Resamplingverfahren in der Statistik

Vorlesungsskript

Thorsten Dickhaus
Humboldt-Universität zu Berlin
Sommersemester 2011
Version: 21. Juli 2011

Vorbemerkungen

Das Material aus Kapitel 1 dieses Skripts ist im Wesentlichen aus den Vorlesungsskripten über Statistik I und II von Prof. Arnold Janssen, den Artikeln von Janssen and Pauls (2003) und Janssen (2005) sowie den Dissertationen von Thorsten Pauls und Markus Pauly übernommen. Teile von Kapitel 2 stammen aus dem Skript von Prof. Gerhard Dikta über Bootstrapverfahren in der Statistik. Arnold Janssen und Gerhard Dikta gilt mein herzlicher Dank für die vielen guten Lehrveranstaltungen, die ich bei ihnen hören durfte. Sollten sich in den Kapiteln 1 und 2 Fehler finden, so bin dafür natürlich ich verantwortlich. Lob und positive Kritik gebührt indes den Original-Autoren. Abschnitt 1.2 findet sich in leicht anderer Form in meiner Master-Arbeit.

Für die Manuskripterstellung danke ich Mareile Große Ruse.

Übungsaufgaben und R-Programme zu diesem Kurs stelle ich auf Anfrage gerne zur Verfügung. Einige Referenzen dazu finden sich im Text an den zugehörigen Stellen.

Verzeichnis der Abkürzungen und Symbole

$B(p,q)$	Betafunktion, $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$
$\lceil x \rceil$	Kleinste ganze Zahl größer oder gleich x
χ^2_ν	Chi-Quadrat Verteilung mit ν Freiheitsgraden
$\complement M$	Komplement der Menge M
δ_a	Dirac-Maß im Punkte a
$\stackrel{\mathcal{D}}{=}$	Gleichheit in Verteilung
F_X	Verteilungsfunktion einer reellwertigen Zufallsvariable X
FDR	False Discovery Rate
FWER	Family Wise Error Rate
$\lfloor x \rfloor$	Größte ganze Zahl kleiner oder gleich x
$\Gamma(\cdot)$	Gammafunktion, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt,\ x > 0$
im(X)	Bildbereich einer Zufallsgröße X
iid.	independent and identically distributed
$\mathbf{1}_M$	Indikatorfunktion einer Menge M
$\mathcal{L}(X)$	Verteilungsgesetz einer Zufallsvariable X
LFC	Least Favorable Configuration
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
Φ	Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung

$\varphi(\cdot)$	Verteilungsdichte der $\mathcal{N}(0,1)$-Verteilung
supp(F)	Träger der Verteilungsfunktion F
UNI$[a,b]$	Gleichverteilung auf dem Intervall $[a,b]$

Inhaltsverzeichnis

1 Einführung, Beispiele und allgemeine Theorie **1**

1.1 Grundlagen aus der Statistik 1

1.2 Motivation und Beispiele 6

1.3 L_1-Differenzierbarkeit und lokal beste Tests 10

1.4 Einige Rangtests 14

1.5 Allgemeine Theorie von Resamplingtests 21

2 Spezielle Resamplingverfahren für unabhängige Daten **28**

2.1 Mehrstichprobenprobleme, Permutationstests 28

2.2 Einstichprobenprobleme, Bootstraptests 32

2.3 Bootstrapverfahren für lineare Modelle 36

3 Resamplingverfahren für multiple Testprobleme **46**

3.1 Subset pivotality, Westfall und Young 46

3.2 Dudoit, van der Laan, Pollard 46

4 Resamplingverfahren für Zeitreihen und abhängige Daten **47**

5 Statistisches Lernen, Klassifikationstheorie **48**

Tabellenverzeichnis **49**

Abbildungsverzeichnis **50**

Literaturverzeichnis **51**

Kapitel 1

Einführung, Beispiele und allgemeine Theorie

1.1 Grundlagen aus der Statistik

Bezeichne X eine Zufallsgröße, die den möglichen Ausgang eines Experimentes beschreibt.[1]

Sei Ω der zu X gehörige Stichprobenraum, d. h., die Menge aller möglichen Realisierungen von X und $\mathcal{A} \subseteq 2^\Omega$ eine σ-Algebra über Ω. Die Elemente von $\mathcal{A}$ heißen messbare Teilmengen von Ω oder Ereignisse.

Bezeichne $\mathbb{P}^X$ die Verteilung von X. Es gelte $\mathbb{P}^X \in \mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$.

Definition 1.1 (Statistisches Experiment / Modell)
Ein Tripel $(\Omega, \mathcal{A}, \mathcal{P})$ mit $\Omega \neq \emptyset$ eine nichtleere Menge, $\mathcal{A} \subseteq 2^\Omega$ eine σ-Algebra über Ω und $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathcal{A}$ heißt statistisches Experiment bzw. statistisches Modell.

Falls $\Theta \subseteq \mathbb{R}^k, k \in \mathbb{N}$, so heißt $(\Omega, \mathcal{A}, \mathcal{P})$ parametrisches statistisches Modell, $\vartheta \in \Theta$ Parameter und Θ Parameterraum.

Statistische Inferenz beschäftigt sich damit, Aussagen über die wahre Verteilung $\mathbb{P}^X$ bzw. den wahren Parameter ϑ zu gewinnen. Speziell: Entscheidungsprobleme, insbesondere Testprobleme.

Testprobleme: Gegeben zwei disjunkte Teilmengen $\mathcal{P}_0, \mathcal{P}_1$ von $\mathcal{P}$ mit $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ ist eine Entscheidung darüber gesucht, ob $\mathbb{P}^X$ zu $\mathcal{P}_0$ oder $\mathcal{P}_1$ gehört. Falls $\mathcal{P}$ durch ϑ eineindeutig identifiziert ist, kann die Entscheidungsfindung auch vermittels ϑ und Teilmengen Θ_0 und Θ_1 von Θ mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$ formalisiert werden.

[1] Witting (1985): „Wir denken uns das gesamte Datenmaterial zu einer „Beobachtung“ x zusammengefasst.“

Formale Beschreibung des Testproblems:

$$H_0 : \vartheta \in \Theta_0 \quad \text{versus} \quad H_1 : \vartheta \in \Theta_1 \quad \text{oder}$$

$$H_0 : \mathbb{P}^X \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : \mathbb{P}^X \in \mathcal{P}_1.$$

Die $H_i, i = 1, 2$ nennt man Hypothesen. H_0 heißt Nullhypothese, H_1 Alternativhypothese / Alternative. Oft interpretiert man H_0 und H_1 auch direkt selbst als Teilmengen des Parameterraums, d. h., $H_0 \cup H_1 = \Theta$ und $H_0 \cap H_1 = \emptyset$. Zwischen H_0 und H_1 ist nun aufgrund von $x \in \Omega$ eine Entscheidung zu treffen. Dazu benötigt man eine Entscheidungsregel. Diese liefert ein statistischer Test.

Definition 1.2 (Statistischer Test)
Ein (nicht-randomisierter) statistischer Test ist eine messbare Abbildung

$$\varphi : (\Omega, \mathcal{A}) \to (\{0, 1\}, 2^{\{0,1\}}).$$

Konvention:

$$\varphi(x) = 1 \iff \textit{Nullhypothese wird verworfen, Entscheidung für } H_1,$$

$$\varphi(x) = 0 \iff \textit{Nullhypothese wird nicht verworfen.}$$

$\{x \in \Omega : \varphi(x) = 1\}$ *heißt Ablehnbereich (oder auch kritischer Bereich) von* φ*, kurz:* $\{\varphi = 1\}$.
$\{x \in \Omega : \varphi(x) = 0\}$ *heißt Annahmebereich von* φ*, kurz:* $\{\varphi = 0\} = \complement\{\varphi = 1\}$.

Problem: Testen beinhaltet mögliche Fehlentscheidungen.
Fehler 1. Art (α-Fehler, type I error): Entscheidung für H_1, obwohl H_0 wahr ist.
Fehler 2. Art (β-Fehler, type II error): Nicht-Verwerfung von H_0, obwohl H_1 wahr ist.
In der Regel ist es nicht möglich, die Wahrscheinlichkeiten für die Fehler 1. und 2. Art gleichzeitig zu minimieren. Daher: Asymmetrische Betrachtungsweise von Testproblemen.

(i) Begrenzung der Fehlerwahrscheinlichkeit 1. Art durch eine vorgegebene obere Schranke α (Signifikanzniveau, englisch: level),

(ii) Unter der Maßgabe (i) Minimierung der Wahrscheinlichkeit für Fehler 2. Art $\Rightarrow$ „optimaler“ Test.

Eine (zum Niveau α) statistisch abgesicherte Entscheidung kann also immer nur zu Gunsten von H_1 getroffen werden $\Rightarrow$ Merkregel: „Was nachzuweisen ist stets als Alternative H_1 formulieren!“.

Bezeichnungen 1.3

(i) $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta\big[\varphi\big] = \mathbb{P}_\vartheta(\varphi(X) = 1) = \int_\Omega \varphi d\mathbb{P}_\vartheta$ *bezeichnet die Ablehnwahrscheinlichkeit eines vorgegebenen Tests* φ *in Abhängigkeit von* $\vartheta \in \Theta$*. Für* $\vartheta \in \Theta_1$ *heißt* $\beta_\varphi(\vartheta)$ *Gütefunktion von* φ *an der Stelle* ϑ*. Für* $\vartheta \in \Theta_0$ *ergibt* $\beta_\varphi(\vartheta)$ *die Typ I-Fehlerwahrscheinlichkeit von* φ *unter* $\vartheta \in \Theta_0$*.*

Für $\alpha \in (0, 1)$ *vorgegeben heißt*

(ii) ein Test φ *mit* $\beta_\varphi(\vartheta) \leq \alpha$ *für alle* $\vartheta \in H_0$ *Test zum Niveau* α*,*

(iii) ein Test φ *zum Niveau* α *unverfälscht, falls* $\beta_\varphi(\vartheta) \geq \alpha$ *für alle* $\vartheta \in H_1$*.*

(iv) ein Test φ_1 *zum Niveau* α *besser als ein zweiter Niveau-*α *Test* φ_2*, falls* $\beta_{\varphi_1}(\vartheta) \geq \beta_{\varphi_2}(\vartheta)$ *für alle* $\vartheta \in H_1$ *und* $\exists \vartheta^* \in H_1$ *mit* $\beta_{\varphi_1}(\vartheta^*) > \beta_{\varphi_2}(\vartheta^*)$*.*

Eine wichtige Teilklasse von Tests sind die Tests vom Neyman-Pearson Typ.

Definition 1.4
Sei $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *ein statistisches Modell und sei* φ *ein Test für das Hypothesenpaar* $\emptyset \neq H \subset \Theta$ *versus* $K = \Theta \setminus H$*, der auf einer Prüfgröße* $T : \Omega \to \mathbb{R}$ *basiert. Genauer sei* φ *charakterisiert durch die Angabe von Ablehnbereichen* $\Gamma_\alpha \subset \mathbb{R}$ *für jedes Signifikanzniveau* $\alpha \in (0, 1)$*, so dass* $\varphi(x) = 1 \iff T(x) \in \Gamma_\alpha$ *für* $x \in \Omega$ *gilt. Sei nun die Teststatistik* $T(X)$ *derart, dass die Monotoniebedingung*

$$\forall \vartheta_0 \in H : \forall \vartheta_1 \in K : \forall c \in \mathbb{R} : \mathbb{P}_{\vartheta_0}(T(X) > c) \leq \mathbb{P}_{\vartheta_1}(T(X) > c) \tag{1.1}$$

gilt. Dann heißt φ *ein Test vom (verallgemeinerten) Neyman-Pearson Typ, falls für alle* $\alpha \in (0, 1)$ *eine Konstante* c_α *existiert, so dass*

$$\varphi(x) = \begin{cases} 1, & T(x) > c_\alpha, \\ 0, & T(x) \leq c_\alpha. \end{cases}$$

Bemerkung 1.5

(a) Die Monotoniebedingung (1.1) wird häufig so umschrieben, dass „die Teststatistik unter Alternativen zu größeren Werten neigt“.

(b) Die zu einem Test vom Neyman-Pearson (N-P) Typ gehörigen Ablehnbereiche sind gegeben als $\Gamma_\alpha = (c_\alpha, \infty)$*.*

(c) Die Konstanten c_α werden in der Praxis bestimmt über $c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}^*(T(X) > c) \leq \alpha\}$, wobei das Wahrscheinlichkeitsmaß $\mathbb{P}^*$ so gewählt ist, dass

$$\mathbb{P}^*(T(X) \in \Gamma_\alpha) = \sup_{\vartheta \in H} \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha)$$

gilt, falls H eine zusammengesetzte Nullhypothese ist („am Rande der Nullhypothese"). Ist H einelementig und $\mathbb{P}_H$ stetig, so gilt $c_\alpha = F_T^{-1}(1-\alpha)$, wobei F_T die Verteilungsfunktion von $T(X)$ unter H bezeichnet.

(d) Fundamentallemma der Testtheorie von Neyman und Pearson: Unter (leicht verschärftem) (1.1) ist ein Test vom N-P Typ gleichmäßig (über alle $\vartheta_1 \in K$) bester Test für H versus K.

Es gibt Dualitäten zwischen Testproblemen / Tests und (Bereichs-)Schätzproblemen / Konfidenzintervallen.

Definition 1.6

Gegeben sei ein statistisches Modell $(\Omega, \mathcal{A}, \mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\})$. Dann heißt $\mathcal{C} = (C(x) : x \in \Omega)$ mit $C(x) \subseteq \Theta \, \forall \, x \in \Omega$ eine Familie von Konfidenzbereichen zum Konfidenzniveau $1-\alpha$ für $\vartheta \in \Theta :\Longleftrightarrow \forall\, \vartheta \in \Theta : \mathbb{P}_\vartheta(\{x : C(x) \ni \vartheta\}) \geq 1-\alpha$.

Satz 1.7 (Korrespondenzsatz, siehe z.B. Lehmann and Romano (2005) oder Witting, 1985)

(a) Liegt für jedes $\vartheta \in \Theta$ ein Test φ_ϑ zum Niveau α vor und wird $\varphi = (\varphi_\vartheta,\ \vartheta \in \Theta)$ gesetzt, so ist $\mathcal{C}(\varphi)$, definiert über $C(x) = \{\vartheta \in \Theta : \varphi_\vartheta(x) = 0\}$, eine Familie von Konfidenzbereichen zum Konfidenzniveau $1-\alpha$.

(b) Ist $\mathcal{C}$ eine Familie von Konfidenzbereichen zum Konfidenzniveau $1-\alpha$ und definiert man $\varphi = (\varphi_\vartheta,\ \vartheta \in \Theta)$ über $\varphi_\vartheta(x) = 1 - \mathbf{1}_{C(x)}(\vartheta)$, so ist φ ein Test zum allgemeinen lokalen Niveau α, d. h., zum Niveau α für jedes $\vartheta \in \Theta$.

Beweis:

Sowohl in (a) als auch in (b) erhält man $\forall\, \vartheta \in \Theta \quad \forall\, x \in \Omega : \varphi_\vartheta(x) = 0 \Longleftrightarrow \vartheta \in C(x)$. Also ist φ ein Test zum allgemeinen lokalen Niveau α genau dann, wenn

$$\begin{aligned} & \forall\, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta(\{\varphi_\vartheta = 0\}) \geq 1-\alpha \\ \Leftrightarrow \quad & \forall\, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta(\{x : C(x) \ni \vartheta\}) \geq 1-\alpha \\ \Leftrightarrow \quad & \mathcal{C} \text{ ist Familie von Konfidenzbereichen zum Konfidenzniveau } 1-\alpha. \end{aligned}$$

■

Bemerkung 1.8

(a) Die Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$ lässt sich schön grafisch veranschaulichen, falls Ω und Θ eindimensional sind.

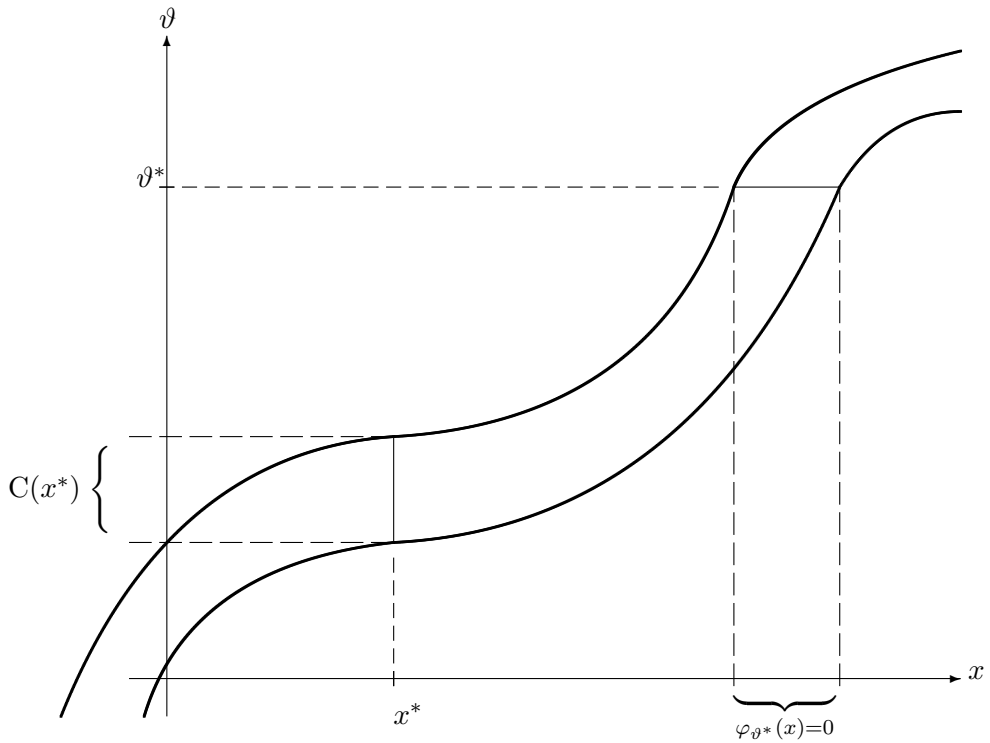

Abbildung 1.1: Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$

(b) Ein einzelner Test φ zum Niveau α für eine Hypothese H kann interpretiert werden als $(1-\alpha)$-Konfidenzbereich. Setze dazu

$$C(x) = \begin{cases} \Theta\,, & \textit{falls } \varphi(x) = 0, \\ K = \Theta \backslash H\,, & \textit{falls } \varphi(x) = 1. \end{cases}$$

Umgekehrt liefert jeder Konfidenzbereich $C(x)$ einen Test zum Niveau α für eine Hypothese $H \subset \Theta$.
Setze hierzu $\varphi(x) = \mathbf{1}_K(C(x))$, wobei

$$\mathbf{1}_B(A) := \begin{cases} 1\,, & \textit{falls } A \subseteq B, \\ 0\,, & \textit{sonst.} \end{cases}$$

für beliebige Mengen A und B.

Abschließend noch ein maßtheoretischer Satz, der sich einige Male für technische Beweise in den nachfolgenden Abschnitten in Kapitel 1 als nützlich erweisen wird.

Satz 1.9 (Satz von Vitali, siehe Witting (1985), Satz 1.181)
Sei $(\Omega, \mathcal{A}, \mu)$ ein σ-endlicher Messraum. Für $n \in \mathbb{N}_0$ seien $f_n : \Omega \to \mathbb{R}$ messbare Abbildungen.

Ist $f_n \to f_0$ μ-fast überall konvergent und ist

$$\limsup_{n\to\infty} \int |f_n|^p d\mu \leq \int |f_0|^p d\mu < \infty \text{ für ein } p \geq 1,$$

so folgt $\int |f_n - f_0|^p d\mu \to 0$ für $n \to \infty$. Ist μ ein Wahrscheinlichkeitsmaß, so genügt die Voraussetzung μ-stochastischer Konvergenz von f_n gegen f_0 anstelle der Konvergenz μ-fast überall.

1.2 Motivation und Beispiele

Ein Hauptproblem der statistischen Testtheorie ist das Testen des Erwartungswertes von Zufallsgrößen, die als Modell für eine erhobenen Stichprobe im experimentativen Umfeld vom Umfang n benutzt werden. Wir betrachten also n Zufallsvariablen $X_1, \ldots, X_n$, wobei die X_i im einfachsten Fall als i.i.d. angenommen werden. Das statistische Testproblem lautet nun häufig

$$H_0 : \mathbb{E}[X_1] = 0 \text{ versus } H_1 : \mathbb{E}[X_1] > 0.$$

Dieses Testproblem ergibt sich zum Beispiel beim Testen der mittleren Wirksamkeit eines neuen Medikamentes im Vergleich mit einem bereits etablierten Produkt zum Zwecke der Zulassung des neuen Präparates.
Als Teststatistik für dieses Problem findet bei bekannter Varianz $\sigma^2 = \mathrm{Var}(X_1)$ das arithmetische Mittel $T_n = \frac{1}{n}\sum_{i=1}^n X_i =: \bar{X}_n$ Verwendung; diese Teststatistik ist suffizient und vollständig für das zu Grunde liegende Testproblem. Ist (wie in den meisten Anwendungsfällen) σ^2 indes unbekannt, so bildet sich die geeignete Teststatistik als $\tilde{T}_n = \sqrt{n} \cdot \bar{X}_n / V_n^{\frac{1}{2}}$, wobei hier für die unbekannte Varianz σ^2 der erwartungstreue Schätzer $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ eingesetzt wird. Will man nun einen Niveau α-Test

$$\varphi_n = \begin{cases} 1 & > \\ & \tilde{T}_n \quad c_n(\alpha) \\ 0 & \leq \end{cases}$$

konstruieren, stellt sich das Problem, den richtigen kritischen Wert $c_n(\alpha)$ zu ermitteln. Lässt sich für die zur Modellierung herangezogenen Zufallsgrößen die Normalverteilungsannahme rechtfertigen, so ist dieses Problem bereits gelöst und das Ergebnis ist der sogenannte Gaußtest für T_n bzw. der Studentische t-Test für $\tilde{T}_n$, bei welchem die kritischen Werte als die Quantile der Standardnormalverteilung bzw. t-Verteilung mit $(n-1)$ Freiheitsgraden gewählt werden. Ist die Normalverteilungsannahme jedoch nicht gerechtfertigt und ist insbesondere keine Information über die Verteilung von $X_1, \ldots, X_n$ verfügbar, so gibt es keine Theorie für die exakte Bestimmung von $c_n(\alpha)$. Der t-Test ist in Fällen, in denen die X_i nicht normalverteilt sind nicht zu empfehlen, da er das Niveau α schlecht einhält. Eine erste Möglichkeit, auch in diesem Fall einen Test anzugeben,

stammt aus dem Zentralen Grenzwertsatz. Dieser besagt, dass, mit $\mu = E[X_1]$,

$$\mathcal{L}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) \to \mathcal{N}(0,1),\ n \to \infty.$$

Zusammen mit dem Satz von Slutsky lässt sich hieraus ein asymptotischer Niveau α-Test für das obige Testproblem konstruieren, nämlich

$$\varphi_n^{as} = \begin{cases} 1 & > \\ & \tilde{T}_n \qquad \Phi^{-1}(1-\alpha)\ . \\ 0 & \leq \end{cases}$$

Allerdings ist bei diesem Vorgehen die Approximationsgüte für kleine Stichprobenumfänge n häufig nicht hinreichend gut, siehe unten.

Eine Lösungsmöglichkeit der angedeuteten Problematik stellt der sogenannte **bootstrap**, eine Resamplingmethode, dar. Sei dazu im Einstichprobenproblem $X = (X_1, \ldots, X_n)$. Das statistische Modell sei gegeben durch $(\Omega^n, \mathcal{A}^n, (\mathbb{P}_\vartheta^n)_{\vartheta\in\Theta})$. Hierbei ist also $\mathbb{P}_\vartheta = \mathcal{L}(X_i), \Omega \subseteq \mathbb{R}$ der Bildraum von X_i und X_i „lebt“ auf $(\Omega^{-1}, \mathcal{F}, \mathbb{P})$, $i = 1, \ldots, n$. Es sei

$$\begin{aligned} T : \{Q : Q \text{ Verteilung auf } \Omega\} &\to \mathbb{R} \\ Q &\mapsto T(Q) \end{aligned}$$

ein interessierendes Funktional (häufig: Kennzahl einer Verteilung) vom Bildraum Ω der X_i in die reellen Zahlen. Ein Schätzer für das Wahrscheinlichkeitsmaß $\mathbb{P}_\vartheta$ ist dann das empirische Maß $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_{X_i}$ (Gleichverteilung auf den Daten). Daraus lässt sich ein (plug-in) Schätzer $T(\hat{\mathbb{P}}_n)$ für das Funktional $T(\mathbb{P}_\vartheta)$ gewinnen, der im Allgemeinen nicht erwartungstreu ist. Gesucht ist deshalb die Verteilung

$$\mathbb{P}(T(\hat{\mathbb{P}}_n) - T(\mathbb{P}_\vartheta) \leq t),\ t \in \mathbb{R} \tag{1.2}$$

des Fehlers, um beispielsweise Konfidenzintervalle zu konstruieren oder Tests durchzuführen.
Die **bootstrap Idee** besteht nun darin, den ursprünglichen Wahrscheinlichkeitsraum $(\Omega^n, \mathcal{A}^n, \mathbb{P}_\vartheta^n)$ durch eine empirische Version $(\Omega^n, \mathcal{A}^n, (\hat{\mathbb{P}}_n)^n)$ zu ersetzen.
Dazu konstruiert man eine iid. **bootstrap Stichprobe** $X_1^*, \ldots, X_n^*$ mit $X_i^* : (\Omega^*, \mathcal{A}^*, \mathbb{P}^*) \to (\Omega, \mathcal{A})$, für die gilt:

$$\mathbb{P}^{*X_1^*|(X_1,\ldots,X_n)} = \hat{\mathbb{P}}_n.$$

Auf Grund der Definition von $\hat{\mathbb{P}}_n$ ist unmittelbar klar, dass das Ziehen der bootstrap Stichprobe dem Ziehen mit Zurücklegen von n Größen aus der Ausgangsstichprobe entspricht.
Man berechnet dann den Ausdruck (1.2) in dem bootstrap Modell, bestimmt also

$$\mathbb{P}(T(\hat{\mathbb{P}}_n^*) - T(\hat{\mathbb{P}}_n) \leq t),\ t \in \mathbb{R}. \tag{1.3}$$

Der Ausdruck (1.3) ist der bootstrap Schätzer für (1.2) und ist (im Prinzip) genau berechenbar, da er nur von den beobachteten Daten abhängt. Zum Beispiel lassen sich unmittelbar die (theoretischen!) bedingten Momente von Bootstrap-Zufallsvariablen ausrechnen.

Satz 1.10 (Bedingte Momente von bootstrap Größen)
Es sei $X = (X_1, \ldots, X_n)$ ein Vektor von i.i.d. Original-Variablen. Dann gilt bedingt unter X:

$$\mathbb{E}^*[X_1^*|X] = \frac{1}{n}\sum_{i=1}^{n} X_i =: \bar{X}_n \tag{1.4}$$

$$\mathbb{E}^*[\frac{1}{m(n)}\sum_{i=1}^{m(n)} X_i^*|X] = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n \tag{1.5}$$

$$\mathbb{E}^*[{X_1^*}^2|X] = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 \tag{1.6}$$

$$\mathit{Var}\,(X_1^*|X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{1.7}$$

$$\mathit{Var}\left(\frac{1}{m(n)}\sum_{i=1}^{m(n)} X_i^*|X\right) = \frac{1}{n\cdot m(n)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{1.8}$$

$$\mathbb{E}^*[{X_1^*}^3|X] = \frac{1}{n}\sum_{i=1}^{n} {X_i}^3 \tag{1.9}$$

Beweis: Zur Übung. ■

Betrachten wir zur Komplettierung der Motivation von Bootstrapverfahren nun die Konvergenzrate im Zentralen Grenzwertsatz, um zu einer Aussage über die zu erreichende Approximationsgenauigkeit des asymptotischen Tests φ_n^{as} zu gelangen.

Satz 1.11 (Satz von Berry-Esséen)
Seien $(X_i)_{i\in\mathbb{N}}$ stochastisch unabhängige, reellwertige Zufallsvariablen mit $0 < \mathit{Var}\,(X_i) < \infty$ für alle $i \in \mathbb{N}$. Bezeichne F_n die Verteilungsfunktion der standardisierten Summe

$$\frac{\sum_{i=1}^{n}(X_i - \mathbb{E}\,[X_i])}{\sqrt{\sum_{i=1}^{n}\mathit{Var}\,(X_i)}}.$$

Dann gilt:

$$\sup_{x\in\mathbb{R}}|F_n(x) - \Phi(x)| \le \frac{6}{s_n^3}\cdot\sum_{i=1}^{n}\mathbb{E}\left[|X_i|^3\right],$$

wobei Φ die Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung bezeichnet und $s_n^2 = \sum_{i=1}^{n}\mathit{Var}\,(X_i)$ gilt.

Liegen iid. Variablen X_i vor, so ergibt sich damit die folgende Abschätzung:

$$\sup_{x\in\mathbb{R}}|F_n(x) - \Phi(x)| \le \frac{6}{\sqrt{n}\cdot\mathit{Var}\,(X_1)^{\frac{3}{2}}}\cdot\mathbb{E}\left[|X_1|^3\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Beweis: Klassisches Resultat, siehe, z. B., Gaenssler and Stute (1977). ■

Bemerkung 1.12

Damit ein bootstrap Test dem asymptotischen Test φ_n^{as} in Sachen Niveaueinhaltung überlegen ist, muss die Konvergenzgeschwindigkeit der bootstrap Verteilung in gewisser Weise schneller sein als die „worst case" Rate $\sqrt{n}$ im zentralen Grenzwertsatz. Dies ist auch tatsächlich der Fall, wie das Buch von Hall (1992) mit Hilfe von asymptotischen (Edgeworth-)Entwicklungen nachweist. Hall argumentiert, dass durch den bootstrap eine automatische Bias-Korrektur vorgenommen wird. Technisch bedeutet das, dass der Term, der durch die dritte Kumulante von X_1 bestimmt wird, in der Edgeworth-Entwicklung der bootstrap Verteilungsfunktion verschwindet.

Für Zweistichprobenprobleme kann man sich eine andere Überlegung zu Nutze machen, um zu einer Resamplingmethode zu gelangen. Dazu betrachten wir wieder stochastisch unabhängige Zufallsvariablen $(X_1, \ldots, X_n)$. Wir nehmen an, dass (für eine festgelegte Zahl $2 \leq n_1 \leq n-2$) die $(X_i)_{i=1,\ldots,n_1}$ identisch nach der Verteilung mit Verteilungsfunktion F_1 (Gruppe 1) und die $(X_j)_{j=n_1+1,\ldots,n}$ identisch nach der Verteilung mit Verteilungsfunktion F_2 (Gruppe 1) verteilt sind. Das interessierende (nichtparametrische) Testproblem ist dann gegeben als $H_0 : F_1 = F_2$ gegen $H_1 : F_1 \neq F_2$. Unter der Nullhypothese H_0 sollten sich nun wichtige gruppenspezifische Charakteristika einer empirisch erhobenen Stichprobe, die sich als eine Realisierung unter dem vorstehenden Modell beschreiben lässt, nicht zu stark ändern, wenn die Gruppenzugehörigkeit zufällig „ausgewürfelt" wird, also jedem beobachteten Wert aus $(x_1, \ldots, x_n)$ ein zufälliger Gruppenindikator angeheftet wird. Halten wir wie zuvor angedeutet die Plätze $i = 1, \ldots, n_1$ für die Gruppe 1 fest, so entspricht dieses „label shuffling" offensichtlich einem zufälligen Ziehen ohne Zurücklegen aus $(x_1, \ldots, x_n)$ und Verteilung der Werte auf die Plätze von 1 bis n. Mathematisch ist dies äquivalent zu einer Permutation der Werte $(x_1, \ldots, x_n)$. Genau diese Idee liegt den sogenannten Permutationstests zu Grunde. Betrachtet man zum Beispiel speziell Lageparametermodelle (Gruppe 1 ist unter der Alternative bezüglich eines gewissen Kriteriums besser als Gruppe 2), so kann ein Permutationstest z. B. die Differenz der arithmetischen Gruppenmittel der Original-Stichprobe als Teststatistik benutzen und sie mit einem emprischen Quantil der Differenzen von arithmetischen Resampling-Gruppenmittelwerten vergleichen, die durch das Ausführen von einer festgelegten Anzahl B von Permuationen $\sigma \in \mathcal{S}_n$ zu Stande kommen.

Das Ziel der folgenden Abschnitte dieses Kapitels ist es, die vorgenannten heuristischen Überlegungen zu Bootstrap- und Permutationstests auf eine solide mathematische Grundlage zu stellen. Kapitel 2 stellt dann die praktische Umsetzbarkeit der resultierenden Methoden in den Vordergrund. Die Kapitel 3 bis 5 gehen auf spezielle nicht-Standard Probleme ein, die mit Resamplingverfahren bearbeitet werden können.

1.3 L_1-Differenzierbarkeit und lokal beste Tests

Das Testen von zusammengesetzten Nullhypothesen bzw. Alternativen ist ein nicht-triviales Problem in der Inferenzstatistik. Nur in Spezialfällen (z.B. monotoner Dichtequotient, verallgemeinerte Neyman-Pearson-Theorie) ist eine zufriedenstellende generelle Methodik verfügbar, die zu gleichmäßig (über $\vartheta \in H_1$) besten Niveau-α-Tests führt.
Ist die „Geometrie“ des Parameterraums indes komplizierter, so kann die Typ-II-Fehlerwahrscheinlichkeit (unter Maßgabe der Einhaltung des Signifikanzniveaus) typischerweise nicht gleichmäßig minimiert werden und es ist eine Auswahl an konkurrierenden Testverfahren notwendig. Oftmals kommt es entscheidend darauf an, gegen welche Art von Alternativen man sich bestmöglich absichern möchte, d.h., gegen welche „Regionen“ von H_1 man größtmögliche Trennschärfe anstrebt.
Eine Klasse von Verfahren bilden die sogenannten lokal besten Tests. Hierbei wird Trennschärfemaximierung in Regionen „nahe bei H_0“ angestrebt. Zu ihrer Anwendbarkeit benötigt man das Konzept der L_1-Differenzierbarkeit von statistischen Modellen.

Definition 1.13 (L_1-Differenzierbarkeit)
Sei $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$. Die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ sei dominiert, d.h. $\forall\vartheta \in \Theta : \mathbb{P}_\vartheta \ll \mu$ für ein Maß μ auf $(\Omega, \mathcal{A})$. Dann heißt $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ L_1-differenzierbar in $\vartheta_0 \in \mathring{\Theta}$, falls $\exists g \in L_1(\mu)$ mit

$$\left\| t^{-1}\left(\frac{d\mathbb{P}_{\vartheta_0+t}}{d\mu} - \frac{d\mathbb{P}_{\vartheta_0}}{d\mu}\right) - g \right\|_{L_1(\mu)} \longrightarrow 0 \quad \text{für } t \to 0.$$

Die Funktion g heißt $L_1(\mu)$-Ableitung von $\vartheta \mapsto \mathbb{P}_\vartheta$ in ϑ_0.

Zur Vereinfachung der Notation sei von nun an oft ohne explizite Erwähnung und o.B.d.A $\vartheta_0 \equiv 0$.

Satz 1.14 (§18 in Hewitt and Stromberg (1975), Satz 1.183 in Witting (1985))
Unter den Voraussetzungen von Definition 1.13 sei $\vartheta_0 = 0$ und seien $f_\vartheta(x) := \frac{d\mathbb{P}_\vartheta}{d\mu}(x)$ Versionen der Dichten mit folgenden Eigenschaften:

(a) Es gibt eine offene Umgebung $\mathcal{U}$ von 0, so dass für μ-fast alle x die Abbildung $\mathcal{U} \ni \vartheta \mapsto f_\vartheta(x)$ absolut stetig ist, d.h., es existiert eine integrierbare Funktion $\tau \mapsto \dot{f}(x,\tau)$ auf $\mathcal{U}$ mit

$$\int_{\vartheta_1}^{\vartheta_2} \dot{f}(x,\tau)d\tau = f_{\vartheta_2}(x) - f_{\vartheta_1}(x), \quad \vartheta_1 < \vartheta_2$$

und es sei $\frac{\partial}{\partial\vartheta} f_\vartheta(x)|_{\vartheta=0} = \dot{f}(x,0)$ μ-fast überall.

(b) Für $\vartheta \in \mathcal{U}$ sei $x \mapsto \dot{f}(x,\vartheta)$ μ-integrierbar mit

$$\int \left|\dot{f}(x,\vartheta)\right| d\mu(x) \overset{\vartheta\to 0}{\longrightarrow} \int \left|\dot{f}(x,0)\right| d\mu(x).$$

Dann ist $\vartheta \mapsto \mathbb{P}_\vartheta$ in 0 $L_1(\mu)$-differenzierbar mit $g = \dot{f}(\cdot, 0)$.

Grob gesagt erhält man also im absolutstetigen Fall die L_1-Ableitung einfach durch analytisches Differenzieren der Dichte nach dem Parameter. Eine andere wichtige Anwendung von Satz 1.14 ist die Bearbeitung von Lageparametermodellen wie in Beispiel 1.16.

Satz 1.15 (Satz und Definition)
Unter den Voraussetzungen von Definition 1.13 seien die Dichten $\vartheta \mapsto f_\vartheta$ im Nullpunkt ($\vartheta_0 = 0$) $L_1(\mu)$-differenzierbar mit einer Ableitung g.

(a) Dann konvergiert für $\vartheta \to 0$ $\quad \vartheta^{-1} \log \frac{f_\vartheta}{f_0}(x) = \vartheta^{-1}(\log f_\vartheta(x) - \log f_0(x)) \quad$ $\mathbb{P}_0$-stochastisch gegen (sagen wir) $\dot{L}(x)$.
$\dot{L}$ heißt Ableitung des (logarithmischen) Dichtequotienten bzw. Score-Funktion. Ferner gilt $\dot{L}(x) = \frac{g(x)}{f_0(x)}$.

(b) $\int \dot{L} d\mathbb{P}_0 = 0$ und $\{f_0 = 0\} \subseteq \{g = 0\}$ $\mathbb{P}_0$-fast sicher.

Beweis:

(a) $\vartheta^{-1}(\frac{f_\vartheta}{f_0} - 1) \longrightarrow \frac{g}{f_0}$ konvergiert in $L_1(\mathbb{P}_0)$ und daher $\mathbb{P}_0$-stochastisch. Die Kettenregel liefert das Resultat.

(b) Nach dem Satz von Vitali (Satz 1.9 hier im Skript) folgt ($\vartheta \to 0$ entlang einer geeignet gewählten Teilfolge), dass $\int (f_\vartheta - f_0) d\mu = 0$ gilt. Damit folgt $\int \dot{L} d\mathbb{P}_0 = \int g d\mu = 0$.

■

Beispiel 1.16 *(a) Lageparametermodell*
Sei $X = \vartheta + Y, \vartheta \geq 0$, und habe Y die Dichte f, wobei f absolutstetig bezüglich des Lebesguemaßes λ und ϑ-frei sei. Dann sind die Dichten $\vartheta \mapsto f(x - \vartheta)$ von X unter ϑ $L_1(\lambda)$-differenzierbar in 0 mit Scorefunktion $\dot{L}(x) = -\frac{f'(x)}{f(x)}$ (Differentiation nach x).

(b) Skalenparametermodell
Sei $X = \exp(\vartheta) Y$, Y habe absolutstetige ϑ-freie Dichte f und es gelte $\int |x f'(x)| dx < \infty$. Dann sind die Dichten $\vartheta \mapsto \exp(-\vartheta) f(x \exp(-\vartheta))$ von X unter ϑ $L_1(\lambda)$-differenzierbar in 0 mit Score-Funktion $\dot{L}(x) = -(1 + \frac{x f'(x)}{f(x)})$.

Beides folgt sofort aus den Sätzen 1.14 und 1.15 zusammen mit der Translationsäquivarianz des Lebesguemaßes.
Beachte: $\vartheta^{-1}(f(x - \vartheta) - f(x)) \overset{\vartheta \to 0}{\longrightarrow} -f'(x)$ *λ-fast überall.*

Lemma 1.17
Seien $\vartheta \mapsto \mathbb{P}_\vartheta$ eine $L_1(\mu)$-differenzierbare Familie mit Score-Funktion $\dot{L}$ in $\vartheta_0 = 0$ und c_i, $1 \leq i \leq n$ reelle Konstanten. Dann ist auch $\vartheta \mapsto \bigotimes_{i=1}^n \mathbb{P}_{c_i \vartheta}$ im Nullpunkt $L_1(\mu)$-differenzierbar mit Scorefunktion $(x_1, \ldots, x_n) \mapsto \sum_{i=1}^n c_i \dot{L}(x_i)$.

Beweis: Zur Übung. ■

Anmerkung: Ist das Modell L_2-differenzierbar, so liegt $\dot{L}$ in $L_2(\mathbb{P}_0)$ und wird auch Tangentialvektor oder Einflusskurve genannt (vgl. auch Abschnitt 3.5 Mathematische Statistik, Markus Reiß).

Definition 1.18 (Score-Test)
Sei $\vartheta \mapsto \mathbb{P}_\vartheta$ L_1-differenzierbar in ϑ_0 mit Score-Funktion $\dot{L}$. Dann heißt jeder Test ψ von der Form

$$\psi(x) = \begin{cases} 1, & \text{falls } \dot{L}(x) > \tilde{c} \\ \gamma, & \text{falls } \dot{L}(x) = \tilde{c} \\ 0, & \text{falls } \dot{L}(x) < \tilde{c} \end{cases}$$

ein Score-Test. Dabei ist $\gamma \in [0,1]$ eine Randomisationskonstante.

Definition 1.19 (Lokal bester Test)
Sei $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta \subseteq \mathbb{R}$ L_1-differenzierbar in $\vartheta_0 \in \mathring{\Theta}$. Ein $\{\vartheta_0\}$ α-ähnlicher Test φ^* heißt lokal bester $\{\vartheta_0\}$ α-ähnlicher Test für $\tilde{H} = \{\vartheta_0\}$ gegen $K = \Theta \cap \{\vartheta > \vartheta_0\}$, falls gilt

$$\frac{d}{d\vartheta}\mathbb{E}_\vartheta\left[\varphi^*\right]|_{\vartheta=\vartheta_0} \geq \frac{d}{d\vartheta}\mathbb{E}_\vartheta\left[\varphi\right]\big|_{\vartheta=\vartheta_0}$$

für alle $\{\vartheta_0\}$ α-ähnlichen Tests φ, d.h. für alle Tests φ mit $\mathbb{E}_{\vartheta_0}\left[\varphi\right] = \alpha$.

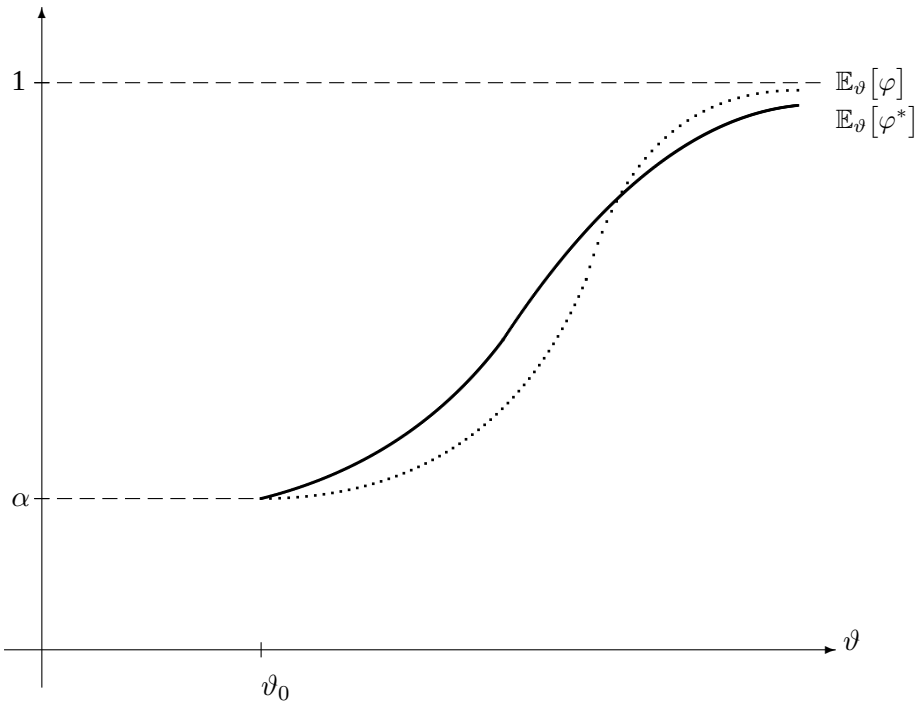

Abbildung 1.2: Lokal bester $\{\vartheta_0\}$ α-ähnlicher Test φ^*

Anmerkung: Lokal beste Tests können für ϑ-Werte, die weit entfernt von ϑ_0 liegen, schlechte Eigenschaften haben.

Satz 1.20 (Satz 2.44 in Witting (1985))
Unter den Voraussetzungen von Definition 1.19 ist der Score-Test

$$\psi(x) = \begin{cases} 1, & \textit{falls } \dot{L}(x) > c(\alpha) \\ \gamma, & \textit{falls } \dot{L}(x) = c(\alpha), \gamma \in [0,1] \\ 0, & \textit{falls } \dot{L}(x) < c(\alpha) \end{cases}$$

mit $\mathbb{E}_{\vartheta_0}[\psi] = \alpha$ ein $\{\vartheta_0\}$ α-ähnlicher, lokal bester Test für $\tilde{H} = \{\vartheta_0\}$ gegen $K = \{\vartheta > \vartheta_0\}$.

Zumindest lokal um ϑ_0 sind die Score-Tests also ein vernünftiger „Ersatz“ für Neyman-Pearson Tests, wenn kein monotoner Dichtequotient vorliegt. Für Einstichprobenprobleme ist die Anwendung sofort einsichtig.
Liege eine Stichprobe $(x_1, \ldots, x_n)$ vor, die als Realisierung von $(X_1, \ldots, X_n)$ iid mit f_ϑ als Dichte von X_1 aufgefasst werde, also $f_\vartheta(x) = \frac{d\mathbb{P}_\vartheta}{d\mu(x)}$.
Das Produktexperiment mit Produktmaß $\mathbb{P}_\vartheta^n$ hat nach Lemma 1.17 die Score-Funktion $(x_1, \ldots, x_n) \mapsto \sum_{i=1}^n \dot{L}(x_i)$.
Sind wir am einseitigen Test $\tilde{H} = \{\vartheta_0\}$ gegen $K = \{\vartheta \in \Theta : \vartheta > \vartheta_0\}$ interessiert, so lehnen wir $\tilde{H}$ ab, falls $\sum_{i=1}^n \dot{L}(x_i) > c(\alpha)$ gilt.
Für Mehrstichprobenprobleme ($k \geq 2$ Gruppen) betrachten wir die nichtparametrische Hypothese

$$H_0 : \{\mathbb{P}^{X_1} = \mathbb{P}^{X_2} = \ldots = \mathbb{P}^{X_n} : \mathbb{P}^{X_1} \text{ stetig}\} \tag{1.10}$$

Die Idee ist nun, zunächst einparametrige Kurven $\vartheta \mapsto \mathbb{P}_{n,\vartheta}$ zu studieren, die nur für $\vartheta = 0$ in H_0 liegen ($\mathbb{P}_{n,0} \in H_0$). Für $\vartheta \neq 0$ besteht $\mathbb{P}_{n,\vartheta}$ im Allgemeinen aus einem Produktmaß mit nicht identischen Faktoren.

Beispiel 1.21 *(a) Regressionmodell für einen Lageparameter*

Seien $X_i = c_i\vartheta + Y_i, 1 \leq i \leq n, \vartheta \geq 0$. Die Y_i seien iid mit einer Lebesgue-Dichte f (ϑ-frei!). Für das Zweistichprobenproblem z.B. setzen wir nun $c_1 = c_2 = \cdots = c_{n_1} = 1$ und $c_i = 0 \, \forall n_1 + 1 \leq i \leq n$. Damit unterscheidet sich die erste Gruppe (Plätze $1, \ldots, n_1$) von der zweiten Gruppe unter Alternativen ($\vartheta > 0$) durch einen positiven Shift.

(b) Regressionsmodell für einen Skalenparameter

Seien c_i reelle Regressionskoeffizienten, $X_i = \exp(c_i\vartheta)Y_i, 1 \leq i \leq n, \vartheta \in \mathbb{R}$. Die Y_i seien iid mit der ϑ-freien Lebesguedichte f. Dann ist

$$\frac{d\mathbb{P}_{n,\vartheta}}{d\lambda^n}(x) = \prod_{i=1}^n \exp(-c_i\vartheta) f(x_i \exp(-c_i\vartheta)).$$

Unter $\vartheta_0 = 0$ liegt obiges Produktmaß offenbar in H_0, unter Alternativen nicht.

(c) *Allgemeines Modell*

Sei $\vartheta \mapsto \mathbb{P}_\vartheta$ eine einparametrige Kurve von Verteilungen mit reellem Parameter ϑ. Setze $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^n \mathbb{P}_{c_i\vartheta}$.

1.4 Einige Rangtests

Satz 1.22

Sei $\vartheta \mapsto \mathbb{P}_\vartheta$ $L_1(\mu)$-differenzierbar im Nullpunkt ($\vartheta_0 = 0$) mit Score-Funktion $\dot{L}$. Ferner sei $S : \Omega \to \Omega'$ eine Statistik. Dann ist $\vartheta \mapsto \mathbb{P}_\vartheta^S$ (Bildmaß unter S) $L_1(\mu^S)$-differenzierbar mit Score-Funktion $y \mapsto \mathbb{E}_{\mathbb{P}_0}\big[\dot{L} \mid S = y\big]$.

Beweis: O.B.d.A. sei μ ein Wahrscheinlichkeitsmaß und es gelte

$$\vartheta^{-1}(f_\vartheta - f_0) \longrightarrow g \quad \text{ in } L_1(\mu) \quad \text{ für } \vartheta \to 0. \tag{1.11}$$

Allgemein gilt (Stochastik II):

$$Q \ll P \implies \frac{dQ^T}{dP^T}(t) = \mathbb{E}_P\big[\frac{dQ}{dP} \mid T = t\big]$$

für Wahrscheinlichkeitsmaße P und Q und eine Statistik T. Also haben wir

$$\frac{d\mathbb{P}_\vartheta^S}{d\mu^S}(y) = \mathbb{E}_\mu\big[f_\vartheta \mid S = y\big].$$

Damit gilt

$$\begin{aligned}
&\int \big|\vartheta^{-1}(\frac{d\mathbb{P}_\vartheta^S}{d\mu^S} - \frac{d\mathbb{P}_0^S}{d\mu^S}) - \mathbb{E}_\mu\big[g \mid S = y\big]\big| d\mu^S(y) \\
&\qquad = \int \big|\mathbb{E}_\mu\big[\vartheta^{-1}(f_\vartheta - f_0) - g \mid S\big]\big| d\mu && \text{(Linearität von } \mathbb{E}_\mu\big[\cdot \mid S\big]\text{)} \\
&\qquad \le \int \mathbb{E}_\mu\big[\big|\vartheta^{-1}(f_\vartheta - f_0) - g\big| \mid S\big] d\mu && \text{(Dreiecksungleichung)} \\
&\qquad \overset{\vartheta\to 0}{\longrightarrow} 0 && \text{((1.11), Satz von Vitali)}
\end{aligned}$$

Also besitzt $\mathbb{P}_\vartheta^S$ die Score-Funktion $y \mapsto \frac{\mathbb{E}_\mu\left[g|S=y\right]}{\mathbb{E}_\mu\left[\frac{d\mathbb{P}_0}{d\mu}|S=y\right]}$ nach der Kettenregel ($\frac{d}{dx}\ln(f(x)) = \frac{f'(x)}{f(x)}$). Nach Satz 1.14 (a) gilt zudem $g = \dot{L}\frac{d\mathbb{P}_0}{d\mu}$. Es bleibt zu zeigen:

$$\mathbb{E}_\mu\big[\dot{L}\frac{d\mathbb{P}_0}{d\mu} \mid S\big] = \mathbb{E}_{\mathbb{P}_0}\big[\dot{L} \mid S\big]\mathbb{E}_\mu\big[\frac{d\mathbb{P}_0}{d\mu} \mid S\big] \qquad \mu\text{ -fast sicher.}$$

Dazu sei $A \subset \Omega'$ eine beliebige messbare Menge. Wir rechnen nach (von rechts nach links):

$$\begin{aligned}
\int \mathbf{1}_A(S)\mathbb{E}_{\mathbb{P}_0}\big[\dot{L} \mid S\big]\mathbb{E}_{\mu}\big[\frac{d\mathbb{P}_0}{d\mu} \mid S\big]d\mu &= \int \mathbf{1}_A(S)\mathbb{E}_{\mathbb{P}_0}\big[\dot{L} \mid S\big]\frac{d\mathbb{P}_0}{d\mu}d\mu \qquad \text{(tower equation)} \\
&= \int \mathbf{1}_A(S)\mathbb{E}_{\mathbb{P}_0}\big[\dot{L} \mid S\big]d\mathbb{P}_0 \\
&= \int \mathbf{1}_A(S)\dot{L}d\mathbb{P}_0 \qquad \text{(tower equation)} \\
&= \int \mathbf{1}_A(S)\dot{L}\frac{d\mathbb{P}_0}{d\mu}d\mu.
\end{aligned}$$

■

Wir werden Satz 1.22 benutzen, um von den parametrischen Kurven $\mathbb{P}_{n,\vartheta}$ wie in Beispiel 1.21 auf Rangtests zu kommen. Es wird sich zeigen, dass die Vergröberung der Information (nur Ränge, nicht die Werte der X_i fließen in die Datenanalyse ein) zu einer einfachen Struktur der Score-Teststatistiken führt (einfache lineare Rangstatistik). Ferner haben Ränge den Vorteil, robuster gegenüber Modell-Fehlspezifikationen zu sein. Oftmals sind auch nur Ränge beobachtbar oder vertrauenswürdig.
Es bleibt natürlich der Kritikpunkt, dass man bei tatsächlichem Vorliegen eines parametrischen Modells einen Verlust an Trennschärfe in Kauf nehmen muss, also höhere Stichprobenumfänge für gleiche Güte benötigt. Effizienzrechnungen können die zu erwartenden Stichprobenumfangserhöhungen quantifizieren.
Zur Vorbereitung sammeln wir Basiswissen zu Rang- und Orderstatistiken. Wir verzichten auf Beweise und verweisen auf §1 und §2 in Janssen (1998) oder andere einschlägige Literatur.

Definition 1.23
Sei $x = (x_1, \ldots, x_n)$ ein Punkt im $\mathbb{R}^n$, die x_i seien paarweise verschieden. Seien $x_{1:n} < x_{2:n} < \ldots < x_{n:n}$ die geordneten Werte der x_i.

(a) Für $1 \le i \le n$ heißt $r_i \equiv r_i(x) := \#\{j \in \{1, \ldots, n\} : x_j \le x_i\}$ der Rang von x_i (in x). Der Vektor $r(x) := (r_1(x), \ldots, r_n(x)) \in \mathcal{S}_n$ heißt Rangvektor von x.
($\mathcal{S}_n$: symmetrische Gruppe)

(b) Die inverse Permutation $d(x) := [r(x)]^{-1}$ heißt der Antirangvektor von x, $d(x) =: (d_1(x), \ldots, d_n(x))$, die Zahl $d_i(x)$ heißt der Antirang von i (Index, der zur i-ten kleinsten Beobachtung gehört)

Seien nun $X_1, \ldots, X_n$ mit $X_i : \Omega_i \to \mathbb{R}$ stochastisch unabhängige, stetig verteilte Zufallsvariablen. Bezeichne $\mathbb{P}$ die gemeinsame Verteilung von $(X_1, \ldots, X_n)$.

(c) Da $\mathbb{P}(\bigcup_{i \ne j}\{X_i = X_j\}) = 0$ gilt, können wir $\mathbb{P}$-fast sicher eindeutig die folgenden Größen definieren:
$X_{i:n}$ heißt i-te Orderstatistik von $X = (X_1, \ldots, X_n)$,

$R_i(X) := n\hat{F}_n(X_i) = r_i(X_1, \ldots, X_n)$ *heißt Rang von* X_i,
$D_i(X) := d_i(X_1, \ldots, X_n)$ *heißt Antirang von* i *bezüglich* X *und*
$D(X) := d(X)$ *heißt Antirangvektor zu* X.

Lemma 1.24

Voraussetzungen wir unter Definition 1.23.

(a) $i = r_{d_i} = d_{r_i}, \quad x_i = x_{r_i:n}, \quad x_{i:n} = x_{d_i}$

(b) Sind $X_1, \ldots, X_n$ *austauschbar (gilt natürlich speziell bei iid.), so ist*

$$R(X) : \bigtimes_{i=1}^{n} \Omega_i =: \Omega \to \mathcal{S}_n$$

gleichverteilt auf $\mathcal{S}_n$*, also* $\mathbb{P}(R(X) = (r_1, \ldots, r_n)) = \frac{1}{n!}$ *für alle* $\sigma = (r_1, \ldots, r_n) \in \mathcal{S}_n$.

(c) Sind $U_1, \ldots, U_n$ *iid. mit* $U_1 \sim$ *UNI[0,1], und ist* $X_i = F^{-1}(U_i) \quad \forall 1 \leq i \leq n$*, dann gilt* $X_{i:n} = F^{-1}(U_{i:n})$.
Ist die Verteilungsfunktion F *von* X_1 *stetig, so gilt* $R(X) = R(U)$.

(d) Sind $(X_1, \ldots, X_n)$ *iid. mit Verteilungsfunktion* F *von* X_1*, so gilt:*

(i) $\mathbb{P}(X_{i:n} \leq x) = \sum_{j=i}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$

(ii) $\frac{d\mathbb{P}^{X_{i:n}}}{d\mathbb{P}^{X_1}}(x) = n\binom{n-1}{i-1} F(x)^{i-1}(1 - F(x))^{n-i}$.
Besitzt $\mathbb{P}^{X_1}$ *Lebesgue-Dichte* f*, so besitzt* $\mathbb{P}^{X_{i:n}}$ *Lebesguedichte* $f_{i:n}$*, gegeben durch*

$$f_{i:n}(x) = n\binom{n-1}{i-1} F(x)^{i-1}(1 - F(x))^{n-i} f(x)$$

(iii) Sei $\mu := \mathbb{P}^{X_1}$*. Dann besitzt* $(X_{i:n})_{i \leq n}$ *die gemeinsame* μ^n*-Dichte*

$$(x_1, \ldots, x_n) \mapsto n!\, \mathbf{1}_{\{x_1 < x_2 < \ldots < x_n\}}.$$

Besitzt μ *die Lebesguedichte* f*, so besitzt* $(X_{i:n})_{1 \leq i \leq n}$ *die* λ^n*-Dichte*

$$(x_1, \ldots, x_n) \mapsto n! \prod_{i=1}^{n} f(x_i)\, \mathbf{1}_{\{x_1 < x_2 < \ldots < x_n\}}.$$

Bemerkung 1.25

Lemma 1.24(c) (Quantilstransformation) zeigt die besondere Bedeutung der Verteilung der Orderstatistiken von iid. UNI[0,1]-verteilten Zufallsvariablen $U_1, \ldots, U_n$.
$U_{i:n}$ *besitzt nach Lemma 1.24(d) eine Beta*$(i, n - i + 1)$*-Verteilung mit* $\mathbb{E}[U_{i:n}] = \frac{i}{n+1}$ *und Var* $(U_{i:n}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}$.
Für die Berechnung der gemeinsamen Verteilungsfunktion von $(U_{1:n}, \ldots, U_{n:n})$ *existieren effiziente rekursive Algorithmen, inbesondere die Bolshev-Rekursion und die Steck-Rekursion (Shorack and Wellner (1986), S.362 ff.).*

Satz 1.26

Seien $X_1, \ldots, X_n$ reelle iid. Zufallsvariablen mit stetigem $\mu = \mathbb{P}^{X_1}$. Sei $X = (X_1, \ldots, X_n)$.

(a) $R(X)$ und $(X_{i:n})_{1 \le i \le n}$ sind stochastisch unabhängig.

(b) Sei $T : \mathbb{R}^n \to \mathbb{R}$ eine Statistik. Die Statistik $T(X)$ sei integrierbar. Für $\sigma = (r_1, \ldots, r_n) \in \mathcal{S}_n$ gilt

$$\mathbb{E}\big[T(X) \mid R(X) = \sigma\big] = \mathbb{E}\big[T((X_{r_i:n})_{1 \le i \le n})\big]$$

Beweis: zu (a): Seien $\sigma = (r_1, \ldots, r_n) \in \mathcal{S}_n$ und $A_i \in \mathcal{B}(\mathbb{R})$ für $1 \le i \le n$ beliebig gewählt. $(d_1, \ldots, d_n) := \sigma^{-1}$.
Wir beachten

$$X_{d_i} = X_{i:n} \in A_i \iff X_i \in A_{r_i} \quad \text{und} \quad R(X) = \sigma \iff X_{d_1} < X_{d_2} < \ldots X_{d_n}.$$

Es sei $B := \{x \in \mathbb{R}^n : x_1 < x_2 < \ldots < x_n\}$. Dann ergibt sich für die gemeinsame Verteilung von Rängen und Orderstatistiken:

$$\begin{aligned}
\mathbb{P}\big(R(X) = \sigma, X_{i:n} \in A_i \,\forall 1 \le i \le n\big) &= \mathbb{P}\big(\forall 1 \le i \le n : X_{d_i} \in A_i, (X_{d_i})_{1 \le i \le n} \in B\big), \\
&= \int_{\times_{i=1}^n A_{r_i}} \mathbf{1}_B(x_{d_1}, \ldots, x_{d_n}) d\mu^n(x_1, \ldots, x_n) \\
&= \int_{\times_{i=1}^n A_{r_i}} \mathbf{1}_B(x_1, \ldots, x_n) d\mu^n(x_1, \ldots, x_n),
\end{aligned}$$

da wegen Austauschbarkeit μ^n invariant unter der Transformation $(x_1, \ldots, x_n) \mapsto (x_{d_1}, \ldots, x_{d_n})$ ist. Summiert man über alle $\sigma \in \mathcal{S}_n$, so folgt

$$\mathbb{P}\big(X_{i:n} \in A_i \,\forall 1 \le i \le n\big) = \int_{\times_{i=1}^n A_{r_i}} n!\, \mathbf{1}_B(x_1, \ldots, x_n) d\mu^n(x_1, \ldots, x_n).$$

Wegen Lemma 1.24(b) ist demnach

$$\mathbb{P}\big(R(X) = \sigma, X_{i:n} \in A_i \,\forall 1 \le i \le n\big) = \mathbb{P}\big(R(X) = \sigma\big)\, \mathbb{P}\big(\forall 1 \le i \le n : X_{i:n} \in A_i\big).$$

zu (b):

$$\begin{aligned}
\mathbb{E}\big[T(X) \mid R(X) = \sigma\big] &= \int_{\{R(X)=\sigma\}} \frac{T(X)}{\mathbb{P}(R(X) = \sigma)} d\mathbb{P} \\
&= \mathbb{E}\big[T((X_{r_i:n})_{1 \le i \le n}) \mid R(X) = \sigma\big] \qquad (*) \\
&= \mathbb{E}\big[T((X_{r_i:n})_{1 \le i \le n})\big] \qquad (\text{(a)})
\end{aligned}$$

$(*)$ gilt, da auf der Menge $\{R(X) = \sigma\}$ offenbar die Beziehung $X = (X_{r_i:n})_{i=1}^n$ gilt. ■

Nach diesem längeren Exkurs kehren wir zurück zu den Score-Tests.

Korollar 1.27 (zu Satz 1.22 mit Lemma 1.17)
Sei $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta \subseteq \mathbb{R}$ eine Familie von im Nullpunkt $L_1(\mu)$-differenzierbaren Verteilungen (μ dominierendes Maß von $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$) mit Score-Funktion $\dot{L}$ in $\vartheta_0 = 0$. Sei $X = (X_1, \ldots, X_n)$ nach $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^n \mathbb{P}_{c_i\vartheta}$ verteilt. Dann besitzt $\mathbb{P}_{n,\vartheta}^R$ die Score-Funktion

$$\begin{aligned}
\sigma = (r_1, \ldots, r_n) \longmapsto \quad & \mathbb{E}_{\mathbb{P}_{n,0}}\big[\sum_{i=1}^n c_i \dot{L}(X_i) \mid R(X) = \sigma\big] \\
&= \sum_{i=1}^n c_i \mathbb{E}_{\mathbb{P}_{n,0}}\big[\dot{L}(X_i) \mid R(X) = \sigma\big] \\
&= \sum_{i=1}^n c_i \mathbb{E}_{\mathbb{P}_{n,0}}\big[\dot{L}(X_{r_i:n})\big] \qquad \text{(Satz 1.26(b))} \\
&=: \sum_{i=1}^n c_i a(r_i)
\end{aligned}$$

mit $a(i) = \mathbb{E}_{\mathbb{P}_{n,0}}\big[\dot{L}(X_{i:n})\big]$.

Bemerkung 1.28 *(a) Die Gewichte $a(i)$ heißen „Scores“ (entsprechen Punktzahlen in sportlichen Wettbewerben).*

(b) Die nichtparametrische Hypothese H_0 aus (1.10) *führt unter $R(X)$ zu einer einelementigen Nullhypothese auf $\mathcal{S}_n$, nämlich der Gleichverteilung auf $\mathcal{S}_n$ (siehe Lemma 1.24(b)). Damit können die kritischen Werte $c(\alpha)$ für den resultierenden Rangtest $\psi \equiv \psi(R(X))$, gegeben durch*

$$\psi(x) = \begin{cases} 1, \text{ falls } \sum_{i=1}^n c_i a(R_i(x)) > c(\alpha), \\ \gamma, \text{ falls } \sum_{i=1}^n c_i a(R_i(x)) = c(\alpha), \\ 0, \text{ falls } \sum_{i=1}^n c_i a(R_i(x)) < c(\alpha), \end{cases} \tag{1.12}$$

durch diskrete Erwartungswertbildung ermittelt werden. Für großes n kann $c(\alpha)$ approximiert werden, indem eine Zahl $B < n!$ festgesetzt wird und nur B zufällig ausgewählte Permutationen $\sigma \in \mathcal{S}_n$ traversiert werden.

(c) Die Teststatistik $T \equiv T(R(X)) = \sum_{i=1}^n c_i a(R_i(X))$ heißt einfache lineare Rangstatistik.

(d) Für die Scores gilt $\sum_{i=1}^n a(i) = 0$ (zur Übung, einfach).
Ist $\dot{L}$ isoton, so gilt $a(1) \le a(2) \le \ldots \le a(n)$.

(e) Wegen $X_{i:n} \overset{\mathcal{D}}{=} F^{-1}(U_{i:n})$ werden die Scores häufig in der Form $a(i) = \mathbb{E}\big[\dot{L} \circ F^{-1}(U_{i:n})\big]$ angegeben und man nennt $\dot{L} \circ F^{-1}$ Score-erzeugende Funktion. Für große n kann man approximativ mit $b(i) := \dot{L} \circ F^{-1}(\frac{i}{n+1})$ (vgl. $\mathbb{E}\big[U_{i:n}\big] = \frac{1}{n+1}$ aus Bemerkung 1.25) oder $\tilde{b}(i) = n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \dot{L} \circ F^{-1}(u)du$ gearbeitet werden.

Lemma 1.29

Sei $\tilde{T}$ eine einfache lineare Rangstatistik von der Form wie in Bemerkung 1.28(c), aber mit allgemeinen deterministischen Scores $a(i)$. Sei $\bar{c} := n^{-1}\sum_{i=1}^n c_i$ und $\bar{a} = n^{-1}\sum_{i=1}^n a(i)$. Unter H_0 aus (1.10) *gilt dann*

$$\mathbb{E}\big[\tilde{T}\big] = n\,\bar{c}\,\bar{a} \qquad \textit{und}$$

$$\textit{Var}\left(\tilde{T}\right) = \frac{1}{n-1}\sum_{i=1}^n (c_i-\bar{c})^2 \sum_{i=1}^n (a(i)-\bar{a})^2.$$

Beweis: $R_i(X)$ ist gleichverteilt auf $\{1,\dots,n\}$, also

$$\mathbb{E}\big[a(R_i(X))\big] = \sum_{i=1}^n a(i)n^{-1} = \bar{a} \qquad \text{und}$$

$$\mathbb{E}\big[\tilde{T}\big] = \sum_{i=1}^n c_i \mathbb{E}\big[a(R_i(X))\big] = \sum_{i=1}^n c_i\bar{a}.$$

Aus $\sum_{i=1}^n a(i)$ =const. folgt (mit $R_i := R_i(X)\,\forall 1 \le i \le n$)

$$0 = \text{Var}\left(\sum_{i=1}^n a(i)\right) = \text{Var}\left(\sum_{i=1}^n a(R_i)\right) = \sum_{i=1}^n \text{Var}\left(a(R_i)\right) + 2\sum_{1\le i<j\le n} \text{Cov}\left(a(R_i),a(R_j)\right).$$

Wegen Austauschbarkeit ist $\mathbb{P}^{R_i,R_j} = \mathbb{P}^{R_k,R_l}$ für $i \neq j, k \neq l$. Damit ist

$$0 = n\text{Var}\left(a(R_1)\right) + n(n-1)\text{Cov}\left(a(R_1),a(R_2)\right)$$

$$\Leftrightarrow \quad \text{Cov}\left(a(R_1),a(R_2)\right) = -\frac{1}{n-1}\text{Var}\left(a(R_1)\right).$$

Ferner ergibt sich

$$\text{Var}\left(a(R_1)\right) = \mathbb{E}\left[(a(R_1)-\bar{a})^2\right] = \sum_{j=1}^n \frac{(a(j)-\bar{a})^2}{n}$$

und mit weiteren Routinerechnungen die Varianz von $\tilde{T}$ wie angegeben. ■

Anwendung: Normalapproximation zur Ermittlung kritischer Werte für ψ.

Lemma 1.30

Sei ψ wie in (1.12) *lokal bester Rangtest im Modell $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^n \mathbb{P}_{c_i,\vartheta}$ für $\{\vartheta = 0\}$ gegen $\{\vartheta > 0\}$, vgl. Satz 1.20 zusammen mit Lemma 1.17. Ist $S : \mathbb{R} \to \mathbb{R}$ eine streng isotone Funktion, so ist ψ lokal optimal für $\bigotimes_{i=1}^n \mathbb{P}^S_{c_i\vartheta}$.*

Beweis: $\forall 1 \le i \le n$ gilt $R_i((S(X_1),\dots,S(X_n))) = R_i(X)$. ■

Lemma 1.31 (Stochastisch größer-Alternativen, Lemma 4.4 in Janssen (1998))
Sei $a(1) \leq a(2) \leq \ldots \leq a(n)$ *(vgl. Bemerkung 1.28) und sei* ψ *ein Rangtest zum Niveau* α *unter* H_0 *aus* (1.10)*, d.h.* $\mathbb{E}_{H_0}\big[\psi\big] = \alpha$*. Es seien* $X_1, \ldots, X_{n_1}$ *iid. mit Verteilungsfunktion* F_1 *und* $X_{n_1+1}, \ldots, X_n$ *iid. mit Verteilungsfunktion* F_2*,* $X = (X_1, \ldots, X_n)$*.*

(a) Gilt $F_1 \geq F_2$*, so folgt* $\mathbb{E}_{\vartheta_0}\big[\psi(R(X))\big] \leq \alpha$*.*

(b) Gilt $F_1 < F_2$*, so folgt* $\mathbb{E}_{\vartheta_0}\big[\psi(R(X))\big] \geq \alpha$*.*

Anmerkung: Für Lageparametermodelle (siehe Beispiel 1.16(a)) ist die Score-Funktion genau dann isoton, wenn die Dichte f von Y unimodal ist. Dazu abschließend einige Beipsiele.

Beispiel 1.32
[Zweistichproben-Rangtests in Lageparametermodellen für stochastisch größer-Alternativen]

(a) Fisher-Yates-Test
Sei f *die Dichte von* $\mathcal{N}(0,1)$*. Dann gilt* $\dot{L}(x) = x$ *und es ergibt sich*

$$T = \sum_{i=1}^{n_1} a(R_i) \quad \textit{mit} \quad a(i) = \mathbb{E}\big[X_{i:n}\big]$$

Dabei ist $X_{i:n}$ *die* i*-te Orderstatistik von* $X_1, \ldots, X_n$ *iid. mit* $X_1 \sim \mathcal{N}(0,1)$*.*

(b) Van der Waerden-Test
Sei f *wieder die Dichte von* $\mathcal{N}(0,1)$*. Die Score-erzeugende Funktion ist gegeben durch* $u \mapsto \Phi^{-1}(u)$*. Damit sind nach Bemerkung 1.28(e)* $b(i) = \Phi^{-1}(\frac{i}{n+1})$ *approximative Scores und es ergibt sich*

$$T = \sum_{i=1}^{n_1} \Phi^{-1}(\frac{R_i}{n+1})$$

(c) Wilcoxon rank sum Test
Sei $f(x) = \exp(-x)(1+\exp(-x))^{-2}$ *die Dichte der logistischen Verteilung mit Verteilungsfunktion* $F(x) = (1+\exp(-x))^{-1}$*. Die Score-erzeugende Funktion berechnet sich zu* $u \mapsto 2u - 1$*. Damit ist*

$$a(i) = \mathbb{E}\big[\dot{L} \circ F^{-1}(U_{i:n})\big] = \frac{2i}{n+1} - 1$$

Die Scores lassen sich affin linear auf die Identität transformieren und es ergibt sich (vgl. Lemma 1.30)

$$T = \sum_{i=1}^{n_1} R_i(X),$$

die Rangsumme aus Gruppe 1.

(d) <u>*Median-Test*</u>

Die doppelte Exponentialverteilung (auch: Laplace-Verteilung) hat die Dichte $f(x) = \frac{1}{2}\exp(-|x|)$ *und die Score-erzeugende Funktion* $u \mapsto \operatorname{sgn}(\ln(2u)) = \operatorname{sgn}(2u-1)$*. Approximative Scores ergeben sich damit zu*

$$b(i) = \dot{L} \circ F^{-1}(\frac{i}{n+1}) = \begin{cases} 1, & \text{falls } i > \frac{n+1}{2} \\ 0, & \text{falls } i = \frac{n+1}{2} \\ -1, & \text{falls } i < \frac{n+1}{2} \end{cases}$$

Zum Schluss noch der beliebte Savage-Test (auch: log-rank Test) als Beispiel für einen Skalentest (vgl. Beispiel 1.16(b)).
Sei im Skalenparametermodell Y *exponentialverteilt mit* $f(x) = \exp(-x)\,\mathbf{1}_{(0,\infty)}(x)$*. Dann gilt für* $x > 0$*, dass*

$$\dot{L}(x) = -(1 + x\frac{f'(x)}{f(x)}) = x - 1.$$

<u>*Übungsaufgabe:*</u> *Zeige, dass für* $X_1, \ldots, X_n$ *ii. standard-exponentialverteilt gilt:* $\mathbb{E}\big[X_{i:n}\big] = \sum_{j=1}^{i} \frac{1}{n+1-j}$*.*
Damit ergeben sich exakte Scores zu

$$a(i) = \sum_{j=1}^{i} \frac{1}{n+1-j} - 1.$$

Da Y *fast-sicher positiv ist, kann* $X = \exp(\vartheta)Y$ *auf ein Lageparametermodell* $\log(X) = \vartheta + \log(Y)$ *zurückgeführt werden. Gilt* $Y \sim \operatorname{Exp}(1)$*, so genügt* $\log(Y)$ *einer gespiegelten Gumbelverteilung mit*

$$\mathbb{P}\big(\log(Y) \le x\big) = 1 - \exp(-\exp(x)), \quad x > 0.$$

Die Rangtests haben starke Analogie zu Permutationstests, da kritische Werte im finiten Fall durch Traversieren aller $\sigma \in \mathcal{S}_n$ und nachfolgender Bestimmung des $(1-\alpha)$-Quantils ermittelt werden. Wir verallgemeinern die Theorie nun noch auf (nahezu) beliebige Abbildungen $g(X)$ anstelle von $R(X)$ und zufällige Scores. Wir erhalten lineare Resamplingstatistiken.

1.5 Allgemeine Theorie von Resamplingtests

Wir orientieren uns am Artikel von Janssen and Pauls (2003) bzw. den Doktorarbeiten von Pauls (2003) und Pauly (2009).

Definition 1.33

Sei $(Y_{n,i})_{1 \le i \le k(n)}$ *ein Dreiecksschema von Zufallsvariablen auf einem beliebigen Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, \mathbb{P})$*. Dabei sei* $n \in \mathbb{N}$ *und* $k(n) \in \mathbb{N}$*. In vielen relevanten Beispielen wird* $k(n) \equiv n$ *sein.*

Ferner sei $(W_{n,i})_{1\le i\le k(n)}$ ein Dreiecksschema zufälliger Gewichtsfunktionen auf einem Wahrscheinlichkeitsraum $(\tilde\Omega,\tilde{\mathcal{A}},\tilde{\mathbb{P}})$. Die Zufallsvariablen $(Y_{n,i})_{1\le i\le k(n)}$ und $(W_{n,i})_{1\le i\le k(n)}$ seien bezüglich des Produktmaßes $\mathbb{P}\otimes\tilde{\mathbb{P}}$ stochastisch unabhängig. Die Gewichte mögen die folgenden Generalvoraussetzungen erfüllen:

[*GV1*] *Für alle $n\in\mathbb{N}$ seien $(W_{n,1},\dots,W_{n,k(n)})$ austauschbar.*

[*GV2*] $\max\limits_{1\le i\le k(n)}\left|W_{n,i}-\overline{W}_n\right|\longrightarrow 0$ $\tilde{\mathbb{P}}$*-stochastisch für* $n\to\infty$.

[*GV3*] $\sum\limits_{i=1}^{n}(W_{n,i}-\overline{W}_n)^2\longrightarrow C\in\mathbb{R}$ $\tilde{\mathbb{P}}$*-stochastisch für* $n\to\infty$.

Dann heißt

$$T_n^*=\sqrt{k(n)}\sum_{i=1}^{k(n)}W_{n,i}(Y_{n,i}-\overline{Y}_n)$$

lineare Resamplingstatistik mit Gewichtsfunktionen $(W_{n,i})_{1\le i\le k(n)}$.

Bemerkung 1.34
Gilt $\overline{W}_n=k(n)^{-1}\sum_{i=1}^{k(n)}W_{n,i}=0$ $\tilde{\mathbb{P}}$-fast sicher, so folgt

$$T_n^*=\sqrt{k(n)}\sum_{i=1}^{k(n)}W_{n,i}Y_{n,i}\quad\tilde{\mathbb{P}}\text{-fast sicher}$$

Beispiel 1.35 *(a) Einfache lineare Rangstatistiken (vgl. Abschnitt 1.4)*

Sei n beliebig, aber fest vorgegeben und $R_n(X)=(R_{n,i}(X))_{1\le i\le n}$ für $X=(X_1,\dots,X_n)$ ein Vektor von Rangstatistiken wie in Abschnitt 1.4 betrachtet. Seien $a_n(i)=\mathbb{E}_{\mathbb{P}_{n,0}}\left[\dot L(X_{i:n})\right]$ (oder auch $b(i)$ bzw. $\tilde b(i)$ aus Bemerkung 1.28). Scores und $(c_{n,i})_{1\le i\le n}$ Regressionskoeffizienten. Setze $k(n)\equiv n$.

Dann hat die einfache lineare Rangstatistik

$$T_n(R_n(X))=\sum_{i=1}^{n}(c_{n,i}-\bar c_n)a_n(R_{n,i}(X))$$

die Struktur einer linearen Resamplingstatistik (beachte auch Bemerkung 1.28(d)). Wähle dazu

$$W_{n,i}:=\frac{a_n(R_{n,i}(X))}{\sqrt{n}},\quad 1\le i\le n\quad\text{und}$$

$$Y_{n,i}:=(c_{n,i}-\bar c_n),\quad 1\le i\le n$$

und rechne [GV1] bis [GV3] nach.

Beachte für den Nachweis der Gültigkeit von [GV1] die stochastische Unabhängigkeit von

Rängen und Ordnungsstatistiken (Satz 1.26(a)).

Ein Resamplingverfahren kommt nun wie zuvor diskutiert zustande, indem die linearen Resamplingstatistiken $T_n^* := T_n(\sigma)$ für zufällige Permutationen $\sigma \in \mathcal{S}_n$ betrachtet werden. Unter H_0 aus (1.10) *ist jedes $\sigma \in \mathcal{S}_n$ gleichwahrscheinlich für $R_n(X)$ und durch diese Überlegung wird der kritische Wert für den Score-Test (der als Rangtest durchgeführt wird) aus den Quantilen der Verteilung der $(T_n(\sigma))_{\sigma\in\mathcal{S}_n}$ bezüglich Gleichverteilung auf $\mathcal{S}_n$ gewonnen.*

(b) *Einfache lineare Permutationsstatistiken*

In Anlehnung an das zweite motivierende Eingangsbeispiel aus Abschnitt 1.2 seien nunmehr die Werte der $(X_i)_{1\le i\le n}$ im Mehrstichprobenproblem selbst vertrauenswürdig, aber kein konkretes parametrisches Modell. Dann sind sinnvolle lineare Permutationsstatistiken gegeben durch

$$T_n^* = \sqrt{k(n)} \sum_{i=1}^{k(n)} c_{n,\sigma(i)}(Y_{n,i} - \overline{Y}_n)$$

mit

$$Y_{n,i} := \frac{X_{n,i}}{\sqrt{k(n)}} \qquad \forall\, n \in \mathbb{N}, 1 \le i \le k(n), \tag{1.13}$$

(nicht notwendigerweise fest vorgegebenen) Regressionskoeffizienten $(c_{n,i})_{1\le i\le k(n)}$ und resultierenden Gewichten $W_{n,i} := c_{n,\sigma(i)}$. Dabei ist $\sigma \in \mathcal{S}_{k(n)}$ wieder eine zufällige (gleichverteilte) Permutation der Zahlen $1, \ldots, k(n)$.

Offenbar erfüllen diese Gewichte [GV1] bis [GV3] genau dann, wenn es die Regressionskoeffizienten tun.

Greifen wir das zweite motivierende Eingangsbeispiel (Zweistichprobenproblem) aus Abschnitt 1.2 konkret wieder auf, so sind sinnvolle Regressionskoeffizienten für $k(n) = n_1 + n_2$ gegeben durch

$$c_{n,i} = \sqrt{\frac{n_1 n_2}{k(n)}} \quad \cdot \quad \begin{Bmatrix} \frac{1}{n_1} & , i \le n_1 \\ -\frac{1}{n_2} & , n_1 < i \le k(n) \end{Bmatrix}$$

Übungsaufgabe: Rechne nach, dass für $\sigma = id$ die Originalstatistik $T_n = \sqrt{\frac{n_1 n_2}{n_1+n_2}}(\overline{X}_{n_1} - \overline{X}_{n_2})$ entsteht.

(c) *Bootstrap-Statistiken*

Für Einstichprobenprobleme (z.B. erstes motivierendes Eingangsbeispiel aus Abschnitt 1.2) sind Permutationsverfahren inadequat, da unter der Nullhypothese die Summenstatistik permutationsinvariant ist.

Es bietet sich hier vielmehr ein auf „Ziehen mit Zurücklegen“ basierendes Resamplingschema an.

Seien dazu zum Beispiel (Klassischer Bootstrap von ?? (efr)) $(M_{n,1}, \ldots, M_{n,k(n)})$ *multinomial verteilte Zufallsvariablen zum Stichprobenumfang* $k(n) = \sum_{i=1}^{k(n)} M_{n,i}$ *und Auswahlwahrscheinlichkeiten* $p_{n,i} \equiv \frac{1}{k(n)}\, \forall n$. *Dann sind Bootstrap-Gewichte gegeben durch*

$$W_{n,i} := k(n)^{-\frac{1}{2}}(M_{n,i} - 1).$$

Mit $Y_{n,i}$ *wie in* (1.13) *hat eine lineare Bootstrap-Resamplingstatistik also die Form (nachrechnen zur Übung!)*

$$T_n^* = \sqrt{k(n)}\Big(\sum_{i=1}^{k(n)} \frac{M_{n,i}}{k(n)} X_{n,i} - \overline{X}_n\Big)$$

Die folgenden zwei Sätze 1.37 und 1.40 zeigen, dass der Formalismus in Definition 1.33 so allgemein ist, dass für eine sehr große Klasse von Resamplingverfahren asymptotische Äquivalenz des bedingten (auf die Daten) Resamplingtests und eines unbedingten Niveau α-Tests basierend auf der Original-Teststatistik T_n gezeigt werden kann. Als Vorbereitung prägen wir die vorgenannten Begriffe mathematisch exakt.

Definition 1.36

(a) <u>*Bedingter Resampling-Test*</u>
Unter den Voraussetzungen von Definition 1.33 habe für alle $n \in \mathbb{N}$ *die bedingte Verteilung* $\mathcal{L}(T_n^*|Y_{n,1}, \ldots, Y_{n,k(n)})$ *die (bedingte) Verteilungsfunktion* F_n^*.
Bezeichne

$$c_n^*(\alpha) \equiv c_n^*(\alpha|Y_{n,1}, \ldots, Y_{n,k(n)}) := (F_n^*)^{-1}(1-\alpha)$$

das $(1-\alpha)$*-Quantil von* F_n^*. *Sei* T_n *eine reelle Statistik. Dann ist ein nicht-randomisierter bedingter* T_n*-(Resampling-)Test definiert durch* $\varphi_{n,\alpha}^* := \mathbf{1}_{(c_n^*(\alpha),\infty)}(T_n)$.

(b) <u>*Asymptotische Äquivalenz von Testfolgen*</u>
Seien $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ *und* $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ *zwei Folgen von Tests für das selbe Testproblem* $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$.
Sei $\vartheta_0 \in H_0$. *Dann heißen* $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ *und* $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ *asymptotisch äquivalent unter* ϑ_0, *falls*

$$\mathbb{E}_{\vartheta_0}\big[\big|\varphi_{n,\alpha} - \varphi_{n,\alpha}^*\big|\big] \overset{n\to\infty}{\longrightarrow} 0 \qquad \forall \alpha \in (0,1).$$

Satz 1.37

Es gelten [GV1] bis [GV3] und in [GV3] gelte o.B.d.A. $C = 1$. *Sei* $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ *eine Folge von Tests für* $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$ *und sei* ϑ_0 *ein beliebiges Element aus* H_0. *Die Folge* $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ *habe die folgenden Eigenschaften:*

[E1] Für jedes $n \in \mathbb{N}$ *sei* $\varphi_{n,\alpha}$ *charakterisiert durch eine reellwertige Statistik* $T_n : \Omega \to \mathbb{R}$ *und einen unbedingten kritischen Wert* $c_n(\alpha)$, *so dass gilt*

$$\varphi_{n,\alpha} = \mathbf{1}_{(c_n(\alpha),\infty)}(T_n), \qquad \textit{wobei} \quad \mathbb{E}_{\vartheta_0}\big[\varphi_{n,\alpha}\big] \overset{n\to\infty}{\longrightarrow} \alpha$$

(Asymptotischer unbedingter Niveau-α*-Test basierend auf* T_n*).*

[E2] Die Statistik T_n aus $[E1]$ konvergiere in Verteilung gegen eine Zufallsvariable T. Die Verteilungsfunktion F_T von T sei stetig und streng monoton steigend auf ihrem Träger supp(F_T). (Unbedingte Konvergenz)

Sei $(\varphi^*_{n,\alpha})_{n\in\mathbb{N}}$ eine Folge von bedingten (Resampling-)Tests für $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$. Dann sind $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ und $(\varphi^*_{n,\alpha})_{n\in\mathbb{N}}$ genau dann asymptotisch äquivalent unter ϑ_0, wenn

$$d(\mathcal{L}(T_n), \mathcal{L}(T_n^*|Y_{n,1}, \ldots, Y_{n,k(n)})) \longrightarrow 0 \quad \mathbb{P}_{\vartheta_0}\text{ -stochastisch für } n \to \infty. \tag{1.14}$$

Dabei bezeichnet $d(\cdot,\cdot)$ eine Metrik, die die schwache Konvergenz auf dem Raum der Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ metrisiert, z.B. die Lévy-Metrik $d_L(\cdot,\cdot)$, definiert durch

$$d_L(F,G) := \inf\{\varepsilon > 0 :\ F(x-\varepsilon) - \varepsilon \le G(x) \le F(x+\varepsilon) + \varepsilon\, \forall x \in \mathbb{R}\}$$

für zwei Verteilungsfunktionen F und G auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Beweis: Dass die Konvergenz (1.14) die asymptotische Äquivalenz von $(\varphi_{n,\alpha})$ und $(\varphi^*_{n,\alpha})$ impliziert, leuchtet intuitiv sofort ein. Ein technisch sauberer Beweis findet sich in Witting and Nölle (1970), S. 58.
Der Beweis der Rückrichtung nutzt Stetigkeitsannahman aus und argumentiert technisch mit dem Teilfolgenkriterium. Selbststudium wird empfohlen (Pauls (2003), Beweis von Lemma 3.4). ■

Definition 1.38
Gilt die Aussage von Satz 1.37 für alle $\vartheta_0 \in H_0$, so heißt $\varphi^*_{n,\alpha}$ asymptotisch effektiv in Bezug auf $\varphi_{n,\alpha}$.

Bemerkung 1.39

(a) Die „hin"-Richtung von Satz 1.37 ist enorm wichtig. Sie mahnt den/die Praktiker/in zur Vorsicht und zeigt, dass ein allzu blindes „Herumrühren" in den Daten oft zu nichts Gutem (keinen validen Testprozeduren) führen kann. Vielmehr muss das Resamplingschema sorgfältig so ausgewählt werden, dass die Verteilungseigenschaften der Original-Teststatistik T_n unter der Nullhypothese möglichst exakt durch die Konstruktion von T_n^* abgebildet werden. Ein „beliebter" Fehler ist es z.B., unter den Gegebenheiten von Beispiel 1.35(b) im Falle von $n_1 \neq n_2$ (unbalancierte Stichprobenumfänge) „blindlings" zu permutieren, also $c_{n,i} \propto n^{-1/2} \cdot \mathbf{1}_{\{i\le n_1\}}$ zu wählen. Man rechnet (unter asymptotischer Normalität von $T_n = \sqrt{\frac{n_1 n_2}{n_1+n_2}}(\overline{X}_{n_1} - \overline{X}_{n_2})$) leicht nach, dass dieses falsche Resamplingschema keine valide Testprozedur im Sinne von Definition 1.36(b) ergibt.
Andere klassische Gegenbeispiele sind das „blinde" Bootstrappen der Maximumsstatistik oder von korrelierten Daten (vgl. Abschnitt 2.3.1 der Dissertation von Pauly (2009)).

(b) Ist T_n eine normalisierte Summenstatistik unabhängiger Ausgangsvariablen und T_n^* eine lineare Resamplingstatistik, so liefern nach Satz 1.37 bedingte zentrale Grenzwertsätze die

asymptotische Effektivität von $\varphi^*_{n,\alpha}$ bezüglich $\varphi_{n,\alpha}$. Ein recht allgemeiner solcher bedingter zentraler Grenzwertsatz soll dieses Kapitel beschließen.

Satz 1.40 (Bedingter Zentraler Grenzwertsatz, Satz 3.3. in Pauly (2009), vgl. Theorem 2.1 in Janssen (2005))
Es sei T^*_n eine lineare Resampling-Statistik. Es gelten die Generalvoraussetzungen [GV1] bis [GV3] an die $W_{n,i}$, wobei o.B.d.A. $C = 1$ in [GV3] gelte. Ferner gelte $\sqrt{k(n)}(W_{n,1} - \overline{W}_n) \xrightarrow{\mathcal{D}} W_1$, W_1 Zufallsvariable mit Var $(W_1) = 1$. Die $Y_{n,i}$ sollen die folgenden Regularitätsvoraussetzungen erfüllen:

$$[R1] \qquad \overline{Y}_n \longrightarrow 0 \quad \mathbb{P}\text{-stochastisch}$$

$$[R2] \qquad \max_{1\le i\le k(n)} \left|Y_{n,i} - \overline{Y}_n\right| \longrightarrow 0 \quad \mathbb{P}\text{-stochastisch}$$

$$[R3] \qquad \sum_{i=1}^{k(n)} \left(Y_{n,i} - \overline{Y}_n\right)^2 \xrightarrow{\mathcal{D}} V^2, \quad V^2 \text{ nicht-negative Zufallsvariable}$$

Dann folgt

$$d(\mathcal{L}(T^*_n|(Y_{n,i})_{1\le i\le k(n)}), \mathcal{L}(Z)) \longrightarrow 0 \quad \mathbb{P}\text{-stochastisch},$$

wobei Z eine Zufallsvariable auf $\Omega \times \Omega'$ mit $Z(\omega, \cdot) \sim \mathcal{N}(0, V^2(\omega))$ bezeichnet.

Anmerkung: Die $Y_{n,i}$ aus Beispiel 1.35(a) erfüllen (unskaliert) [R1] bis [R3] in aller Regel nicht. Da Rangtests jedoch invariant gegenüber isotonen Transformationen sind (vgl. Lemma 1.30), lassen sie sich entsprechend reskalieren.
Übung: Betrachte den Wilcoxon Rangsummentest aus Beispiel 1.32(c). Wie müssen die Regressionskoeffizienten sinnvollerweise gewählt werden, um asymptotische Normalität der einfachen linearen Rangstatistik zeigen zu können? Wie sind die Gewichte zu transformieren?

Bemerkung 1.41 (Bemerkung 3.4 in Pauly (2009))

(a) *Die Bedingungen [R1] und [R2] sind nach Dreiecksungleichung zusammengenommen äquivalent zur $\mathbb{P}$-stochastischen Konvergenz von $\max_{1\le i\le k(n)} \left|Y_{n,i}\right|$ gegen 0.*

(b) *Ist $V^2 \equiv \sigma^2 > 0$ in [R3] konstant und positiv, so konvergiert die bedingte Verteilungsfunktion F^*_n (vgl. Definition 1.36(a)) nach dem Satz von Polya (Witting and Müller-Funk (1995), Satz 5.75) sogar gleichmäßig, also es gilt*

$$\sup_{x\in\mathbb{R}} \left|F^*_n(x) - \Phi(\frac{x}{\sigma})\right| \longrightarrow 0 \quad \mathbb{P}\text{-stochastisch}$$

(c) *Gilt für die Ausgangsvariablen $\mathbb{P}\big(\sum_{i=1}^{k(n)}(X_{n,i} - \overline{X}_n)^2 > 0\big) \longrightarrow 1$ für $n \to \infty$, so können studentisierte $Y_{n,i}$ der Form*

$$Y_{n,i} := \frac{X_{n,i} - \overline{X}_n}{\sqrt{\sum_{i=1}^{k(n)}(X_{n,i} - \overline{X}_n)^2}} \mathbf{1}_{\{\sum_{i=1}^{k(n)}(X_{n,i}-\overline{X}_n)^2>0\}}$$

benutzt werden. Diese erfüllen offensichtlich [R3] mit $V^2 \equiv 1$.
Damit werden t-Test Analoga behandelbar (Janssen (2005)).

Kapitel 2

Spezielle Resamplingverfahren für unabhängige Daten

2.1 Mehrstichprobenprobleme, Permutationstests

Anknüpfend an das zweite motivierende Eingangsbeispiel aus Abschnitt 1.2 und in enger Analogie zur Theorie der Rangverfahren in Abschnitt 1.4 geben wir zuerst einen Aufriss der Theorie von Zweistichproben-Permutationstests. Greifen wir das entsprechende Beispiel aus 1.2 wieder auf, so erhalten wir formal die folgende Problemstellung.

Modell 2.1 (Zweistichprobenproblem)
Seien $(X_i)_{1\leq i\leq n}$ *reellwertige, stochastisch unabhängige Zufallsvariablen.* $X_1, \ldots, X_{n_1}$ *seien iid. mit* $X_1 \sim F_1$ *und* $X_{n_1+1}, \ldots, X_n$ *seien iid. mit* $X_{n_1+1} \sim F_2$*. Sei* $n_2 := n - n_1$ *und* $\overline{X}_{n_1} := n_1^{-1} \sum_{i=1}^{n_1} X_i$, $\overline{X}_{n_2} := n_2^{-1} \sum_{j=n_1+1}^{n} X_j$*, wobei wir fordern, dass* $0 < n_1 < n$ *ist.*

Das interessierende Testproblem sei gegeben durch

$$H = \{F_1 = F_2\} \qquad \text{versus} \qquad K = \{F_1 \neq F_2\} \qquad [**]$$

In dem Lehrbuch von Lehmann und Romano (2005, ca. 800 Seiten) werden genau fünf parametrische Situationen genannt, in denen das Problem $[**]$ (einigermaßen) befriedigend bearbeitet werden kann:

(1) $F_1 = \mathcal{N}(\mu_1, \sigma^2)$, $F_2 = \mathcal{N}(\mu_2, \sigma^2)$, $\sigma^2 > 0$ bekannt. Hier wird der Zweistichproben-Gaußtest durchgeführt mit Teststatistik $T_1 = \left|\overline{X}_{n_1} - \overline{X}_{n_2}\right|$ und kritischem Wert als Normalverteilungsquantil.

(2) F_1 und F_2 wie zuvor, aber $\sigma^2 > 0$ unbekannt. Man führt den Zweistichproben-t-Test durch mit Teststatistik $T_2 = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \frac{T_1}{S}$, wobei S die gepoolte Stichprobenstreuung bezeichnet. Quantile der t-Verteilung mit $n-2$ Freiheitsgraden dienen als kritische Werte.

(3) $F_1 = \text{Bernoulli}(p_1), F_2 = \text{Bernoulli}(p_2)$: Exponentialfamilie, UMPU-Theorie

(4) $F_1 = \text{Poisson}(\lambda_1), F_2 = \text{Poisson}(\lambda_2)$: Exponentialfamilie, UMPU-Theorie

(5) $F_1 = \text{Exp}(\lambda_1), F_2 = \text{Exp}(\lambda_2)$. Es existiert ein UMPU-Test für das Problem $\frac{\lambda_1}{\lambda_2} \leq \Lambda$ versus $\frac{\lambda_1}{\lambda_1} > \Lambda$. Führt man diesen mit vertauschten Rollen von λ_1 und λ_2 aus und adjustiert für Multiplizität, so kann man [$**$] damit bearbeiten.

Wir untersuchen zunächst den Fall, dass sowohl F_1 als auch F_2 stetige Verteilungsfunktionen sind und betrachten Teststatistiken der Form

$$T = \sum_{i=1}^{n} c_i g(X_i) = \sum_{i=1}^{n} c_{D_i(X)} g(X_{i:n})$$

für eine Statistik $g : \mathbb{R} \to \mathbb{R}$. Die zweite Darstellung zeigt schon die Verbindung zu Rangtests auf. Zum Beispiel geht $\left|T\right|$ in T_1 über, falls $g = id$ und $c_i = n_1^{-1}$ für $i \leq n_1$ und $c_j = -n_2^{-1}$ für $j > n_1$ angesetzt wird. Unter H aus [$**$] sind die Antiränge $D(X) = (D_i(X)))_{1\leq i\leq n}$ und die Orderstatistiken $(X_{i:n})_{1\leq i\leq n}$ stochastisch unabhängig nach Satz 1.26.
Der nichtparametrische Test basierend auf T wird nun als Permutationstest bzw. Rangtest gemäß dem folgenden Resamplingschema durchgeführt.

Schema 2.2 (Resamplingschema)
Das Resamplingschema wird hier für die stochastisch-größer Alternative angegeben, der zweiseitige Fall verläuft analog.

(A) Halte $(X_{i:n})_{1\leq i\leq n}$ fest und betrachte $a(i) := g(X_{i:n})$ als zufällige Scores.

(B) Sei $\tilde{D} = (\tilde{D}_i)_{1\leq i\leq n} : \tilde{\Omega} \to \mathcal{S}_n$ eine auf $\mathcal{S}_n$ gleichverteilte Zufallsgröße. Bezeichne $c = c(\alpha, (X_{i:n})_{1\leq i\leq n})$ das $(1-\alpha)$-Quantil der diskreten Zufallsvariable $\tilde{D} \mapsto \sum_{i=1}^{n} c_{\tilde{D}_i} a(i)$.

(C) Führe den Rangtest

$$\varphi(D(X)) = \begin{cases} 1, & T > c, \\ \gamma, & T = c, \\ 0 & T < c \end{cases}$$

aus.

(D) Insgesamt erhält man einen bedingten Test $\varphi = \varphi(D(X), (X_{i:n})_{1\leq i\leq n})$.

Bemerkung 2.3
Wählt man $g = id$ und $(c_j)_{1\leq j\leq n}$ so, dass $\left|T\right| = T_1$ gilt, so heißt der Test gemäß Resamplingschema 2.2 Pitmanscher Permutationstest und wurde bereits von Pitman (1937) hergeleitet.

Das Permutationstest-Prinzip lässt sich auch auf die nichtparametrische Nullhypothese

$$H_0 : X_1, \ldots, X_n \quad \text{sind iid.} \qquad [***]$$

verallgemeinern. Dazu müssen die $X_j, 1 \leq j \leq n$, noch nicht einmal reellwertig sein. In diesem Fall muss das Resamplingschema indes leicht modifiziert werden.

Schema 2.4 (Modifiziertes Resamplingschema)

(A) Gegeben seien n Zufallsgrößen $X_j : \Omega \to \Omega'$, $1 \leq j \leq n$ und eine reellwertige Teststatistik $T(X_1, \ldots, X_n)$.

(B) Betrachte gleichverteilte Permutationen und $\mathcal{S}_n \to \mathbb{R}$, $\pi \mapsto T(X_{\pi(1)}, \ldots, X_{\pi(n)})$, wobei die $\pi \in \mathcal{S}_n$ unabhängig von $X_1, \ldots, X_n$ gewählt werden.

(C) Bezeichne Q_0 die Gleichverteilung auf $\mathcal{S}_n$ und bezeichne $c = c(\omega)$ das $(1-\alpha)$-Quantil der Verteilung $x \mapsto Q_0(\{\pi \in \mathcal{S}_n : T(X_{\pi(1)}, \ldots, X_{\pi(n)}) \leq x\})$.

(D) Führe den bedingten Test

$$\tilde{\varphi}(\omega) = \begin{cases} 1, & T(X_1, \ldots, X_n) > c(\omega), \\ \gamma, & T(X_1, \ldots, X_n) = c(\omega), \\ 0, & T(X_1, \ldots, X_n) < c(\omega) \end{cases}$$

durch.

Satz 2.5

Sowohl der bedingte Test $\varphi(D(X), (X_{i:n})_{1 \leq i \leq n})$ aus Resamplingschema 2.2 als auch der bedingte Test $\tilde{\varphi}(\omega)$ aus Resamplingschema 2.4 haben unter H aus [$**$] bzw. H_0 aus [$***$] die Typ-I-Fehlerwahrscheinlichkeit exakt gleich α für jedes feste $n \in \mathbb{N}$ (unter den jeweils angegebenen Voraussetzungen).

Beweis: Bedingt auf die Orderstatistiken (Schema 2.2) bzw. auf die Daten selbst (Schema 2.4) wird der jeweilige kritische Wert so eingestellt, dass

$$\mathbb{E}_{\mathcal{L}(\tilde{D})}\big[\varphi(D(X))\big] = \mathbb{E}_{Q_0}\big[\tilde{\varphi}\big] = \alpha$$

gilt. Zudem sind Antiränge unter H aus [$**$] stochastisch unabhängig von den Orderstatistiken bzw. werden die Permutationen π unabhängig von $(X_1, \ldots, X_n)$ gewählt. Das Resultat liefern folgende Rechenregeln für bedingte Erwartungen:

$$X \perp Y \Rightarrow \mathbb{E}\big[h(X,Y) \mid X = x\big] = \mathbb{E}\left[h(x,Y)\right] = \int h(x,y)\mathbb{P}_Y(dy) \qquad \text{und}$$

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\mathbb{E}\big[Y \mid X\big]\right] = \int \mathbb{E}\big[Y \mid X = x\big]\mathbb{P}_X(dx)$$

■

An sich ist damit unter Nullhypothesen eine befriedigende Theorie von Permutationstests entwickelt. Für kleine n ist ein explizites Ausrechnen der kritischen Werte durch Traversieren aller $n!$ möglichen Permutationen möglich. Für moderat große n ist ein Traversieren von $B \leq n!$ zufällig ausgewählten Permutationen eine Approximationsmethode (Monte Carlo-kritische Werte).
Für sehr große n kommt indes eher eine Normalapproximation der kritischen Werte in Frage, sofern die Teststatistik Summengestalt hat.
Hierzu benötigen wir bedingte und unbedingte zentrale Grenzwertsätze.

Modellvoraussetzungen 2.6

Seien $(X_i)_{i\geq 1}$ iid. Zufallsgrößen mit $X_1 : (\Omega, \mathcal{A}, \mathbb{P}) \to \Omega'$ und sei $h : \Omega' \to \mathbb{R}$ eine Statistik mit

[1] $\int h^2(X_1) d\mathbb{P} < \infty$.

Seien Regressionskoeffizienten c_{ni} gegeben mit

[2] $\sum_{i=1}^n c_{ni} = 0 \, \forall n \in \mathbb{N}$

[3] $\lim_{n\to\infty} \sum_{i=1}^n c_{ni}^2 = c^2$ *mit* $c > 0$

[4] $\forall \varepsilon > 0 \exists M = M(\varepsilon) > 0$ *mit* $\sum_{i=1}^n c_{ni}^2 \mathbf{1}_{[M,\infty)}(\left|\sqrt{n} c_{ni}\right|) \leq \varepsilon \, \forall n \in \mathbb{N}$ *(gleichgradige Integrierbarkeit).*
Sei ferner

[5] $\sigma^2 := c^2 \int \{h(X_1) - \mathbb{E}\,[h(X_1)]\}^2 d\mathbb{P} > 0$

Satz 2.7

Setze unter den Voraussetzungen aus 2.6 $T_n = \sum_{i=1}^n c_{ni} h(X_i)$. Dann gilt

$$\textit{(a)} \quad \mathcal{L}(T_n) \xrightarrow{w} \mathcal{N}(0, \sigma^2) \textit{ für } n \to \infty.$$

Wähle unabhängig von $(\Omega, \mathcal{A}, \mathbb{P})$ eine gleichverteilte Zufallsgröße $\tau_n = (\tau_{ni})_{1\leq i\leq n} : (\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbb{P}}) \to \mathcal{S}_n$ von Permutationen. Für festes $\omega \in \Omega$ bezeichne $F_{n,\omega}(\cdot)$ die Verteilungsfunktion von $\tilde{\omega} \mapsto T_n((X_{\tau_{ni}(\tilde{\omega})}(\omega))_{1\leq i\leq n})$. Dann gilt

$$\textit{(b)} \quad \sup_{t\in\mathbb{R}} \left|F_{n,\cdot}(t) - \Phi(\frac{t}{\sigma})\right| \to 0 \quad \mathbb{P}\textit{-stochastisch.}$$

Beweis: (a) Zentraler Grenzwertsatz von Lindeberg-Feller.

(b) Wir wenden Satz 1.40 mit Bemerkung 1.41 an. Dazu stellen wir die Resamplingstatistik für $\tau_n = \pi$ dar als

$$T_n((X_{\pi(j)})_{1\leq j\leq n}) = \sum_{i=1}^n c_{ni} h(X_{\pi(i)}) = \sqrt{n} \sum_{i=1}^n c_{n,\pi^{-1}(i)} \frac{h(X_i)}{\sqrt{n}}$$

$$= \sqrt{n} \sum_{i=1}^n W_{n,i}(Y_{n,i} - \overline{Y}_n)$$

mit $Y_{n,i} := c\frac{h(X_i)}{\sqrt{n}}$ und $W_{n,i} := \frac{c_{n,\pi^{-1}(i)}}{c}$, wobei wir o.B.d.A. die $h(X_i), 1 \leq i \leq n$ als zentriert an ihrem arithmetischen Mittel annehmen können, beachte [2] und Bemerkung 1.34.

Die Regularitätsannahme $\max_{1\leq i\leq n} \left|Y_{n,i}\right| \xrightarrow{\mathbb{P}} 0$ ist offenbar erfüllt.

Die Annahme [R3] mit $V^2 \equiv \sigma^2 > 0$ folgt aus Voraussetzung [5].

Die Annahme [GV1] folgt daraus, dass $\tau_n = \pi$ zufällig gleichverteilt auf $\mathcal{S}_n$ ist.

Voraussetzung [GV3] mit $C = 1$ folgt aus den Annahmen [2] und [3].

Es bleibt, [GV2] zu prüfen, also $\max_{1\leq i\leq n} c_{ni} \xrightarrow{\tilde{\mathbb{P}}} 0$.

Da die $c_{n,\pi^{-1}(i)}$ austauschbar sind und $\lim_{n\to\infty} \sum_{i=1}^n c_{ni}^2$ =konst. [3] gilt, muss $c_{n,\pi^{-1}(i)} = \mathcal{O}_{\tilde{\mathbb{P}}}\left(\frac{1}{\sqrt{n}}\right)$ für große n für alle Indizes i gelten.

■

Selbstverständlich gilt die asymptotische Normalität von $\sum_{i=1}^n c_{ni}h(X_{\tau_{ni}})$ erst recht, wenn auf $\vec{X} = \vec{x}$ bedingt wird. Wir berechnen die Permutationsvarianz wie folgt (o.B.d.A.: $\mathbb{E}\left[h(X_1)\right] = 0$):

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^n c_{ni}h(X_{\tau_{ni}})|\vec{X} = \vec{x}\right) &= \text{Var}\left(\sum_{i=1}^n c_{n,\tau_{ni}^{-1}}h(x_i)\right) && \text{(Unabhängigkeit)}\\ &= \frac{1}{n-1}\sum_{i=1}^n (c_{ni} - \overline{c})^2 \cdot \sum_{i=1}^n (h(x_i) - n^{-1}\sum_{j=1}^n h(x_j))^2\end{aligned}$$

Die letzte Gleichung erhält man wie im Beweis zu Lemma 1.29.

2.2 Einstichprobenprobleme, Bootstraptests

Wir verallgemeinern (leicht) die Situation des ersten motivierenden Eingangsbeispiels aus Abschnitt 1.2 und beschäftigen uns hier mit dem Testen linearer Funktionale (Erwartungswerte) im iid.-Fall.

Modell 2.8

Seien $X_1, \ldots, X_n$ stochastisch unabhängige, identisch verteilte Zufallsgrößen, $X_1 : (\Omega^{-1}, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{A})$, $g : \Omega \to \mathbb{R}$ eine Abbildung mit der Eigenschaft $0 < \sigma^2 := Var_{\mathbb{P}}\left(g(X_1)\right) < \infty$ und

$$T(\mathbb{P}^{X_1}) = \int g(X_1)d\mathbb{P} = \mathbb{E}\left[g(X_1)\right]$$

das uns interessierende statistische Funktional. Sei $\hat{\mathbb{P}}_n = n^{-1}\sum_{i=1}^n \varepsilon_{X_i}$ das empirische Maß und

$$\hat{\sigma}_n^2 = n^{-1}\sum_{j=1}^n \left(g(X_j) - n^{-1}\sum_{i=1}^n g(X_i)\right)^2$$

die (unkorrigierte) Stichprobenvarianz von g. Abkürzend schreiben wir $Z_i := g(X_i), 1 \leq i \leq n$, $\overline{Z}_n := n^{-1}\sum_{j=1}^n Z_j = T(\hat{\mathbb{P}}_n)$ und $\hat{\sigma}_n := \sqrt{\hat{\sigma}_n^2}$.

Lemma 2.9

(a) $\mathcal{L}\left(\sqrt{n}\frac{T(\hat{\mathbb{P}}_n)-T(\mathbb{P}^{X_1})}{\sigma}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1)$

(b) $\mathcal{L}\left(\sqrt{n}\frac{T(\hat{\mathbb{P}}_n)-T(\mathbb{P}^{X_1})}{\hat{\sigma}_n}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1)$

Beweis:

(a) Zentraler Grenzwertsatz für die standardisierte Summe der $Z_j, j = 1, \ldots, n$.

(b) Teil (a) plus Lemma von Slutzky und Gesetz der großen Zahlen. ■

Damit sind Gaußtests für Hypothesen, die sich auf $T(\mathbb{P}^{X_1})$ beziehen, unter Verwendung von $\overline{Z}_n$ als Teststatistik asymptotisch valide.
Drei Bootstraptests für die Hypothese

$$H_0 : T(\mathbb{P}^{X_1}) = \mu_0 \quad \text{versus} \quad H_1 : T(\mathbb{P}^{X_1}) \neq \mu_0$$

sind gegeben durch die Resamplingschemata (2.10), (2.11) und (2.12). Wir beschränken uns dabei auf den praxisrelevanten Fall, dass σ^2 unbekannt ist.

Schema 2.10 (Resamplingschema (Bootstraptest))

(A) Sei $X = (X_1, \ldots, X_n)$ gemäß Modell (2.8) *gegeben.*

(B) Sei $X^* = (X_1^*, \ldots, X_n^*)$ ein Vektor von iid. Zufallsgrößen mit $X_i^* : (\Omega^*, \mathcal{A}^*, \mathbb{P}^*) \to (\Omega, \mathcal{A})$ für alle $1 \leq i \leq n$ und mit (gemeinsamer) bedingter Verteilung $\mathcal{L}(X^*|X) = (\hat{\mathbb{P}}_n)^n$.

(C) Bezeichne $\hat{\mathbb{P}}_n^* = n^{-1}\sum_{i=1}^n \varepsilon_{X_i^*}, \overline{Z}_n^* = T(\hat{\mathbb{P}}_n^*) = n^{-1}\sum_{i=1}^n g(X_i^*)$ und q_β das (untere) β-Quantil der bedingten Verteilungsfunktion $x \mapsto \mathbb{P}^*(\overline{Z}_n^* - \overline{Z}_n \leq x|X)$.

(D) Lehne H_0 genau dann ab, wenn $\overline{Z}_n \notin [\mu_0 + q_{\alpha/2}, \mu_0 + q_{1-\alpha/2}]$.

Schema 2.11 (Resamplingschema (Monte Carlo bootstrap, Efron (1977),Efron (1979))

(A) Sei $Z = (Z_1, \ldots, Z_n)$ gemäß Modell (2.8) *gegeben.*

(B) Sei eine Zahl $B \in \mathbb{N}$ fest vorgegeben. Generiere am Computer B bootstrap-Stichproben $\left((Z_{b,1}^*, \ldots, Z_{b,n}^*)\right)_{b=1,\ldots,B}$. Dabei werden alle $Z_{b,j}^*$ für $b = 1, \ldots, B$ und $j = 1, \ldots, n$ unabhängig gleichverteilt (mit Zurücklegen) aus den ursprünglichen $(Z_1, \ldots, Z_n)$ gezogen.

(C) Berechne die bootstrap-Teststatistiken $T_{n,b}^* = \sqrt{n}\left|\frac{\overline{Z}_{n,b}^* - \overline{Z}_n}{\hat{\sigma}_n}\right|, b = 1, \ldots, B$.

(D) Lehne H_0 genau dann ab, wenn $\sqrt{n}\left|\frac{\overline{Z}_n - \mu_0}{\hat{\sigma}_n}\right|$ größer als die $\{(1-\alpha)\cdot B\}$-te Orderstatistik des Vektors $(T_{n,b}^*)_{b=1,\ldots,B}$ ist.

Schema 2.12 (Verbessertes Resamplingschema (siehe z.B. Hall and Wilson (1991)))
Identisch mit (2.11)*, jedoch wird in Schritt (C) die Studentisierung mit in das Bootstrapverfahren einbezogen, also $\hat{\sigma}_n$ durch $\hat{\sigma}_{n,b}^*$ ersetzt, wobei $\hat{\sigma}_{n,b}^* = \sqrt{n^{-1}\sum_{j=1}^n (Z_{b,j}^* - \overline{Z}_{n,b}^*)^2}$ gelte.*

Wir zeigen die asymptotische Effektivität des durch (2.10) gegebenen Bootstraptests in Bezug auf den Gaußtest wieder mit Hilfe des bedingten zentralen Grenzwertsatzes 1.40 für allgemeine lineare Resamplingstatistiken.
Seien dazu $\forall 1 \leq i \leq n$:

$$Y_i = \frac{Z_i - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n} \quad \text{und} \quad Y_i^* = \frac{g(X_i^*) - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n}.$$

Wir erhalten

$$T_n^* := \sqrt{n}\frac{\overline{Z}_n^* - \overline{Z}_n}{\hat{\sigma}_n} = \sum_{j=1}^{n} \left(Y_j^* - \overline{Y}_n\right) = \sum_{j=1}^{n} Y_j^* - n\overline{Y}_n \tag{2.1}$$

und offenbar besteht eine eineindeutige Zuordnung der Y_i^* zu den X_i^*, so dass wir uns den Resamplingmechanismus auch vermittels der Y_i^* vorstellen können.
Betrachten wir nun $\sum_{j=1}^n Y_j^*$. Für jedes $\omega = (\omega_1, \ldots, \omega_n)$ werden aus den Werten $x_i = X_i(\omega)$ jeweils genau $m_{n,i}$ Replikate ($1 \leq i \leq n$) zur Bildung dieser Summe herangezogen, wobei die $m_{n,i}$ als Realisierungen eines multinomialverteilten Zufallsvektors $M_n = (M_{n,1}, \ldots, M_{n,n})$ zum Stichprobenumfang $n = \sum_{i=1}^n M_{n,i}$ und mit den Auswahlwahrscheinlichkeiten $p_{n,i} \equiv n^{-1}\, \forall i = 1, \ldots, n$ aufgefasst werden können. Damit ist $\sum_{j=1}^n Y_j^* = \sum_{j=1}^n M_{n,j} Y_j$. Setzen wir dies in (2.1) ein, so können wir schließlich schreiben

$$T_n^* = \sum_{j=1}^{n} M_{n,j} Y_j - \sum_{i=1}^{n} Y_i = \sum_{j=1}^{n} (M_{n,j} - 1) Y_j = \sqrt{n} \sum_{j=1}^{n} W_{n,j} (Y_j - \overline{Y}_n)$$

mit $W_{n,j} = \frac{M_{n,j}-1}{\sqrt{n}}$ für alle $j = 1, \ldots, n$.
Im Hinblick auf die Anwendbarkeit von Satz 1.40 haben wir T_n^* also so geschickt umgeformt, dass T_n^* sich als lineare Resamplingstatistik erweist sowie [R1] und [R2] und [R3] mit $V^2 \equiv 1$ erfüllt sind, vgl. Bemerkung 1.41(c).
Bleibt noch, alle Voraussetzungen an die Gewichte zu prüfen. Dazu beachten wir das folgende Lemma.

Lemma 2.13 (Lemma 20.2 in Janssen (1998))
Sei $M = (M_1, \ldots, M_n)$ multinomialverteilt mit Stichprobenumfang n und Auswahlwahrscheinlichkeit $p_i = \frac{1}{n}\, \forall 1 \leq i \leq n$. Dann gilt

- *(a)* $\forall 1 \leq i \leq n$: $M_i \overset{\mathcal{D}}{=} \sum_{k=1}^n \mathbf{1}_{\{i\}}(Z_k)$, *wobei* $Z_1, \ldots, Z_n$ *auf* $\{1, \ldots, n\}$ *gleichverteilte Zufallsvariablen sind.*
- *(b)* $\forall 1 \leq i \leq n$: $\mathbb{E}[M_i] = 1$, $\mathrm{Var}(M_i) = \frac{n-1}{n}$
- *(c)* $\sum_{j=1}^n M_j \equiv n \Rightarrow \overline{M}_n \equiv 1$
- *(d)* $\mathbb{P}\left(\sqrt{n}\max_{1\leq i\leq n}\left|\frac{M_i}{n} - \frac{1}{n}\right| \geq \varepsilon\right) \to 0 \quad \forall \varepsilon > 0$ *für* $n \to \infty$

(e) *Var* $\left(\sum_{j=1}^{n}\left[\frac{M_j-1}{\sqrt{n}}\right]^2\right) = \frac{(n-1)^2}{n^3} \to 0$ *für* $n \to \infty$.

Korollar 2.14

Seien $W_{n,j} = \frac{M_{n,j}-1}{\sqrt{n}}$, $1 \leq j \leq n$ *die Bootstrapgewichte für multinomialverteiltes* $M_n = (M_{n,1}, \ldots, M_{n,n})$ *wie in Lemma 2.13.*

1. *Die* $W_{n,j}$ *erfüllen [GV1] wegen 2.13 (a).*
2. *Die* $W_{n,j}$ *erfüllen [GV2] wegen 2.13 (d).*
3. *Die* $W_{n,j}$ *erfüllen [GV3] mit* $C = 1$, *d.h.* $S := \sum_{j=1}^{n}(W_{n,j} - \overline{W}_n)^2 \longrightarrow 1$ *stochastisch für* $n \to \infty$, *denn:*
   Nach 2.13 (b) und (c) ist $\mathbb{E}[S] = Var(M_{n,1}) = \frac{n-1}{n} \longrightarrow 1$ *für* $n \to \infty$ *und*
   $Var(S) = Var\left(\sum_{i=1}^{n}\frac{(M_{n,j}-1)^2}{n}\right) \longrightarrow 0$ *für* $n \to \infty$ *nach 2.13 (e).*
4. *Schließlich gilt* $\sqrt{n}(W_{n,1} - \overline{W}_n) = M_{n,1} - 1$ *und* $Var(M_{n,1}) \to 1$ *für* $n \to \infty$.

Korollar 2.14 liefert zusammen mit Satz 1.40, dass

$$\mathcal{L}\left(\sqrt{n}\frac{\overline{Z}_n^* - \overline{Z}_n}{\hat{\sigma}_n}|X\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1).$$

Da unter H_0 unbedingt $\mathcal{L}\left(\sqrt{n}\frac{\overline{Z}_n-\mu_0}{\hat{\sigma}_n}\right)$ für $n \to \infty$ schwach gegen $\mathcal{N}(0,1)$ konvergent ist, folgt nach Satz 1.37 die asymptotische Effektivität des Bootstraptests aus (2.10).
Zum Abschluss geben wir noch die Begründung, warum das Schema (2.12) dem Schema (2.11) vorzuziehen ist.

Satz 2.15

Es liege die Situation aus Modell 2.8 vor; $\sigma^2 > 0$ *sei unbekannt. Wir schreiben abkürzend* $\mu := T(\mathbb{P}^{X_1})$.

1. *Falls* $\mathbb{E}\left[Z_1^4\right] < \infty$, *so ist*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}(\overline{Z}_n-\mu) \leq x\right) - \mathbb{P}^*\left(\sqrt{n}(\overline{Z}_n^* - \overline{Z}_n) \leq x\right)\right| = O\left(\sqrt{\frac{\ln(\ln(n))}{n}}\right) \quad \textit{fast sicher.}$$

2. *Falls* $\mathbb{E}\left[Z_1^6\right] < \infty$, *so ist*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\frac{\overline{Z}_n - \mu}{\hat{\sigma}_n} \leq x\right) - \mathbb{P}^*\left(\sqrt{n}\frac{\overline{Z}_n^* - \overline{Z}_n}{\sigma_n^*} \leq x\right)\right| = o(n^{-1/2}) \quad \textit{fast sicher.}$$

Beweis:

1. Singh (1981), Theorem 1.B
2. Hall (1988) ■

2.3 Bootstrapverfahren für lineare Modelle

In diesem Abschnitt benutzen wir multivariate (bedingte) zentrale Grenzwertsätze, um die Konsistenz von Bootstrapapproximationen der Verteilung von Schätzern für Regressionskoeffizienten in linearen Modellen zu zeigen.
Wir beginnen mit der Betrachtung fixer Designs.

Modell 2.16
Wir betrachten den Stichprobenraum $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *und modellieren die Beobachtungen* $(y_1, \ldots, y_n)$ *als Realisierungen von reellwertigen stochastisch unabhängigen Zufallsvariablen* $(Y_1, \ldots, Y_n)$ *mit*

$$\forall 1 \leq i \leq n: \quad Y_i = \sum_{k=1}^{p} \beta_k x_{i,k} + \varepsilon_i \tag{2.2}$$

Der Vektor $\beta = (\beta_1, \ldots, \beta_p)^t$ *ist der Parameter von Interesse. Die* $(x_{i,k})_{1 \leq i \leq n, 1 \leq k \leq p}$ *seien fest vorgegebene, uns bekannte reelle Zahlen („Messstellen"). Über die Verteilung der iid. „Messfehler", sagen wir* $\mathbb{P}^{\varepsilon_1}$ *induziert durch* F*, sei lediglich bekannt, dass* $\mathbb{E}[\varepsilon_1] = 0$ *und*
$0 < \sigma^2 := Var(\varepsilon_1) < \infty$ *gilt, also insbesondere Homoskedastizität. Die unbekannte Verteilungsfunktion* F *sei ein Störparameter, also nicht selbst Ziel der statistischen Inferenz.*
Wir kürzen ab:

$$Y(n) \equiv Y := (Y_1, \ldots, Y_n)^t \in \mathbb{R}^n : \qquad \textit{Response-Vektor}$$

$$x(n) \equiv x := \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p} : \qquad \textit{Design-Matrix}$$

$$\varepsilon(n) \equiv \varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^t \in \mathbb{R}^n : \qquad \textit{Residuenvektor}$$

$$\beta \equiv (\beta_1, \ldots, \beta_p)^t \in \mathbb{R}^p : \qquad \textit{Parametervektor}$$

und erhalten als Matrixschreibweise von (2.2)

$$Y(n) = x(n)\beta + \varepsilon(n) \quad \textit{bzw.} \quad Y = x\beta + \varepsilon \tag{2.3}$$

Die Designmatrix habe maximalen Rang, so dass $x^t x \in \mathbb{R}^{p \times p}$ *positiv definit und invertierbar ist.*

Als Verlustfunktion für eine Punktschätzung von β unter Modell 2.16 wählen wir den quadratischen (L_2-) Verlust. Damit ist $\hat{\beta}(n) \equiv \hat{\beta}$ die L_2-Projektion von Y auf den Vektorraum

$\{z \in \mathbb{R}^n : z = x\gamma, \gamma \in \mathbb{R}^p\}$ und kann somit charakterisiert werden durch

$$\begin{aligned} & \forall \gamma \in \mathbb{R}^p \quad \langle Y - x\hat{\beta}, x\gamma \rangle_{\mathbb{R}^n} = 0 \\ \Leftrightarrow \quad & \forall \gamma \in \mathbb{R}^p \qquad Y^t x\gamma = \hat{\beta}^t x^t x\gamma \qquad \text{(Bilinearität von } \langle \cdot, \cdot \rangle_{\mathbb{R}^n}\text{)} \\ \Leftrightarrow \quad & \qquad\qquad Y^t x = \hat{\beta}^t x^t x \end{aligned}$$

Multiplikation von rechts mit $(x^t x)^{-1}$ liefert

$$Y^t x (x^t x)^{-1} = \hat{\beta}^t$$

und folglich

$$\hat{\beta} = (x^t x)^{-1} x^t Y,$$

da $(x^t x)^{-1}$ symmetrisch ist.
Setzen wir (2.3) in $\hat{\beta}$ ein, so ergibt sich außerdem

$$\begin{aligned} \hat{\beta} &= (x^t x)^{-1} x^t (x\beta + \varepsilon) \\ &= \beta + (x^t x)^{-1} x^t \varepsilon \end{aligned}$$

$$\text{bzw.} \qquad \hat{\beta} - \beta = (x^t x)^{-1} x^t \varepsilon \tag{2.4}$$

Gleichung (2.4) ist hilfreich bei der (asymptotischen) Analyse der L_2-basierten statistischen Inferenz über β.
Berechnen wir zunächst die ersten beiden Momente von $\hat{\beta}$ im finiten Fall.

Satz 2.17
Unter Modell 2.16 sei $\hat{\beta}(n) \equiv \hat{\beta} = (x^t x)^{-1} x^t Y$. Dann gilt

(i) $\mathbb{E}[\hat{\beta}] = \beta$

(ii) $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (x^t x)^{-1}$

Beweis: Wir benutzen die alternative Darstellung

$$\hat{\beta} = \beta + (x^t x)^{-1} x^t \varepsilon.$$

Linearität des Erwartungswertoperators liefert (i). Ferner ist damit

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t\right] \\ &= (x^t x)^{-1} x^t \mathbb{E}\left[\varepsilon\varepsilon^t\right] x (x^t x)^{-1} \\ &= \sigma^2 (x^t x)^{-1}, \qquad\qquad \text{da } \mathbb{E}\left[\varepsilon\varepsilon^t\right] = \sigma^2 I_p \text{ ist} \end{aligned}$$

■

Widmen wir uns nun der asymptotischen Betrachtung. Unser Ziel ist ein unbedingter multivariater zentraler Grenzwertsatz für $\hat{\beta}(n)$. Zunächst ein vorbereitendes Resultat.

Lemma 2.18

Es sei $a^t = (a_1, \ldots, a_p)$ ein fest vorgegebener Vektor im $\mathbb{R}^p$. Unter Modell 2.16 nehmen wir an, dass

(i) $n^{-\frac{1}{2}} \max_{1 \leq i \leq n, 1 \leq k \leq p} \left| x_{i,k} \right| \longrightarrow 0$ *für* $n \to \infty$.

(ii) $n^{-1} x^t x \longrightarrow V$ *für eine positiv-definite, symmetrische Matrix* $V \in \mathbb{R}^{p \times p}$.

Dann gilt mit $\rho^2 = \sigma^2 a^t V a$, dass

$$\mathcal{L}\left(n^{-\frac{1}{2}} a^t x^t \varepsilon \right) \xrightarrow[n \to \infty]{w} \mathcal{N}(0, \rho^2)$$

Beweis: Sei $S_n := a^t x^t \varepsilon$. Wir beachten, dass

$$S_n = \sum_{k=1}^{p} \left(a_k \sum_{i=1}^{n} x_{i,k} \varepsilon_i \right) = \sum_{i=1}^{n} \varepsilon_i \left(\sum_{k=1}^{p} a_k x_{i,k} \right) =: \sum_{i=1}^{n} b_i \varepsilon_i 1/2$$

eine Summe stochastisch unabhängiger, zentrierter Zufallsvariablen ist. Ferner gilt

$$\begin{aligned} \text{Var}\,(S_n) = \sigma^2 \sum_{i=1}^{n} b_i^2 &= \sigma^2 \sum_{i=1}^{n} \sum_{j,k=1}^{p} a_j a_k x_{i,j} x_{i,k} \\ &= \sigma^2 \sum_{j,k=1}^{p} a_j a_k (x^t x)_{j,k} \\ &= \sigma^2 a^t (x^t x) a. \end{aligned}$$

Damit folgt

$$\text{Var}\left(n^{-\frac{1}{2}} S_n \right) = n^{-1} \sigma^2 a^t x^t x a \longrightarrow \rho^2 = \sigma^2 a^t V a \quad \text{für } n \to \infty$$

Bleibt, die Lindeberg-Bedingung zu überprüfen, also zu zeigen, dass $\forall \delta > 0$ gilt:

$$n^{-1} \sum_{i=1}^{n} \left[b_i^2 \int_{\{|\varepsilon_i| \geq \delta \sqrt{n} / |b_i|\}} \varepsilon_i^2 d\mathbb{P} \right] \longrightarrow 0 \quad \text{für } n \to \infty.$$

Annahme (i) liefert, dass $\forall 1 \leq i \leq n$:

$$\frac{\sqrt{n}}{|b_i|} \geq \frac{\sqrt{n}}{\max_{1 \leq i \leq n} |b_i|} =: c_n \longrightarrow \infty \quad \text{für } n \to \infty.$$

Damit lässt sich für alle $1 \leq i \leq n$ abschätzen:

$$\int_{\{|\varepsilon_i| \geq \delta \sqrt{n} / |b_i|\}} \varepsilon_i^2 d\mathbb{P} \leq \int_{\{|\varepsilon_i| \geq \delta c_n\}} \varepsilon_1^2 d\mathbb{P} \longrightarrow 0 \quad \text{für } n \to \infty,$$

da ε_1 ein endliches zweites Moment besitzt.

Da ferner $n^{-1} \sum_{i=1}^{n} b_i^2$ gegen den Wert $a^t V a$ konvergent ist (siehe oben), ist damit alles gezeigt.

■

Satz 2.19 (Multivariater zentraler Grenzwertsatz)
Unter Modell 2.16 seien die Voraussetzungen (i) und (ii) aus Lemma 2.18 erfüllt. Dann gilt für $\hat{\beta}(n)$ *wie in Satz 2.17, dass*

$$\mathcal{L}\left(\sqrt{n}\,[\hat{\beta}(n)-\beta]\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0,\sigma^2 V^{-1}\right).$$

Beweis: Wir beachten, dass

$$\sqrt{n}[\hat{\beta}(n)-\beta] = \frac{1}{\sqrt{n}}(n^{-1}x^t x)^{-1}x^t\varepsilon.$$

Nach Cramér-Wold device (siehe z.B. Shorack and Wellner (1986), Seite 862) gilt

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}x^t\varepsilon\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0,\sigma^2 V\right).$$

Da nach Annahme (ii) ferner $(n^{-1}x^t x)^{-1}$ gegen V^{-1} konvergent ist, gilt insgesamt

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}(n^{-1}x^t x)^{-1}x^t\varepsilon\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0,\sigma^2 V^{-1}\right).$$

■

Anders als im $\mathbb{R}^1$ ist ein Glivenko-Cantelli-artiges Verhalten empirischer Größen im $\mathbb{R}^p$ keineswegs garantiert (vgl. Vapnik-Chervonenkis Theorie, z.B. Kapitel 12 in DasGupta (2008)). In Modell 2.16 lassen sich indes die folgenden Aussagen zeigen.

Satz 2.20
Unter Modell 2.16 seien Annahmen (i) und (ii) aus Lemma 2.18 erfüllt. Dann gilt

(a) $n^{-1}(x(n))^t\varepsilon(n) \longrightarrow 0$ *für* $n\to\infty$ *fast sicher,*

(b) $\hat{\beta}(n) \longrightarrow \beta$ *für* $n\to\infty$ *fast sicher.*

Bezeichne

$$\hat{\varepsilon} = (\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n)^t = Y - x\hat{\beta} = x(\beta-\hat{\beta})+\varepsilon \tag{2.5}$$

den Vektor der geschätzten Residuen unter Verwendung des LSE $\hat{\beta}=\hat{\beta}(n)$ *für die Regressionskoeffizienten und sei* $\hat{F}_n$ *die empirische Verteilungsfunktion von* $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n$*. Dann gilt*

(c) $\hat{\mu}_n = \int z\hat{F}_n(dz) = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i \xrightarrow[n\to\infty]{} 0$ *fast sicher,*

(d) $\hat{\sigma}_n^2 = \int z^2\hat{F}_n(dz) = \frac{(\hat{\varepsilon})^t\hat{\varepsilon}}{n} \xrightarrow[n\to\infty]{} \sigma^2$ *fast sicher.*

Beweis: a) und b): Freedman (1981), Lemma 2.3.
c): Starkes Gesetz der großen Zahlen.
d): Beweis von Formel (2.10) in Freedman (1981) ■

Anmerkung: Nach Satz 2.20 c) bleibt die Konvergenz in Teil d) von Satz 2.20 richtig, falls die Residuen an $\hat{\mu}_n$ zentriert werden.

Ein Bootstrapverfahren zur Schätzung der Verteilung von $\hat{\beta}(n)$ kombiniert nun zufällig die geschätzten Residuen mit Zeilen der Designmatrix.

Schema 2.21 (Resamplingschema: Bootstrap für Modell 2.16)

(A) Berechne den LSE $\hat{\beta}(n) = (x^t x)^{-1} x^t Y$ basierend auf dem original Response-Vektor $Y = (Y_1, \ldots, Y_n)^t$.

(B) Bestimme die geschätzten Residuen (vgl. (2.5)) $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ sowie $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i$. Bezeichne $\tilde{F}_n$ die empirische Verteilungsfunktion der zentrierten Residuen $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ mit $\forall 1 \leq j \leq n : \tilde{\varepsilon}_j = \hat{\varepsilon}_j - \hat{\mu}_n$.

(C) Sei $\varepsilon_1^*, \ldots, \varepsilon_n^*$ eine iid. bootstrap-Stichprobe, deren bedingte Verteilung gegeben Y durch $\varepsilon_1^* | Y \sim \tilde{F}_n$ charakterisiert ist.
Setze $Y_j^* = x_j^t \hat{\beta}(n) + \varepsilon_j^*, 1 \leq j \leq n$, wobei x_j die j-te Zeile von x bezeichnet, und $Y^* = (Y_1^*, \ldots, Y_n^*)$.

(D) Berechne $\hat{\beta}^*(n) = (x^t x)^{-1} x^t Y^*$ und benutze die (bedingte) Verteilung von $\sqrt{n}\left(\hat{\beta}^*(n) - \hat{\beta}(n)\right)$ als bootstrap-Approximation der Verteilung von $\sqrt{n}\left(\hat{\beta}(n) - \beta\right)$.

Ein bedingter multivariater zentraler Grenzwertsatz zeigt die Konsistenz von Resamplingschema 2.21. Dazu vorbereitend zwei Lemmata.

Lemma 2.22

Unter Modell 2.16 mit Annahmen (i) und (ii) aus Lemma 2.18 gilt mit den Bezeichnungen aus Resamplingschema 2.21

(a) $n^{-1} \|\hat{\varepsilon} - \varepsilon\|^2 = n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i - \varepsilon_i)^2 \xrightarrow[n\to\infty]{} 0$ *fast sicher.*

(b) $n^{-1} \|\tilde{\varepsilon} - \varepsilon\|^2 = n^{-1} \sum_{i=1}^n (\tilde{\varepsilon}_i - \varepsilon_i)^2 \xrightarrow[n\to\infty]{} 0$ *fast sicher.*

Beweis: Aus (2.5) folgt $\hat{\varepsilon} - \varepsilon = x(\beta - \hat{\beta})$. Damit ist

$$\|\hat{\varepsilon} - \varepsilon\|^2 = (\beta - \hat{\beta})^t x^t x (\beta - \hat{\beta})$$

und Satz 2.20 b) liefert unter Annahme (ii) die Aussage unter (a).
Teil (b) folgt unter zusätzlicher Beachtung von Satz 2.20 c). ■

Lemma 2.23

Unter den Voraussetzungen von Lemma 2.22 gilt

$$\tilde{F}_n \xrightarrow[n\to\infty]{w} F \qquad \text{mit Wahrscheinlichkeit } 1.$$

Beweis: Sei Ψ eine beschränkte, Lipschitz-stetige Funktion mit Lipschitzkonstante K. Dann ist

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\left|\Psi(\tilde{\varepsilon}_i)-\Psi(\varepsilon_i)\right| &\leq \frac{K}{n}\sum_{i=1}^{n}\left|\tilde{\varepsilon}_i-\varepsilon_i\right| \\ &\leq K\left(n^{-1}\sum_{i=1}^{n}(\tilde{\varepsilon}_i-\varepsilon_i)^2\right)^{\frac{1}{2}} \qquad \text{(Cauchy-Schwarz)} \\ &\xrightarrow[n\to\infty]{} 0 \quad \text{fast sicher} \end{aligned}$$

wegen Lemma 2.22 (b). Also gilt

$$\int \Psi(x)\tilde{F}_n(dx)-\int \Psi(x)F_n(dx)\xrightarrow[n\to\infty]{} 0 \quad \text{fast sicher,}$$

wobei F_n die empirische Verteilungsfunktion der wahren Residuen bezeichnen möge. Eine leichte Abwandlung des Satzes von Vitali (vgl. Lemma 8.4 in Bickel and Freedman (1981)) liefert das Resultat. ■

Satz 2.24 (Bedingter multivariater zentraler Grenzwertsatz)
Es gelte Modell 2.16 mit Annahmen (i) und (ii) aus Lemma 2.18. Dann gilt mit Wahrscheinlichkeit 1, dass

$$\mathcal{L}\left(\sqrt{n}\left[\hat{\beta}^*(n)-\hat{\beta}(n)\right]|Y\right)\xrightarrow[n\to\infty]{w}\mathcal{N}_p\left(0,\sigma^2V^{-1}\right).$$

Beweis: Wir beachten, dass

$$x^tx\left[\hat{\beta}^*(n)-\hat{\beta}(n)\right]=x^t\varepsilon^*$$

ist, da $Y^*-Y=\varepsilon^*$ gilt.
Wir verfahren nun wie in Lemma 2.18 und Satz 2.19 und beachten zur Überprüfung der Lindeberg-Bedingung die Lemmata 2.22 und 2.23. ■

Wenden wir uns nun zufälligen Designs zu.

Modell 2.25 (Lineares Modell mit zufälligem Design)
Wir betrachten den Stichprobenraum $(\mathbb{R}^{n(p+1)},\mathcal{B}(\mathbb{R}^{n(p+1)}))$. Die Beobachtungen werden modelliert als Realisierungen von iid. Tupeln $(X_i,Y_i)_{1\leq i\leq n}$, wobei $X_1\equiv x_1\in\mathbb{R}^p$ und $Y_1=y_1\in\mathbb{R}$ gelte. Wir nehmen den folgenden linearen Zusammenhang an:

$$\forall 1\leq i\leq n: \quad Y_i=\sum_{k=1}^{p}\beta_kX_{i,k}+\varepsilon_i=X_i^t\beta+\varepsilon_i \qquad (*)$$

Die Matrix

$$X(n)\equiv X=\begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & & \vdots \\ X_{n,1} & \dots & X_{n,p}\end{pmatrix}$$

von Zufallsvariablen heißt zufällige Design-Matrix und die Matrixschreibweise von $(*)$ lautet

$$Y = X\beta + \varepsilon, \qquad (**)$$

wobei der Index n überall zur notationellen Vereinfachung weggelassen wurde. Wir machen die folgenden (Regularitäts-) Annahmen.

(i) Die Matrix $\Sigma = \mathbb{E}\left[X_1 X_1^t\right] \in \mathbb{R}^{p\times p}$ ist endlich (alle Einträge sind endlich) und positiv definit.

(ii) Bezeichnet μ die $(p+1)$-dimensionale Wahrscheinlichkeitsverteilung von (X_1, Y_1), so ist die Verteilung von ε_1 durch μ bereits voll spezifiziert, denn $\varepsilon_1 = Y_1 - X_1^t\beta$. Insbesondere sind die $(\varepsilon_j)_{j=1,\dots,n}$ iid.

(iii) Der Parametervektor β ist definiert über die Eigenschaft, dass er $\mathbb{E}\left[\left\|Y_1 - X_1^t\beta\right\|^2\right]$ minimiert. Daraus folgt, dass $Y_1 - X_1^t\beta = \varepsilon_1$ senkrecht auf X_1 steht, also $\forall 1 \le j \le p$: $\mathbb{E}\left[X_{1,j}\varepsilon_1\right] = 0 \Rightarrow \beta = \Sigma^{-1}\mathbb{E}\left[X_1 Y_1\right]$.

(iv) Die Matrix $M = (M_{j,k})_{1\le j,k\le p}$ mit $M_{j,k} = \mathbb{E}\left[X_{1,j}X_{1,k}\varepsilon_1^2\right]$ existiert in $\mathbb{R}^{p\times p}$. Diese Annahme ist erfüllt, wenn $\mathbb{E}\left[\|(X_1, Y_1)\|_2^4\right] < \infty$ ist.

Lemma 2.26

Annahmen wie unter Modell 2.25. Die Matrix $n^{-1}X^tX = n^{-1}\left(\sum_{i=1}^n X_{i,j}X_{i,k}\right)_{j,k=1,\dots,p}$ mit Werten in $\mathbb{R}^{p\times p}$ konvergiert μ^n-fast sicher gegen $\Sigma \in \mathbb{R}^{p\times p}$ für $n \to \infty$.

Beweis: Starkes Gesetz der großen Zahlen. ■

Der LSE für $\beta \in \mathbb{R}^p$ ist analog zum Fall mit fixem Design gegeben durch $\hat\beta(n) \equiv \hat\beta = (X^tX)^{-1}X^tY$. Wir rechnen

$$(X^tX)(\hat\beta - \beta) = (X^tX)[(X^tX)^{-1}X^tY - \beta] = X^tY - (X^tX)\beta = X^t\varepsilon.$$

Lemma 2.27

Unter den Voraussetzungen unter Modell 2.25 gilt

$$n^{-\frac{1}{2}}(X^tX)(\hat\beta - \beta) = n^{-\frac{1}{2}}X^t\varepsilon \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p\left(0, M\right).$$

Beweis: Wir beachten $X^t\varepsilon = \left(\sum_{i=1}^n X_{i,j}\varepsilon_i\right)_{j=1,\dots,p}^t$ und Annahmen (iii) bis (iv) und folgern die Aussage mittels multivariatem zentralen Grenzwertsatz analog zur Herleitung im Falle fixer Designs. ■

Nehmen wir Lemmata 2.26 und 2.27 zusammen, so ergibt sich ein (unbedingter) multivariater zentraler Grenzwertsatz.

Satz 2.28 (Multivariater zentraler Grenzwertsatz)
Unter den Voraussetzungen von Modell 2.25 gilt

$$\sqrt{n}\left(\hat{\beta}(n)-\beta\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p\left(0, \Sigma^{-1}M\Sigma^{-1}\right).$$

Außerdem gilt fast sichere Konvergenz von $\hat{\beta}(n)$ gegen β für $n \to \infty$.

Beweis:

$$\hat{\beta}(n) = (X^tX)^{-1}X^tY = \beta + (X^tX)^{-1}X^t\varepsilon = \beta + (n^{-1}X^tX)^{-1}(n^{-1}X^t\varepsilon)$$

und nach Lemma 2.26 ist $n^{-1}(X^tX)$ fast sicher gegen Σ konvergent. Damit liefert die Annahme (iii) mit dem starken Gesetz der großen Zahlen, dass $\hat{\beta}(n) \to \beta$ fast sicher für $n \to \infty$. ■

Da über die genaue Gestalt des Daten-generierenden Wahrscheinlichkeitsmaßes μ^n keine genauen (parametrischen) Annahmen gemacht wurden, bietet sich auch hier für festes n eine bootstrap-Approximation von $\mathcal{L}\left(\sqrt{n}\left[\hat{\beta}(n)-\beta\right]\right)$ an.

Schema 2.29 (Resamplingschema: Bootstrap für Modell 2.25)

(A) Seien iid. Daten $((X_i = x_i, Y_i = y_i))_{1\le i\le n}$ gemäß Modell 2.25 gegeben. Wir bezeichnen mit $\hat{\beta} \equiv \hat{\beta}(n)$ den LSE basierend auf dieser Stichprobe, also $\hat{\beta} = (X^tX)^{-1}X^tY$ mit zufälliger Design-Matrix X und Response-Vektor Y.
Ferner bezeichnen wir die $(p+1)$-variate empirische Verteilung der Daten mit $\hat{\mathbb{P}}_n$.

(B) Bezeichne $((X_i^* = x_i^*, Y_i^* = y_i^*))_{1\le i\le n}$ eine iid. bootstrap-Stichprobe. Dabei ist (klassisch) $(X_1^*, Y_1^*) : (\Omega^*, \mathcal{A}^*, \mathbb{P}^*) \to (\mathbb{R}^{p+1}, \mathcal{B}(\mathbb{R}^{p+1}))$ mit der Eigenschaft $\mathbb{P}^{*(X_1^*,Y_1^*)|Daten} = \hat{\mathbb{P}}_n$ (zufälliges gleichverteiltes Ziehen aus den original beobachteten Datentupeln mit Zurücklegen).

(C) Berechne den LSE der bootstrap-Stichprobe, also $\hat{\beta}^*(n) \equiv \hat{\beta}^* = (X^{*t}X^*)^{-1}X^{*t}Y^*$ und setze $\varepsilon^* := Y^* - X^*\hat{\beta}$ mit Werten in $\mathbb{R}^n$.

(D) Approximiere $\mathcal{L}\left(\sqrt{n}\left[\hat{\beta}(n)-\beta\right]\right)$ durch $\mathcal{L}\left(\sqrt{n}\left[\hat{\beta}^*(n)-\hat{\beta}(n)\right]|Daten\right)$.

Die Konsistenz dieser Bootstrapapproximation wurde von Stute (1990) durch Nachahmung der Beweisschritte gezeigt, die zum unbedingten multivariaten zentralen Grenzwertsatz geführt haben.

Lemma 2.30
Unter Resamplingschema 2.29 gilt

$$\forall 1 \le i \le n: \quad \forall 1 \le j \le p: \quad \mathbb{E}^*\left[X_{i,j}^*\varepsilon_i^*|Daten\right] = 0$$

Beweis: Da Erwartungswertbildung bezüglich $\mathbb{P}^*$ gegeben die Daten einer diskreten Summe mit uniformen Gewichten entspricht, erhalten wir unter Beachtung der Definition von ε^*

$$\mathbb{E}^*\left[X^*_{i,j}\varepsilon^*_i|\text{Daten}\right] = n^{-1}\sum_{k=1}^{n} X_{k,j}(Y_k - X_k^t\hat{\beta}) = n^{-1}\langle X_j, Y - X\hat{\beta}\rangle_{\mathbb{R}^n} = 0$$

nach Konstruktion von $\hat{\beta}$. ■

Lemma 2.31

Unter Resamplingschema 2.29 gilt μ^n-fast sicher für alle $\delta > 0$

$$\mathbb{P}^*\left(\left\|n^{-1}X^{*t}X^* - \Sigma\right\| > \delta\right) \xrightarrow[n\to\infty]{} 0$$

Beweis: Es ist $X^{*t}X^* = \left(\sum_{i=1}^n X^*_{i,j}X^*_{i,k}\right)_{j,k=1,\ldots,p}$. Ferner ist

$$\begin{aligned}\mathbb{E}^*\left[\sum_{i=1}^n X^*_{i,j}X^*_{i,k}|\text{Daten}\right] &= \sum_{i=1}^n \mathbb{E}^*\left[X^*_{i,j}X^*_{i,k}|\text{Daten}\right] = n\mathbb{E}^*\left[X^*_{1,j}X^*_{1,k}|\text{Daten}\right]\\ &= nn^{-1}\sum_{l=1}^n X_{l,j}X_{l,k} = \sum_{l=1}^n X_{l,j}X_{l,k}\end{aligned}$$

und damit ist

$$\begin{aligned}\mathbb{E}\left[n^{-1}\sum_{i=1}^n X^*_{i,j}X^*_{i,k}\right] &= \mathbb{E}\left[n^{-1}\mathbb{E}^*[\sum_{i=1}^n X^*_{i,j}X^*_{i,k}|\text{Daten}]\right] = n^{-1}\mathbb{E}\left[\sum_{l=1}^n X_{l,j}X_{l,k}\right]\\ &= \mathbb{E}\left[X_{1,j}X_{1,k}\right] = \Sigma_{j,k} < \infty\end{aligned}$$

für alle $1 \le j,k \le p$.

Ferner erfüllt das Modell das „Degenerate Convergence Criterion“ (siehe Loève (1977), Seite 329), was den Beweis komplettiert. ■

Lemma 2.32

Unter Resamplingschema 2.29 gilt μ^n-fast sicher

$$n^{-\frac{1}{2}}a^tX^{*t}\varepsilon^* \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, a^tMa\right) \quad \textit{für alle } a \in \mathbb{R}^p.$$

Beweis: Wir kürzen ab:

$$S^*_n := n^{-\frac{1}{2}}a^tX^{*t}\varepsilon^* = n^{-\frac{1}{2}}\sum_{k=1}^n\sum_{j=1}^p a_jX^*_{k,j}\varepsilon^*_k$$

und stellen fest, dass S^*_n eine normalisierte Summe von iid. Zufallsvariablen

$$(Z^*_k)_{1\le k\le n} := (\sum_{j=1}^p a_jX^*_{k,j}\varepsilon^*_k)_{1\le k\le n}$$

ist.

Z_1 ist nach Lemma 2.30 zentriert, also auch S_n^*. Bleibt die Varianz von S_n^* zu berechnen.

$$\begin{aligned}\text{Var}\left(S_n^*|\text{Daten}\right) &= \mathbb{E}^*\left[(Z_1^*)^2|\text{Daten}\right] = \sum_{l=1}^{p}\sum_{j=1}^{p} a_l a_j \mathbb{E}^*\left[X_{1,l}^* \varepsilon_1^* X_{1,j}^* \varepsilon_1^*|\text{Daten}\right] \\ &= \sum_{l=1}^{p}\sum_{j=1}^{p} a_l a_j \left[n^{-1}\sum_{i=1}^{n} X_{i,l}X_{i,j}(Y_i - X_i^t\hat{\beta})^2\right].\end{aligned}$$

Nach Satz 2.28 konvergiert $\hat{\beta} = \hat{\beta}(n)$ fast sicher gegen β für $n \to \infty$. Damit ist
$\forall 1 \leq i \leq n \quad (Y_i - X_i^t\hat{\beta})^2$ fast sicher gegen ε_i^2 konvergent und nach dem starken Gesetz der großen Zahlen strebt damit $n^{-1}\sum_{i=1}^{n} X_{i,l}X_{i,j}(Y_i - X_i^t\hat{\beta})^2$ fast sicher gegen $M_{l,j}$, vgl. Annahme (iv).
Zusammengefasst ergibt sich damit

$$\text{Var}^*\left(S_n^*|\text{Daten}\right) \xrightarrow[n\to\infty]{} \sum_{l=1}^{p}\sum_{j=1}^{p} a_l a_j M_{l,j} = a^t M a \quad \mu^n\text{ -fast sicher.}$$

■

Schließlich ergibt sich damit die Konsistenz der Bootstrapapproximation gemäß Resamplingschema 2.29.

Satz 2.33 (Bedingter multivariater zentraler Grenzwertsatz)
Unter Resamplingschema 2.29 gilt μ^n-fast sicher

$$\mathcal{L}\left(\sqrt{n}\left[\hat{\beta}^*(n) - \hat{\beta}(n)\right]|\textit{Daten}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0, \Sigma^{-1}M\Sigma^{-1}\right)$$

Beweis:

$$\begin{aligned}\sqrt{n}\left[\hat{\beta}^*(n) - \hat{\beta}(n)\right] &= n^{\frac{1}{2}}\left[(X^{*t}X^*)^{-1}X^{*t}(X^*\hat{\beta} + \varepsilon^*) - \hat{\beta}\right] = n^{\frac{1}{2}}(X^{*t}X^*)^{-1}X^{*t}\varepsilon^* \\ &= (n^{-1}X^{*t}X^*)^{-1}(n^{-\frac{1}{2}}X^{*t}\varepsilon^*).\end{aligned}$$

Nach Lemma 2.32 und Cramér-Wold device konvergiert die bedingte Verteilung von $n^{-\frac{1}{2}}X^{*t}\varepsilon^*$ für μ^n-fast alle Beobachtungen gegen $\mathcal{N}_p(0, M)$. Ferner liegt nach Lemma 2.31 stochastische Konvergenz von $n^{-1}X^{*t}X^*$ gegen die invertierbare Matrix Σ vor. Damit ist alles gezeigt. ■

Kapitel 3

Resamplingverfahren für multiple Testprobleme

3.1 Subset pivotality, Westfall und Young

3.2 Dudoit, van der Laan, Pollard

Kapitel 4

Resamplingverfahren für Zeitreihen und abhängige Daten

Kapitel 5

Statistisches Lernen, Klassifikationstheorie

Die statistische Lerntheorie befasst sich mit Methoden, um anhand einer bivariaten mathematischen Stichprobe $(X_i, Y_i)_{i=1,\dots,n}$ einen funktionalen Zusammenhang, also eine Funktion $f : D \to W$, $D \ni x \mapsto y \in W$, nur anhand der „Trainigsbeispiele“ $(x_i, y_i)_{i=1,\dots,n}$, d. h., ohne Formulierung eines statistischen Modells, zu „erlernen“. Dabei ist typischerweise $D \subseteq \mathbb{R}^p$ für $p >> 1$ (Matrizen und Tensoren werden häufig, wenn nötig, gecastet) und $W \subseteq \mathbb{R}$. Genauer wird in einer vorgegebenen Funktionenklasse $\mathcal{M}$ diejenige Funktion $f \in \mathcal{M}$ gesucht, die ein vorgegebenes Fehlermaß (empirisch) minimiert. Je nach Kardinalität von W unterscheidet man unterschiedliche Teildisziplinen des statistischen Lernens. Ist W überabzählbar, so spricht man von einer Regressionsaufgabe, für $W \subseteq \mathbb{N}$ von einer Klassifikationsaufgabe (englisch auch: pattern recognition problem). Kann das „Label“ Y speziell nur genau zwei diskrete Werte annehmen (sagen wir $+1$ und -1), so,liegt ein binäres Klassifikationsproblem vor. Die binäre Klassifikation hat besondere Bedeutung in der (angewandten) Informatik, da im Computer sämtliche Information binär codiert wird. So können binäre Klassifikationsalgorithmen sogar dazu benutzt werden, Computerprogramme durch externen Input (z. B. EEG-Signale) zu steuern (siehe **Referenz einfügen: BBCI-Kapitel im BCI-Buch von Graimann et al. (Hrsg.)**).

Tabellenverzeichnis

Abbildungsverzeichnis

1.1 Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$ 5
1.2 Lokal bester $\{\vartheta_0\}$ α-ähnlicher Test φ^* 12

Literaturverzeichnis

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol. 57*(1), 289–300.

Bickel, P. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics 9*, 1196–1217.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability.* Springer Texts in Statistics. New York, NY: Springer.

Dudoit, S. and M. J. van der Laan (2008). *Multiple testing procedures with applications to genomics.* Springer Series in Statistics. Springer, New York.

Efron, B. (1977, July). Bootstrap methods: Another look at the jackknife. Technical Report 37, Department of Statistics, Stanford University.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap.* Monographs on Statistics and Applied Probability. 57. New York, NY: Chapman & Hall.

Finner, H. (1994). *Testing Multiple Hypotheses: General Theory, Specific Problems, and Relationships to Other Multiple Decision Procedures*. Habilitationsschrift. Fachbereich IV, Universität Trier.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh and London.

Freedman, D. A. (1981). Bootstrapping Regression Models. *Annals of Statistics 9*, 1218–1228.

Gaenssler, P. and W. Stute (1977). *Wahrscheinlichkeitstheorie.* Hochschultext. Berlin-Heidelberg-New York: Springer-Verlag.

Hall, P. (1988). Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics 16*(3), 927–953.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* Springer Series in Statistics, New York.

Hall, P. and S. R. Wilson (1991). Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics 47*(2), 757–762.

Hewitt, E. and K. Stromberg (1975). *Real and abstract analysis. A modern treatment of the theory of functions of a real variable. 3rd printing.* Graduate Texts in Mathematics. 25. New York - Heidelberg - Berlin: Springer-Verlag.

Janssen, A. (1998). *Zur Asymptotik nichtparametrischer Tests, Lecture Notes. Skripten zur Stochastik Nr. 29.* Gesellschaft zur Förderung der Mathematischen Statistik, Münster.

Janssen, A. (2005). Resampling Student's t-type statistics. *Ann. Inst. Stat. Math. 57*(3), 507–529.

Janssen, A. and T. Pauls (2003). How do bootstrap and permutation tests work? *Ann. Stat. 31*(3), 768–806.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses. 3rd ed.* Springer Texts in Statistics. New York, NY: Springer.

Loève, M. (1977). *Probability theory I. 4th ed.* Graduate Texts in Mathematics. 45. New York - Heidelberg - Berlin: Springer-Verlag. XVII, 425 p. DM 45.00; $ 19.80 .

Pauls, T. (2003). *Resampling-Verfahren und ihre Anwendungen in der nichtparametrischen Testtheorie.* Books on Demand GmbH, Norderstedt.

Pauly, M. (2009). *Eine Analyse bedingter Tests mit bedingten Zentralen Grenzwertsätzen für Resampling-Statistiken.* Ph. D. thesis, Heinrich Heine Universität Düsseldorf.

Pitman, E. (1937). Significance Tests Which May be Applied to Samples From any Populations. *Journal of the Royal Statistical Society 4*(1), 119–130.

Shorack, G. R. and J. A. Wellner (1986). *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics. New York, NY: Wiley.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics 9*(6), 1187–1195.

Stute, W. (1990). Bootstrap of the linear correlation model. *Statistics 21*(3), 433–436.

Westfall, P. H. and S. Young (1992). *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, New York.

Witting, H. (1985). *Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang.* Stuttgart: B. G. Teubner.

Witting, H. and U. Müller-Funk (1995). *Mathematische Statistik II. Asymptotische Statistik: Parametrische Modelle und nichtparametrische Funktionale.* Stuttgart: B. G. Teubner.

Witting, H. and G. Nölle (1970). *Angewandte Mathematische Statistik. Optimale finite und asymptotische Verfahren.* Leitfäden der angewandten Mathematik und Mechanik. Bd. 14. Stuttgart: B.G. Teubner.