Nichtparametrische Modelle für faktorielle Diagnosestudien

Diplomarbeit

vorgelegt von

Katharina Lange

aus Detmold

angefertigt am
Institut für Mathematische Stochastik
der Georg-August-Universität Göttingen
2008

Danksagung

Für die Vergabe und die engagierte Betreuung meiner Diplomarbeit bedanke ich mich bei Herrn Prof. Dr. E. Brunner, der mich mit wertvollen Ratschlägen und Hinweisen bei der Anfertigung dieser Arbeit begleitete und durch die Bereitstellung der hervorragenden Arbeitsmöglichkeiten in der Abteilung Medizinische Statistik die Entstehung der Arbeit in der vorliegenden Form ermöglichte.
Außerdem gebührt mein Dank Herrn Prof. M. Denker, der trotz Zeitknappheit das Koreferat für die Arbeit übernommen hat.
Des Weiteren bedanke ich mich bei Herrn Frank Konietschke, der mir durch seine stete Diskussionsbereitschaft eine große Hilfe war und beim Korrekturlesen dieser Arbeit den einen oder anderen wertvollen Tipp für mich hatte. Bedanken möchte ich mich auch für die Hinweise der Korrekturleser Frau Inga Knorr und Herrn Christian Böge.
Für die –nicht nur finanzielle– Unterstützung während meines gesamten Studiums möchte ich meinen Eltern danken, insbesondere meiner Mutter, die mir bei der Beseitigung der sprachlichen Fehler dieser Arbeit eine besondere Hilfe war.

Göttingen, den 06.08.2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Evidenzbasierte Medizin „*ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten*“ (Sackett u. a., 1996). Im Rahmen dieser Disziplin bildet die biometrische Statistik eines der wichtigsten Werkzeuge medizinischer Forschung. Neben dem Vergleich verschiedener Therapiemethoden rückt zunehmend auch die Suche nach effizienten Verfahren zur Diagnose von Krankheiten in den Fokus, denn die Feststellung des Gesundheitszustandes eines Patienten bildet die Grundlage jeder weiteren klinischen Forschung. Daher ist es eine zentrale Aufgabe der Biometrie, valide Verfahren für die Auswertung von Diagnosestudien zu entwickeln.
Die Planung von Diagnosestudien sieht häufig vor, dass – insbesondere bei bildgebenden Diagnoseverfahren – mehrere Personen (sog. Reader) die erhobenen Daten auswerten und eine Diagnose stellen. In Abhängigkeit von der Frage, ob für die Auswertung verschiedener Diagnoseverfahren unterschiedliches Fachpersonal erforderlich ist, und von der ethischen Vertretbarkeit der Anwendung unterschiedlicher Verfahren am gleichen Patienten, entstehen dabei vier Konstellationsmöglichkeiten für Studien. Die hieraus resultierenden faktoriellen Designs sind Thema der vorliegenden Arbeit.
Die von Brunner u. a. (2002) erarbeitete Lösung zum sog. multivariaten nichtparametrischen Behrens-Fisher Problem bildet dabei das theoretische Fundament zur Auswertung der vier Studiendesigns: Bei den in Diagnosestudien untersuchten gesunden und kranken Patienten kann nicht davon ausgegangen werden, dass den erhobenen Messwerten in beiden Kollektiven die gleiche Verteilung zu Grunde liegt; insbesondere muss auch von einer Ungleichheit der Varianzen ausgegangen werden. Da außerdem die als Gütemaß eines diagnostischen Verfahrens herangezogene Fläche unter der ROC-Kurve dem durch die Mann-Whitney Statistik geschätzten relativen Effekt entspricht, liegt statistisch gesehen die Situation des nichtparametrischen Behrens-Fisher Problems vor. Lösungen hierzu findet man für den bivariaten Fall bei Brunner und Munzel (2000), für den multivariaten Fall bei Brunner u. a. (2002).
Somit bildet diese Arbeit die statistische Grundlage für die Auswertung bestimmter Diagnosestudien und könnte daher zukünftig einen wichtigen Beitrag für die medizinische Forschung leisten.

1.2 Aufbau der Arbeit

Die Arbeit beginnt mit einem Überblick über das Gebiet der Diagnosestudien: Die Verwendung des Gütemaßes AUC wird erklärt und das Zustandekommen der vier Studiendesigns erläutert. Im dritten Kapitel wird zunächst die Lösung des multivariaten nichtparametrischen Behrens-Fischer Problems nach Brunner u. a. (2002) vorgestellt und um ein Verfahren zur Berechnung bereichserhaltender Konfidenzintervalle erweitert. Dieses Lösungskonzept wird in den anschließenden Abschnitten an die vier faktoriellen Designs angepasst; dabei werden die Verteilungen von Schätzern hergeleitet, Statistiken (ANOVA- und Wald-Typ-Statistik) für das Testen standardmäßiger Hypothesen entwickelt und Konfidenzintervalle konstruiert. Eine kurze Vorstellung der im Rahmen dieser Arbeit entstandenen SAS-Makros erfolgt im nachfolgenden Kapitel. Schließlich werden in Kapitel 5 die Möglichkeiten und Grenzen der praktischen Anwendung der entwickelten Verfahren an Hand von Simulationen ausgelotet. Die Arbeit schließt mit einem Ausblick auf mögliche Erweiterungen und Entwicklungen der in dieser Arbeit dargelegten Konzepte.

2 Diagnosestudien

Bevor die Effizienz verschiedener diagnostischer Verfahren miteinander verglichen werden kann, muss die Frage geklärt werden, wie die Güte eines Diagnoseverfahrens zu einer messbaren, mathematisch handhabbaren Größe wird. Die Aufgabe eines diagnostischen Tests liegt darin, durch positive und negative Testergebnisse kranke und gesunde Patienten gesichterten Gesundheitszustandes möglichst gut zu unterscheiden. Dabei können folgende Fälle eintreten:

Tabelle 2.1: Vierfeldertafel möglicher diagnostischer Testergebnisse

	Patient krank	Patient gesund
Test positiv	richtig positiv (RP)	falsch positiv (FP)
Test negativ	falsch negativ (FN)	richtig negativ (RN)

In Tabelle 2.1 kann man erkennen, dass es zwei Werte gibt, die Einfluss auf die Größe, die die Güte eines diagnostischen Verfahren misst, haben sollten: der Anteil der korrekterweise als krank diagnostizierten Patienten an allen kranken Patienten, die sog. *Sensitivität* (#RP / (#RP + #FN)), sowie der Anteil der korrekterweise als gesund diagnostizierten Patienten an allen gesunden Patienten, die sog. *Spezifität* (#RN / (#RN + #FP)) (vgl. z.B. Werner, 2006).
Da nun der wahre Gesundheitszustand des Patienten unbekannt ist, wird dieser durch den Goldstandard ersetzt, der auf Grundlage des besten derzeit zur Verfügung stehenden Verfahrens ermittelt wird. Somit ist der Goldstandard das Maß, an dem sich die zu testenden Diagnoseverfahren orientieren, denn er tritt stellvertretend für den wahren Gesundheitszustand des Patienten ein. Zur Festsetzung des Goldstandards kann es erforderlich sein, mehrere bereits bekannte Verfahren zu kombinieren, posthum eine Biopsie durchzuführen oder auf invasive Verfahren zurückzugreifen. Dabei ist darauf zu achten, dass es bei der Ermittlung des Goldstandards zu einem Bias kommen kann, der z.B. auf die Unmöglichkeit der Durchführung einer Biopsie oder eines stark invasiven Verfahrens an gesunden Patienten zurückzuführen ist. Um dieses zu vermeiden, sollten bereits bei der Studienplanung die Möglichkeiten zur Festsetzung des Goldstandards gründlich geprüft werden.
In dieser Arbeit wird davon ausgegangen, dass der Goldstandard mit hoher Wahrscheinlichkeit den wahren Gesundheitszustand des Patienten widerspiegelt und somit als adäquater Maßstab für die Bestimmung der Güte anderer Verfahren dienen kann.

2.1 ROC-Kurven

Das Ergebnis X eines Diagnoseverfahrens ist häufig nicht dichotom (gesund/krank), sondern quantitativ – z.B. bei der Messung von Konzentrationen gewisser Substanzen im Blut – oder qualitativ – beispielsweise in Form von ordinalen Scores. Betrachtet man nun die Dichtefunktionen der Messergebnisse der gesunden Patienten und die Dichtefunktion der Messwerte der kranken Patienten, so beschreibt die Größe der Fläche, auf der sich beide Funktionen überlappen, die Fähigkeit des Diagnoseverfahrens, zwischen den beiden Gesundheitszuständen zu unterscheiden. Nach allgemeiner Konvention bezeichnet man einen Wert c als Schwellenwert, wenn ein Patient k als krank gilt, falls $X_k \geq c$ und als gesund, wenn $X_k < c$. Somit entsteht für jeden möglichen Schwellenwert c eine Vierfeldertafel wie in Tabelle 2.1, mit deren Hilfe Sensitivität und Spezifität in Abhängigkeit von c berechnet werden können. Steigt c an, so steigt auch die Sensitivität, wohingegen die Spezifität fällt, d.h. es gibt einen Zusammenhang zwischen Sensitivität und Spezifität in Abhängigkeit des Schwellenwertes c. Dieser wird durch die *Receiver Operation Characteristic Curve* (ROC-Kurve) dargestellt, welche auf der Abszisse 1-Spezifität und auf der Ordinate die Sensitivität darstellt.
Graphisch wird der Zusammenhang zwischen Dichtefunktion und ROC-Kurve in Abbildung 2.1 verdeutlicht.

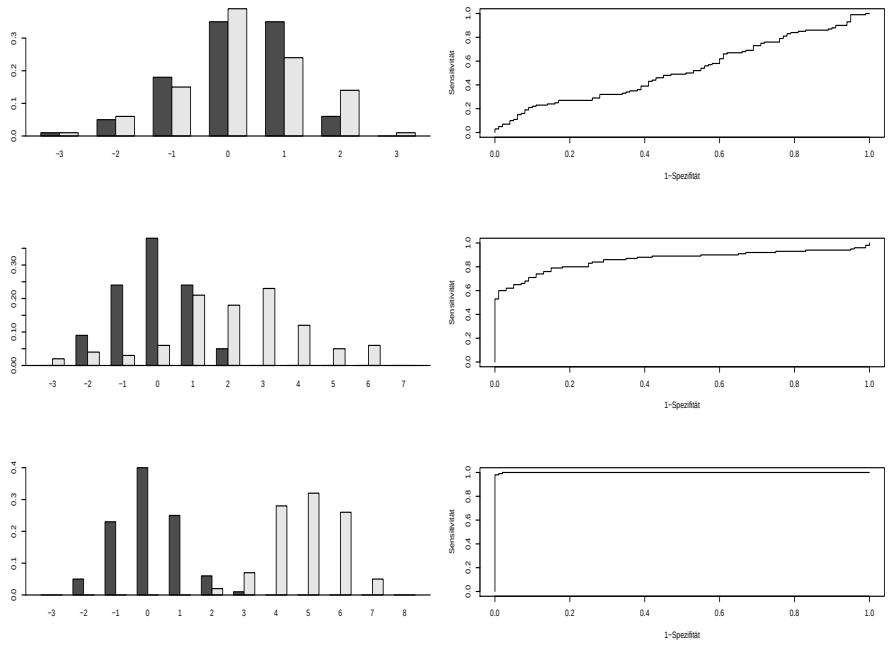

Abbildung 2.1: Zusammenhang zwischen Dichtefunktion und ROC-Kurve

In hellgrau ist das Kollektiv der kranken, in dunkelgrau das Kollektiv der gesunden Patienten dargestellt. Zu beachten ist dabei, dass, je weiter die Dichtefunktionen auseinander

liegen, desto größer ist der Abstand der ROC-Kurve zur Hauptwinkelhalbierenden.
Dieses Konzept der ROC-Kurven wurde ursprünglich im Bereich der Signalerkennung in der Nachrichtentechnik entwickelt, wo die ROC-Kurve den Zusammenhanghang zwischen korrekt und falsch erkannten Signalen darstellt (Peterson u. a., 1954).

2.2 Die AUC

Um nun die Güte eines Diagnoseverfahrens zu bestimmen, besteht die Möglichkeit, Sensitivität und Spezifität für einen festen Wert c zu betrachten. Wird der Schwellenwert hierbei allerdings falsch festgesetzt, so kann das Urteil über ein sonst gutes diagnostisches Verfahren negativ ausfallen. Daher wird als Index für die Treffsicherheit eines Diagnoseverfahrens die Fläche unter der ROC-Kurve (die sog. *Area Under the Curve* oder AUC) verwendet, denn bei diesem Gütemaß werden Sensitivität und Spezifität für alle möglichen Werte c berücksichtigt. Dieser Index hat außerdem eine anschauliche Interpretation: Er gibt die Wahrscheinlichkeit an, dass bei je einem zufällig gewählten kranken und gesunden Patienten $X_{krank} > X_{gesund}$ gilt, d.h. dass bei dem kranken Patienten ein größerer Wert als bei dem gesunden gemessen wird. Somit gibt die AUC Auskunft über die Trennschärfe (Accuracy) eines Diagnoseverfahrens über den gesamten Wertebereich. Berechnet wird die AUC dabei durch:

$$AUC = \int_{-\infty}^{\infty} F_{gesund}(x) dF_{krank}(x) = P(X_{gesund} < X_{krank}) + \frac{1}{2} P(X_{gesund} = X_{krank})$$

Damit entspricht die AUC dem durch die Mann-Whitney Statistik geschätzten relativen Effekt und ist hiermit ein unabhängig von der Verteilung der Messwerte definierter Index. Bereits Bamber (1975) und Hanley und McNeil (1982) erkannten durch diese Feststellung die Verknüpfung zwischen nichtparametrischer Statistik und der Theorie der ROC-Kurven.

2.3 Reader-Methoden-Kombinationen: die verschiedenen faktoriellen Designs

Diagnostische Studien dienen dem Zweck der Evaluation neuer Verfahren im Vergleich zu bekannten Methoden oder der Gegenüberstellung verschiedener, bereits existierender Verfahren, wie beipielsweise dem Vergleich

- von Röntgenaufnahmen mit und ohne Kontrastmittel,
- von Röntgenaufnahmen mit einem etablierten und einem neuen Kontrastmittel,
- eines neuen Verfahrens zur Feststellung von Karzinomen mit einer Standardmethode.

Dabei werden insbesondere bei bildgebenden Verfahren die Ergebnisse häufig von mehr als einem Reader ausgewertet, um die Qualität des Diagnoseverfahrens unabhängig von der des Readers beurteilen zu können. Bei dieser Vorgehensweise erhält man für jede Reader-Methoden-Kombination eine ROC-Kurve und somit pro Kombination eine AUC, mit deren Hilfe sich schließlich die Hypothesen formulieren lassen.

Nun kann es erforderlich sein, dass bei verschiedenen Diagnoseverfahren unterschiedliches Fachpersonal für das Stellen der Diagnose erforderlich ist, wodurch sich folgende Möglichkeiten ergeben:

- Verschiedene Verfahren werden von den gleichen Readern ausgewertet.
- Verschiedene Verfahren werden von verschiedenen Readern ausgewertet.

Des Weiteren kann es auf Grund ethischer Einwände oder der medizinischen Unvereinbarkeit zweier Verfahren erforderlich sein, dass für verschiedene Diagnoseverfahren unterschiedliche Patientenkollektive erforderlich sind, d.h. auch hier ergeben sich zwei Konstellationen:

- Verschiedene Verfahren werden an den gleichen Patienten getestet.
- Verschiedene Verfahren werden an unterschiedlichen Patienten getestet.

Durch diese Unterscheidungen ergeben sich nun vier mögliche faktorielle Designs.

2.3.1 Design 1

Design 1 betrachtet den einfachsten Fall, in welchem an allen Patienten alle Methoden getestet werden und jede Methode auch von jedem Reader ausgewertet werden kann. Die Ergebnisse eines solchen Designs lassen sich in folgender Tabelle übersichtlich darstellen (vgl. Brunner, 2002):

Tabelle 2.2: Schematische Darstellung Design 1

	Methode	1			2		
Zustand	Reader	1	2	3	1	2	3
gesund	Pat. 1	X	X	X	X	X	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Pat. n_0	X	X	X	X	X	X
krank	Pat. 1	X	X	X	X	X	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Pat. n_1	X	X	X	X	X	X
AUC		w_{11}	w_{12}	w_{13}	w_{21}	w_{22}	w_{23}

Dieses Design tritt dann auf, wenn es dem Patienten bedenkenlos zugemutet werden kann, mit allen Diagnosemethode untersucht zu werden. Dabei müssen die auswertenden Reader allerdings geschult sein, die Ergebnisse aller Verfahren zu bewerten. Ein Beispiel

hierzu wäre der Vergleich verschiedener Aufnahmegeschwindigkeiten und Auflösungen eines Gerätes (CT oder MRT).

2.3.2 Design 2

Die Annahme verschiedener Reader für verschiedene Diagnoseverfahren, aber gleicher Patienten für alle Verfahren führt zum zweiten faktoriellen Design, welches sich gleichermaßen als Tabelle darstellen lässt (vgl. ebenfalls Brunner, 2002).

Tabelle 2.3: Schematische Darstellung Design 2

	Methode	1			2		
Zustand	Reader	1	2	3	4	5	6
gesund	Pat. 1	X	X	X	X	X	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Pat. n_0	X	X	X	X	X	X
krank	Pat. 1	X	X	X	X	X	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Pat. n_1	X	X	X	X	X	X
AUC		w_{11}	w_{12}	w_{13}	w_{24}	w_{25}	w_{26}

Wie auch Design 1 tritt dieses Design auf, wenn es ethisch vertretbar ist, jeden Patienten mit allen Methoden zu untersuchen. Allerdings sind hier die auswertenden Ärzte der verschiedenen Methoden nicht die gleichen, wie es beispielsweise beim Vergleich einer MRT- mit einer Ultraschalluntersuchung der Fall sein kann.

2.3.3 Design 3

Das dritte Design liegt vor, wenn unterschiedliche Diagnoseverfahren, welche von den gleichen Readern ausgewertet werden, an verschiedenen Patienten getestet werden. Dieses lässt sich durch folgende Tabelle darstellen (vgl. Brunner, 2002):

Tabelle 2.4: Schematische Darstellung Design 3

Methode 1								Methode 2							
	gesund			krank					gesund			krank			
	Patient			Patient					Patient			Patient			
Reader	1	⋯	n_{01}	1	⋯	n_{11}	AUC	Reader	1	⋯	n_{02}	1	⋯	n_{12}	AUC
1	X	⋯	X	X	⋯	X	w_{11}	1	X	⋯	X	X	⋯	X	w_{21}
2	X	⋯	X	X	⋯	X	w_{12}	2	X	⋯	X	X	⋯	X	w_{22}
3	X	⋯	X	X	⋯	X	w_{13}	3	X	⋯	X	X	⋯	X	w_{23}

Ein Design dieser Form liegt beispielsweise dann vor, wenn verschiedene Kontrastmittel miteinander verglichen werden sollen, wobei diese auf Grund ethischer Einwände nicht an den gleichen Patienten getestet werden dürfen.

2.3.4 Design 4

Werden verschiedene Methoden an unterschiedlichen Patienten getestet und die einzelnen Verfahren dabei nicht von den gleichen Readern ausgewertet, liegt das vierte und letzte Design vor (vgl. auch hier Brunner, 2002):

Tabelle 2.5: Schematische Darstellung Design 4

Methode 1								Methode 2							
	gesund			krank					gesund			krank			
	Patient			Patient					Patient			Patient			
Reader	1	$\cdots$	n_{01}	1	$\cdots$	n_{11}	AUC	Reader	1	$\cdots$	n_{02}	1	$\cdots$	n_{12}	AUC
1	X	$\cdots$	X	X	$\cdots$	X	w_{11}	4	X	$\cdots$	X	X	$\cdots$	X	w_{24}
2	X	$\cdots$	X	X	$\cdots$	X	w_{12}	5	X	$\cdots$	X	X	$\cdots$	X	w_{25}
3	X	$\cdots$	X	X	$\cdots$	X	w_{13}	6	X	$\cdots$	X	X	$\cdots$	X	w_{26}

Der Vergleich eines Kontrastmittels für ein MRT mit einem Kontrastmittel für eine Ultraschalluntersuchung liefert ein Beispiel für eine derartige Versuchsstruktur.

3 Theorie

Im Folgenden werden die vier vorgestellten Versuchsanordungen statistisch modelliert; es werden Statistiken zum Testen standardmäßiger Hypothesen sowie Konfidenzintervalle für die AUCs hergeleitet. Hierbei wird ausgenutzt, dass die durch die AUC geschätzte Accuracy eines Diagnoseverfahrens dem durch die Mann-Whitney Statistik geschätzten relativen Effekt entspricht. Daher können die aus der nichtparametrischen Statistik bekannten Resultate für relative Effekte angewendet werden.
Dabei wird auf eine Einführung in die Standardnotationen aus den Bereichen Nichtparametrik und lineare Modelle an dieser Stelle verzichtet. Dieses wird zu Gunsten einer besseren Lesbarkeit in den Anhang verschoben (siehe Abschnitt A.1).

3.1 Das nichtparametrische Behrens-Fisher Problem

Da die Lösungskonzepte für die verschiedenen faktoriellen Designs bei Diagnosestudien auf der Lösung des multivariaten nichtparametrischen Behrens-Fisher Problems von Brunner u. a. (2002) basieren, werden zunächst die dort erarbeiteten Ergebnisse vorgestellt. Danach werden diese Resultate zur Herleitung der jeweiligen Ergebnisse in den vier faktoriellen Designs angewendet.

3.1.1 Modell und Notation

Voraussetzung 3.1
Gegeben seien $N = n_0 + n_1$ unabhängige Zufallsvektoren $\mathbf{X}_{i,k} = (X_{i,k}^{(1)}, ..., X_{i,k}^{(d)})'$, $X_{i,k}^{(l)} \sim F_i^{(l)}$, $i = 0, 1$, $k = 1, ..., n_i$, $l = 1, ..., d$, wobei die Funktionen $F_i^{(l)}$ beliebige Verteilungsfunktionen (in normalisierter Version) mit Ausnahme des trivialen Falles einer Ein-Punkt-Verteilung seien. Weiter gelte:

1. Für alle $l, r = 1, ..., d$ sei die bivariate Verteilung von $(X_{i,k}^{(l)}, X_{i,k}^{(r)})$ gleich für alle $k = 1, ..., n_i$ in der jeweiligen Gruppe $i = 0, 1$.
2. Für N gelte: $N \to \infty$, derart, dass $\frac{N}{n_i} \leq N_0 < \infty$, $i = 0, 1$.

Weiter sei $w^{(l)} = \int F_0^{(l)} \, dF_1^{(l)}$ der relative Effekt der l-ten Komponente ($l = 1, ..., d$) und $\hat{w}^{(l)} = \frac{1}{n_0} \left[\frac{1}{n_1} \sum_{k=1}^{n_1} R_{1,k}^{(l)} - \frac{n_0+1}{2} \right]$ der nach Satz A.2 konsistente Schätzer für $w^{(l)}$. (Hierbei sei $R_{1,k}^{(l)}$ der Rang von $X_{1,k}^{(l)}$ unter allen N Beobachtungen innerhalb der Komponente l.)

3.1.2 Asymptotische Verteilung des Schätzers

Satz 3.1 *[Asymptotischer Äquivalenzsatz] Unter Voraussetzung 3.1 gilt:*

$$\sqrt{N}\left(\hat{w}^{(l)} - w^{(l)}\right) \doteq \sqrt{N}\left(\frac{1}{n_1}\sum_{k=1}^{n_1} F_0^{(l)}(X_{1,k}^{(l)}) - \frac{1}{n_0}\sum_{k=1}^{n_0} F_1^{(l)}(X_{0,k}^{(l)}) + 1 - 2w^{(l)}\right) =: \sqrt{N}B^{(l)}$$

Beweis: siehe Brunner u. a. (2002) □

Sei nun $\mathbf{B} = (B^{(1)}, ..., B^{(d)})'$ und λ_{min} der kleinste Eigenwert von $Cov(\sqrt{N}\mathbf{B})$. Zusätzlich gelte folgende Voraussetzung:

Voraussetzung 3.2
Es sei $\lambda_{min} \geq \lambda_0 > 0$ eine beliebige Konstante.

Dann lässt sich zeigen:

Satz 3.2 *[Asymptotische Verteilung] Unter Voraussetzung 3.1 gilt:*

1. *Die Zufallsvariablen $\sqrt{N}B^{(l)}$, $l = 1, ..., d$ sind gleichmäßig beschränkt.*

Gilt zusätzlich Voraussetzung 3.2, so folgt außerdem:

2. *Die Statistik $\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w})$ ist asymptotisch multivariat normalverteilt mit Erwartungswert $\mathbf{0}$ und Kovarianzmatrix $\mathbf{V}_N = Cov(\sqrt{N}\mathbf{B})$.*

Beweis: siehe Brunner u. a. (2002) □

3.1.3 Schätzung der Kovarianzmatrix

Es sei

$$\begin{aligned}\hat{Y}_{0,k}^{(l)} &= \hat{F}_1^{(l)}(X_{0,k}^{(l)}) = \frac{1}{n_1}(R_{0,k}^{(l)} - R_{0,k}^{(l|0)}) \\ \hat{Y}_{1,k}^{(l)} &= \hat{F}_0^{(l)}(X_{1,k}^{(l)}) = \frac{1}{n_0}(R_{1,k}^{(l)} - R_{1,k}^{(l|1)})\end{aligned}$$

wobei $R_{i,k}^{(l)}$ den Rang von $X_{i,k}^{(l)}$ unter allen N Beobachtungen innerhalb der l-ten Komponente bezeichne, außerdem sei $R_{i,k}^{(l|i)}$ der Rang von $X_{i,k}^{(l)}$ unter den n_i Beobachtungen der i-ten Stichprobe innerhalb der l-ten Komponente ($i = 0, 1$ und $l = 1, ..., d$).
Weiter seien $\mathbf{R}_{i,k} = (R_{i,k}^{(1)}, ..., R_{i,k}^{(d)})'$ und $\mathbf{R}_{i,k}^{(i)} = (R_{i,k}^{(1|i)}, ..., R_{i,k}^{(d|i)})'$ für $i = 0, 1$. Außerdem seien

$$\overline{\mathbf{R}}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{R}_{i,k} \quad \text{und} \quad \overline{\mathbf{R}}_{i\cdot}^{(i)} = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{R}_{i,k}^{(i)} = \frac{n_i+1}{2}\mathbf{1}_d, \quad i = 0, 1.$$

Schließlich sei $\mathbf{Z}_{i,k} = \mathbf{R}_{i,k} - \mathbf{R}_{i,k}^{(i)}$ und $\overline{\mathbf{Z}}_{i\cdot} = \overline{\mathbf{R}}_{i\cdot} - \frac{1}{2}(n_i+1)\mathbf{1}_d$.

Satz 3.3 *[Schätzung der Kovarianzmatrix] Unter den obigen Voraussetzungen ist* $\hat{\mathbf{V}}_N = \hat{\mathbf{V}}_{N,0} + \hat{\mathbf{V}}_{N,1}$ *mit*

$$\hat{\mathbf{V}}_{N,i} = \frac{N}{(N-n_i)^2 n_i (n_i-1)} \sum_{k=1}^{n_i} (\mathbf{Z}_{i,k} - \overline{\mathbf{Z}}_{i\cdot})(\mathbf{Z}_{i,k} - \overline{\mathbf{Z}}_{i\cdot})' \qquad i = 0,1$$

ein konsistenter Schätzer für $\mathbf{V}_N$ *in dem Sinne, dass* $||\hat{\mathbf{V}}_N - \mathbf{V}_N||_2 \to 0$.

Beweis: siehe Brunner u. a. (2002) □

3.1.4 Teststatistiken

Um Hypothesen der Form $H_0 : \mathbf{Cw} = \mathbf{0}$ zu testen, wobei $\mathbf{C}$ eine geeignete Hypothesenmatrix im zu Grunde liegenden faktoriellen Modell sei, gibt es zwei auf quadratischen Formen basierende Teststatistiken – die Wald-Typ- und die ANOVA-Typ-Statistik:

Die Wald-Typ-Statistik

Satz 3.4 *[Wald-Typ-Statistik] Es gelten die Voraussetzungen 3.1 und 3.2, weiter sei* $\mathbf{C} \in \mathbb{R}^{(m \times d)}, m \in \mathbb{N}$*, dann gilt unter* $H_0 : \mathbf{Cw} = \mathbf{0}$*:*

$$Q_N^{WTS}(\mathbf{C}) = N \hat{\mathbf{w}}' \mathbf{C}' \left[\mathbf{C}\hat{\mathbf{V}}_N \mathbf{C}'\right]^+ \mathbf{C}\hat{\mathbf{w}} \quad \dot{\sim} \quad \chi^2_{rang(\mathbf{C})}$$

Beweis: analog zu Brunner u. a. (1999) □

In Simulationen zeigt sich nun, dass für die Konvergenz der Wald-Typ-Statistik ein großer Stichprobenumfang nötig ist, was Anlass dazu gibt, eine weitere Teststatistik vorzustellen.

Die ANOVA-Typ-Statistik

Satz 3.5 *[ANOVA-Typ-Statistik] Es gelten die Voraussetzungen von Satz 3.4, weiter sei* $\mathbf{T} = \mathbf{C}'(\mathbf{CC}')^- \mathbf{C}$*, dann gilt unter* $H_0 : \mathbf{Cw} = \mathbf{0}$*:*

$$Q_N^{ATS}(\mathbf{T}) = N \frac{Sp(\mathbf{T}\hat{\mathbf{V}}_N)}{Sp(\mathbf{T}\hat{\mathbf{V}}_N \mathbf{T}\hat{\mathbf{V}}_N)} \hat{\mathbf{w}}' \mathbf{T} \hat{\mathbf{w}} \quad \dot{\sim} \quad \chi^2_{\hat{f}} \quad \textit{mit} \quad \hat{f} = \frac{[Sp(\mathbf{T}\hat{\mathbf{V}}_N)]^2}{Sp(\mathbf{T}\hat{\mathbf{V}}_N \mathbf{T}\hat{\mathbf{V}}_N)}.$$

Beweis: Es gilt $\mathbf{Tw} = \mathbf{0} \Leftrightarrow \mathbf{Cw} = \mathbf{0}$, denn $\mathbf{Tw} = (\mathbf{C}'(\mathbf{CC}')^- \mathbf{C})\mathbf{w} = \mathbf{0}$ ist gerade die Standardformulierung der Hypothese $\mathbf{Cw} = \mathbf{0}$. D.h. unter $H_0 : \mathbf{Cw} = \mathbf{0}$ gilt nach Satz 3.2

$$\sqrt{N}\mathbf{T}\hat{\mathbf{w}} \dot{\sim} N(\mathbf{0}, \mathbf{T}'\mathbf{V}_N\mathbf{T}).$$

Nach dem Satz über die Verteilung einer Quadratform (vgl. Satz A.4) gilt somit

$$Q_N^*(\mathbf{T}) = N(\mathbf{T}\hat{\mathbf{w}})'\mathbf{T}\mathbf{T}\hat{\mathbf{w}} = N\hat{\mathbf{w}}'\mathbf{T}\hat{\mathbf{w}} \quad \dot{\sim} \quad \sum_{i=1}^{d} \lambda_i U_i^2,$$

wobei U_i standardnormalverteilte Zufallsvariablen sind und die λ_i die Eigenwerte von $\mathbf{T}'\mathbf{V}_N\mathbf{T}$.
$U = \sum_{i=1}^{d} \lambda_i U_i^2$ wird nun durch eine mit g gestreckte χ_f^2-Verteilung approximiert und zwar derart, dass die ersten beiden Momente von U und der $g \cdot \chi_f^2$-Verteilung gleich sind, d.h.

$$Sp(\mathbf{T}\mathbf{V}_N) = \sum_{i=1}^{d} \lambda_i = E(U) \quad \stackrel{!}{=} \quad E(g \cdot \chi_f^2) = g \cdot f$$

$$Sp(\mathbf{T}\mathbf{V}_N\mathbf{T}\mathbf{V}_N) = 2 \cdot \sum_{i=1}^{d} \lambda_i^2 = Var(U) \quad \stackrel{!}{=} \quad Var(g \cdot \chi_f^2) = 2 \cdot g^2 f$$

Daraus folgt für f und g

$$f = \frac{Sp(\mathbf{T}\mathbf{V}_N)^2}{Sp(\mathbf{T}\mathbf{V}_N\mathbf{T}\mathbf{V}_N)} \qquad g = \frac{Sp(\mathbf{T}\mathbf{V}_N\mathbf{T}\mathbf{V}_N)}{Sp(\mathbf{T}\mathbf{V}_N)};$$

damit gilt

$$\frac{f}{g \cdot f} \cdot N\hat{\mathbf{w}}'\mathbf{T}\hat{\mathbf{w}} \quad \dot{\sim} \quad \chi_f^2. \tag{3.1}$$

Setzt man in Gleichung 3.1 die entsprechenden Werte für f und g ein und ersetzt $\mathbf{V}_N$ nun durch seinen konsistenten Schätzer $\hat{\mathbf{V}}_N$, so gilt nach dem Satz von SLUTZKY (vgl. Satz A.3):

$$Q_N^{ATS}(\mathbf{T}) = N\frac{Sp(\mathbf{T}\hat{\mathbf{V}}_N)}{Sp(\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T}\hat{\mathbf{V}}_N)}\ \hat{\mathbf{w}}'\mathbf{T}\hat{\mathbf{w}} \quad \dot{\sim} \quad \chi_{\hat{f}}^2 \quad \text{mit} \quad \hat{f} = \frac{[Sp(\mathbf{T}\hat{\mathbf{V}}_N)]^2}{Sp(\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T}\hat{\mathbf{V}}_N)}$$

□

Die Idee der Approximation einer Verteilung durch Gleichsetzen der ersten beiden Momente geht auf Box (1954) zurück, die Verwendung dieser Teststatistik in der Nichtparametrik auf Brunner u. a. (1997). Zahlreiche Niveausimulationen haben gezeigt, dass insbesondere bei kleinen Stichproben die ANOVA-Typ-Statistik der Wald-Typ-Statistik überlegen ist, weswegen diese vorzuziehen ist.

3.1.5 Konfidenzintervalle

Nach den Sätzen 3.2 und 3.3 gilt

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \dot{\sim} \quad N(\mathbf{0}, \hat{\mathbf{V}}_N).$$

Hieraus ergeben sich durch Anwendung der Pivotmethode auf die studentisierten Teststatistiken $t^{(l)} = \sqrt{N}\left(\hat{w}^{(l)} - w^{(l)}\right) / \sqrt{\hat{v}_N^{(l,l)}}$ die klassischen Konfidenzintervallgrenzen

$$\hat{w}_{klass,U}^{(l)} = \hat{w}^{(l)} - \frac{\sqrt{\hat{v}_N^{(l,l)}}u_{1-\alpha/2}}{\sqrt{N}} \qquad \hat{w}_{klass,O}^{(l)} = \hat{w}^{(l)} + \frac{\sqrt{\hat{v}_N^{(l,l)}}u_{1-\alpha/2}}{\sqrt{N}}, \tag{3.2}$$

wobei $\hat{v}_N^{(l,l)}$ das l-te Diagonalelement von $\hat{\mathbf{V}}_N$ bezeichne (vgl. z.B. Brunner, 2005).
Wählt man diese klassische Variante zur Berechnung der Konfidenzintervalle, so treten zwei Probleme auf:

1. Es gilt $w^{(l)} \in [0,1]$ für alle $l = 1,...,d$; für das auf klassische Art und Weise geschätzte Konfidenzintervall KI_{klass} muss aber nicht gelten $KI_{klass} \subseteq [0,1]$. In Anlehnung an Efron und Tibshirani (1993, Kapitel 13.6) bezeichnet man diese Konfidenzintervalle als nicht bereichserhaltend.
2. Die empirische Varianz $\hat{v}^{(l,l)}$ des Schätzers für den relativen Effekt $\hat{w}^{(l)}$ ist 0, wenn $\hat{w}^{(l)} \in \{0,1\}$, d.h. in diesem Fall hat das Konfidenzintervall die Länge 0.

Die Problematik, dass die klassischen Konfidenzintervalle nicht immer bereichserhaltend sind, ist bereits in der Arbeit von Domhof (2001) aufgegriffen, welcher als Lösung die sog. Transformationsmethode vorschlägt. Diese ist ein Verfahren, bei welchem mit Hilfe stetiger Transformationen und unter Anwendung der δ- Methode (siehe Satz A.7) Statistiken modifiziert werden. Die Grundidee der Transformationsmethode beruht auf folgendem Prinzip: Durch eine Abbildung $g : (0,1) \to \mathbb{R}$ wird das offene Einheitsintervall auf die Menge aller reellen Zahlen abgebildet. Unter

Voraussetzung 3.3

1. g sei streng monoton steigend,
2. g sei differenzierbar,
3. g sei bijektiv,

ist $\sqrt{N}\,(g(\hat{\mathbf{w}}) - g(\mathbf{w}))$ ebenfalls asymptotisch normalverteilt. Es sei $\tilde{w}^l := g(\hat{w}^l)$, dann lässt sich das transformierte Konfidenzintervall $\left[\tilde{w}_{g;U}^{(l)}, \tilde{w}_{g;O}^{(l)}\right]$ für $g(w^{(l)})$ klassisch wie in Gleichung 3.2 berechnen. Aus der Monotonie und Umkehrbarkeit von g folgt schließlich, dass $g^{-1}\left(\left[\tilde{w}_{g;U}^{(l)}, \tilde{w}_{g;O}^{(l)}\right]\right) = \left[\hat{w}_{g;U}^{(l)}, \hat{w}_{g;O}^{(l)}\right]$ wieder ein Intervall ist, und da $g^{-1}(\mathbb{R}) = (0,1)$, ist dieses Intervall ein bereichserhaltendes Konfidenzintervall für $w^{(l)}$.
Graphisch wird die Vorgehensweise bei der Transformationsmethode in Abbildung 3.1 (in Anlehnung an Konietschke, 2006) verdeutlicht.

$$
\begin{array}{cccccc}
(0,1) & \ni & w^{(l)} & \xrightarrow{g} & g(w^{(l)}) & \in \mathbb{R} \\
 & & & & \downarrow & \\
(0,1) & \supset & [\hat{w}_{g,U}^{(l)}, \hat{w}_{g,O}^{(l)}] & \xleftarrow{g^{-1}} & [\tilde{w}_{g,U}^{(l)}, \tilde{w}_{g,O}^{(l)}] & \subset \mathbb{R}
\end{array}
$$

Abbildung 3.1: Darstellung der Transformationsmethode

Da die beobachteten Werte für $\hat{w}^{(l)}$ immer zwischen 0 und 1 liegen, ist die Verteilungsfunktion von $\hat{w}^{(l)}$ für sehr hohe oder sehr niedrige Werte von w^l nicht symmetrisch um

$E(\hat{w}^{(l)}) = w^{(l)}$. Neben der Bereichserhaltung ist ein weiterer Vorteil der Transformationsmethode, dass durch Wahl einer geeigneten Transformationsfunktion g die Dichtefunktion des transformierten Schätzers symmetrischer ist als die nicht transformierte Dichtefunktion. Somit ist die Normalapproximation der transformierten Verteilungsfunktion von $g(\hat{w}^{(l)})$ besser als die Approximation der ursprünglichen Verteilungsfunktion von $\hat{w}^{(l)}$ an die Normalverteilung. Abbildung 3.2 verdeutlicht diesen Zusammenhang.

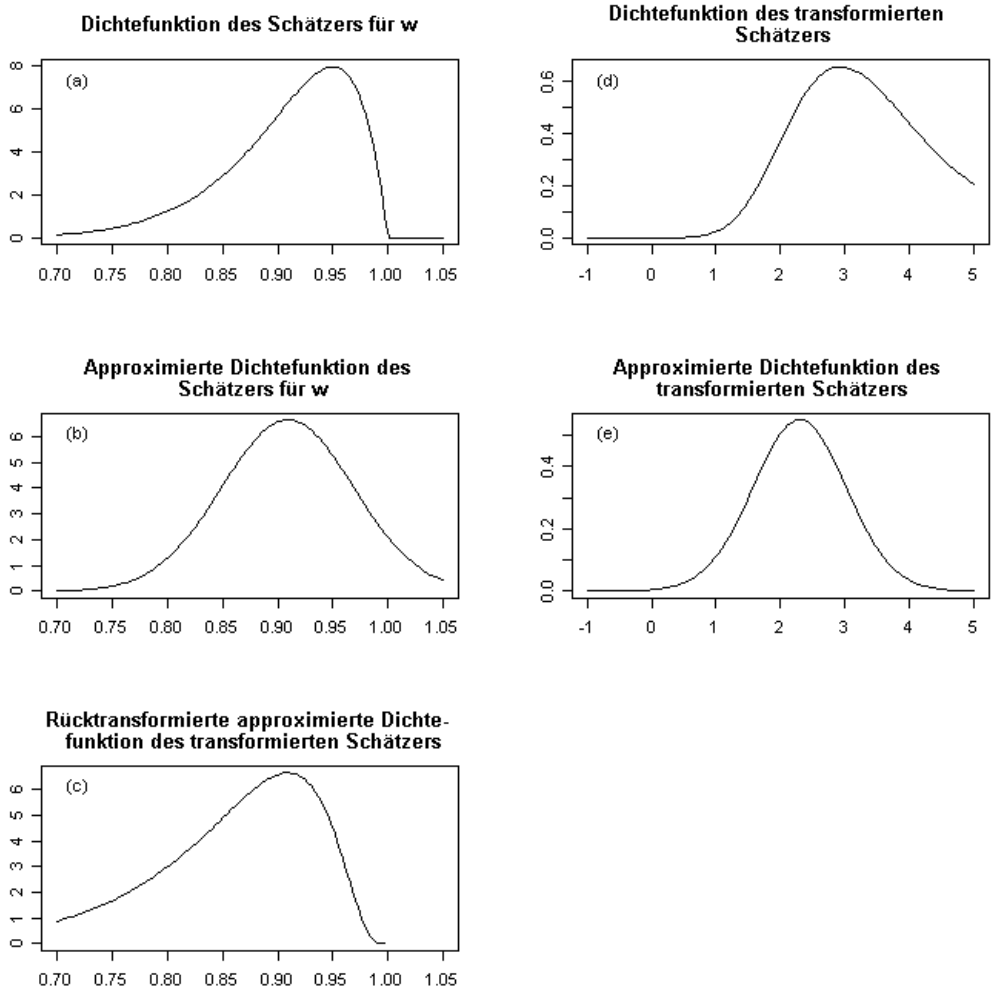

Abbildung 3.2: Dichtefunktion, transformierte Dichtefunktion und deren Approximationen

Zwei in der Statistik häufig verwendeten Transformationsfunktionen sind die logit-Funktion und die probit-Funktion. Die auf Grund ihrer guten Handhabbarkeit zu großer Popularität gelangte logit-Funktion wird in dieser Arbeit verwendet:

$$\begin{aligned} logit : (0,1) &\rightarrow \mathbb{R} \\ x &\mapsto \log\left(\frac{x}{1-x}\right), \end{aligned}$$

deren Ableitung und Umkehrfunktion

$$logit'(x) \quad = \quad \frac{1}{x(1-x)}$$

und

$$logit^{-1}(x) = \quad expit(x) \quad = \frac{\exp(x)}{\exp(x)+1}$$

sind. Hieraus ergeben sich mit

$$\begin{aligned} \tilde{w}^{(l)} &= logit(\hat{w}^{(l)}) \\ \tilde{w}^{(l)}_{logit,U} &= \tilde{w}^{(l)} - \frac{\sqrt{\hat{v}^{(l,l)}_N} \cdot u_{1-\alpha/2}}{logit'(\hat{w}^{(l)})\sqrt{N}} \\ \tilde{w}^{(l)}_{logit,O} &= \tilde{w}^{(l)} + \frac{\sqrt{\hat{v}^{(l,l)}_N} \cdot u_{1-\alpha/2}}{logit'(\hat{w}^{(l)})\sqrt{N}} \end{aligned}$$

nun die logit-Konfidenzintervalle (vgl. Brunner und Munzel, 2002, Abschnitt 2.1.4):

$$\hat{w}^{(l)}_{logit,U} = logit^{-1}(\tilde{w}^{(l)}_{logit,U}) \quad \text{und} \quad \hat{w}^{(l)}_{logit,O} = logit^{-1}(\tilde{w}^{(l)}_{logit,O}) \tag{3.3}$$

Sollen die klassischen Konfidenzintervalle ihre Coverageprobability einhalten, so sind – wie Simulationen zeigen – sehr große Stichprobenumfänge nötig. Die logit-transformierten Konfidenzintervalle halten das Niveau schon bei deutlich geringen Stichprobenumfängen ein, sind allerdings konservativ. Anhand von Abbildung 3.2 lässt sich eine mögliche Ursache für die Nachteile der Verfahren festmachen. Während Grafik (a) die unbekannte tatsächliche Dichtefunktion des Schätzers darstellt, zeigt Abbildung (b) die Dichtefunktion, auf deren Grundlage die klassischen Konfidenzintervalle berechnet werden. In Abbildung (c) ist die Dichtefunktion, die durch Rücktransformation der approximierten transformierten Dichtefunktion entstanden ist, dargestellt, welche die Grundlage zur Berechnung der Konfidenzintervalle durch die Transformationsmethode darstellt. Deutlich zu erkennen ist, dass bei der Dichtefunktion in (b), im Vergleich zur tatsächlichen Dichtefunktion, zu viel Verteilungsmasse rechts vom Erwartungswert liegt, d.h. die klassischen Konfidenzintervalle liegen zu weit rechts. Bei der transformierten Dichtefunktion in (c) dreht sich das Bild, und die logit-Konfidenzintervalle liegen somit zu weit links. Legt man andere tatsächliche Verteilungsfunktionen für den Schätzer des relativen Effektes zu Grunde, ergeben sich ähnliche Bilder, was Anlass dazu gibt, ein Verfahren zu entwickeln, welches beide Methoden miteinander kombiniert:

Satz 3.6 *Seien X_N, Y_N, Z_N ($N \in \mathbb{N}$) Folgen von Zufallsvariablen, für die gilt:*

1. $\sqrt{N}(X_N - \phi) \doteq \sqrt{N}Y_N \sim F_N$
2. $\sqrt{N}(X_N - \phi) \doteq \sqrt{N}Z_N \sim G_N$

Weiter sei $KI_{F_N} = [c_U^{F_N}, c_O^{F_N}]$ *ein mit Hilfe der Verteilungsfunktion* F_N *konstruiertes asymptotisches* $(1-\alpha)$*-Konfidenzintervall für* ϕ*, wobei* KI_{F_N} *symmetrisch sei, in dem Sinne dass* $P(\phi > c_O^{F_N}) = P(\phi < c_U^{F_N}) = \frac{\alpha}{2}$.
$KI_{G_N} = [c_U^{G_N}, c_O^{G_N}]$ *sei ein entsprechendes mit Hilfe von* G_N *konstruiertes Konfidenzintervall.*
Unter diesen Voraussetzungen ist

$$KI_{neu} \quad = \quad \left[(c_U^{F_N} + c_U^{G_N})/2, (c_O^{F_N} + c_O^{G_N})/2\right]$$

ebenfalls ein asymptotisches $(1-\alpha)$*-Konfidenzintervall für* ϕ*.*

Beweis: Es gilt $\sqrt{N}(X_N - \phi) \doteqdot \sqrt{N}Y_N$ und $\sqrt{N}(X_N - \phi) \doteqdot \sqrt{N}Z_N$. Daraus folgt $\sqrt{N}Y_N \doteqdot \sqrt{N}Z_N$. Daher konvergieren die Verteilungen von $\sqrt{N}Y_N$ und $\sqrt{N}Z_N$ schwach gegeneinander (vgl. Bauer, 2002, S.35 ff.), d.h.

$$\lim_{N\to\infty} F_N(x) \quad = \quad \lim_{N\to\infty} G_N(x).$$

Das bedeutet, dass die beiden Verteilungen asymptotisch gleich sind. Daher sind auch die Konfidenzintervalle asymptotisch gleich, in dem Sinne, dass $P(\sqrt{N}|c_U^{F_N} - c_U^{G_N}| > 0) \to 0$ für $N \to \infty$ (analog für die obere Grenze) . Damit ist nun auch jede Konvexkombination von Intervallgrenzen asymptotisch gleich, also ist KI_{neu} ebenfalls ein asymptotisches $(1-\alpha)$-Konfidenzintervall. □

Hieraus folgt, dass eine Kombination wie in Satz 3.6 der mit Hilfe der Transformationsmethode berechneten Konfidenzintervalle und der klassischen Konfidenzintervalle asymptotisch ebenfalls ein Konfidenzintervall für $w^{(l)}$ ist. Problematisch bleibt, dass diese Konfidenzintervalle nicht bereichserhaltend sein müssen. Simulationen zeigen jedoch, dass in diesem Fall einseitige Intervalle eine gute Alternative darstellen, sodass man schließlich folgende Konfidenzintervalle enthält:
Sei $KI_{klass} = [\hat{w}_{klass,U}^{(l)}, \hat{w}_{klass,O}^{(l)}]$ das nach Gleichung 3.2 berechnete klassische Konfidenzintervall, außerdem seien $KI_{logit} = [\hat{w}_{logit,U}^{(l)}, \hat{w}_{logit,O}^{(l)}]$, $KI_{logit\ links} = [\hat{w}_{logit\ links,U}^{(l)}, 1)$ und $KI_{logit\ rechts} = (0, \hat{w}_{logit\ rechts,O}^{(l)}]$ die nach Gleichung 3.3 mit Hilfe der logit-Transformation berechneten ein- und zweiseitigen $(1-\alpha)$-Konfidenzintervalle. Zur Vereinfachung der Notation sei $\hat{w}_{kombi,U}^{(l)} = \frac{\hat{w}_{klass,U}^{(l)} + \hat{w}_{logit,U}^{(l)}}{2}$, sowie $\hat{w}_{kombi,O}^{(l)} = \frac{\hat{w}_{klass,O}^{(l)} + \hat{w}_{logit,O}^{(l)}}{2}$, dann ist

$$KI_{neu} \quad = \quad \begin{cases} \left[\hat{w}_{kombi,U}^{(l)}, \hat{w}_{kombi,O}^{(l)}\right] & : \quad \hat{w}_{kombi,U}^{(l)} \geq 0,\ \hat{w}_{kombi,O}^{(l)} \leq 1 \\ \left(0, \hat{w}_{logit\ rechts,O}^{(l)}\right] & : \quad \hat{w}_{kombi,U}^{(l)} < 0 \\ \left[\hat{w}_{logit\ links,U}^{(l)}, 1\right) & : \quad \hat{w}_{kombi,O}^{(l)} > 1 \end{cases} \tag{3.4}$$

nach Satz 3.6 ein asymptotisches, bereichserhaltendes $(1-\alpha)$-Konfidenzintervall für $w^{(l)}$.

Zu behandeln bleibt noch das angesprochene zweite Problem, wenn $\hat{w}^{(l)} \in \{0,1\}$. In diesem Fall kann keine Transformation durchgeführt werden, da die logit-Funktion nur auf dem offenen Einheitsintervall definiert ist. Berechnet man das einseitige, klassische Konfidenzintervall für $\hat{w}^{(l)}$, wobei man den in diesem Fall den Wert 0 annehmenden Varianzschätzer $\hat{v}_N^{(l,l)}$ durch $\hat{\sigma}^2 = \max_{l=1,\dots d} \hat{v}_N^{(l,l)}$ ersetzt, so erhält man ein echtes Konfidenzintervall mit Länge ungleich 0. Diese Vorgehensweise ist legitim, denn für $n \to \infty$ wird $\hat{w}^{(l)}$ nur dann 0 oder 1, wenn $w^{(l)}$ den Wert 0 oder 1 hat, und daher tritt diese Problematik asymptotisch nicht mehr auf.
Simulationen zeigen, dass die so konstruierten Konfidenzintervalle etwas kürzer als die logit-Konfidenzintervalle sind und dabei schon bei kleinen Stichprobenumfängen das Niveau besser einhalten, weswegen sie den einfachen logit-Konfidenzintervallen vorzuziehen sind (vgl. Abschnitt 5.5).

Nachdem nun die Lösung des nichtparametrischen Behrens-Fisher Problems skizziert ist, kann die Anwendung der hier dargestellten Resultate auf die vier faktoriellen Designs bei Diagnosestudien erfolgen. Sofern nicht anders vermerkt, ist dabei in den folgenden Abschnitten die Notation an die hier eingeführte angelehnt.

3.2 Design 1

Im ersten Design werden alle Patienten mit allen Diagnoseverfahren untersucht, und alle Verfahren werden von allen Readern ausgewertet.

3.2.1 Modell und Notation

Es sei $X_{i,k}^{(s,r)}$ das Ergebnis des s-ten $(s = 1,\dots,S)$ Diagnoseverfahrens ausgewertet vom r-ten Reader $(r = 1,\dots,R)$, erhoben am k-ten Patienten $(k = 1,\dots,n_i)$, welcher gesund $(i = 0)$ oder krank $(i = 1)$ (nach Goldstandard) ist. Es bezeichne weiter $\mathbf{X}_{i,k} = (X_{i,k}^{(1,1)},\dots,X_{i,k}^{(1,R)},\dots,X_{i,k}^{(S,1)},\dots,X_{i,k}^{(S,R)})'$ den Vektor der Messergebnisse des k-ten Patienten.
Damit ergibt sich folgendes Modell:

Voraussetzung 3.4
Seien $\mathbf{X}_{i,k} = (X_{i,k}^{(1,1)},\dots,X_{i,k}^{(S,R)})$, $X_{i,k}^{(s,r)} \sim F_i^{(s,r)}$ $N = n_0 + n_1$ unabhängige Zufallsvektoren, wobei die $F_i^{(s,r)}$ beliebige Verteilungen (in normalisierter Version) mit Ausnahme der Ein-Punkt-Verteilung seien. Weiter fordert man:

1. Für alle $s,t = 1,\dots,S$ und alle $r,u = 1,\dots,R$ sei die bivariate Verteilung von $(X_{i,k}^{(s,r)}, X_{i,k}^{(t,u)})$ gleich für alle $k = 1,\dots,n_i$ in der jeweiligen Gruppe $i = 0,1$.
2. Für N gelte, $N \to \infty$, derart, dass $\frac{N}{n_i} \leq N_0 < \infty$, $i = 0,1$.

Die Bedingung, dass die $F_i^{(s,r)}$ keine Ein-Punkt-Verteilungen seien, besagt, dass es bei der Messung/Erhebung der $X_{i,k}^{(s,r)}$, $k = 1, ..., n_i$ eine Varianz gibt, d.h. dass nicht alle $X_{i,k}^{(s,r)}$ den gleichen Wert haben.
Die erste Voraussetzung bedeutet, dass die Abhängigkeit, die zwischen zwei Ergebnissen besteht, die am selben Patienten erhoben wurden, die gleiche bei allen Patienten ist, welches die Annahme widerspiegelt, dass unterschiedliche Patienten gleichen Gesundheitszustandes Messwiederholungen sind. Die zweite Voraussetzung stellt sicher, dass die Stichprobenumfänge im Kollektiv der Gesunden und der Kranken nicht zu unterschiedlich sind.
Eine Übersicht über die Notation befindet sich in Tabelle 3.1.

Tabelle 3.1: Notationsübersicht Design 1

	Methode	1			2			
Zustand	Reader	1	2	3	1	2	3	
gesund	Pat. 1	$(X_{0,1}^{(1,1)},$	$X_{0,1}^{(1,2)},$	$X_{0,1}^{(1,3)},$	$X_{0,1}^{(2,1)},$	$X_{0,1}^{(2,2)},$	$X_{0,1}^{(2,3)})$	$= \mathbf{X}'_{0,1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Pat. n_0	$(X_{0,n_0}^{(1,1)},$	$X_{0,n_0}^{(1,2)},$	$X_{0,n_0}^{(1,3)},$	$X_{0,n_0}^{(2,1)},$	$X_{0,n_0}^{(2,2)},$	$X_{0,n_0}^{(2,3)})$	$= \mathbf{X}'_{0,n_0}$
krank	Pat. 1	$(X_{1,1}^{(1,1)},$	$X_{1,1}^{(1,2)},$	$X_{1,1}^{(1,3)},$	$X_{1,1}^{(2,1)},$	$X_{0,1}^{(2,2)},$	$X_{1,1}^{(2,3)})$	$= \mathbf{X}'_{1,1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Pat. n_1	$(X_{1,n_1}^{(1,1)},$	$X_{1,n_1}^{(1,2)},$	$X_{1,n_1}^{(1,3)},$	$X_{1,n_1}^{(2,1)},$	$X_{1,n_1}^{(2,2)},$	$X_{1,n_1}^{(2,3)})$	$= \mathbf{X}'_{1,n_1}$
AUC		$(w_{11},$	$w_{12},$	$w_{13},$	$w_{21},$	$w_{22},$	$w_{23})$	$= \mathbf{w}'$

3.2.2 Asymptotische Verteilung des Schätzers und Schätzung der Kovarianzmatrix

Definiert man eine bijektive Abbildung $\phi : \{1, \ldots, S\} \times \{1, \ldots, R\} \to \{1, \ldots, S \cdot R\}$, $(s, r) \mapsto l$, so erkennt man, dass das Modell dem nichtparametrischen Behrens-Fisher Problem nach Brunner u. a. (2002) entspricht, und da die Voraussetzungen 3.4 den Voraussetzungen 3.1 entsprechen, sind die Ergebnisse von Brunner u. a. (2002) anwendbar, d.h. für den Vektor der Accuracies $\mathbf{w} = (w^{(1,1)}, ..., w^{(S,R)})'$ – der dem Vektor der relativen Effekte im nichtparametrischen Behrens-Fisher Problem entspricht – gilt nach Satz 3.1:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \doteqdot \quad \sqrt{N}\mathbf{B},$$

wobei $\mathbf{B}$ wie in Satz 3.1 definiert ist und $\hat{\mathbf{w}}$ – wie im nichtparametrischen Behrens-Fisher Problem – der konsistente Schätzer für $\mathbf{w}$ ist. Fordert man zusätzlich,

Voraussetzung 3.5
das Minimum der Eigenwerte λ_{min} von $Cov(\sqrt{N}\mathbf{B})$ ist durch eine Konstante $\lambda_0 > 0$ nach unten beschränkt, d.h. $\lambda_{min} \geq \lambda_0 > 0$,

so folgt nach Satz 3.2:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \dot{\sim} \quad N(\mathbf{0}, \hat{\mathbf{V}}_N),$$

wobei $\hat{\mathbf{V}}_N$ der nach Satz 3.3 konsistente Schätzer für $\mathbf{V}_N$ ist.
Sei nun $\mathbf{C} \in \mathbb{R}^{m \times SR}$, $m \in \mathbb{N}$ eine Hypothesenmatrix, dann gilt unter $H_0 : \mathbf{Cw} = \mathbf{0}$:

$$\sqrt{N}\mathbf{C}\hat{\mathbf{w}} \quad \dot{\sim} \quad N(\mathbf{0}, \mathbf{C}\hat{\mathbf{V}}_N\mathbf{C}').$$

3.2.3 Hypothesen und Teststatistiken

Wie in faktoriellen Designs üblich, werden auch hier die Hypothesen in Matrixnotation formuliert. Eine Übersicht der zu diesem Zweck verwendeten Matrizenoperationen sowie der Definitionen und Bezeichnungen für speziell im Zusammenhang mit faktoriellen Designs auftretende Matrizen befindet sich im Anhang A.1.

In diesem Fall liegt ein gekreuztes Design der beiden Faktoren Reader und Methode vor, somit lassen sich Hypothesen auf keinen Methodeneffekt, auf keinen Readereffekt und auf keinen Wechselwirkungseffekt testen.
Zur Vereinfachung der Schreibweise wird folgende Notation eingeführt:

$$\overline{w}^{(s,\cdot)} = \frac{1}{R}\sum_{r=1}^{R} w^{(s,r)}, \quad \overline{w}^{(\cdot,r)} = \frac{1}{S}\sum_{s=1}^{S} w^{(s,r)}, \quad \overline{w} = \frac{1}{S \cdot R}\sum_{r=1}^{R}\sum_{s=1}^{S} w^{(s,r)}$$

Einfluss des Faktors Methode
Zu beantworten ist die Frage, ob sich die Methoden in ihrer Accuracy unterscheiden. Diese Fragestellung wird durch die Hypothese $\mathrm{H}_0^M : \overline{w}^{(s,\cdot)} = \overline{w}^{(t,\cdot)}\ \forall s,t = 1,\ldots S,\ s \neq t$ bzw. die hierzu äquivalente Hypothese $\mathrm{H}_0^M : \overline{w}^{(s,\cdot)} - \overline{w} = 0\ \forall s = 1,\ldots S$ mathematisch dargestellt. In vektorieller Schreibweise wird diese Hypothese wie folgt formuliert:

$$\mathrm{H}_0^M : \mathbf{C}_M\mathbf{w} = \mathbf{0}, \quad \text{wobei} \quad \mathbf{C}_M = \mathbf{P}_S \otimes \frac{1}{R}\mathbf{1}_R'$$

Einfluss des Faktors Reader
An dieser Stelle wird die Frage nach der Gleichheit der diagnostischen Fähigkeiten aller Reader durch die Hypothese $\mathrm{H}_0^{RD} : \overline{w}^{(\cdot,r)} - \overline{w} = 0\ \forall r = 1,\ldots R$ gestellt, welche sich vektoriell analog zu der des Faktors Methode darstellen lässt (hierbei sei RD die Kurzschreibweise für den Faktor Reader):

$$\mathrm{H}_0^{RD} : \mathbf{C}_{RD}\mathbf{w} = \mathbf{0}, \quad \text{wobei} \quad \mathbf{C}_{RD} = \frac{1}{S}\mathbf{1}_S' \otimes \mathbf{P}_R$$

Einfluss der Wechselwirkung
Zu klären ist, ob die Qualität der Beurteilungen der Reader homogen in allen Methoden

ist, welches durch die Hypothese $\mathsf{H}_0^{M\times RD}: w^{(s,r)} - \overline{w}^{(s,\cdot)} - \overline{w}^{(\cdot,r)} + \overline{w} = 0 \ \forall s = 1,\ldots S$, $r = 1,\ldots R$ bzw. die äquivalente vektoriell formulierte Hypothese

$$\mathsf{H}_0^{M\times RD}: \mathbf{C}_{M\times RD}\mathbf{w} = \mathbf{0}, \quad \text{wobei} \quad \mathbf{C}_{M\times RD} = \mathbf{P}_S \otimes \mathbf{P}_R,$$

getestet wird. Wendet man die Sätze 3.4 und 3.5 auf diese Hypothesenmatrizen an, so erhält man die entsprechenden Wald-Typ- bzw. ANOVA-Typ-Statistiken zum Testen der Hypothesen.

3.2.4 Konfidenzintervalle

Die Anwendung der in Abschnitt 3.1.5 vorgestellten Methode zur Berechnung der Konfidenzintervalle für den relativen Effekt liefert in diesem Fall als Ergebnis die Konfidenzintervalle für die Accuracy einer einzelnen Reader-Methoden-Kombination.
In einigen Fällen interessieren zusätzlich zu Konfidenzintervallen für die einzelnen Accuracies auch die Konfidenzintervalle für bestimmte Linearkombinationen $\mathbf{c}'\mathbf{w}$ der Trennschärfen. Da $\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w})$ asymptotisch $N(\mathbf{0}, \hat{\mathbf{V}}_N)$-verteilt ist, folgt:

$$\sqrt{N}\mathbf{c}'(\hat{\mathbf{w}} - \mathbf{w}) \stackrel{\cdot}{\sim} N(\mathbf{0}, \mathbf{c}'\hat{\mathbf{V}}_N\mathbf{c}), \tag{3.5}$$

wobei $\mathbf{c} \in \mathbb{R}^{R\cdot S}$ ein beliebiger Kontrastvektor sei. Durch Anwendung der Resultate aus Abschnitt 3.1.5 auf dieses Ergebnis erhält man für das Konfidenzintervall von $\mathbf{c}'\mathbf{w}$ die klassischen Konfidenzintervallgrenzen

$$\mathbf{c}'\hat{w}^{(l)} \mp \frac{\sqrt{\mathbf{c}'\hat{\mathbf{V}}_N\mathbf{c}} \cdot u_{1-\alpha/2}}{\sqrt{N}}, \tag{3.6}$$

sowie unter Verwendung des Satzes von CRAMER (siehe Satz A.6) die Grenzen für das logit-Konfidenzintervall

$$expit\left(logit(\mathbf{c}'\hat{w}^{(l)}) \mp \frac{\sqrt{\mathbf{c}'\mathbf{D}_{logit}\hat{\mathbf{V}}_N\mathbf{D}'_{logit}\mathbf{c}} \cdot u_{1-\alpha/2}}{\sqrt{N}}\right), \tag{3.7}$$

wobei $\mathbf{D}_{logit} = diag\left(\frac{1}{(\hat{w}^{(s,r)}(1-\hat{w}^{(s,r)})}\right)_{s=1,\ldots S,\ r=1,\ldots R}$ die Jacobimatrix der partiellen Ableitungen der logit-Funktion in vektorieller Form ist. Eine Kombination beider Intervalle, wie in Abschnitt 3.1.5 präsentiert, ist hier ebenfalls möglich.

Konfidenzintervalle für die mittlere Accuracy der Methoden

Wenn keine Interaktion vorliegt, so ist es zur Darstellung und Einordnung der Relevanz statistisch signifikanter Ergebnisse sinnvoll, Konfidenzintervalle für die mittlere Accuracy der einzelnen Methoden zu berechnen: Bezeichnet $\mathbf{e}_s$ den s-ten S-dimensionalen Einheitsvektor, so erhält man durch Verwendung des Kontrastvektors $\mathbf{c}_s = \mathbf{e}_s \otimes \frac{1}{R}\mathbf{1}_R$ und unter

Anwendung der oben genannten Methode ein Konfidenzintervall für die s-te Methode gemittelt über die R Reader.

Konfidenzintervalle für die Differenz zweier Methoden
Wählt man $\mathbf{c}_{s;t} = (\mathbf{e}_s - \mathbf{e}_t) \otimes \frac{1}{R}\mathbf{1}_R$, so erhält man ein Konfidenzintervall für die Differenz der mittleren Accuracies der Methoden s und t. Bei Paarvergleichen dieser Art ist jedoch darauf zu achten, dass zur Einhaltung des multiplen Niveaus eine Adjustierung (beispielsweise nach BONFERRONI) der zur Berechnung benötigten Quantile vorgenommen werden muss.

3.3 Design 2

Im zweiten Design werden alle Diagnoseverfahren an allen Patienten getestet, allerdings ist nun für die Auswertung der Resultate unterschiedliches Fachpersonal erforderlich, d.h. eine Methode s wird nur von den ihr zugeordneten Readern $r(s)$, $r(s) = 1,...,R_s$ ($s = 1,...,S$) ausgewertet. Diese Anordnung führt zu folgendem statistischen Modell.

3.3.1 Modell und Notation

Es sei $X_{i,k}^{(s,r(s))}$ das Ergebnis des s-ten $(s = 1,...,S)$ Diagnoseverfahrens ausgewertet vom $r(s)$-ten Reader $(r(s) = 1,...,R_s)$ der s-ten Methode, erhoben am k-ten Patienten $(k = 1,...,n_i)$, welcher gesund $(i = 0)$ oder krank $(i = 1)$ ist. Es sei weiter $\mathbf{X}_{i,k} = (X_{i,k}^{(1,1)},...,X_{i,k}^{(1,R_1)},...,X_{i,k}^{(S,1)},...,X_{i,k}^{(S,R_S)})'$, der Vektor der Diagnoseergebnisse des k-ten Patienten.
Damit ergibt sich folgendes Modell:

Voraussetzung 3.6
Seien $\mathbf{X}_{i,k} = (X_{i,k}^{(1,1)},...,X_{i,k}^{(S,R_S)})'$, $X_{i,k}^{(s,r(s))} \sim F_i^{(s,r(s))}$ $N = n_0 + n_1$ unabhängige Zufallsvektoren, wobei die $F_i^{(s,r)}$ beliebige Verteilungen (in normalisierter Version) mit Ausnahme der Ein-Punkt-Verteilung seien. Weiter fordert man:

1. Für alle $s,t = 1,...,S$ und alle $r(s) = 1,...,R_s,\ u(t) = 1,...,R_t$ sei die bivariate Verteilung von $(X_{i,k}^{(s,r(s))}, X_{i,k}^{(t,u(t))})$ gleich für alle $k = 1,...,n_i$ in der jeweiligen Gruppe $i = 0,1$.

2. Für N gelte, $N \to \infty$, derart, dass $\frac{N}{n_i} \leq N_0 < \infty$, $i = 0,1$.

Diese Forderungen sind analog zu denen im ersten Design zu interpretieren.

Übersichtlich stellt Tabelle 3.2 die Notation dar.

Tabelle 3.2: Notationsübersicht Design 2

	Methode	1			2			
Zustand	Reader	1	2	3	4	5	6	
gesund	Pat. 1	$(X_{0,1}^{(1,1)},$	$X_{0,1}^{(1,2)},$	$X_{0,1}^{(1,3)},$	$X_{0,1}^{(2,4)},$	$X_{0,1}^{(2,5)},$	$X_{0,1}^{(2,6)})$	$=\mathbf{X}'_{0,1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Pat. n_0	$(X_{0,n_0}^{(1,1)},$	$X_{0,n_0}^{(1,2)},$	$X_{0,n_0}^{(1,3)},$	$X_{0,n_0}^{(2,4)},$	$X_{0,n_0}^{(2,5)},$	$X_{0,n_0}^{(2,6)})$	$=\mathbf{X}'_{0,n_0}$
krank	Pat. 1	$(X_{1,1}^{(1,1)},$	$X_{1,1}^{(1,2)},$	$X_{1,1}^{(1,3)},$	$X_{1,1}^{(2,4)},$	$X_{0,1}^{(2,5)},$	$X_{1,1}^{(2,6)})$	$=\mathbf{X}'_{1,1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Pat. n_1	$(X_{1,n_1}^{(1,1)},$	$X_{1,n_1}^{(1,2)},$	$X_{1,n_1}^{(1,3)},$	$X_{1,n_1}^{(2,4)},$	$X_{1,n_1}^{(2,5)},$	$X_{1,n_1}^{(2,6)})$	$=\mathbf{X}'_{1,n_1}$
AUC		$(w_{11},$	$w_{12},$	$w_{13},$	$w_{24},$	$w_{25},$	$w_{26})$	$=\mathbf{w}'$

3.3.2 Asymptotische Verteilung des Schätzers und Schätzung der Kovarianzmatrix

Es sei $R = \sum_{s=1}^{S} R_s$ die Gesamtanzahl der Reader. Da bei jedem Patienten jeder Reader genau ein Diagnoseergebnis auswertet bzw. erhebt, gilt also $\mathbf{X}_{i,k} \in \mathbb{R}^R$ ($i = 0,1$, $k = 1,...,n_i$).
Definiert man nun eine bijektive Abbildung $\phi : \{(s,r(s)) : s = 1,...,S, r(s) = 1,...,R_s\} \to \{1,...,R\}$, $(s,r(s)) \mapsto l$, so kann man analog zur Lösung von Design 1 auch hier die Lösung durch einfache Umstrukturierung des Index auf die Lösung des nichtparametrischen Behrens-Fisher Problems zurückführen. D.h. nach den Satz 3.1 und 3.2 gilt

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad = \quad \sqrt{N}\mathbf{B} \stackrel{\cdot}{\sim} N(\mathbf{0}, \mathbf{V}_N),$$

wenn die Regularitätsvoraussetzung

Voraussetzung 3.7
$\lambda_{min} \geq \lambda_0 > 0$, wobei λ_{min} das Minimum der Eigenwerte von $Cov(\sqrt{N}\mathbf{B})$ sei,

erfüllt ist. Es folgt weiter:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \stackrel{\cdot}{\sim} \quad N(\mathbf{0}, \hat{\mathbf{V}}_N),$$

wobei die Berechnung des Schätzers $\hat{\mathbf{V}}_N$ analog zu der Berechnung von $\hat{\mathbf{V}}_N$ im nichtparametrischen Behrens-Fisher Problem (Satz 3.3) durchgeführt wird und $\mathbf{B}$ wie in Satz 3.2 definiert ist.

3.3.3 Hypothesen und Teststatistiken

Diesem faktoriellen Design liegt ein zweifaktorieller hierarchischer Versuchsplan (CRHF-RD(M)) zu Grunde. Somit lassen sich Hypothesen auf keinen Haupteffekt des Faktors Methode (M) und keinen Subkategorieeffekt des Faktors Reader (RD(M)), der unter dem Faktor Methode verschachtelt ist, testen. Es sei

$$\overline{w}^{(s,\cdot)} = \frac{1}{R_s} \sum_{r(s)=1}^{R_s} w^{(s,r(s))} \quad \overline{w} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r(s)=1}^{R_s} w^{(s,r(s))}.$$

Einfluss des Faktors Methode

Es stellt sich die Frage, ob die Methoden alle die gleiche Accuracy haben, d.h getestet wird $\mathsf{H}_0^M : \overline{w}^{(s,\cdot)} = \overline{w}^{(t,\cdot)}\ \forall s,t = 1,\dots S,\ s \neq t$ bzw. die äquivalente Hypothese $\mathsf{H}_0^M : \overline{w}^{(s,\cdot)} - \overline{w} = 0$, $\forall s = 1,\dots S$, welche vektoriell durch

$$\mathsf{H}_0^M : \mathbf{C}_M \mathbf{w} = \mathbf{0} \quad \text{wobei} \quad \mathbf{C}_M = \mathbf{P}_S \cdot \bigoplus_{s=1}^{S} \frac{1}{R_s} \mathbf{1}'_{R_s}$$

dargestellt wird. Zu beachten ist hierbei jedoch, dass die Accuracy einer einzelnen Methode immer mit den diagnostischen Fähigkeiten ihrer jeweiligen Reader zusammenhängt, daher sollte schon bei der Planung einer Studie darauf geachtet werden, dass die Reader auf ihrem jeweiligen Fachgebiet vergleichbare Fähigkeiten aufweisen. Sonst kann es zu einem Bias bei der Schätzung des Methodeneffekts kommen, welcher darauf zurückzuführen ist, dass in diesem Design ein Wechselwirkungseffekt zwischen Reader und Methode nicht modelliert werden kann.

Einfluss des Faktors Reader(Methode)

Beantwortet wird hierbei die Frage, ob es einen Unterschied in den Accuracies gibt, der nicht auf den Einfluss des Kategorieeffektes Methode zurückzuführen ist. Dabei wird die Hypothese $\mathsf{H}_0^{RD(M)} : w^{(s,r(s))} - \overline{w}^{(s,\cdot)} = 0\ \forall r(s) = 1,...,R_s,\ s = 1,...,S$ bzw. die entsprechende Hypothese in Matrixnotation

$$\mathsf{H}_0^{RD(M)} : \mathbf{C}_{RD(M)} \mathbf{w} = \mathbf{0}, \quad \text{wobei} \quad \mathbf{C}_{RD} = \bigoplus_{s=1}^{S} \mathbf{P}_{R_s},$$

getestet. Analog zum ersten Design erhält man durch Anwendung der Sätze 3.4 und 3.5 die Teststatistiken zum Testen der Hypothesen sowie deren Verteilungen.

3.3.4 Konfidenzintervalle

Die in Abschnitt 3.1.5 hergeleiteten Konfidenzintervalle für die relativen Effekte beim nichtparametrischen Behrens-Fisher Problem entsprechen in diesem Design den Konfidenzintervallen für die Accuracy eines einzelnen Readers (verschachtelt unter dem Faktor Methode).

Auch hier kann es interessant sein, Konfidenzintervalle für gewisse Linearkombinationen $\mathbf{c}'\mathbf{w}$ zu betrachten. Die Konfidenzintervallgrenzen für eine solche Linearkombination entsprechend denen im ersten Design (vgl. Gleichung 3.6 und 3.7), sodass gelegentlich die Kontrastvektoren angepasst werden müssen.

Konfidenzintervalle für die mittlere Accuracy der Methoden
Auch in diesem Design kann es sinnvoll sein, Konfidenzintervalle für die mittlere Accuracy der einzelnen Methoden zu berechnen. Dabei sollten jedoch auch hier obige Anmerkungen über einen möglichen Bias beim Schätzen der Methodeneffekte beachtet werden. Der in diesem Fall verwendete Kontrastvektor zur Berechnung der mittleren Accuracy der Methode ist $\mathbf{c}'_s = (0 \cdot \mathbf{1}'_{R_1}, ..., 0 \cdot \mathbf{1}'_{R_{s-1}}, \frac{1}{R_s} \cdot \mathbf{1}'_{R_s}, 0 \cdot \mathbf{1}'_{R_{s+1}}, ..., 0 \cdot \mathbf{1}'_{R_S})$.

Konfidenzintervalle für die Differenz zweier Methoden
Der Kontrastvektor zur Berechnung des Konfidenzintervalls der Differenz zweier Methoden ist $\mathbf{c}'_{s;t} = (0 \cdot \mathbf{1}'_{R_1}, ..., \frac{1}{R_s} \cdot \mathbf{1}'_{R_s}, ..., -\frac{1}{R_t} \cdot \mathbf{1}'_{R_t}, ..., 0 \cdot \mathbf{1}'_{R_S})$. Auch hier ist, wie bei den entsprechenden Konfidenzintervallen im ersten Design, auf die Einhaltung des multiplen Niveaus zu achten.

3.4 Design 3

Im dritten Design werden wieder alle diagnostischen Verfahren von allen Readern ausgewertet, allerdings werden die verschiedenen Methoden an unterschiedlichen Patientenkollektiven getestet.

3.4.1 Modell und Notation

Es sei $X_{i,k}^{(s,r)}$ das Ergebnis des s-ten $(s = 1, ..., S)$ Diagnoseverfahrens ausgewertet vom r-ten Reader $(r = 1, ..., R)$, erhoben am k-ten Patienten $(k = 1, ..., n_{is})$, welcher gesund $(i = 0)$ oder krank $(i = 1)$ ist. Es sei weiter $\mathbf{X}_{i,k}^{(s)} = (X_{i,k}^{(s,1)}, ..., X_{i,k}^{(s,R)})'$, der Vektor der Ergebnisse des k-ten Patienten, an welchem das s-te Diagnoseverfahren getestet wurde.
In dieser Notation erhält man folgendes statistische Modell:

Voraussetzung 3.8
Für alle Methoden $s = 1, ..., S$ seien $\mathbf{X}_{i,k}^{(s)} = (X_{i,k}^{(s,1)}, ..., X_{i,k}^{(s,R)})'$, $X_{i,k}^{(s,r)} \sim F_i^{(s,r)}$ $N_s = n_{0s} + n_{1s}$ unabhängige Zufallsvektoren, wobei die $F_i^{(s,r)}$ beliebige Verteilungen (in normalisierter Version) mit Ausnahme der Ein-Punkt-Verteilung seien. Weiter gelte:

1. Bei allen Methoden $s = 1, ..., S$ sei für alle $r, u = 1, ..., R$ die bivariate Verteilung von $(X_{i,k}^{(s,r)}, X_{i,k}^{(s,u)})$ gleich für alle $k = 1, ..., n_{is}$ in der jeweiligen Gruppe $i = 0, 1$.

2. Für die Stichprobenumfänge $N_s = n_{0s} + n_{1s}$ und $N = \sum_{s=1}^{S} N_s$ gelte

   (a) $N \to \infty$, derart, dass $\frac{N}{N_s} \leq M_0 < \infty$ für alle $s = 1, ..., S$.
   (b) Für alle $s = 1, ..., S$ gelte $N_s \to \infty$, derart, dass $\frac{N_s}{n_{is}} \leq N_{0s} < N_0 < \infty$, $s = 0, 1$.

Die erste Voraussetzung stellt analog zu denen in den ersten beiden Designs sicher, dass unterschiedliche Patienten gleichen Gesundheitszustandes Messwiederholungen sind, die zweite Voraussetzung sichert zweierlei: Zum einen, dass die Patientenkollektive der verschiedenen Behandlungsgruppen sich in ihrer Größe nicht zu stark unterscheiden (a) und zum anderen, dass in jeder einzelnen Methodengruppe der Unterschied zwischen der Anzahl der gesunden und kranken Patienten nicht zu groß ausfällt (b).
Eine Übersicht über die eingeführte Notation befindet sich in Tabelle 3.3.

Tabelle 3.3: Notationsübersicht Design 3

Methode 1							
	gesund			krank			
	Patient			Patient			
Reader	1	$\cdots$	n_{01}	1	$\cdots$	n_{11}	AUC
1 2 3	$\begin{pmatrix} X_{0,1}^{(1,1)} \\ X_{0,1}^{(1,2)} \\ X_{0,1}^{(1,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{0,n_{01}}^{(1,1)} \\ X_{0,n_{01}}^{(1,2)} \\ X_{0,n_{01}}^{(1,3)} \end{pmatrix}$	$\begin{pmatrix} X_{1,1}^{(1,1)} \\ X_{1,1}^{(1,2)} \\ X_{1,1}^{(1,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{1,n_{11}}^{(1,1)} \\ X_{1,n_{11}}^{(1,2)} \\ X_{1,n_{11}}^{(1,3)} \end{pmatrix}$	$\begin{pmatrix} w_{11} \\ w_{12} \\ w_{13} \end{pmatrix}$
	$= \mathbf{X}_{0,1}^{(1)}$	$\cdots$	$= \mathbf{X}_{0,n_{01}}^{(1)}$	$= \mathbf{X}_{1,1}^{(1)}$	$\cdots$	$= \mathbf{X}_{1,n_{11}}^{(1)}$	$= \mathbf{w}^{(1)}$

Methode 2							
	gesund			krank			
	Patient			Patient			
Reader	1	$\cdots$	n_{02}	1	$\cdots$	n_{12}	AUC
1 2 3	$\begin{pmatrix} X_{0,1}^{(2,1)} \\ X_{0,1}^{(2,2)} \\ X_{0,1}^{(2,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{0,n_{02}}^{(2,1)} \\ X_{0,n_{02}}^{(2,2)} \\ X_{0,n_{02}}^{(2,3)} \end{pmatrix}$	$\begin{pmatrix} X_{1,1}^{(2,1)} \\ X_{1,1}^{(2,2)} \\ X_{1,1}^{(2,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{1,n_{12}}^{(2,1)} \\ X_{1,n_{12}}^{(2,2)} \\ X_{1,n_{12}}^{(2,3)} \end{pmatrix}$	$\begin{pmatrix} w_{21} \\ w_{22} \\ w_{23} \end{pmatrix}$
	$= \mathbf{X}_{0,1}^{(2)}$	$\cdots$	$= \mathbf{X}_{0,n_{02}}^{(2)}$	$= \mathbf{X}_{1,1}^{(2)}$	$\cdots$	$= \mathbf{X}_{1,n_{12}}^{(2)}$	$= \mathbf{w}^{(2)}$

3.4.2 Asymptotische Verteilung des Schätzers und Schätzung der Kovarianzmatrix

Sei $\bar{s} \in 1, ..., S$ beliebig, aber fest. Die Vektoren $\mathbf{X}_{i,k}^{(\bar{s})}$ erfüllen dann die Voraussetzung für das nichtparametrische Behrens-Fisher Problem, d.h. für den Vektor $\mathbf{w}^{(\bar{s})} = (w^{(\bar{s},1)}, ..., w^{(\bar{s},R)})'$ der Accuracies der einzelnen Reader bei Methode $\bar{s}$ gelten folgende Aussagen:

1. **Schätzung der Accuracies**
   Die einzelnen Accuracies $w^{(\bar{s},r)} = \int {F_0}^{(\bar{s},r)} d{F_1}^{(\bar{s},r)}$ werden erwartungstreu und konsistent durch
   $$\hat{w}^{(\bar{s},r)} = \frac{1}{n_{0\bar{s}}} \left[\frac{1}{n_{1\bar{s}}} \sum_{k=1}^{n_{1\bar{s}}} R_{1,k}^{(\bar{s},r)} - \frac{n_{0\bar{s}}+1}{2} \right]$$
   geschätzt, wobei $R_{1,k}^{(\bar{s},r)}$ der Rang von $X_{1,k}^{\bar{s},r}$ unter allen $N_{\bar{s}}$ Diagnoseergebnissen des r-ten Readers der $\bar{s}$-ten Methode ist (vgl. Satz A.2).

2. **Asymptotische Verteilung des Schätzers**
   Es gilt nach Satz 3.1
   $$\sqrt{N_{\bar{s}}}(\hat{\mathbf{w}}^{(\bar{s})} - \mathbf{w}^{(\bar{s})}) \doteq \sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})},$$
   wobei
   $$\begin{aligned} \sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})} &= (\sqrt{N_{\bar{s}}}B^{(\bar{s},1)}, ..., \sqrt{N_{\bar{s}}}B^{(\bar{s},R)})' \quad \text{und} \\ \sqrt{N_{\bar{s}}}B^{(\bar{s},r)} &= \sqrt{N_{\bar{s}}} \left(\frac{1}{n_{1\bar{s}}} \sum_{k=1}^{n_{1\bar{s}}} F_0^{(\bar{s},r)}(X_{1,k}^{(\bar{s},r)}) - \frac{1}{n_{0\bar{s}}} \sum_{k=1}^{n_{0\bar{s}}} F_1^{(\bar{s},r)}(X_{0,k}^{(\bar{s},r)}) + 1 - 2w^{(\bar{s},r)} \right), \\ & \quad r = 1, ..., R. \end{aligned}$$
   Die Zufallsvariablen $\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}$ sind gleichmäßig beschränkt; fordert man nun zusätzlich, dass für das Minimum der Eigenwerte von $Cov(\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}) = \mathbf{V}_{N_{\bar{s}}}$, welches mit $\lambda_{min}^{\bar{s}}$ bezeichnet werde, gilt: $\lambda_{min}^{\bar{s}} \geq \lambda_0 > 0$, dann ist $\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}$ asymptotisch $N(\mathbf{0}, \mathbf{V}_{N_{\bar{s}}})$ verteilt (vgl. Satz 3.2).

3. **Schätzung der Kovarianzmatrix**
   Es sei $R_{i,k}^{(\bar{s},r)}$ der Rang von $X_{i,k}^{\bar{s},r}$ unter allen $N_{\bar{s}}$ Diagnoseergebnissen des r-ten Readers der $\bar{s}$-ten Modalität und $\mathbf{R}_{i,k}^{(\bar{s})} = (R_{i,k}^{(\bar{s},1)}, ..., R_{i,k}^{(\bar{s},R)})'$.
   Weiter sei $R_{i,k}^{(\bar{s},r|i)}$ der Rang von $X_{i,k}^{\bar{s},r}$ unter allen $n_{i\bar{s}}$ Diagnoseergebnissen der Patienten im Zustand i des r-ten Readers der $\bar{s}$-ten Methode. Auch hier sei $\mathbf{R}_{i,k}^{(\bar{s}|i)} = (R_{i,k}^{(\bar{s},1|i)}, ..., R_{i,k}^{(\bar{s},R|i)})'$.
   $\overline{\mathbf{R}}_{i\cdot}^{(\bar{s})} = (\overline{R}_{i\cdot}^{(\bar{s},1)}, ..., \overline{R}_{i\cdot}^{(\bar{s},R)})'$ sei der Vektor der über die Patienten gemittelten Ränge.
   Mit $\mathbf{Z}_{i,k}^{(\bar{s})} = \mathbf{R}_{i,k}^{(\bar{s})} - \mathbf{R}_{i,k}^{(\bar{s}|i)}$ und $\overline{\mathbf{Z}}_{i\cdot}^{(\bar{s})} = \overline{\mathbf{R}}_{i\cdot}^{(\bar{s})} - \left(\frac{n_{i\bar{s}}+1}{2} \cdot \mathbf{1}_R \right)$ definiert man schließlich:
   $$\hat{\mathbf{V}}_{N_{\bar{s}},i} = \frac{N_{\bar{s}}}{(N_{\bar{s}} - n_{i\bar{s}})^2 n_{i\bar{s}}(n_{i\bar{s}} - 1)} \sum_{k=1}^{n_{i\bar{s}}} (\mathbf{Z}_{i,k}^{(\bar{s})} - \overline{\mathbf{Z}}_{i\cdot}^{(\bar{s})})(\mathbf{Z}_{i,k}^{(\bar{s})} - \overline{\mathbf{Z}}_{i\cdot}^{(\bar{s})})' \quad i = 0, 1$$
   Mit dieser Notation ist $\hat{\mathbf{V}}_{N_{\bar{s}}} = \hat{\mathbf{V}}_{N_{\bar{s}},0} + \hat{\mathbf{V}}_{N_{\bar{s}},1}$ ein konsistenter Schätzer für $\mathbf{V}_{N_{\bar{s}}}$ (siehe Satz 3.3).

Fordert man nun für alle $s = 1, ..., S$ die Regularitätsbedingung

Voraussetzung 3.9

für alle $s = 1, ..., S$ gelte für das Minimum der Eigenwerte λ_{min}^{s} der Kovarianzmatrix $Cov(\sqrt{N_s}\mathbf{B}^{(s)}) = \mathbf{V}_{N_s}$: $\lambda_{min}^{s} \geq \lambda_0 > 0$,

so gelten obige Resultate für alle $s = 1,...,S$, d.h. es gibt für alle s einen zu $\sqrt{N_s}(\hat{\mathbf{w}}^{(s)} - \mathbf{w}^{(s)})$ asymptotisch äquivalenten Vektor von Zufallsvariablen $\sqrt{N_s}\mathbf{B}^{(s)}$, welcher asymptotisch $N(\mathbf{0}, \mathbf{V}_{N_s})$-verteilt ist.
Da die Diagnoseergebnisse unterschiedlicher Verfahren von verschiedenen Patienten stammen, sind die Zufallsvektoren $\sqrt{N_s}(\hat{\mathbf{w}}^{(s)} - \mathbf{w}^{(s)})$ und $\sqrt{N_t}(\hat{\mathbf{w}}^{(t)} - \mathbf{w}^{(t)})$, $s,t = 1,..,S, s \neq t$ und somit auch die Zufallsvektoren $\sqrt{N_s}\mathbf{B}^{(s)}$ und $\sqrt{N_t}\mathbf{B}^{(t)}$, $s,t = 1,..,S, s \neq t$ paarweise unabhängig, d.h. es gilt:

$$\begin{pmatrix} \sqrt{N_1}(\hat{\mathbf{w}}^{(1)} - \mathbf{w}^{(1)}) \\ \vdots \\ \sqrt{N_S}(\hat{\mathbf{w}}^{(S)} - \mathbf{w}^{(S)}) \end{pmatrix} \doteqdot \begin{pmatrix} \sqrt{N_1}\mathbf{B}^{(1)} \\ \vdots \\ \sqrt{N_S}\mathbf{B}^{(S)} \end{pmatrix} \dot{\sim} N(\mathbf{0}, \bigoplus_{s=1}^{S} \mathbf{V}_{N_s})$$

Es sei $\hat{\mathbf{w}} = (\hat{\mathbf{w}}^{(1)'}, ..., \hat{\mathbf{w}}^{(S)'})'$, der Vektor der Accuracies aller Reader-Methoden-Kombinationen und entsprechend seien $\mathbf{w} = (\mathbf{w}^{(1)'}, ..., \mathbf{w}^{(S)'})'$ und $\mathbf{B} = (\mathbf{B}^{(1)'}, ..., \mathbf{B}^{(S)'})'$ definiert. Nach Voraussetzung 2(a) ist $\frac{N}{N_s}$ und somit auch $\sqrt{\frac{N}{N_s}}$ gleichmäßig beschränkt für alle $s = 1,...,S$, d.h. aus $\sqrt{N_s}(\hat{\mathbf{w}}^{(s)} - \mathbf{w}^{(s)}) \doteqdot \sqrt{N_s}\mathbf{B}^{(s)}$ folgt nach Satz A.8:

$$\sqrt{\frac{N}{N_s}} \cdot \sqrt{N_s}(\hat{\mathbf{w}}^{(s)} - \mathbf{w}^{(s)}) \doteqdot \sqrt{\frac{N}{N_s}} \cdot \sqrt{N_s}\mathbf{B}^{(s)}.$$

In vektorieller Schreibweise bedeutet dieses:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) = \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_R \cdot \begin{pmatrix} \sqrt{N_1}(\hat{\mathbf{w}}^{(1)} - \mathbf{w}^{(1)}) \\ \vdots \\ \sqrt{N_S}(\hat{\mathbf{w}}^{(S)} - \mathbf{w}^{(S)}) \end{pmatrix} \doteqdot \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_R \cdot \begin{pmatrix} \sqrt{N_1}\mathbf{B}^{(1)} \\ \vdots \\ \sqrt{N_S}\mathbf{B}^{(S)} \end{pmatrix} = \sqrt{N}\mathbf{B}$$

Nun gilt:

- Abgesehen von einer Konstanten ist $\sqrt{N}B^{(s,r)}$ die Differenz der Mittelwerte der unabhängigen Zufallsvariablen $F_0^{(\bar{s},r)}(X_{1,k}^{(s,r)})$ und $F_1^{(\bar{s},r)}(X_{0,k}^{(s,r)})$.
- Da das Minimum der Eigenwerte von $Cov(\sqrt{N_s}\mathbf{B}^{(s)})$ für alle $s = 1,...,S$ stets größer ist als λ_0 folgt, dass auch das Minimum der Eigenwerte von $Cov(\sqrt{N}\mathbf{B})$ größer als λ_0 ist.
- Da die Folge der Zufallsvariablen $\sqrt{N_s}B^{(s,r)}$ gleichmäßig beschränkt ist, ist wegen Voraussetzung 2(a) auch $\sqrt{N}B^{(s,r)} = \sqrt{\frac{N}{N_s}}\sqrt{N_s}B^{(s,r)} \leq M_0\sqrt{N_s}B^{(s,r)}$ gleichmäßig beschränkt.

Daher lässt sich die asymptotische Normalität von $\sqrt{N}\mathbf{B}$ analog zu der in Satz 3.2 mit Hilfe des zentralen Grenzwertsatzes nach LINDEBERG zeigen (vgl. Brunner u. a., 2002), d.h. es gilt:

$$\begin{aligned} \sqrt{N}\mathbf{B} &\dot{\sim} N\left(\mathbf{0}, \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_R \cdot \left[\bigoplus_{s=1}^{S} \mathbf{V}_{N_s}\right] \cdot \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_R\right) \\ \Rightarrow \quad \sqrt{N}\mathbf{B} &\dot{\sim} N(\mathbf{0}, \underbrace{\bigoplus_{s=1}^{S} \frac{N}{N_s}\mathbf{V}_{N_s}}_{=:\mathbf{V}_N}) \end{aligned}$$

Ersetzt man die Kovarianzmatrizen $\mathbf{V}_{N_s}$ nun durch ihre konsistenten Schätzer $\hat{\mathbf{V}}_{N_s}$, so erhält man einen konsistenten Schätzer $\hat{\mathbf{V}}_N$ für $\mathbf{V}_N$, und nach dem Satz von SLUTZKY (Satz A.3) somit die Aussage:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \dot{\sim} \quad N(\mathbf{0}, \hat{\mathbf{V}}_N)$$

3.4.3 Hypothesen, Teststatistiken und Konfidenzintervalle

Auch wenn sich in diesem Design die Schätzung der Kovarianzmatrix von der im ersten Design unterscheidet, liegt hier trotzdem das gleiche faktorielle Design – ein gekreuztes Design der Faktoren Methode und Reader – zu Grunde, sodass die Hypothesenmatrizen, die daraus folgenden Teststatistiken sowie die Kontrastvektoren für die Konfidenzintervalle identisch sind. Es wird einzig eine andere Kovarianzmatrix verwendet, sodass an dieser Stelle darauf verzichtet wird, die Ergebnisse der Abschnitte 3.2.3 und 3.2.4 zu rezitieren.

3.5 Design 4

Im vierten Design werden wie im dritten unterschiedliche Diagnoseverfahren an verschiedenen Patientenkollektiven getestet. Wie im zweiten Design sind dabei unterschiedliche Reader für die Auswertung der verschiedenen Methoden nötig.

3.5.1 Modell und Notation

Es sei $X_{i,k}^{(s,r(s))}$ das Ergebnis des s-ten $(s = 1,...,S)$ Diagnoseverfahrens ausgewertet vom $r(s)$-ten Reader $(r = 1,...,R_s)$, erhoben am k-ten Patienten $(k = 1,...,n_{is})$, welcher gesund $(i = 0)$ oder krank $(i = 1)$ ist. Es sei weiter $\mathbf{X}_{i,k}^{(s)} = (X_{i,k}^{(s,1)},...,X_{i,k}^{(s,R_s)})'$ der Vektor der Ergebnisse des k-ten Patienten, an welchem das s-te Diagnoseverfahren getestet wurde. Statistisch wird das Modell somit wie folgt beschrieben:

Voraussetzung 3.10
Für alle Methoden $s = 1,...,S$ seien $\mathbf{X}_{i,k}^{(s)} = (X_{i,k}^{(s,1)},...,X_{i,k}^{(s,R_s)})$, $X_{i,k}^{(s,r(s))} \sim F_i^{(s,r(s))}$ $N_s = n_{0s} + n_{1s}$ unabhängige Zufallsvektoren, wobei die $F_i^{(s,r(s))}$ beliebige Verteilungen (in normalisierter Version) mit Ausnahme der Ein-Punkt-Verteilung seien. Weiter gelte:

1. Für jede Methoden $s = 1,...,S$ sei für alle $r(s), u(s) = 1,...,R_s$ die bivariate Verteilung von $(X_{i,k}^{(s,r(s))}, X_{i,k}^{(s,u(s))})$ gleich für alle $k = 1,...,n_{is}$ in der jeweiligen Gruppe $i = 0,1$.
2. Für die Stichprobenumfänge $N_s = n_{0s} + n_{1s}$ und $N = \sum_{s=1}^{S} N_s$ gelte
   (a) $N \to \infty$, derart, dass $\frac{N}{N_s} \leq M_0 < \infty$, $s = 1,...,S$.
   (b) Für alle $s = 1,...,S$ gilt $N_s \to \infty$, derart, dass $\frac{N_s}{n_{is}} \leq N_{0s} < N_0 < \infty$, $s = 0,1$.

Die Voraussetzungen sind analog zu denen des dritten Designs zu interpretieren. Zur Verbesserung der Übersichtlichkeit ist die Notation in Tabelle 3.5.1 dargestellt:

Tabelle 3.4: Notationsübersicht Design 4

Methode 1							
	gesund			krank			
	Patient			Patient			
Reader	1	$\cdots$	n_{01}	1	$\cdots$	n_{11}	AUC
1 2 3	$\begin{pmatrix} X_{0,1}^{(1,1)} \\ X_{0,1}^{(1,2)} \\ X_{0,1}^{(1,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{0,n_{01}}^{(1,1)} \\ X_{0,n_{01}}^{(1,2)} \\ X_{0,n_{01}}^{(1,3)} \end{pmatrix}$	$\begin{pmatrix} X_{1,1}^{(1,1)} \\ X_{1,1}^{(1,2)} \\ X_{1,1}^{(1,3)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{1,n_{11}}^{(1,1)} \\ X_{1,n_{11}}^{(1,2)} \\ X_{1,n_{11}}^{(1,3)} \end{pmatrix}$	$\begin{pmatrix} w_{11} \\ w_{12} \\ w_{13} \end{pmatrix}$
	$= \mathbf{X}_{0,1}^{(1)}$	$\cdots$	$= \mathbf{X}_{0,n_{01}}^{(1)}$	$= \mathbf{X}_{1,1}^{(1)}$	$\cdots$	$= \mathbf{X}_{1,n_{11}}^{(1)}$	$= \mathbf{w}^{(1)}$

Methode 2							
	gesund			krank			
	Patient			Patient			
Reader	1	$\cdots$	n_{02}	1	$\cdots$	n_{12}	AUC
4 5 6	$\begin{pmatrix} X_{0,1}^{(2,4)} \\ X_{0,1}^{(2,5)} \\ X_{0,1}^{(2,6)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{0,n_{02}}^{(2,4)} \\ X_{0,n_{02}}^{(2,5)} \\ X_{0,n_{02}}^{(2,6)} \end{pmatrix}$	$\begin{pmatrix} X_{1,1}^{(2,4)} \\ X_{1,1}^{(2,5)} \\ X_{1,1}^{(2,6)} \end{pmatrix}$	$\cdots$	$\begin{pmatrix} X_{1,n_{12}}^{(2,4)} \\ X_{1,n_{12}}^{(2,5)} \\ X_{1,n_{12}}^{(2,6)} \end{pmatrix}$	$\begin{pmatrix} w_{24} \\ w_{25} \\ w_{26} \end{pmatrix}$
	$= \mathbf{X}_{0,1}^{(2)}$	$\cdots$	$= \mathbf{X}_{0,n_{02}}^{(2)}$	$= \mathbf{X}_{1,1}^{(2)}$	$\cdots$	$= \mathbf{X}_{1,n_{12}}^{(2)}$	$= \mathbf{w}^{(2)}$

3.5.2 Asymptotische Verteilung des Schätzers und Schätzung der Kovarianzmatrix

Sei $\bar{s} \in \{1,...,S\}$ beliebig, aber fest. Für festes $\bar{s}$ erfüllt der Vektor $\mathbf{X}_{i,k}^{(\bar{s})}$ die Voraussetzungen für das nichtparametrische Behrens-Fisher Problem. Lehnt man nun die Notation dieses Designs an die des dritten Designs an; ändert dabei nur den im Design 3 von s unabhängigen Index r in den von s abhängigen Index $r(s)$, so gilt hier für den Vektor $\mathbf{w}^{(\bar{s})} = (w^{(\bar{s},1)},...,w^{(\bar{s},R_s)})'$ der Accuracies der $\bar{s}$-ten Methode :

1. **Schätzung der Accuracies**
   Die einzelnen Accuracies $w^{(\bar{s},r(s))} = \int F_0{}^{(\bar{s},r(s))} dF_1{}^{(\bar{s},r(s))}$ werden erwartungstreu und konsistent durch

$$\hat{w}^{(\bar{s},r(\bar{s}))} = \frac{1}{n_{0\bar{s}}} \left[\frac{1}{n_{1\bar{s}}} \sum_{k=1}^{n_{1\bar{s}}} R_{1,k}^{(\bar{s},r(\bar{s}))} - \frac{n_{0\bar{s}}+1}{2} \right]$$

   geschätzt.

2. **Asymptotische Verteilung des Schätzers**
   Es gilt

$$\sqrt{N_{\bar{s}}}(\hat{\mathbf{w}}^{(\bar{s})} - \mathbf{w}^{(\bar{s})}) \doteq \sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}.$$

Die Zufallsvariablen $\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}$ sind dabei gleichmäßig beschränkt und durch die zusätzliche Forderung, dass für das Minimum der Eigenwerte von $Cov(\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}) = \mathbf{V}_{N_{\bar{s}}}$, welches mit $\lambda^{\bar{s}}_{min}$ bezeichnet werde, gilt: $\lambda^{\bar{s}}_{min} \geq \lambda_0 > 0$, ergibt sich, dass $\sqrt{N_{\bar{s}}}\mathbf{B}^{(\bar{s})}$ asymptotisch $N(\mathbf{0}, \mathbf{V}_{N_{\bar{s}}})$ verteilt ist.

3. **Schätzung der Kovarianzmatrix**
   Mit

$$\hat{\mathbf{V}}_{N_{\bar{s}},i} = \frac{N_{\bar{s}}}{(N_{\bar{s}} - n_{i\bar{s}})^2 n_{i\bar{s}}(n_{i\bar{s}} - 1)} \sum_{k=1}^{n_{i\bar{s}}} (\mathbf{Z}^{(\bar{s})}_{i,k} - \overline{\mathbf{Z}}^{(\bar{s})}_{i\cdot})(\mathbf{Z}^{(\bar{s})}_{i,k} - \overline{\mathbf{Z}}^{(\bar{s})}_{i\cdot})' \quad i = 0, 1,$$

   ist $\hat{\mathbf{V}}_{N_{\bar{s}}} = \hat{\mathbf{V}}_{N_{\bar{s}},0} + \hat{\mathbf{V}}_{N_{\bar{s}},1}$ ein konsistenter Schätzer für $\mathbf{V}_{N_{\bar{s}}}$.

Fordert man also analog zum dritten Design für alle $s = 1, ..., S$ die Regularitätsbedingung,

Voraussetzung 3.11
für alle $s = 1, ..., S$ gelte für das Minimum der Eigenwerte λ^{s}_{min} der Kovarianzmatrix $Cov(\sqrt{N_s}\mathbf{B}^{(s)}) = \mathbf{V}_{N_s}$: $\lambda^{\bar{s}}_{min} \geq \lambda_0 > 0$,

so gibt es auch hier für alle s einen zu $\sqrt{N_s}(\hat{\mathbf{w}}^{(s)} - \mathbf{w}^{(s)})$ asymptotisch äquivalenten Vektor von Zufallsvariablen $\sqrt{N_s}\mathbf{B}^{(s)}$, welcher asymptotisch $N(\mathbf{0}, \mathbf{V}_{N_s})$ verteilt ist.
Aus der paarweisen Unabhängigkeit von $\sqrt{N_s}\mathbf{B}^{(s)}$ und $\sqrt{N_t}\mathbf{B}^{(t)}$, $s,t = 1, .., S, s \neq t$ folgt also auch hier:

$$\begin{pmatrix} \sqrt{N_1}(\hat{\mathbf{w}}^{(1)} - \mathbf{w}^{(1)}) \\ \vdots \\ \sqrt{N_S}(\hat{\mathbf{w}}^{(S)} - \mathbf{w}^{(S)}) \end{pmatrix} \doteq \begin{pmatrix} \sqrt{N_1}\mathbf{B}^{(1)} \\ \vdots \\ \sqrt{N_S}\mathbf{B}^{(S)} \end{pmatrix} \dot{\sim} N(\mathbf{0}, \bigoplus_{s=1}^{S} \mathbf{V}_{N_s})$$

Mit $\hat{\mathbf{w}} = (\hat{\mathbf{w}}^{(1)'}, ..., \hat{\mathbf{w}}^{(S)'})'$, $\mathbf{w} = (\mathbf{w}^{(1)'}, ..., \mathbf{w}^{(S)'})'$ und $\mathbf{B} = (\mathbf{B}^{(1)'}, ..., \mathbf{B}^{(S)'})'$ folgt aus Voraussetzung 2(a) und Satz A.8:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) = \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_{R_s} \cdot \begin{pmatrix} \sqrt{N_1}(\hat{\mathbf{w}}^{(1)} - \mathbf{w}^{(1)}) \\ \vdots \\ \sqrt{N_S}(\hat{\mathbf{w}}^{(S)} - \mathbf{w}^{(S)}) \end{pmatrix} \doteq \bigoplus_{s=1}^{S} \sqrt{\frac{N}{N_s}}\mathbf{I}_{R_s} \cdot \begin{pmatrix} \sqrt{N_1}\mathbf{B}^{(1)} \\ \vdots \\ \sqrt{N_S}\mathbf{B}^{(S)} \end{pmatrix} = \sqrt{N}\mathbf{B}$$

Die asymptotische Normalität

$$\sqrt{N}\mathbf{B} \quad \dot{\sim} \quad N(\mathbf{0}, \bigoplus_{s=1}^{S} \frac{N}{N_s}\mathbf{V}_{N_s})$$

folgt analog zu der von $\sqrt{N}\mathbf{B}$ im dritten Design wie in Satz 3.2. Ersetzt man die Kovarianzmatrizen $\mathbf{V}_{N_s}$ nun durch ihre konsistenten Schätzer $\hat{\mathbf{V}}_{N_s}$, so erhält man nach dem Satz von SLUTZKY (Satz A.3) das Resultat:

$$\sqrt{N}(\hat{\mathbf{w}} - \mathbf{w}) \quad \dot{\sim} \quad N(\mathbf{0}, \bigoplus_{s=1}^{S} \frac{N}{N_s}\hat{\mathbf{V}}_{N_s})$$

3.5.3 Hypothesen, Teststatistiken und Konfidenzintervalle

Da wie im zweiten Design auch hier ein zweifaktorieller, hierarchischer Versuchsplan (CRHF-RD(M)) zu Grunde liegt, wird an dieser Stelle auf die Abschnitte 3.3.3 und 3.3.4 verwiesen, wo die entsprechenden Hypothesenmatrizen, Teststatistiken und Kontrastvektoren für zweifaktorielle, hierarchische Versuchspläne angegeben sind. Dabei ist die dort angegebene Kovarianzmatrix durch die in diesem Abschnitt hergeleitete zu ersetzen.

4 Umsetzung der Theorie in SAS

Um Anwendern die Möglichkeit zu geben, die vorgestellten Methoden ohne großen Programmieraufwand einzusetzen, entstanden vier SAS Makros:

- **design1.sas**
- **design2.sas**
- **design3.sas**
- **design4.sas**

Alle Programme sind auf gleiche Art und Weise konzipiert, sodass an **design1.sas** exemplarisch die Verwendung der Makros erklärt wird. Stellvertretend für alle Makros befindet sich der Quellcode zu **design2.sas** im Anhang (B).
Für die Makros wird als Eingabe ein Datensatz (`data`) folgender Struktur benötigt:

Tabelle 4.1: Datenstruktur der SAS Makros

Patienten ID `subject`	Goldstandard `state`	Reader `rater`	Methode `method`	Messwert `var`
1	0	1	1	$X_{0,1}^{(1,1)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	R	S	$X_{1,n_1}^{(R,S)}$

Mit dem Befehl

```
%DESIGN1(DATA= ,
         VAR= ,
         STATE= ,
         RATER= ,
         METHOD= ,
         SUBJECT= ,
         ALPHA=0.05);
```

wird das Programm aufgerufen.
Abbildung 4.1 zeigt ein beispielhaftes Output, welches mit Hilfe simulierter Daten erzeugt wurde.

Ergebnisse der Analyse Design 1

Stichprobenumfänge

gesund	krank
20	20

Konfidenzintervalle

Meth	Read.	AUC	unten	oben
1	1	0.75	0.63	0.854
1	2	0.65	0.544	0.749
2	1	0.935	0.828	0.997
2	2	0.816	0.665	0.932
3	1	0.625	0.526	0.719
3	2	0.555	0.428	0.679

Konfidenzintervalle Methodeneffekte

Methode	mittlere AUC	unten	oben
1	0.7	0.641	0.756
2	0.876	0.787	0.944
3	0.59	0.521	0.658

ANOVA-Typ-Statistik

	Q_n	df	p-Wert
Methode	36.5684	1.88965	0
Reader	2.6859	1	0.10124
Wechselwirkung	0.1512	1.6114	0.87057

Wald-Typ-Statistik

	Q_n	df	p-Wert
Methode	32.3373	2	0
Reader	2.6859	1	0.10124
Wechselwirkung	0.3511	2	0.839

Abbildung 4.1: Beispielhafte Ausgabe von **design1.sas**

Die Ausgabe enthält sowohl die Werte der einfachen Punktschätzer für die Accuracies als auch die Werte der Schätzer für die gemittelten Methodeneffekte. Zusätzlich gibt das Programm für alle Accuracies Konfidenzintervalle an, welche mit Hilfe der in Abschnitt 3.1.5 vorgestellten Methode berechnet wurden. Das Niveau ist hierbei per Default auf 5% gesetzt, kann aber manuell über den Parameter `alpha` geändert werden.
Schließlich enthält die Ausgabe die Werte der Wald-Typ- und der ANOVA-Typ-Statistik (`Q_n`), sowie deren Freiheitsgerade (`df`) und p-Werte (`p-Wert`) für die im entsprechenden Design vorliegenden Effekte.

5 Simulationsergebnisse

Um Möglichkeiten und Grenzen der praktischen Anwendung der entwickelten Verfahren aufzuzeigen, werden Simulationen durchgeführt. Von entscheidener Bedeutung sind dabei die Einhaltung des Niveaus sowie das Verhalten der Teststatistiken unter bestimmten Alternativen. Auf Grund der zahlreichen Konstellationsmöglichkeiten für die Anzahl der Faktorstufen und die Größe der Stichprobenumfänge muss die Anzahl der simulierten Fälle auf einige typische Kombinationen und Datenlagen beschränkt werden.
Zur Simulation des Niveaus werden in jedem der vier Designs die empirischen Niveaus (bei einem nominellen Niveau von 5%) in Abhängigkeit vom Stichprobenumfang und der Größe der AUC berechnet. Hierbei werden Stichprobenumfänge von n_i bzw. $n_{is} \in \{20, 30, 50\}$ verwendet. Die simulierten Verteilungen F_0 und F_1 werden dabei so gegeneinander verschoben, dass Accuracies $AUC \in \{0.5, 0.7, 0.85, 0.9, 0.95\}$ entstehen.
Die Powersimulation untersucht das Verhalten der Teststatistiken für zwei mögliche Alternativen:

1. Die erste Möglichkeit ist eine *Ein-Punkt-Alternative*, bei der zur Powersimulataion der Teststatistik des Faktors Methode eine Accuracy $\overline{w}^{(t,\cdot)},\ t \in \{1, ..., S\}$ gegenüber den übrigen $\overline{w}^{(s,\cdot)},\ s = 1, ..., S,\ s \neq t$ um δ verschoben wird.
2. Die zweite Alternative bildet eine ansteigende Folge von AUCs, deren Steigung durch einen Parameter δ gekennzeichnet ist, d.h. der Verschiebungsvektor $(1, .., S)$ wird mit ansteigendem δ multipliziert.

In der Arbeit von Werner (2006) wird darauf hingewiesen, dass beide Verfahren ähnliche Resultate liefern, weswegen an dieser Stelle nur die erste Variante zur Powersimulation verwendet wird.
Für die Konfidenzintervalle werden Coverageprobability und Länge in Abhängigkeit vom Stichprobenumfang $n_i \in \{20, 30, 50\}$ und von der Größe der $AUC \in \{0.5, 0.7, 0.85, 0.9, 0.95\}$ simuliert.
Bei den hier dargestellten Simulationsergebnissen sind die simulierten Verteilungen F_0 und F_1 normalverteilt. Simulationen mit Scoredaten zeigen kaum Unterschiede, sodass zu Gunsten einer besseren Lesbarkeit auf eine Präsentation letzterer Simulationsergebnisse verzichtet wird.

5.1 Design 1

Bei einem nominellen Niveau von 5% wird das empirische Niveau in Abhängigkeit von Stichprobenumfängen und Größe der AUC ermittelt. Simuliert werden das Niveau der ANOVA- und der Wald-Typ-Statistik für die beiden Haupteffekte Methode und Reader und für die Wechselwirkung:

Tabelle 5.1: Niveausimulation Design 1 (nominelles Niveau: 5%)

		Anzahl der		Test-		AUC				
n_0	n_1	Methoden	Reader	statistik		0,5	0,7	0,85	0,9	0,95
20	20	2	2	ATS/WTS	Methode	0,0563	0,0559	0,052	0,0475	0,0376
					Reader	0,0563	0,0545	0,0523	0,0471	0,0383
					Wechselw.	0,0562	0,0542	0,0511	0,0472	0,0375
30	30	2	2	ATS/WTS	Methode	0,0537	0,0523	0,0509	0,0494	0,0425
					Reader	0,0543	0,0538	0,051	0,0487	0,0429
					Wechselw.	0,0541	0,0529	0,0504	0,0488	0,0435
50	50	2	2	ATS/WTS	Methode	0,0525	0,0502	0,0497	0,0489	0,0461
					Reader	0,0521	0,0521	0,0506	0,0494	0,0461
					Wechselw.	0,0521	0,0528	0,0496	0,0487	0,0466
20	20	3	2	ATS	Methode	0,0519	0,0489	0,0424	0,0361	0,0247
					Reader	0,0552	0,0544	0,0527	0,0484	0,0433
					Wechselw.	0,0506	0,0501	0,0473	0,0425	0,0309
20	20	3	2	WTS	Methode	0,066	0,0641	0,0634	0,0609	0,0581
					Reader	0,0552	0,0544	0,0527	0,0484	0,0433
					Wechselw.	0,0651	0,0637	0,0584	0,0503	0,0309
30	30	3	2	ATS	Methode	0,0506	0,0502	0,0456	0,0403	0,0305
					Reader	0,0531	0,054	0,0511	0,0486	0,0447
					Wechselw.	0,0505	0,0498	0,0474	0,0441	0,038
30	30	3	2	WTS	Methode	0,0594	0,06	0,0597	0,058	0,0556
					Reader	0,0531	0,054	0,0511	0,0486	0,0447
					Wechselw.	0,0595	0,0584	0,0541	0,0499	0,0395
50	50	3	2	ATS	Methode	0,0499	0,0488	0,0469	0,0439	0,0377
					Reader	0,0516	0,0524	0,0497	0,0503	0,0454
					Wechselw.	0,0517	0,0488	0,0485	0,0454	0,0403
50	50	3	2	WTS	Methode	0,0547	0,0549	0,055	0,0547	0,054
					Reader	0,0516	0,0524	0,0497	0,0503	0,0454
					Wechselw.	0,0573	0,0539	0,0525	0,0496	0,0416
20	20	3	4	ATS	Methode	0,0523	0,0495	0,0448	0,0428	0,0338
					Reader	0,047	0,0464	0,0407	0,0353	0,0252
					Wechselw.	0,0419	0,0378	0,032	0,0274	0,0179
20	20	3	4	WTS	Methode	0,0667	0,0634	0,0614	0,0626	0,0577
					Reader	0,0741	0,0767	0,0742	0,0751	0,0768
					Wechselw.	0,1252	0,1208	0,111	0,1021	0,0787
30	30	3	4	ATS	Methode	0,0508	0,0498	0,0462	0,044	0,0379
					Reader	0,0484	0,0476	0,0433	0,0395	0,0317
					Wechselw.	0,043	0,0437	0,037	0,0333	0,0246
30	30	3	4	WTS	Methode	0,06	0,0593	0,057	0,056	0,0548
					Reader	0,0666	0,067	0,0669	0,0663	0,0682
					Wechselw.	0,0953	0,0943	0,0867	0,823	0,0691
50	50	3	4	ATS	Methode	0,0502	0,0508	0,0487	0,0459	0,0424
					Reader	0,0504	0,0482	0,045	0,0432	0,0384
					Wechselw.	0,0467	0,0459	0,0425	0,0392	0,0325
50	50	3	4	WTS	Methode	0,0555	0,056	0,0549	0,0538	0,0537
					Reader	0,0613	0,0592	0,0587	0,059	0,0611
					Wechselw.	0,0758	0,0743	0,0716	0,0692	0,0635

Die Ergebnisse in Tabelle 5.1 zeigen, dass ab einer AUC von $0,90$ die Approximation zusammenbricht; sogar die Wald-Typ-Statistik wird konservativ. Ursächlich hierfür ist die in Kapitel 3.1.5 bereits vorgestellte Problematik am Rand des Definitionsbereichs, wo die Verteilung des Schätzers für die AUC nicht mehr symmetrisch ist und die Normalapproximation somit erst bei hohen Stichprobenumfängen eine gute Annäherung bildet.

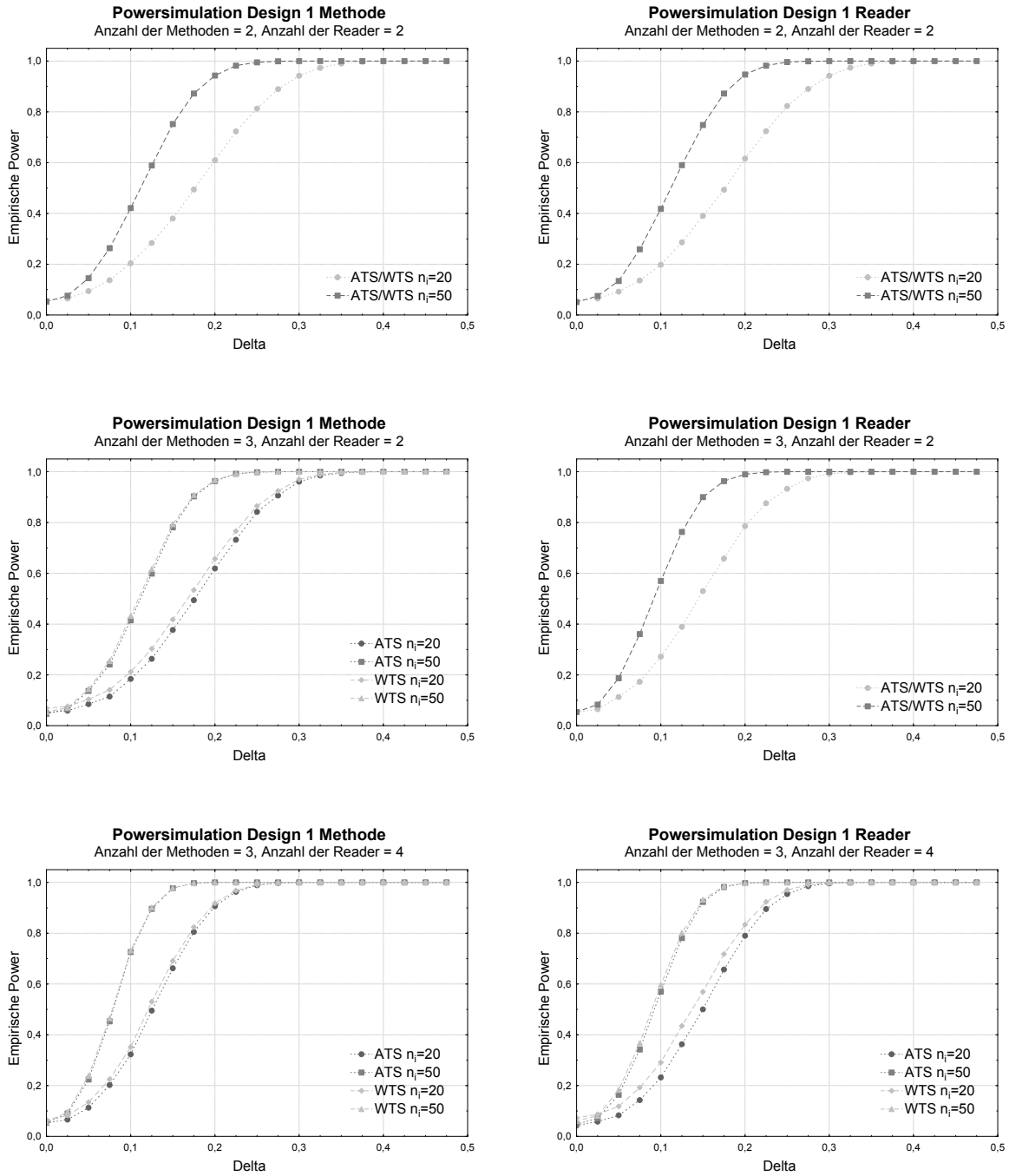

Abbildung 5.1: Powersimulation Design 1

Für hohe Accuracies werden sowohl ANOVA- als auch Wald-Typ-Statistik konservativ, so-

dass die sonst liberale Wald-Typ-Statistik hier eine bessere Approximation bildet. Folgendes sollte aber beachtet werden: Bei kleinen Stichproben und geringer Readeranzahl wird die Kovarianzmatrix $\mathbf{C}_M\hat{\mathbf{V}}_N\mathbf{C}_M'$ singulär und die Wald-Typ-Statistik lehnt somit die Hypothese auf keinen Methodeneffekt immer ab; Analoges gilt bei einer kleinen Methodenanzahl für den Readereffekt. Daher ist bei der Verwendung der Wald-Typ-Statistik auf eine große Fallzahl zu achten.
Die in Abbildung 5.1 dargestellte Powersimulation zeigt, dass es zwischen der ANOVA-Typ-Statistik und der Wald-Typ-Statistik keine nennenswerten Unterschiede gibt. Daher sollte die oben diskutierte Einhaltung des Niveaus über die Wahl des Verfahrens entscheiden.

5.2 Design 2

Tabelle 5.2: Niveausimulation Design 2 (nominelles Niveau: 5%)

		Anzahl der		Test-		AUC				
n_0	n_1	Methoden S	Reader $(r_s)_{s=1,\ldots,S}$	statistik		0,5	0,7	0,85	0,9	0,95
20	20	2	$(2,2)$	ATS	Methode	0,0557	0,0559	0,05	0,0476	0,0383
					Reader(Meth.)	0,0517	0,0524	0,044	0,038	0,0217
20	20	2	$(2,2)$	WTS	Methode	0,0557	0,0559	0,05	0,0476	0,0383
					Reader(Meth.)	0,0667	0,066	0,526	0,0404	0,0181
30	30	2	$(2,2)$	ATS	Methode	0,0532	0,0529	0,0505	0,0503	0,0427
					Reader(Meth.)	0,0506	0,0511	0,0454	0,0408	0,0316
30	30	2	$(2,2)$	WTS	Methode	0,0532	0,0529	0,0505	0,0503	0,0427
					Reader(Meth.)	0,0603	0,0597	0,0519	0,0442	0,0286
50	50	2	$(2,2)$	ATS	Methode	0,0518	0,0498	0,051	0,0501	0,0473
					Reader(Meth.)	0,0503	0,0494	0,0471	0,045	0,0389
50	50	2	$(2,2)$	WTS	Methode	0,0518	0,0498	0,051	0,0501	0,0473
					Reader(Meth.)	0,0559	0,0552	0,0515	0,0475	0,0375
20	20	3	$(2,3,2)$	ATS	Methode	0,0524	0,0502	0,0425	0,0388	0,0274
					Reader(Meth.)	0,0474	0,0435	0,0352	0,0292	0,0162
20	20	3	$(2,3,2)$	WTS	Methode	0,0654	0,0665	0,0644	0,0647	0,0629
					Reader(Meth.)	0,0936	0,0905	0,0788	0,0677	0,039
30	30	3	$(2,3,2)$	ATS	Methode	0,0509	0,0492	0,045	0,0417	0,0338
					Reader(Meth.)	0,0472	0,0453	0,0409	0,0345	0,0248
30	30	3	$(2,3,2)$	WTS	Methode	0,0597	0,0596	0,0596	0,0596	0,0609
					Reader(Meth.)	0,0766	0,0745	0,0703	0,0639	0,0471
50	50	3	$(2,3,2)$	ATS	Methode	0,0505	0,0492	0,047	0,0447	0,0382
					Reader(Meth.)	0,0493	0,0486	0,0438	0,0399	0,0317
50	50	3	$(2,3,2)$	WTS	Methode	0,0568	0,0553	0,0555	0,0568	0,0555
					Reader(Meth.)	0,0653	0,0664	0,062	0,0579	0,0506
20	20	3	$(4,3,4)$	ATS	Methode	0,0508	0,0495	0,0456	0,0404	0,0404
					Reader(Meth.)	0,0384	0,0345	0,0261	0,0218	0,0218
20	20	3	$(4,3,4)$	WTS	Methode	0,0644	0,0636	0,0631	0,0611	0,0611
					Reader(Meth.)	0,174	0,177	0,1848	0,1855	0,1855
30	30	3	$(4,3,4)$	ATS	Methode	0,0504	0,0491	0,0472	0,0404	0,0431
					Reader(Meth.)	0,0428	0,0389	0,0321	0,0218	0,0275
30	30	3	$(4,3,4)$	WTS	Methode	0,0592	0,0589	0,0585	0,0611	0,0575
					Reader(Meth.)	0,1242	0,125	0,1293	0,1855	0,135
50	30	3	$(4,3,4)$	ATS	Methode	0,05	0,05	0,0488	0,0431	0,0456
					Reader(Meth.)	0,0442	0,0434	0,0388	0,0275	0,0352
50	50	3	$(4,3,4)$	WTS	Methode	0,0553	0,0553	0,0551	0,0575	0,0543
					Reader(Meth.)	0,0891	0,091	0,0932	0,135	0,0985

Da im zweiten Design die Kovarianzmatrix wie im ersten Design geschätzt wird, unterscheidet sich dieses Modell vom ersten allein durch die Hypothesenmatrix $\mathbf{C}_{RD(M)}$. Es überrascht daher nicht, dass die Simulationsergebnisse (vgl. Tabelle 5.2) vergleichbar zu denen des ersten Designs sind.
Da man in diesem Design allerdings keinen Readereffekt isolieren kann, sind die Simulationsergebnisse auch für den Haupteffekt Methode etwas schlechter. Bei einer großen Anzahl an Faktorstufen brechen beide Statistiken für den Subkategorieeffekt Reader(Methode) zusammen: Die Wald-Typ-Statistik wird sehr liberal, die ANOVA-Typ-Statistik konservativ. Dieses ist jedoch auf Grund der hohen Dimension der Daten nicht verwunderlich und durch große Stichprobenumfänge auszugleichen.
Eine Betrachtung der Power (s. Abbildung 5.2) lässt analoge Schlüsse wie im ersten Design zu – ANOVA- und Wald-Typ-Statistik sind kaum unterschiedlich – sodass auch hier die Wahl der Statistik von der Einhaltung des Niveaus abhängig gemacht werden sollte.

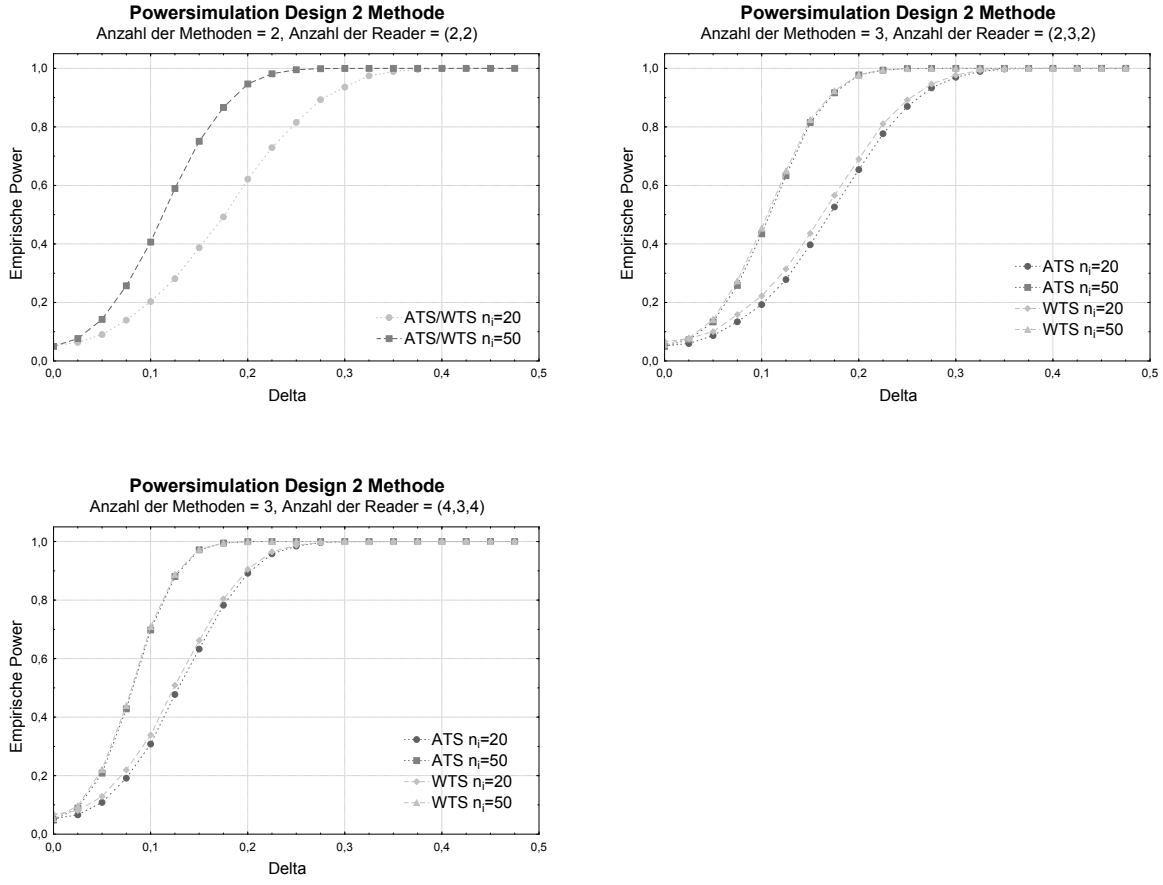

Abbildung 5.2: Powersimulation Design 2

5.3 Design 3

Tabelle 5.3: Niveausimulation Design 3 (nominelles Niveau: 5%)

$n_{si} = n$ $\forall i,s$	Anzahl der Methoden	Anzahl der Reader	Test-statistik		AUC 0,5	0,7	0,85	0,9	0,95
20	2	2	ATS/WTS	Methode	0,052	0,053	0,048	0,044	0,035
				Reader	0,052	0,053	0,047	0,044	0,035
				Wechselw.	0,052	0,051	0,048	0,044	0,034
30	2	2	ATS/WTS	Methode	0,052	0,051	0,047	0,046	0,041
				Reader	0,052	0,052	0,049	0,046	0,041
				Wechselw.	0,053	0,05	0,048	0,045	0,041
50	2	2	ATS/WTS	Methode	0,051	0,051	0,049	0,049	0,044
				Reader	0,05	0,051	0,048	0,048	0,044
				Wechselw.	0,052	0,051	0,049	0,048	0,045
20	3	2	ATS	Methode	0,051	0,05	0,041	0,035	0,023
				Reader	0,051	0,05	0,047	0,044	0,038
				Wechselw.	0,053	0,05	0,045	0,039	0,03
20	3	2	WTS	Methode	0,057	0,058	0,054	0,053	0,05
				Reader	0,051	0,05	0,047	0,044	0,038
				Wechselw.	0,058	0,055	0,047	0,038	0,023
30	3	2	ATS	Methode	0,051	0,049	0,044	0,04	0,03
				Reader	0,052	0,05	0,049	0,045	0,044
				Wechselw.	0,051	0,05	0,046	0,042	0,036
30	3	2	WTS	Methode	0,054	0,053	0,053	0,052	0,051
				Reader	0,052	0,05	0,049	0,045	0,044
				Wechselw.	0,054	0,053	0,048	0,043	0,033
50	3	2	ATS	Methode	0,051	0,05	0,048	0,043	0,038
				Reader	0,05	0,05	0,048	0,049	0,046
				Wechselw.	0,05	0,051	0,048	0,046	0,041
50	3	2	WTS	Methode	0,053	0,052	0,053	0,051	0,051
				Reader	0,05	0,05	0,048	0,049	0,046
				Wechselw.	0,052	0,053	0,049	0,047	0,04
20	3	4	ATS	Methode	0,05	0,049	0,045	0,042	0,031
				Reader	0,048	0,046	0,039	0,034	0,024
				Wechselw.	0,046	0,043	0,036	0,03	0,019
20	3	4	WTS	Methode	0,055	0,054	0,053	0,053	0,049
				Reader	0,057	0,058	0,058	0,059	0,062
				Wechselw.	0,077	0,074	0,064	0,056	0,036
30	3	4	ATS	Methode	0,049	0,049	0,046	0,044	0,038
				Reader	0,048	0,047	0,042	0,038	0,03
				Wechselw.	0,047	0,045	0,04	0,036	0,027
30	3	4	WTS	Methode	0,053	0,053	0,052	0,051	0,05
				Reader	0,054	0,055	0,056	0,055	0,057
				Wechselw.	0,066	0,064	0,06	0,055	0,043
50	3	4	ATS	Methode	0,052	0,048	0,049	0,046	0,042
				Reader	0,049	0,048	0,044	0,044	0,039
				Wechselw.	0,049	0,048	0,042	0,041	0,034
50	3	4	WTS	Methode	0,053	0,05	0,052	0,05	0,05
				Reader	0,053	0,053	0,052	0,053	0,055
				Wechselw.	0,06	0,06	0,055	0,053	0,048

Im dritten wie später auch im vierten Design sind deutlich größere Stichprobenumfänge vonnöten, da die Kovarianzmatrix für jede Methode separat geschätzt werden muss. Die Stichprobenumfänge n_{is} beziehen sich daher auf die Anzahl der Patienten der s-ten Methode (im Gesundheitszustand i). Auch hier lassen sich zu Design 1 vergleichbare Resultate

beobachten: ab einer AUC von $0,90$ bricht die Approximation zusammen und sogar die Wald-Typ-Statistik wird konservativ (vgl. Tabelle 5.3).

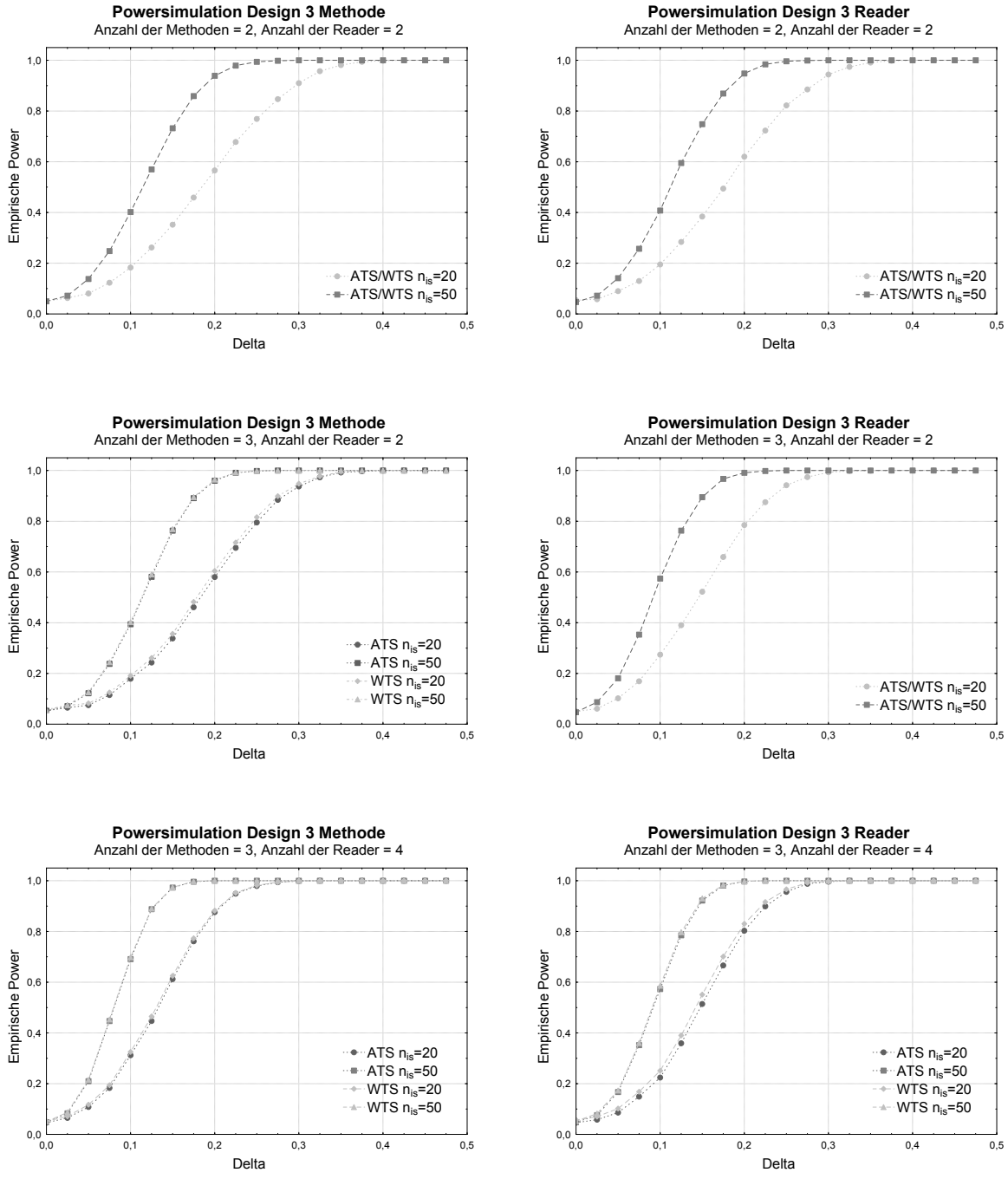

Abbildung 5.3: Powersimulation Design 3

Auch die Powersimulationen liefern im Vergleich zu 5.1 keine großen Unterschiede (vgl. Abbildung 5.3), sodass auch hier die Einhaltung des Niveaus über die Wahl der Statistik entscheiden sollte.

5.4 Design 4

Die Ergebnise der Niveau- (vgl. Tabelle 5.4) und Powersimulationen (vgl. Abbildung 5.4) des vierten Designs sind der Vollständigkeit halber ebenfalls angegeben. Sie entsprechen denen des zweiten Designs, allerdings ist auch hier zu beachten, dass der angegebene simulierte Stichprobenumfang pro Methode gilt.
Zwar werden die Komponenten der Kovarianzmatrix auf Grundlage des gleichen Stichprobenumfangs wie im zweiten Design geschätzt (vgl. Tabelle 5.2), trotzdem wird das Niveau für den Subkategorieeffekt besser eingehalten, da nicht alle Komponenten der Kovarianzmatrix geschätzt werden müssen, sondern manche als 0 bekannt sind.

Tabelle 5.4: Niveausimulation Design 4 (nominelles Niveau: 5%)

$n_{si} = n$	Anzahl der		Test-		AUC				
$\forall i,s$	Methoden S	Reader $(r_s)_{s=1,\ldots,S}$	statistik		0,5	0,7	0,85	0,9	0,95
20	2	$(2,2)$	ATS	Methode	0,0518	0,051	0,0463	0,0437	0,0351
				Reader(Meth.)	0,0544	0,0515	0,0448	0,0388	0,021
20	2	$(2,2)$	WTS	Methode	0,0518	0,051	0,0463	0,0437	0,0351
				Reader(Meth.)	0,0625	0,0583	0,0465	0,0337	0,0127
30	2	$(2,2)$	ATS	Methode	0,0511	0,0492	0,0484	0,0452	0,0417
				Reader(Meth.)	0,0524	0,0506	0,0464	0,0428	0,032
30	2	$(2,2)$	WTS	Methode	0,0511	0,0492	0,0484	0,0452	0,0417
				Reader(Meth.)	0,0581	0,0553	0,0482	0,0413	0,0241
50	2	$(2,2)$	ATS	Methode	0,0516	0,0511	0,0492	0,0481	0,0437
				Reader(Meth.)	0,0528	0,0506	0,0479	0,0455	0,0384
50	2	$(2,2)$	WTS	Methode	0,0516	0,0511	0,0492	0,0481	0,0437
				Reader(Meth.)	0,0554	0,0534	0,0489	0,0451	0,0345
20	3	$(2,3,2)$	ATS	Methode	0,0509	0,0491	0,0427	0,0378	0,0264
				Reader(Meth.)	0,0511	0,0469	0,039	0,0327	0,0182
20	3	$(2,3,2)$	WTS	Methode	0,0566	0,056	0,0559	0,056	0,0556
				Reader(Meth.)	0,074	0,0697	0,0595	0,0485	0,0221
30	3	$(2,3,2)$	ATS	Methode	0,051	0,0493	0,0449	0,0404	0,0315
				Reader(Meth.)	0,0496	0,0478	0,0419	0,0373	0,0257
30	3	$(2,3,2)$	WTS	Methode	0,054	0,0539	0,0539	0,0535	0,0534
				Reader(Meth.)	0,0641	0,0626	0,0567	0,0509	0,0364
50	3	$(2,3,2)$	ATS	Methode	0,0507	0,0486	0,0467	0,0445	0,0389
				Reader(Meth.)	0,0491	0,0482	0,0449	0,0407	0,033
50	3	$(2,3,2)$	WTS	Methode	0,053	0,0514	0,0525	0,0521	0,0516
				Reader(Meth.)	0,0572	0,0573	0,054	0,0508	0,0443
20	3	$(4,3,4)$	ATS	Methode	0,0511	0,0481	0,0441	0,0399	0,0332
				Reader(Meth.)	0,0468	0,0419	0,0315	0,0242	0,012
20	3	$(4,3,4)$	WTS	Methode	0,0556	0,054	0,0534	0,0524	0,0518
				Reader(Meth.)	0,1014	0,1015	0,1065	0,1037	0,0687
30	3	$(4,3,4)$	ATS	Methode	0,0484	0,0495	0,0455	0,0423	0,0371
				Reader(Meth.)	0,047	0,0447	0,0364	0,0313	0,0188
30	3	$(4,3,4)$	WTS	Methode	0,0518	0,0527	0,0516	0,051	0,506
				Reader(Meth.)	0,0802	0,0819	0,0859	0,089	0,0803
50	3	$(4,3,4)$	ATS	Methode	0,0507	0,0497	0,048	0,0446	0,0415
				Reader(Meth.)	0,0485	0,0457	0,0409	0,0377	0,0281
50	3	$(4,3,4)$	WTS	Methode	0,0529	0,0516	0,0522	0,0506	0,05
				Reader(Meth.)	0,0668	0,0672	0,0713	0,0743	0,0789

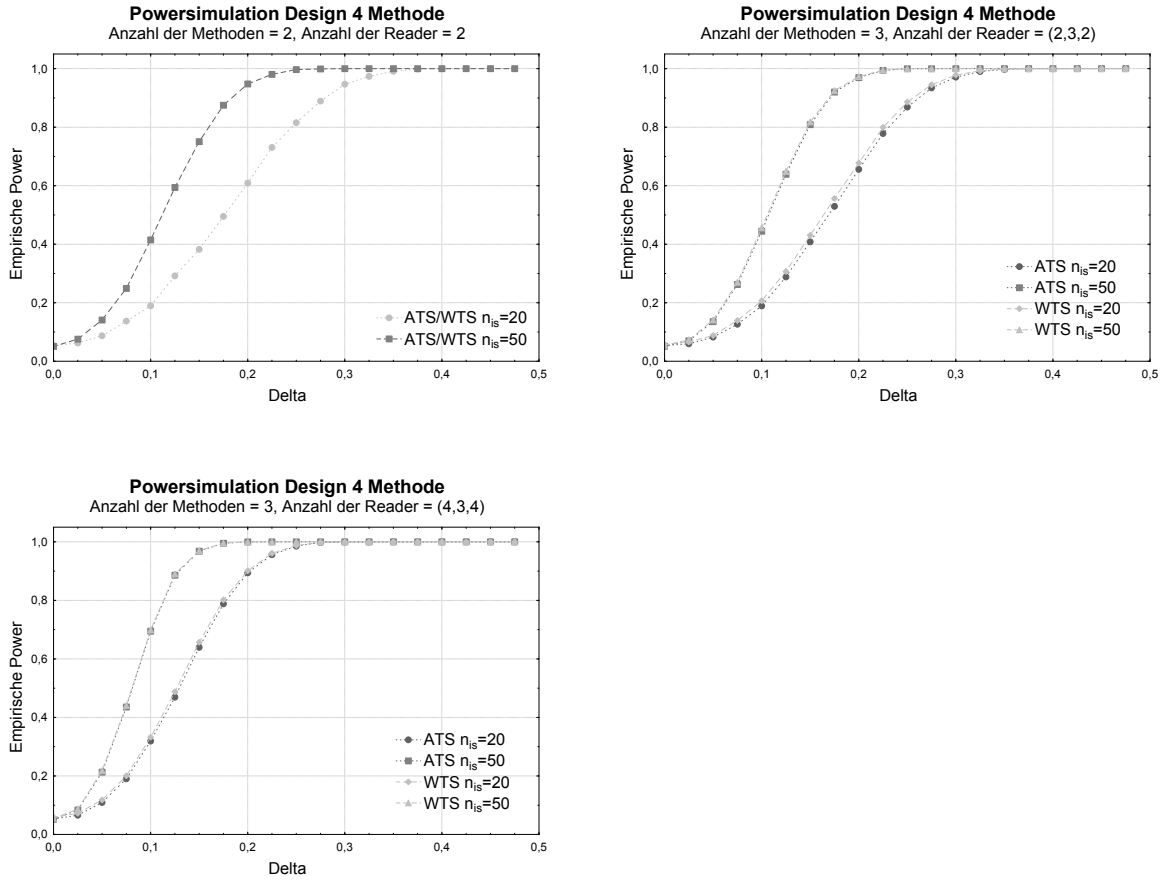

Abbildung 5.4: Powersimulation Design 4

5.5 Konfidenzintervalle

Für die Konfidenzintervalle wurden Länge und Coverageprobability simuliert. Mit steigendem Wert für die AUC werden die Konfidenzintervalle (auf Grund der kleineren Varianz) kürzer.

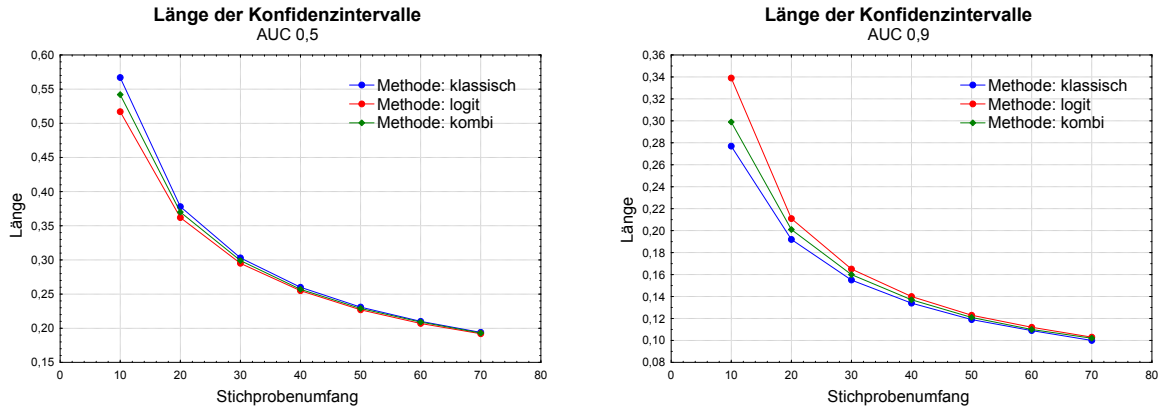

Abbildung 5.5: Länge der Konfidenzintervalle

Die Form der Funktion, welche die Länge in Abhängigkeit vom Stichprobenumfang darstellt, ist allerdings für alle Werte der AUC gleich, sodass sie nur exemplarisch für AUCs von $0,5$ und $0,9$ dargestellt wird (s. Abbildung 5.5).

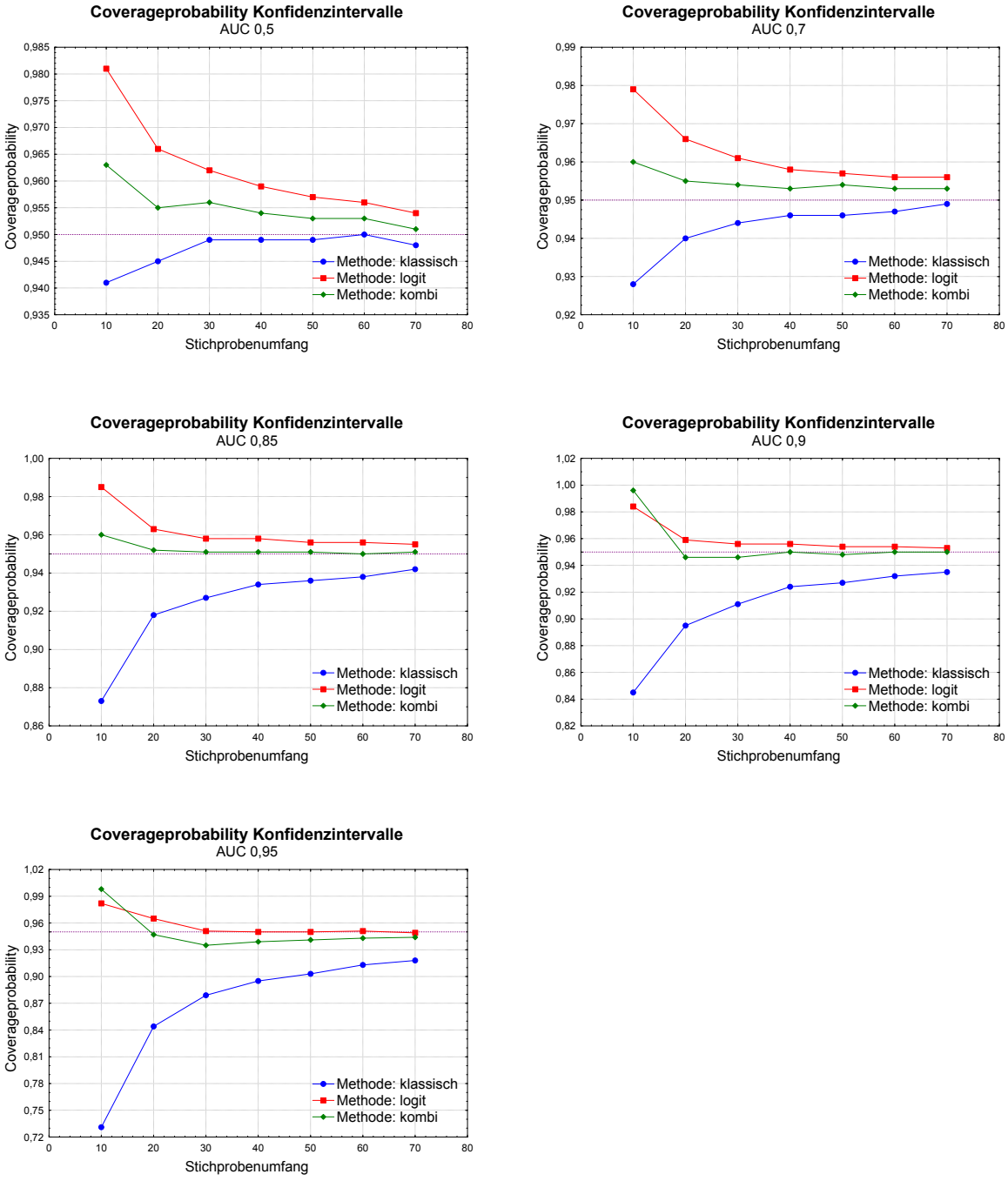

Abbildung 5.6: Simulation Coverageprobability der Konfidenzintervalle

Im Bezug auf die Coverageprobability zeigt sich, dass die kürzeren, klassischen Konfidenzintervalle antikonservativ und die mit Hilfe der logit-Transformation berechneten Konfidenzintervalle konservativ sind. Die kombinierten Konfidenzintervalle sind noch immer

leicht konservativ, halten das Niveau bei kürzerer Länge dabei allerdings besser ein als die logit Konfidenzintervalle (s. Abbildung 5.6), weswegen diese bevorzugt verwendet werden sollten.

6 Zusammenfassung und Ausblick

Für die im Rahmen von Diagnosestudien häufig auftretenden vier faktoriellen Designs wurden in dieser Arbeit nichtparametrische Modelle entwickelt, mit deren Hilfe zukünftig die Evaluation diagnostischer Verfahren erfolgen kann. Hierfür wurde die Lösung des multivariaten nichtparametrischen Behrens-Fisher Problems nach Brunner u. a. (2002) auf faktorielle Designs in Diagnosestudien erweitert. Das Ziel dieser Arbeit, den potenziellen Anwendern eine Verwendung der vorgestellten Methoden zu ermöglichen, wurde durch die programmtechnische Umsetzung der Theorie in SAS erreicht.

Zur besseren Diagnostik werden in der Praxis häufig mehrere Beobachtungseinheiten eines Patienten betrachtet, die unterschiedliche Gesundheitszustände aufweisen können. Dieses Problem der sog. Clusterdaten wird in der vorliegenden Arbeit nicht behandelt. Für das erste Design findet man hierfür Lösungsvorschläge bei Werner (2006). Eine Erweiterung dieser Lösungskonzepte auf die übrigen Designs wäre bei der Auswertung zukünftiger Diagnosestudien von großem Nutzen.
Bei hohen Accuracies treten Schwierigkeiten bei der Approximation der Verteilungsfunktion des Schätzers für die AUC auf. Für Konfidenzintervalle wurden die daraus resultierenden Probleme durch eine Kombination der klassischen und der logit-Konfidenzintervalle gelöst. Das Konzept der Transformation lässt sich in ähnlicher Art und Weise auch auf Teststatistiken erweitern, sodass sich die Frage stellt, ob für hohe relative Effekte die logit-Transformation zu einer Verbesserung der hier vorgestellten Verfahren führen würde. In der Arbeit von Konietschke (2006) zeigt diese Methode vielversprechende Simulationsergebnisse.

Es ist wünschenswert, die in dieser Arbeit erstellten Konzepte wie oben dargelegt zu erweitern, um so die möglichen Anwendungsbereiche zu vergrößern und auf diese Art und Weise eine präzise, statistisch fundierte Basis für die nichtparametrische Auswertung von Diagnosestudien zu legen.

A Definitionen, Sätze und Notationen

A.1 Matrizenrechnung

Matrizen und Vektoren werden stets **fett** geschrieben. Eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ hat dabei die Gestalt:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

Die zu $\mathbf{A}$ transponierte Matrix wird mit $\mathbf{A}'$ bezeichnet.

Definition A.1 *[Spezielle Matrizen]*

1. ***Einheitsmatrix***
$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}$$
2. ***Einservektor***
$$\mathbf{1}_n = (1, ..., 1)'_{1 \times n}$$
3. $\mathbf{n} \times \mathbf{n}$ ***-Einser-Matrix***
$$\mathbf{J}_n = \mathbf{1}_n \mathbf{1}'_n = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$
4. ***zentrierende Matrix***
$$\mathbf{P}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & & \vdots \\ \vdots & & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \dots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix}$$

Definition A.2 *[Spur] Für eine quadratische Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ heißt $Sp(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$ die Spur von $\mathbf{A}$.*

Definition A.3 *[Rang] Für eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ heißt die Anzahl der linear unabhänigen Zeilenvektoren $r(\mathbf{A})$ der Rang von $\mathbf{A}$.*

Definition A.4 *[Kronecker-Summe] Für beliebige Matrizen $\mathbf{A}$ und $\mathbf{B}$ heißt*

$$\mathbf{A} \oplus \mathbf{B} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{B} \end{array} \right)$$

die Kronecker-Summe von $\mathbf{A}$ und $\mathbf{B}$.

Definition A.5 *[Kronecker-Produkt] Für beliebige Matrizen $\mathbf{A} \in \mathbb{R}^{m \times n}$ und $\mathbf{B} \in \mathbb{R}^{p \times q}$ heißt*

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & & a_{mn}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{mp \times nq}$$

Kronecker-Produkt von $\mathbf{A}$ und $\mathbf{B}$.

Definition A.6 *[Verallgemeinerte Inverse] Für eine beliebige Matrix $\mathbf{A}$ heißt $\mathbf{A}^-$ verallgemeinerte Inverse zu $\mathbf{A}$, falls $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ ist. Weiter heißt $\mathbf{A}$ reflexive verallgemeinerte Inverse zu $\mathbf{A}$, falls $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$ gilt.*

Definition A.7 *[MOORE-PENROSE-Inverse] Eine Matrix $\mathbf{A}^+$ mit den Eigenschaften*

1. $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$,
2. $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$,
3. $(\mathbf{A}\mathbf{A}^+)' = \mathbf{A}\mathbf{A}^+$ *und*
4. $(\mathbf{A}^+\mathbf{A})' = \mathbf{A}^+\mathbf{A}$

heißt MOORE-PENROSE-*Inverse zu $\mathbf{A}$.*

A.2 Wahrscheinlichkeitstheorie

Definition A.8 *[Verteilungsfunktion] Für eine Zufallsvariable X heißt*

$$\begin{aligned} F^-(x) &= P(X < x) \quad \textit{links-stetige} \\ F^+(x) &= P(X \leq x) \quad \textit{rechts-stetige} \\ F(x) &= \frac{1}{2}[F^+(x) + F^-(x)] \quad \textit{normalisierte} \end{aligned}$$

Version der Verteilungsfunktion.

Definition A.9 *[Zählfunktion] Die Funktion*

$$\begin{aligned} c^-(x) &= \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \text{ heißt links-stetige} \\ c^+(x) &= \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \text{ heißt rechts-stetige} \\ c(x) &= \frac{1}{2}[c^+(x) + c^-(x)] \text{ heißt normalisierte} \end{aligned}$$

Version der Zählfunktion.

Definition A.10 *[Mittelränge] Es sei $c(x)$ definiert wie in Definition A.9, ferner seien $x_1, ..., x_N$ beliebige reelle Zahlen, dann heißt*

$$r_i = \frac{1}{2} + \sum_{i=1}^{N} c(x_i - x_j)$$

der Mittelrang von x_i unter allen Zahlen $x_1, ..., x_N$.

Satz A.1 *[Normalisierte empirische Verteilungsfunktion] Für eine Stichprobe $X_1, ..., X_N$ von Beobachtungen $X_k \sim F$, $k = 1, ..., N$ heißt die Funktion*

$$\hat{F}_i(x) = \frac{1}{N} \sum_{k=1}^{N} c(x - X_k)$$

normalisierte empirische Verteilungsfunktion von $X_1, ..., X_k$. Sie ist ein erwartungstreuer und konsistenter Schätzer für die normalisierte Verteilungsfunktion F.

Definition A.11 *[Relativer Effekt] Für zwei unabhänige Zufallsvariablen $X_0 \sim F_0$ und $X_1 \sim F_1$ heißt die Wahrscheinlichkeit*

$$w = P(X_0 < X_1) + \frac{1}{2} P(X_0 = X_1) = \int F_0 \, dF_1$$

relativer Effekt von X_1 zu X_2.

Satz A.2 *Die Zufallsvariablen X_{ij} seien u.i.v. $\sim F_i$, $i = 1, 2$ und $j = 1, ..., n_i$. $\hat{F}_i(x)$ bezeichne die empirische Verteilungsfunktion von $X_{i1}, ..., X_{n_i}$ und R_{ij} den Rang von X_{ij} in der gesamten Stichprobe $\{X_{11}, ..., X_{2n_2}\}$, d.h. unter allen $N = n_1 + n_2$ Beobachtungen, dann gilt:*

1. $\hat{w}_N = \int \hat{F}_1 d\hat{F}_2 = \frac{1}{n_1} \left(\frac{1}{n_2} \sum_{j=1}^{n_2} R_{2j} - \frac{n_2+1}{2} \right)$
2. *$\hat{w}_N$ ist erwartungstreu für w*
3. *$\hat{w}_N$ ist konsistent für w, falls $min(n_1, n_2) \to \infty$*

Beweis: siehe z.B. Brunner und Munzel (2002, Abschnitt 4.2.2) □

Definition A.12 *[Asymptotische Äquivalenz] Zwei Folgen von Zufallsvariablen X_N, Y_N heißen asymptotisch äquivalent ($X_N \doteqdot Y_N$), wenn $\forall \varepsilon > 0$ gilt $\lim_{N\to\infty} P(|X_N - Y_N| > \varepsilon) = 0$. Gilt $Y_N \sim F_Y$ und $X_N \doteqdot Y_N$ so wird verkürzend $X_N \dot{\sim} F_Y$ geschrieben.*

Satz A.3 *[SLUTZKY]*

1. *Sei $\mathbf{X}_n \in \mathbb{R}^k$, $n \geq 1$, eine Folge von Zufallsvariablen mit $\mathbf{X} \xrightarrow{p} a$, wobei $\mathbf{a} \in \mathbb{R}^k$ konstant ist und sei ferner $g(\cdot)$ stetig in $\mathbf{a}$, dann gilt*

$$g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{a})$$

2. *Sei $\mathbf{X}_n \in \mathbb{R}^k$, $n \geq 1$, eine Folge von Zufallsvariablen mit $\mathbf{X} \xrightarrow{\mathcal{L}} \mathbf{X} \sim \mathbf{F}(\mathbf{x})$ und sei $g(\cdot)$ ein F-fast-überall stetige Funktion*

$$g(\mathbf{X}_n) \xrightarrow{\mathcal{L}} g(\mathbf{X}) \sim \mathbf{F}_g(\mathbf{x})$$

Beweis: siehe Ferguson (1996, Kapitel 6, Theorem 6)

Satz A.4 *[Verteilung einer quadratischen Form] Sei $\mathbf{A}_{n\times n} = \mathbf{A}'$ eine symmetrische Matrix und $\mathbf{X} = (X_1, ..., X_n)' \sim N(\mathbf{0}, \mathbf{V})$, mit $r(\mathbf{V}) = r \leq n$. Dann gilt*

$$\mathbf{X}'\mathbf{A}\mathbf{X} \sim \sum_{i=1}^{n} \lambda_i C_i$$

wobei $C_i \sim \chi^2_1$, $i = 1, ..., n$ u.i.v. Zufallsvariablen und λ_i die Eigenwerte von $\mathbf{AV}$ sind.

Beweis: siehe Mathai und Provost (1992, Kapitel 3.1, Representation 3.1a.1) □

Satz A.5 *[OGASAWARI-TAKAHASHI] Sei $\mathbf{X} = (X_1, ..., X_n)' \sim N(\mu, \mathbf{V})$ mit $r(\mathbf{V}) = r \leq n$. Sei ferner $\mathbf{V}^-$ eine symmetrische reflexive verallgemeinerte Inverse zu $\mathbf{V}$. Dann hat die quadratische Form $\mathbf{X}'\mathbf{V}^-\mathbf{X}$ eine nicht zentrale χ^2_f-Verteilung mit $f = r(\mathbf{V})$ Freiheitsgraden und Nichtzentralitätsparameter $\delta = \mu'\mathbf{V}^-\mu$. Falls $\mu = \mathbf{0}$ ist, hat die quadratische Form $\mathbf{X}'\mathbf{V}^-\mathbf{X}$ für jede beliebige Wahl einer g-Inversen $\mathbf{V}^-$ eine zentrale χ^2_f-Verteilung mit $f = r(\mathbf{V})$ Freiheitsgraden.*

Beweis: siehe Rao und Mitra (1971) □

Satz A.6 *[CRAMER] Die Abbildung $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$ sei auf einer Umgebung von $\mu \in \mathbb{R}^d$ stetig differenzierbar. Es sei $\mathbf{X}_n$ eine Folge von d-dimensionalen Zufallsvektoren mit $\sqrt{n}(\mathbf{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathbf{X}$. Dann gilt $\sqrt{n}[\mathbf{g}(\mathbf{X_n}) - \mathbf{g}(\mu)] \xrightarrow{\mathcal{L}} \mathbf{g}'(\mu)\mathbf{X}$. Insbesondere gilt, falls $\sqrt{n}(\mathbf{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathbf{U} \sim N(\mathbf{0}, \Sigma)$:*

$$\sqrt{n}[\mathbf{g}(\mathbf{X_n}) - \mathbf{g}(\mu)] \xrightarrow{\mathcal{L}} \mathbf{V} \sim N(\mathbf{0}, \mathbf{g}'(\mu)\Sigma[\mathbf{g}'(\mu)]')$$

Beweis: siehe Ferguson (1996, Kapitel 7, Theorem 7) □

Satz A.7 *[δ-Satz] Es sei $\phi \in \mathbb{R}$ eine Konstante und T_n eine Folge von Zufallsvariablen, sodass für $n \to \infty$ gilt $\sqrt{n}(T_n - \phi) \xrightarrow{\mathcal{L}} T$. Ferner sei die Funktion $g(t)$ an der Stelle ϕ stetig differenzierbar. Dann gilt:*

$$\sqrt{n}\,[g(T_n) - g(\phi)] \xrightarrow{\mathcal{L}} g'(\phi) \cdot T$$

Beweis: Der Beweis ist ein Spezialfall des Satzes von CRAMER. □

Satz A.8 *Es seien X_N und Y_N zwei asymptotisch äquivalente Folgen und sei weiter M_N mit $0 < M_N < M_0$ eine beschränkte Folge, dann gilt:*

$$M_N \cdot X_N \doteqdot M_N \cdot Y_N$$

Beweis: Da $X_N \doteqdot Y_N$ gibt es $\forall \varepsilon' > 0,\ \forall \delta' > 0$ ein $N_0 \in \mathbb{N}$ sodass $\forall n > N_0$ gilt $P(|X_n - Y_n| > \varepsilon') < \delta'$. Seien $\varepsilon, \delta > 0$ beliebig, wähle $N_0 \in \mathbb{N}$, so, dass $\forall n > N_0$ gilt $P(|X_n - Y_n| > \frac{\varepsilon}{M_0}) < \delta$. Dann gilt für alle $n > N_0$:

$$\begin{aligned} P(|M_n \cdot X_n - M_n \cdot Y_n| > \varepsilon) &= P(M_n \cdot |X_n - Y_n| > \varepsilon) \leq P(M_0 \cdot |X_n - Y_n| > \varepsilon) \\ &= P(|X_n - Y_n| > \frac{\varepsilon}{M_0}) < \delta \end{aligned}$$

Also gilt $M_N \cdot X_N \doteqdot M_N \cdot Y_N$. □

B Quellcode

```
/*************************************************************************/
/*                                                                       */
/*                              Design 2                                 */
/*-verschiedene Methoden werden an gleichen Patienten getestet          */
/*-verschiedene Methodne werden von verschiedenen Readern ausgewertet   */
/*                                                                       */
/*************************************************************************/

/*************************************************************************/
/* %DESIGN2(DATA=Data,                                                   */
/*                      VAR=Var,                                         */
/*                      STATE=Goldstandard,                              */
/*                      RATER=Reader,                                    */
/*                      METHOD=Modality,                                 */
/*                      SUBJECT=Subject,                                 */
/*                      ALPHA=0.05);                                     */
/* mit:                                                                  */
/* data: Name des Datensatzes                                            */
/* var: Name der Variablen die die Beobachtungen enthält                 */
/* goldstandard: goldstandard 0=gesund, 1=krank                          */
/* Modality: Methode (1,...,S)                                           */
/* Reader: Reader (1,...,r_1) für Methode 1,                             */
/*                              (r_1+1,...,r_2) für Methode 2, u.s.w.    */
/* Subject: Beobachtungsobjekt ID (1,...,N)                              */
/* alpha: Level für das Konfidenzintervall                               */
/*************************************************************************/;


%MACRO design2(data=_last_, var=, state=, rater=, method=, subject=, alpha=0.05);

***Datensatz sortieren***;
proc sort data=&data;
by &state &subject &method &rater;
run;


proc iml;
***Datensatz einlesen***;
use &DATA;
read all var{&state}  into state;
read all var{&rater}  into rater;
read all var{&method} into method;
read all var{&subject}into subject;
read all var{&VAR}    into X_vec;
close &DATA;

***Bestimme die Anzahl der Methoden und der Reader***;
stat=unique(state);
rat=unique(rater);
met=unique(method);
```

```
sub=unique(subject);

g=ncol(stat);
r_sum=ncol(rat);
s=ncol(met);
datalength=nrow(x_vec);
d=r_sum;

***Bestimme die Anzahl der Reader pro Methode***;
r_s = J(s,1,0);
do i=1 to s;
        index = loc(method=met[i]);
        r_s[i]= ncol(unique(rater[index]));
end;

***Bestimme Anzahl n0 und n1 der gesunden und kranken Patienten***;
n_i = J(g,1,0);
do i=1 to g;
   index = loc(state=stat[i]);
   n_i[i] = ncol(unique(subject[index]));
end;
n0 = n_i[1];
n1 = n_i[2];
n = sum(n_i);

***Test ob die Bedingungen des Designs erfüllt sind***;
if (d*n)^=datalength then
        print 'Nicht jeder Patient wurde von allen Readern ausgewertet! Abbruch!';
else do;

print 'Ergebnisse der Analyse Design 2';
Stichprobe = n0 || n1;
print Stichprobe[c={'gesund' 'krank'} label='Stichprobenumfänge'];
Design = j(2,s,0);
do i=1 to s;
        Design[1,i] = i;
        Design[2,i] =r_s[i];
end;
print Design[r = {'Methode' '#Reader'} label ='Anzahl der Reader pro Methode'];


***Bringe Datensatz in adäquate Form für die Analyse***;
x=(shape (x_vec,n,d))';
x0= x[,1:n0];
x1= x[,(n0+1):n];


***Ränge berechnen***;
***r0i, r1i Ränge innerhalb der jeweiligen Stp., r0, r1 Globalränge;
r0i = j(d,n0, 0);
r1i = j(d,n1, 0);
rx = j(d,n,0);
do j=1 to d;
        r0i[j,] = ranktie(x0[j,]);
        r1i[j,] = ranktie(x1[j,]);
        rx[j,]  = ranktie(x[j,]);
end;
r0 = rx[,1:n0];
r1 = rx[,n0+1:n];
```

```
***Schätzer für die AUC***;
AUC = 1/n0*(1/n1*r1[,+]-(n1+1)/2);

***Schätzer für die Kovarianzmatrix***;
z0 = r0-r0i;
z1 = r1-r1i;

*Mittelwerte;
z0m = z0[,+]/n0;
z1m = z1[,+]/n1;

*zentrieren;
z0c = z0-(z0m * j(1, n0, 1));
z1c = z1-(z1m * j(1, n1, 1));

vn0 = n/((n-n0)*(n-n0)*n0*(n0-1))*(z0c*z0c');
vn1 = n/((n-n1)*(n-n1)*n1*(n1-1))*(z1c*z1c');

vn = vn0+vn1;


***Aufstellen der Kontrastmatrixen***;
Jsd= j(s,d,1);
CR = j(d,d,0);
CS = j(s,d,0);
q = j(s+1,1,0);
p = j(s+1,1,0);
**Indesvektoren;
do i=1 to s;
        q[i+1]=r_s[i];
end;
do i=2 to s+1;
        p[i]= sum(q[1:i,1]);
end;

do i=1 to s;
        ri = r_s[i];
        Jri = j(ri,ri,1);
        Pri = 1/ri*Jri;
        CR[p[i]+1:p[i+1],p[i]+1:p[i+1]] = Pri;
        CS[i, p[i]+1:p[i+1]] = 1/ri * j(1,ri,1);
end;

CS = CS - 1/d*Jsd;  ***Matrix für den Methodeneffekt;
CR = i(d) - CR;          ***Matrix für den Readereffekt;


***Berechnung der Teststatistiken und der Freiheitsgrade der Chi-Quadratverteilung der
***Teststatistiken;

***Funktion zur Berechnung des p-Wertes bei der ANOVA-Typ-Statistik;
start ANOVA_TYP (n, AUC, C, vn, anova, df, p);
        T  = C' * ginv(C*C') * C;
        TV = T*vn;
        sp_TV = trace(TV);
        sp_TVTV = trace(TV*TV);
        anova = round((n*sp_TV) / (sp_TVTV) * AUC' * T *AUC, 0.0001) ;
        df = round((sp_TV * sp_TV) / sp_TVTV, 0.00001);
        p = round(1 - probchi(anova,df),0.00001);
finish;
```

```
***Funktion zur Berechnung des p-Wertes der Wald-Type-Statistik;
start Wald_TYPE (n, AUC, C, vn, wald, df, p);
        wald = round(n*AUC'*C'*ginv(C*vn*C')*C*AUC, 0.0001);
        df = round(round(trace(ginv(C)*C)), 0.00001);
        p = round(1-probchi(wald,df), 0.00001);
finish;


***Funktion zur Berechnung der Konfidenzintervalle;
start KI(AUC_i, vn_i, vn_max, n, upper_neu_KI_i, lower_neu_KI_i);

        u = probit(1-&alpha/2);
        u2= probit(1-&alpha);
        **Konfidenzintervalle klassisch;
        upper_KI_klassisch_i = AUC_i + sqrt(1/n * vn_i)*u;
        lower_KI_klassisch_i = AUC_i - sqrt(1/n * vn_i)*u;


        **Konfidenzintervall logit;
        *wenn AUC<1;
        if AUC_i<1 then do;
                *Transformation der AUC;
                AUC_logit_i = log(AUC_i /(1-AUC_i));
                AUC_logit_diff_i = 1/(AUC_i*(1-AUC_i));
                vn_logit = AUC_logit_diff_i*AUC_logit_diff_i*(vn_i);
                *Berechnung der logit Konfidenzintervalle;
                upper_logit_i = AUC_logit_i + sqrt(1/n * vn_logit) * u;
                lower_logit_i = AUC_logit_i - sqrt(1/n * vn_logit) * u;
                upper_KI_logit_i = (exp(upper_logit_i))/(1+exp(upper_logit_i));
                lower_KI_logit_i = (exp(lower_logit_i))/(1+exp(lower_logit_i));

        end;
        *Konfidenzintervall wenn Schätzer gleich 1 ist;
        else do;
                upper_KI_i = 1;
                lower_KI_i = AUC_i - sqrt(1/n * vn_max)*u2;
        end;


        **Konfidenzintervalle kombiniert;
        if AUC_i<1 then do;
                upper_neu_KI_i=(upper_KI_klassisch_i+upper_KI_logit_i)/2;
                lower_neu_KI_i=(lower_KI_klassisch_i+lower_KI_logit_i)/2;
                if upper_neu_KI_i > 1 then do;
                        upper_neu_KI_i = 1;
                        lower_logit_i = AUC_logit_i - sqrt(1/n * vn_logit) * u2;
                        lower_neu_KI_i = (exp(lower_logit_i))/(1+exp(lower_logit_i));
                end;
        end;
        else do;
                upper_neu_KI_i = 1;
                lower_neu_KI_i = AUC_i - sqrt(1/n * vn_max)*u2;
        end;
        upper_neu_KI_i  = round(upper_neu_KI_i, 0.001);
        lower_neu_KI_i  = round(lower_neu_KI_i, 0.001);

finish;
```

```
***Matrixen in die die Ergebnisse geschrieben werden;
anova_box = j(2,1,0);
p_anova_box = j(2,1,0);
df_anova_box = j(2,1,0);

wald_box = j(2,1,0);
p_wald_box = j(2,1,0);
df_wald_box = j(2,1,0);

*ANOVA - Methodeneffekt;
run ANOVA_TYP(n, AUC, CS, vn, anova, df, p);
anova_box[1] = anova; df_anova_box[1] = df ; p_anova_box[1] = p;
*ANOVA - Readereffekt;
run ANOVA_TYP(n, AUC, CR, vn, anova, df, p);
anova_box[2] = anova; df_anova_box[2] = df ; p_anova_box[2] = p;

*Wald-Type - Methodeneffekt;
run Wald_TYPE(n, AUC, CS, vn, wald, df, p);
wald_box[1] = wald; df_wald_box[1] = df ; p_wald_box[1] = p;
*Wald-Type - Readereffekt;
run Wald_TYPE(n, AUC, CR, vn, wald, df, p);
wald_box[2] = wald; df_wald_box[2] = df ; p_wald_box[2] = p;


*Konfidenzintervalle für die einzelnen relativen Effekte;
***Index für die Reader, Methoden;
confidence = j(d,2,1);
do i=2 to s;
        sum1 = sum(r_s[1:(i-1),1]) + 1;
        sum2 = sum(r_s[1:i,1]);
        confidence[sum1:sum2,1] = i;
end;

do i=1 to r_sum;
        confidence[i,2] = i;
end;

variances = vecdiag(vn);
vn_max = max(variances);

upper_KI = AUC;
lower_KI = AUC;

do i=1 to d;
        AUC_i = AUC[i];
        vn_i = vn[i,i];
        upper_KI_i = 0;
        lower_KI_i = 0;
        run  KI(AUC_i, vn_i, vn_max, n, upper_KI_i, lower_KI_i);
        upper_KI[i]=upper_KI_i;
        lower_KI[i]=lower_KI_i;
end;

Konfidenzintervalle = confidence||round(AUC,0.001)||lower_KI||upper_KI;
print Konfidenzintervalle[c={'Meth' 'Read.' ' AUC' 'unten ' ' oben '}
label='Konfidenzintervalle'];

*Konfidenzintervall für den mittleren Methodeneffekt;

Methodeneffekt = j(s,4,0);
```

```
do meth=1 to s;
        kontrast = j(1,r_sum,0);
        rs = r_s[meth];
        meth_s_upper_KI=0;
        meth_s_lower_KI=0;
        if meth=1 then do; sum1 = 1; sum2 = r_s[1]; end;
        else do;
        sum1 = sum(r_s[1:(meth-1),1]) + 1;
        sum2 = sum(r_s[1:meth,1]);
        end;
        kontrast[1,sum1:sum2] = j(1,rs,1/rs);
        vn_s = kontrast*vn*kontrast';
        vn_max = vn_s;
        AUC_s = kontrast*AUC;
        run KI(AUC_s, vn_s, vn_max, n, meth_s_upper_KI, meth_s_lower_KI);
        Methodeneffekt[meth, ] = meth || round(AUC_s, 0.001) || meth_s_lower_KI ||
meth_s_upper_KI;
end;
run;
print Methodeneffekt[c={'Methode' 'mittlere AUC' 'unten' 'oben'}
                                        label='Konfidenzintervalle Methodeneffekte'];


anova=anova_box||df_anova_box||p_anova_box;
print anova[c={'Q_n' 'df' 'p-Wert'} r={'Methode' 'Reader(Methode)'}
                        label='ANOVA-Typ-Statistik'];

wald_set=wald_box||df_wald_box||p_wald_box;
print wald_set[c={'Q_n' 'df' 'p-Wert'} r={'Methode' 'Reader(Methode)'}
                        label='Wald-Typ-Statistik'];
end;

quit; ***beende IML;

%MEND design2;
```

Literaturverzeichnis

[Bamber 1975] BAMBER, D.: The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. In: *Journal of Mathematical Psychology* 12 (1975), S. 387–415

[Bauer 2002] BAUER, H.: *Wahrscheinlichkeitstheorie*. de Gruyter, 2002

[Box 1954] BOX, G.E.P.: Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. In: *The Annals of Mathematical Statistics* 25 (1954), S. 290–302

[Brunner 2002] BRUNNER, E.: *Nonparametric Methods for Analyzing the Accuracy of Diagnostic Tests with Multiple Readers*. 2002. – Konferenzvortrag: Schering AG, Berlin

[Brunner 2005] BRUNNER, E.: *Angewandte Statistik I*. Oktober 2005. – Vorlesungsskript

[Brunner u. a. 1997] BRUNNER, E. ; DETTE, H. ; MUNK, A.: Box-Type Approximations in Nonparametric Factorial Designs. In: *Journal of the American Statistical Association* 92 (1997), S. 1494–1502

[Brunner und Munzel 2000] BRUNNER, E. ; MUNZEL, U.: The nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. In: *Biometrical Journal* 42 (2000), S. 1–9

[Brunner und Munzel 2002] BRUNNER, E. ; MUNZEL, U.: *Nichtparametrische Datenanalyse*. Springer Verlag, 2002

[Brunner u. a. 1999] BRUNNER, E. ; MUNZEL, U. ; PURI, M. L.: Rank-Score Tests in Factorial Designs with Repeated Measures. In: *Journal of Multivariate Analysis* 70 (1999), S. 286–317

[Brunner u. a. 2002] BRUNNER, E. ; MUNZEL, U. ; PURI, M. L.: The multivariate nonparametric Behrens-Fisher problem. In: *Journal of Statistical Planning and Inference* 108 (2002), S. 37–53

[Domhof 2001] DOMHOF, S.: *Nichtparametrische relative Effekte*, Georg-August-Universität Göttingen, Dissertation, 2001

[Efron und Tibshirani 1993] EFRON, E ; TIBSHIRANI, R.J.: *An Introduction to Bootstrap*. Chapman & Hall, 1993

[Ferguson 1996] FERGUSON, T.S.: *A Course in Large Sample Theory*. Chapman & Hall, 1996

[Hanley und McNeil 1982] HANLEY, J.A. ; MCNEIL, B.J.: The Meaning and the Use of the Area under a Receiver Operating Characteristic Curve. In: *Radiology* 143 (1982), S. 29–36

[Konietschke 2006] KONIETSCHKE, F.: *Konstruktion einfacher Konfidenzintervalle für Linearkombinationen von Erfolgswahrscheinlichkeiten*, Georg-August-Universität Göttingen, Diplomarbeit, 2006

[Mathai und Provost 1992] MATHAI, A.M. ; PROVOST, S.B.: *Quadratic Forms in Random Variables. Theory and Applications*. Marcel Dekker, Inc., 1992

[Peterson u. a. 1954] PETERSON, W.W. ; BIRDSALL, T.G. ; FOX, W.C.: The Theory of Signal Detectability. In: *Transactions of the IRE Professional Group on Information Theory* 4 (1954), S. 171–212

[Rao und Mitra 1971] RAO, C.R. ; MITRA, S.K.: *Generalized Inverse of Matrices and its Applications*. Wiley, New York, 1971

[Sackett u. a. 1996] SACKETT, D. L. ; ROSENBERG, W. M. C. ; GRAY, J. A. M. ; HAYNES, R.B. ; RICHARDSON, W. S.: Evidence based medicine: what it is and what it isn't. In: *BMJ (British Medical Journal)* 312 (1996), S. 71–72

[Werner 2006] WERNER, C.: *Nichtparametrische Analyse von diagnostischen Tests*, Georg-August-Universität Göttingen, Dissertation, 2006