen die Null mit ein und lehnen also die Hypothese $H_0: \theta = 0$ nicht ab.

Literaturverzeichnis

Babu, G.J., & Padmanabhan, A.R. 2002. Resampling Methods For The Nonparametric Behrens-Fisher Problem. *Sankhyā, Series A*, **64**, 678–692.

Bauer, D.F. 1972. Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, **67**, 687–690.

Brunner, E., & Munzel, U. 2000. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory und a Small-Sample Approximation. *Biometrical Journal*, **42**, 17–25.

Brunner, E., & Munzel, U. 2002. *Nichtparametrische Datenanalyse - Unverbundene Stichproben.* Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Tokio: Springer.

Cantelli, F.P. 1933. Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, **4**, 421–424.

Chen, M., & Kianifard, F. 2000. A Nonparametric Procedure Associated with a Clinically Meaningful Efficacy Measure. *Biostatistics*, **1**, 293–298.

Cochran, W.B. 1964. Approximate Significance Levels of the Behrens-Fisher Test. *Biometrics*, **20**, 191–195.

Devroye, L., & Lugosi, G. 2001. *Combinatorial Methods in Density Estimation.* Springer.

Dwass, M. 1957. Modified Randomization Tests for Nonparametric Hypotheses. *Annals of Mathematical Statistics*, **28**, 181–187.

Eden, T., & Yates, B.A. 1933. On the Validity of Fisher's z Test when Applied to an Actual Example on Non-Normal Data. *Journal of Agricultural Science*, **23**, 6–17.

EMEA. 1998 (Feb.). *ICH Topic E 9 - Statistical Principles for Clinical Trials.* The European Agency for the Evaluation of Medical Products. CPMP/ICH/363/96.

Fligner, M.A., & Policello, G.E. 1981. Robust Rank Procedures for the Behrens-Fisher Problem. *Journal of the American Statistical Association*, **76**, 162–178.

Glivenko, V. 1933. Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, **4**, 92–99.

Hájek, J., & Ŝidák, Z. 1967. *Theory of Rank Tests.* New York: Academic Press.

Hodges, J.L., & Lehmann, E.L. 1963. Estimates of Location Based on Rank Tests. *Annals of Mathematical Statistics*, **34**, 598–611.

Janssen, A. 1997. Studentized Permutation Tests for Non-i.i.d. Hypotheses and the Generalized Behrens-Fischer Problem. *Statistics & Probability Letters*, **36**, 9–21.

Lehmann, E.L. 1963. Nonparametric Confidence Intervals for a Shift Parameter. *Annals of Mathematical Statistics*, **34**, 1507–1512.

Lumley, T. 1996. Generalized Estimating Equations for Ordinal Data: A Note on Working Correlation Structures. *Biometrics*, **52**, 354–361.

Mann, H.B., & Whitney, D.R. 1947. On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other. *Annals of Mathematical Statistics*, **18**, 50–60.

Moser, B.K., & Stevens, G.R. 1992. Homogeneity of Variance in the Two-Sample Means Test. *The American Statistican*, **46**, 19–21.

Neubert, K., & Brunner, E. 2006. *A Studentized Permutation Test for the Nonparametric Behrens-Fisher Problem.* submitted to Computational Statistics and Data Analysis.

Neuhaus, G. 1993. Conditional Rank Test for the Two-Sample Problem under Random Censorship. *Annals of Statistics*, **21**, 1760–1779.

PERM_BF. 2006. URL: http://www.ams.med.uni-goettingen.de.

PERM_KI. 2006. URL: http://www.ams.med.uni-goettingen.de.

Pesarin, F. 2001. *Multivariate Permutation Tests.* Chichester, New York, Weinheim, Brisbane, Singapore, Toronto: Wiley.

Pratt, J.W. 1964. Robustness of Some Procedures for the Two-Sample Location Problem. *Journal of the American Statistical Association*, **59**, 665–680.

Reiczigel, J., Zakariás, I., & Rózsa, L. 2005. A Bootstrap Test of Stochastic Equality of Two Populations. *The American Statistician*, **59**, 156–161.

Ruymgaart, F.H. 1980. A Unified Approach to the Asymptotic Distribution Theory of Certain Midrank Statstics. *In: Raoult, J.P. (Ed.), Statistique non Parametrique Asymptotique, Lecture Notes on Mathematics, Springer, Berlin*, **821**, 1–18.

Satterthwaite, F.E. 1946. An Approximate Distribution of Estimates of Variance Components. *Biometrical Bulletin*, **2**, 110–114.

Smith, H.F. 1936. The Problem of Comparing the Results of Two Experiments With Unequal Error. *Journal of the Council for Scientific and Industrial Research*, **9**, 211–212.

Stange, K. 1970. *Angewandte Statistik, Teil 1.* Berlin, Heidelberg, New York: Springer.

Streitberg, B., & Röhmel, J. 1986. Exact Distribution for Permutation and Rank Tests: An Introduction to some Recently Published Algorithms. *Statistical Software Newsletter*, **12**, 10–17.

Troendle, J.F. 2002. A Likelihood Ratio Test for the Nonparametric Behrens-Fisher Problem. *Biometrical Journal*, **44**, 813–824.

Welch, B.L. 1937. The Significance of the Difference Between Two Means When the Population Variances are Unequal. *Biometrika*, **29**, 350–362.

Wiegand, A., I., Müller, Schnapp, J.D., Werner, C., & T., Attin. 2005. *Impact of Fluoride, Milk and Water Rinsing on Surface Rehardening of Acid Softened Enamel and Dentin - An In-Situ-Study.* submitted to Caries Research.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics*, **1**, 80–83.

Lebenslauf

Persönliche Daten

Adresse	Karin Neubert Gosslerstr. 33a/99 37075 Göttingen Telefon: 05 51 - 50 06 55 62 E-Mail: karin.neubert@medizin.uni-goettingen.de
Geburt	02. August 1977 in Karl-Marx-Stadt
	ledig, deutsch

Schulische Ausbildung und Studium

1992 – 1996	Georgius-Agricola-Gymnasium Chemnitz, Leistungskurse in Mathematik und Physik, Abitur mit Gesamtnote 1,8
1997 – 2003	Studium der Mathematik an der Universität Heidelberg mit Vertiefungsgebiet Statistik und Nebenfach Volkswirtschaftslehre
Juli 1999	Diplomvorprüfung in Mathematik mit Gesamtnote „gut“
März 2003	Diplom in Mathematik mit Gesamtnote „sehr gut“
seit Mai 2003	Promotion in Medizinischer Statistik an der Universität Göttingen über „Das Behrens-Fisher-Problem: ein studentisierter Permutationstest und ein robustes Konfidenzintervall für den Shift-Effekt“, Doktorvater: Herr Prof. Dr. E. Brunner

seit Oktober 2003	Mitglied im Promotionsstudiengang „Angewandte Statistik und Empirische Methoden“ der Universität Göttingen

Praktische Erfahrungen und Tätigkeiten

März 1999 – März 2001	Beschäftigung als wissenschaftliche Hilfskraft an der Fakultät für Mathematik, Leitung von Übungsgruppen und Korrektur von Übungsaufgaben
Februar – April 2000	Praktikum bei der Deutschen Bank AG in Frankfurt am Main im Bereich Risk Management, Kalibrierung eines Ratingmodell für regulierte Fonds ohne Leverage
seit Mai 2003	Wissenschaftliche Angestellte in der Abteilung Medizinische Statistik des Universitätsklinikums Göttingen, Statistische Beratung und Analyse von medizinischen Studien