Thorsten Dickhaus

Theory of Nonparametric Tests

Springer

Theory of Nonparametric Tests

Thorsten Dickhaus

Theory of Nonparametric Tests

Springer

Thorsten Dickhaus
Institute for Statistics
University of Bremen
Bremen, Germany

ISBN 978-3-319-76314-9 ISBN 978-3-319-76315-6 (eBook)
https://doi.org/10.1007/978-3-319-76315-6

Library of Congress Control Number: 2018935297

Mathematics Subject Classification (2010): 62-01, 62G10, 62G09, 62G30, 62J05, 60F05, 60G15

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my family

Preface

This book originated from lecture notes, and its main purpose is to provide the basis for a one-semester graduate course on nonparametric statistical tests. Indeed, I have given such courses at the Humboldt University of Berlin in the winter term 2014/2015 and at the University of Bremen in the winter term 2016/2017, and I thank the students for many constructive comments. Special thanks for help with TeXing are due to Mareile Große Ruse. In the case that the book is used for self-studies, the reader should have knowledge in measure-theoretic probability theory on the level of an introductory course. Some advanced concepts from probability theory, in particular the concept of conditional expectations, are developed in Chap. 1 of the present work. On the other hand, also researchers may find some of the material valuable when designing nonparametric tests for specific applications.

My own interest in nonparametric test theory started during my diploma studies at Aachen University of Applied Sciences, Campus Jülich, when attending lectures given by Prof. Gerhard Dikta in the years 2001–2003. This was followed up by M. Sc. studies at Heinrich Heine University Düsseldorf from 2003 to 2005, under the supervision of Prof. Arnold Janssen. I am very grateful to these teachers of mine for the many excellent lectures which I have been attending as their student. Also, some of the teaching material from back then has been used when writing this book. In particular, the German textbook by Janssen (1998),[1] together with some original research articles of him, has been a valuable source throughout, and parts of Chap. 6 of the present work have their origins in lecture notes of Gerhard Dikta on Bootstrap Methods in Statistics.

It is impossible to cover the entire spectrum of nonparametric tests within the scope of a one-semester course. Hence, it is a matter of fact that even some very popular nonparametric tests are not explicitly covered in this work. For example, we will not work out rank tests for k-sample problems with $k > 2$ groups, although the theory derived in Chap. 4 is general enough to deduce, for example, the

[1] Janssen A (1998) Zur Asymptotik nichtparametrischer Tests, Lecture notes. Skripten zur Stochastik Nr. 29. Gesellschaft zur Förderung der Mathematischen Statistik, Münster.

Kruskal-Wallis test for the comparison of k groups. Also, we will not carry out detailed power analyses of the derived test procedures, but mainly study their behavior under the null. Narrowing the focus in this manner allows, on the other hand, for explaining in a rather detailed manner the underlying mathematical foundations and statistical principles. I consider this more important than providing comprehensive lists of test procedures in the style of "cook recipes." Once the general principles have been understood, specific procedures follow easily. A conceptual overview of the material and the underlying general ideas is provided in Sect. 1.3.

To facilitate the usage of the book for teaching purposes, exercises are provided at the end of each chapter. In general, there will be four different types of exercises, namely (1) theoretical exercises involving proofs, (2) application-oriented exercises involving real data, (3) programming exercises in `R`, mostly in terms of simulation studies, and (4) multiple select exercises.

Bremen, Germany Thorsten Dickhaus
January 2018

Contents

Acronyms

$(\Omega, \mathscr{F}, \mathbb{P})$	Probability space
$\mathscr{B}(\mathscr{Y})$	Some σ-field over $\mathscr{Y}$
$(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), \mathscr{P})$	Statistical model
ANOVA	Analysis of variance
Beta(a, b)	Beta distribution with parameters a and b
Bin(n, p)	Binomial distribution with parameters n and p
χ^2_ν	Chi-square distribution with ν degrees of freedom
$\xrightarrow{\mathscr{D}}$	Convergence in distribution
cdf	Cumulative distribution function
det(A)	Determinant of the matrix A
diag$(\ldots)$	Diagonal matrix, the diagonal elements of which are given by $\ldots$
ecdf	Empirical cumulative distribution function
ELR	Empirical likelihood ratio
$\bar{Y}_n$	Empirical mean of n observables
$\stackrel{\mathscr{D}}{=}$	Equality in distribution
Exp(λ)	Exponential distribution with intensity parameter $\lambda > 0$
F_{ν_1,ν_2}	Fisher's F-distribution with ν_1 and ν_2 degrees of freedom
$\Gamma(\cdot)$	Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt,\ x > 0$
$\lfloor x \rfloor$	Largest integer smaller than or equal to x
GEE	Generalized estimating equation
GLM	Generalized linear model
I_n	Identity matrix in $\mathbb{R}^{n \times n}$
i.i.d.	independent and identically distributed
$\mathbf{1}_A$	Indicator function of the set A
$\mathring{\Theta}$	The interior of Θ
$\mathscr{L}(Y)$	Law (or distribution) of the random variate Y
LSE	Least squares estimator
MLE	Maximum likelihood estimator
A^+	Moore-Penrose pseudo inverse of the matrix A

$\mathscr{M}_c(n, p)$	Multinomial distribution with c categories, sample size n and vector of probabilities p
NPMLE	Nonparametric maximum likelihood estimator
$\mathscr{N}(\mu, \sigma^2)$	Normal distribution on $\mathbb{R}$ with mean μ and variance σ^2
$\mathscr{N}_k(\mu, \Sigma)$	Normal distribution on $\mathbb{R}^k$ with mean vector μ and covariance matrix Σ
pdf	Probability density function
Φ	Cumulative distribution function of the standard normal law on $\mathbb{R}$
ϕ	Lebesgue density of the standard normal law on $\mathbb{R}$
pmf	Point mass function
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, +\infty\}$
sgn	Sign function
$X \perp\!\!\!\perp Y$	X is stochastically independent of Y
$\mathscr{S}_n$	Symmetric group of order n
t_ν	Student's t-distribution with ν degrees of freedom
UNI$[a, b]$	Uniform distribution on the interval $[a, b]$
$\xrightarrow{w}$	Weak convergence
$W_m(\nu, \Sigma)$	Wishart distribution with parameters m, ν and Σ
w. l. o. g.	Without loss of generality

Chapter 1
Introduction and Examples

1.1 Basics from Statistics

Let Y denote a random quantity (which may be, depending on the context, a real-valued random variable, an $\mathbb{R}^d$-valued random vector, $d > 1$, an $\mathbb{R}^{n\times k}$-valued random matrix, where n denotes a sample size and k the number of groups in a k-sample problem, etc.), which describes the possible outcome of an experiment.[1]

Let $\mathscr{Y}$ denote the sample space corresponding to Y, i.e., the set of all possible realizations of Y, and let $\mathscr{B}(\mathscr{Y}) \subseteq 2^{\mathscr{Y}}$ denote an appropriate σ-field over $\mathscr{Y}$. The elements of $\mathscr{B}(\mathscr{Y})$ are called measurable subsets of $\mathscr{Y}$ or, synonymously, events. Denote by $\mathbb{P}^Y$ the distribution of Y, assuming that Y is defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The latter probability space will typically not be given explicitly, because the fundamental Definition 1.1 below does not require such an explicit definition. However, one should always be aware of Ω in practice, meaning that we should know for which universe (or: population) our sample is representative. We assume that we do not know the distribution $\mathbb{P}^Y$, but that we can provide a family $\mathscr{P}$ of probability measures on the measurable space $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ for which we are sure that $\mathbb{P}^Y \in \mathscr{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$.

Definition 1.1 (Statistical Experiment/Model) A triple $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), \mathscr{P})$ consisting of a non-empty set $\mathscr{Y}$, a σ-field $\mathscr{B}(\mathscr{Y}) \subseteq 2^{\mathscr{Y}}$ over $\mathscr{Y}$, and a family $\mathscr{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ of probability measures on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ is called a statistical experiment or a statistical model, respectively. If $\Theta \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$, then we call $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), \mathscr{P})$ a parametric statistical model, where $\vartheta \in \Theta$ is called the parameter, and Θ is called the parameter space.

[1] Witting (1985): "We think of all the data material summarized as one "observation" [...]." (translation by the author, the observation will be denoted as $Y = y$).

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_1

Remark 1.2 We will mainly be concerned with nonparametric statistical models. For example, letting $\mathscr{Y} = \mathbb{R}$ and denoting by $\mathscr{B}(\mathbb{R})$ the system of Borel sets of $\mathbb{R}$, we may define

$$\Theta = \{F : F \text{ is a cumulative distribution function on } \mathbb{R}\}. \tag{1.1}$$

In (1.1), Θ is a function space of infinite dimension, and $\vartheta = F$ indexes *all* distributions on the real line by means of their cumulative distribution functions (cdfs). The latter model will be of much importance throughout the remainder.

The goal of statistical inference is to derive assertions about the true, but unknown distribution $\mathbb{P}^Y$ or, equivalently, about the true, but unknown and unobservable value of ϑ on the basis of the data (i.e., observation) $Y = y$. Formally, many different types of statistical inference problems can be specified as statistical decision problems. Here, we will mainly be concerned with test problems, which constitute an important class of statistical decision problems.

Definition 1.3 (Statistical Test Problem, Statistical Test) Assume that two non-empty and disjoint subsets $\mathscr{P}_0$ and $\mathscr{P}_1$ of $\mathscr{P}$ are given, such that $\mathscr{P}_0 \cup \mathscr{P}_1 = \mathscr{P}$. Our goal is to decide, on the basis of the data $Y = y$, if $\mathbb{P}^Y \in \mathscr{P}_0$ or if $\mathbb{P}^Y \in \mathscr{P}_1$ holds true. If there is a one-to-one correspondence between the elements of $\mathscr{P}$ and the value of ϑ, we can equivalently ask if $\vartheta \in \Theta_0$ or if $\vartheta \in \Theta_1$ holds true, where the non-empty, disjoint subsets Θ_0 and Θ_1 of Θ correspond to $\mathscr{P}_0$ and $\mathscr{P}_1$ in the sense that $\mathbb{P}^Y \in \mathscr{P}_0$ if and only if $\vartheta \in \Theta_0$. In the latter case, we define the null hypothesis H_0 by

$$H_0 : \vartheta \in \Theta_0 \Longleftrightarrow \mathbb{P}^Y \in \mathscr{P}_0 \tag{1.2}$$

and the alternative hypothesis by

$$H_1 : \vartheta \in \Theta_1 \Longleftrightarrow \mathbb{P}^Y \in \mathscr{P}_1. \tag{1.3}$$

Often, one directly interprets H_0 and H_1 themselves as subsets of Θ, i.e., one considers sets H_0 and H_1 such that $H_0 \cup H_1 = \Theta$ and $H_0 \cap H_1 = \emptyset$.

A (non-randomized) statistical test φ is a measurable mapping

$$\varphi : (\mathscr{Y}, \mathscr{B}(\mathscr{Y})) \to (\{0, 1\}, 2^{\{0,1\}})$$

with the convention that

$$\varphi(y) = 1 \Longleftrightarrow \text{Rejection of the null hypothesis } H_0, \text{ decision in favor of } H_1,$$
$$\varphi(y) = 0 \Longleftrightarrow \text{Non-rejection of } H_0.$$

The subset $\{y \in \mathscr{Y} : \varphi(y) = 1\}$ is called the rejection region or, synonymously, the critical region of φ, $\{\varphi = 1\}$ for short. Its complement $\{y \in \mathscr{Y} : \varphi(y) = 0\}$ is called the acceptance region of φ, $\{\varphi = 0\} = \complement\{\varphi = 1\}$ for short.

The randomness of the data $Y = y$ implies the possibility of making an error when carrying out the test. One says that an error of the first kind (type I error) occurs, if the decision in favor of H_1 is taken, although actually H_0 holds true. Analogously, an error of the second kind (type II error) occurs if H_0 is not rejected, although H_1 holds true. Typically, it is not possible to minimize the two error probabilities (type I error probability and type II error probability) at the same time for a given statistical model and fixed hypotheses H_0 and H_1. The classical (Neyman-Pearson) approach to the statistical test problem therefore treats the two errors asymmetrically. The type I error is considered as the more severe error, and its probability is bounded by a pre-defined constant $\alpha \in (0, 1)$, which is called the significance level. A standard choice is to take $\alpha = 5\%$. In the class of all tests φ for which the type I error probability does not exceed α, one then tries to find the best test in terms of minimal type II error probability. Equivalently, this means that one aims at maximizing the test power (which is the probability of rejecting the null hypothesis, if H_1 is true) under the constraint of keeping the significance level α. Notice, however, that the test power typically depends on the location of $\vartheta \in H_1$, if H_1 is a composite alternative hypothesis (i.e., $|H_1| > 1$). Hence, one typically has to decide against which regions in Θ_1 one aims at optimal power. Let us summarize these concepts formally.

Notation 1.4 *Consider the framework of Definition 1.3.*

(i) *The quantity* $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi] = \mathbb{P}_\vartheta(\varphi = 1) = \int_{\mathscr{Y}} \varphi d\mathbb{P}_\vartheta$ *denotes the rejection probability of a given test* φ *as a function of* $\vartheta \in \Theta$. *For* $\vartheta \in \Theta_1$ *we call* $\beta_\varphi(\vartheta)$ *the power of* φ *in the point* ϑ. *For* $\vartheta \in \Theta_0$, $\beta_\varphi(\vartheta)$ *is the type I error probability of* φ *under* $\vartheta \in \Theta_0$.

For fixed $\alpha \in (0, 1)$, *we call*

(ii) *a test* φ *with* $\beta_\varphi(\vartheta) \leq \alpha$ *for all* $\vartheta \in H_0$ *a* level α test,

(iii) *a level* α *test* φ unbiased, *if* $\beta_\varphi(\vartheta) \geq \alpha$ *for all* $\vartheta \in H_1$,

(iv) *a level* α *test* φ_1 better *than another level* α *test* φ_2, *if* $\beta_{\varphi_1}(\vartheta) \geq \beta_{\varphi_2}(\vartheta)$ *for all* $\vartheta \in H_1$ *and* $\exists \vartheta^* \in H_1$ *with* $\beta_{\varphi_1}(\vartheta^*) > \beta_{\varphi_2}(\vartheta^*)$.

Remark 1.5 Notice that, under the framework of Notation 1.4, a statistically safeguarded decision (namely, keeping the significance level α) can only be taken in favor of the alternative H_1. This implies the standard rule that one should express the scientific claim that one is interested to gain evidence for as the alternative hypothesis H_1. From a decision-theoretic viewpoint, one may interpret $\varphi(y) = 0$ as a decision in favor of Θ (the null hypothesis cannot be rejected, meaning that the union of both H_0 and H_1, which is whole Θ, is "compatible" with the data y).

An important subclass of tests is constituted by tests of (generalized) Neyman-Pearson type.

Definition 1.6 Let $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be a statistical model and φ a test for the pair of hypotheses $\emptyset \neq H_0 \subset \Theta$ versus $H_1 = \Theta \setminus H_0$. Assume that φ relies on

a test statistic $T : \mathscr{Y} \to \mathbb{R}$. More precisely, assume that the decision rule of φ is characterized by means of rejection regions $\Gamma_\alpha \subset \mathbb{R}$ for each significance level $\alpha \in (0, 1)$, such that $\varphi(y) = 1 \iff T(y) \in \Gamma_\alpha$, for data $y \in \mathscr{Y}$. Assume that the test statistic fulfills the monotonicity condition

$$\forall \vartheta_0 \in H_0 : \forall \vartheta_1 \in H_1 : \forall c \in \mathbb{R} : \mathbb{P}_{\vartheta_0}(T > c) \leq \mathbb{P}_{\vartheta_1}(T > c). \tag{1.4}$$

Then we call φ a test of (generalized) Neyman-Pearson type, if for every $\alpha \in (0, 1)$ there exists a constant c_α, such that

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

Remark 1.7

(a) The monotonicity condition (1.4) means that $T(Y)$ tends under alternatives to larger values than under the null.
(b) The rejection regions corresponding to a test of Neyman-Pearson (N-P) type are given by $\Gamma_\alpha = (c_\alpha, \infty)$.
(c) The constants c_α are determined in practice via $c_\alpha = \inf\{c \in \mathbb{R} : \bar{\mathbb{P}}(T > c) \leq \alpha\}$, where the probability measure $\bar{\mathbb{P}}$ is chosen such that

$$\bar{\mathbb{P}}(T \in \Gamma_\alpha) = \sup_{\vartheta_0 \in H_0} \mathbb{P}_{\vartheta_0}(T \in \Gamma_\alpha)$$

holds true, if H_0 is a composite null hypothesis ("the test is calibrated at the boundary of the null hypothesis"). If H_0 is simple (meaning that $|H_0| = 1$) and the distribution of T under $\mathbb{P}_{H_0}$ is absolutely continuous, then it holds $c_\alpha = F_T^{-1}(1 - \alpha)$, where F_T denotes the cdf of $T(Y)$ under H_0.
(d) A version of the fundamental lemma of test theory by Neyman and Pearson yields that under (1.4) and further mild conditions N-P type tests are uniformly (over all $\vartheta_1 \in H_1$) best level α tests for H_0 versus H_1.

There are dualities between test problems/tests and confidence estimation problems/confidence regions in the sense of the following definition.

Definition 1.8 Let a statistical model $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be given. Then we call $\mathscr{C} = (C(y) : y \in \mathscr{Y})$, where $C(y) \subseteq \Theta$ for all $y \in \mathscr{Y}$, a family of confidence regions at confidence level $1 - \alpha$ for $\vartheta \in \Theta$, if

$$\forall\, \vartheta \in \Theta : \mathbb{P}_\vartheta\left(\{y \in \mathscr{Y} : C(y) \ni \vartheta\}\right) \geq 1 - \alpha.$$

Theorem 1.9 (Correspondence Theorem, see, e.g., Aitchison, 1964) *Let a statistical model $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be given.*

(a) If for every $\vartheta \in \Theta$ a level α test φ_ϑ for the point null hypothesis $\{\vartheta\}$ is available, and we let $\varphi = (\varphi_\vartheta : \vartheta \in \Theta)$, then $\mathscr{C} \equiv \mathscr{C}(\varphi)$, defined by $C(y) = \{\vartheta \in \Theta : \varphi_\vartheta(y) = 0\}$, constitutes a family of confidence regions at confidence level $1 - \alpha$ for $\vartheta \in \Theta$.

(b) If $\mathscr{C}$ is a family of confidence regions at confidence level $1 - \alpha$ for $\vartheta \in \Theta$, and if we define $\varphi = (\varphi_\vartheta : \vartheta \in \Theta)$ via $\varphi_\vartheta(y) = 1 - \mathbf{1}_{C(y)}(\vartheta)$, then φ is a (multiple) test at local level α, meaning that, for every $\vartheta \in \Theta$, the test φ_ϑ is a level α test for the point null hypothesis $\{\vartheta\}$.

Proof In both parts of Theorem 1.9, we have that for all $\vartheta \in \Theta$ and for all $y \in \mathscr{Y}$ the relationship $\varphi_\vartheta(y) = 0 \iff C(y) \ni \vartheta$ holds true. Hence, φ is a (multiple) test at local level α, if and only if

$$\begin{aligned} & \forall\, \vartheta \in \Theta : \mathbb{P}_\vartheta\left(\{y \in \mathscr{Y} : \varphi_\vartheta(y) = 0\}\right) \geq 1 - \alpha \\ \Leftrightarrow \quad & \forall\, \vartheta \in \Theta : \mathbb{P}_\vartheta\left(\{y \in \mathscr{Y} : C(y) \ni \vartheta\}\right) \geq 1 - \alpha \\ \Leftrightarrow \quad & \mathscr{C} \text{ is a family of confidence regions at confidence level } 1 - \alpha \text{ for } \vartheta \in \Theta. \end{aligned}$$

Remark 1.10

(a) Figure 1.1 illustrates the duality $\varphi_\vartheta(y) = 0 \iff \vartheta \in C(y)$ graphically for the case, that both $\mathscr{Y}$ and Θ are one-dimensional.

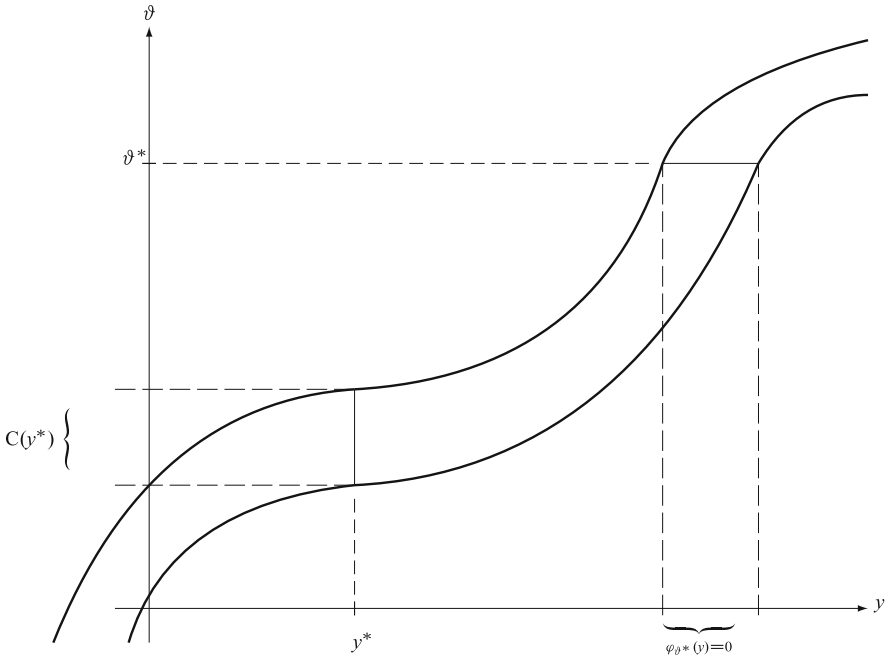

Fig. 1.1 Graphical illustration of the duality $\varphi_\vartheta(y) = 0 \iff \vartheta \in C(y)$

(b) A single level α test φ for an (arbitrary) null hypothesis H_0 can also be interpreted as a $(1-\alpha)$-confidence region by setting

$$C(y) = \begin{cases} \Theta\,, & \text{if } \varphi(y) = 0, \\ H_1 = \Theta \backslash H_0\,, & \text{if } \varphi(y) = 1, \end{cases}$$

cf. Remark 1.5. Conversely, every single confidence region $C(y)$ induces a level α test φ for a null hypothesis $H_0 \subset \Theta$ versus the alternative hypothesis $H_1 = \Theta \setminus H_0$ by setting $\varphi(y) = \mathbf{1}_{H_1}(C(y))$, where

$$\mathbf{1}_B(A) := \begin{cases} 1\,, & \text{if } A \subseteq B, \\ 0\,, & \text{otherwise,} \end{cases}$$

for arbitrary sets A and B.

Let us end this section with a theorem from measure theory which we will occasionally use for some technical proofs in the forthcoming sections.

Theorem 1.11 (Vitali's Theorem, see Witting (1985), Satz 1.181) *Let $(\Omega, \mathscr{F}, \mu)$ be a σ-finite measure space. Assume that $f_n : \Omega \to \mathbb{R}$ is a measurable mapping for each $n \in \mathbb{N}_0$. If $f_n \to f_0$ μ-almost everywhere and*

$$\limsup_{n\to\infty} \int |f_n|^p d\mu \le \int |f_0|^p d\mu < \infty \text{ for some } p \ge 1,$$

then it follows that $\int |f_n - f_0|^p d\mu \to 0$ for $n \to \infty$. If μ is a probability measure, then the assumption of μ-almost everywhere convergence of f_n to f_0 can be replaced by convergence in μ-probability of f_n to f_0.

1.2 Conditional Distributions and Expectations

Definition and Theorem 1.12 *Let X and Y be two real-valued,* jointly absolutely continuously distributed *random variables, which are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Denote the (joint) Lebesgue density of (X, Y) by $f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}_{\ge 0}$. Then the following assertions hold true.*

(a) The function f_Y, given by $f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)dx$, is a marginal Lebesgue density of Y.
(b) The function $f_{Y|X}(\cdot|\cdot)$, given by

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)} \text{ for } x, y \in \mathbb{R},$$

is a conditional Lebesgue density of Y with respect to X, where we let $f_{Y|X}(y|x) := 0$ whenever $f_X(x) = 0$.

(c) Let $\mathscr{B}(\mathbb{R})$ denote the Borel σ-algebra on $\mathbb{R}$. For $x \in \mathbb{R}$ with $f_X(x) > 0$, the set function

$$\mathscr{B}(\mathbb{R}) \ni B \mapsto \mathbb{P}(Y \in B|X = x) := \int_B f_{Y|X}(y|x)dy$$

is called the conditional distribution of Y given $X = x$.

(d) Calculation rules:

(i) $\mathbb{P}(X \in A, Y \in B) = \int_A \mathbb{P}(Y \in B|X = x) f_X(x)dx$.

(ii) $\mathbb{P}(Y \in B) = \int_{-\infty}^{\infty} \mathbb{P}(Y \in B|X = x) f_X(x)dx$.

(iii) $\mathbb{P}((X, Y) \in C) = \int_{-\infty}^{\infty} \mathbb{P}(Y \in C(x)|X = x) f_X(x)dx$ *for $C \in \mathscr{B}(\mathbb{R}^2)$ and by defining $C(x) = \{y \in \mathbb{R} : (x, y) \in C\}$.*

(iv) If A and B are Borel sets of $\mathbb{R}$ with $\mathbb{P}(X \in A) > 0$, then the elementary conditional probability *of $\{Y \in B\}$ given $\{X \in A\}$ is defined by*

$$\mathbb{P}(Y \in B|X \in A) = \frac{\mathbb{P}(X \in A, Y \in B)}{P(X \in A)}.$$

Proof All assertions follow immediately by elementary properties of the Lebesgue integral, and by Fubini's Theorem.

The goal of this section is to generalize the concepts in Definition and Theorem 1.12 to more general types of (joint) distributions of (X, Y), and to derive a notion of conditional probability which is still well-defined if the condition has probability zero.

Definition 1.13 Let $(\Omega_1, \mathscr{A}_1)$ and $(\Omega_2, \mathscr{A}_2)$ be two measurable spaces. A mapping $q : \Omega_1 \times \mathscr{A}_2 \to [0, 1]$ is called *transition probability distribution (Markov kernel)* from Ω_1 to Ω_2 (or to $\mathscr{A}_2$, respectively): $\Leftrightarrow$

(i) $A' \mapsto q(x, A')$ is a probability measure on $(\Omega_2, \mathscr{A}_2)$ for all $x \in \Omega_1$.
(ii) $x \mapsto q(x, A')$ is $(\mathscr{A}_1, \mathscr{B}([0, 1]))$-measurable for all $A' \in \mathscr{A}_2$.

Definition and Theorem 1.14 *Let $(\Omega_i, \mathscr{A}_i)$, $i = 1, 2$, be two measurable spaces. Let μ be a probability measure on $(\Omega_1, \mathscr{A}_1)$ and q a Markov kernel from Ω_1 to Ω_2.*

a) The mapping $\mu \otimes q$, defined by

$$\mu \otimes q(A_1 \times A_2) := \int_{A_1} q(x, A_2)\mu(dx),\ A_i \in \mathscr{A}_i,\ i = 1, 2,$$

is a probability measure on $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$.

b) For $C \in \mathscr{A}_1 \otimes \mathscr{A}_2$, it holds that

$$\mu \otimes q(C) = \int_{\Omega_1} q(x, C(x))\mu(dx).$$

Proof Define $Q(C) := \int_{\Omega_1} q(x, C(x))\mu(dx)$ for $C \in \mathscr{A}_1 \otimes \mathscr{A}_2$. Exercise 1.2 yields that Q is normalized and σ-additive, hence, Q is a probability measure on $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$. Now, consider a Cartesian product $A \times B \in \mathscr{A}_1 \times \mathscr{A}_2$. We verify that

$$Q(A\times B) = \int_{\Omega_1} q(x, (A\times B)(x))\mu(dx) = \int_{\Omega_1} \mathbf{1}_A(x) q(x, B)\mu(dx) = \int_A q(x, B)\mu(dx).$$

Notice that the system of all such Cartesian products is a $\cap$-stable generating system of $\mathscr{A}_1 \otimes \mathscr{A}_2$. Thus, uniqueness of measures yields that $Q =: \mu \otimes q$ is uniquely determined.

Example 1.15

a) Let $q(x, B) \equiv \nu(B)$, where ν is a probability measure on $(\Omega_2, \mathscr{A}_2)$. Then we have that

$$\begin{aligned}\mu \otimes q(A_1 \times A_2) &= \int_{A_1} q(x, A_2)\mu(dx) = \int_{A_1} \nu(A_2)\mu(dx) = \mu(A_1)\nu(A_2)\\ &= \mu \times \nu(A_1 \times A_2),\end{aligned}$$

such that $\mu \otimes q$ equals the "classical" product measure of μ and ν in this case, where we denoted the latter with the symbol $\times$ for notational convenience here.

b) Let X and Y be two stochastically independent random variables on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in Ω_1 and Ω_2, respectively. Then we have that $\mathbb{P}^{(X,Y)} = \mathbb{P}^X \times \mathbb{P}^Y$. Utilizing part a), it follows that

$$\mathbb{P}^{(X,Y)} = \mathbb{P}^X \times \mathbb{P}^Y = \mathbb{P}^X \otimes \mathbb{P}^Y, \text{ i.e., } q(t, A_2) = \mathbb{P}(Y \in A_2)$$

is a version of the conditional distribution $\mathbb{P}^{Y|X=t}$, for all $t \in \Omega_1$; cf. Definition 1.18 below.

Theorem 1.16 (Fubini's Theorem for Markov Kernels) *Under the notational framework of Theorem 1.14 let $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ denote a measurable mapping. Then we have*

$$\int_{\Omega_1\times\Omega_2} f\, d(\mu \otimes q) = \int_{\Omega_1} \left[\int_{\Omega_2} f(x, y) q(x, dy)\right] \mu(dx),$$

if one of the following conditions holds true.

(i) $f \geq 0$.
(ii) f *is quasi-integrable with respect to* $\mu \otimes q$.

Proof Satz 14.29 in Klenke (2008).

Remark 1.17

(a) If ν is a probability measure on Ω_2 and $q(x, B) :\equiv \nu(B)$, then Theorem 1.16 reduces to the classical version of Fubini's Theorem for product measures, namely,

$$\int_{\Omega_1 \times \Omega_2} f d(\mu \times \nu) = \int_{\Omega_1} \left[\int_{\Omega_2} f d\nu \right] d\mu = \int_{\Omega_2} \left[\int_{\Omega_1} f d\mu \right] d\nu$$

under the assumptions of Theorem 1.16.

(b) For the proof of Theorem 1.16, it is important that the function $h : \Omega_1 \to \overline{\mathbb{R}}$, defined by $h(x) = \int_{\Omega_2} f(x, y) q(x, dy)$, is measurable. This can be shown by algebraic induction (left to the reader).

Definition 1.18 Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let X, Y be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in $(\Omega_1, \mathscr{A}_1)$ and $(\Omega_2, \mathscr{A}_2)$, respectively. Then we call a Markov kernel q from Ω_1 to Ω_2 with the property that

$$\mathbb{P}(X \in A_1, Y \in A_2) = \int_{A_1} q(x, A_2) \mathbb{P}^X(dx) \text{ for all } A_i \in \mathscr{A}_i, i = 1, 2,$$

a *regular version of the conditional distribution of Y with respect to X*.

Shortcut notation: $\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes q = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}$.

If $(\Omega_2, \mathscr{A}_2) = (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ with $d \in \mathbb{N}$, then there always exists a regular version of $\mathbb{P}^{Y|X}$.

Definition 1.19 Under the assumptions of Definition 1.18, let $T : (\Omega_2, \mathscr{A}_2) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a measurable function which is such that $T(Y) \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$ holds true.

Then we call

$$\mathbb{E}[T(Y)|X = x] := \int T(y) q(x, dy) =: g(x)$$

a *version of the conditional expected value of $T(Y)$ under the hypothesis that $X = x$*.

Remark 1.20 Under the assumptions of Definition 1.19, the following assertions hold true.

(i) There always exists a version of $\mathbb{E}[T(Y)|X = x]$.

(ii) All versions of $\mathbb{E}[T(Y)|X = x]$ are measurable and $\mathbb{P}^X$-integrable mappings $g : \Omega_1 \to \mathbb{R}$.

Definition and Theorem 1.21 *Assume that the conditions of Definition 1.19 are fulfilled with $T = id$.*

a) The random variable $\mathbb{E}[Y|X] := g(X) = g \circ X$, which takes the value $g(x) = \mathbb{E}[Y|X = x] = \int y q(x, dy)$ in case of $X(\omega) = x$, is called conditional expectation of Y with respect to X.

b) Denote by

$$\sigma(X) = X^{-1}(\mathscr{A}_1) = \{X^{-1}(B) | B \in \mathscr{A}_1\} = \{A \in \mathscr{F} \mid \exists B \in \mathscr{A}_1 : X^{-1}(B) = A\}$$

the sub-σ-algebra of $\mathscr{F}$ generated by $X : (\Omega, \mathscr{F}) \to (\Omega_1, \mathscr{A}_1)$. Let $A \in \sigma(X)$ and let $B \in \mathscr{A}_1$ such that $X^{-1}(B) = A$. Then it holds that

$$\begin{aligned}\int_A Y d\mathbb{P} &= \int_\Omega \mathbf{1}_B(X) Y d\mathbb{P} = \int_{\Omega_1 \times \mathbb{R}} \mathbf{1}_B(x) y d\mathbb{P}^{(X,Y)}(x, y) \\ &= \int_{\Omega_1} \mathbf{1}_B(x) \left[\int_{\mathbb{R}} y q(x, dy)\right] \mathbb{P}^X(dx) = \int_{\Omega_1} \mathbf{1}_B(x) g(x) \mathbb{P}^X(dx) \\ &= \int_B g(x) \mathbb{P}^X(dx) = \int_\Omega \mathbf{1}_B(X) g(X) d\mathbb{P} = \int_A g \circ X d\mathbb{P} \\ &= \int_A \mathbb{E}[Y|X] d\mathbb{P}.\end{aligned}$$

c) Let, more generally, $\mathscr{C}$ be any sub-σ-algebra of $\mathscr{F}$. Then, a conditional expectation $Z \in \mathbb{E}[Y|\mathscr{C}]$ (where we will typically write $Z = \mathbb{E}[Y|\mathscr{C}]$ instead) is characterized by the following two properties.

(i) Z is $(\mathscr{C}, \mathscr{B}(\mathbb{R}))$-measurable.
(ii) $\forall C \in \mathscr{C} : \int_C Z d\mathbb{P} = \int_C Y d\mathbb{P}$.

Formally, one may write such a sub-σ-algebra $\mathscr{C} \subseteq \mathscr{F}$ as $\sigma(X)$ for a suitable random variable X.

Example 1.22

a) Assume that characters in a transmission channel get disrupted with an unknown disruption probability. We model this unknown disruption probability as a random variable X with values in $(0, 1)$. Assume that for a given realization $X = p$ the disruptions occur independently, each with the same probability p. Let the random variable Y denote the waiting time until the first disruption occurs, measured in the number of transmitted characters. We are interested in the mean "time" until the first disruption.

Solution: A version of $\mathbb{P}^{Y|X=p}$ is the geometric distribution with parameter p, hence $\mathbb{P}(Y = k|X = p) = p(1 - p)^k$, $k \geq 0$. It follows that

$$\mathbb{E}[Y|X = p] = \sum_{k=0}^{\infty} kp(1 - p)^k = \frac{1 - p}{p} = g(p).$$

Now, assume that the (random) disruption probability X has a two-point distribution, such that $\mathbb{P}(X = \frac{1}{2}) =: a$ and $\mathbb{P}(X = \frac{3}{4}) = 1 - a$, for some

$a \in (0, 1)$. Then we obtain that

$$\mathbb{E}[Y|X] = \frac{1 - X}{X} =: Z,$$

where Z is a random variable with $\mathbb{P}(Z = 1) = a = 1 - \mathbb{P}(Z = \frac{1}{3})$.

b) Let Y be a real-valued, integrable random variable and X a discrete random variable with values in $\mathbb{N}_0$, where X and Y are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then, $g(i) := \mathbb{E}[Y|X = i]$ for $i \in \mathbb{N}_0$ with $\mathbb{P}(X = i) > 0$ can be calculated as follows. From the formula for elementary conditional probabilities, we obtain that

$$\begin{aligned}\mathbb{P}(Y \in B|X = i) &= \frac{\mathbb{P}(Y \in B, X = i)}{\mathbb{P}(X = i)} \\ &= [\mathbb{P}(X = i)]^{-1} \int \mathbf{1}_{\{Y \in B\}} \mathbf{1}_{\{X=i\}} d\mathbb{P}.\end{aligned}$$

This leads to

$$g(i) = \frac{\mathbb{E}[Y \mathbf{1}_{\{X=i\}}]}{\mathbb{P}(X = i)}, i \in \mathbb{N}_0.$$

For a detailed verification of the latter result, see Exercise 1.3. For example, we get for $X := \lfloor Y \rfloor$ that

$$\mathbb{E}[Y|X = i] = \frac{\mathbb{E}[Y \mathbf{1}_{\{i \le Y < i+1\}}]}{\mathbb{P}(i \le Y < i + 1)} = g(i), i \in \mathbb{N}_0.$$

Remark 1.23 (Evocative Interpretation of the Conditional Expectation) A conditional expectation $Z = \mathbb{E}[Y|X]$ (more precisely: $Z \in \mathbb{E}[Y|X]$) has the following properties.

(i) Z is defined on the same probability space as Y.
(ii) The mean of Z equals the mean of Y, when restricted to sets of the form $X^{-1}(B)$.
(iii) Due to $Z = g(X)$ the random variable Z varies only as strongly as X. If, for instance, X can only take finitely many values, then the same holds true for $Z = \mathbb{E}[Y|X]$, even if the image of Ω under Y is an uncountable set. In this sense, the conditional expectation smoothens Y along X.
(iv) We obtain the graphical illustration displayed in Fig. 1.2.
(v) If Y lies in $\mathscr{L}_2(\Omega, \mathscr{F}, \mathbb{P})$, then $\mathbb{E}[Y|X]$ yields the best L_2-approximation of Y among all functions of the form $h(X)$, where $h : \Omega_1 \to \mathbb{R}$. This means that the L_2-distance between Y and any (deterministic) L_2-transformation of X is minimum for $\mathbb{E}[Y|X]$. In other words, $\mathbb{E}[Y|X]$ is the projection of Y onto $\mathscr{L}_2(\Omega, \sigma(X), \mathbb{P})$.

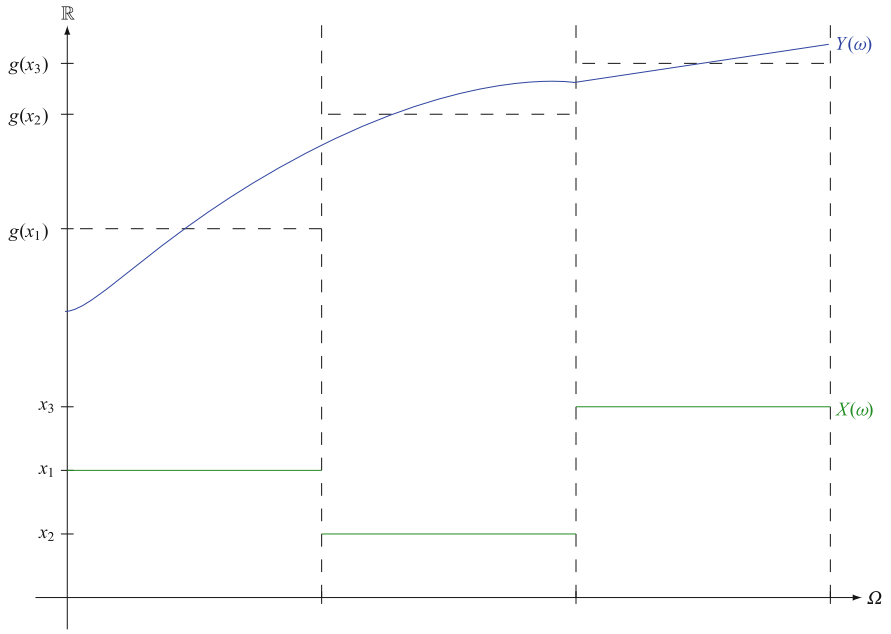

Fig. 1.2 Graphical illustration of $\mathbb{E}[Y|X]$

Let us end this section with important calculation rules for conditional expectations.

Theorem 1.24 (Calculation Rules for Conditional Expectations) *Under the assumptions of Definition 1.19 the following calculation rules hold true $\mathbb{P}$-almost surely.*

a) Linearity of the conditional expectation:

$$\mathbb{E}[\alpha Y_1 + \beta Y_2|X] = \alpha\mathbb{E}[Y_1|X] + \beta\mathbb{E}[Y_2|X].$$

b) Law of iterated expectations:

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right] = \int_{\Omega_1} \mathbb{E}[Y|X = x]\mathbb{P}^X(dx).$$

c) Let $h : \Omega_1 \times \mathbb{R} \to \mathbb{R}$ be such that $h(X, Y)$ is integrable. Then it holds:

(i) $\mathbb{E}[h(X, Y)|X = x] = \mathbb{E}[h(x, Y)|X = x] = \int h(x, y)\mathbb{P}^{Y|X=x}(dy)$.
(ii) $X \perp\!\!\!\perp Y \Rightarrow \mathbb{E}[h(X, Y)|X = x] = \mathbb{E}[h(x, Y)] = \int h(x, y)\mathbb{P}^Y(dy)$.

d) Let $h : \Omega_1 \to \mathbb{R}$ be measurable and such that $Y \cdot h(X)$ is integrable. Then it holds that

$$\mathbb{E}[Y \cdot h(X)|X] = h(X) \cdot \mathbb{E}[Y|X].$$

e) Let $g : (\Omega_1, \mathscr{A}_1) \to (\Omega', \mathscr{A}')$. *Then it holds that*

$$\mathbb{E}\left[\mathbb{E}[Y|X]|g(X)\right] = \mathbb{E}[Y|g(X)] = \mathbb{E}\left[\mathbb{E}[Y|g(X)]|X\right].$$

f) Tower equation: Let $\mathscr{B}_1 \subset \mathscr{B}_2$ *be sub-*σ*-algebras of* $\mathscr{F}$*, and assume that* $Y \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$. *Then it holds that*

$$\mathbb{E}\left[\mathbb{E}[Y|\mathscr{B}_1]|\mathscr{B}_2\right] = \mathbb{E}[Y|\mathscr{B}_1] = \mathbb{E}\left[\mathbb{E}[Y|\mathscr{B}_2]|\mathscr{B}_1\right].$$

Notice: *Sub-*σ*-algebras can be interpreted as levels of information!*

Proof All assertions follow immediately from properties of the Lebesgue integral (cf. measure and integration theory) or can be verified by algebraic induction (the reader may, for instance, verify part c) for indicator functions).

1.3 Overview and Motivating Examples

For concreteness, let us consider a sample of real-valued, stochastically independent observables $Y = (Y_1, \ldots, Y_n)^\top$ on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where n denotes the sample size. We are not willing to assume a parametric model for the distribution $\mathbb{P}^Y$ of Y, but rather consider (at least in the first place) the family $\mathscr{P}$ of all product measures on $\mathbb{R}^n$ as a statistical model for Y. Certain restrictions of $\mathscr{P}$ will however be inherent depending on the type of problem. For example, in a so-called one-sample problem it will be assumed that all Y_i have the same (marginal) distribution $\mathbb{P}^{Y_i} \equiv \mathbb{P}^{Y_1}$, $1 \leq i \leq n$. If, in this context, the mean of Y_1 shall be tested, then one has to restrict $\mathscr{P}$ further by assuming that the first moment of Y_1 actually exists. Of course, the advantage of considering $\mathscr{P}$ is that the issue of model misspecification, which is often problematic in parametric models, is avoided. On the other hand, many familiar concepts from the parametric case, for example the likelihood ratio approach to testing hypotheses, do not apply straightforwardly anymore in the nonparametric setting, which requires new strategies for hypothesis testing.

In this work, the leading strategy will be the substitution principle. Restricting attention for the moment to one-sample problems, we have that the model can be expressed as

$$\mathscr{P} = \{\mathbb{P}^Y = P^{\otimes n} : P \text{ is a probability measure on } \mathbb{R}\},$$

where we set $P := \mathbb{P}^{Y_1}$ for notational convenience. If P would be known, the data-generating distribution $\mathbb{P}^Y$ would be known as well. Hence, statistical hypotheses can be formalized in terms of P. Let Θ be the set of all probability measures on $\mathbb{R}$, and Θ_0 some subset of Θ. Typically, Θ_0 will be characterized in terms of certain properties of P, for example the property that P has mean zero. Now, assume that

we can define some distance measure $\rho : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$ fulfilling $\rho(P, Q) = 0$ iff $P = Q$. Then, we may equivalently express the null hypothesis $H_0 : P \in \Theta_0$ in terms of ρ, namely,

$$H_0 : \inf_{Q \in \Theta_0} \rho(P, Q) = 0. \tag{1.5}$$

Having re-written H_0 in the form (1.5), the substitution principle works as follows. We substitute the unknown data-generating distribution P by the empirical measure $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$. This probability measure simply assigns the mass $1/n$ to each observation in the sample. Notice, however, that $\hat{P}_n$ is itself random, meaning that the value of $\hat{P}_n(A)$ depends on the data $Y = y$, where A is some Borel set on the real line.

Now, a suitable test statistic for testing (1.5) is given by

$$T_n = \inf_{Q \in \Theta_0} \rho(\hat{P}_n, Q), \tag{1.6}$$

and we will reject H_0 for large values of T_n. The calibration of the test based on T_n with respect to type I error control requires knowledge about the stochastic properties of the random fluctuations of $\hat{P}_n$ around P. Restricting (w. l. o. g.) attention to special sets A of the form $A = (-\infty, x]$ for $x \in \mathbb{R}$, this immediately leads to studying properties of the empirical process

$$\sqrt{n} \left(\hat{F}_n - F \right),$$

where F is the cdf corresponding to P and $\hat{F}_n$, given by

$$\hat{F}_n(y) = \hat{P}_n((-\infty, y]) = n^{-1} \sum_{i=1}^n \mathbf{1}_{(-\infty, y]}(Y_i),$$

denotes the empirical cdf of $Y = (Y_1, \ldots, Y_n)^\top$. A brief review of results from the theory of empirical processes will therefore be provided in Chap. 2.

Presumably, the most stringent application of the substitution principle is the problem of testing for goodness-of-fit, where in the simplest case H_0 is a simple null hypothesis, meaning that $\Theta_0 = \{P_0\}$ for some given probability distribution P_0. Different choices of ρ lead to classical goodness-of-fit tests like the Kolmogorov-Smirnov test or the Cramér-Smirnov-von Mises test, cf. Chap. 3. The case that H_0 is a composite null hypothesis, where Θ_0 is characterized via the value of a functional $\kappa : \Theta \to \mathbb{R}$, $Q \mapsto \kappa(Q) \in \mathbb{R}$, leads to projection tests which we will study in Chap. 7.

The connection to rank tests (see Chap. 4) can be drawn by observing that $n\hat{F}_n(Y_i)$ equals the rank (i.e., the position in the ordered sample) of Y_i among $Y_1, \ldots, Y_n$, at least in the absence of ties (identical observations). Furthermore, a

direct connection by exploiting the substitution principle in the context of canonical gradients of statistical functionals is given in Sect. 4.3. As we will explain in Chap. 4, the null distribution of a rank test is typically given by the uniform distribution on the symmetric group $\mathscr{S}_n$, which is the set of all permutations of $\{1, \ldots, n\}$. This provides the link to permutation tests (see Sects. 1.3.2 and 5.2.1). Finally, as we will see in Chap. 5, permutation tests can be embedded into the large class of resampling tests, of which also Efron's bootstrap is a prominent member. In the case of the bootstrap, the substitution principle is applied in a slightly modified manner. Namely, the unknown or intractable theoretical null distribution of an (almost) arbitrary test statistic T_n, which does not necessarily have to be of the form as in (1.6), is approximated by an empirical version, where the substitution of P by $\hat{P}_n$ is in a certain sense performed in two different layers, namely in an estimator of the functional of interest and in the approximation of its null distribution.

The remaining parts of this section provide a more concrete outlook on the bootstrap and permutation test approaches.

1.3.1 Bootstrap Tests for One-Sample Problems

Let us consider the problem of constructing a nonparametric test for a univariate mean based on an independent and identically distributed (i.i.d.) sample $Y_1, \ldots, Y_n$, where all Y_i are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and each Y_i takes values in $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, $1 \leq i \leq n$. For ease of exposition, consider the one-sided pair of hypotheses

$$H_0 : \mathbb{E}[Y_1] = 0 \text{ versus } H_1 : \mathbb{E}[Y_1] > 0,$$

assuming that $\mathbb{E}[Y_1]$ exists.

Suitable test statistics for this test problem are based on the arithmetic mean $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$, which is an estimator for $\mathbb{E}[Y_1]$ that is based on the substitution principle. Indeed, notice that

$$\mathbb{E}[Y_1] = \int y d\mathbb{P}^{Y_1}(y), \quad \bar{Y}_n = \int y d\hat{P}_n(y).$$

Assuming that $\sigma^2 := \mathrm{Var}(Y_1) \in (0, \infty)$ is unknown (which is often the case in practice), one typically chooses the test statistic

$$T_n = \sqrt{n} \cdot \bar{Y}_n / V_n^{\frac{1}{2}} \text{ with } V_n = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

For the calibration of the resulting test φ_n with respect to the significance level α, it is necessary to determine the critical value $c_n(\alpha)$, such that $\{\varphi_n = 1\} = \{T_n >$

$c_n(\alpha)\}$. Obviously, this requires knowledge about the distribution of T_n under H_0. If we can assume that Y_1 is normally distributed, then Student (1908) provided the solution to this problem. Namely, the null distribution of T_n is then equal to Student's t-distribution with $n-1$ degrees of freedom, t_{n-1} for short, and we choose $c_n(\alpha) = F_{t_{n-1}}^{-1}(1-\alpha)$, which is the upper α-quantile of t_{n-1}. However, in almost all other cases the exact null distribution of T_n is either intractable or, if no concrete distributional assumption can be made for Y_1, simply unknown.

One way out of this dilemma is to rely on asymptotics as $n \to \infty$. Namely, according to the central limit theorem for i.i.d. random variables, we have that

$$\mathscr{L}\left(\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{w} \mathscr{N}(0,1)$$

as $n \to \infty$, with $\mu = \mathbb{E}[Y_1]$. Furthermore, it is easy to show that V_n is a consistent estimator of σ^2. Hence, application of Slutsky's lemma yields that the choice $c_n(\alpha) = \Phi^{-1}(1-\alpha)$ leads to asymptotic type I error control at level α of φ_n, where Φ denotes the cdf of the standard normal distribution on $\mathbb{R}$. Unfortunately, this asymptotic approximation of the null distribution of T_n is often not accurate enough if n is small or moderate.

The nonparametric bootstrap yields a finite-sample approximation of the null distribution of T_n which is based on the substitution principle. Namely, we replace $P^{\otimes n}$ by $\hat{P}_n^{\otimes n}$ in the construction of the null distribution of T_n. To this end, consider a new probability space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ and random variables $Y_1^*, \ldots, Y_n^*$ with $Y_i^* : (\Omega^*, \mathscr{F}^*, \mathbb{P}^*) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ for $1 \le i \le n$, such that

$$\mathbb{P}^{*(Y_1^*,\ldots,Y_n^*)^\top | (Y_1,\ldots,Y_n)^\top} = \hat{P}_n^{\otimes n}.$$

One may think of $(Y_1^*, \ldots, Y_n^*)^\top$ as a pseudo-sample which is constituted by n independent drawings with replacement from the original observables $Y_1, \ldots, Y_n$. This justifies to call the bootstrap a *resampling* procedure. Based on this construction, we replace the theoretical null distribution, i.e., the distribution

$$\mathbb{P}(\bar{Y}_n - \mu \le t), \ t \in \mathbb{R}, \tag{1.7}$$

of the estimation error of $\bar{Y}_n$ by its bootstrap analogue

$$\mathbb{P}^*\left(\bar{Y}_n^* - \bar{Y}_n \le t | (Y_1, \ldots, Y_n)^\top\right), \ t \in \mathbb{R}, \tag{1.8}$$

which we evaluate at our actually observed data $Y_1 = y_1, \ldots, Y_n = y_n$. Since (1.8) is purely empirical, it is (for given data) exactly computable, at least in principle. To see this, notice that, given the original observables, there are only finitely many possible realizations of $(Y_1^*, \ldots, Y_n^*)^\top$, namely, n^n of them, which are all equally likely. Hence, we can evaluate all n^n possible realizations of $\bar{Y}_n^* - \bar{Y}_n$ (where $\bar{Y}_n = \bar{y}_n$ is kept fixed given the original data), and simply count how many of these values

do not exceed our argument $t \in \mathbb{R}$. Analogously, the bootstrap approximation of the critical value for the test φ_n is given by the $(1-\alpha)$-quantile of the discrete bootstrap distribution of T_n.

Remark 1.25 If n is too large in order to traverse all n^n possible bootstrap resamples, one can use a Monte Carlo variant of the bootstrap, where only $B < n^n$ randomly chosen bootstrap resamples are traversed.

As an illustration of the fact that the (conditional to the data) bootstrap distribution is explicitly given, let us calculate (conditional) moments of bootstrap random variables.

Theorem 1.26 (Conditional Moments of Bootstrap Variates) *Let* $Y = (Y_1, \ldots, Y_n)^\top$ *denote the vector of i.i.d. real-valued original observables, and* $\{m(n)\}_{n\in\mathbb{N}}$ *a sequence of integers. Then the following assertions hold true.*

$$\mathbb{E}^*[Y_1^*|Y] = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n, \tag{1.9}$$

$$\mathbb{E}^*\left[\frac{1}{m(n)}\sum_{i=1}^{m(n)} Y_i^*|Y\right] = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n, \tag{1.10}$$

$$\mathbb{E}^*[Y_1^{*2}|Y] = \frac{1}{n}\sum_{i=1}^{n} {Y_i}^2, \tag{1.11}$$

$$Var\left(Y_1^*|Y\right) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2, \tag{1.12}$$

$$Var\left(\frac{1}{m(n)}\sum_{i=1}^{m(n)} Y_i^*|Y\right) = \frac{1}{n\cdot m(n)}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2, \tag{1.13}$$

$$\mathbb{E}^*[Y_1^{*3}|Y] = \frac{1}{n}\sum_{i=1}^{n} {Y_i}^3. \tag{1.14}$$

Proof This is Exercise 1.6.

In order to finish our motivation of the bootstrap technique, recall that the approximation error of the normal approximation is in the general case of asymptotic order $O\left(n^{-1/2}\right)$, which is often denoted as "correctness of the first order". On the other hand, as proved by Hall (1992) by means of asymptotic (Edgeworth) expansions, the bootstrap is, under mild conditions, second-order correct, meaning that its approximation error achieves the asymptotic order $O\left(n^{-1}\right)$; cf. Hall (1988, 1992). Technically, the argumentation is based on the concept of "bias correction" or "skewness correction", meaning that certain terms in the Edgeworth expansions which refer to the third cumulant of Y_1 vanish.

In the general case of a one-sample problem with i.i.d. observables $Y_1, \ldots, Y_n$, the bootstrap method can be characterized as follows. We consider a statistical functional

$$\kappa : \{Q : Q \text{ distribution on } \mathrm{supp}(Y_1)\} \to \mathbb{R}$$
$$Q \mapsto \kappa(Q).$$

Typically, $\kappa(Q)$ will be chosen as some kind of quantitative characteristic of Q. Since $P = \mathbb{P}^{Y_1}$ is unknown, we estimate $\kappa(P)$ by plug-in, i.e., by applying the substitution principle. Thus, the estimator is simply given by $\kappa(\hat{P}_n)$. For the construction of level α-tests or $(1-\alpha)$-confidence regions for $\kappa(P)$, we need information about the error distribution

$$\mathbb{P}(\kappa(\hat{P}_n) - \kappa(P) \le t), \ t \in \mathbb{R}, \tag{1.15}$$

or at least an approximation of it. The (nonparametric) bootstrap approximation of (1.15) is provided by evaluating the bootstrap analogue of (1.15), which is given by

$$\mathbb{P}^* \left(\kappa(\hat{P}_n^*) - \kappa(\hat{P}_n) \le t | (Y_1, \ldots, Y_n)^\top \right), \ t \in \mathbb{R}, \tag{1.16}$$

where $\kappa(\hat{P}_n^*)$ means that we evaluate the functional on the bootstrap resample.

1.3.2 Permutation Tests for Two-Sample Problems

For two-sample problems, another idea can be used in order to derive a resampling method for the comparison of the two groups. We consider again stochastically independent random variables $Y_1, \ldots, Y_n$ which are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, each taking values in the same space $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$. However, we now do not assume that all Y_i are identically distributed in general, but that there exists a number $2 \le n_1 \le n-2$ such that $Y_1, \ldots, Y_{n_1}$ belong to the first group with $\mathbb{P}^{Y_1} = \mathbb{P}^{Y_2} = \ldots \mathbb{P}^{Y_{n_1}} = P$ and $Y_{n_1+1}, \ldots, Y_n$ belong to the second group with $\mathbb{P}^{Y_i} = Q$ for all $n_1 + 1 \le i \le n$. Hence, the model is characterized by

$$\Theta = \{(P, Q) : P \text{ and } Q \text{ are probability distributions on } (\mathscr{Y}, \mathscr{B}(\mathscr{Y}))\},$$

where again certain further restrictions may be inherent to the actual problem at hand.

Often, one will be concerned with the null hypothesis $H_0 : P = Q$ of equality of the group-specific distributions, with corresponding alternative $H_1 : P \ne Q$. In particular, $Y_1, \ldots, Y_n$ are exchangeable under H_0, meaning that in this case $\mathbb{P}^{(Y_{i_1}, \ldots, Y_{i_k})^\top} = \mathbb{P}^{(Y_{\pi(i_1)}, \ldots, Y_{\pi(i_k)})^\top}$ holds true for any $1 \le k \le n$, $(i_1, \ldots, i_k) \subseteq$

$\{1, \ldots, n\}$, and $\pi \in \mathscr{S}_n$ (permutation invariance of all k-variate marginal distributions). This property implies that the null distribution of any test statistic T_n for testing H_0 versus H_1 remains the same if we apply any $\pi \in \mathscr{S}_n$ to $Y_1, \ldots, Y_n$, meaning that $\forall \pi \in \mathscr{S}_n : \mathscr{L}(T_n(Y_1, \ldots, Y_n)) = \mathscr{L}(T_n(Y_{\pi(1)}, \ldots, Y_{\pi(n)}))$ under H_0, while this property is typically violated under the alternative H_1, at least for reasonable choices of T_n. Therefore, a conditional (to the observed data) exact null distribution of T_n, which can be used for calibrating a test for H_0 versus H_1 with respect to type I error control, is given by the permutation distribution of T_n. We simply traverse all $\pi \in \mathscr{S}_n$, calculate the value of $T_n(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ on our data sample, and store all these $n!$ values. Now, if, for example, T_n tends to larger values under H_1, we take the upper α-quantile of the discrete distribution which puts a point mass of $1/n!$ in each $\pi \in \mathscr{S}_n$ as the (conditional) critical value for the permutation test of H_0 versus H_1. Again, if n is too large for carrying out all $n!$ possible permutations, a Monte Carlo variant may be employed instead, such that only $B < n!$ randomly chosen permutations $\pi \in \mathscr{S}_n$ are traversed. This permutation test approach is also based on resampling, but in contrast to the bootstrap, here the resamples of size n are drawn without replacement from the original observables $Y_1, \ldots, Y_n$, which is equivalent to permuting $Y_1, \ldots, Y_n$. The advantage of drawing without replacement is that the permutation test exactly keeps the significance level α, for any finite sample size n, if exchangeability of $Y_1, \ldots, Y_n$ under H_0 holds true. On the other hand, notice that the permutation approach is inappropriate in one-sample problems as discussed in Sect. 1.3.1, because in that case exchangeability of $Y_1, \ldots, Y_n$ also holds under H_1, implying that the rejection probability of the permutation test is bounded from above by α under the alternative as well.

1.4 Notes on the Literature

Overviews of nonparametric test methods are provided by Krishnaiah and Sen (Eds.) (1984), Hollander et al. (2014), Wilcox (2012) and, with an emphasis on computation, Neuhäuser (2012).

Theory and practice of rank tests are described by Hájek et al. (1999), Lehmann (2006), Büning and Trenkler (1994), Büning (1991), Brunner and Munzel (2013), Puri (Ed.) (1970), and Duller (2008).

Randomization and resampling tests are treated by Edgington and Onghena (2007), Good (2005, 2006), Berry et al. (2016), Pesarin and Salmaso (2010), Efron and Tibshirani (1993), Politis et al. (1999), Mammen (1992), Davison and Hinkley (1997), Barbe and Bertail (1995), Lahiri (2003), and LePage and Billard (Eds.) (1992).

Projection tests have been developed by Owen (2001), Basu et al. (2011), and Pardo (2006).

A valuable source for theory and application of goodness-of-fit tests is the book by D'Agostino and Stephens (Eds.) (1986).

1.5 Exercises

Exercise 1.1 (Taken from Küchler (2016)) *Let $(X_1, X_2)^\top$ denote a bivariate random vector on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in $\mathbb{R}^2$. Assume that $(X_1, X_2)^\top$ possesses the (joint) Lebesgue density f_{X_1,X_2}, given by*

$$f_{X_1,X_2}(x_1, x_2) = \frac{2}{\pi r^2} \mathbf{1}_{H_r}(x_1, x_2), \quad (x_1, x_2)^\top \in \mathbb{R}^2, \text{ where}$$
$$H_r = \{(x_1, x_2)^\top \in \mathbb{R}^2 : x_2 \in [0, r], x_1^2 + x_2^2 \leq r^2\}$$

for a given radius $r > 0$.

(a) Derive (the canonical version of) the marginal density f_{X_1} of X_1.
(b) Derive the conditional density $f_{X_2|X_1=x_1}$ of X_2 given that $X_1 = x_1 \in (-r, r)$.
(c) Compute the conditional expected value $\mathbb{E}[X_2|X_1 = x_1]$ of X_2 given that $X_1 = x_1 \in (-r, r)$.
(d) Derive the conditional expectation $\mathbb{E}[X_2|X_1]$ of X_2 with respect to X_1.

Exercise 1.2 *Check that the set function Q defined in the proof of Theorem 1.14 is normalized and σ-additive.*

Exercise 1.3 *Verify the formula for $g(i)$ in part b) of Example 1.22 by checking the characterizing integral formula for conditional expectations.*

Exercise 1.4 *Let X and Y denote two discrete random variables defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this, let $X(\omega) \in \{x_1, x_2, \ldots\}$ and $Y(\omega) \in \{y_1, y_2, \ldots\}$, where $\mathbb{P}(X = x_i) > 0$ for all i. Furthermore, assume that $Y \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$.*

(a) The elementary formula for conditional probabilities yields that $g(i) := \mathbb{E}[Y|X = x_i] = \sum_{j\geq 1} y_j \mathbb{P}(Y = y_j|X = x_i)$. Use this result to derive a version of $\mathbb{E}[Y|X]$.
(b) Verify by elementary calculations that $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$ holds true.

Exercise 1.5 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) Let $\mathbf{X} = (X_1, X_2)^\top$ be a random vector with values in $\mathbb{R}^2$ which is (jointly) normally distributed with a positive definite covariance matrix $\Sigma = Cov(\mathbf{X})$. Then, the conditional distribution of X_1 given $X_2 = x_2$ is a normal distribution on $\mathbb{R}$ for every fixed value $x_2 \in \mathbb{R}$.
(b) Let X and Y be two stochastically independent random variables which are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this, assume that $Y \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$. Then, the conditional expectation $\mathbb{E}[Y|X]$ of Y with respect to X is $\mathbb{P}$-almost surely equal to Y.
(c) Let X and Y be two random variables which are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this, assume that $Y \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$. Then, the cardinality

of the support of the conditional expectation $\mathbb{E}[Y|X]$ of Y with respect to X is upper-bounded by the cardinality of the support of X.

(d) *Let $(X_i)_{i\geq 1}$ be a sequence of stochastically independent, real-valued random variables which are all defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this, assume that $\mathbb{E}[X_i] = 0$ for all $i \geq 1$. For $n \in \mathbb{N}$, denote by $S_n := \sum_{i=1}^{n} X_i$ the n-th partial sum of the X_i. Then, it holds $\mathbb{P}$-almost surely for all $n \in \mathbb{N}$ and all $m > n$, that $\mathbb{E}[S_m|S_n] = S_n$.*

Exercise 1.6 *Prove Theorem 1.26.*

Exercise 1.7 (Bootstrap Inclusion Probabilities)

(a) *Under the general assumptions of Sect. 1.3.1, determine the (conditional) probability for the event that a bootstrap pseudo sample $(Y_1^*, \ldots, Y_n^*)^\top$ contains the data point $Y_1 = y_1$, i.e., calculate*

$$p(n) := \mathbb{P}^* \left(\exists 1 \leq i \leq n : Y_i^* = Y_1 \mid (Y_1, \ldots, Y_n)^\top \right).$$

(b) *Calculate $p_\infty := \lim_{n\to\infty} p(n)$. Provide the numerical value of p_∞, rounded to three decimal places.*

Exercise 1.8 (Programming Exercise)

(a) *Implement the nonparametric bootstrap for the mean of a univariate sample, which has been introduced in Sect. 1.3.1, in* R.

(b) *Evaluate how accurately this procedure keeps the significance level α. To this end, carry out a Monte Carlo simulation study for $\alpha = 5\%$ and sample sizes $n \in \{20, 50, 100, 500, 1000\}$ under the null hypothesis. For each n, perform 10,000 Monte Carlo simulation runs and assess the relative rejection frequency of the nonparametric bootstrap.*

Exercise 1.9 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) *If, under the general assumptions of Sect. 1.3.1, the alternative hypothesis holds true, then the considered bootstrap procedure approximates the distribution of $\kappa(\hat{P}_n)$ under the alternative.*

(b) *If, under the general assumptions of Sect. 1.3.2, the considered permutation test is carried out as a randomized test, then its rejection probability under the null hypothesis exactly equals the specified significance level, for any fixed sample size $n \in \mathbb{N}$.*

(c) *Let $X_1, \ldots, X_d$, $d \geq 2$, be stochastically independent and identically distributed, real-valued random variables, such that X_1 possesses the centered normal distribution on $\mathbb{R}$ with variance $\sigma^2 > 0$. Moreover, let ν be a given positive integer and S a further real-valued random variable, stochastically independent of the X_i's, such that $\nu S^2/\sigma^2$ possesses the chi-square distribution*

with ν degrees of freedom. Then, the random variables $(Y_i)_{1\le i\le d}$ with $Y_i := X_i/S$ are exchangeable, but not stochastically independent.

(d) *The order statistics of stochastically independent random variables $Y_1, \ldots, Y_n$, $n \ge 2$, are stochastically independent.*

References

Aitchison J (1964) Confidence-region tests. J R Stat Soc Ser B 26:462–476

Barbe P, Bertail P (1995) The weighted bootstrap. Springer, Berlin

Basu A, Shioya H, Park C (2011) Statistical inference. The minimum distance approach. CRC Press, Boca Raton, FL. https://doi.org/10.1201/b10956

Berry KJ, Mielke PW Jr, Johnston JE (2016) Permutation statistical methods. An integrated approach. Springer, Cham. https://doi.org/10.1007/978-3-319-28770-6

Brunner E, Munzel U (2013) Nichtparametrische Datenanalyse. Unverbundene Stichproben, 2nd edn. Springer Spektrum, Heidelberg. https://doi.org/10.1007/978-3-642-37184-4

Büning H (1991) Robuste und adaptive Tests. de Gruyter, Berlin

Büning H, Trenkler G (1994) Nichtparametrische statistische Methoden, 2nd edn. de Gruyter, Berlin

D'Agostino RB, Stephens MA (eds) (1986) Goodness-of-fit techniques. Statistics: textbooks and monographs, vol 68. Marcel Dekker, New York, NY

Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Duller C (2008) Einführung in die nichtparametrische Statistik mit SAS und R. Ein anwendungsorientiertes Lehr- und Arbeitsbuch. Physica, Heidelberg

Edgington ES, Onghena P (2007) Randomization tests. With CD-ROM, 4th edn. Chapman & Hall/CRC, Boca Raton, FL

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Monographs on statistics and applied probability, vol 57. Chapman & Hall, New York, NY

Good P (2005) Permutation, parametric and bootstrap tests of hypotheses, 3rd edn. Springer, New York, NY. https://doi.org/10.1007/b138696

Good PI (2006) Resampling methods. A practical guide to data analysis, 3rd edn. Birkhäuser, Boston, MA

Hájek J, Šidák Z, Sen PK (1999) Theory of rank tests, 2nd edn. Elsevier/Academic Press, Orlando, FL

Hall P (1988) Theoretical comparison of bootstrap confidence intervals. Ann Stat 16(3):927–953

Hall P (1992) The bootstrap and Edgeworth expansion. Springer series in statistics. Springer, New York, NY

Hollander M, Wolfe DA, Chicken E (2014) Nonparametric statistical methods, 3rd updated edn. John Wiley & Sons, Hoboken, NJ

Klenke A (2008) Probability theory (Wahrscheinlichkeitstheorie), 2nd revised edn. Springer, Berlin

Krishnaiah P, Sen P (eds) (1984) Nonparametric methods. Handbook of statistics, vol 4. North-Holland, Amsterdam

Küchler U (2016) Maßtheorie für Statistiker. Grundlagen der Stochastik. Springer Spektrum, Heidelberg. https://doi.org/10.1007/978-3-662-46375-8

Lahiri S (2003) Resampling methods for dependent data. Springer, New York, NY

Lehmann EL (2006) Nonparametrics. Statistical methods based on ranks. With the special assistance of H. J. M. D'Abrera, 1st revised edn. Springer, New York, NY

LePage R, Billard L (eds) (1992) Exploring the limits of bootstrap: papers presented at a special topics meeting, East Lansing, UK, May 1990. Wiley, New York, NY

Mammen E (1992) When does bootstrap work? Asymptotic results and simulations. Springer, New York, NY
Neuhäuser M (2012) Nonparametric statistical tests. A computational approach. CRC Press, Boca Raton, FL
Owen AB (2001) Empirical likelihood. Chapman & Hall/CRC, Boca Raton, FL
Pardo L (2006) Statistical inference based on divergence measures. Chapman & Hall/CRC, Boca Raton, FL
Pesarin F, Salmaso L (2010) Permutation tests for complex data: theory, applications and software. Wiley series in probability and statistics. John Wiley & Sons, Chichester
Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer, New York, NY
Puri M (ed) (1970) Nonparametric techniques in statistical inference. In: Proceedings of the first international symposium on nonparametric techniques held at Indiana University, June 1969. University Press, Cambridge
Student (1908) The probable error of a mean. Biometrika 6:1–25
Wilcox R (2012) Introduction to robust estimation and hypothesis testing, 3rd edn. Elsevier/Academic Press, Amsterdam
Witting H (1985) Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang. B. G. Teubner, Stuttgart

Chapter 2
Empirical Measures, Empirical Processes

In this preparatory chapter, we gather some results from the theory of empirical measures and empirical processes.

Throughout the chapter we assume an independent and identically distributed sample $Y = (Y_1, \ldots, Y_n)^\top$ which is defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. As in Chap. 1, we denote the marginal distribution of Y_1 by $P = \mathbb{P}^{Y_1}$ and notice that the joint distribution of Y is given by $\mathbb{P}^Y = P^{\otimes n}$. Hence, $\mathbb{P}^Y$ is already identified by P.

2.1 Properties of Empirical Measures

Assume that Y_1 is real-valued. Let $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$ denote the empirical measure pertaining to the sample Y. Notice that, for any Borel set A of $\mathbb{R}$, we have

$$\hat{P}_n(A) = n^{-1} \sum_{i=1}^n \mathbf{1}_A(Y_i). \tag{2.1}$$

More generally, we get for any measurable function $g : \mathbb{R} \to \mathbb{R}$ that

$$\hat{E}_n[g] = \mathbb{E}_{\hat{P}_n}[g] = \int_{\mathbb{R}} g(y) \hat{P}_n(dy) = n^{-1} \sum_{i=1}^n g(Y_i), \tag{2.2}$$

meaning that taking the expectation of g with respect to $\hat{P}_n$ is equivalent to averaging the values of g over the observational units. Of course, (2.1) is a special case of (2.2), where we take $g = \mathbf{1}_A$. If we restrict our attention further to Borel sets of the form $A = (-\infty, y]$, $y \in \mathbb{R}$, we obtain the empirical cumulative distribution function

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_2

(ecdf) $\hat{F}_n$, given by

$$\begin{aligned}\hat{F}_n(y) &= \hat{P}_n((-\infty, y]) = \hat{E}_n\left[\mathbf{1}_{(-\infty,y]}\right] \\ &= n^{-1}\sum_{i=1}^{n}\mathbf{1}_{(-\infty,y]}(Y_i) = \frac{|\{1 \le i \le n : Y_i \le y\}|}{n}. \end{aligned} \tag{2.3}$$

Notice that the right-hand sides of (2.1)–(2.3) are random variables. Interpreting these random variables as estimators, we will now see that they possess desirable estimation properties, such as unbiasedness, consistency, and asymptotic normality.

Theorem 2.1 *For any Borel set A of $\mathbb{R}$, the following assertions hold true.*

(a) $\mathbb{E}[\hat{P}_n(A)] = P(A)$, *where* $\mathbb{E}$ *refers to* $\mathbb{P}$.
(b) $Var\left(\hat{P}_n(A)\right) = n^{-1}\sigma_A^2$, *where* $\sigma_A^2 = P(A)[1 - P(A)]$.
(c) For all $y \in \mathbb{R}$, $n\hat{F}_n(y) \sim Bin(n, F(y))$, *where* F *is the cdf pertaining to* P, *i.e.,* $F(y) = \mathbb{P}(Y_1 \le y)$.
(d) $\hat{P}_n(A) \to P(A)$ $\mathbb{P}$*-almost surely as* $n \to \infty$.
(e) $\sqrt{n}\left\{\hat{P}_n(A) - P(A)\right\} \to \mathcal{N}(0, \sigma_A^2)$ *in distribution as* $n \to \infty$.

Proof This is Exercise 2.1.

The results of Theorem 2.1 can be extended to certain functionals of the distribution P.

Theorem 2.2 *Let $g : \mathbb{R} \to \mathbb{R}$ be a function on the real line such that*

$$\int_{-\infty}^{\infty} g^2(y)P(dy) < \infty. \tag{2.4}$$

Consider the functional

$$\kappa : \{Q : Q \text{ is a probability distribution on } \mathbb{R}\} \to \mathbb{R},$$

which is given by $\kappa(Q) = \mathbb{E}_Q[g]$. *Denote* $\kappa_0 = \kappa(P) = \mathbb{E}[g(Y_1)] = \int_{\mathbb{R}} g(y)P(dy)$ *and consider the (plug-in) estimator* $\hat{\kappa}_n = \kappa(\hat{P}_n) = n^{-1}\sum_{i=1}^{n} g(Y_i)$.

Then the following assertions hold true.

(a) For all $n \in \mathbb{N}$, $\mathbb{E}[\hat{\kappa}_n] = \kappa_0$.
(b) $Var(\hat{\kappa}_n) = \sigma_\kappa^2/n$, *where*

$$\sigma_\kappa^2 = \int_{\mathbb{R}} g^2(y)P(dy) - \kappa_0^2 = \int_{\mathbb{R}} [g(y) - \kappa_0]^2\, P(dy).$$

(c) $\hat{\kappa}_n \to \kappa_0$ $\mathbb{P}$*-almost surely as* $n \to \infty$.
(d) Assuming that $\sigma_\kappa^2 > 0$, $\sqrt{n}\left(\hat{\kappa}_n - \kappa_0\right) \to \mathcal{N}(0, \sigma_\kappa^2)$ *in distribution as* $n \to \infty$.

(e) Let $h : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function on the real line, and assume that $h'(\kappa_0) \neq 0$. Then, as $n \to \infty$,

$$h(\hat{\kappa}_n) \to h(\kappa_0) \ \mathbb{P}\text{-almost surely}, \tag{2.5}$$

$$\sqrt{n}\left\{h(\hat{\kappa}_n) - h(\kappa_0)\right\} \to \mathcal{N}(0, \sigma_h^2) \text{ in distribution}, \tag{2.6}$$

where $\sigma_h^2 = |h'(\kappa_0)|^2 \sigma_\kappa^2$.

Proof Introduce, for notational convenience, the new random variables $\xi_i = g(Y_i)$, $1 \leq i \leq n$. We observe that $(\xi_i : 1 \leq i \leq n)$ are i.i.d. with $\mathbb{E}[\xi_1] = \kappa_0$ and $\mathrm{Var}(\xi_1) = \sigma_\kappa^2$. This immediately entails parts (a) and (b) by linearity of expectation operators. Furthermore, part (c) follows by applying the strong law of large numbers to $(\xi_i : 1 \leq i \leq n)$. In order to prove part (d), notice that condition (2.4) ensures finiteness of σ_κ^2. Hence, we can apply the central limit theorem to $(\xi_i : 1 \leq i \leq n)$, and the result follows. Finally, assertion (2.5) is a consequence of the Continuous Mapping Theorem (see, e.g., Theorem 1.14 in DasGupta 2008), and (2.6) is a consequence of the Delta method (see, e.g., Theorem 3.6 of DasGupta 2008).

Remark 2.3 The results of Theorem 2.2 can be extended to the case of a vector-valued function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^p$, that is, $\mathbf{g}(y) = \big(g_1(y), \ldots, g_p(y)\big)^\top$ for $y \in \mathbb{R}$; see Theorem 2.1.4 of Spokoiny and Dickhaus (2015).

Let us conclude this section with an important statistical property of the empirical cdf $\hat{F}_n$. Namely, Theorem 2.5 shows that choosing $\hat{F}_n$ as the estimator of F actually follows from the statistical principle of likelihood maximization.

Definition 2.4 (Nonparametric Likelihood Function) Let $Y_1, \ldots, Y_n$ be real-valued i.i.d. random variables, $y = (y_1, \ldots, y_n)^\top$ a realization of $(Y_1, \ldots, Y_n)^\top$, and $\mathscr{F}$ the set of all cdfs on $\mathbb{R}$.

Then we call $Z : \mathbb{R}^n \times \mathscr{F} \to [0, 1]$, given by

$$\begin{aligned}(y, F) \mapsto Z(y, F) &= \prod_{i=1}^{n} [F(y_i) - F(y_i-)] \\ &= \prod_{i=1}^{n} \mathbb{P}_F(\{y_i\}),\end{aligned}$$

where $\mathbb{P}_F$ denotes the probability measure on $\mathbb{R}$ induced by F, the *nonparametric likelihood function* for the data y.

Theorem 2.5 *Under the i.i.d. model for $Y = (Y_1, \ldots, Y_n)^\top$, the empirical cdf $\hat{F}_n$ is the nonparametric maximum likelihood estimator (NPMLE) of the cdf F of Y_1, i.e., $\hat{F}_n$ maximizes $Z(y, \cdot)$ over $\mathscr{F}$, for every data point $Y = y = (y_1, \ldots, y_n)^\top$.*

Proof It is immediately clear that the NPMLE $\hat{F}_{\mathrm{NPMLE}}$ is necessarily such that $\mathbb{P}_{\hat{F}_{\mathrm{NPMLE}}}$ is a discrete probability measure which puts its entire mass in the

observation points $y_1, \ldots, y_n$. In other words, the NPMLE is such that $\mathbb{P}_{\hat{F}_{\text{NPMLE}}}$ is absolutely continuous with respect to the empirical measure $\hat{P}_n$. Thus, the NPMLE is characterized by the n-tuple $(p_i : 1 \leq i \leq n)$, where $p_i = \mathbb{P}_{\hat{F}_{\text{NPMLE}}}(\{y_i\})$, such that $p_i \geq 0$ for all $1 \leq i \leq n$ and $\sum_{i=1}^n p_i = 1$. This shows that the optimization of $Z(y, \cdot)$ over $\mathscr{F}$ is actually equivalent to a finite-dimensional optimization problem over the unit simplex in $\mathbb{R}^n$ regarding $(p_i : 1 \leq i \leq n)$, although $\mathscr{F}$ is an infinite-dimensional function space. Hence, we have to solve the following constrained optimization problem.

$$\max_{p_1,\ldots,p_n} \prod_{i=1}^n p_i \text{ subject to } p_i \geq 0 \,\forall 1 \leq i \leq n, \sum_{i=1}^n p_i = 1. \tag{2.7}$$

The solution of (2.7) can easily be obtained by using the Lagrange multiplier method, and it is given by $p_i = 1/n$ for all $1 \leq i \leq n$; cf. Exercise 2.2. This completes the proof.

2.2 The Principle of Quantile Transformation

This section follows Section 1.1 of Shorack and Wellner (1986).

Definition 2.6 (Generalized Inverse, Quantile Function) Let F be any cdf on $\mathbb{R}$. Then we call the function

$$\begin{aligned} F^{-1} : [0, 1] &\to \bar{\mathbb{R}} \\ t &\mapsto F^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\} \end{aligned}$$

the (left-continuous) *generalized inverse* of F or the *quantile function* corresponding to F, respectively.

If F is strictly isotone and continuous, then F^{-1} coincides with the (ordinary) inverse of F.

Theorem 2.7 (Quantile Transformation) *Let $U \sim UNI[0, 1]$ and F be any cdf on $\mathbb{R}$. Then, the random variable $X := F^{-1}(U)$ possesses the cdf F. We will use the notation: $X = F^{-1}(U) \sim F$.*

Furthermore, even the following stronger assertions hold true for all $x \in \mathbb{R}$.

$$\{X \leq x\} = \{U \leq F(x)\}, \tag{2.8}$$

$$\mathbf{1}_{\{X \leq x\}} = \mathbf{1}_{\{U \leq F(x)\}}. \tag{2.9}$$

Proof The definition of F^{-1} implies that

$$U \leq F(x) \Rightarrow X = F^{-1}(U) = \inf\{z | F(z) \geq U\} \leq x.$$

On the other hand, we can conclude from $X = F^{-1}(U) \leq x$ that $\forall \varepsilon > 0 : F(x + \varepsilon) \geq U$. Exploiting the right-continuity of F and letting $\varepsilon \to 0$, it follows that $F(x) \geq U$. This proves (2.8), which immediately implies (2.9).

The assertion about the distribution of X follows by noticing that

$$\mathbb{P}(X \leq x) \underset{(2.8)}{=} \mathbb{P}(U \leq F(x)) \underset{U \sim \text{UNI}[0,1]}{=} F(x).$$

Remark 2.8 We get from (2.8) that the following assertions hold true for any cdf F on $\mathbb{R} \ni x$ and for any $t \in]0, 1[$.

(a) $F(x) \geq t \Leftrightarrow F^{-1}(t) \leq x$.
(b) $F(x) < t \Leftrightarrow F^{-1}(t) > x$.
(c) $F(x_1) < t \leq F(x_2) \Leftrightarrow x_1 < F^{-1}(t) \leq x_2$.

Theorem 2.9 *For any cdf F on $\mathbb{R}$, it holds that*

$$\forall 0 \leq t \leq 1 : \ (F \circ F^{-1})(t) \geq t. \tag{2.10}$$

Equality holds in (2.10), if and only if t belongs to the range of F on $\bar{\mathbb{R}}$, meaning that there exists an argument $x \in \bar{\mathbb{R}}$ such that $F(x) = t$. If F is continuous, then we have that $F \circ F^{-1} = id_{[0,1]}$.

Proof Considering the special case $x = F^{-1}(t)$ in part (a) of Remark 2.8, we get that $F(F^{-1}(t)) \geq t \Leftrightarrow F^{-1}(t) \leq F^{-1}(t)$, and the latter statement is always true.

Now, assume that t belongs to the range of F on $\bar{\mathbb{R}}$. It follows that

$$F^{-1}(t) = \inf\{x | F(x) \geq t\} = \inf\{x | F(x) = t\} \Rightarrow F(F^{-1}(t)) = t.$$

On the other hand, if t does not belong to the range of F on $\bar{\mathbb{R}}$, we obtain that

$$F^{-1}(t) = \inf\{x | F(x) \geq t\} = \inf\{x | F(x) > t\} \Rightarrow F(F^{-1}(t)) > t.$$

Theorem 2.10 (Probability Integral Transformation) *Let X be a real-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with cdf F. Then it holds that*

$$\mathbb{P}(F(X) \leq t) \leq t \ \text{ for all } \ 0 \leq t \leq 1. \tag{2.11}$$

Equality in (2.11) holds, if and only if t belongs to the closure of the range of F. Thus, if F is continuous, it holds that $U := F(X) \sim UNI[0, 1]$.

Proof Assume that t belongs to the closure of the range of F. Then, Theorem 2.9 yields that

$$\mathbb{P}(F(X) \leq t) = \mathbb{P}(X \leq F^{-1}(t)) = F(F^{-1}(t)) = t.$$

On the other hand, assume now that t does not belong to the closure of the range of F. Then we can choose $\varepsilon > 0$ such that $t - \varepsilon$ belongs to the closure of the range of F, but $t - \varepsilon + \delta$ does not, for every $\delta > 0$. We obtain that

$$\begin{aligned}\mathbb{P}(F(X) \leq t) &= \mathbb{P}(F(X) \leq t - \varepsilon) \\ &= \mathbb{P}(X \leq F^{-1}(t - \varepsilon)) \\ &= F(F^{-1}(t - \varepsilon)) = t - \varepsilon < t.\end{aligned}$$

Theorem 2.11 *Let F be any cdf on $\mathbb{R}$. Then it holds that*

$$\forall x \in \mathbb{R} : \ (F^{-1} \circ F)(x) \leq x,$$

with equality failing if and only if there exists an $\varepsilon > 0$ with $F(x - \varepsilon) = F(x)$.

It follows that $\mathbb{P}((F^{-1} \circ F)(X) \neq X) = 0$, where $X \sim F$.

Proof This is Exercise 2.3.

Corollary 2.12

(i) A cdf F on $\mathbb{R}$ is continuous, if and only if F^{-1} is strictly isotone.
(ii) A cdf F on $\mathbb{R}$ is strictly isotone, if and only if F^{-1} is continuous.

Theorem 2.13 *Let F be a continuous cdf on $\mathbb{R}$ and assume that $U := F(X) \sim UNI[0, 1]$, where X is some real-valued random variable. Then we can conclude, that $X \sim F$.*

Proof For any cdf F, it holds that $\{X \leq x\} \subseteq \{F(X) \leq F(x)\}$, because F is isotone, but not necessarily strictly isotone. Hence, we have that

$$\mathbb{P}(X \leq x) \leq \mathbb{P}(F(X) \leq F(x)) = \mathbb{P}(U \leq F(x)) = F(x)$$

by our assumptions.

On the other hand, it holds that

$$\begin{aligned}F(x) &= \mathbb{P}(U \leq F(x)) \\ &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \mathbb{P}((F^{-1} \circ F)(X) \leq x) = \mathbb{P}(X \leq x),\end{aligned}$$

unless there exists an $\varepsilon > 0$ with $F(x - \varepsilon) = F(x)$; cf. Theorem 2.11.

Since F is continuous, we conclude that $x \mapsto \mathbb{P}(X \leq x)$ coincides with F.

2.3 Some Results from the Theory of Empirical Processes

Theorem 2.14 (Glivenko-Cantelli) *It holds that*

$$\sup_{y\in\mathbb{R}}\left|\hat{F}_n(y)-F(y)\right|\to 0\ \mathbb{P}\text{-almost surely as } n\to\infty. \tag{2.12}$$

Before we prove Theorem 2.14, recall the following important properties of monotonic functions.

Lemma 2.15 (Theorem 8.19 in Hewitt and Stromberg 1975) *Let F be a real-valued, non-decreasing function defined on $\mathbb{R}$. Then, F has finite right- and left-hand limits in all points of $\mathbb{R}$, and F is continuous except at a countable set of points of $\mathbb{R}$.*

Proof We now prove Theorem 2.14. In this, we follow the argumentation of Einmahl and de Haan (1999–2000), pp. 2–3. Due to Lemma 2.15, it suffices to show that

$$\forall m\in\mathbb{N}:\limsup_{n\to\infty}\sup_{y\in\mathbb{R}}\left|\hat{F}_n(y)-F(y)\right|\le\frac{1}{m+1}\ \mathbb{P}\text{-almost surely.} \tag{2.13}$$

To this end, choose for fixed $m\in\mathbb{N}$ a grid

$$-\infty=y_{0,m}<y_{1,m}\le\ldots\le y_{m,m}<y_{m+1,m}=+\infty$$

such that

$$\forall 1\le k\le m: F(y_{k,m}-)\le\frac{k}{m+1}\le F(y_{k,m}). \tag{2.14}$$

Now, observe the basic fact that for all $y\in\mathbb{R}$ we have $\left|\hat{F}_n(y)-F(y)\right|=\max\{\hat{F}_n(y)-F(y),F(y)-\hat{F}_n(y)\}$. We analyze both values separately.

For any $1\le k\le m+1$ and any $y\in[y_{k-1,m},y_{k,m})$, we have that

$$\begin{aligned}\hat{F}_n(y)-F(y)&\le\hat{F}_n(y_{k,m}-)-F(y_{k-1,m})\\&=\hat{F}_n(y_{k,m}-)-F(y_{k,m}-)+F(y_{k,m}-)-F(y_{k-1,m})\\&\le\left|\hat{F}_n(y_{k,m}-)-F(y_{k,m}-)\right|+\frac{1}{m+1}\end{aligned}$$

due to (2.14). Analogously, we get

$$F(y)-\hat{F}_n(y)\le\left|\hat{F}_n(y_{k-1,m})-F(y_{k-1,m})\right|+\frac{1}{m+1}.$$

Because of $\hat{F}_n(-\infty) - F(-\infty) = \hat{F}_n(\infty) - F(\infty) = 0$ this yields

$$\sup_{y\in\mathbb{R}} \left|\hat{F}_n(y) - F(y)\right| \le \max_{1\le k\le m} \left|\hat{F}_n(y_{k,m}-) - F(y_{k,m}-)\right| + \max_{1\le k\le m} \left|\hat{F}_n(y_{k,m}) - F(y_{k,m})\right| + \frac{1}{m+1}. \quad (2.15)$$

From Theorem 2.1.(d), we have point-wise almost sure convergence, i.e.,

$$\hat{F}_n(y_{k,m}-) \to F(y_{k,m}-) \text{ and } \hat{F}_n(y_{k,m}) \to F(y_{k,m})$$

$\mathbb{P}$-almost surely as $n \to \infty$. Since the maximization in (2.15) involves only a finite number of random variables, assertion (2.13) follows.

Definition 2.16 ((Reduced) Empirical Process) Under the general assumptions of this chapter, we call the random function $\sqrt{n}\left(\hat{F}_n - F\right)$ the *empirical process* pertaining to $Y_1, \ldots, Y_n$. In the special case that $Y_1 \sim \text{UNI}[0,1]$, we write $U_1, \ldots, U_n$ instead of $Y_1, \ldots, Y_n$. In the latter case, we furthermore define for $0 \le u \le 1$ the quantities

$$\hat{G}_n(u) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{U_i \le u\}},$$

$$\mathbb{U}_n(u) = \sqrt{n}\left(\hat{G}_n(u) - u\right),$$

and call the random function $\mathbb{U}_n(\cdot)$ a *reduced empirical process*.

Lemma 2.17 *Let $n \in \mathbb{N}$, $U_1, \ldots, U_n$ i.i.d. with $U_1 \sim UNI[0,1]$, and define $Y_i = F^{-1}(U_i)$, for a given cdf F on $\mathbb{R}$. Then, the following assertions hold true.*

(a) $\forall y \in \mathbb{R} : \hat{F}_n(y) = \hat{G}_n(F(y))$.

(b) $\forall y \in \mathbb{R} : \sqrt{n}\left(\hat{F}_n(y) - F(y)\right) = \mathbb{U}_n(F(y))$.

(c) If F is continuous, then $\mathbb{U}_n(u) = \sqrt{n}\left(\hat{F}_n(F^{-1}(u)) - u\right)$ for all $u \in [0,1]$.

Proof To prove part (a), we straightforwardly calculate

$$\hat{G}_n(F(y)) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{U_i \le F(y)\}} = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \le y\}} = \hat{F}_n(y), \quad (2.16)$$

where the left-hand side of (2.16) follows from (2.9). Part (b) is an immediate consequence of part (a). For proving part (c), we substitute $y = F^{-1}(u)$ in part (b) and recall from Theorem 2.9 that $F \circ F^{-1} = \mathrm{id}_{[0,1]}$ for continuous F.

Remark 2.18

(a) The function F, although typically unknown in a statistical context, is a fixed, deterministic function. Hence, part (b) of Lemma 2.17 shows that the stochastic behavior of a (general) empirical process is already determined by that of the reduced empirical process.
(b) If one is only interested in probabilistic statements about some i.i.d. random variables $Y_1, \ldots, Y_n$ with cdf F of Y_1, then Theorem 2.7 (quantile transformation) yields that $Y_i = F^{-1}(U_i)$ for $1 \leq i \leq n$ can be assumed without loss of generality.

Definition 2.19

(a) A *stochastic process* is an indexed collection $(X_t)_{t \in \mathcal{T}}$ of random variables, where $\mathcal{T}$ denotes an (arbitrary) index set.
(b) A real-valued stochastic process $(X_t)_{t \in \mathcal{T}}$ is called a *Gaussian process*, if for any $m \in \mathbb{N}$ and any m-tuple $(t_1, \ldots, t_m) \subseteq \mathcal{T}$ the random vector $(X_{t_1}, \ldots, X_{t_m})^\top$ is (jointly) normally distributed.
(c) A real-valued stochastic process $(B_t)_{t \geq 0}$, which is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is called a *Brownian motion*, if the following conditions are fulfilled.
    (i) The process $(B_t)_{t \geq 0}$ is a Gaussian process with $\mathbb{E}[B_t] = 0$ for all $t \geq 0$ and $\mathrm{Cov}(B_s, B_t) = s \wedge t$ for all $s, t \geq 0$.
    (ii) The path $t \mapsto B_t(\omega)$ is continuous for $\mathbb{P}$-almost all $\omega \in \Omega$.
(d) Assume that $(B_t)_{0 \leq t \leq 1}$ is a Brownian motion. Then, the process $(B_t^0)_{0 \leq t \leq 1}$, defined by $B_t^0 = B_t - tB_1$, $0 \leq t \leq 1$, is called a *Brownian bridge*.

Remark 2.20

(a) The existence of Brownian motion has first been shown by Wiener (1923).
(b) If $(B_t)_{t \geq 0}$ is a Brownian motion, then $B_0 = 0$ almost surely, because $\mathbb{E}[B_0] = \mathrm{Var}(B_0) = 0$.

Lemma 2.21 *A Brownian bridge $(B_t^0)_{0 \leq t \leq 1}$ is a centered Gaussian process with $Cov(B_s^0, B_t^0) = s \wedge t - st$ for all $0 \leq s, t \leq 1$. In particular, $B_0^0 = B_1^0 = 0$ almost surely.*

Proof Gaussianity of $(B_t^0)_{0 \leq t \leq 1}$ holds due to linearity of normal distributions. Since B_t is centered for all $0 \leq t \leq 1$, so is B_t^0. The covariance structure of $(B_t^0)_{0 \leq t \leq 1}$ is calculated as follows.

$$\begin{aligned}
\mathrm{Cov}(B_s^0, B_t^0) &= \mathrm{Cov}(B_s - sB_1, B_t - tB_1) \\
&= \mathrm{Cov}(B_s, B_t) - s\mathrm{Cov}(B_1, B_t) - t\mathrm{Cov}(B_s, B_1) + st\mathrm{Var}(B_1) \\
&= s \wedge t - st - st + st = s \wedge t - st, \quad 0 \leq s, t \leq 1.
\end{aligned}$$

Theorem 2.22 *All finite-dimensional marginal distributions pertaining to the reduced empirical process* $(\mathbb{U}_n(u))_{0\le u\le 1}$ *converge weakly to the corresponding marginal distributions of a Brownian bridge* $(B^0_u)_{0\le u\le 1}$ *as* $n\to\infty$.

Proof Let $m\in\mathbb{N}$ and $0\le u_1\le u_2\le\ldots\le u_m\le 1$ be arbitrary, but fixed. We notice that

$$(\mathbb{U}_n(u_1),\ldots,\mathbb{U}_n(u_m))^\top = n^{-\frac{1}{2}}\sum_{i=1}^n \left(\mathbf{1}_{\{U_i\le u_1\}}-u_1,\ldots,\mathbf{1}_{\{U_i\le u_m\}}-u_m\right)^\top$$

is a normalized sum of i.i.d. centered m-dimensional random vectors with finite covariance matrix Σ (say). Hence, the multivariate central limit theorem for i.i.d. random vectors yields that $(\mathbb{U}_n(u_1),\ldots,\mathbb{U}_n(u_m))^\top \xrightarrow{\mathscr{D}} Z$, where Z is a centered m-dimensional Gaussian random vector with the same covariance matrix Σ. It remains to calculate Σ. This is done in Exercise 2.5. We obtain that $\Sigma_{j,k}=u_j-u_ju_k$, $1\le j\le k\le m$. Hence, the covariance structure of Z is the same as that of the Brownian bridge, completing the proof.

Corollary 2.23 (Donsker 1952) *The reduced empirical process* $(\mathbb{U}_n(u))_{0\le u\le 1}$ *converges in distribution to* $(B^0_u)_{0\le u\le 1}$ *as* $n\to\infty$.

2.4 Exercises

Exercise 2.1 *Prove Theorem 2.1.*

Exercise 2.2 *Solve the constrained optimization problem (2.7) using the Lagrange multiplier method.*
Hints:

(i) It suffices to take only the equality constraint of (2.7) into account.
(ii) It may be easier to work with $L(y,F)=\log(Z(y,F))$ *instead of* $Z(y,F)$. *Since the logarithm is a strictly increasing function, the (constrained) maximizers of* $L(y,\cdot)$ *and* $Z(y,\cdot)$ *coincide.*

Exercise 2.3

(a) Prove Theorem 2.11.
(b) Illustrate the assertion of Corollary 2.12 graphically.
Hint: *Choose* F *such that* $\mathrm{supp}(F)=[0,1]$ *and that* F *possesses on* $[0,1]$ *exactly one jump and exactly one "plateau."*
(c) Elucidate the essential assertions of Sect. 2.2 with the help of your graph from part (b) of this exercise.

Exercise 2.4 (Programming Exercise)

(a) Write an R *program which simulates and plots (equidistantly discretized) paths of a Brownian motion on* $[0, 1]$. *To this end, show first that the increments of a Brownian motion are normally distributed and stochastically independent.*
(b) Add to your graph from part (a) of this exercise the path of the corresponding Brownian bridge.

Exercise 2.5 *Calculate the covariance matrix* Σ *appearing in the proof of Theorem 2.22.*

Exercise 2.6 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) If, under the general assumptions of Sect. 2.1, the variance $\sigma^2 := Var(Y_1)$ *is finite, then the plug-in estimator for* σ^2 *is unbiased.*
(b) Assume that the random variable Z *is standard normally distributed on* $\mathbb{R}$, *and let* $\lambda > 0$ *be a given real constant. Then, the random variable* $Y := -\lambda^{-1} \log(\Phi(Z))$ *is exponentially distributed with intensity parameter* λ, *where* Φ *denotes the cdf of* $\mathcal{N}(0, 1)$.
(c) Under the assumptions of Theorem 2.14, the supremum in Eq. (2.12) is always attained at a point of discontinuity of $\hat{F}_n$ *or at a point of discontinuity of* F.
(d) The distribution of a Gaussian process $(X_t)_{t \in \mathcal{T}}$ *is uniquely determined by the specification of a mean function* $\mu : \mathcal{T} \to \mathbb{R}$ *with* $\mu(t) = \mathbb{E}[X_t]$ *and the specification of a covariance function* $\Sigma : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ *with* $\Sigma(s, t) = Cov(X_s, X_t)$.

References

DasGupta A (2008) Asymptotic theory of statistics and probability. Springer texts in statistics. Springer, New York, NY

Donsker MD (1952) Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. Ann Math Stat 23:277–281. https://doi.org/10.1214/aoms/1177729445

Einmahl JHJ, de Haan L (1999–2000) Empirical processes and statistics of extreme values. Lecture notes, AIO course

Hewitt E, Stromberg K (1975) Real and abstract analysis. A modern treatment of the theory of functions of a real variable, 3rd printing. Graduate texts in mathematics, vol 25. Springer, New York, NY

Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley series in probability and mathematical statistics. Wiley, New York, NY

Spokoiny V, Dickhaus T (2015) Basics of modern mathematical statistics. Springer, Heidelberg

Wiener N (1923) Differential-space. J Math Phys 2:131–174

Chapter 3
Goodness-of-Fit Tests

As outlined in Sect. 1.3, we will construct goodness-of-fit tests by making use of the substitution principle. Let $\mathscr{P} = \{Q : Q \text{ is a probability measure on } (\mathbb{R}, \mathscr{B}(\mathbb{R}))\}$. We assume a sample $Y = (Y_1, \ldots, Y_n)^\top$ of i.i.d. real-valued random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, such that $P = \mathbb{P}^{Y_1} \in \mathscr{P}$. Clearly, P already determines $\mathbb{P}^Y = P^{\otimes n}$.

Our aim is to test certain null hypotheses H_0 corresponding to subsets $\mathscr{P}_0 \subset \mathscr{P}$. To this end, assume that we have some distance ρ defined on $\mathscr{P} \times \mathscr{P}$. Then, we can equivalently express the null hypothesis as $H_0 : \inf_{Q \in \mathscr{P}_0} \rho(P, Q) = 0$, cf. (1.5). Hence, the substitution principle leads to the test statistic $T_n = \inf_{Q \in \mathscr{P}_0} \rho(\hat{P}_n, Q)$, where $\hat{P}_n$ denotes the empirical measure pertaining to $Y_1, \ldots, Y_n$; cf. (1.6).

In order to implement this idea into practical test procedures, we need to

(1) define appropriate distances on $\mathscr{P} \times \mathscr{P}$, and
(2) derive or approximate the null distribution of the resulting test statistics T_n.

In this chapter, we will cover two types of null hypotheses, namely

(1) simple null hypotheses of the form $\mathscr{P}_0 = \{P_0\}$, where P_0 is completely specified, and
(2) (composite) null hypotheses consisting of a parametric family of the form $\mathscr{P}_0 = \{P_\vartheta : \vartheta \in \Theta \subseteq \mathbb{R}^p, p \in \mathbb{N}\}$.

Other types of (composite) null hypotheses, relating to statistical functionals, will be treated in Chap. 7.

Definition 3.1 (See Section 14.2 in Deza and Deza 2016) Let Q_1 and $Q_2 \in \mathscr{P}$ with corresponding cdfs F_1 and F_2. Then we call

(i) ρ_{KS}, given by

$$\rho_{\mathrm{KS}}(Q_1, Q_2) = \sup_{y \in \mathbb{R}} |F_1(y) - F_2(y)|,$$

the *Kolmogorov-Smirnov metric* (or: uniform metric) on $\mathscr{P} \times \mathscr{P}$.

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_3

(ii) ρ_{CvM}, given by

$$\rho_{\mathrm{CvM}}(Q_1, Q_2) = \int_{\mathbb{R}} [F_1(y) - F_2(y)]^2 \, dF_2(y),$$

the *Cramér-von Mises distance* on $\mathscr{P} \times \mathscr{P}$.

We will also write $\rho(F_1, F_2)$ instead of $\rho(Q_1, Q_2)$.

3.1 Simple Null Hypotheses

Theorem 3.2 *Let $Y_1, \ldots, Y_n$ be i.i.d. real-valued random variables with continuous cdf F of Y_1, $\hat{F}_n$ the empirical cdf pertaining to $Y_1, \ldots, Y_n$, and let $\mathbb{U}_n$ be the reduced empirical process defined in Definition 2.16, where $U_i := F(Y_i)$ for all $1 \le i \le n$. Then it holds:*

$$\sqrt{n} \cdot \rho_{KS}(\hat{F}_n, F) = \sup_{0 \le u \le 1} |\mathbb{U}_n(u)|, \tag{3.1}$$

$$n \cdot \rho_{CvM}(\hat{F}_n, F) = \int_0^1 \mathbb{U}_n^2(u) du. \tag{3.2}$$

Proof To prove (3.1), recall that

$$\sqrt{n} \cdot \rho_{\mathrm{KS}}(\hat{F}_n, F) = \sup_{y \in \mathbb{R}} \left| \sqrt{n} \{ \hat{F}_n(y) - F(y) \} \right|. \tag{3.3}$$

Now, substitute $y = F^{-1}(u)$ in (3.3). We obtain that

$$\begin{aligned} \sqrt{n} \cdot \rho_{\mathrm{KS}}(\hat{F}_n, F) &= \sup_{0 \le u \le 1} \left| \sqrt{n} \{ \hat{F}_n(F^{-1}(u)) - F(F^{-1}(u)) \} \right| \\ &= \sup_{0 \le u \le 1} \left| \sqrt{n} \{ \hat{F}_n(F^{-1}(u)) - u \} \right| \\ &= \sup_{0 \le u \le 1} |\mathbb{U}_n(u)|, \end{aligned}$$

due to Lemma 2.17.(c).

Assertion (3.2) can be proved analogously; cf. Exercise 3.1.

Corollary 3.3 *Under the assumptions of Theorem 3.2, it holds that*

$$\sqrt{n} \cdot \rho_{KS}(\hat{F}_n, F) \xrightarrow{\mathscr{D}} \sup_{0 \le t \le 1} |B_t^0|, \tag{3.4}$$

$$n \cdot \rho_{CvM}(\hat{F}_n, F) \xrightarrow{\mathscr{D}} \int_0^1 \left[B_t^0 \right]^2 dt =: W^2, \tag{3.5}$$

where $(B_t^0)_{0\le t\le 1}$ *denotes a Brownian bridge.*

Proof The assertions follow from Theorem 3.2 by applying the continuous mapping theorem.

Theorem 3.4 (Kolmogorov Distribution) *Let* $(B_t^0)_{0\le t\le 1}$ *denote a Brownian bridge defined on some probability space* $(\Omega, \mathscr{F}, \mathbb{P})$*. Then it holds that*

$$\forall x > 0 : \mathbb{P}\left(\sup_{0\le t\le 1} |B_t^0| \le x\right) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\left(-2k^2x^2\right) =: L(x). \tag{3.6}$$

Proof See Kolmogorov (1933).

Application 3.5 (Kolmogorov-Smirnov Test) *Let* $Y_1, \ldots, Y_n$ *denote i.i.d. real-valued random variables defined on some probability space* $(\Omega, \mathscr{F}, \mathbb{P})$ *with unknown cdf* F *of* Y_1*. Consider the test problem*

$$H_0 : \{F = F_0\} \text{ versus } H_1 : \{F \ne F_0\} \tag{3.7}$$

for some given cdf F_0 *on* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$*. Then we call*

$$D_n := \rho_{KS}(\hat{F}_n, F_0) = \sup_{z\in\mathbb{R}} \left|\hat{F}_n(z) - F_0(z)\right|$$

the Kolmogorov-Smirnov test statistic *and* φ_{KS} *with*

$$\varphi_{KS}(y) = 1 \iff D_n(y) > c_n^{KS}(\alpha)$$

the Kolmogorov-Smirnov test *at significance level* α *for* (3.7)*, where* $\hat{F}_n$ *denotes the empirical cdf pertaining to* $Y_1, \ldots, Y_n$ *and* $y = (y_1, \ldots, y_n)^\top$ *the observed data.*

If F_0 *is continuous, then* φ_{KS} *is* distribution-free*, meaning that the critical value* $c_n^{KS}(\alpha)$ *only depends on* n *and* α*, and not on* F_0*. If* n *is large, then* $\sqrt{n}c_n^{KS}(\alpha)$ *can be approximated by* $L^{-1}(1-\alpha)$*, where* L *is as in* (3.6)*.*

Theorem 3.6 *Let* $(B_t^0)_{0\le t\le 1}$ *denote a Brownian bridge, and let* W^2 *be as in* (3.5)*. Then it holds that*

$$W^2 \overset{\mathscr{D}}{=} \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2} Z_j^2,$$

where $(Z_j)_{j\ge 1}$ *denotes a sequence of i.i.d. random variables such that* $Z_1 \sim \mathscr{N}(0,1)$*.*

Proof See Theorem 5.3.1 in Shorack and Wellner (1986).

Remark 3.7 The cdf of W^2 is tabulated in Table 1 of Anderson and Darling (1952); see also Table 3.8.4 in Shorack and Wellner (1986).

Application 3.8 (Cramér-von Mises Test) *Under the assumptions of Application 3.5, we call*

$$\omega_n^2 := \rho_{CvM}(\hat{F}_n, F_0) = \int_{\mathbb{R}} \left[\hat{F}_n(z) - F_0(z)\right]^2 dF_0(z)$$

the Cramér-von Mises test statistic *and* φ_{CvM} *with*

$$\varphi_{CvM}(y) = 1 \iff \omega_n^2(y) > c_n^{CvM}(\alpha)$$

the Cramér-von Mises test *at significance level* α *for* (3.7).

If F_0 *is continuous, then* φ_{CvM} *is distribution-free. If* n *is large, then* $nc_n^{CvM}(\alpha)$ *can be approximated by the* $(1-\alpha)$*-quantile of the distribution of* W^2*; cf. Theorem 3.6 and Remark 3.7.*

In view of (3.2), the following result is useful for carrying out a Cramér-von Mises test in practice.

Lemma 3.9 *It holds that*

$$W_n^2 := \int_0^1 \mathbb{U}_n^2(u)du = \sum_{k=1}^n \left(U_{k:n} - \frac{2k-1}{2n}\right)^2 + \frac{1}{12n},$$

where $U_{1:n} < \ldots < U_{n:n}$ *are the order statistics of the uniformly distributed random variables* $U_1, \ldots, U_n$ *which define the reduced empirical process* $\mathbb{U}_n$*, where* $\mathbb{U}_n(u) = \sqrt{n}\left(\hat{G}_n(u) - u\right)$ *for* $0 \le u \le 1$*; cf. Definition 2.16.*

Proof Let $u_{1:n} < u_{2:n} < \ldots < u_{n:n}$ denote arbitrary realizations of $U_{1:n} < U_{2:n} < \ldots < U_{n:n}$, and define $u_{0:n} := 0$ as well as $u_{n+1:n} := 1$. For these realizations, obviously $\hat{G}_n(u)$ takes the value i/n for $u \in [u_{i:n}, u_{i+1:n})$, $0 \le i \le n$. Thus, we get that

$$\begin{aligned} W_n^2 &= n\int_0^1 \left[\hat{G}_n(u) - u\right]^2 du \\ &= n\sum_{i=0}^n \int_{u_{i:n}}^{u_{i+1:n}} \left(\frac{i}{n} - u\right)^2 du = n\sum_{i=0}^n \left[\frac{\left(u - \frac{i}{n}\right)^3}{3}\right]_{u_{i:n}}^{u_{i+1:n}} \\ &= n\left[\sum_{i=0}^{n-1} \frac{\left(u_{i+1:n} - \frac{i}{n}\right)^3}{3} - \sum_{i=1}^n \frac{\left(u_{i:n} - \frac{i}{n}\right)^3}{3}\right] \end{aligned}$$

$$= n\left[\sum_{k=1}^{n}\frac{\left(u_{k:n}-\frac{k-1}{n}\right)^3}{3}-\sum_{i=1}^{n}\frac{\left(u_{i:n}-\frac{i}{n}\right)^3}{3}\right]$$

$$= n\left[\sum_{k=1}^{n}\frac{\left(u_{k:n}-\frac{k-1}{n}\right)^3-\left(u_{k:n}-\frac{k}{n}\right)^3}{3}\right]. \tag{3.8}$$

An elementary calculation now yields that

$$\left(u_{k:n}-\frac{k-1}{n}\right)^3-\left(u_{k:n}-\frac{k}{n}\right)^3=\frac{3u_{k:n}^2}{n}-3u_{k:n}\left(\frac{2k-1}{n^2}\right)+\frac{3k^2-3k+1}{n^3}.$$

Hence, the right-hand side of (3.8) equals

$$\sum_{k=1}^{n}\left\{u_{k:n}^2-\frac{2k-1}{n}u_{k:n}+\frac{3k^2-3k+1}{3n^2}\right\}=\sum_{k=1}^{n}\left\{u_{k:n}^2-\frac{2k-1}{n}u_{k:n}\right\}+\frac{n}{3}, \tag{3.9}$$

because $\sum_{k=1}^{n}(3k^2-3k+1)=n^3$. Finally, notice that the right-hand side of (3.9) equals

$$\sum_{k=1}^{n}\left\{\left(u_{k:n}-\frac{k-1/2}{n}\right)^2-\frac{(k-1/2)^2}{n^2}\right\}+\frac{n}{3}=\sum_{k=1}^{n}\left(u_{k:n}-\frac{2k-1}{2n}\right)^2+\frac{1}{12n}$$

as desired, because $\sum_{k=1}^{n}(k-1/2)^2=n^3/3-n/12$.

If, under the assumptions of Application 3.8, F_0 is continuous and the null hypothesis is true (meaning that $F=F_0$), we have that $n\omega_n^2=n\rho_{\mathrm{CvM}}(\hat{F}_n,F_0)$ behaves like W_n^2, where $U_i=F_0(Y_i)$ for $1\le i\le n$, because of (3.2). Hence, we can also in the general case compute $Z_i:=F_0(Y_i)$ for $1\le i\le n$ and compare the test statistic

$$T_n=\sum_{k=1}^{n}\left(Z_{k:n}-\frac{2k-1}{2n}\right)^2+\frac{1}{12n}$$

with the appropriate critical value $nc_n^{\mathrm{CvM}}(\alpha)$.

3.2 Tests for Parametric Families

Parts of this section are based on the lecture notes of Rui Castro from his Applied Statistics Lectures 2 and 3 in the year 2013.[1]

A reasonable approach for testing the null hypothesis

$$H_0 : P \in \{P_\vartheta : \vartheta \in \Theta \subseteq \mathbb{R}^p, p \in \mathbb{N}\} \tag{3.10}$$

consists in a two-step strategy:

1. Estimate ϑ by $\hat{\vartheta} = \hat{\vartheta}(Y_1, \ldots, Y_n)$.
2. Apply one of the tests treated in Sect. 3.1 with F_0 replaced by $F_{\hat{\vartheta}}$ (the cdf pertaining to $P_{\hat{\vartheta}}$).

However, the fact that $\hat{\vartheta} = \hat{\vartheta}(Y_1, \ldots, Y_n)$ depends on the data yields that the "estimated empirical process" $\sqrt{n}\left(\hat{F}_n - F_{\hat{\vartheta}}\right)$ will, even under the null hypothesis, have a different stochastic behavior than $\sqrt{n}\left(\hat{F}_n - F\right)$, due to the additional variability which is contributed by $\hat{\vartheta}$. In addition, the tests of Kolmogorov-Smirnov- and Cramér-von Mises-type based on $\sqrt{n}\left(\hat{F}_n - F_{\hat{\vartheta}}\right)$ are unfortunately not distribution-free anymore. In particular model classes, though, they are at least *parameter-free*, meaning that their null distribution depends on the family $(P_\vartheta)_{\vartheta \in \Theta}$, but not on the value of ϑ.

Definition 3.10 A family $\{P_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$ of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with corresponding cdfs $F_{\mu,\sigma}$ is called a *location-scale family*, if there exists a cdf $F_{0,1}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that

$$\forall \mu \in \mathbb{R} : \forall \sigma > 0 : \forall y \in \mathbb{R} : F_{\mu,\sigma}(y) = F_{0,1}\left(\frac{y - \mu}{\sigma}\right).$$

In other words: If the real-valued random variable Y possesses the cdf $F_{\mu,\sigma}$ belonging to a location-scale family, then the standardized random variable $(Y - \mu)/\sigma$ possesses the cdf $F_{0,1}$. The parameter μ is called the *location parameter* and the parameter σ is called the *scale parameter.*

Example 3.11 The family $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ of univariate normal distributions constitutes a location-scale family with the expectation μ as the location parameter and the standard deviation σ as the scale parameter. Here, the cdf $F_{0,1} = \Phi$ is the cdf of the standard normal distribution on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Definition 3.12 Let $Y_1, \ldots, Y_n$ denote some real-valued random variables and let $T : \mathbb{R}^n \to \mathbb{R}$ be a measurable transformation. Then we call T

[1] These lecture notes are available at http://www.win.tue.nl/~rmcastro/AppStat2013/.

(i) *location-scale invariant in distribution*, if

$$\forall a > 0 : \forall b \in \mathbb{R} : T(aY_1 + b, \ldots, aY_n + b) \stackrel{\mathscr{D}}{=} aT(Y_1, \ldots, Y_n) + b.$$

(ii) *scale invariant in distribution*, if

$$\forall a > 0 : \forall b \in \mathbb{R} : T(aY_1 + b, \ldots, aY_n + b) \stackrel{\mathscr{D}}{=} aT(Y_1, \ldots, Y_n).$$

Lemma 3.13 *Let $\{P_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$ be a location-scale family of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with corresponding cdfs $F_{\mu,\sigma}$. Let $Y_1, \ldots, Y_n$ denote an i.i.d. sample from this family. Assume that the maximum likelihood estimators $\hat{\mu} = \hat{\mu}(Y_1, \ldots, Y_n)$ of μ and $\hat{\sigma} = \hat{\sigma}(Y_1, \ldots, Y_n)$ of σ exist. Then $\hat{\mu}$ is location-scale invariant in distribution and $\hat{\sigma}$ is scale invariant in distribution.*

Proof If $Y_1 \sim P_{\mu,\sigma}$ and $a > 0$ and $b \in \mathbb{R}$ are given constants, then

$$\begin{aligned} \mathbb{P}(aY_1 + b \leq y) &= \mathbb{P}\left(Y_1 \leq \frac{y-b}{a}\right) \\ &= F_{\mu,\sigma}\left(\frac{y-b}{a}\right) = F_{0,1}\left(\frac{y-b-a\mu}{a\sigma}\right). \end{aligned} \tag{3.11}$$

On the other hand, if $Y_1 \sim P_{a\mu+b,a\sigma}$, we have that

$$\mathbb{P}(Y_1 \leq y) = F_{0,1}\left(\frac{y-b-a\mu}{a\sigma}\right). \tag{3.12}$$

Obviously, the right-hand sides of (3.11) and (3.12) coincide, implying that

$$\hat{\mu}(aY_1 + b, \ldots, aY_n + b) \stackrel{\mathscr{D}}{=} \widehat{a\mu + b}(Y_1, \ldots, Y_n).$$

However, due to the parametrization invariance of the maximum likelihood method (cf., e.g., Zehna 1966 or Section 2.8 of Barndorff-Nielsen and Cox 1994), we have that $\widehat{a\mu + b} = a\hat{\mu} + b$. Hence, $\hat{\mu}$ is location-scale invariant in distribution. The assertion for $\hat{\sigma}$ follows analogously.

Corollary 3.14 *Under the assumptions of Lemma 3.13, the statistics $(\hat{\mu} - \mu)/\sigma$ and $\hat{\sigma}/\sigma$ are pivotal, meaning that their distributions do not depend on μ or σ.*

Proof This is Exercise 3.5.

Theorem 3.15 *Under the assumptions of Lemma 3.13, assume additionally that $F_{\mu,\sigma}$ is continuous for all $\mu \in \mathbb{R}$ and all $\sigma > 0$. Then, the distribution of the "estimated empirical process" $\sqrt{n}\left(\hat{F}_n - F_{\hat{\mu},\hat{\sigma}}\right)$ does not depend on μ or σ.*

Proof Recall that

$$\hat{F}_n(y) - F_{\hat{\mu},\hat{\sigma}}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le y\} - F_{\hat{\mu},\hat{\sigma}}(y)$$

for all $y \in \mathbb{R}$. Now, substitute $y := F^{-1}_{\hat{\mu},\hat{\sigma}}(u)$ for $0 \le u \le 1$. This yields that

$$\begin{aligned}\hat{F}_n(y) - F_{\hat{\mu},\hat{\sigma}}(y) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le F^{-1}_{\hat{\mu},\hat{\sigma}}(u)\} - u \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{F_{\hat{\mu},\hat{\sigma}}(Y_i) \le u\} - u,\end{aligned}$$

because $F_{\hat{\mu},\hat{\sigma}}$ is continuous. Finally, notice that, for all $1 \le i \le n$, the distribution of

$$Z_i := F_{\hat{\mu},\hat{\sigma}}(Y_i) = F_{0,1}\left(\frac{Y_i - \hat{\mu}}{\hat{\sigma}}\right) = F_{0,1}\left(\frac{\frac{Y_i-\mu}{\sigma} - \frac{\hat{\mu}-\mu}{\sigma}}{\hat{\sigma}/\sigma}\right) \tag{3.13}$$

does not depend on μ or σ, due to Corollary 3.14.

Remark 3.16 The right-hand side of (3.13) reveals that, even under the null hypothesis, the distribution of Z_i will typically not be uniform, $1 \le i \le n$, because $(Y_i - \mu)/\sigma \sim F_{0,1}$ and the statistics $(\hat{\mu} - \mu)/\sigma$ and $\hat{\sigma}/\sigma$ will typically have non-degenerate distributions, at least for finite values of n.

Application 3.17 *The result of Theorem 3.15 suggests the following procedure for calibrating a test for the null hypothesis (3.10), provided that the assumptions of Theorem 3.15 are fulfilled.*

(i) Choose a suitable test statistic T_n, for example D_n from Application 3.5 or ω_n^2 from Application 3.8, where F_0 is replaced by $F_{\hat{\mu},\hat{\sigma}}$ in both cases.

(ii) Generate a (pseudo-) random sample $\tilde{Y}_1, \ldots, \tilde{Y}_n$ from $F_{0,1}$ on the computer.

(iii) Compute the value of the test statistic T_n on the pseudo-sample $\tilde{Y}_1, \ldots, \tilde{Y}_n$, by computing the values of $\hat{\mu} = \hat{\mu}(\tilde{Y}_1, \ldots, \tilde{Y}_n)$, $\hat{\sigma} = \hat{\sigma}(\tilde{Y}_1, \ldots, \tilde{Y}_n)$, and $\tilde{Z}_i = F_{0,1}\left((\tilde{Y}_i - \hat{\mu})/\hat{\sigma}\right)$ for $1 \le i \le n$.

(iv) Repeat steps (ii) and (iii) B times, store the B computed values of the test statistic, and approximate the critical value for the test based on $T_n(Y_1, \ldots, Y_n)$ by the $(1-\alpha)$-quantile of the B stored values.

3.3 Exercises

Exercise 3.1 *Prove assertion (3.2) in Theorem 3.2 under the assumptions that P is absolutely continuous with Lebesgue density f and that F is strictly isotone on its support.*

Exercise 3.2 *Compute the expectation and the variance of the statistic W^2 appearing in (3.5) and in Theorem 3.6.*

Exercise 3.3 (Programming Exercise)

(a) Write an R *program which simulates paths of the reduced empirical process $\mathbb{U}_n$ on $[0,1]$ for $n \in \{20, 50, 100, 500, 1000\}$, respectively. For each value of n, perform $B = 10{,}000$ simulation runs. On the basis of these simulations, approximate the critical values $c_n^{KS}(0.05)$, $n \in \{20, 50, 100, 500, 1000\}$, of the Kolmogorov-Smirnov test at significance level $\alpha = 5\%$ in the case of continuous F_0 (cf. Application 3.5).*

(b) Verify by means of Theorem 3.4 that $c_n^{KS}(0.05)$ can be approximated by $1.358/\sqrt{n}$ for large values of n. Compare your approximated values from part (a) of this exercise with the latter approximation.

Exercise 3.4 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) Under the assumptions of Definition 3.1, ρ_{CvM} is symmetric, i.e., for all $Q_1, Q_2 \in \mathscr{P}$ it holds that $\rho_{CvM}(Q_1, Q_2) = \rho_{CvM}(Q_2, Q_1)$.

(b) Under the assumptions of Theorem 3.4 it holds that

$$L(x) = \sum_{m=-\infty}^{\infty} (-1)^m \exp(-2m^2x^2).$$

(c) If, under the assumptions of Application 3.5, the cdf F_0 is not continuous, then the critical value $c_n^{KS}(0.05)$ can nevertheless be approximated by means of a computer simulation (with in principle arbitrary precision).

(d) The statistic W^2 appearing in (3.5) and in Theorem 3.6 is in distribution equal to a weighted sum of chi-square-distributed random variables.

Exercise 3.5 *Prove Corollary 3.14.*

Exercise 3.6 (Programming Exercise) *Let $Y_1, \ldots, Y_n$ be real-valued, stochastically independent and identically normally distributed random variables, where both $\mu = \mathbb{E}[Y_1]$ and $\sigma^2 = Var(Y_1)$ are unknown.*

Write an R *program which simulates paths of the "estimated empirical process"*

$$\sqrt{n}\left(\hat{F}_n - F_{\hat{\mu},\hat{\sigma}}\right)$$

for $n \in \{20, 50, 100, 500, 1000\}$*; cf. Application 3.17. For every value of* n*, perform* $B = 10{,}000$ *simulation runs. By means of these simulations, approximate the critical values* $c_n^{KS2}(0.05)$*,* $n \in \{20, 50, 100, 500, 1000\}$*, of the two-stage Kolmogorov-Smirnov test at significance level* $\alpha = 5\%$ *according to the construction principle considered in Sect. 3.2.*

Exercise 3.7 (Tests for Exponentiality) *Let* $\{Exp(\lambda) : \lambda > 0\}$ *denote the family of exponential distributions on* $(\mathbb{R}_{\geq 0}, \mathscr{B}(\mathbb{R}_{\geq 0}))$ *with intensity parameter* $\lambda > 0$*, and let* $\{F_\lambda : \lambda > 0\}$ *denote the corresponding cdfs.*

(a) Show that $\{Exp(\lambda) : \lambda > 0\}$ *constitutes a scale parameter family, and determine the scale parameter.*
Hint: *A family* $\{P_\sigma : \sigma > 0\}$ *of probability distributions on* $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ *with corresponding cdfs* $\{F_\sigma : \sigma > 0\}$ *is called a scale parameter family with scale parameter* σ*, if there exists a cdf* H *on* $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ *such that for all* $y \in \mathscr{Y}$ *and all* $\sigma > 0$*, it holds that* $F_\sigma(y) = H(y/\sigma)$*.*

(b) Now, let $n \in \mathbb{N}$*, and* $Y_1, \ldots, Y_n$ *be stochastically independent and identically distributed random variables with* $Y_1 \sim Exp(\lambda)$ *for unknown* $\lambda > 0$*. Denote by* $\hat{\lambda} = \hat{\lambda}(Y_1, \ldots, Y_n)$ *the maximum likelihood estimator for* λ *based on* $Y_1, \ldots, Y_n$*. Show that under these assumptions the distribution of* $\hat{\lambda}/\lambda$ *is pivotal.*

(c) Show that under the assumptions of part (b) the distribution of the "estimated empirical process" $\sqrt{n}\left(\hat{F}_n - F_{\hat{\lambda}}\right)$ *does not depend on the value of* λ*. Discuss implications of the latter result for goodness-of-fit tests for exponentiality.*

References

Anderson T, Darling D (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Stat 23:193–212. https://doi.org/10.1214/aoms/1177729437

Barndorff-Nielsen O, Cox D (1994) Inference and asymptotics. Chapman and Hall, London

Deza MM, Deza E (2016) Encyclopedia of distances, 4th edn. Springer, Berlin. https://doi.org/10.1007/978-3-662-52844-0

Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione. G Ist Ital Attuari 4:83–91

Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley series in probability and mathematical statistics. Wiley, New York, NY

Zehna P (1966) Invariance of maximum likelihood estimators. Ann Math Stat 37:744. https://doi.org/10.1214/aoms/1177699475

Chapter 4
Rank Tests

4.1 Parametric Score Tests

Test problems involving composite null and/or alternative hypotheses are often highly non-trivial. Only in special cases (e.g., for models with isotone likelihood ratio in which the Neyman–Pearson theory is applicable) there exists a satisfactory methodology allowing for uniformly (over $\vartheta \in H_1$) best level α-tests. If the "geometry" of the statistical model is more complicated, then it is typically impossible to minimize the type II error probability under the level constraint uniformly. This issue is ubiquitous in nonparametric test problems. Hence, in the latter case the researcher has to decide for which kinds of alternatives the test procedure shall be most sensitive, meaning towards which "regions" of H_1 optimal power are targeted.

In the one-parametric case, one class of procedures is constituted by locally best tests. They are targeted towards regions of H_1 which are "close to H_0". Derivation of locally best tests requires the concept of L_1-differentiability or differentiability in the mean.

Definition 4.1 (Differentiability in the Mean) Let $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ denote a statistical model and assume that $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ is a dominated (by μ) family of measures, where $\Theta \subseteq \mathbb{R}$. Then, $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ is called differentiable in the mean in $\vartheta_0 \in \mathring{\Theta}$, if a function $g \in L_1(\mu)$ exists with

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

The function g is called $L_1(\mu)$-derivative of $\vartheta \mapsto \mathbb{P}_\vartheta$ in ϑ_0. In the sequel, we choose w.l.o.g. $\vartheta_0 \equiv 0$.

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_4

Theorem 4.2 (§18 in Hewitt and Stromberg (1975), Satz 1.183 in Witting (1985)) *Under the assumptions of Definition 4.1 let $\vartheta_0 = 0$ and let densities be given by $f_\vartheta(y) = d\mathbb{P}_\vartheta/(d\mu)(y)$. Assume that there exists an open neighborhood $\mathscr{U}$ of 0 such that for μ-almost all y the mapping $\mathscr{U} \ni \vartheta \mapsto f_\vartheta(y)$ is absolutely continuous, i.e., it exists an integrable function $\tau \mapsto \dot{f}(y,\tau)$ on $\mathscr{U}$ with*

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

and assume that $\frac{\partial}{\partial\vartheta} f_\vartheta(y)|_{\vartheta=0} = \dot{f}(y,0)$ μ-almost everywhere. Furthermore, assume that for $\vartheta \in \mathscr{U}$ the function $y \mapsto \dot{f}(y,\vartheta)$ is μ-integrable with

$$\int |\dot{f}(y,\vartheta)|d\mu(y) \to \int |\dot{f}(y,0)|d\mu(y), \quad \vartheta \to 0.$$

Then, $\vartheta \mapsto \mathbb{P}_\vartheta$ is differentiable in the mean in 0 with $g = \dot{f}(\cdot,0)$.

Roughly speaking, Theorem 4.2 yields that in the case of absolute continuity the L_1-derivative can be computed by usual differentiation of the density with respect to the parameter. Further applications of Theorem 4.2 are considered in Example 4.4.

Definition and Theorem 4.3 *Under the assumptions of Definition 4.1 assume that the densities $\vartheta \mapsto f_\vartheta$ are differentiable in the mean in 0 with $L_1(\mu)$-derivative g. Then,*

$$\vartheta^{-1}\log(f_\vartheta(y)/f_0(y)) = \vartheta^{-1}(\log f_\vartheta(y) - \log f_0(y))$$

converges for $\vartheta \to 0$ to $\dot{L}(y)$ (say) in $\mathbb{P}_0$-probability. We call $\dot{L} : \mathscr{Y} \to \mathbb{R}$ the derivative of the logarithmic likelihood ratio or score function. *It holds*

$$\dot{L}(y) = g(y)/f_0(y) \text{ and } \int_{\mathscr{Y}} \dot{L}d\mathbb{P}_0 = 0. \tag{4.1}$$

Proof $\vartheta^{-1}(f_\vartheta/f_0 - 1) \to g/f_0$ converges in $L_1(\mathbb{P}_0)$ and, consequently, in $\mathbb{P}_0$-probability. The chain rule yields that $\dot{L}(y) = g(y)/f_0(y)$. Noting that $\int_{\mathscr{Y}} (f_\vartheta - f_0)d\mu = 0$ for all $\vartheta \in \Theta$ we conclude that

$$\int_{\mathscr{Y}} \dot{L}d\mathbb{P}_0 = \int_{\mathscr{Y}} g d\mu = 0.$$

Example 4.4

(a) *Location parameter model:*
Let $Y = \vartheta + X$, $\vartheta \geq 0$, and assume that X has a density f which is absolutely continuous with respect to the Lebesgue measure and does not depend on ϑ.

Then, the densities $\vartheta \mapsto f(y-\vartheta)$ of Y under ϑ are differentiable in the mean in zero with score function $\dot{L}$, given by $\dot{L}(y) = -f'(y)/f(y)$ (where the prime indicates differentiation with respect to y).

(b) *Scale parameter model:*
Let $Y = \exp(\vartheta)X$ and assume again that X has density f with the properties stated in part (a). Moreover, assume that $\int |xf'(x)|dx < \infty$. Then, the densities $\vartheta \mapsto \exp(-\vartheta) f(y\exp(-\vartheta))$ of Y under ϑ are differentiable in the mean in zero with score function $\dot{L}$, given by $\dot{L}(y) = -[1 + yf'(y)/f(y)]$.

Lemma 4.5 *Assume that the family $\vartheta \mapsto P_\vartheta$ is differentiable in the mean with score function $\dot{L}$ in $\vartheta_0 = 0$ and that c_i, $1 \le i \le n$, are real constants. Then, also $\vartheta \mapsto \bigotimes_{i=1}^n P_{c_i\vartheta}$ is differentiable in the mean in zero, with score function*

$$(y_1, \ldots, y_n)^\top \mapsto \sum_{i=1}^n c_i \dot{L}(y_i).$$

Proof This is Exercise 4.1.

Definition 4.6 (Score Test) Let $\vartheta \mapsto \mathbb{P}_\vartheta$ be differentiable in the mean in ϑ_0 with score function $\dot{L}$. Then, every test φ of the form

$$\varphi(y) = \begin{cases} 1, & \text{if } \dot{L}(y) > c, \\ \gamma, & \text{if } \dot{L}(y) = c, \\ 0, & \text{if } \dot{L}(y) < c, \end{cases}$$

is called a score test. In this, $\gamma \in [0, 1]$ denotes a randomization constant.

Definition 4.7 (Locally Best Test) Let $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ with $\Theta \subseteq \mathbb{R}$ denote a family which is differentiable in the mean in $\vartheta_0 \in \Theta$. A test φ^* with $\mathbb{E}_{\vartheta_0}[\varphi^*] = \alpha$ is called locally best test among all tests with expectation α under ϑ_0 for the test problem $H_0 = \{\vartheta_0\}$ versus $H_1 = \{\vartheta > \vartheta_0\}$ if

$$\frac{d}{d\vartheta}\mathbb{E}_\vartheta[\varphi^*]\Big|_{\vartheta=\vartheta_0} \ge \frac{d}{d\vartheta}\mathbb{E}_\vartheta[\varphi]\Big|_{\vartheta=\vartheta_0}$$

for all tests φ with $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$.

Figure 4.1 illustrates the situation considered in Definition 4.7 graphically.

Remark 4.8 As indicated by Fig. 4.1, it is possible that locally best tests have suboptimal power properties for values of $\vartheta > \vartheta_0$ which have a large distance to ϑ_0.

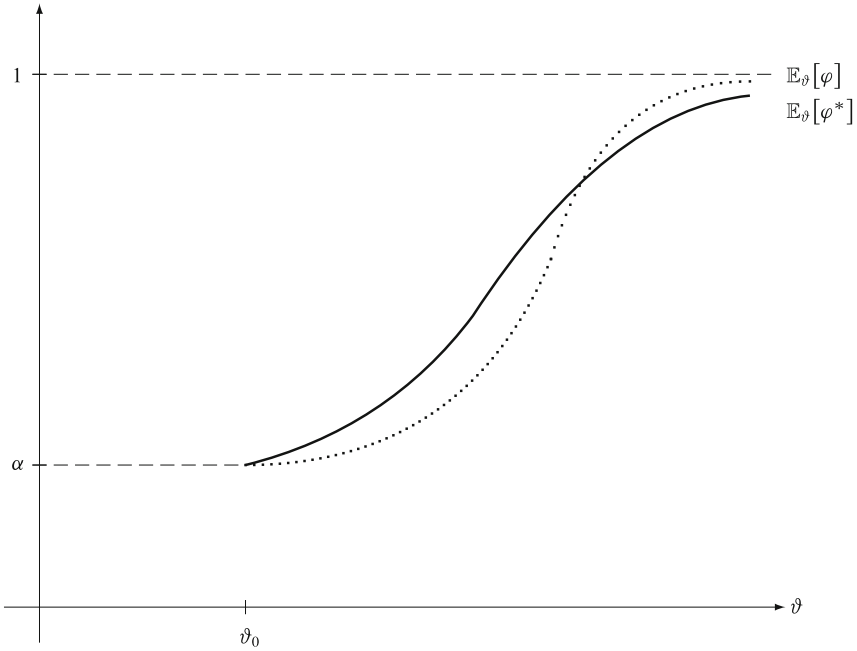

Fig. 4.1 Locally best test φ^* with expectation α under ϑ_0

Theorem 4.9 (Satz 2.44 in Witting (1985)) *Under the assumptions of Definition 4.7, the score test φ, given by*

$$\varphi(y) = \begin{cases} 1, & \text{if } \dot{L}(y) > c(\alpha) \\ \gamma, & \text{if } \dot{L}(y) = c(\alpha), \quad \gamma \in [0,1] \\ 0, & \text{if } \dot{L}(y) < c(\alpha) \end{cases} \tag{4.2}$$

with $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$, is a locally best test for testing $H_0 = \{\vartheta_0\}$ against $H_1 = \{\vartheta > \vartheta_0\}$ among all tests with expectation α under ϑ_0.

Proof We notice that for any test φ and any $\vartheta \in \Theta$, it holds

$$\mathbb{E}_\vartheta[\varphi] = \int_{\mathscr{Y}} \varphi(y) f_\vartheta(y) \mu(dy)$$

and hence,

$$\begin{aligned}\frac{d}{d\vartheta}\mathbb{E}_\vartheta[\varphi]\Big|_{\vartheta=\vartheta_0} &= \int_{\mathscr{Y}} \varphi(y)g(y)\mu(dy)\\ &= \int_{\mathscr{Y}} \varphi(y)\dot{L}(y)f_{\vartheta_0}(y)\mu(dy)\\ &= \int_{\mathscr{Y}} \varphi(y)\dot{L}(y)\mathbb{P}_{\vartheta_0}(dy) = \mathbb{E}_{\vartheta_0}[\varphi\dot{L}].\end{aligned}$$

Thus, we have to maximize $\mathbb{E}_{\vartheta_0}[\varphi\dot{L}]$ with respect to φ under the condition that $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$. To this end, for the sake of simplicity, assume that the distribution of $\dot{L}$ under ϑ_0 is absolutely continuous, and let c_α be such that

$$\mathbb{P}_{\vartheta_0}(\dot{L} > c_\alpha) = \alpha.$$

We get that

$$\begin{aligned}\mathbb{E}_{\vartheta_0}[\varphi\dot{L}] - \alpha c_\alpha &= \mathbb{E}_{\vartheta_0}[\varphi\dot{L}] - c_\alpha\mathbb{E}_{\vartheta_0}[\varphi]\\ &= \mathbb{E}_{\vartheta_0}[(\dot{L} - c_\alpha)\varphi]\\ &\le \mathbb{E}_{\vartheta_0}[(\dot{L} - c_\alpha)_+]\end{aligned}$$

with equality for $\varphi = \mathbf{1}\{\dot{L} > c_\alpha\}$.

The general case can be treated with the usual modifications, leading to the randomized version of the score test as given in (4.2).

Theorem 4.9 shows that in the theory of locally best tests the score function $\dot{L}$ takes the role that the likelihood ratio has in the Neyman–Pearson theory. Notice that, for an i.i.d. sample $Y = (Y_1, \ldots, Y_n)^\top$ with $Y_1 \sim P_\vartheta$, the joint distribution of Y is the product measure $\mathbb{P}_\vartheta = P_\vartheta^{\otimes n}$. It has the score function $(y_1, \ldots, y_n)^\top \mapsto \sum_{i=1}^n \dot{L}(y_i)$, where $\dot{L}$ is the score function pertaining to $(P_\vartheta : \vartheta \in \Theta)$ in ϑ_0, according to Lemma 4.5. Hence, Theorem 4.9 can be applied to test $H_0 = \{\vartheta_0\}$ against $H_1 = \{\vartheta > \vartheta_0\}$ based on Y in the following manner: We reject H_0, iff $\sum_{i=1}^n \dot{L}(y_i)$ exceeds its critical value.

Moreover, for k-sample problems with $k \ge 2$ groups and n jointly independent observables, Lemma 4.5 can be utilized to test the homogeneity hypothesis

$$H_0 = \big\{\mathbb{P}^{Y_1} = \mathbb{P}^{Y_2} = \ldots = \mathbb{P}^{Y_n} \colon \mathbb{P}^{Y_1} \text{ absolutely continuous}\big\}. \tag{4.3}$$

To this end, one considers auxiliary parametric families $\vartheta \mapsto \mathbb{P}_{n,\vartheta}$ which belong to H_0 only in case of $\vartheta = 0$, i.e., $\mathbb{P}_{n,0} \in H_0$ and $\mathbb{P}_{n,\vartheta} \notin H_0$, $\vartheta \neq 0$. For $\vartheta \neq 0$, $\mathbb{P}_{n,\vartheta}$ is a product measure with non-identical factors.

Example 4.10

(a) *Regression model for a location parameter*
Let $Y_i = c_i\vartheta + X_i$, $1 \le i \le n$, where $\vartheta \ge 0$. In this, assume that the X_i are i.i.d. with Lebesgue density f which is independent of ϑ. Now, for a two-sample problem with n_1 observations in the first group and $n_2 = n - n_1$ observations in the second group, we set $c_1 = c_2 = \cdots = c_{n_1} = 1$ and $c_i = 0$ for all $n_1 + 1 \le i \le n$. Under alternatives, the observations in the first group are shifted by $\vartheta > 0$.

(b) *Regression model for a scale parameter*
Let c_i, $1 \le i \le n$, denote real regression coefficients and consider the model $Y_i = \exp(c_i\vartheta)X_i$, $1 \le i \le n$, $\vartheta \in \mathbb{R}$, where we assume again that the X_i are i.i.d. with Lebesgue density f which is independent of ϑ. Then, it holds

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

Under $\vartheta_0 = 0$, the product measure $\mathbb{P}_{n,0}$ belongs to H_0, while under alternatives it does not.

(c) *General model*
Let $\vartheta \mapsto P_\vartheta$ be a one-parametric curve of probability distributions depending on a real-valued parameter ϑ. Define $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^{n} P_{c_i\vartheta}$ with real constants $c_1, \ldots, c_n$ as the model for the sample $Y_1, \ldots, Y_n$.

In the nonparametric case, H_0 from (4.3) shall of course be tested without utilizing a specific density f in the calibration of the test. This can be achieved by conditioning on the ranks of the observations.

4.2 Deriving Rank Tests by Conditioning

Theorem 4.11 *Let $\vartheta \mapsto \mathbb{P}_\vartheta$ denote a parametric family which is $L_1(\mu)$-differentiable in $\vartheta_0 = 0$ with score function $\dot{L}$. Furthermore, let $S : \mathscr{Y} \to \mathscr{S}$ be a statistic. Then, $\vartheta \mapsto \mathbb{P}_\vartheta^S$ is $L_1(\mu^S)$-differentiable in zero with score function $s \mapsto \mathbb{E}_0[\dot{L} \,|\, S = s]$.*

Proof First, we show that the $L_1(\mu^S)$-derivative of $\vartheta \mapsto \mathbb{P}_\vartheta^S$ is given by $s \mapsto \mathbb{E}_\mu[g \,|\, S = s]$, where g is the $L_1(\mu)$-derivative of $\vartheta \mapsto \mathbb{P}_\vartheta$, meaning that

$$\vartheta^{-1}(f_\vartheta - f_0) \longrightarrow g \ \text{ in } \ L_1(\mu), \ \vartheta \to 0. \tag{4.4}$$

To this end, notice that (see Satz 1.121.b) in Witting (1985))

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

for probability measures P and Q and a statistic T. Applying this result to our situation yields that

$$\frac{d\mathbb{P}_\vartheta^S}{d\mu^S}(s) = \mathbb{E}_\mu[f_\vartheta \mid S = s], \quad \text{where} \quad f_\vartheta = \frac{d\mathbb{P}_\vartheta}{d\mu}$$

as before. Linearity of $\mathbb{E}_\mu[\cdot \mid S = s]$ and transformation of measures leads to

$$\begin{aligned}
&\int \left| \vartheta^{-1} \left(\frac{d\mathbb{P}_\vartheta^S}{d\mu^S}(s) - \frac{d\mathbb{P}_0^S}{d\mu^S}(s) \right) - \mathbb{E}_\mu[g \mid S = s] \right| d\mu^S(s) \\
&= \int \left| \vartheta^{-1} \left(\mathbb{E}_\mu[f_\vartheta \mid S = s] - \mathbb{E}_\mu[f_0 \mid S = s] \right) - \mathbb{E}_\mu[g \mid S = s] \right| d\mu^S(s) \\
&= \int \left| \mathbb{E}_\mu[\vartheta^{-1}(f_\vartheta - f_0) - g \mid S = s] \right| d\mu^S(s) \\
&= \int \left| \mathbb{E}_\mu[\vartheta^{-1}(f_\vartheta - f_0) - g \mid S] \right| d\mu.
\end{aligned}$$

Applying Jensen's inequality and Vitali's Theorem 1.11, we conclude that $s \mapsto \mathbb{E}_\mu[g \mid S = s]$ is $L_1(\mu^S)$-derivative of $\vartheta \mapsto \mathbb{P}_\vartheta^S$.

Now, the chain rule yields that the score function of $\mathbb{P}_\vartheta^S$ in zero is given by

$$s \mapsto \frac{\mathbb{E}_\mu[g \mid S = s]}{\mathbb{E}_\mu[f_0 \mid S = s]} = \frac{\mathbb{E}_\mu[\dot{L} f_0 \mid S = s]}{\mathbb{E}_\mu[f_0 \mid S = s]}$$

and the assertion follows by verifying that

$$\mathbb{E}_0[\dot{L} \mid S]\, \mathbb{E}_\mu[\frac{d\mathbb{P}_0}{d\mu} \mid S] = \mathbb{E}_\mu[\dot{L} \frac{d\mathbb{P}_0}{d\mu} \mid S] \quad \mu\text{-almost surely,} \tag{4.5}$$

which is Exercise 4.3.

As outlined before, we will employ Theorem 4.11 in order to derive (nonparametric) rank tests based on the parametric curves $\mathbb{P}_{n,\vartheta}$ considered in Example 4.10. It will turn out that the coarsening of the information (only the ranks of the observables are utilized, not their actual values) leads to a simple structure of the score test statistics (namely, it leads to linear rank statistics). Furthermore, ranks have the advantage of being robust against model misspecification. In some applications, only the ranks of the observations are trustworthy (or available).

As a preparation, we will first gather some basic results about ranks and order statistics. Detailed proofs for these results can be found, e.g., in §1 and §2 of Janssen (1998), in Reiss (1989), or in other textbooks on the subject.

Definition 4.12 Let $y = (y_1, \ldots, y_n)^\top$ be a point in $\mathbb{R}^n$. Assume that the y_i are pairwise distinct and denote their ordered values by $y_{1:n} < y_{2:n} < \ldots < y_{n:n}$.

(a) For $1 \le i \le n$, the integer $r_i \equiv r_i(y) := \#\{j \in \{1, \ldots, n\} : y_j \le y_i\}$ is called the rank of y_i (in y). The permutation $r(y) := (r_1(y), \ldots, r_n(y))^\top \in \mathscr{S}_n$ is called rank vector of y.
(b) The inverse permutation $d(y) = (d_1(y), \ldots, d_n(y))^\top := [r(y)]^{-1}$ is called the vector of antiranks of y, and the integer $d_i(y)$ is called antirank of i (the index that corresponds to the i-th smallest observation in y).

Now, let $Y_1, \ldots, Y_n$ with $Y_i : \Omega_i \to \mathbb{R}$ be stochastically independent, absolutely continuously distributed random variables, all driven by the same probability measure $\mathbb{P}$.

(c) Because of $\mathbb{P}(\bigcup_{i \ne j}\{Y_i = Y_j\}) = 0$ the following objects are almost surely uniquely defined: $Y_{i:n}$ is called i-th order statistic of $Y = (Y_1, \ldots, Y_n)^\top$, $R_i(Y) := n\hat{F}_n(Y_i) = r_i(Y_1, \ldots, Y_n)$ is called rank of Y_i, $R(Y) := (R_1(Y), \ldots, R_n(Y))^\top$ is called vector of rank statistics of Y, $D_i(Y) := d_i(Y_1, \ldots, Y_n)$ is called antirank of i with respect to Y and $D(Y) := d(Y)$ is called vector of antiranks of Y.

Lemma 4.13 *Under the assumptions of Definition 4.12, it holds*

(a) $i = r_{d_i} = d_{r_i}, \quad y_i = y_{r_i:n}, \quad y_{i:n} = y_{d_i}$.
(b) If $Y_1, \ldots, Y_n$ are exchangeable random variables, then $R(Y)$ is uniformly distributed on $\mathscr{S}_n$, i.e., $\mathbb{P}(R(Y) = \sigma) = 1/n!$ for all permutations $\sigma = (r_1, \ldots, r_n)^\top \in \mathscr{S}_n$.
(c) If $U_1, \ldots, U_n$ are i.i.d. with $U_1 \sim UNI[0, 1]$, and $Y_i = F^{-1}(U_i)$, $1 \le i \le n$, for some c.d.f. F, then it holds $Y_{i:n} = F^{-1}(U_{i:n})$. If F is continuous, then it holds $R(Y) = R(U_1, \ldots, U_n)$.
(d) If $(Y_1, \ldots, Y_n)$ are i.i.d. with c.d.f. F of Y_1, then we have

(i) $\mathbb{P}(Y_{i:n} \le y) = \sum_{j=i}^n \binom{n}{j} F(y)^j (1 - F(y))^{n-j}$.

(ii) $\frac{d\mathbb{P}^{Y_{i:n}}}{d\mathbb{P}^{Y_1}}(y) = n\binom{n-1}{i-1} F(y)^{i-1}(1 - F(y))^{n-i}$. *If $\mathbb{P}^{Y_1}$ has Lebesgue density f, then $\mathbb{P}^{Y_{i:n}}$ has Lebesgue density $f_{i:n}$, given by* $f_{i:n}(y) = n\binom{n-1}{i-1} F(y)^{i-1}(1 - F(y))^{n-i} f(y)$.

(iii) Letting $P = \mathbb{P}^{Y_1}$, $(Y_{i:n})_{1 \le i \le n}$ has the joint $P^{\otimes n}$-density

$$(y_1, \ldots, y_n) \mapsto n! \mathbf{1}_{\{y_1 < y_2 < \ldots < y_n\}}.$$

If P has Lebesgue density f, then $(Y_{i:n})_{1 \le i \le n}$ has Lebesgue density

$$(y_1, \ldots, y_n) \mapsto n! \prod_{i=1}^n f(y_i) \mathbf{1}_{\{y_1 < y_2 < \ldots < y_n\}}.$$

Remark 4.14

(a) Part (c) of Lemma 4.13 (quantile transformation of order statistics) shows the special importance of the distribution of order statistics of i.i.d. UNI[0, 1]-distributed random variables $U_1, \ldots, U_n$. According to part (d) of Lemma 4.13, the order statistic $U_{i:n}$ has a Beta$(i, n - i + 1)$ distribution with $\mathbb{E}[U_{i:n}] = i/(n+1)$ and $\mathrm{Var}(U_{i:n}) = [i(n-i+1)]/[(n+1)^2(n+2)]$.

(b) For computing the joint cdf of $(U_{1:n}, \ldots, U_{n:n})^\top$, efficient recursive algorithms exist, for instance Bolshev's recursion and Steck's recursion (see Shorack and Wellner 1986, p. 362 ff.).

Definition 4.15 (Sufficient Statistic) Let $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical model, and let $Y \sim \mathbb{P}_\vartheta$ denote a random variable with values in $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$. Then, a statistic $S : (\mathscr{Y}, \mathscr{B}(\mathscr{Y})) \to (\mathscr{S}, \mathscr{B}(\mathscr{S}))$ is called *sufficient* (for ϑ), if there exists for all $\vartheta \in \Theta$ a regular version of the conditional distribution of Y with respect to S which does not depend on ϑ, i.e.,

$$\exists h : \forall \vartheta \in \Theta, \forall B \in \mathscr{B}(\mathscr{Y}) : \mathbb{E}_\vartheta\big[\mathbf{1}_B|S\big] = \mathbb{P}_\vartheta(B|S) = h(B, S) \quad \mathbb{P}_\vartheta - \text{almost surely.}$$

Theorem 4.16 *Let $Y = (Y_1, \ldots, Y_n)^\top$ be a random vector with values in $(\mathscr{Y}^n, \mathscr{B}(\mathscr{Y}^n))$ such that $Y \sim P_\vartheta^{\otimes n}$ for an absolutely continuous distribution P_ϑ depending on $\vartheta \in \Theta$. Then, the vector $(Y_{i:n} : 1 \le i \le n)$ of the order statistics of $Y_1, \ldots, Y_n$ is sufficient for ϑ.*

Proof This is Exercise 4.4.

Definition 4.17 Under the assumptions of Definition 4.15, the statistic S is called *complete* for $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, if for every measurable function g and for every $\vartheta \in \Theta$, it holds

$$\mathbb{E}_\vartheta[g(S(Y))] = 0 \Rightarrow \mathbb{P}_\vartheta(g(S(Y)) = 0) = 1. \tag{4.6}$$

The statistic S is said to be *boundedly complete*, if the implication in (4.6) holds for every measurable function g which is also bounded.

Theorem 4.18 (Basu's Theorem) *Let $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical model, $S : (\mathscr{Y}, \mathscr{B}(\mathscr{Y})) \to (\mathscr{S}, \mathscr{B}(\mathscr{S}))$ a boundedly complete sufficient statistic for ϑ, and $T : (\mathscr{Y}, \mathscr{B}(\mathscr{Y})) \to (\mathscr{T}, \mathscr{B}(\mathscr{T}))$ ancillary to ϑ (i.e., a pivotal statistic). Then, S and T are stochastically independent.*

Proof Let $B \in \mathscr{B}(\mathscr{T})$ be any event. We have that

$$\mathbb{P}_\vartheta^T(B) = \int_{\mathscr{S}} \mathbb{P}_\vartheta^T(B|S = s)\mathbb{P}_\vartheta^S(ds).$$

Now, notice that $\mathbb{P}_\vartheta^T(B)$ does not depend on ϑ, because T is ancillary to ϑ. On the other hand, also $\mathbb{P}_\vartheta^T(\cdot|S = s)$ does not depend on ϑ, because S is sufficient for ϑ.

Hence,

$$\int_{\mathscr{S}} \left[\mathbb{P}_\vartheta^T(B|S=s) - \mathbb{P}_\vartheta^T(B)\right] \mathbb{P}_\vartheta^S(ds) =: \int_{\mathscr{S}} g_B(s)\mathbb{P}_\vartheta^S(ds) = 0,$$

where the integrand $g_B(\cdot)$ is independent of ϑ. Due to bounded completeness of S, we conclude that $\mathbb{P}_\vartheta^T(B|S=s) = \mathbb{P}_\vartheta^T(B)$ for almost all $s \in \mathscr{S}$, thus $S \perp\!\!\!\perp T$.

Corollary 4.19 *Let $Y = (Y_1, \ldots, Y_n)^\top$ be a vector of real-valued i.i.d. random variables with absolutely continuous distribution $P = \mathbb{P}^{Y_1}$. Then, the following assertions hold true.*

(a) The random vectors $R(Y)$ and $(Y_{i:n})_{1\le i\le n}$ are stochastically independent.
(b) Let $T : \mathbb{R}^n \to \mathbb{R}$ denote a mapping such that the statistic $T(Y)$ is integrable. Then, for any $\sigma = (r_1, \ldots, r_n)^\top \in \mathscr{S}_n$, it holds that

$$\mathbb{E}[T(Y) \,\big|\, R(Y) = \sigma] = \mathbb{E}[T((Y_{r_i:n})_{1\le i\le n})]. \tag{4.7}$$

Proof Due to part (b) of Lemma 4.13, the statistic $R(Y)$ is pivotal. Furthermore, Theorem 4.16 yields that $(Y_{i:n})_{1\le i\le n}$ is sufficient for $P^{\otimes n}$, and it is easy to verify that it is also boundedly complete. Thus, the assertion of part (a) can be concluded by applying Basu's Theorem. Part (b) is an immediate consequence of part (a) by virtue of part c).(ii) of Theorem 1.24.

Now we are ready to apply Theorem 4.11 to vectors of rank statistics.

Corollary 4.20 (cf. Theorem 4.11, Corollary 4.19 and Lemma 4.5) *Let $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ with $\Theta \subseteq \mathbb{R}$ denote an $L_1(\mu)$-differentiable family with score function $\dot{L}$ in $\vartheta_0 = 0$. Let $Y = (Y_1, \ldots, Y_n)^\top$ denote a sample from $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^n P_{c_i\vartheta}$. Then, $\mathbb{P}_{n,\vartheta}^R$ has score function*

$$\begin{aligned}
\sigma = (r_1, \ldots, r_n)^\top \mapsto \mathbb{E}_{n,0}\left[\sum_{i=1}^n c_i \dot{L}(Y_i) \,\big|\, R(Y) = \sigma\right] \\
&= \sum_{i=1}^n c_i \mathbb{E}_{n,0}[\dot{L}(Y_i) \,\big|\, R(Y) = \sigma] \\
&= \sum_{i=1}^n c_i \mathbb{E}_{n,0}[\dot{L}(Y_{r_i:n})] = \sum_{i=1}^n c_i a(r_i)
\end{aligned}$$

with $\mathbb{E}_{n,0}$ denoting the expectation with respect to $\mathbb{P}_{n,0}$ and scores $a(i) := \mathbb{E}_{n,0}[\dot{L}(Y_{i:n})]$.

Remark 4.21

(a) The test statistic

$$T(Y) = \sum_{i=1}^{n} c_i a(R_i(Y)) \tag{4.8}$$

is called a linear rank statistic.

(b) The homogeneity hypothesis H_0 from (4.3) leads under conditioning on $R(Y)$ to a simple null hypothesis on $\mathscr{S}_n$, namely, the discrete uniform distribution on $\mathscr{S}_n$, see part (b) of Lemma 4.13. Therefore, the critical value $c(\alpha)$ for the rank test $\varphi = \varphi(R(Y))$, given by

$$\varphi(y) = \begin{cases} 1, & \text{if } T(y) > c(\alpha), \\ \gamma, & \text{if } T(y) = c(\alpha), \\ 0, & \text{if } T(y) < c(\alpha), \end{cases} \tag{4.9}$$

can be computed by traversing all possible permutations $\sigma \in \mathscr{S}_n$ and thereby determining the discrete permutation distribution of $T(Y)$ under H_0. The test φ from (4.9) is a locally best rank test at level α for the test problem $\{\vartheta = 0\}$ versus $\{\vartheta > 0\}$ in the chosen auxiliary parametric model $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^{n} P_{c_i\vartheta}$, cf. Example 4.10. For large n, we can approximate $c(\alpha)$ by traversing only $B < n!$ randomly chosen permutations $\sigma \in \mathscr{S}_n$.

(c) For the scores $a(i) = \mathbb{E}_{n,0}[\dot{L}(Y_{i:n})]$, it holds that $\sum_{i=1}^{n} a(i) = 0$, because $\dot{L}$ is centered under $\vartheta = 0$; cf. (4.1). If $\dot{L}$ is isotone, then the scores fulfill $a(1) \le a(2) \le \ldots \le a(n)$.

(d) Due to the relation $Y_{i:n} \overset{\mathscr{D}}{=} F^{-1}(U_{i:n})$, the scores are often given in the form $a(i) = \mathbb{E}[\dot{L} \circ F^{-1}(U_{i:n})]$ and the function $\dot{L} \circ F^{-1}$ is called *score-generating function*. For large n, one can approximately work with $b(i) = \dot{L} \circ F^{-1}(\frac{i}{n+1})$ (since $\mathbb{E}[U_{i:n}] = 1/(n+1)$, see Remark 4.14) or with $\tilde{b}(i) = n \int_{(i-1)/n}^{i/n} \dot{L} \circ F^{-1}(u)du$ instead of $a(i)$.

If the sample size n is large, an alternative method to approximate the critical value $c(\alpha)$ from (4.9) is a normal approximation of the null distribution of $T(Y)$. To this end, it is helpful to compute the first two moments of $T(Y)$ under H_0.

Lemma 4.22 *Let $T = T(Y) = \sum_{i=1}^{n} c_i a(R_i(Y))$ denote a linear rank statistic of the form given in (4.8), but with general deterministic scores $(a(i) : 1 \le i \le n)$ which do not necessarily sum up to zero. Let $\bar{c} := n^{-1} \sum_{i=1}^{n} c_i$ and $\bar{a} = n^{-1} \sum_{i=1}^{n} a(i)$. Then, it holds under H_0 from (4.3) that*

$$\mathbb{E}\big[T\big] = n\,\bar{c}\,\bar{a} \quad \textit{and} \quad \textit{Var}\,(T) = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (a(i) - \bar{a})^2.$$

Proof For any $1 \leq i \leq n$, the rank $R_i(Y)$ is under H_0 uniformly distributed on $\{1, \ldots, n\}$. Thus, we have that $\mathbb{E}\big[a(R_i(Y))\big] = \sum_{i=1}^n a(i) n^{-1} = \bar{a}$ and, consequently,

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

For computing the null variance of T, notice that $\sum_{i=1}^n a(i) = n\bar{a}$ is a constant, because the scores are deterministic. Hence, letting $R_i := R_i(Y)$ for all $1 \leq i \leq n$, we have that

$$\begin{aligned} 0 &= \mathrm{Var}\left(\sum_{i=1}^n a(i)\right) = \mathrm{Var}\left(\sum_{i=1}^n a(R_i)\right) \\ &= \sum_{i=1}^n \mathrm{Var}\left(a(R_i)\right) + 2 \sum_{1 \leq i < j \leq n} \mathrm{Cov}\left(a(R_i), a(R_j)\right). \end{aligned} \tag{4.10}$$

Due to exchangeability under H_0, it holds that $\mathbb{P}^{R_i} = \mathbb{P}^{R_1}$ for all $1 \leq i \leq n$ as well as $\mathbb{P}^{R_i, R_j} = \mathbb{P}^{R_1, R_2}$ for all $1 \leq i \neq j \leq n$. Utilizing these facts in (4.10), we get that $0 = n\mathrm{Var}\left(a(R_1)\right) + n(n-1)\mathrm{Cov}\left(a(R_1), a(R_2)\right)$ or, equivalently,

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

Furthermore, it holds that

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

Finally, the variance of T computes as

$$\begin{aligned} \mathrm{Var}(T) &= \mathrm{Var}\left(\sum_{i=1}^n c_i a(R_i(Y))\right) \\ &= \mathrm{Var}\left(a(R_1)\right) \sum_{i=1}^n c_i^2 - \frac{\mathrm{Var}\left(a(R_1)\right)}{n-1} \sum_{1 \leq i \neq j \leq n} c_i c_j \\ &= \sum_{i=1}^n \frac{(a(i) - \bar{a})^2}{n} \left[\sum_{j=1}^n c_j^2 - \frac{1}{n-1} \sum_{1 \leq i \neq j \leq n} c_i c_j\right], \end{aligned}$$

which leads after further elementary simplifications to the asserted expression for $\mathrm{Var}(T)$.

As shown by Janssen (1998), the rank test φ from (4.9) is, under certain assumptions, also an unbiased level α test for the (extended) two-sample test problem

$$H_0^* : \{F_1 \geq F_2\} \quad \text{(group 1 does not tend to larger values than group 2) versus}$$
$$H_1^* : \{F_1 < F_2\} \quad \text{(group 1 tends to larger values than group 2).}$$

Lemma 4.23 ("Stochastically Larger" Alternatives, Lemma 4.4 in Janssen (1998)) *Assume that $a(1) \leq a(2) \leq \ldots \leq a(n)$ (cf. Remark 4.21) and let $\varphi = \varphi(R(Y))$ be a rank test of the form (4.9) with $\mathbb{E}_{H_0}[\varphi] = \alpha$, where H_0 denotes the homogeneity hypothesis from (4.3). Let $Y_1, \ldots, Y_{n_1}$ be i.i.d. with cdf F_1 of Y_1 and $Y_{n_1+1}, \ldots, Y_n$ i.i.d. with cdf F_2 of Y_{n_1+1}. Let $Y = (Y_1, \ldots, Y_n)^\top$. Then, the following assertions hold true.*

(a) If $F_1 \geq F_2$, then $\mathbb{E}[\varphi(R(Y))] \leq \alpha$ (meaning that φ has level α on whole H_0^*).
(b) If $F_1 < F_2$, then $\mathbb{E}[\varphi(R(Y))] \geq \alpha$ (meaning that φ is unbiased on whole H_1^*).

Remark 4.24 In location parameter models (cf. part (a) of Example 4.4), the score function is isotone if and only if the density f of X is strongly unimodal; see, e.g., the discussion on page 81 of Klaassen (2003) which is based on Ibragimov (1956).

Before we discuss specific examples, let us note that rank tests are invariant under strictly isotone transformations.

Lemma 4.25 *Let φ as in (4.9) denote a locally best rank test in the model $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^n P_{c_i\vartheta}$ for testing $\{\vartheta = 0\}$ versus $\{\vartheta > 0\}$, cf. Theorem 4.9 in connection with Lemma 4.5. Moreover, let $S : \mathbb{R} \to \mathbb{R}$ denote a strictly isotone function. Then, φ is also locally optimal for $\bigotimes_{i=1}^n P^S_{c_i\vartheta}$.*

Proof The assertion follows immediately from the fact that $R_i(S(Y_1), \ldots, S(Y_n)) = R_i(Y)$ for all $1 \leq i \leq n$, due to the strict isotonicity of S.

Example 4.26 (Two-Sample Rank Tests in Location Parameter Models) Throughout, we consider the location parameter model introduced in part (a) of Example 4.10 for $Y = (Y_1, \ldots, Y_n)^\top$, and we abbreviate $R_i = R_i(Y)$ for all $1 \leq i \leq n$.

(i) *Fisher–Yates test:*
Let f be the Lebesgue density of $\mathcal{N}(0,1)$. Then it holds $\dot{L}(y) = y$ and we obtain that

$$T = \sum_{i=1}^{n_1} a(R_i) \quad \text{with} \quad a(i) = \mathbb{E}[Y_{i:n}].$$

In this, $Y_{i:n}$ denotes the i-th order statistic of i.i.d. random variables $Y_1, \ldots, Y_n$ with $Y_1 \sim \mathcal{N}(0,1)$.

(ii) *Van der Waerden test:*
Let f be as in part (i). The score-generating function is given by $u \mapsto \Phi^{-1}(u)$, where Φ denotes the cdf of $\mathcal{N}(0,1)$. Following part (d) of Remark 4.21, approximate scores are given by $b(i) = \Phi^{-1}(i/(n+1))$, leading to the test statistic

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

(iii) *Wilcoxon's rank sum test:*
Let f be the density of the standard logistic distribution, given by $f(x) = \exp(-x)(1+\exp(-x))^{-2}$ with corresponding cdf $F(x) = (1+\exp(-x))^{-1}$. The score-generating function is in this case given by $u \mapsto 2u-1$, leading to the scores given by

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

These scores are an affine transformation of the identity and therefore, the test can equivalently be carried out by means of the test statistic

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

which is the sum of the ranks in the first group.

(iv) *Median test:*
The Lebesgue density of the double-exponential (Laplace) distribution is given by $f(x) = \exp(-|x|)/2$, with induced score-generating function $u \mapsto \operatorname{sgn}(\ln(2u)) = \operatorname{sgn}(2u-1)$. Approximate scores are therefore given by

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

We conclude this section with the Savage test (or log-rank test), which is an example for a scale parameter test, cf. part (b) of Example 4.4.

Example 4.27 Under the scale parameter model considered in part (b) of Example 4.4, assume that X is exponentially distributed with Lebesgue density $x \mapsto f(x) = \exp(-x)\mathbf{1}_{(0,\infty)}(x)$. Then we obtain for $y > 0$ the score function $\dot{L}$, given by

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

According to Exercise 4.9, it holds that

$$\mathbb{E}[Y_{i:n}] = \sum_{j=1}^{i} \frac{1}{n+1-j}. \tag{4.11}$$

for i.i.d. random variables $Y_1, \ldots, Y_n$ with $Y_1 \sim \text{Exp}(1)$.

Making use of (4.11), exact scores are given by

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

Since X is almost surely positive, the model $Y = \exp(\vartheta)X$ can be transformed into the location parameter model $\log(Y) = \vartheta + \log(X)$. For $X \sim \text{Exp}(1)$, it holds that $\log(X)$ possesses a reflected Gumbel distribution, satisfying

$$\mathbb{P}\big(\log(X) \le x\big) = 1 - \exp(-\exp(x)), \quad x \in \mathbb{R}.$$

4.3 Justification of Rank Tests via Statistical Functionals

This section closely follows the derivations of Janssen (1999).

From the nonparametric viewpoint, it is incoherent to derive the scores appearing in (4.8) by means of an auxiliary parametric model $\mathbb{P}_{n,\vartheta} = \bigotimes_{i=1}^{n} P_{c_i\vartheta}$. It would be more coherent to derive the test statistic by achieving optimal local power with respect to a *nonparametric criterion*, i.e., by considering a statistical functional with respect to which the two groups differ under alternatives.

To this end, let again $Y_1, \ldots, Y_n$ denote stochastically independent, real-valued observables such that $Y_1, \ldots, Y_{n_1}$ belong to group 1 with $\mathbb{P}^{Y_i} = P$ for all $1 \le i \le n_1$ and $Y_{n_1+1}, \ldots, Y_n$ belong to group 2 with $\mathbb{P}^{Y_j} = Q$ for all $n_1 + 1 \le j \le n$. Obviously, this nonparametric model depends on (P, Q) only, and we consider an appropriate set

$$\Theta \subseteq \{(P, Q) : P \text{ and } Q \text{ are probability distributions on } (\mathbb{R}, \mathscr{B}(\mathbb{R}))\}$$

of pairs of probability distributions for the mathematical formalization of the model. Let

$$\kappa : \Theta \to \mathbb{R}, \quad (P, Q) \mapsto \kappa(P, Q) \in \mathbb{R} \tag{4.12}$$

denote a real-valued binary statistical functional. Within the nonparametric approach, it is appealing to formalize the one-sided two-groups comparison problem

for a chosen κ as follows:

$$H_0 = \{(P, P) \in \Theta : \kappa(P, P) = c\} \text{ versus } H_1 = \{(P, Q) \in \Theta : \kappa(P, Q) > c\} \tag{4.13}$$

for a known value $c \in \mathbb{R}$. The functional κ formalizes a nonparametric aspect with respect to which P and Q differ under H_1. One may think of κ as some kind of quantification of "how much better" group 1 is than group 2 under H_1.

Example 4.28

(a) Let $\Xi \subseteq \{P : P \text{ is a probability distribution on } (\mathbb{R}, \mathscr{B}(\mathbb{R}))\}$ and $\kappa_1 : \Xi \to \mathbb{R}$ denote a statistical functional of one variable (which is a probability distribution). Let $\kappa : \Xi \times \Xi \to \mathbb{R}$ be defined by $\kappa(P, Q) = \kappa_1(P) - \kappa_1(Q)$. One particularly relevant example is the difference of medians, i.e., $\kappa(P, Q) = \text{median}(P) - \text{median}(Q)$ for absolutely continuous distributions P and Q. In this case, $\kappa_1(P) = \text{median}(P) = F_P^{-1}(1/2)$, where F_P denotes the cdf pertaining to P. Of course, the constant c from (4.13) equals zero here. Under alternatives, the first group has a larger median than the second group.

(b) Let $h : (0, 1) \to \mathbb{R}$ be a function fulfilling $\int_0^1 h^2(u)du < \infty$, and define

$$\kappa_h(P, Q) = \int h(F_Q(x))dP(x), \tag{4.14}$$

where $\Theta \ni (P, Q)$ is chosen such that F_Q is continuous and $\kappa_h(P, Q)$ exists in $\mathbb{R}$. Substitution yields that $c = \int_0^1 h(u)du$ here. One particularly relevant example is the identity $h = id$, leading to

$$\kappa_{id}(P, Q) = P \otimes Q\left(\{(x, y)^\top \in \mathbb{R}^2 : y \leq x\}\right) \tag{4.15}$$

and $c = 1/2$.

Let X and Y be stochastically independent, real-valued random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with $X \sim P$ and $Y \sim Q$. Notice that the right-hand side of (4.15) can then be written as

$$\begin{aligned}\mathbb{P}(Y \leq X) &= \int \mathbb{P}(Y \leq X | X = x)d\mathbb{P}^X(x) \\ &= \int \mathbb{P}(Y \leq x)d\mathbb{P}^X(x) \\ &= \int F_Q(x)dP(x) = \kappa_{id}(P, Q).\end{aligned}$$

Now, assume for the moment that we are able to define a gradient $\dot{\kappa}_{P_0} \in L_2(P_0)$ of $P \mapsto \kappa(P, P_0)$ in $P_0 \in \Xi$. Then, for fixed P_0 and in analogy to the considerations for the score function, the test statistic

$$T_n = \sum_{i=1}^{n_1} \dot{\kappa}_{P_0}(Y_i) \tag{4.16}$$

would yield some kind of locally optimal discrepancy between $\kappa(P_0, P_0) = c$ and $\kappa(P, P_0) > c$.

For functionals of expectation type, the definition of a gradient is straightforward.

Definition 4.29 (Gradient for von Mises Functionals) Let $P_0 \in \Xi$ denote a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, and $h \in L_2(P_0)$.
Let $\Xi = \{P : P \ll P_0, \int h^2 dP < \infty\}$. Define $\kappa : \Xi \to \mathbb{R}$ by

$$\kappa(P) = \int h dP. \tag{4.17}$$

Then, the (right-sided, centered) gradient $\dot{\kappa}_{P_0}$ of $P \mapsto \kappa(P)$ in P_0 is given by

$$\dot{\kappa}_{P_0} = h - \kappa(P_0). \tag{4.18}$$

Corollary 4.30 *Let P_0 with corresponding cdf F_0 denote a fixed probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let the binary statistical functional κ_h be defined as in part (b) of Example 4.28 with Q replaced by P_0, i.e.,*

$$\kappa_h(P, P_0) = \int h(F_0(x)) dP(x).$$

Furthermore, let Ξ be as in Definition 4.29, with h replaced by $h \circ F_0$. Then, the gradient $\dot{\kappa}_{h,P_0}$ of $P \mapsto \kappa_h(P, P_0)$ in P_0 is given by

$$\dot{\kappa}_{h,P_0} = \left(h - \int_0^1 h(u)du\right) \circ F_0. \tag{4.19}$$

Example 4.31 (Example 4.28 Continued)

(a) For the functional κ_{id} from part (b) of Example 4.28, we get that

$$\dot{\kappa}_{id,P_0} = F_0 - 1/2. \tag{4.20}$$

(b) Let $h := \mathbf{1}_{(1/2,1]} + \mathbf{1}_{\{1/2\}}/2$. Then we get that $\int_0^1 h(u)du = \int_{1/2}^1 du = 1/2$, hence,

$$\begin{aligned}\dot{\kappa}_{h,P_0}(y) &= \mathbf{1}_{(1/2,1]}(F_0(y)) + \mathbf{1}_{\{1/2\}}(F_0(y))/2 - \frac{1}{2} \\ &= \begin{cases} -1/2, & F_0(y) < 1/2, \\ 0, & F_0(y) = 1/2, \\ 1/2, & F_0(y) > 1/2 \end{cases} \\ &= \frac{1}{2}\operatorname{sgn}\left(F_0(y) - \frac{1}{2}\right).\end{aligned}$$

(c) Notice that $\text{median}(P_0) = F_0^{-1}(\kappa_h(P_0, P_0))$, where h is as in part (b), and the formula

$$\frac{d}{du}F_0^{-1}(u) = \frac{1}{f_0(F_0^{-1}(u))},$$

where f_0 denotes the Lebesgue density of P_0. Thus, the chain rule yields that the gradient of the functional κ from part (a) of Example 4.28, given by $P \mapsto \text{median}(P) - \text{median}(P_0)$, at P_0 is given by

$$\dot{\kappa}_{P_0}(y) = [2f_0(\text{median}(P_0))]^{-1}\operatorname{sgn}\left(F_0(y) - \frac{1}{2}\right). \tag{4.21}$$

Unfortunately, the expressions in (4.19)–(4.21) depend on the cdf F_0 which is unspecified in the nonparametric context. Hence, they cannot directly be plugged into the expression for the test statistic T_n given in (4.16). However, we can again make use of the *substitution principle* and replace F_0 by $\hat{F}_n$. In doing so, we recover the rank tests considered in Example 4.26.

Example 4.32 (Example 4.31 Continued)

(a) Replacing F_0 by $\hat{F}_n$ in (4.20), the test statistic T_n given in (4.16) becomes

$$T_n = \sum_{i=1}^{n_1} \hat{F}_n(Y_i) - \frac{n_1}{2} \text{ so that } nT_n + \frac{nn_1}{2} = \sum_{i=1}^{n_1} R_i(Y).$$

Hence, the statistical functional κ_{id} from part (b) of Example 4.28 leads to Wilcoxon's rank sum test considered in part (iii) of Example 4.26.

(b) Replacing F_0 by $\hat{F}_n$ in (4.21) and ignoring the constant factor $\left[2 f_0(\text{median}\,(P_0))\right]^{-1}$, the test statistic T_n given in (4.16) becomes

$$T_n = \sum_{i=1}^{n_1} \operatorname{sgn}\left(\hat{F}_n(Y_i) - \frac{1}{2}\right). \tag{4.22}$$

For every summand in (4.22), we obtain that

$$\operatorname{sgn}\left(\hat{F}_n(Y_i) - \frac{1}{2}\right) = \begin{cases} 1, & R_i(Y) > n/2, \\ 0, & R_i(Y) = n/2, \\ -1, & R_i(Y) < n/2. \end{cases}$$

Hence, choosing the difference of the medians as the nonparametric criterion for the comparison of the groups leads to the median test considered in part (iv) of Example 4.26.

Remark 4.33 There are obvious similarities between rank tests and permutation tests (see Sect. 1.3.2), because the critical value $c(\alpha)$ can in both cases be determined by evaluating the test statistic for all permutations $\sigma \in \mathscr{S}_n$ and determining the resulting $(1-\alpha)$-quantile of the $n!$ computed values.

In the next chapter, we generalize the theory by considering more general mappings $g(Y)$ instead of $R(Y)$, and by considering potentially random scores. This will lead to linear resampling statistics.

4.4 Exercises

Exercise 4.1 *Prove Lemma 4.5.*

Exercise 4.2 (Score Test in the Gaussian Shift Model) *Let $n \in \mathbb{N}$ and $\sigma^2 > 0$ be given numbers. Consider the statistical model $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), (\mathcal{N}(\vartheta, \sigma^2)^{\otimes n} : \vartheta \in \mathbb{R}))$, which is typically referred to as the Gaussian shift model. This means that we can observe real-valued, stochastically independent and identically normally distributed random variables $Y_1, \ldots, Y_n$, where we know the variance σ^2 of Y_1, but not its expected value $\vartheta = \mathbb{E}[Y_1]$.*

Show that under this model the score test at significance level $\alpha \in (0,1)$ for the test problem

$$H_0 : \{\vartheta = 0\} \text{ versus } H_1 : \{\vartheta > 0\} \tag{4.23}$$

coincides with the usual Z-Test φ at significance level α for (4.23), which is given by the rejection region $\{\varphi = 1\} = \{\sqrt{n}\bar{Y}_n/\sigma > \Phi^{-1}(1-\alpha)\}$. In this, $\bar{Y}_n$ denotes the arithmetic mean of $Y_1, \ldots, Y_n$, and Φ denotes the cdf of the univariate standard normal distribution.

Hint: *Compute the score function of the model and exploit the invariance of a statistical test of Neyman–Pearson type with respect to strictly isotone transformations of the corresponding test statistic.*

Exercise 4.3 *Verify formula* (4.5) *appearing in Theorem 4.11.*

Exercise 4.4 *Prove Theorem 4.16.*

Exercise 4.5 (Wilcoxon's Rank Sum Test in Practice) *Assume that* $n = 16$ *randomly chosen calves, all from the same target population and of roughly equal age, are randomly divided into two groups of equal size. The calves in the first group receive normal feed, while the calves in the second group receive a special, concentrated feed. After a fixed time span, the weight gain (in kilograms) of each of the* 16 *calves is assessed, and the following measurements are obtained.*

Group 1 (normal feed)	8.7	9.3	10.6	11.1	11.2	12.4	12.7	14.0
Group 2 (concentrated feed)	9.7	10.9	11.8	12.9	13.5	14.1	14.3	15.6

Utilize Wilcoxon's rank sum test at significance level $\alpha = 5\%$ *to test the null hypothesis that the (random) weight gain values under concentrated feed do not tend to larger values than under normal feed against the (one-sided) alternative that they do.*
Hint:
Calibrate the critical value under the homogeneity hypothesis (implying that the rank statistic of all $n = 16$ *(random) measurements is uniformly distributed on the symmetric group* $\mathscr{S}_{16}$*) approximately by means of the Monte Carlo method employing* $B = 10{,}000$ *randomly chosen permutations* $\sigma \in \mathscr{S}_{16}$.

Exercise 4.6 (Mann–Whitney U**-Test)** *Consider the statistical model of a nonparametric two-groups comparison with stochastically independent, real-valued observables. In this, the random variables* $Y_1, \ldots, Y_{n_1}$ *are identically distributed with absolutely continuous distribution* P *of* Y_1 *and* $Y_{n_1+1}, \ldots, Y_n$ *are identically distributed with absolutely continuous distribution* Q *of* Y_n. *We let* $n_2 := n - n_1$.

(a) Show that Wilcoxon's rank sum test is equivalent to the Mann–Whitney U*-test, which is based on the test statistic*

$$U_{n_1,n_2} = \left|\{(i,j) \in \{1,\ldots,n\}^2 : i \le n_1 < j, Y_i > Y_j\}\right| = \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \mathbf{1}_{(0,\infty)}(Y_i - Y_j).$$

Hint: *Show that* $U_{n_1,n_2} = \sum_{i=1}^{n_1} R_i(Y) - n_1(n_1+1)/2$.

(b) *Compute the expected value and the variance of U_{n_1,n_2} under the homogeneity hypothesis H_0 from (4.3). (In particular, the rank statistic is under H_0 uniformly distributed on the symmetric group of order n.)*
(c) *Now, consider the special case of $n_1 = 3$ and $n_2 = 5$. Compile a table with the (approximated by a computer simulation, if necessary) values $\mathbb{P}_{H_0}(U_{n_1,n_2} \leq u)$ for $u \in \{0, 1, 2, 3, \ldots, 15\}$.*

Exercise 4.7 (Recursion Formula for Wilcoxon's Rank Sum Statistic) *Let $W_{n_1,n_2} = \sum_{i=1}^{n_1} R_i(Y)$ denote Wilcoxon's rank sum statistic (see part (iii) of Example 4.26), where $n_2 = n - n_1$ denotes the number of observational units in the second group.*

(a) *Show that*

$$supp(W_{n_1,n_2}) = \left\{ \frac{n_1(n_1+1)}{2}, \ldots, \frac{n_1(n_1+2n_2+1)}{2} \right\}.$$

(b) *Let $w \in supp(W_{n_1,n_2})$ be fixed. Show that under the homogeneity hypothesis (under which the rank statistic is uniformly distributed on the symmetric group of order n) the following recursive formula for $p_{n_1,n_2}(w) = \mathbb{P}(W_{n_1,n_2} = w)$ holds true:*

$$np_{n_1,n_2}(w) = n_1 p_{n_1-1,n_2}(w-n) + n_2 p_{n_1,n_2-1}(w)$$

(c) *Tabulate the discrete distribution of $W_{3,5}$ under the homogeneity hypothesis. Compare your tabulated values with your result from part (c) of Exercise 4.6.*

Exercise 4.8 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) *Under the assumptions of Exercise 4.6, the statistic U_{n_1,n_2} possesses under the homogeneity hypothesis H_0 the same distribution as the statistic U_{n_2,n_1}.*
(b) *Under the assumptions of Exercise 4.6, the statistic U_{n_1,n_2} has support $\{0, \ldots, n_1 n_2\}$.*
(c) *Under the assumptions of Exercise 4.6, the statistic U_{n_1,n_2} has under the homogeneity hypothesis H_0 a smaller variance than Wilcoxon's rank sum statistic W_{n_1,n_2}.*
(d) *Wilcoxon's rank sum statistic W_{n_1,n_2} possesses for all $n_1, n_2 \in \mathbb{N}$ both under the homogeneity hypothesis H_0 and under the entire alternative hypothesis moments of any order, even if the random variables $Y_1, \ldots, Y_n$ are not integrable.*

Exercise 4.9 *Let $Y_1, \ldots, Y_n$ be stochastically independent and identically distributed random variables, all defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, such that Y_1 possesses the standard exponential distribution (i.e., the exponential*

distribution with intensity parameter $\lambda = 1$*). Furthermore, let* $W_1 := Y_{1:n}$ *and* $W_j := Y_{j:n} - Y_{j-1:n}$ *for* $2 \le j \le n$.

(a) *Compute the (joint) Lebesgue density of* $(Y_{1:n}, \ldots, Y_{n:n})^\top$.
(b) *Show that* $W_1, \ldots, W_n$ *are stochastically independent random variables fulfilling that* $\mathscr{L}(W_j) = Exp(n - j + 1)$ *for all* $1 \le j \le n$.
Hint: *Transformation formula for Lebesgue densities.*
(c) *Prove assertion* (4.11) *in Example 4.27.*

Exercise 4.10 *Show that utilizing* $h(u) = -\ln(1-u)$ *in part (b) of Example 4.28 leads to (an approximate version of) the log-rank test from Example 4.27.*
Hint: *Show and exploit that*

$$\sum_{j=1}^{r} \frac{1}{n+1-j} \approx \int_1^r \frac{1}{n+1-j} dj = \ln\left(\frac{n}{n+1-r}\right), \ r \ge 1.$$

Exercise 4.11 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) *Under the assumptions of part (b) of Lemma 4.13, it holds that*

$$\mathbb{P}(R_1(Y) = i, R_2(Y) = j) = \frac{1}{n^2} \text{ for all } i \ne j \in \{1, \ldots, n\}.$$

(b) *Under the Gaussian shift model considered in Exercise 4.2, the arithmetic mean* $\bar{Y}_n$ *is sufficient for* ϑ.
(c) *Under the assumptions of Corollary 4.20 and Remark 4.21, the linear rank statistic* $T = T(Y)$ *tends to larger values under alternatives (in the sense of* (1.4)*) than under the null hypothesis* H_0 *from* (4.3).
(d) *Under the assumptions of part (c) of this exercise,* $T = T(Y)$ *is asymptotically (as* $n \to \infty$*) normally distributed.*

References

Hewitt E, Stromberg K (1975) Real and abstract analysis. A modern treatment of the theory of functions of a real variable, 3rd printing. Graduate texts in mathematics, vol 25. Springer, New York, NY

Ibragimov IA (1956) On the composition of unimodal distributions. Teor Veroyatnost i Primenen 1:283–288

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

Janssen A (1999) Testing nonparametric statistical functionals with applications to rank tests. J Stat Plann Inference 81(1):71–93. https://doi.org/10.1016/S0378-3758(99)00009-9

Klaassen CAJ (2003) Asymptotically most accurate confidence intervals in the semiparametric symmetric location model. In: Mathematical statistics and applications: Festschrift for Constance van Eeden. IMS lecture notes monograph series, vol 42. Institute of Mathematical Statistics, Beachwood, OH, pp 65–84

Reiss RD (1989) Approximate distributions of order statistics. With applications to nonparametric statistics. Springer series in statistics. Springer, New York, NY
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley series in probability and mathematical statistics. Wiley, New York, NY
Witting H (1985) Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang. B. G. Teubner, Stuttgart

Chapter 5
Asymptotics of Linear Resampling Statistics

This chapter mainly follows the work of Janssen and Pauls (2003) and the Ph. D. theses of Pauls (2003) and Pauly (2009).

5.1 General Theory

Definition 5.1 Let $(\xi_{n,i})_{1\le i\le k(n)}$ denote a triangular array of random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where $n \in \mathbb{N}$ and $k(n) \in \mathbb{N}$. In many relevant examples we will choose $k(n) \equiv n$. Furthermore, let $(W_{n,i})_{1\le i\le k(n)}$ denote a triangular array of (random) weight functions defined on a further probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathbb{P}})$. We assume that $(\xi_{n,i})_{1\le i\le k(n)}$ and $(W_{n,i})_{1\le i\le k(n)}$ are stochastically independent with respect to the product measure $\mathbb{P}\otimes\tilde{\mathbb{P}}$. The weights shall fulfill the following three general assumptions.

(GA1) For all $n \in \mathbb{N}$, the random variables $W_{n,1}, \ldots, W_{n,k(n)}$ are exchangeable.

(GA2) $\max_{1\le i\le k(n)} |W_{n,i} - \overline{W}_n| \to 0$ in $\tilde{\mathbb{P}}$-probability as $n \to \infty$, where $\overline{W}_n = k(n)^{-1} \sum_{i=1}^{k(n)} W_{n,i}$.

(GA3) $\sum_{i=1}^{k(n)} (W_{n,i} - \overline{W}_n)^2 \to C \in \mathbb{R}$ in $\tilde{\mathbb{P}}$-probability as $n \to \infty$.

Then we call

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

a *linear resampling statistic* with weight functions $(W_{n,i})_{1\le i\le k(n)}$.

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_5

Remark 5.2 If $\overline{W}_n = 0$ $\tilde{\mathbb{P}}$-almost surely, then it follows that

$$T_n^* = \sqrt{k(n)} \sum_{i=1}^{k(n)} W_{n,i} \xi_{n,i} \quad \tilde{\mathbb{P}}\text{-almost surely.}$$

Example 5.3

(a) *Linear rank statistics (cf. Chap. 4):*
Let $n \in \mathbb{N}$ and $R_n(Y) = (R_{n,i}(Y))_{1 \le i \le n}$ for $Y = (Y_1, \ldots, Y_n)^\top$ be a rank vector as considered in Chap. 4. Consider scores $a_n(i) = \mathbb{E}_{\mathbb{P}_{n,0}}\big[\dot{L}(Y_{i:n})\big]$ (or $b_n(i)$ or $\tilde{b}_n(i)$ from part (d) of Remark 4.21) and regression coefficients $(c_{n,i})_{1 \le i \le n}$. Let $k(n) \equiv n$.

Then, the linear rank statistic

$$T_n(R_n(Y)) = \sum_{i=1}^{n} (c_{n,i} - \overline{c}_n) a_n(R_{n,i}(Y))$$

has the structure of a linear resampling statistic (recall part (c) of Remark 4.21). To see this, define

$$W_{n,i} := \frac{a_n(R_{n,i}(Y))}{\sqrt{n}}, \quad 1 \le i \le n \quad \text{and}$$
$$\xi_{n,i} := (c_{n,i} - \overline{c}_n), \quad 1 \le i \le n,$$

and verify (GA1)–(GA3).

A resampling procedure based on $T_n(R_n(Y))$ utilizes the uniform distribution of $R_n(Y)$ on $\mathscr{S}_n$ under the homogeneity hypothesis H_0 from (4.3). Every $\sigma \in \mathscr{S}_n$ is equally probable for $R_n(Y)$ under this homogeneity hypothesis. Based on this, the critical value for the score test (which is carried out as a rank test) is obtained by a quantile of the distribution of $(T_n(\sigma) : \sigma \in \mathscr{S}_n)$ with respect to the uniform distribution on $\mathscr{S}_n$.

(b) *Linear permutation statistics:*
Returning to the considerations in Sect. 1.3.2, we now assume that the values of $(Y_i)_{1 \le i \le n}$ are themselves trustworthy in a multi-sample problem. However, we do not assume a parametric model. Then, a reasonable linear permutation statistic is given by

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

with

$$\xi_{n,i} := \frac{Y_{n,i}}{\sqrt{k(n)}}, \ 1 \le i \le k(n), \tag{5.1}$$

regression coefficients $(c_{n,i})_{1\le i\le k(n)}$, and resulting weights $W_{n,i} := c_{n,\sigma(i)}$. In this, $\sigma \in \mathscr{S}_{k(n)}$ again denotes a random (uniformly distributed) permutation of $1, \dots, k(n)$. It is clear that these weights fulfill the general assumptions (GA1)–(GA3) if and only if the regression coefficients fulfill them.

For example, consider a two-sample problem, and let $k(n) = n$ as well as $n_2 = n - n_1$. Then, reasonable regression coefficients for the detection of location alternatives are given by

$$c_{n,i} = \sqrt{\frac{n_1 n_2}{n}} \cdot \begin{cases} \frac{1}{n_1}, & 1 \le i \le n_1, \\ -\frac{1}{n_2}, & n_1 + 1 \le i \le n. \end{cases}$$

For $\sigma = id$ we obtain the original test statistic $T_n = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{Y}_{n_1} - \overline{Y}_{n_2})$. This test statistic is also used in the pooled two-sample t-test for the comparison of the (theoretical) group means under the assumption of normally distributed data.

(c) *Bootstrap statistics:*
For one-sample problems (cf. Sect. 1.3.1), permutation methods (relying on drawings without replacement) are inadequate, as discussed at the end of Sect. 1.3. Instead, we will consider resampling methods based on drawings with replacement, namely, bootstrap methods. For example, the classical nonparametric bootstrap by Efron (1979) considers a vector $(M_{n,1}, \dots, M_{n,k(n)})^\top$ of multinomially distributed random variables, where the total sample size equals $k(n) = \sum_{i=1}^{k(n)} M_{n,i}$ and the cell probabilities are given by $p_{n,i} \equiv k(n)^{-1}$, for all $1 \le i \le k(n)$. The bootstrap weights are then given by

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

Letting $\xi_{n,i}$ as in (5.1), a linear bootstrap statistic thus takes the form

$$T_n^* = \sqrt{k(n)} \left(\sum_{i=1}^{k(n)} \frac{M_{n,i}}{k(n)} Y_{n,i} - \overline{Y}_n \right). \tag{5.2}$$

Exercise 5.1 shows that the construction of T_n^* is equivalent to the bootstrap procedure for the mean functional which has been discussed in Sect. 1.3.1.

Theorems 5.5 and 5.8 will demonstrate that Definition 5.1 is general enough to show asymptotic effectiveness of conditional (on the observed data) resampling tests with respect to an unconditional level α test based on the original test statistic T_n, for broad classes of resampling procedures. The advantage of the resampling approach is that the conditional distribution of T_n^* given the data only depends on the distribution of the weights, which is under the control of the statistician.

Definition 5.4

(a) *Conditional resampling test*
Under the assumptions of Definition 5.1, denote the cdf of the conditional distribution $\mathscr{L}(T_n^*|\xi_{n,1},\ldots,\xi_{n,k(n)})$ by F_n^*.
Let

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

denote the $(1-\alpha)$-quantile of F_n^*. Let T_n be a real-valued statistic. Then, a non-randomized conditional resampling test based on T_n and T_n^* is defined by $\varphi_{n,\alpha}^* := \mathbf{1}_{(c_n^*(\alpha),\infty)}(T_n)$.

(b) *Asymptotic equivalence of sequences of tests*
Let $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ and $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ denote two sequences of tests for the same test problem $(\mathscr{Y}^n, \mathscr{B}(\mathscr{Y}^n), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$, and let $\vartheta_0 \in H_0$. Then, $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ and $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ are called asymptotically equivalent under ϑ_0, if

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

Theorem 5.5 *Assume that (GA1)–(GA3) hold true, where w. l . o. g. $C = 1$ in (GA3). Let $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ be a sequence of tests for $(\mathscr{Y}^n, \mathscr{B}(\mathscr{Y}^n), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$, and let ϑ_0 be an arbitrary element of H_0. Assume that the sequence $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ has the following two properties.*

(T1) For every $n\in\mathbb{N}$, the decision rule of $\varphi_{n,\alpha}$ is defined via a real-valued test statistic $T_n : \mathscr{Y}^n \to \mathbb{R}$ and an (unconditional) critical value $c_n(\alpha)$, such that

$$\varphi_{n,\alpha} = \mathbf{1}_{(c_n(\alpha),\infty)}(T_n),\ \text{where } \mathbb{E}_{\vartheta_0}\big[\varphi_{n,\alpha}\big] \to \alpha,\ n\to\infty.$$

(Asymptotic unconditional level α-test based on T_n)

(T2) The test statistic T_n from (T1) converges under ϑ_0 in distribution to a random variable T. The cdf F_T of T is continuous and strictly isotone on its support $\mathrm{supp}(F_T)$.
(Unconditional convergence)

Now, let $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ be a sequence of (conditional) resampling tests for the same test problem $(\mathscr{Y}^n, \mathscr{B}(\mathscr{Y}^n), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$. Then, $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$ and $(\varphi_{n,\alpha}^*)_{n\in\mathbb{N}}$ are asymptotically equivalent under ϑ_0, if and only if

$$d(\mathscr{L}(T_n), \mathscr{L}(T_n^*|\xi_{n,1},\ldots,\xi_{n,k(n)})) \to 0 \text{ in } \mathbb{P}_{\vartheta_0}\text{-probability as } n\to\infty. \tag{5.3}$$

In (5.3), $d(\cdot,\cdot)$ denotes any metric which metrizes the weak convergence on the space of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. For instance, the Lévy metric $d_L(\cdot,\cdot)$ is defined by

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

for two cdfs F and G on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Proof A proof for the fact that the convergence in (5.3) implies the asymptotic equivalence of $(\varphi_{n,\alpha})$ and $(\varphi^*_{n,\alpha})$ can be found on page 58 of Witting and Nölle (1970). A proof for the reverse implication can be found in Pauls (2003) (see the proof of Lemma 3.4 there).

Definition 5.6 If the assertion of Theorem 5.5 holds for all $\vartheta_0 \in H_0$, then $(\varphi^*_{n,\alpha})_{n\in\mathbb{N}}$ is called *asymptotically effective* with respect to $(\varphi_{n,\alpha})_{n\in\mathbb{N}}$.

Remark 5.7

(a) The "only if" part of Theorem 5.5 is very important for the design of concrete resampling methods for practical data analysis. It implies that the resampling scheme has to be chosen carefully, such that the distributional properties of the original test statistic T_n under the null hypothesis are reproduced as closely as possible by the construction of T^*_n. For example, consider again part (b) of Example 5.3, and assume an unbalanced design, meaning that $n_1 \neq n_2$. Then, it would not be valid to choose $c_{n,i} \propto n^{-1/2} \cdot \mathbf{1}_{\{i\leq n_1\}}$. Under asymptotic Gaussianity of $T_n = \sqrt{\frac{n_1 n_2}{n_1+n_2}}(\overline{Y}_{n_1} - \overline{Y}_{n_2})$ one can easily verify that (5.3) is violated if T^*_n is based on these inappropriate regression coefficients.

Further classical counterexamples comprise bootstrapping the maximum statistic by choosing $k(n) = n$ or naive bootstrapping of correlated data (cf. Section 2.3.1 of Pauly (2009)).

(b) If T_n is a normalized sum of stochastically independent original observables and T^*_n is a linear resampling statistic, then a combination of an *unconditional central limit theorem for T_n under the null* and a *conditional central limit theorem for T^*_n given the data* is typically employed to establish asymptotic effectiveness of $\varphi^*_{n,\alpha}$ with respect to $\varphi_{n,\alpha}$ by means of Theorem 5.5.

We end this section with a rather general conditional central limit theorem for linear resampling statistics.

Theorem 5.8 (Conditional Central Limit Theorem for Linear Resampling Statistics) *Let T^*_n be a linear resampling statistic. Assume that the weights $W_{n,i}$ fulfill the general assumptions (GA1)–(GA3), where $C = 1$ w. l. o. g. in (GA3). Furthermore, assume that $\sqrt{k(n)}(W_{n,1} - \overline{W}_n) \xrightarrow{\mathscr{D}} W_1$, where W_1 is a random variable with $Var(W_1) = 1$. Finally, assume that the $\xi_{n,i}$ fulfill the following three regularity assumptions.*

(R1) $\overline{\xi}_n \to 0$ *in* $\mathbb{P}$*-probability.*

(R2) $\max_{1\leq i\leq k(n)} |\xi_{n,i} - \overline{\xi}_n| \to 0$ *in* $\mathbb{P}$*-probability.*

(R3) $\sum_{i=1}^{k(n)} \left(\xi_{n,i} - \overline{\xi}_n\right)^2 \xrightarrow{\mathscr{D}} V^2$, *where V^2 is a non-negative random variable.*

Then it holds that

$$d(\mathscr{L}(T^*_n|(\xi_{n,i})_{1\leq i\leq k(n)}), \mathscr{L}(Z)) \to 0 \text{ in } \mathbb{P}\text{-probability},$$

where Z denotes a random variable on $\Omega \times \tilde{\Omega}$ with $Z(\omega, \cdot) \sim \mathscr{N}(0, V^2(\omega))$.

Proof See Satz 3.3 of Pauly (2009); cf. also Theorem 2.1 of Janssen (2005).

Notice that the $\xi_{n,i}$ from part (a) of Example 5.3 do typically not fulfill (R1)–(R3). However, rank tests are invariant with respect to strictly isotone transformations (see Lemma 4.25). Hence, one may re-scale the $\xi_{n,i}$ in order to fulfill (R1)–(R3); see Exercise 5.2.

Remark 5.9 (Bemerkung 3.4 of Pauly (2009))

(a) The combination of conditions (R1) and (R2) is equivalent to the convergence of $\max_{1\le i\le k(n)}|\xi_{n,i}|$ to zero in $\mathbb{P}$-probability, by virtue of the triangular inequality.
(b) If $V^2 \equiv \sigma^2 > 0$ in (R3) is a positive constant, then the conditional cdf F_n^* (cf. part (a) of Definition 5.4) even converges uniformly, due to Polya's Theorem (see Satz 5.75 of Witting and Müller-Funk (1995)), meaning that

$$\sup_{y\in\mathbb{R}}|F_n^*(y) - \Phi(\frac{y}{\sigma})| \to 0 \text{ in } \mathbb{P}\text{-probability.}$$

(c) If the original observables fulfill $\mathbb{P}\left(\sum_{i=1}^{k(n)}(Y_{n,i} - \overline{Y}_n)^2 > 0\right) \to 1$ as $n \to \infty$, then one can use Studentized variates $\xi_{n,i}$ of the form

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

Obviously, they fulfill (R3) with $V^2 \equiv 1$. This leads to resampling-based analogues of Student's t-test (see Janssen 2005 for a detailed treatment).

5.2 Application to Special Resampling Procedures

5.2.1 *Multi-Sample Problems, Permutation Tests*

We return to the setup introduced in Sect. 1.3.2 and analyze permutation tests. As outlined in Remark 4.33, we will find close analogies between permutation tests and rank tests, because in both cases the resampling distribution is based on the uniform distribution on the set of all permutations of $1, \ldots, n$ for the involved weight functions. First, we consider two-sample problems.

Model 5.10 (Two-Sample Problem) *Let $(Y_i)_{1\le i\le n}$ be real-valued, stochastically independent random variables. The variates $Y_1, \ldots, Y_{n_1}$ are assumed to be i.i.d. with $Y_1 \sim F_1$ and $Y_{n_1+1}, \ldots, Y_n$ are assumed to be i.i.d. with $Y_{n_1+1} \sim F_2$. Let $n_2 := n - n_1$ and assume that $0 < n_1 < n$. The test problem of interest is given by*

$$H_0 = \{F_1 = F_2\} \text{ versus } H_1 = \{F_1 \neq F_2\}. \tag{5.4}$$

Example 5.11 (Two-Sample Problem in a Gaussian Location Parameter Model) Under the assumptions of Model 5.10, consider the special case that F_1 and F_2 are Gaussian cdfs which only differ in their means. In this case, one compares the empirical group means to test (5.4). More specifically, we define the group means by $\overline{Y}_{n_1} := n_1^{-1} \sum_{i=1}^{n_1} Y_i$ and $\overline{Y}_{n_2} := n_2^{-1} \sum_{j=n_1+1}^{n} Y_j$. The test statistic of the resulting two-sample Z-test is then given by $\tilde{T} := |\overline{Y}_{n_1} - \overline{Y}_{n_2}|$ and the test for (5.4) can easily be calibrated by noticing that $\overline{Y}_{n_1} - \overline{Y}_{n_2}$ is again normally distributed under H_0.

However, in the case of general F_1 and F_2, exact distributional results for $\tilde{T}$ are difficult to obtain. Assuming that F_1 and F_2 are continuous, we consider more general statistics of the form

$$T = \sum_{i=1}^{n} c_i \, g(Y_i) = \sum_{i=1}^{n} c_{D_i(Y)} g(Y_{i:n}) \tag{5.5}$$

for a given function $g : \mathbb{R} \to \mathbb{R}$ and real numbers $(c_i)_{1 \le i \le n}$.

The representation of T on the right-hand side of (5.5) establishes the connection to rank tests. For example, $|T|$ equals $\tilde{T}$ from Example 5.11 if we choose $g = id$, $c_i = n_1^{-1}$ for $i \le n_1$ and $c_i = -n_2^{-1}$ for $i > n_1$. Under H_0 from (5.4), the vector of antiranks $D(Y) = (D_i(Y))_{1 \le i \le n}$ and the order statistics $(Y_{i:n})_{1 \le i \le n}$ are stochastically independent, see Theorem 4.19. Due to this property, the two-sample homogeneity test based on T can be carried out as a permutation test (or as a rank test with random scores) according to the following resampling scheme.

Scheme 5.12 (Resampling Scheme for Problem (5.4)) *The following resampling scheme is appropriate for a one-sided "stochastically larger" alternative. The two-sided case is obtained by obvious modifications. Furthermore, we have to assume here that the Y_i, $1 \le i \le n$, possess absolutely continuous distributions.*

(A) Consider the order statistics $(Y_{i:n})_{1 \le i \le n}$ and regard $a(i) := g(Y_{i:n})$ as random scores.

(B) Denote by $\tilde{D} = (\tilde{D}_i)_{1 \le i \le n}$ a random vector which is uniformly distributed on the symmetric group $\mathcal{S}_n$ and let $c = c(\alpha, (Y_{i:n})_{1 \le i \le n})$ denote the $(1-\alpha)$-quantile of the discretely distributed random variable $\tilde{D} \mapsto \sum_{i=1}^{n} c_{\tilde{D}_i} a(i)$.

(C) The permutation test φ for testing (5.4) is then given by

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

where T is as in (5.5), and $\gamma \in [0, 1]$ denotes a randomization constant.

The steps (A)–(C) lead to a conditional test φ, where the critical value $c = c(\alpha, (Y_{i:n})_{1 \le i \le n})$ is calibrated conditionally to the observed order statistics $Y_{i:n}$, $1 \le i \le n$.

Remark 5.13 If we choose $g = id$ and $(c_j)_{1\le j\le n}$ as in Example 5.11, leading to $|T| = \tilde{T}$, then the test φ from Scheme 5.12 is called *Pitman's permutation test*; see Pitman (1937).

The permutation test principle can be adapted to test the more general null hypothesis

$$H_0 \colon Y_1, \dots, Y_n \quad \text{are i.i.d.} \tag{5.6}$$

In the generalized form, the $Y_j \colon 1 \le j \le n$ are not even restricted to be real-valued. The modified resampling scheme is given as follows.

Scheme 5.14 (Modified Resampling Scheme for General Permutation Tests)

(A) Consider n random variates Y_j, $1 \le j \le n$, each taking values in some space $\mathscr{Y}$, and a real-valued test statistic $T = T(Y_1, \dots, Y_n)$.
(B) In the remainder, consider permutations π with values in $\mathscr{S}_n$, which are independent of $Y_1, \dots, Y_n$.
(C) Denote by Q_0 the uniform distribution on $\mathscr{S}_n$ and let $c = c(\alpha, Y_1, \dots, Y_n)$ denote the $(1-\alpha)$-quantile of $t \mapsto Q_0(\{\pi \in \mathscr{S}_n : T(Y_{\pi(1)}, \dots, Y_{\pi(n)}) \le t\})$.
(D) The modified (conditional) permutation test $\tilde{\varphi}$ for testing (5.6) is then given by

$$\tilde{\varphi} = \begin{cases} 1, & T > c, \\ \gamma, & T = c, \\ 0, & T < c. \end{cases}$$

Theorem 5.15 *Under the respective assumptions, the permutation test φ defined in Scheme 5.12 and the modified permutation test $\tilde{\varphi}$ defined in Scheme 5.14 are under the null hypothesis H_0 from (5.4) or (5.6), respectively, tests of* exact level α *for any fixed $n \in \mathbb{N}$.*

Proof Conditionally to the order statistics (Scheme 5.12) or to the data themselves (Scheme 5.14), the critical value c and the randomization constant γ are chosen such that

$$\mathbb{E}_{\mathscr{L}(\tilde{D})}[\varphi \,|\, Y = y] = \mathbb{E}_{Q_0}[\tilde{\varphi} \,|\, Y = y] = \alpha$$

holds true. Furthermore, the antiranks $D(Y)$ are under H_0 from (5.4) stochastically independent of the order statistics. Analogously, the random permutations π are chosen stochastically independent of $Y_1, \dots, Y_n$ in the case of $\tilde{\varphi}$. The result of the theorem follows by averaging with respect to the distribution of Y and exploiting part c) of Theorem 1.24.

In principle, Theorem 5.15 provides a very satisfying assertion for the behavior of permutation tests under the null hypothesis of distributional homogeneity. For small values of n it is possible to calculate (conditional) critical values explicitly

by traversing all $n!$ possible permutations. For moderate values of n, one can approximate the critical value by traversing $B < n!$ randomly chosen permutations (Monte Carlo-variant of the permutation test). For large values of n, however, one may also consider a normal approximation if the test statistic has the form of an (appropriately normalized) sum. In order to establish asymptotic effectiveness of the latter approach, we need unconditional (under H_0) and conditional (given the data) central limit theorems, cf. part (b) of Remark 5.7.

Assumptions 5.16 *Let $(Y_i)_{i\geq 1}$ be i.i.d. random variates with $Y_1 : (\Omega, \mathscr{F}, \mathbb{P}) \to \mathscr{Y}$, and let $g : \mathscr{Y} \to \mathbb{R}$ a measurable mapping fulfilling*

$$\int g^2(Y_1) d\mathbb{P} < \infty. \tag{5.7}$$

Assume that regression coefficients $(c_{ni})_{1\leq i\leq n}$ are given such that the following four conditions are fulfilled.

$$\forall n \in \mathbb{N} : \sum_{i=1}^{n} c_{ni} = 0. \tag{5.8}$$

$$\lim_{n\to\infty} \sum_{i=1}^{n} c_{ni}^2 = c^2 > 0. \tag{5.9}$$

$$\forall \varepsilon > 0 \, \exists M = M(\varepsilon) > 0 : \forall n \in \mathbb{N} : \sum_{i=1}^{n} c_{ni}^2 \mathbf{1}_{[M,\infty)}(|\sqrt{n} c_{ni}|) \leq \varepsilon. \tag{5.10}$$

$$\sigma^2 := c^2 \int \{g(Y_1) - \mathbb{E}[g(Y_1)]\}^2 d\mathbb{P} > 0. \tag{5.11}$$

Theorem 5.17 *Under Assumptions 5.16, let $T_n = \sum_{i=1}^{n} c_{ni} g(Y_i)$.*

(a) It holds that $\mathscr{L}(T_n) \xrightarrow{w} \mathscr{N}(0, \sigma^2)$ for $n \to \infty$.

(b) Let $\tau_n = (\tau_{ni})_{1\leq i\leq n} : (\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathbb{P}}) \to \mathscr{S}_n$ denote a uniformly distributed random variate taking its values in $\mathscr{S}_n$, where τ_n is stochastically independent of $(Y_i)_{i\geq 1}$. For fixed $\omega \in \Omega$ let $F_{n,\omega}(\cdot)$ denote the cdf of $\tilde{\omega} \mapsto T_n((Y_{\tau_{ni}(\tilde{\omega})}(\omega))_{1\leq i\leq n})$. Then it holds that

$$\sup_{t\in\mathbb{R}} |F_{n,\cdot}(t) - \Phi(\frac{t}{\sigma})| \to 0 \text{ in } \mathbb{P}\text{-probability.}$$

Proof For proving part (a), we apply the central limit theorem of Lindeberg-Feller, noticing that assumption (5.10) yields uniform integrability of the summands of T_n.

For proving part (b), we apply Theorem 5.8 in connection with Remark 5.9. To this end, re-write the resampling statistic for fixed $\tau_n = \pi = (\pi(1), \ldots, \pi(n))$ in the form

$$T_n((Y_{\pi(j)})_{1\le j\le n}) = \sum_{i=1}^n c_{ni}\, g(Y_{\pi(i)}) = \sqrt{n}\sum_{i=1}^n c_{n,\pi^{-1}(i)}\frac{g(Y_i)}{\sqrt{n}}$$

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

where

$$\xi_{n,i} := c\frac{g(Y_i)}{\sqrt{n}} \quad \text{and} \quad W_{n,i} := \frac{c_{n,\pi^{-1}(i)}}{c}.$$

In this, notice that we may assume w. l. o. g. that the $g(Y_i)$, $1 \le i \le n$, are centered at their arithmetic mean, cf. assumption (5.8) in connection with Remark 5.2.

It remains to check all assumptions of Theorem 5.8. The regularity assumption $\max_{1\le i\le n}|\xi_{n,i}| \xrightarrow{\mathbb{P}} 0$ from part (a) of Remark 5.9 is obviously fulfilled. Validity of (R3) with $V^2 \equiv \sigma^2 > 0$ follows from assumption (5.11). Assumption (GA1) is fulfilled, because $\tau_n = \pi$ is uniformly distributed on $\mathscr{S}_n$. Validity of (GA3) with $C = 1$ follows from assumptions (5.8) and (5.9). It remains to check (GA2), i.e., $\max_{1\le i\le n} c_{ni} \xrightarrow{\tilde{\mathbb{P}}} 0$ as $n \to \infty$. To this end, we argue as follows. Since the $c_{n,\pi^{-1}(i)}$ are exchangeable and $\lim_{n\to\infty}\sum_{i=1}^n c_{ni}^2 = c^2$ (which is a constant) due to (5.9), it must hold that $c_{n,\pi^{-1}(i)} = O_{\tilde{\mathbb{P}}}\left(\frac{1}{\sqrt{n}}\right)$ for all indices i, for eventually all large n.

Remark 5.18 For the application of Theorem 5.17 in practice, the permutation variance of $\sum_{i=1}^n c_{ni}\, g(Y_{\tau_{ni}})$ (i.e., its conditional variance given the data $Y = y$ with respect to the distribution of τ_n) is needed. Assuming w. l. o. g. that $\mathbb{E}[g(Y_1)] = 0$, this permutation variance is computed as follows:

$$\mathrm{Var}\left(\sum_{i=1}^n c_{ni}\, g(Y_{\tau_{ni}})|Y = y\right) = \mathrm{Var}\left(\sum_{i=1}^n c_{n,\tau_{ni}^{-1}} g(y_i)\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^n (c_{ni} - \overline{c})^2 \cdot \sum_{i=1}^n \left[g(y_i) - n^{-1}\sum_{j=1}^n g(y_j)\right]^2,$$

where the first equality follows from part c) of Theorem 1.24, and the second equality is obtained as in the proof of Lemma 4.22.

5.2.2 One-Sample Problems, Bootstrap Tests

Here, we return to testing (linear) statistical functionals in one-sample problems (cf. Sect. 1.3.1).

Model 5.19 *Let $Y_1, \ldots, Y_n$ be stochastically independent and identically distributed random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Y_1 takes values in $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$. Let $g : \mathscr{Y} \to \mathbb{R}$ be a measurable mapping fulfilling $0 < \sigma^2 := Var(g(Y_1)) < \infty$ and let*

$$\kappa(\mathbb{P}^{Y_1}) = \int g(Y_1) d\mathbb{P} = \mathbb{E}[g(Y_1)]$$

be the statistical functional of interest. Denote by $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$ the empirical measure pertaining to $Y_1, \ldots, Y_n$ and by

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

the (uncorrected) sample variance of g. We abbreviate $Z_i := g(Y_i)$, $1 \le i \le n$, and $\overline{Z}_n := n^{-1} \sum_{j=1}^n Z_j = \kappa(\hat{P}_n)$.

Lemma 5.20 *Under Model 5.19, the following two assertions hold true.*

(a) $\mathscr{L}\left(\sqrt{n} \frac{\kappa(\hat{P}_n) - \kappa(\mathbb{P}^{Y_1})}{\sigma} \right) \xrightarrow{w} \mathscr{N}(0, 1)$ *as* $n \to \infty$.

(b) $\mathscr{L}\left(\sqrt{n} \frac{\kappa(\hat{P}_n) - \kappa(\mathbb{P}^{Y_1})}{\hat{\sigma}_n} \right) \xrightarrow{w} \mathscr{N}(0, 1)$ *as* $n \to \infty$.

Proof Part (a) is an application of the central limit theorem to the i.i.d. variables $Z_1, \ldots, Z_n$, and part (b) follows from part (a) in combination with Slutzky's lemma, since $\hat{\sigma}_n^2$ estimates σ^2 consistently due to the law of large numbers.

Now, assume that we are interested in testing

$$H_0 : \kappa(\mathbb{P}^{Y_1}) = \mu_0 \quad \text{versus} \quad H_1 : \kappa(\mathbb{P}^{Y_1}) \neq \mu_0 \tag{5.12}$$

for some fixed value $\mu_0 \in \mathbb{R}$. Lemma 5.20 yields that a Z-test based on $\bar{Z}_n$ is asymptotically valid. However, as argued in Sect. 1.3.1, the normal approximation of the null distribution of $\bar{Z}_n$ may be inaccurate for finite sample sizes.

We will discuss three bootstrap tests for (5.12), which are given by the resampling Schemes 5.21, 5.22, and 5.23, respectively. In this, we restrict our attention to the case of an unknown variance σ^2 which has a high relevance for practical applications.

Scheme 5.21 (Bootstrap Test)

(A) Let $Y = (Y_1, \ldots, Y_n)^\top$ *be as in Model 5.19.*
(B) Let $Y^* = (Y_1^*, \ldots, Y_n^*)^\top$ *be a vector of random variables defined on some further probability space* $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$*, each taking values in* $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$*, such that* $\mathscr{L}(Y^*|Y) = (\hat{P}_n)^{\otimes n}$.
(C) Let $\hat{P}_n^* = n^{-1} \sum_{i=1}^n \delta_{Y_i^*}$, $\overline{Z}_n^* = \kappa(\hat{P}_n^*) = n^{-1} \sum_{i=1}^n g(Y_i^*)$*, and* q_β *be the* β*-quantile of the conditional cdf* $z \mapsto \mathbb{P}^*(\overline{Z}_n^* - \overline{Z}_n \le z|Y)$.
(D) Reject H_0*, if and only if* $\overline{Z}_n \notin [\mu_0 + q_{\alpha/2}, \mu_0 + q_{1-\alpha/2}]$.

For reasonable sample sizes n, the exact determination of the critical values $q_{\alpha/2}$ and $q_{1-\alpha/2}$ will often be computationally too demanding in practice. Therefore, a Monte Carlo variant of Scheme 5.21 can be performed as follows.

Scheme 5.22 (Monte Carlo Bootstrap, cf. Efron (1977, 1979))

(A) Let $Y = (Y_1, \ldots, Y_n)^\top$ *be as in Model 5.19. Compute from this sample the transformed sample* $Z = (Z_1, \ldots, Z_n)^\top = (g(Y_1), \ldots, g(Y_n))^\top$.
(B) Fix a number $B \in \mathbb{N}$ *of Monte Carlo repetitions, ideally such that* $(1-\alpha) \cdot B$ *is an integer. Generate* B *bootstrap pseudo-samples* $\left((Z_{b,1}^*, \ldots, Z_{b,n}^*)\right)_{b=1,\ldots,B}$. *In this, all* $Z_{b,j}^*$ *for* $1 \le b \le B$ *und* $1 \le j \le n$ *are drawn independently and uniformly with replacement from the original sample units* $Z_1, \ldots, Z_n$.
(C) Compute the bootstrap test statistics

$$T_{n,b}^* = \sqrt{n} \left| \frac{\overline{Z}_{n,b}^* - \overline{Z}_n}{\hat{\sigma}_n} \right| \quad \text{for } 1 \le b \le B.$$

(D) Reject H_0*, if and only if* $\sqrt{n} \left| \dfrac{\overline{Z}_n - \mu_0}{\hat{\sigma}_n} \right|$ *exceeds the* $\{(1-\alpha) \cdot B\}$*-th order statistic of the vector* $(T_{n,b}^*)_{b=1,\ldots,B}$.

Scheme 5.22 can be improved by including the Studentization into the resampling mechanism, leading to Scheme 5.23 which is recommended for practical data analysis.

Scheme 5.23 (Improved Resampling Scheme (See, e.g., Hall and Wilson 1991))
Proceed as in Scheme 5.22, but replace in step (C) the original estimate $\hat{\sigma}_n$ *of the standard deviation by its resampling counterpart* $\hat{\sigma}_{n,b}^*$*, given by* $\hat{\sigma}_{n,b}^* = \sqrt{n^{-1} \sum_{j=1}^n (Z_{b,j}^* - \overline{Z}_{n,b}^*)^2}$.

According to the guidelines by Hall and Wilson (1991), Scheme 5.23 should be preferred over Scheme 5.22. Theoretical justifications for this guideline regarding the speed of convergence have been established by Hall (1988); cf. also Singh (1981).

In the remainder of this section, we will utilize the conditional central limit theorem for general linear resampling statistics (i.e., Theorem 5.8) in order to

establish the asymptotic effectiveness of the bootstrap test defined by Scheme 5.21 with respect to the Z-test based on $\overline{Z}_n$. To this end, define $Z_i^* = g(Y_i^*)$,

$$\xi_i = \frac{Z_i - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n} \text{ and } \xi_i^* = \frac{Z_i^* - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n}, \text{ for all } 1 \le i \le n.$$

This leads to the (Studentized) resampling statistic

$$T_n^* := \sqrt{n}\frac{\overline{Z}_n^* - \overline{Z}_n}{\hat{\sigma}_n} = \sum_{j=1}^n \left(\xi_j^* - \overline{\xi}_n\right) = \sum_{j=1}^n \xi_j^* - n\overline{\xi}_n, \tag{5.13}$$

because

$$\begin{aligned}\sum_{j=1}^n \left(\xi_j^* - \overline{\xi}_n\right) &= \sum_{j=1}^n \left(\frac{Z_j^* - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n} - \frac{1}{n}\sum_{i=1}^n \frac{Z_i - \overline{Z}_n}{\sqrt{n}\hat{\sigma}_n}\right) \\ &= \frac{1}{\sqrt{n}\hat{\sigma}_n}\sum_{j=1}^n \left(Z_j^* - \overline{Z}_n - \overline{Z}_n + \overline{Z}_n\right) \\ &= \frac{1}{\sqrt{n}\hat{\sigma}_n}\left(n\overline{Z}_n^* - n\overline{Z}_n\right) = \sqrt{n}\frac{\overline{Z}_n^* - \overline{Z}_n}{\hat{\sigma}_n}.\end{aligned}$$

Now, we analyze the sum $\sum_{j=1}^n \xi_j^*$. For fixed $\omega \in \Omega$ and fixed $1 \le i \le n$, we observe that the value $y_i = Y_i(\omega)$ will exactly $m_{n,i}$ times be used in this sum, where the counts $(m_{n,i})_{1\le i\le n}$ can be regarded as realizations of a multinomially distributed random vector $M_n = (M_{n,1}, \ldots, M_{n,n})^\top$, where the total sample size in this multinomial distribution equals $n = \sum_{i=1}^n M_{n,i}$, and the cell probabilities are given by $p_{n,i} \equiv n^{-1}$ for all $1 \le i \le n$. This leads to the stochastic representation $\sum_{j=1}^n \xi_j^* = \sum_{j=1}^n M_{n,j}\xi_j$. Substituting the latter representation in (5.13), we obtain the linear resampling statistic

$$T_n^* = \sum_{j=1}^n M_{n,j}\xi_j - \sum_{i=1}^n \xi_i = \sum_{j=1}^n (M_{n,j} - 1)\xi_j = \sqrt{n}\sum_{j=1}^n W_{n,j}(\xi_j - \overline{\xi}_n)$$

with weights $W_{n,j} = n^{-1/2}(M_{n,j} - 1)$ for all $1 \le j \le n$.

Parts (a) and (c) of Remark 5.9 immediately yield that (R1)–(R3) are fulfilled for $(\xi_i)_{1\le i\le n}$, where $V^2 \equiv 1$ in (R3). It remains to verify the general assumptions for the weights. To this end, consider the following lemma.

Lemma 5.24 (Lemma 20.2 in Janssen (1998)) *Let $M = (M_1, \ldots, M_n)^\top$ be a multinomially distributed random vector, where the total sample size in this multinomial distribution equals $n = \sum_{i=1}^n M_i$, and the cell probabilities are given by $p_i \equiv n^{-1}$ for all $1 \le i \le n$. Then, the following assertions hold true.*

(a) $\forall 1 \le i \le n$: $M_i \overset{\mathscr{D}}{=} \sum_{k=1}^{n} \mathbf{1}_{\{i\}}(\zeta_k)$, *where* $\zeta_1, \ldots, \zeta_n$ *are i.i.d. and uniformly distributed on* $\{1, \ldots, n\}$.
(b) $\forall 1 \le i \le n$: $\mathbb{E}[M_i] = 1$, $Var(M_i) = \frac{n-1}{n}$.
(c) $\sum_{j=1}^{n} M_j \equiv n \Rightarrow \overline{M}_n \equiv 1$.
(d) $\forall \varepsilon > 0 : \mathbb{P}\left(\sqrt{n} \max_{1 \le i \le n} |\frac{M_i}{n} - \frac{1}{n}| \ge \varepsilon\right) \to 0$ *as* $n \to \infty$.
(e) $Var\left(\sum_{j=1}^{n} \left[\frac{M_j - 1}{\sqrt{n}}\right]^2\right) = \frac{(n-1)^2}{n^3} \to 0$ *as* $n \to \infty$.

Corollary 5.25 *Let* $W_{n,j} = n^{-1/2}(M_{n,j} - 1)$, $1 \le j \le n$, *denote bootstrap weights, where* $M_n = (M_{n,1}, \ldots, M_{n,n})^\top$ *is distributed as* M *in Lemma 5.24. Then, the following assertions hold true.*

(a) *The* $W_{n,j}$ *fulfill (GA1) because of part (a) of Lemma 5.24.*
(b) *The* $W_{n,j}$ *fulfill (GA2) because of part (d) of Lemma 5.24.*
(c) *The* $W_{n,j}$ *fulfill (GA3) with* $C = 1$, *meaning that* $S_W := \sum_{j=1}^{n} (W_{n,j} - \overline{W}_n)^2 \to 1$ *in probability as* $n \to \infty$, *because of the following argumentation. First, we have that*

$$S_W = \sum_{j=1}^{n} W_{n,j}^2 = \sum_{j=1}^{n} \frac{(M_{n,j} - 1)^2}{n},$$

because $\overline{W}_n = 0$. *According to part (b) of Lemma 5.24, we have* $\mathbb{E}[S_W] = Var\left(M_{n,1}\right) = (n-1)/n \to 1$ *as* $n \to \infty$. *Moreover,* $Var(S_W) \to 0$ *as* $n \to \infty$ *due to part (e) of Lemma 5.24.*
(d) *Finally, it holds* $\sqrt{n}(W_{n,1} - \overline{W}_n) = M_{n,1} - 1$ *and* $Var\left(M_{n,1}\right) \to 1$ *as* $n \to \infty$.

Combining Corollary 5.25 and Theorem 5.8, we have that

$$\mathscr{L}\left(\sqrt{n} \frac{\overline{Z}_n^* - \overline{Z}_n}{\hat{\sigma}_n} \,\middle|\, Y\right) \overset{w}{\to} \mathscr{N}(0, 1)$$

in probability. On the other hand, we have unconditional weak convergence of $\mathscr{L}\left(\sqrt{n} \frac{\overline{Z}_n - \mu_0}{\hat{\sigma}_n}\right)$ to $\mathscr{N}(0, 1)$ under H_0 as $n \to \infty$; see Lemma 5.20. Hence, Theorem 5.5 yields asymptotic effectiveness of the bootstrap test from Scheme 5.21 with respect to the Z-test based on $\overline{Z}_n$. The Schemes 5.22 and 5.23 can be analyzed analogously.

5.3 Non-exchangeability, Studentization

We have seen in Theorem 5.15 that exchangeability (i.e., distributional invariance with respect to permutations) of all n observational units under the null hypothesis is sufficient for a permutation test to keep the significance level α exactly for any

finite sample size. However, the properties of permutation tests may also be analyzed if this invariance assumption is violated (cf. Romano 1990). For concreteness, consider again the "pooled t-type" test statistic

$$T_n = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{Y}_{n_1} - \overline{Y}_{n_2}).$$

Assuming finite first moments of the observables, one may employ this test statistic in a permutation test to test the null hypothesis

$$H_0 = \{\mu_1 = \mu_2\} \text{ versus } H_1 = \{\mu_1 \neq \mu_2\}, \tag{5.14}$$

where $\mu_1 = \mathbb{E}[Y_1]$ denotes the (theoretical) mean in the first group and $\mu_2 = \mathbb{E}[Y_{n_1+1}]$ denotes the mean in the second group. Notice that the observables are not necessarily exchangeable under H_0 from (5.14), because it may well be true that the (theoretical) group means coincide, while their higher moments (e.g., the group variances) are different.

Theorem 5.26 *Under the general assumptions from Model 5.10, assume that testing H_0 from (5.14) is of interest, and that the observables possess non-trivial, finite variances $\sigma_1^2 = Var(Y_1)$ and $\sigma_2^2 = Var(Y_{n_1+1})$. We assume that n_2/n tends to $\lambda \in (0, 1)$ for $n \to \infty$. Then, the following assertions hold true.*

(a) Letting $\sigma^2(T_n|H_0)$ denote the (unconditional) sampling variance of T_n under H_0, we have that

$$\lim_{n\to\infty} \sigma^2(T_n|H_0) = \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2. \tag{5.15}$$

(b) Letting $\sigma^2(T_n^*|Y)$ denote the (conditional) permutation variance of T_n^* given the data, we have that

$$\sigma^2(T_n^*|Y) \to (1-\lambda)\sigma_1^2 + \lambda\sigma_2^2 \tag{5.16}$$

almost surely as $n \to \infty$.

(c) The right-hand sides of (5.15) and (5.16) coincide if and only if $\sigma_1^2 = \sigma_2^2$ or $\lambda = 1/2$.

Proof To prove part (a), we straightforwardly calculate that

$$\begin{aligned}\sigma^2(T_n|H_0) &= \frac{n_1 n_2}{n}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \\ &= \frac{n_2}{n}\sigma_1^2 + \frac{n_1}{n}\sigma_2^2 \\ &\to \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2, \ n \to \infty.\end{aligned}$$

For showing the validity of (5.16), we employ Remark 5.18. Noticing that the regression coefficients $(c_{n,i})_{1\le i\le n}$ given in part (b) of Example 5.3 are centered, elementary calculations yield that

$$\sigma^2(T_n^*|Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2,$$

which is the pooled sample variance. It follows from the strong law of large numbers that $\sigma^2(T_n^*|Y)$ converges almost surely to the pooled population variance, which is given by

$$\mathrm{Var}\left((1-\lambda)\mathbb{P}^{Y_1} + \lambda\mathbb{P}^{Y_{n_1+1}}\right) = (1-\lambda)\sigma_1^2 + \lambda\sigma_2^2.$$

Part (c) follows immediately from parts (a) and (b).

In view of Theorem 5.5, the permutation test based on T_n can hence only be asymptotically effective for testing H_0 from (5.14) if $\sigma_1^2 = \sigma_2^2$ (in particular, of course, in the case of exchangeability under H_0) or if $\lambda = 1/2$ (balanced sample sizes). However, there exists a technique (namely, Studentization) such that the Studentized version of the permutation test based on T_n is asymptotically effective for testing H_0 from (5.14) even if the latter assumptions may be violated.

Theorem 5.27 (Part (a) of Theorem 2.1 of Janssen (1997)) *Let the pooled sample variance V_n be defined by*

$$V_n = \frac{n_1 n_2}{n}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right),$$

where

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(Y_i - \bar{Y}_{n_1})^2 \text{ and } S_2^2 = \frac{1}{n_2-1}\sum_{j=n_1+1}^{n}(Y_j - \bar{Y}_{n_2})^2.$$

Define the Studentized variant of the test statistic T_n by $T_n^{Stud.} = T_n / V_n^{1/2}$.

Then, under the general assumptions of Model 5.10 and additionally assuming non-trivial, finite variances σ_1^2 and σ_2^2, the permutation test based on $T_n^{Stud.}$ is of asymptotic ($\min(n_1, n_2) \to \infty$) size α for all (joint) data distributions $\mathbb{P}^Y$ such that H_0 from (5.14) holds true, where $Y = (Y_1, \ldots, Y_n)^\top$.

Remark 5.28 The argumentation of Janssen (1997) is similar to the proof of Theorem 5.17 and our argumentation in Sect. 5.2.2. Namely, it is shown that both the unconditional null distribution of $T_n^{\mathrm{Stud.}}$ and the (conditional) permutation distribution of $T_n^{\mathrm{Stud.}}$ given the data are asymptotically standard normal (where the latter convergence holds in $\mathbb{P}$-probability).

5.4 Exercises

Exercise 5.1 *Show that the linear bootstrap statistic T_n^* from (5.2) is equivalent to the bootstrap method for the mean functional which has been discussed in Sect. 1.3.1.*
Hint: *Show that T_n^* is a scaled version of the difference $\bar{Y}_n^* - \bar{Y}_n$ appearing in (1.8).*

Exercise 5.2 *Consider Wilcoxon's rank sum statistic from part (iii) of Example 4.26 and regard this statistic (or its centered and scaled version, respectively) as a linear resampling statistic as described in part (a) of Example 5.3. How can one define the regression coefficients and the weights such that the assumptions of Theorem 5.8 are fulfilled?*

Exercise 5.3 (Permutation Test in Practice) *Assume that six cell cultures of the same kind are grown in Petri dishes. Three randomly chosen cultures are treated with vitamin E, while the other three cell cultures do not receive this treatment. After three weeks it is assessed how many cells per cell culture still have the ability to grow. The obtained data are summarized in the following table.*

Group 1 (Treatment with vitamin E)	121	118	110
Group 2 (No treatment with vitamin E)	34	22	12

Test at significance level $\alpha = 5\%$ the null hypothesis that the vitamin E treatment does not lead to an increase in the (random) number of cells with ability to grow after three weeks against its (one-sided) alternative that it does, by means of a permutation test. Employ as test statistic the sum of the number of cells with the ability to grow after three weeks in group 1 (treated with vitamin E).
Hint: *You do not have to explicitly traverse all $6! = 720$ possible permutations, because many of them lead to identical values of the test statistic. Thus, consider first how many and which permutations have to be traversed.*

Exercise 5.4 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) For testing the simple null hypothesis $\{F_1 = F_2\}$ against the ("two-sided") alternative $\{F_1 \neq F_2\}$ under a two-sample problem, one needs two critical values for the test statistic T_n of the permutation test considered in part (b) of Example 5.3.
(b) The weights $W_{n,i}$ from part (c) of Example 5.3 are centered for all $n \in \mathbb{N}$ and all $1 \leq i \leq k(n)$.
(c) The critical value $c_n^*(\alpha)$ from part (a) of Definition 5.4 only depends on n and α, and can hence be tabulated.
(d) For every $n \in \mathbb{N}$, let X_n be a (real-valued) random variable with distribution $\mathcal{N}(0, \sigma_n^2)$. Assume that the variances $\sigma_n^2 > 0$ converge to zero for $n \to \infty$. Denote by F_n the cdf of X_n and by G the cdf of the Dirac distribution with point mass 1 in zero. Then it holds for the Lévy metric $d_L(\cdot, \cdot)$ introduced in Theorem 5.5, that $d_L(F_n, G) \to 0$ for $n \to \infty$.

Exercise 5.5 (Generalized Weighted Bootstrap) *Let us construct a weighted bootstrap procedure or the mean functional, where we assume that $k(n) \equiv n$ for ease of notation.*

To this end, let $(\zeta_\ell)_{\ell \geq 1}$ be a sequence of real-valued, stochastically independent and identically distributed, almost surely positive random variables, such that ζ_1 possesses a finite second moment. For $n \in \mathbb{N}$ we let $M_{n,i} = \zeta_i / \sum_{j=1}^n \zeta_j$, $1 \leq i \leq n$. Finally, in analogy to part (c) of Example 5.3, we let $W_{n,i} = b_n(M_{n,i} - a_n)$ for $1 \leq i \leq n$ and real numbers a_n and b_n.

Construct sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ such that the general assumptions (GA1)–(GA3) are fulfilled for these weights.

Exercise 5.6 (Programming Exercise) *Compare, under the general assumptions of Sect. 5.3, the permutation tests based on T_n and on $T_n^{Stud.}$, respectively, with respect to the accurateness with which they keep the significance level in a computer simulation. Try out varying ratios n_1/n_2 of the group-specific sample sizes as well as varying ratios σ_1^2/σ_2^2 of the group-specific variances.*

Exercise 5.7 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) Efron's bootstrap can also be applied to two-sample problems.
(b) In the case of Pitman's permutation test from Remark 5.13, one does not have to explicitly consider those permutations of the antiranks which leave all observational units in their original groups.
(c) Scheme 5.14 remains valid in the case where there are ties among the Y_i, $1 \leq i \leq n$.
(d) Under the assumptions of Lemma 5.24, the covariance of M_i and M_j converges to zero as $n \to \infty$, for every pair $1 \leq i < j \leq n$ of indices.

References

Efron B (1977) 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. Ann Stat 7:1–26. https://doi.org/10.1214/aos/1176344552

Hall P (1988) Theoretical comparison of bootstrap confidence intervals. Ann Stat 16(3):927–953

Hall P, Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. Biometrics 47(2): 757–762

Janssen A (1997) Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. Stat Probab Lett 36(1):9–21. https://doi.org/10.1016/S0167-7152(97)00043-6

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. https://doi.org/10.1007/BF02509237

Janssen A, Pauls T (2003) How do bootstrap and permutation tests work? Ann Stat 31(3):768–806. https://doi.org/10.1214/aos/1056562462

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

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

Pitman E (1937) Significance tests which may be applied to samples from any populations. J R Stat Soc 4(1):119–130

Romano JP (1990) On the behavior of randomization tests without a group invariance assumption. J Am Stat Assoc 85(411):686–692. https://doi.org/10.2307/2290003

Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann. Stat. 9(6):1187–1195

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

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

Chapter 6
Bootstrap Methods for Linear Models

In this chapter, we employ *multivariate* central limit theorems for establishing consistency of bootstrap approximations of the distribution of estimators of vectors of regression coefficients in linear models. As argued by Freedman (1981), the choice of appropriate resampling schemes crucially depends on whether the design matrix is deterministic or random.

6.1 Deterministic Design

Model 6.1 *We consider the sample space* $(\mathscr{Y}, \mathscr{B}(\mathscr{Y})) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. *The observations* $y_1, \ldots, y_n$ *are modeled as realizations of real-valued, stochastically independent random variables* $Y_1, \ldots, Y_n$ *fulfilling*

$$\forall 1 \leq i \leq n: \quad Y_i = \sum_{k=1}^{p} \beta_k x_{i,k} + \varepsilon_i. \tag{6.1}$$

The vector $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ *is the parameter of interest. We assume that the real numbers* $(x_{i,k})_{1 \leq i \leq n, 1 \leq k \leq p}$ *are fixed and known. The random variables* $\varepsilon_1, \ldots, \varepsilon_n$ *are called error terms. They are assumed to be i.i.d. and defined on some probability space* $(\Omega, \mathscr{F}, \mathbb{P})$. *We assume that* $\mathbb{E}[\varepsilon_1] = 0$ *and that* $0 < \sigma^2 := Var(\varepsilon_1) < \infty$ *holds true. However, we do not assume a parametric family for the cdf* F *of* ε_1. *Thus,* F *is an infinite-dimensional nuisance parameter, leading*

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_6

to a semiparametric model. In abbreviated notation, we have the following model quantities.

$$
\begin{aligned}
Y(n) \equiv Y &:= (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n && \textit{response vector} \\
x(n) \equiv x &:= \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p} && \textit{design matrix} \\
\varepsilon(n) \equiv \varepsilon &:= (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n && \textit{vector of error terms} \\
\beta &= (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p && \textit{parameter vector}
\end{aligned}
$$

In matrix notation, we can express (6.1) as follows:

$$Y(n) = x(n)\beta + \varepsilon(n) \quad \textit{or} \quad Y = x\beta + \varepsilon. \tag{6.2}$$

Finally, we assume that the design matrix has full rank, such that $x^\top x \in \mathbb{R}^{p \times p}$ is positive definite and hence invertible.

Lemma 6.2 *Under Model 6.1, the least squares estimator (LSE) $\hat{\beta}$ of β is given by*

$$\hat{\beta}(n) \equiv \hat{\beta} \equiv \hat{\beta}(Y) = (x^\top x)^{-1} x^\top Y. \tag{6.3}$$

Furthermore, the estimation error can be represented as

$$\hat{\beta} - \beta = (x^\top x)^{-1} x^\top \varepsilon. \tag{6.4}$$

Proof The LSE $\hat{\beta}(n) \equiv \hat{\beta}$ is the L_2-projection of Y onto the vector space $\{z \in \mathbb{R}^n : z = x\gamma, \gamma \in \mathbb{R}^p\}$. Hence, it can be characterized by the following property.

$$
\begin{aligned}
&\forall \gamma \in \mathbb{R}^p : \langle Y - x\hat{\beta}, x\gamma \rangle_{\mathbb{R}^n} = 0 \\
\Leftrightarrow\ &\forall \gamma \in \mathbb{R}^p : (Y - x\hat{\beta})^\top x\gamma = 0 \\
\Leftrightarrow\ &\forall \gamma \in \mathbb{R}^p : Y^\top x\gamma - \hat{\beta}^\top x^\top x\gamma = 0 \\
\Leftrightarrow\ &\forall \gamma \in \mathbb{R}^p : Y^\top x\gamma = \hat{\beta}^\top x^\top x\gamma \\
\Leftrightarrow\ &Y^\top x = \hat{\beta}^\top x^\top x.
\end{aligned}
$$

Multiplication from the right by $(x^\top x)^{-1}$ yields

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

because $(x^\top x)^{-1} \in \mathbb{R}^{p \times p}$ is a symmetric matrix.

Now, substituting the right-hand side of (6.2) in the representation for $\hat{\beta}$, we obtain

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

or, equivalently,

$$\hat{\beta} - \beta = (x^\top x)^{-1} x^\top \varepsilon,$$

as desired.

Equation (6.4) is a helpful tool for the (asymptotic) analysis of least squares-based statistical inference methods for β. The following theorem yields the first two moments of $\hat{\beta}$ for any finite sample size.

Theorem 6.3 *Under Model 6.1, let* $\hat{\beta}(n) \equiv \hat{\beta} = (x^\top x)^{-1} x^\top Y$. *Then, the following two assertions hold true.*

(i) $\mathbb{E}_\beta[\hat{\beta}] = \beta$.
(ii) $Cov(\hat{\beta}) = \sigma^2 (x^\top x)^{-1}$.

Proof We exploit the representation

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

Linearity of expectation operators yields (i), because ε is centered. Furthermore, we get

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

because $\mathbb{E}\left[\varepsilon \varepsilon^\top\right] = \sigma^2 I_n$.

Our next goal is to prove an (unconditional) multivariate central limit theorem for $\hat{\beta}(n)$. To this end, we first have the following auxiliary result.

Lemma 6.4 *Let* $a^\top = (a_1, \ldots, a_p)$ *be an arbitrary, but fixed vector in* $\mathbb{R}^p$. *Under Model 6.1, assume that*

(i) $n^{-\frac{1}{2}} \max_{1 \le i \le n, 1 \le k \le p} |x_{i,k}| \to 0$ *as* $n \to \infty$,
(ii) $n^{-1} x^\top x \to V$ *as* $n \to \infty$, *for a positive definite, symmetric matrix* $V \in \mathbb{R}^{p \times p}$.

Then it holds that

$$\mathscr{L}\left(n^{-\frac{1}{2}} a^\top x^\top \varepsilon\right) \xrightarrow{w} \mathscr{N}(0, \rho^2) \text{ as } n \to \infty,$$

where $\rho^2 = \sigma^2 a^\top V a$.

Proof Let $S_n := a^\top x^\top \varepsilon$ and notice that

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

can be written as a sum of stochastically independent, centered random variables. Furthermore, we have that

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

Consequently,

$$\mathrm{Var}\left(n^{-\frac{1}{2}} S_n \right) = n^{-1} \sigma^2 a^\top x^\top x a \to \rho^2 = \sigma^2 a^\top V a \text{ as } n \to \infty.$$

It remains to check the Lindeberg condition, meaning that

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

To this end, notice that assumption (i) implies that

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

Thus, we get for all $1 \leq i \leq n$ that

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

because ε_1 possesses a finite second moment. Furthermore, as shown before, $n^{-1} \sum_{i=1}^{n} b_i^2$ converges to the fixed value $a^\top V a$ as $n \to \infty$, completing the proof.

Theorem 6.5 (Multivariate Central Limit Theorem) *Under Model 6.1, assume that assumptions (i) and (ii) from Lemma 6.4 are fulfilled. Then, letting $\hat{\beta}(n)$ be as in Theorem 6.3, we have the convergence*

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

Proof Notice that

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

Now, due to the Cramér-Wold device (see, e.g., page 862 of Shorack and Wellner (1986)), it holds that

$$\mathscr{L}\left(\frac{1}{\sqrt{n}}x^\top \varepsilon\right) \xrightarrow{w} \mathscr{N}_p\left(0, \sigma^2 V\right) \text{ as } n \to \infty.$$

Furthermore, assumption (ii) implies that $(n^{-1}x^\top x)^{-1}$ converges to V^{-1} as $n \to \infty$. Altogether, this entails that

$$\mathscr{L}\left(\frac{1}{\sqrt{n}}(n^{-1}x^\top x)^{-1}x^\top \varepsilon\right) \xrightarrow{w} \mathscr{N}_p\left(0, \sigma^2 V^{-1}\right) \text{ as } n \to \infty,$$

as desired.

Theorem 6.6 (cf. Freedman (1981)) *Under the assumptions of Theorem 6.5, the following assertions hold true.*

(a) $n^{-1}(x(n))^\top \varepsilon(n) \to 0$ *almost surely as* $n \to \infty$.
(b) $\hat{\beta}(n) \to \beta$ *almost surely as* $n \to \infty$.

Let

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top = Y - x\hat{\beta} = x(\beta - \hat{\beta}) + \varepsilon \tag{6.5}$$

denote the vector of residuals, and let $\hat{F}_n$ *denote the ecdf pertaining to* $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. *Then we have that*

(c) $\bar{\hat{\varepsilon}}_n = \int z\hat{F}_n(dz) = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i \to 0$ *almost surely as* $n \to \infty$,
(d) $\hat{\sigma}_n^2 = \int z^2\hat{F}_n(dz) = n^{-1}(\hat{\varepsilon})^\top\hat{\varepsilon} \to \sigma^2$ *almost surely as* $n \to \infty$.
(e) Due to part (c), the convergence in part (d) remains the same if the residuals are centered at $\bar{\hat{\varepsilon}}_n$.

The bootstrap procedure given in Scheme 6.7 for the approximation of $\mathscr{L}(\hat{\beta}(n))$ randomly combines the centered residuals with rows of the design matrix.

Scheme 6.7 (Bootstrap for Model 6.1)

(A) Compute the LSE $\hat{\beta}(n) = (x(n)^\top x(n))^{-1}x(n)^\top Y(n)$ *based on the originally observed response vector* $Y(n) \equiv Y = (Y_1, \ldots, Y_n)^\top$.
(B) Compute the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ *defined in (6.5), as well as their mean* $\bar{\hat{\varepsilon}}_n = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i$. *Denote by* $\tilde{F}_n$ *the ecdf of the centered residuals* $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$, *where* $\tilde{\varepsilon}_j = \hat{\varepsilon}_j - \bar{\hat{\varepsilon}}_n$, $1 \le j \le n$.

(C) *Denote by* $\varepsilon_1^*, \ldots, \varepsilon_n^*$ *a bootstrap pseudo sample which is conditionally (to* Y*) i.i.d. and such that* $\varepsilon_1^* | Y \sim \tilde{F}_n$*. Let* $Y_j^* = x_j \hat{\beta}(n) + \varepsilon_j^*$*,* $1 \leq j \leq n$*, where* x_j *denotes the* j*-th row of* x*, and* $Y^* = (Y_1^*, \ldots, Y_n^*)^\top$*.*

(D) *Let* $\hat{\beta}^*(n) = (x^\top x)^{-1} x^\top Y^*$ *and take the (conditional) distribution of* $\sqrt{n}\left(\hat{\beta}^*(n) - \hat{\beta}(n)\right)$ *given* Y *as a bootstrap approximation of the (unconditional) distribution of* $\sqrt{n}\left(\hat{\beta}(n) - \beta\right)$*.*

The consistency of the bootstrap approximation defined by Scheme 6.7 will be shown by the Conditional Multivariate Central Limit Theorem 6.10. To this end, we need two preparatory lemmas.

Lemma 6.8 *Under Model 6.1, let assumptions (i) and (ii) from Lemma 6.4 be fulfilled. Then, with the notation introduced in Scheme 6.7, the following two assertions hold true.*

(a) $n^{-1} \left\| \hat{\varepsilon} - \varepsilon \right\|^2 = n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i - \varepsilon_i)^2 \to 0$ *almost surely as* $n \to \infty$*.*

(b) $n^{-1} \left\| \tilde{\varepsilon} - \varepsilon \right\|^2 = n^{-1} \sum_{i=1}^n (\tilde{\varepsilon}_i - \varepsilon_i)^2 \to 0$ *almost surely as* $n \to \infty$*.*

Proof Representation (6.5) implies that $\hat{\varepsilon} - \varepsilon = x(\beta - \hat{\beta})$, hence

$$\left\| \hat{\varepsilon} - \varepsilon \right\|^2 = (\beta - \hat{\beta})^\top x^\top x (\beta - \hat{\beta}).$$

Now, part (b) of Theorem 6.6 yields assertion (a), because $n^{-1} x^\top x$ converges to the fixed matrix V due to assumption (ii). Assertion (b) follows by additionally utilizing part (c) of Theorem 6.6.

Lemma 6.9 *Under the assumption of Lemma 6.8, it holds that*

$$\tilde{F}_n \to F \text{ with probability } 1 \text{ as } n \to \infty,$$

where the convergence is in the sense of weak convergence of the corresponding probability measures.

Proof Let Ψ denote a bounded, Lipschitz continuous function with Lipschitz constant K. Then we have that

$$\begin{aligned} n^{-1} \sum_{i=1}^n |\Psi(\tilde{\varepsilon}_i) - \Psi(\varepsilon_i)| &\leq \frac{K}{n} \sum_{i=1}^n |\tilde{\varepsilon}_i - \varepsilon_i| \\ &\leq K \left(n^{-1} \sum_{i=1}^n (\tilde{\varepsilon}_i - \varepsilon_i)^2 \right)^{\frac{1}{2}}, \end{aligned} \tag{6.6}$$

because for any vector $a \in \mathbb{R}^n$ it holds that $\|a\|_1 \leq \sqrt{n}\|a\|_2$. Now, Part (b) of Lemma 6.8 yields that the right-hand side of (6.6) tends to zero almost surely as $n \to \infty$. Hence, we have that

$$\int \Psi(x)\tilde{F}_n(dx) - \int \Psi(x)\hat{F}_n(dx) \to 0 \text{ almost surely as } n \to \infty,$$

where $\hat{F}_n$ denotes the ecdf pertaining to $\varepsilon_1, \ldots, \varepsilon_n$. A slight variation of Vitali's Theorem (see Lemma 8.4 in Bickel and Freedman (1981)) yields the assertion.

Theorem 6.10 (Conditional Multivariate Central Limit Theorem) *Under Model 6.1, let assumptions (i) and (ii) from Lemma 6.4 be fulfilled. Then, the convergence*

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

holds with probability 1.

Proof We notice that

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

because

$$\begin{aligned}\hat{\beta}^*(n) &= (x^\top x)^{-1}x^\top Y^* \\ &= (x^\top x)^{-1}x^\top(x\hat{\beta}(n) + \varepsilon^*) \\ &= \hat{\beta}(n) + (x^\top x)^{-1}x^\top\varepsilon^*.\end{aligned}$$

Now, we can proceed as in the proofs of Lemma 6.4 and Theorem 6.5, where Lemmas 6.8 and 6.9 can be exploited in order to establish the validity of the Lindeberg condition.

6.2 Random Design

Model 6.11 (Linear Model with Random Design, Linear Correlation Model) *We consider the sample space* $(\mathbb{R}^{n(p+1)}, \mathscr{B}(\mathbb{R}^{n(p+1)}))$. *The observations are modeled as realizations of i.i.d. tuples* $(X_i, Y_i)_{1\leq i\leq n}$, *defined on some probability space* $(\Omega, \mathscr{F}, \mathbb{P})$, *where* $X_1 = x_1 \in \mathbb{R}^p$ *and* $Y_1 = y_1 \in \mathbb{R}$. *For ease of exposition, assume that both* X_1 *and* Y_1 *are centered.*

We are interested in the strength of the linear association between X_1 and Y_1. Thus, we may write

$$\forall 1 \le i \le n: \quad Y_i = \sum_{k=1}^{p} \beta_k X_{i,k} + \varepsilon_i = X_i^\top \beta + \varepsilon_i, \tag{6.7}$$

and thereby introduce "error terms" $\varepsilon_1, \ldots, \varepsilon_n$. The random matrix

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

is called the random design matrix *of the model, leading to the matrix representation*

$$Y = X\beta + \varepsilon$$

of (6.7), where we omitted the index n for ease of notation.

Under Model (6.7), we make the following assumptions.

(i) The (covariance) matrix $\Sigma = \mathbb{E}\left[X_1 X_1^\top\right] \in \mathbb{R}^{p\times p}$ is finite and positive definite.

(ii) Let μ denote the $(p+1)$-dimensional probability distribution of (X_1, Y_1). Then μ already fully specifies the distribution of ε_1, because $\varepsilon_1 = Y_1 - X_1^\top \beta$. In particular, $(\varepsilon_j)_{j=1,\ldots,n}$ are (unconditionally) i.i.d. However, it may well be true that $Var(\varepsilon_1 | X_1 = x_1) =: \sigma^2(x_1)$ is a non-constant function of $x_1 \in \mathbb{R}^p$ (heteroscedasticity).

(iii) The parameter vector $\beta = (\beta_1, \ldots, \beta_p)^\top$ is defined as the minimizer of the expected squared prediction error $\mathbb{E}\left[\left(Y_1 - X_1^\top \beta\right)^2\right]$. This implies (see, e.g., Section 5.2 in Whittaker (1990)) that $Y_1 - X_1^\top \beta = \varepsilon_1$ is perpendicular to X_1, meaning that $\forall 1 \le j \le p: \mathbb{E}[X_{1,j}\varepsilon_1] = 0$ or $Cov(\varepsilon_1, X_1) = Cov(Y_1 - \beta^\top X_1, X_1) = 0$, respectively. Hence, $\beta = \Sigma^{-1}\mathbb{E}[X_1 Y_1]$.

(iv) The matrix $M = (M_{j,k})_{1\le j,k\le p}$ with entries $M_{j,k} = \mathbb{E}[X_{1,j} X_{1,k} \varepsilon_1^2]$ exists in $\mathbb{R}^{p\times p}$. This assumption is fulfilled if $\mathbb{E}\left[\left\|(X_1^\top, Y_1)\right\|_2^4\right] < \infty$.

Lemma 6.12 *Under the assumptions of Model 6.11, the matrix*

$$n^{-1} X^\top X = n^{-1} \left(\sum_{i=1}^{n} X_{i,j} X_{i,k} \right)_{j,k=1,\ldots,p}$$

with values in $\mathbb{R}^{p\times p}$ converges $\mu^{\otimes n}$-almost surely to $\Sigma \in \mathbb{R}^{p\times p}$ as $n \to \infty$.

Proof The assertion follows from the strong law of large numbers, applied element-wise.

By the substitution principle, we define the estimator $\hat{\beta}(n) \equiv \hat{\beta}$ of $\beta \in \mathbb{R}^p$ as the minimizer of the empirical squared prediction error $n^{-1} \sum_{i=1}^n \left(Y_i - X_i^\top \beta\right)^2$. The solution is (see Sect. 6.1) the LSE $\hat{\beta}(n) \equiv \hat{\beta} = (X^\top X)^{-1} X^\top Y$. We obtain the stochastic representation

$$(X^\top X)(\hat{\beta} - \beta) = (X^\top X)[(X^\top X)^{-1} X^\top Y - \beta] = X^\top Y - (X^\top X)\beta = X^\top \varepsilon$$

of the (weighted) random estimation error.

Lemma 6.13 *Under the assumptions of Model 6.11, it holds that*

$$n^{-\frac{1}{2}}(X^\top X)(\hat{\beta} - \beta) = n^{-\frac{1}{2}} X^\top \varepsilon \xrightarrow{\mathscr{D}} \mathscr{N}_p(0, M) \text{ as } n \to \infty.$$

Proof Notice that $X^\top \varepsilon = \left(\sum_{i=1}^n X_{i,j} \varepsilon_i\right)_{j=1,\ldots,p}^\top$ is a sum of i.i.d. random vectors. Assumption (iii) from Model 6.11 entails that each of these vectors is centered, and assumption (iv) entails that each of these vectors possesses the (finite) covariance matrix M. This allows us to apply the multivariate central limit theorem for i.i.d. random vectors; see also the proof of Lemma 6.18 below for an analogous argumentation.

Combining Lemmas 6.12 and 6.13, we obtain an (unconditional) multivariate central limit theorem for $\hat{\beta}$.

Theorem 6.14 (Multivariate Central Limit Theorem) *Under the assumptions of Model 6.11, it holds that*

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

Furthermore, $\hat{\beta}(n)$ converges almost surely to β as $n \to \infty$.

Proof We exploit the representation

$$\hat{\beta}(n) = (X^\top X)^{-1} X^\top Y = \beta + (X^\top X)^{-1} X^\top \varepsilon = \beta + (n^{-1} X^\top X)^{-1} (n^{-1} X^\top \varepsilon).$$

Now, Lemma 6.12 yields that $n^{-1}(X^\top X)$ converges almost surely to Σ. Analogously, $n^{-1} X^\top \varepsilon$ converges almost surely to $0 \in \mathbb{R}^p$. Hence, the strong law of large numbers yields $\hat{\beta}(n) \to \beta$ almost surely as $n \to \infty$. The multivariate normality of $\sqrt{n}\left(\hat{\beta}(n) - \beta\right)$ follows from Lemma 6.13 in combination with Slutzky's lemma.

Since we did not make any parametric assumptions regarding the (joint) distribution $\mu^{\otimes n}$ of the data, it appears plausible to approximate $\mathscr{L}\left(\sqrt{n}\,[\hat{\beta}(n) - \beta)]\right)$ for finite n by means of the following bootstrap procedure.

Scheme 6.15 (Bootstrap for Model 6.11)

(A) Assume i.i.d. data $((X_i = x_i, Y_i = y_i))_{1\le i\le n}$ *according to Model 6.11. Denote by* $\hat\beta \equiv \hat\beta(n)$ *the LSE based on this sample, i.e.,* $\hat\beta = (X^\top X)^{-1}X^\top Y$ *with random design matrix* X *and random response vector* Y*. Furthermore, denote by* $\hat{\mathbb{P}}_n$ *the* $(p+1)$*-variate empirical distribution of the data tuples.*

(B) Let $\left((X_i^* = x_i^*, Y_i^* = y_i^*)\right)_{1\le i\le n}$ *denote a bootstrap pseudo sample, the* n *elements of which are conditionally i.i.d. given the original data. In this,* $(X_1^*, Y_1^*) : (\Omega^*, \mathscr{F}^*, \mathbb{P}^*) \to (\mathbb{R}^{p+1}, \mathscr{B}(\mathbb{R}^{p+1}))$ *with* $\mathbb{P}^{*(X_1^*, Y_1^*)|data} = \hat{\mathbb{P}}_n$ *(random uniformly distributed drawings with replacement from the originally observed data tuples).*

(C) Let the LSE on the bootstrap sample be given by $\hat\beta^*(n) \equiv \hat\beta^* = (X^{*\top}X^*)^{-1}X^{*\top}Y^*$*, and define* $\varepsilon^* := Y^* - X^*\hat\beta$ *with values in* $\mathbb{R}^n$*.*

(D) Approximate $\mathscr{L}\left(\sqrt{n}\,[\hat\beta(n) - \beta]\right)$ *by* $\mathscr{L}\left(\sqrt{n}\,[\hat\beta^*(n) - \hat\beta(n)]\,|data\right)$*.*

Stute (1990) showed the consistency of the bootstrap approximation defined by Scheme 6.15 by imitating the steps of argumentation, which have led to the unconditional Central Limit Theorem 6.14, in the bootstrap model.

Lemma 6.16 *Under Scheme 6.15, it holds that*

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

Proof Notice that taking the conditional (given the data) expectation with respect to $\mathbb{P}^*$ leads to a discrete sum with uniform weights for the originally observed data tuples. Taking into account the definition of ε^* from Scheme 6.15, we find that

$$\begin{aligned}\mathbb{E}^*\left[X_{i,j}^*\varepsilon_i^*|\text{data}\right] &= n^{-1}\sum_{k=1}^n X_{k,j}(Y_k - X_k^\top\hat\beta)\\ &= n^{-1}\langle X_j, Y - X\hat\beta\rangle_{\mathbb{R}^n} = 0\end{aligned}$$

for all $1 \le i \le n$, $1 \le j \le p$, by the construction of $\hat\beta$.

Lemma 6.17 *Under Scheme 6.15, it holds* $\mu^{\otimes n}$*-almost surely for all* $\delta > 0$ *that*

$$\mathbb{P}^*\left(\left\|n^{-1}X^{*\top}X^* - \Sigma\right\| > \delta\right) \to 0 \text{ as } n \to \infty.$$

Proof We have that $X^{*\top}X^* = \left(\sum_{i=1}^n X_{i,j}^*X_{i,k}^*\right)_{j,k=1,\dots,p}$. Now,

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

Making use of the tower equation, we conclude that

$$\mathbb{E}\left[n^{-1}\sum_{i=1}^{n} X_{i,j}^* X_{i,k}^*\right] = \mathbb{E}\left[n^{-1}\mathbb{E}^*\left[\sum_{i=1}^{n} X_{i,j}^* X_{i,k}^* | \text{data}\right]\right] = n^{-1}\mathbb{E}\left[\sum_{\ell=1}^{n} X_{\ell,j} X_{\ell,k}\right]$$
$$= \mathbb{E}\left[X_{1,j} X_{1,k}\right] = \Sigma_{j,k} < \infty$$

for all $1 \leq j, k \leq p$. Furthermore, the model fulfills the "Degenerate Convergence Criterion" (see Loève (1977), page 329), completing the proof.

Lemma 6.18 *Under Scheme 6.15, it holds $\mu^{\otimes n}$-almost surely that*

$$\forall a \in \mathbb{R}^p : n^{-\frac{1}{2}} a^\top X^{*\top} \varepsilon^* \xrightarrow{\mathscr{D}} \mathscr{N}\left(0, a^\top M a\right) \text{ as } n \to \infty.$$

Proof We introduce the abbreviation

$$S_n^* := n^{-\frac{1}{2}} a^\top X^{*\top} \varepsilon^* = n^{-\frac{1}{2}} \sum_{k=1}^{n} \sum_{j=1}^{p} a_j X_{k,j}^* \varepsilon_k^*.$$

Notice that S_n^* is a normalized sum of i.i.d. random variables

$$(Z_k^*)_{1 \leq k \leq n} := \left(\sum_{j=1}^{p} a_j X_{k,j}^* \varepsilon_k^*\right)_{1 \leq k \leq n}.$$

By Lemma 6.16, Z_1^* is (conditionally) centered, hence also S_n^*. It remains to calculate the (conditional) variance of S_n^*. We get that

$$\text{Var}^*\left(S_n^* | \text{data}\right) = \mathbb{E}^*\left[(Z_1^*)^2 | \text{data}\right] = \sum_{\ell=1}^{p}\sum_{j=1}^{p} a_\ell a_j \mathbb{E}^*\left[X_{1,\ell}^* \varepsilon_1^* X_{1,j}^* \varepsilon_1^* | \text{data}\right]$$
$$= \sum_{\ell=1}^{p}\sum_{j=1}^{p} a_\ell a_j \left[n^{-1}\sum_{i=1}^{n} X_{i,\ell} X_{i,j} (Y_i - X_i^\top \hat{\beta})^2\right].$$

Now, Theorem 6.14 yields that $\hat{\beta} = \hat{\beta}(n)$ converges almost surely to β as $n \to \infty$. This entails that, for all $1 \leq i \leq n$, the random variable $(Y_i - X_i^\top \hat{\beta})^2$ converges almost surely to ε_i^2. Exploiting assumption (iv) of Model 6.11 and the strong law of large numbers, we conclude that $n^{-1}\sum_{i=1}^{n} X_{i,\ell} X_{i,j} (Y_i - X_i^\top \hat{\beta})^2$ converges almost surely to $M_{\ell,j}$, for all $1 \leq \ell, j \leq p$. Altogether, this entails that

$$\text{Var}^*\left(S_n^* | \text{data}\right) \to \sum_{\ell=1}^{p}\sum_{j=1}^{p} a_\ell a_j M_{\ell,j} = a^\top M a$$

$\mu^{\otimes n}$-almost surely as $n \to \infty$.

Finally, the aforementioned results yield the consistency of the bootstrap approximation defined by Scheme 6.15, by means of Theorem 6.19.

Theorem 6.19 (Conditional Multivariate Central Limit Theorem) *Under Scheme 6.15, it holds $\mu^{\otimes n}$-almost surely that*

$$\mathscr{L}\left(\sqrt{n}\,[\hat{\beta}^*(n) - \hat{\beta}(n)]\,|data\right) \xrightarrow{w} \mathscr{N}_p\left(0,\, \Sigma^{-1} M \Sigma^{-1}\right) \text{ as } n \to \infty.$$

Proof Consider the representation

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

By Lemma 6.18 and the Cramér-Wold device, the conditional distribution of $n^{-\frac{1}{2}}X^{*\top}\varepsilon^*$ converges to $\mathscr{N}_p(0, M)$ for $\mu^{\otimes n}$-almost all observations. Furthermore, according to Lemma 6.17 we have stochastic convergence of $n^{-1}X^{*\top}X^*$ to the invertible matrix Σ for $\mu^{\otimes n}$-almost all observations. The assertion follows by Slutzky's lemma.

6.3 Exercises

Exercise 6.1 (Approximate Confidence Interval for a Regression Coefficient) *Consider Model 6.1 with $n = 50$, $p = 1$, $\sigma^2 = 4$ and design points $x_i \equiv x_{i,1} = i/n$, $1 \le i \le n$.*

(a) Utilize Theorem 6.5 to construct an approximate $(1 - \alpha)$-confidence interval for the regression coefficient $\beta \equiv \beta_1$, $\alpha \in (0, 1)$.

(b) Assess the relative coverage frequency of the confidence interval from part (a) of this exercise by means of a computer simulation with 5000 simulation runs, for $\alpha = 5\%$. For the error distribution $\mathbb{P}^{\varepsilon_1}$, choose the centered normal distribution with variance $\sigma^2 = 4$ and a shifted exponential distribution with expectation zero and variance $\sigma^2 = 4$, respectively.

Exercise 6.2 (Conditional Multivariate Central Limit Theorem) *Complete the proof of Theorem 6.10; i.e., carry out explicitly those steps which are in analogy to Lemma 6.4 and Theorem 6.5.*

Exercise 6.3 (Programming Exercise)

(a) Program Scheme 6.7 in R.

(b) Consider again the model from Exercise 6.1. Construct a bootstrap-based 95%-confidence interval for the regression coefficient β. Assess the relative coverage frequency of the latter confidence interval by means of a computer

simulation with 5000 *simulation runs. Choose the same error distributions as in Exercise 6.1, and compare your results with those from part (b) of Exercise 6.1.*

Exercise 6.4 (Multiple Select) *Which of the following assertions are true and which are false? Provide reasons for your answers.*

(a) Under Model 6.1, the least squares estimator $\hat{\beta}$ *for* β *coincides with the maximum likelihood estimator for* β*, if and only if* $\mathbb{P}^{\varepsilon_1} = \mathcal{N}(0, 1)$.
(b) The Cramér-Wold device can only be applied to normal distributions.
(c) Scheme 6.15 can successfully (i.e., asymptotically effectively) be applied under Model 6.1.
(d) Scheme 6.7 can successfully (i.e., asymptotically effectively) be applied under Model 6.11.

References

Bickel P, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217
Freedman DA (1981) Bootstrapping regression models. Ann Stat 9:1218–1228
Loève M (1977) Probability theory I, 4th edn. Graduate texts in mathematics, vol 45. Springer, New York, NY
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley series in probability and mathematical statistics. Wiley, New York, NY
Stute W (1990) Bootstrap of the linear correlation model. Statistics 21(3):433–436
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley Series in probability and mathematical statistics: probability and mathematical statistics. John Wiley & Sons, Chichester

Chapter 7
Projection Tests

7.1 Empirical Likelihood Ratio Tests for Vector Means

This section mainly follows Chapter 11 of Owen (2001) and Section 2 of Schennach (2007).

Let us return to the statistical motivation of utilizing the empirical measure as a nonparametric maximum likelihood estimator in i.i.d. models; cf. Definition 2.4 and Theorem 2.5. Notice first that the assumption that Y_1 is real-valued was nowhere used in the proof of Theorem 2.5, because we translated the optimization problem regarding the cdf F of Y_1 into an optimization problem regarding the probabilities $p_i = \mathbb{P}_{\hat{F}}(\{y_i\})$, $1 \leq i \leq n$, where $\hat{F}$ is some estimator (candidate) for F and $Y_1 = y_1, \ldots, Y_n = y_n$ is the (observed) i.i.d. sample. Hence, we may apply the very same reasoning to i.i.d. models where Y_1 takes values in $\mathbb{R}^d$ for some $d \in \mathbb{N}$.

In the sequel, with slight abuse of notation, we will write $Z(y, \hat{P})$ instead of $Z(y, \hat{F})$, where $\hat{P} \equiv P_{\hat{F}}$ is the probability measure induced by $\hat{F}$. Theorem 2.5 yields that the empirical measure $\hat{P}_n$ maximizes $Z(y, \cdot)$ over the space $\mathscr{P}$ of all probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. Noticing that

$$Z(y, \hat{P}_n) = \frac{1}{n^n} \text{ as well as}$$

$$Z(y, \hat{P}) = \prod_{i=1}^{n} p_i \leq Z(y, \hat{P}_n),$$

we thus obtain that the ratio

$$\mathscr{R}(\hat{P}) = \frac{Z(y, \hat{P})}{Z(y, \hat{P}_n)} = n^n \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} n p_i$$

takes values in $[0, 1]$.

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_7

Now, assume that testing some null hypothesis H_0 corresponding to a subspace $\mathscr{P}_0 \subset \mathscr{P}$ is of interest. Then, we may optimize the probabilities $(p_i : 1 \leq i \leq n)$ under the constraint that the corresponding probability measure $\hat{P}$ belongs to $\mathscr{P}_0$. We will indicate this constraint by writing $\hat{P}_0$ instead of $\hat{P}$ in such cases. Notice that, since the data are random, the empirical measure will typically not belong to $\mathscr{P}_0$, even if H_0 is true. For example, if $H_0 = \{\mathbb{E}[Y_1] = \mu_0\}$ is considered, where μ_0 is a fixed, given vector in $\mathbb{R}^d$, then $\bar{Y}_n = \int y \hat{P}_n(dy)$ will typically not be equal to μ_0, even if H_0 holds true. One may, however, weight the observations $y_1, \ldots, y_n$ with probabilities $p_1, \ldots, p_n$ such that $\sum_{i=1}^n p_i y_i = \mu_0$, at least if μ_0 happens to lie inside the convex hull of the observations, which should happen with very high probability if H_0 is true. By our argumentation from before, these probabilities will typically not all be equal to $1/n$. Hence we may consider the ratio $\mathscr{R}(\mu_0) \equiv \mathscr{R}(\hat{P}_0)$ as some kind of *empirical discrepancy* between the unconstrained optimizer $\hat{P}_n$ and the constrained optimizer $\hat{P}_0$. If this empirical discrepancy is large, we may take this as an indication that H_0 may be false. This is the general idea of *empirical likelihood ratio (ELR) tests*.

Lemma 7.1 *Under the aforementioned assumptions, maximizing $\mathscr{R}(\hat{P}_0)$ over $\hat{P}_0 \in \mathscr{P}_0$ is equivalent to minimizing the Kullback-Leibler divergence $\mathscr{K}(\hat{P}_n \| \hat{P}_0)$ over $\hat{P}_0 \in \mathscr{P}_0$.*

Proof We have that $\hat{P}_0 \ll \hat{P}_n$, meaning that $\hat{P}_0$ is a discrete probability measure which distributes its mass among the observed data points. Hence, the definition of the Kullback-Leibler divergence for discrete probability measures with common support yields that

$$\mathscr{K}(\hat{P}_n \| \hat{P}_0) = \sum_{i=1}^n \frac{1}{n} \log\left(\frac{1/n}{p_i}\right) = \frac{1}{n} \sum_{i=1}^n \log\left(\frac{1}{np_i}\right).$$

Now, minimizing $\mathscr{K}(\hat{P}_n \| \hat{P}_0)$ is equivalent to maximizing

$$-\mathscr{K}(\hat{P}_n \| \hat{P}_0) = \frac{1}{n} \sum_{i=1}^n \log(np_i) = \frac{1}{n} \log\left(\mathscr{R}(\hat{P}_0)\right). \tag{7.1}$$

Obviously, the right-hand side of (7.1) is a strictly isotone transformation of $\mathscr{R}(\hat{P}_0)$.

By Lemma 7.1, we may regard $\hat{P}_0$ as a Kullback-Leibler projection of $\hat{P}_n$ onto the subspace $\mathscr{P}_0$, and $-\log\left(\mathscr{R}(\hat{P}_0)\right)$ as the (scaled) "length of the difference in Kullback-Leibler geometry".

Remark 7.2 In general, the Kullback-Leibler divergence is not symmetric, meaning that $\mathscr{K}(\hat{P}_n \| \hat{P}_0) \neq \mathscr{K}(\hat{P}_0 \| \hat{P}_n)$ in general.

Returning to null hypotheses regarding the (theoretical) mean of Y_1, one of the most important results about empirical likelihood ratio tests is that they exhibit

a Wilks (1938)-type phenomenon, meaning that $-2\log(\mathscr{R}(\mu_0))$ is, under certain regularity assumptions, asymptotically chi-square distributed under $H_0 = \{\mathbb{E}[Y_1] = \mu_0\}$. To prove this, we need some preparatory results.

Lemma 7.3 *Let $y_1, \ldots, y_n$ with $y_i \in \mathbb{R}^d$, $1 \le i \le n$, be given points, and $\mu_0 \in \mathbb{R}^d$ a further given vector. Assume that μ_0 is located inside the convex hull*

$$\left\{ \sum_{i=1}^{n} p_i y_i : p_i \ge 0 \text{ for all } 1 \le i \le n, \sum_{i=1}^{n} p_i = 1 \right\}$$

of $y_1, \ldots, y_n$.

Then, the solution $(p_1, \ldots, p_n)^\top$ of the constrained optimization problem

$$\text{maximize} \quad \prod_{i=1}^{n} n p_i \tag{7.2}$$

$$\text{subject to} \quad \forall 1 \le i \le n : p_i \ge 0,$$

$$\sum_{i=1}^{n} p_i = 1, \tag{7.3}$$

$$\sum_{i=1}^{n} p_i (y_i - \mu_0) = 0 \in \mathbb{R}^d, \tag{7.4}$$

can be written as follows:

$$\forall 1 \le i \le n : p_i = \frac{1}{n} \cdot \frac{1}{1 + \lambda^\top (y_i - \mu_0)}, \tag{7.5}$$

where $\lambda \equiv \lambda(\mu_0) = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$ satisfies the system of equations

$$\frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \mu_0}{1 + \lambda^\top (y_i - \mu_0)} = 0 \in \mathbb{R}^d. \tag{7.6}$$

Proof We employ the method of Lagrange multipliers. To this end, we need one Lagrange multiplier, say γ, for constraint (7.3) and d Lagrange multipliers $\lambda_1, \ldots, \lambda_d$ for the d constraints imposed by (7.4). Considering the target function (7.2) on the logarithmic scale, a suitable Lagrangian function L is then given by

$$L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma) = \sum_{i=1}^{n} \log(n p_i) - n \lambda^\top \left(\sum_{i=1}^{n} p_i (y_i - \mu_0) \right) + \gamma \left(\sum_{i=1}^{n} p_i - 1 \right).$$

For any $1 \leq i \leq n$, we have that

$$\frac{\partial L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma)}{\partial p_i} = \frac{1}{p_i} - n\lambda^\top (y_i - \mu_0) + \gamma.$$

Hence,

$$\begin{aligned} &\forall 1 \leq i \leq n : \frac{\partial L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma)}{\partial p_i} = 0 \\ \iff &\forall 1 \leq i \leq n : \frac{1}{p_i} - n\lambda^\top (y_i - \mu_0) + \gamma = 0 \qquad (7.7) \\ \iff &\forall 1 \leq i \leq n : 1 - n\lambda^\top p_i (y_i - \mu_0) + p_i \gamma = 0, \qquad (7.8) \end{aligned}$$

because any maximizer of the target function (7.2) is such that $p_i > 0$ for all $1 \leq i \leq n$. Adding the left-hand and the right-hand sides of (7.8), we obtain that

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left[1 - n\lambda^\top p_i (y_i - \mu_0) + p_i \gamma \right] \\ &= n - n\lambda^\top \sum_{i=1}^{n} p_i (y_i - \mu_0) + \gamma \\ &= n + \gamma, \end{aligned}$$

hence $\gamma = -n$. Utilizing this in (7.7), we obtain for all $1 \leq i \leq n$ that

$$\begin{aligned} n &= \frac{1}{p_i} - n\lambda^\top (y_i - \mu_0) \\ \iff np_i &= 1 - np_i \lambda^\top (y_i - \mu_0) \\ \iff 1 &= np_i \left[1 + \lambda^\top (y_i - \mu_0) \right] \\ \iff p_i &= \frac{1}{n} \cdot \frac{1}{1 + \lambda^\top (y_i - \mu_0)}, \end{aligned}$$

which is the representation asserted in (7.5). Finally, plugging (7.5) into (7.4) yields (7.6).

Remark 7.4 In order to fulfill $p_i > 0$ for all $1 \leq i \leq n$, the vector $\lambda = (\lambda_1, \ldots, \lambda_d)^\top$ from (7.5) must fulfill the system of equations

$$\forall 1 \leq i \leq n : 1 + \lambda^\top (y_i - \mu_0) > 0. \qquad (7.9)$$

The set of values of λ for which (7.9) holds is an intersection of n half spaces in $\mathbb{R}^d$. It contains the value $\lambda = 0 \in \mathbb{R}^d$. Hence, it is a non-empty and convex subset of $\mathbb{R}^d$.

Next, we would like to bound the stochastic order of $\|\lambda\|$.

Definition 7.5 (Stochastic Orders; cf., e.g., Section 14.4 in Bishop et al. (2007)) Let $\{X_n\}_{n\geq 1}$ denote a sequence of real-valued random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\{b_n\}_{n\geq 1}$ denote a sequence of real numbers.

(a) We say that $X_n = o_{\mathbb{P}}(1)$, if for every $\varepsilon > 0$ we have that $\lim_{n\to\infty} \mathbb{P}(|X_n| \leq \varepsilon) = 1$. More precisely, $X_n = o_{\mathbb{P}}(1)$, if for every $\varepsilon > 0$ and every $\delta > 0$ there exists an integer $n^*(\varepsilon, \delta)$ such that $n > n^*(\varepsilon, \delta)$ implies $\mathbb{P}(|X_n| \leq \varepsilon) \geq 1 - \delta$. We call $\{X_n\}_{n\geq 1}$ a stochastic null sequence if $X_n = o_{\mathbb{P}}(1)$.
(b) We say that $X_n = o_{\mathbb{P}}(b_n)$, if $X_n/b_n = o_{\mathbb{P}}(1)$ or, equivalently, $X_n = b_n o_{\mathbb{P}}(1)$.
(c) We say that $X_n = O_{\mathbb{P}}(1)$, if for every $\delta > 0$ there exists a constant $K(\delta)$ and an integer $n^*(\delta)$ such that $n > n^*(\delta)$ implies $\mathbb{P}(|X_n| \leq K(\delta)) \geq 1 - \delta$. We call $\{X_n\}_{n\geq 1}$ stochastically bounded if $X_n = O_{\mathbb{P}}(1)$.
(d) We say that $X_n = O_{\mathbb{P}}(b_n)$, if $X_n/b_n = O_{\mathbb{P}}(1)$ or, equivalently, $X_n = b_n O_{\mathbb{P}}(1)$.

Lemma 7.6 *Let $\{\xi_n\}_{n\in\mathbb{N}}$ denote a sequence of i.i.d. real-valued random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, such that $\mathbb{E}[\xi_1^2] < \infty$.*

(a) Let $M_n = \max_{1\leq i\leq n} |\xi_i|$. Then $M_n = o(n^{1/2})$ almost surely.
(b) It holds that $n^{-1}\sum_{i=1}^n |\xi_i|^3 = o(n^{1/2})$ almost surely.

Proof For part (a), see Remark 1.6.2 of Chandra (2012).

For proving part (b), notice that $n^{-1}\sum_{i=1}^n |\xi_i|^3 \leq M_n \cdot n^{-1}\sum_{i=1}^n \xi_i^2$. The result follows by applying part (a) to M_n and the strong law of large numbers to $n^{-1}\sum_{i=1}^n \xi_i^2$.

Lemma 7.7 *Let $Y_1, \ldots, Y_n$ denote i.i.d. random vectors defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, each taking values in $\mathbb{R}^d$ for $d \in \mathbb{N}$. Assume that Y_1 possesses a finite and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Denote the empirical likelihood ratio statistic pertaining to the point null hypothesis $H_0 = \{\mathbb{E}[Y_1] = \mu_0\}$ by $\mathscr{R}(\mu_0)$, where μ_0 is a given point in $\mathbb{R}^d$. Then, considering the Lagrange multiplier λ from Lemma 7.3 as a transformation of $Y_1 = y_1, \ldots, Y_n = y_n$, it holds under H_0 that $\|\lambda\| = O_{\mathbb{P}}(n^{-1/2})$.*

Proof If $\lambda = 0 \in \mathbb{R}^d$, then the assertion is trivially true. Hence, we may assume throughout that $\lambda \neq 0$, hence $\|\lambda\| > 0$.

Write $\lambda = \|\lambda\| u$ for a unit vector u in $\mathbb{R}^d$ such that $u^\top u = 1$, and let $\xi_i = \lambda^\top (Y_i - \mu_0)$, $1 \leq i \leq n$. Let

$$g(\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{Y_i - \mu_0}{1 + \lambda^\top (Y_i - \mu_0)},$$

and notice that $g(\lambda) = 0 \in \mathbb{R}^d$ according to (7.6), implying that $\lambda^\top g(\lambda) = 0 \in \mathbb{R}$.

Now, we have that

$$0 = \lambda^\top g(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda^\top (Y_i - \mu_0)}{1 + \lambda^\top (Y_i - \mu_0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{1+\xi_i} = \frac{1}{n}\sum_{i=1}^{n} \xi_i \left(1 - \frac{\xi_i}{1+\xi_i}\right)$$

or, equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i^2}{1+\xi_i}. \tag{7.10}$$

It is straightforward to show (see Exercise 7.1) that

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i = \|\lambda\| \cdot u^\top \left(\bar{Y}_n - \mu_0\right), \tag{7.11}$$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i^2}{1+\xi_i} = \|\lambda\|^2 \cdot u^\top \tilde{S} u, \text{ where } \tilde{S} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \mu_0)(Y_i - \mu_0)^\top}{1+\xi_i}. \tag{7.12}$$

Thus, we may write (7.10) as

$$\|\lambda\| \cdot u^\top \tilde{S} u = u^\top \left(\bar{Y}_n - \mu_0\right). \tag{7.13}$$

Let

$$S = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_0)(Y_i - \mu_0)^\top.$$

From Remark 7.4 we get that $1 + \xi_i > 0$ for all $1 \le i \le n$. Thus,

$$\begin{aligned} \|\lambda\| \cdot u^\top S u &\le \|\lambda\| \cdot u^\top \tilde{S} u \left(1 + \max_{1 \le i \le n} \xi_i\right) \\ &\le \|\lambda\| \cdot u^\top \tilde{S} u \left(1 + \|\lambda\| M_n\right) \end{aligned}$$

according to the Cauchy-Schwarz inequality, where $M_n = \max_{1 \le i \le n} \|Y_i - \mu_0\|$. Utilizing (7.13), we obtain

$$\|\lambda\| \cdot u^\top S u \le u^\top \left(\bar{Y}_n - \mu_0\right) \cdot \left(1 + \|\lambda\| M_n\right)$$

or, equivalently,

$$\|\lambda\| \left(u^\top S u - M_n u^\top \left(\bar{Y}_n - \mu_0\right)\right) \le u^\top \left(\bar{Y}_n - \mu_0\right). \tag{7.14}$$

Let $\sigma_{\min} > 0$ denote the smallest eigenvalue of Σ, and $\sigma_{\max} \ge \sigma_{\min}$ the largest eigenvalue of Σ. Since $\sigma_{\min} \le u^\top \Sigma u \le \sigma_{\max}$ and S is a consistent estimator of Σ,

we conclude that

$$\sigma_{\min} + o_{\mathbb{P}}(1) \leq u^{\top} S u \leq \sigma_{\max} + o_{\mathbb{P}}(1).$$

Furthermore, due to Lemma 7.6, $M_n = o(n^{1/2})$ almost surely. Also, the central limit theorem implies that $u^{\top}\left(\bar{Y}_n - \mu_0\right) = O_{\mathbb{P}}(n^{-1/2})$. Utilizing these three bounds in (7.14), we find that

$$\|\lambda\| \left(u^{\top} S u + o_{\mathbb{P}}(1)\right) = O_{\mathbb{P}}(n^{-1/2}),$$

hence $\|\lambda\| = O_{\mathbb{P}}(n^{-1/2})$ as desired.

The following theorem is the main result of this section.

Theorem 7.8 *Under the assumptions of Lemma 7.7, the following assertions hold true.*

(a) Under H_0, the statistic $-2\log(\mathscr{R}(\mu_0))$ converges in distribution to χ^2_d as $n \to \infty$.

(b) Let $\alpha \in (0,1)$, and let $c_\alpha = \chi^2_{d;1-\alpha}$ denote the $(1-\alpha)$-quantile of χ^2_d. Then, the set

$$C_\alpha(\mu) = \left\{ \mu_0 = \sum_{i=1}^{n} p_i Y_i \,\middle|\, -2\log(\mathscr{R}(\mu_0)) \leq c_\alpha, \forall 1 \leq i \leq n : p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

constitutes an asymptotic $(1-\alpha)$-confidence region for $\mu = \mathbb{E}[Y_1]$, where $n \to \infty$.

(c) The set $C_\alpha(\mu)$ from part (b) is a convex subset of $\mathbb{R}^d$.

Proof To prove part (a), we continue with the same notation as in the proof of Lemma 7.7.

First, combining part (a) of Lemmas 7.6 and 7.7, and the Cauchy-Schwarz inequality, we find that

$$\max_{1\leq i\leq n} |\xi_i| = \max_{1\leq i\leq n} |\lambda^{\top}(Y_i - \mu_0)| \leq \|\lambda\| \cdot M_n = O_{\mathbb{P}}(n^{-1/2}) \cdot o(n^{1/2}) = o_{\mathbb{P}}(1),$$

hence

$$\forall 1 \leq i \leq n : |1 + \xi_i|^{-1} = O_{\mathbb{P}}(1). \tag{7.15}$$

Now, recall that

$$\begin{aligned} 0 = g(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i - \mu_0}{1 + \lambda^{\top}(Y_i - \mu_0)} \\ &= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_0) \cdot \frac{1}{1+\xi_i} \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)\left[1 - \xi_i + \frac{\xi_i^2}{1+\xi_i}\right]$$

$$= \bar{Y}_n - \mu_0 - S\lambda + \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \mu_0)\xi_i^2}{1+\xi_i}. \tag{7.16}$$

Combining part (b) of Lemma 7.6, Lemma 7.7, and (7.15) yields that the norm of the third summand in (7.16) is bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\|Y_i - \mu_0\|^3 \cdot \|\lambda\|^2 \cdot |1+\xi_i|^{-1} = o(n^{1/2}) \cdot O_{\mathbb{P}}(n^{-1}) \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(n^{-1/2}).$$

Thus, we can write

$$\lambda = S^{-1}(\bar{Y}_n - \mu_0) + \beta, \tag{7.17}$$

where $\|\beta\| = o_{\mathbb{P}}(n^{-1/2})$. Furthermore, considering the Taylor expansion of the logarithm, we may write $\log(1+\xi_i) = \xi_i - \xi^2/2 + \eta_i$, where the $(\eta_i : 1 \le i \le n)$ are such that

$$\lim_{n\to\infty} \mathbb{P}(\forall 1 \le i \le n : |\eta_i| \le B|\xi_i|^3) = 1 \tag{7.18}$$

for some finite constant $B > 0$. Also, recall that $np_i = (1+\xi_i)^{-1}$, $1 \le i \le n$.

Now, we write

$$\begin{aligned} -2\log(\mathscr{R}(\mu_0)) &= -2\sum_{i=1}^{n}\log(np_i) \\ &= 2\sum_{i=1}^{n}\log(1+\xi_i) \\ &= 2\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\xi_i^2 + 2\sum_{i=1}^{n}\eta_i \\ &= 2n\lambda^\top(\bar{Y}_n - \mu_0) - n\lambda^\top S\lambda + 2\sum_{i=1}^{n}\eta_i. \end{aligned} \tag{7.19}$$

Plugging (7.17) into (7.19), we obtain that

$$-2\log(\mathscr{R}(\mu_0)) = n(\bar{Y}_n - \mu_0)^\top S^{-1}(\bar{Y}_n - \mu_0) - n\beta^\top S\beta + 2\sum_{i=1}^{n}\eta_i,$$

because

$$2n\lambda^\top(\bar{Y}_n - \mu_0) = 2n(\bar{Y}_n - \mu_0)^\top S^{-1}(\bar{Y}_n - \mu_0) + 2n\beta^\top(\bar{Y}_n - \mu_0), \tag{7.20}$$

$$n\lambda^\top S\lambda = n(\bar{Y}_n - \mu_0)^\top S^{-1}(\bar{Y}_n - \mu_0) + 2n\beta^\top(\bar{Y}_n - \mu_0) + n\beta^\top S\beta, \tag{7.21}$$

according to Exercise 7.2.

The central limit theorem, together with the continuous mapping theorem and Slutzky's lemma, yields that $n(\bar{Y}_n - \mu_0)^\top S^{-1}(\bar{Y}_n - \mu_0)$ converges in distribution to χ^2_d as $n \to \infty$. Moreover, $n\beta^\top S\beta = o_{\mathbb{P}}(1)$, because $\|\beta\|^2 = o_{\mathbb{P}}(n^{-1})$ and S converges to the fixed (covariance) matrix Σ as $n \to \infty$. Finally, we get from (7.18) that

$$\left|\sum_{i=1}^n \eta_i\right| \le B\|\lambda\|^3 \sum_{i=1}^n \|Y_i - \mu_0\|^3 = O_{\mathbb{P}}(n^{-3/2})o_{\mathbb{P}}(n^{3/2}) = o_{\mathbb{P}}(1),$$

implying the assertion of part (a). Part (b) is an immediate consequence of part (a).

To prove part (c), choose $\mu_1 \in C_\alpha(\mu)$ and $\mu_2 \in C_\alpha(\mu)$ as well as $\tau \in (0,1)$ arbitrarily. We have to show that the convex combination $\mu(\tau) := \tau\mu_1 + (1-\tau)\mu_2$ belongs to $C_\alpha(\mu)$ as well. To this end, we argue as follows. Since μ_1 and μ_2 are elements of $C_\alpha(\mu)$, there exist tuples $(p_{1i} : 1 \le i \le n)$ and $(p_{2i} : 1 \le i \le n)$ of probabilities such that $\mu_k = \sum_{i=1}^n p_{ki}Y_i$ for $k = 1, 2$ and $-2\sum_{i=1}^n \log(np_{ki}) \le c_\alpha$, $k = 1, 2$. Now, define $p_i(\tau) = \tau p_{1i} + (1-\tau)p_{2i}$ for $1 \le i \le n$. It is easy to see that $(p_i(\tau) : 1 \le i \le n)$ are non-negative and sum up to one. Furthermore, $\sum_{i=1}^n p_i(\tau)Y_i = \tau\mu_1 + (1-\tau)\mu_2 = \mu(\tau)$. It remains to show that $-2\sum_{i=1}^n \log(np_i(\tau)) \le c_\alpha$. This can be verified by writing

$$-2\sum_{i=1}^n \log(np_i(\tau)) = -2\sum_{i=1}^n \log\left(\tau np_{1i} + (1-\tau)np_{2i}\right)$$

and exploiting the fact that the logarithm is a concave function, meaning that

$$\log\left\{\tau np_{1i} + (1-\tau)np_{2i}\right\} \ge \tau\log\left(np_{1i}\right) + (1-\tau)\log\left(np_{2i}\right)$$

for all $1 \le i \le n$, leading to

$$\begin{aligned} -2\sum_{i=1}^n \log(np_i(\tau)) &\le -2\left[\tau\sum_{i=1}^n \log\left(np_{1i}\right) + (1-\tau)\sum_{i=1}^n \log\left(np_{2i}\right)\right] \\ &\le \tau c_\alpha + (1-\tau)c_\alpha = c_\alpha, \end{aligned}$$

completing the proof.

Remark 7.9 If $0 < \operatorname{rank}(\Sigma) = q < d$, then Theorem 7.8 holds with χ^2_d replaced by χ^2_q.

7.2 Some Modifications and Generalizations

The following result can be proved in analogy to the calculations in the proof of Lemma 7.1.

Lemma 7.10 *Under the general assumptions of Sect. 7.1, it holds that*

$$\mathscr{K}(\hat{P}_0 \| \hat{P}_n) = \sum_{i=1}^{n} p_i \log(np_i).$$

Lemma 7.11 *Let $y_1, \ldots, y_n$ with $y_i \in \mathbb{R}^d$, $1 \le i \le n$, be given points, and $\mu_0 \in \mathbb{R}^d$ a further given vector. Assume that μ_0 is located inside the convex hull of $y_1, \ldots, y_n$.*

Then, the solution $(p_1, \ldots, p_n)^\top$ of the constrained optimization problem

$$\begin{aligned} \textit{minimize} \quad & \sum_{i=1}^{n} p_i \log(np_i) && (7.22)\\ \textit{subject to} \quad & \forall 1 \le i \le n : p_i \ge 0, \\ & \sum_{i=1}^{n} p_i = 1, && (7.23)\\ & \sum_{i=1}^{n} p_i (y_i - \mu_0) = 0 \in \mathbb{R}^d, && (7.24) \end{aligned}$$

can be written as follows:

$$\forall 1 \le i \le n : p_i \equiv p_i(\lambda) = \frac{\exp\left(\lambda^\top (y_i - \mu_0)\right)}{\sum_{j=1}^{n} \exp\left(\lambda^\top (y_j - \mu_0)\right)} = \frac{\exp\left(\lambda^\top y_i\right)}{\sum_{j=1}^{n} \exp\left(\lambda^\top y_j\right)}, \tag{7.25}$$

where $\lambda \equiv \lambda(\mu_0) = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$ satisfies the system of equations

$$\sum_{i=1}^{n} p_i(\lambda)(y_i - \mu_0) = 0 \in \mathbb{R}^d \tag{7.26}$$

or, equivalently,

$$\sum_{i=1}^{n} \exp\left(\lambda^\top y_i\right)(y_i - \mu_0) = 0 \in \mathbb{R}^d.$$

Proof We proceed as in the proof of Lemma 7.3. Here, a suitable Lagrangian function is given by

$$L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma) = \sum_{i=1}^{n} p_i \log(np_i) - \lambda^\top \left(\sum_{i=1}^{n} p_i(y_i - \mu_0)\right) - \gamma\left(\sum_{i=1}^{n} p_i - 1\right),$$

with partial derivatives given by

$$\frac{\partial L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma)}{\partial p_i} = \log(np_i) + 1 - \lambda^\top(y_i - \mu_0) - \gamma.$$

Now, again assuming that $p_i > 0$ for all $1 \leq i \leq n$,

$$\begin{aligned} \frac{\partial L(p_1, \ldots, p_n, \lambda_1, \ldots, \lambda_d, \gamma)}{\partial p_i} &= 0 \\ \iff \log(np_i) + 1 - \lambda^\top(y_i - \mu_0) - \gamma &= 0 \qquad (7.27) \\ \iff p_i \log(np_i) + p_i - \lambda^\top p_i(y_i - \mu_0) - p_i\gamma &= 0. \qquad (7.28) \end{aligned}$$

Summing the left-hand and the right-hand sides of (7.28), we find that

$$\gamma = \sum_{i=1}^{n} p_i \log(np_i) + 1.$$

Utilizing this result in (7.27) and simplifying, we obtain that

$$\forall 1 \leq i \leq n : \log(p_i) - \sum_{i=1}^{n} p_i \log(p_i) = \lambda^\top(y_i - \mu_0). \qquad (7.29)$$

The representation (7.29) implies that there exists a constant $c > 0$ such that

$$\forall 1 \leq i \leq n : p_i = c \cdot \exp\left(\lambda^\top(y_i - \mu_0)\right).$$

Due to constraint (7.23), $c = \left[\sum_{j=1}^{n} \exp\left(\lambda^\top(y_j - \mu_0)\right)\right]^{-1}$, completing the proof.

Remark 7.12

(a) The probabilities $(p_i : 1 \leq i \leq n)$ from (7.25) are commonly referred to as "exponential tilting weights," see Efron (1981). Their projection properties with respect to $\mathscr{K}(\hat{P}_0 \| \hat{P}_n)$ have already been studied by Csiszar (1975).
(b) The exponential tilting-based projection test can be carried out as an asymptotic chi-square test in analogy to the results in Theorem 7.8; see, e.g., Schennach (2007) and Li et al. (2011).

Definition 7.13 (Cressie and Read (1984)) Let $n \in \mathbb{N}$, γ be a real constant and $p \in (0, 1)$. Then, the Cressie-Read discrepancy h_γ between p and n^{-1} is given by

$$h_\gamma(p) = \frac{(np)^{\gamma+1} - 1}{\gamma(\gamma + 1)}.$$

Lemma 7.14 *Let $p \in (0, 1)$ and $n \in \mathbb{N}$ be given. Then it holds that*

$$\lim_{\gamma \to -1} h_\gamma(p) = -\log(np), \tag{7.30}$$

$$\frac{\partial}{\partial \gamma}(np)^{\gamma+1} = (np)^{\gamma+1} \log(np). \tag{7.31}$$

In view of (7.30) and (7.31), we may define $h_{-1}(p) := -\log(np)$ and $h_0(p) := np \log(np)$ for p close to n^{-1}.

Corollary 7.15 *Under the general assumptions of Sect. 7.1, the following assertions hold true.*

$$\mathscr{K}(\hat{P}_n \| \hat{P}_0) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{1}{np_i}\right) = \frac{1}{n} \sum_{i=1}^{n} h_{-1}(p_i),$$

$$\mathscr{K}(\hat{P}_0 \| \hat{P}_n) = \sum_{i=1}^{n} p_i \log(np_i) = \frac{1}{n} \sum_{i=1}^{n} h_0(p_i).$$

Thus, both the empirical likelihood method and the exponential tilting method can be regarded as special cases of (constrained) empirical Cressie-Read discrepancy minimization.

Remark 7.16

(a) The (constrained) empirical Cressie-Read discrepancy minimization technique can further be generalized to treat functionals which are defined via generalized estimating equations (GEEs). To this end, consider a function $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^q$ and define the functional κ of interest (taking values in $\mathbb{R}^p$) implicitly by $\mathbb{E}[g(Y_1, \kappa)] = 0$. Obviously, the mean ($\kappa = \mu = \mathbb{E}[Y_1]$) is a special case, where $p = q = d$ and $g(Y_1, \kappa) = Y_1 - \kappa$. One may now, for given $\gamma \in \mathbb{R}$, consider

the minimization problem

$$\begin{aligned} \text{minimize} \quad & n^{-1} \sum_{i=1}^{n} h_\gamma(p_i) \\ \text{subject to} \quad & \forall 1 \le i \le n : p_i \ge 0, \\ & \sum_{i=1}^{n} p_i = 1, \\ & \sum_{i=1}^{n} p_i g(y_i, \kappa_0) = 0 \end{aligned}$$

for the null hypothesis $H_0 = \{\kappa = \kappa_0\}$, where $\kappa_0 \in \mathbb{R}^p$ is the hypothesized value of κ.

(b) A different generalization approach, allowing for treating derived parameters, consists of profile ELR tests. For example, the correlation coefficient pertaining to a bivariate i.i.d. sample can be tested with this methodology; see Dickhaus (2015) and the references therein.

7.3 Exercises

Exercise 7.1 *Show that Eqs.* (7.11) *and* (7.12) *appearing in the proof of Lemma 7.7 hold true.*

Exercise 7.2 *Show that Eqs.* (7.20) *and* (7.21) *appearing in the proof of Theorem 7.8 hold true.*

Exercise 7.3 (Programming Exercise)

(a) Write an `R` *program which implements the empirical likelihood ratio test for point null hypotheses regarding a multivariate mean.*
Hint: *Use Theorem 7.8 in connection with duality of tests and confidence regions.*

(b) Assess the accuracy of the chi-square approximation of the null distribution of $-2\log(\mathscr{R}(\mu_0))$ *in a computer simulation.*

References

Bishop YM, Fienberg SE, Holland PW (2007) Discrete multivariate analysis. Theory and practice. With the collaboration of Richard J. Light and Frederick Mosteller. Reprint of the 1975 edition. Springer, Berlin. https://doi.org/10.1007/978-0-387-72806-3

Chandra TK (2012) The Borel-Cantelli lemma. Springer, New York, NY. https://doi.org/10.1007/978-81-322-0677-4

Cressie N, Read TR (1984) Multinomial goodness-of-fit tests. J R Stat Soc Ser B 46:440–464

Csiszar I (1975) I-divergence geometry of probability distributions and minimization problems. Ann Probab 3:146–158. https://doi.org/10.1214/aop/1176996454

Dickhaus T (2015) Self-concordant profile empirical likelihood ratio tests for the population correlation coefficient: a simulation study. In: Stochastic models, statistics and their applications. Collected papers based on the presentations at the 12th workshop, Wrocław, Poland, February 2015, Springer, Cham, pp 253–260. https://doi.org/10.1007/978-3-319-13881-7

Efron B (1981) Nonparametric standard errors and confidence intervals. Can J Stat 9:139–172. https://doi.org/10.2307/3314608

Li X, Chen J, Wu Y, Tu D (2011) Constructing nonparametric likelihood confidence regions with high order precisions. Stat Sin 21(4):1767–1783. https://doi.org/10.5705/ss.2009.117

Owen AB (2001) Empirical likelihood. Chapman & Hall/CRC, Boca Raton, FL

Schennach SM (2007) Point estimation with exponentially tilted empirical likelihood. Ann Stat 35(2):634–672. https://doi.org/10.1214/009053606000001208

Wilks S (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9:60–62. https://doi.org/10.1214/aoms/1177732360

Chapter 8
Some Extensions

8.1 Linear Rank Tests for One-Sample Problems

This section mainly follows Section 4.4 of Büning and Trenkler (1994).

In Chap. 4, we have derived rank tests for multi-sample problems (with stochastically independent observables), exploiting the property that the vector of (pooled) ranks is uniformly distributed on $\mathscr{S}_n$ under the null hypothesis of (distributionally) homogeneous groups. Under certain model assumptions, this idea can be adapted to treat one-sample problems.

Model 8.1 (One-Sample Location Parameter Model) *Let $Y_1, \ldots, Y_n$ denote real-valued i.i.d. observables on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with cdf F_θ of Y_1 depending on a parameter $\theta \in \Theta$. Assume that*

$$\forall y \in \mathbb{R} : F_\theta(y) = F(y - \theta),$$

where F is a continuous cdf which is symmetric about the point $(0, 1/2)$. Hence, θ is a location parameter; cf. Definition 3.10. We let F unspecified, yielding a semi-parametric model.

Under Model 8.1, assume that the point null hypothesis $H_0 = \{\theta = \theta_0\}$ for some given point $\theta_0 \in \mathbb{R}$ is of interest, with possible (two-sided or one-sided) alternatives given by $H_1 = \{\theta \neq \theta_0\}$, $H_1^- = \{\theta < \theta_0\}$, or $H_1^+ = \{\theta > \theta_0\}$, respectively. The idea now is to split the sample into two sub-samples, namely $\{Y_i : Y_i > \theta_0\}$ and $\{Y_i : Y_i < \theta_0\}$. Under H_0, these two sub-samples are homogeneous, whereas they are heterogeneous whenever $\theta \neq \theta_0$. In particular, notice that

$$\forall \theta \in \Theta : \forall 1 \leq i \leq n : \mathbb{P}_\theta(Y_i \leq \theta) = \mathbb{P}_\theta(Y_i - \theta \leq 0) = F(0) = \frac{1}{2} = \mathbb{P}_\theta(Y_i > \theta). \tag{8.1}$$

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6_8

Definition 8.2 Assume that the assumptions of Model 8.1 are fulfilled. Let $D_i = Y_i - \theta_0$, with corresponding absolute value $|D_i| = |Y_i - \theta_0|$, $1 \leq i \leq n$. Furthermore, let $R_i^+ = R(|D_i|)$ denote the rank of $|D_i|$ among all absolute differences $|D_1|, \ldots, |D_n|$, and let $Z_i = \mathbf{1}\{D_i > 0\}$, $1 \leq i \leq n$. Due to (8.1), $\{Z_i\}_{1 \leq i \leq n}$ are i.i.d. under H_0, with $Z_1 \sim \text{Bernoulli}(1/2)$. For testing H_0, define the *linear rank statistic*

$$L_n^+ = \sum_{i=1}^{n} g(R_i^+) Z_i,$$

where $(g(i))_{1 \leq i \leq n}$ are given weights (scores).

Alternatively, we can write

$$L_n^+ = \sum_{i=1}^{n} g(i) V_i,$$

where

$$\forall 1 \leq i \leq n : V_i = \mathbf{1}\{|D|_{i:n} \text{ corresponds to a positive } D_j\},$$

with $|D|_{1:n} < |D|_{2:n} < \ldots < |D|_{n:n}$ denoting the order statistics of $|D_1|, \ldots, |D_n|$.

Example 8.3 Assume that $n = 10$ study participants, which were randomly drawn from a (homogeneous) target population, perform (independently from each other) an IQ test. Assume that the distribution of the IQ in the target population is such that the assumptions of Model 8.1 are fulfilled. Assume that the following ten IQ values are observed.

```
99   131   118   112   128   136   120   107   134   122
```

Then, letting $\theta_0 = 110$, we can compute the values of the quantities defined in Definition 8.2 as displayed in Table 8.1.

Table 8.1 Relevant quantities for Example 8.3

i	y_i	$d_i = y_i - 110$	$\lvert d_i \rvert$	r_i^+	z_i	v_i
1	99	−11	11	5	0	1
2	131	21	21	8	1	0
3	118	8	8	3	1	1
4	112	2	2	1	1	1
5	128	18	18	7	1	0
6	136	26	26	10	1	1
7	120	10	10	4	1	1
8	107	−3	3	2	0	1
9	134	24	24	9	1	1
10	122	12	12	6	1	1

The values v_i for $1 \leq i \leq n$ in the last column of Table 8.1 have been obtained by permuting the values $(z_i)_{1\leq i\leq n}$ according to the ordering of $(r_i^+)_{1\leq i\leq n}$. Hence, the realized value of L_n^+ for this dataset equals

$$\begin{aligned} L_n^+(y_1,\ldots,y_n) &= g(8)+g(3)+g(1)+g(7)+g(10)+g(4)+g(9)+g(6) \\ &= \sum_{i=1}^{n} g(i) - g(2) - g(5). \end{aligned}$$

Theorem 8.4 *Under Model 8.1, the following assertions hold true under* $H_0 = \{\theta = \theta_0\}$.

(i) For all $1 \leq i \leq n$, *we have that*

$$\mathbb{P}_{\theta_0}(V_i = 1) = \frac{1}{2} = \mathbb{P}_{\theta_0}(V_i = 0).$$

(ii) For all binary tuples $(v_1,\ldots,v_n)^\top \in \{0,1\}^n$, *we have that*

$$\mathbb{P}_{\theta_0}(V_1 = v_1,\ldots,V_n = v_n) = \frac{1}{2^n}.$$

Thus, the random vector $(V_1,\ldots,V_n)^\top$ *is under* θ_0 *uniformly distributed on* $\{0,1\}^n$.

(iii) For all $\ell^+ \in supp(L_n^+)$, *we have that*

$$\mathbb{P}_{\theta_0}(L_n^+ = \ell^+) = \frac{a(\ell^+)}{2^n},$$

where $a(\ell^+) = |\{(v_1,\ldots,v_n)^\top \in \{0,1\}^n : \sum_{i=1}^n g(i)v_i = \ell^+\}|$.

(iv) The first two moments of L_n^+ *under* θ_0 *are given by*

$$\mathbb{E}_{\theta_0}[L_n^+] = \frac{1}{2}\sum_{i=1}^{n} g(i),$$

$$Var_{\theta_0}(L_n^+) = \frac{1}{4}\sum_{i=1}^{n} g^2(i).$$

Proof The $(V_i)_{1\leq i\leq n}$ are obtained from $(Z_i)_{1\leq i\leq n}$ by permuting. As argued in Definition 8.2, $(Z_i)_{1\leq i\leq n}$ are under θ_0 i.i.d., hence exchangeable. We conclude that

$$\mathscr{L}_{\theta_0}\left((V_1,\ldots,V_n)^\top\right) = \mathscr{L}_{\theta_0}\left((Z_1,\ldots,Z_n)^\top\right) = (\text{Bernoulli}(1/2))^{\otimes n}.$$

This implies all four assertions, because

$$\mathbb{E}[\text{Bernoulli}(1/2)] = \frac{1}{2} \quad \text{and}$$

$$\text{Var}\,(\text{Bernoulli}(1/2)) = \frac{1}{4}.$$

Example 8.5

(a) The *sign test* employs the weights $g(i) \equiv 1$ for all $1 \le i \le n$. The resulting test statistic is given by $V_n^+ := \sum_{i=1}^{n} V_i$. It equals the number of the D_i, $1 \le i \le n$, with positive sign.

(b) *Wilcoxon's signed rank test* employs the weights $g(i) = i$ for all $1 \le i \le n$. The resulting test statistic is given by $W_n^+ := \sum_{i=1}^{n} R_i^+ Z_i$. It equals the sum of the ranks of those $|D_i|$ for which D_i is positive, $1 \le i \le n$.

Theorem 8.6 *With the notation introduced in Example 8.5, the following assertions hold true under* $H_0 = \{\theta = \theta_0\}$.

(a) Under θ_0, V_n^+ *follows the Bin*$(n, 1/2)$ *distribution, i.e.,*

$$\forall 0 \le k \le n : \mathbb{P}_{\theta_0}(V_n^+ = k) = \binom{n}{k} \left(\frac{1}{2}\right)^n = \frac{\binom{n}{k}}{2^n}.$$

(b) For large sample sizes n, we have that

$$\mathbb{P}_{\theta_0}(W_n^+ \le w) \approx \Phi(z),$$

where Φ *denotes the cdf of the standard normal distribution on* $\mathbb{R}$, *and*

$$z \equiv z(w) = \frac{w - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}.$$

Proof Assertion (a) follows immediately from part (ii) of Theorem 8.4.

To prove assertion (b), we note that, due to part (iv) of Theorem 8.4, it holds that

$$\mathbb{E}_{\theta_0}[W_n^+] = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{2}\frac{n(n+1)}{2} = \frac{n(n+1)}{4},$$

$$\text{Var}_{\theta_0}(W_n^+) = \frac{1}{4}\sum_{i=1}^{n} i^2$$

$$= \frac{1}{4}\left[\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right]$$
$$= \frac{2n^3 + 3n^2 + n}{24} = \frac{n(n+1)(2n+1)}{24}.$$

The assertion follows by the central limit theorem.

Application 8.7

(a) The sign test can be carried out as a binomial test.
(b) For $n > 20$ (according to Büning and Trenkler (1994)), Wilcoxon's signed rank test can be carried out as an approximate Z-test.

8.2 Tied Observations

This section mainly follows the theoretical parts of Brunner and Munzel (2013).

In Chap. 4, we have assumed that there are (with probability one) no tied observations, so that ranks could be assigned (almost surely) unambiguously to the observed data points. However, in practice one often has a non-zero probability for ties, because the measurements cannot be performed with arbitrary precision. For example, human height is typically reported in full centimeters, implying a positive probability for observing the same height more than once in a random sample.

There are several possibilities how to deal with tied observations in rank-based statistical inference methods. In many cases, the most appropriate method is to assign so-called *midranks* to tied observations. One of the main reasons for this choice is the fact that the total (mid-)rank sum then equals the tie-free rank sum $n(n+1)/2$, where n denotes the sample size; see Theorem 8.10 below.

In any case, notice that the (random) vector $R(Y)$ of the ranks of an i.i.d. sample $Y = (Y_1, \ldots, Y_n)^\top$ is not uniformly distributed on $\mathscr{S}_n$ anymore if there is a positive probability for ties. Instead, the exact distribution of $R(Y)$ depends on the (expected) tie structure of the joint distribution of Y, which is often unknown in practice. As a consequence, also the exact null distribution of linear rank statistics as considered in Chap. 4 is often intractable or unknown in the presence of ties. However, central limit theorems (with a modified variance or an appropriate estimator thereof) typically continue to hold true in the tied case; see, e.g., Brunner and Munzel (2013) for a detailed treatment.

Definition 8.8

(a) For a real number x, define

$$c^-(x) = \mathbf{1}\{x > 0\},$$
$$c^+(x) = \mathbf{1}\{x \geq 0\},$$
$$c(x) = \frac{1}{2}\left[c^+(x) + c^-(x)\right].$$

(b) For a sample $Y = (Y_1, \ldots, Y_n)^\top$ of real-valued random variables, and for all $y \in \mathbb{R}$, define

$$\hat{F}_n^-(y) = \frac{1}{n} \sum_{i=1}^{n} c^-(y - Y_i),$$

$$\hat{F}_n^+(y) = \frac{1}{n} \sum_{i=1}^{n} c^+(y - Y_i),$$

$$\hat{F}_n^{\text{norm}}(y) = \frac{1}{n} \sum_{i=1}^{n} c(y - Y_i).$$

We call $\hat{F}_n^-$ the left-continuous, $\hat{F}_n^+$ the right-continuous, and $\hat{F}_n^{\text{norm}}$ the normalized version of the ecdf pertaining to $Y_1, \ldots, Y_n$.

(c) Under the assumptions of part (b), define for all $1 \le i \le n$ the following quantities.

$$R_i^- \equiv R_i^-(Y) = \sum_{j=1}^{n} c^-(Y_i - Y_j) + 1 = n\hat{F}_n^-(Y_i) + 1,$$

$$R_i^+ \equiv R_i^+(Y) = \sum_{j=1}^{n} c^+(Y_i - Y_j) = n\hat{F}_n^+(Y_i),$$

$$R_i \equiv R_i(Y) = \frac{1}{2}\left[R_i^- + R_i^+\right]. \tag{8.2}$$

We call $R_i^-(Y)$ the minimum rank, $R_i^+(Y)$ the maximum rank, and $R_i(Y)$ the midrank of Y_i, $1 \le i \le n$.

Lemma 8.9 *Let $1 \le a \le b$ be two positive integers. Then it holds that*

$$\frac{1}{b-a+1} \sum_{\ell=a}^{b} \ell = \frac{a+b}{2}.$$

Proof First, we notice that

$$\sum_{\ell=1}^{b} \ell = \frac{b(b+1)}{2} \quad \text{and} \quad \sum_{\ell=1}^{a-1} \ell = \frac{a(a-1)}{2}.$$

Hence,

$$\sum_{\ell=a}^{b} \ell = \sum_{\ell=1}^{b} \ell - \sum_{\ell=1}^{a-1} \ell$$
$$= \frac{b(b+1)}{2} - \frac{a(a-1)}{2} = \frac{b^2+b-a^2+a}{2}.$$

On the other hand, we also have that

$$\frac{a+b}{2}(b-a+1) = \frac{ab-a^2+a+b^2-ab+b}{2} = \frac{b^2+b-a^2+a}{2}.$$

Lemma 8.9 justifies the terms "midrank" or "average rank," respectively, of Y_i for the quantity

$$R_i = \frac{R_i^- + R_i^+}{2} = \frac{1}{R_i^+ - R_i^- + 1} \sum_{\ell=R_i^-}^{R_i^+} \ell \tag{8.3}$$

from (8.2). In the case of no ties, we have that $R_i^- = R_i^+ = R_i$ for all $1 \le i \le n$.

Theorem 8.10 *The sum of the n midranks of $Y_1, \ldots, Y_n$ always equals $n(n+1)/2$, no matter the structure of the ties in the data.*

Proof Assume that there are G groups with n_g tied observations in group g, $1 \le g \le G$, such that $\sum_{g=1}^{G} n_g = n$. Denote the minimum rank number in group g by $m(g)$ and the maximum rank number in group g by $M(g)$, and notice that $M(g) - m(g) + 1 = n_g$ is the number of observations in group g, $1 \le g \le G$. Due to (8.3), the midrank of every of the n_g observations in group g equals $n_g^{-1} \sum_{\ell=m(g)}^{M(g)} \ell$. We conclude that

$$\sum_{i=1}^{n} r_i = \sum_{g=1}^{G} \left[n_g \cdot \frac{1}{n_g} \sum_{\ell=m(g)}^{M(g)} \ell \right] = \sum_{g=1}^{G} \sum_{\ell=m(g)}^{M(g)} \ell = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Remark 8.11 Apart from assigning minimum ranks, maximum ranks, or midranks, there are also other possibilities to deal with ties. For example, one may remove tied observations altogether, or one may break the ties by randomly assigning the numbers from R_i^- to R_i^+ to the tied observations in question.

8.3 Exercises

Exercise 8.1 *Continue Example 8.3 by testing the null hypothesis* $H_0 = \{\theta = 110\}$ *against its one-sided alternative* $H_1^+ = \{\theta > 110\}$ *with the sign test at significance level* $\alpha = 5\%$.

Exercise 8.2 *Generate on the computer a (pseudo) random sample of size* $n = 10$ *from some univariate probability distribution. Sketch the graphs of* $\hat{F}_{10}^-$, $\hat{F}_{10}^+$, *and* $\hat{F}_{10}^{norm}$ *from Definition 8.8, evaluated on the generated sample, together in one figure.*

References

Brunner E, Munzel U (2013) Nichtparametrische Datenanalyse. Unverbundene Stichproben, 2nd edn. Springer Spektrum, Heidelberg. https://doi.org/10.1007/978-3-642-37184-4

Büning H, Trenkler G (1994) Nichtparametrische statistische Methoden, 2nd edn. de Gruyter, Berlin

Index

T. Dickhaus, *Theory of Nonparametric Tests*,
https://doi.org/10.1007/978-3-319-76315-6