Selected Topics from

Mathematical Statistics

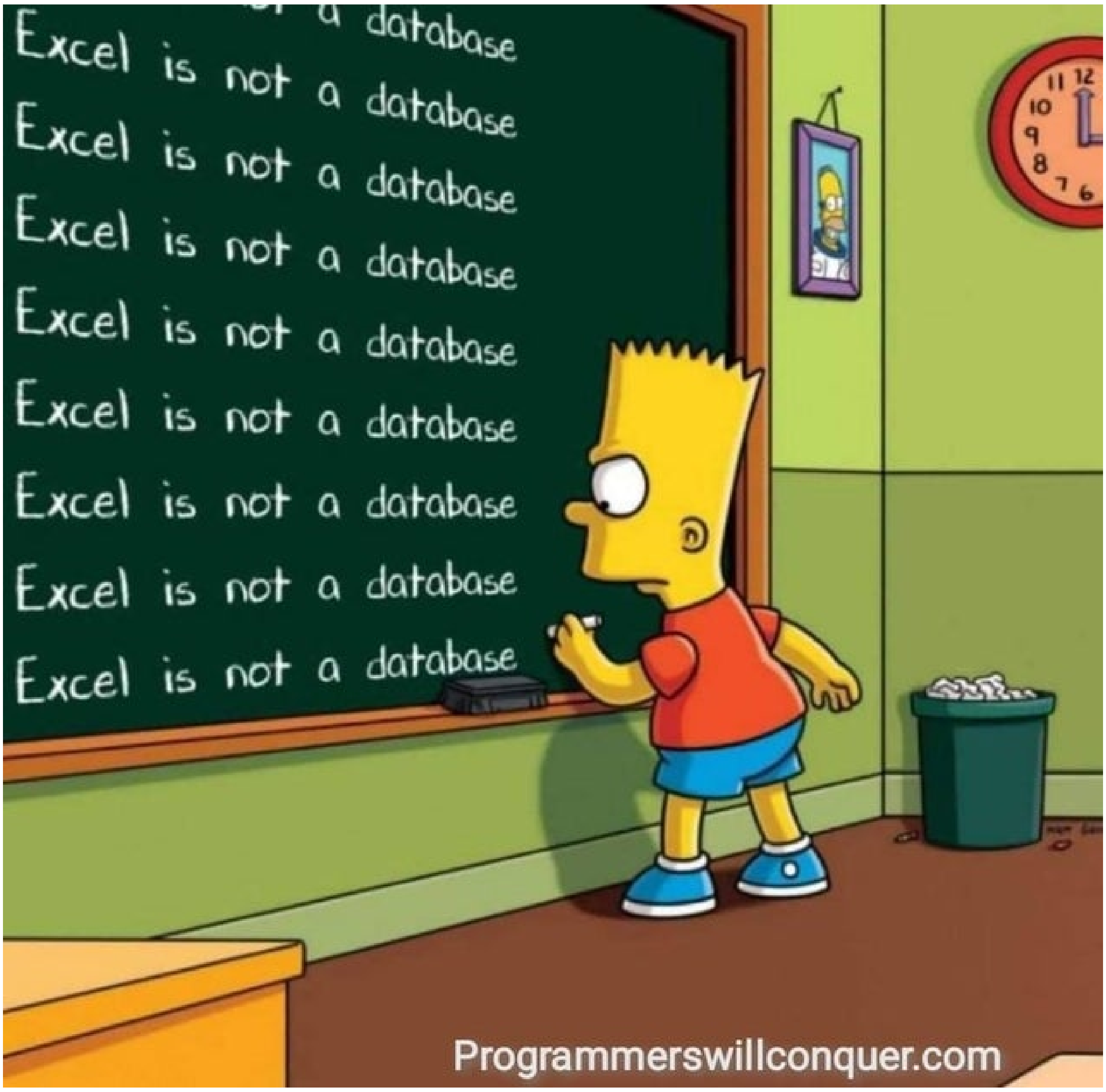
Lecture Notes

Winter 2025/ 2026

Alois Pichler

TECHNISCHE UNIVERSITÄT
CHEMNITZ
Faculty of Mathematics

DRAFT
Version as of October 9, 2025


a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Programmerswillconquer.com

Preface and Acknowledgment

Die Wahrheit wird euch frei machen.

Joh. 8, 32

The purpose of these lecture notes is to facilitate the content of the lecture and the course. From experience it is helpful and recommended to attend and follow the lectures in addition. These lecture notes do not cover the lectures completely.

I am indebted to Prof. Georg Ch. Pflug for numerous discussions in the area, significant support over years and his manuscripts; some particular parts here follow this manuscript “mathematische Statistik” closely.

Important literature for the subject, also employed here, includes Rüschendorf [17], Pruscha [14], Georgii [6], Czado and Schmidt [3], Witting and Müller-Funk [22] and https://www.wikipedia.org. Pruscha [15] collects methods and applications.

Please report mistakes, errors, violations of copyright, improvements or necessary completions. Updated version of these lecture notes:
https://www.tu-chemnitz.de/mathematik/fima/public/mathematischeStatistik.pdf

Complementary introduction and exercises:
https://www.tu-chemnitz.de/mathematik/fima/public/einfuehrung-statistik.pdf

Syllabus and content of the lecture:
https://www.tu-chemnitz.de/mathematik/studium/module/2013/B15.pdf

Prerequisite content known from Abitur:
https://www.tu-chemnitz.de/mathematik/schule/online.php

Contents

1 Preliminaries and Notations

> Nature laughs at the difficulties of integration.
>
> Pierre-Simon Laplace, 1749–1827

1.1 WHERE WE ARE

Descriptive statistics[1] is the process of using and analyzing descriptive statistics, which is a summary statistic that quantitatively describes or summarizes features of a collection of information.

Exploratory data analysis[2] is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

Mathematical statistics[3] is the application of probability theory, a branch of mathematics, to statistics, as opposed to techniques for collecting statistical data.

1.2 NOTATION

Observed variable is a variable that can be observed and directly measured.

Latent variable (from Latin latere (hidden), as opposed to observable variables), is a variable that is not directly observed but rather inferred (through a mathematical model) from other variables that are observed.

Explanatory variable see independent variable

Independent variable The models or experiments investigate how the dependent variables depend on independent variables. Synonyms, especially in statistics, are *predictor variable*, *regressor*, *controlled variable*, *manipulated variable*, *explanatory variable*, *risk factor* (medical statistics), *feature* (in machine learning and pattern recognition) or *input variable.*

Dependent variable Synonyms are *response variable*, *regressand*, *predicted variable*, *measured variable*, *explained variable*, *experimental variable*, *responding variable*, *outcome variable* or *label*.

Categorical variable a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.

Location parameter determines the *location* or shift of a distribution

Scale parameter is a special kind of numerical parameter of a parametric family of probability distributions. The larger the scale parameter, the more spread out the distribution.

[1] Deskriptive, beschreibende oder beurteilende Statistik

[2] Explorative Statistik

[3] Mathematische Statistik

Rate parameter is the reciprocal of a scale parameter.

Nuisance parameter is any parameter which is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest.

Population a set of similar items or events which is of interest for some question or experiment.

Population parameter a statistical parameter is a quantity that indexes a family of probability distributions.

U-statistics unbiased.

L-statistics is a linear combination of order statistics.

L-estimate is a robust estimation technique based on linear combinations of order statistics.

R-estimate robust, based on the rank.

M-estimator is a broad class of robust estimators, which are obtained as the minima of sums of functions of the data (as maximum likelihood, cf. page 107).

Z-estimator (for zero) is a class of estimators which satisfy a system of equations.

Z-score (aka standard or normal score) of a raw score x is $z = (x - \mu)/\sigma$.

1.3 THE EXPECTATION

Let $X\colon \Omega \to \mathcal{X}$ be a random variable (i.e., a measurable function) on a probability space (Ω, Σ, P). Different ways to denote the expectation include

$$\mathbb{E}_P\, X := \int_\Omega X \,\mathrm{d}P = \int_\Omega X(\omega) P(\mathrm{d}\omega) = \int_{\mathcal{X}} x\, P(X \in \mathrm{d}x) \quad (\in \mathcal{X});$$

more generally, for a measurable function $g\colon \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned}\mathbb{E}_P\, g(X) &= \int_\Omega g(X)\,\mathrm{d}P = \int_\Omega g\big(X(\omega)\big)\, P(\mathrm{d}\omega) \\ &= \int_{\mathcal{X}} g(x)\, P(X \in \mathrm{d}x) = \int_{\mathcal{X}} g(x)\, P^X(\mathrm{d}x) = \mathbb{E}_{P^X}\, g \quad (\in \mathcal{Y}),\end{aligned} \tag{1.1}$$

where $P^X := P \circ X^{-1}$ is the *law of* X or the *image measure* (also known as *push forward measure* and occasionally denoted $X_*P := P^X$; recall the notation and identities $P(X \in A) := P\left(X^{-1}(A)\right) = P\left(\{\omega\colon X(\omega) \in A\}\right) = \mathbb{E}\,\mathbb{1}_A$).[4]

Definition 1.1 (Support)**.** The support[5] of a measure P is $\operatorname{supp}(P) := \bigcap\big\{A = \overline{A}\colon P(A^c) = 0\big\}$.

Note, that the support is topologically closed.

[4] We shall also write $X \in \mathcal{X}$ to express that the range of the random variable is $\mathcal{X}$, $X\colon \Omega \to \mathcal{X}$.

[5] Träger, dt.

1.4 MULTIVARIATE RANDOM VARIABLES

Consider multivariate random variables, i.e., random variables with $X\colon \Omega \to \mathcal{X} = \mathbb{R}^n$.

Definition 1.2 (Cumulative distribution function, cdf)**.** The *cumulative distribution function* (*cdf*, for short) is

$$F(x) := P(X \le x) = P(X_1 \le x_1, \dots, X_n \le x_n) = P^X\big((-\infty, x_1] \times \dots \times (-\infty, x_n]\big).$$

We shall also write F_X to associate the distribution function F_X and the random variable X.

For $g\colon \mathbb{R}^n \to \mathcal{Y}$ we have that

$$\mathbb{E}_P\, g(X) = \int_{\mathbb{R}^n} g(x)\,\mathrm{d}F(x), \tag{1.2}$$

where "$\mathrm{d}F(x)$" (often also "$\mathrm{d}^n F(x)$") is understood as a *Riemann–Stieltjes* integral.

Definition 1.3. For a random variables $X\colon \Omega \to \mathbb{R}^d$ on (Ω, Σ, P) and $X'\colon \Omega' \to \mathbb{R}^d$ on (Ω', Σ', P') we shall write

$$X \sim X'$$

if they coincide in distribution, i.e., if $F_X = F_{X'}$.

Definition 1.4 (Probability density function, pdf)**.** The function $f\colon \mathbb{R}^n \to [0, \infty)$ is the *probability density function* (*pdf*, for short, density) with respect to the Lebesgue measure, if

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \dots, t_n)\,\mathrm{d}t_n \cdots \mathrm{d}t_1.$$

A random variable is said to be *continuous*, if it has a density.

Remark 1.5. If X has a density, then

$$f(x_1, \dots, x_n) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_n} F(x_1, \dots, x_n). \tag{1.3}$$

Note that $\mathrm{d}F(x) = f(x)\,\mathrm{d}x$ (often also $\mathrm{d}^n F(x) = f(x)\,\mathrm{d}x^n$ or $f(x) = \frac{\mathrm{d}^n F}{\mathrm{d}x^n}$ in $\mathbb{R}^n$) with $f(\cdot)$ the Radon–Nikodym derivative with respect to the Lebesgue measure $\mathrm{d}x$ (or $\mathrm{d}x^n$), i.e., $f(x)\,\mathrm{d}x = P(X \in \mathrm{d}x)$. The expectation (1.2) thus is

$$\mathbb{E}_P\, g(X) = \int_{\mathbb{R}^n} g(x) f(x)\,\mathrm{d}x \qquad (\in \mathcal{Y}). \tag{1.4}$$

Remark 1.6. For a continuous random variable with density f_X, $P(X \in A) = \int_A f_X(x)\mathrm{d}x$ (where $A \subset \mathbb{R}^n$ is a measurable set). Then the probability that X falls in a small interval around x is approximately $P(x \le X \le x + \mathrm{d}x) \approx f(x)\mathrm{d}x$. So the shorthand physicists or probabilists sometimes write is

$$P(X \in \mathrm{d}x) = f(x)\mathrm{d}x,$$

which is an informal notation meaning "the probability that X lies in an infinitesimal neighborhood of x is $f(x)\mathrm{d}x$."

Definition 1.7 (Variance)**.** The covariance matrix of random vectors $X \in L^2(P; \mathbb{R}^n)$ and $Y \in L^2(P; \mathbb{R}^m)$ is the matrix

$$\operatorname{cov}(X, Y) := \mathbb{E}\left(X Y^\top\right) - (\mathbb{E}\, X)\,(\mathbb{E}\, Y)^\top = \mathbb{E}\,(X - \mathbb{E}\, X)\,(Y - \mathbb{E}\, Y)^\top \in \mathbb{R}^{n \times m};$$

the variance is

$$\operatorname{var} X := \operatorname{cov}(X, X).$$

It holds that $\operatorname{cov}(Y,X)=\operatorname{cov}(X,Y)^\top$.

Definition 1.8 (Precision)**.** The matrix $(\operatorname{var} X)^{-1}$ is the *precision matrix.*

Corollary 1.9 (Linear transformation and scale-invariance, matrix congruence)**.** *It holds that*

$$\operatorname{cov}(c+A\,X,d+B\,Y)=A\cdot\operatorname{cov}(X,Y)\cdot B^\top \text{ and } \operatorname{var}(b+A\,X)=A\cdot\operatorname{var}X\cdot A^\top. \tag{1.5}$$

Proposition 1.10 (Covariance of linear combinations)**.** *Let a, b, c and d be matrices (of adequate dimension) and X, Y, W and V random variables. It holds that*

$$\operatorname{cov}(a\,X+b\,V,c\,Y+d\,W)=a\ \operatorname{cov}(X,Y)\,c^\top+a\ \operatorname{cov}(X,W)\,d^\top+b\ \operatorname{cov}(V,Y)\,c^\top+b\ \operatorname{cov}(V,W)\,d^\top. \tag{1.6}$$

Lemma 1.11. *The covariance matrix is necessarily positive semi-definite.*

Proof. Indeed, define the $\mathbb{R}$-valued random variable $Z:=a^\top X$, then $0\le\operatorname{var}Z=a^\top\cdot\operatorname{var}X\cdot a$ for all vectors a. Hence, $\operatorname{var}X$ is non-negative definite, the result. □

Remark 1.12. The covariance matrix of the multinomial distribution is *singular*, cf. Exercise 1.15.

Remark 1.13 (Approximation towards the delta method). For $g(\cdot)$ and $h(\cdot)$ smooth, we have the Taylor series approximation $g(X)\approx g(\mathbb{E}\,X)+g'(\mathbb{E}\,X)(X-\mathbb{E}\,X)$. It follows from (1.5) that

$$\operatorname{cov}\big(g(X),h(Y)\big)\approx g'(\mathbb{E}\,X)\cdot\operatorname{var}X\cdot h'(\mathbb{E}\,X)^\top,$$

and

$$\operatorname{var}g(X)\approx g'(\mathbb{E}\,X)\cdot\operatorname{var}X\cdot g'(\mathbb{E}\,X)^\top;$$

specifically, $\operatorname{var}g(X)\approx g'(\mathbb{E}\,X)^2\cdot\operatorname{var}X$, if X is $\mathbb{R}$-valued.

Proposition 1.14 (Hoeffding's[6] covariance identity, cf. Lehmann [9])**.** *For X, $Y\in\mathbb{R}$ (recall Footnote 4) it holds that*

$$\operatorname{cov}(X,Y)=\iint_{\mathbb{R}\times\mathbb{R}}F_{X,Y}(x,y)-F_X(x)\cdot F_Y(y)\,\mathrm{d}x\mathrm{d}y,$$

where $F_{X,Y}(\cdot,\cdot)$ is the joint distribution function of (X,Y) and $F_X(\cdot)$ and $F_Y(\cdot)$ its marginals.

Proof. (cf. Lehmann [9][7]). Let $(\tilde X,\tilde Y)$ be an independent copy of (X,Y), i.e., $P(\tilde X\le x,\tilde Y\le y)=P(X\le x,Y\le y)=F_{X,Y}(x,y)$. Then

$$\begin{aligned}2\operatorname{cov}(X,Y)&=2\,\mathbb{E}(X\,Y)-2\,\mathbb{E}(X)\,\mathbb{E}(Y)\\&=\mathbb{E}(X-\tilde X)(Y-\tilde Y)\qquad(1.7)\\&=\mathbb{E}\int_{-\infty}^\infty\int_{-\infty}^\infty\big(\mathbb{1}_{(-\infty,x]}(\tilde X)-\mathbb{1}_{(-\infty,x]}(X)\big)\,\big(\mathbb{1}_{(-\infty,y]}(\tilde Y)-\mathbb{1}_{(-\infty,x]}(Y)\big)\,\mathrm{d}x\mathrm{d}y,\qquad(1.8)\end{aligned}$$

as $b-a=\int_{-\infty}^\infty\mathbb{1}_{(-\infty,x]}(a)-\mathbb{1}_{(-\infty,x]}(b)\,\mathrm{d}x$. We may interchange the integral and the expectation, as all are assumed to be finite. Hence

$$(1.8)=\int_{-\infty}^\infty\int_{-\infty}^\infty 2F(x,y)-2F(x)F(y)\,\mathrm{d}x\mathrm{d}y,$$

from which the assertion follows. □

[6] Wassily Hoeffding, 1914–1991

[7] Erich Leo Lehmann, 1917–2007, American statistician (with German origin)

Proposition 1.15 (Transformation of densities)**.** *Let $g(\cdot)$ be invertible with inverse $g^{-1} \in C^1$ (continuous, with continuous derivative), then*

$$f_{g(X)}(y) = f_X\big(g^{-1}(y)\big) \cdot \underbrace{\left|\det\big(g^{-1}\big)'(y)\right|}_{\text{Jacobian}} = \frac{f_X\big(g^{-1}(y)\big)}{\left|\det g'\big(g^{-1}(y)\big)\right|}. \tag{1.9}$$

Proof. Let $h(\cdot)$ be any test function. By (1.4) and changing the variables,

$$\mathbb{E}\, h\big(g(X)\big) = \int_{\mathcal{X}} h\big(g(x)\big) f_X(x)\,\mathrm{d}x = \int_{\mathcal{Y}} h(y) \cdot f_X\big(g^{-1}(y)\big) \left|\det\big(g^{-1}\big)'(y)\right| \mathrm{d}y;$$

on the other hand,

$$\mathbb{E}\, h\big(g(X)\big) = \int_{\mathcal{Y}} h(y) \cdot f_{g(X)}(y)\,\mathrm{d}y$$

by (1.4) for the random variable $g(X)$. The result follows now by comparing the integrands, as the test functions $h(\cdot)$ are arbitrary. Of course, one may choose the simple test function $h(\cdot) = \mathbb{1}_A(\cdot)$ as well to deduce the result. □

Corollary 1.16 (Linear transformation)**.** *It holds that*

$$f_{b+A\cdot X}(y) = f_X\big(A^{-1}(y-b)\big) \cdot \left|\det A^{-1}\right|.$$

1.5 INDEPENDENCE

Definition 1.17. Two events $E_1, E_2 \in \Sigma$ are *independent*, if $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$.

Definition 1.18. Two random variables X_i are *independent*, if all events $\{X_i \in A\}$ and $\{X_j \in B\}$ are independent for $i \neq j$.

Corollary 1.19. *If X and Y are independent, then* $\operatorname{cov}(X, Y) = 0$.

Proof. Indeed,

$$\begin{aligned}
\mathbb{E}(X \cdot Y^\top) &= \iint x \cdot y^\top\, P(X \in \mathrm{d}x, Y \in \mathrm{d}y) \\
&= \iint x \cdot y^\top\, P(X \in \mathrm{d}x) \cdot P(Y \in \mathrm{d}y) \\
&= \int x\, P(X \in \mathrm{d}x) \cdot \int y^\top\, P(Y \in \mathrm{d}y) \\
&= \mathbb{E}\, X \cdot \mathbb{E}\, Y^\top,
\end{aligned}$$

the assertion. □

It follows from the definition that $X, Y \in \mathbb{R}$ are independent, iff $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$. If X and Y have a density, then further, by (1.3), $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$.

Definition 1.20 (Iid)**.** The random variables X_i, $i = 1, 2, \dots,$ are *independent and identically distributed* (*iid*, for short), if they are independent and they share the same law.

Theorem 1.21 (Existence of independent copies)**.** *Let X be a random variable. Then there exist a probability space together with countably many independent random variables X_i, $i = 1, 2, \dots.$ which share the law of X.*

Proof. Let $\tilde{\Omega} := \Omega^{\mathbb{N}}$ be the space of countable trajectories over Ω and $\tilde{\Sigma}$ its product σ-algebra. Define the pre-measure (product measure) $\tilde{P}(A_1 \times \dots \times A_n \times \Omega \times \Omega \times \dots) := P(A_1) \cdot \ldots \cdot P(A_n)$. By the Carathéodory extension theorem,[8] $\tilde{P}$ extends to a probability measure on $\tilde{\Sigma}$.

The independent copies are $X_i(\tilde{\omega}) := X(\omega_i)$, where $\tilde{\omega} = (\omega_1, \omega_2, \dots) \in \tilde{\Omega}$. □

1.6 DISINTEGRATION AND CONDITIONAL DENSITIES

Let $f_{X,Y}(x,y)$ be the joint density of $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ as outlined in Section 1.4 above (with $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$), i.e.,

$$P(X \in \mathrm{d}x,\ Y \in \mathrm{d}y) = f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

Then the marginal density

$$f_X(x) = \int_{\mathcal{Y}} f_{X,Y}(x,y)\,\mathrm{d}y \tag{1.10}$$

is the density of X.

Definition 1.22. The conditional density is

$$f(y \mid x) := f_{Y|X=x}(y \mid x) := \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{1.11}$$

sometimes also described by $P(Y \in \mathrm{d}y \mid X = x) = f_{Y|X=x}(y \mid X = x)\,\mathrm{d}y$.

It holds that

$$P(X \in A, Y \in B) = \int_A \int_B f(y \mid x)\,\mathrm{d}y\, f_X(x)\,\mathrm{d}x$$

and thus

$$\mathbb{E}\, g(X,Y) = \int_{\mathcal{X}} \underbrace{\int_{\mathcal{Y}} g(x,y) f(y \mid x)\,\mathrm{d}y}_{=\mathbb{E}\,\left(g(x,Y)|X=x\right)} \cdot f_X(x)\,\mathrm{d}x = \mathbb{E}\,\mathbb{E}\,\big(g(X,Y) \mid X\big).$$

The identity $\mathbb{E}\, g(X,Y) = \mathbb{E}\,\mathbb{E}\,\big(g(X,Y) \mid X\big)$ is known as *law of total expectation*.

1.7 UNIVARIATE RANDOM VARIABLES AND QUANTILES

A random variable with range $\mathcal{X} = \mathbb{R}$ is said to be *univariate*.

Definition 1.23. The *quantile function* or *generalized inverse* is

$$F_X^{-1}(\alpha) = \inf\{x \colon P(X \le x) \ge \alpha\}. \tag{1.12}$$

$m \in \mathbb{R}$ is a *median* if

$$P(X \le m) \ge \frac{1}{2} \text{ and } P(X \ge m) \ge \frac{1}{2}; \tag{1.13}$$

in particular, $F_X^{-1}(1/2)$ is a median. The *quartiles* are $F_X^{-1}(1/4)$ and $F_X^{-1}(3/4)$, the *deciles* $F_X^{-1}(i/10)$. The *interquartile range (IQR)* is

$$F_X^{-1}(3/4) - F_X^{-1}(1/4). \tag{1.14}$$

[8] Constantin Carathéodory, 1873–1950, Greek mathematician

If $\mathcal{X} = \mathbb{R}$, then one may change the variables (use (1.2)) to get

$$\mathbb{E}_P\, g(X) = \int_{\mathbb{R}} g(x)\, \mathrm{d}F(x) = \int_0^1 g\big(F^{-1}(\alpha)\big)\, \mathrm{d}\alpha, \tag{1.15}$$

Further, for every $a \in \mathbb{R}$, one may integrate (1.15) by parts so that

$$\mathbb{E}\, g(X) = g(a) - \int_{-\infty}^{a} g'(x) F(x)\, \mathrm{d}x + \int_a^\infty g'(x)\big(1 - F(x)\big)\, \mathrm{d}x. \tag{1.16}$$

Proposition 1.24. *The mean $\mu := \mathbb{E}\, X$ minimizes the squared deviation $f(c) = \mathbb{E}(X - c)^2$.*

Proof. By linearity, the derivative is $0 = f'(c) = 2\,\mathbb{E}(X - c)$ and hence the assertion. □

Proposition 1.25. *The median m_X minimizes the absolute deviation $f(c) = \mathbb{E}\,|X - c|$.*

Proof. Note that $f(c) = \mathbb{E}\,|X - c| = \int_{-\infty}^{c} P(X \le t)\,\mathrm{d}t + \int_c^\infty P(X > t)\,\mathrm{d}t$ by (1.16). It follows with Leibniz's integral rule that $0 = f'(c) = P(X \le c) - P(X > c)$. The assertion follows together with $P(X \le c) + P(X > c) \ge 1$. □

Corollary 1.26. *It holds that $|\mu - m| \le \sigma$.*

Proof. By Jensen's inequality, $|\mu - m| = |\mathbb{E}\, X - m| \le \mathbb{E}\,|X - m|$. From Proposition 1.25 and Hölder's inequality we conclude further that $\mathbb{E}\,|X - m| \le \mathbb{E}\,|X - \mu| = \mathbb{E}\left(1 \cdot |X - \mu|\right) \le \sqrt{\mathbb{E}\, 1^2} \cdot \sqrt{\mathbb{E}(X - \mu)^2} = \sigma$. □

Definition 1.27 (Quantile loss)**.** For $\alpha \in (0, 1)$ define the nonnegative and convex loss function

$$\ell_\alpha(y) := \begin{cases} -(1 - \alpha)\, y & \text{if } y \le 0, \\ \alpha \cdot y & \text{if } y \ge 0 \end{cases} = \left(\alpha - \frac{1}{2}\right) y + \frac{1}{2}\, |y|\,.$$

Proposition 1.28. *The quantiles satisfy the optimal location problem*

$$F_X^{-1}(\alpha) \in \operatorname*{arg\,min}_{c \in \mathbb{R}} \mathbb{E}\, \ell_\alpha(X - c).$$

Proof. With (1.15) and differentiating (recall the Leibniz integral rule) with respect to c to obtain the first order condition for the minimum it follows that

$$\begin{aligned} 0 &= \frac{\partial}{\partial c}\, \mathbb{E}\, \ell_\alpha(X - c) \\ &= \frac{\partial}{\partial c} \left((1 - \alpha) \int_{-\infty}^{c} c - x\, \mathrm{d}F_X(x) + \alpha \int_c^\infty x - c\, \mathrm{d}F_X(x) \right) \\ &= (1 - \alpha) \int_{-\infty}^{c} \mathrm{d}F_X(x) - \alpha \int_c^\infty \mathrm{d}F_X(x) \\ &= (1 - \alpha) F_X(c) - \alpha\big(1 - F_X(c)\big) \\ &= F_X(c) - \alpha \end{aligned}$$

and hence the result. □

Definition 1.29 (Expectiles, cf. [12])**.** For $\alpha \in [0, 1]$, define the scoring function (loss function) $\ell_\alpha(y) := \begin{cases} (1 - \alpha)\, y^2 & \text{if } y \le 0, \\ \alpha \cdot y^2 & \text{if } y \ge 0, \end{cases}$ The *expectiles* are

$$e_\alpha(X) := \operatorname*{arg\,min}_{c \in \mathbb{R}} \mathbb{E}\, \ell_\alpha(X - c).$$

For $\alpha \geq 1/2$, the expectiles constitute the unique solution of the equation

$$\alpha\ \mathbb{E}(X-c)_+ = (1-\alpha)\,\mathbb{E}(c-X)_+,$$

Definition 1.30 (Huber[9] loss function)**.** The function

$$\ell_\delta(x) := \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leq \delta, \\ \delta\left(|x| - \frac{1}{2}\delta\right) & \text{for } |x| \geq \delta \end{cases}$$

is called *Huber loss function*.

The Huber loss function combines properties of $x \mapsto x^2$ and $x \mapsto |x|$ from above.

1.8 PROBLEMS

Exercise 1.1. *The cdf of a (standard) uniform random variable U is $P(U \leq u) = u$, $u \in [0,1]$. Show that $\mathbb{E}\,u = \frac{1}{2}$ and $\operatorname{var} U = \frac{1}{12}$. Give the pdf, cdf, $\mathbb{E}$ and var for a random variable which is uniform in $[a,b]$.*

Exercise 1.2. *Give the cdf, pdf, expectation and variance of the random variable $\tilde{U} := a + (b-a)U$.*

Exercise 1.3. *Verify the linear transformation, Corollary 1.16.*

Exercise 1.4. *Show that $f_{1/X}(y) = f_X\left(\frac{1}{y}\right) \cdot \frac{1}{y^2}$.*

Exercise 1.5. *Verify (1.16), which assumptions on $g(\cdot)$ do we need?*

Exercise 1.6. *Verify the identity in Definition 1.7.*

Exercise 1.7 (Uniform ratio distribution)**.** *Let U_1 and $U_2 \sim U[0,\,1]$ be independent and uniformly distributed on $[0,\,1]$. Show that $P\left(\frac{U_2}{U_1} \in \mathrm{d}x\right) = f_{U_2/U_1}(x)\,\mathrm{d}x$, where the density of the random quotient $\frac{U_2}{U_1}$ is $f_{U_2/U_1}(x) = \begin{cases} 1/2 & \text{if } x \leq 1, \\ 1/2x^2 & \text{if } x > 1. \end{cases}$*

Exercise 1.8. *Let $U_i \sim U[0,\,1]$ be independent uniforms. Show that*

$$P\big(U_1 \cdot \ldots \cdot U_n \in \mathrm{d}z\big) = \frac{(-\log z)^{n-1}}{(n-1)!}\mathrm{d}z,$$

$z \in (0,1)$.

Exercise 1.9. *Verify the integration by parts formula for Riemann–Stieltjes integrals,*

$$\int_0^1 f(x)\,\mathrm{d}g(x) = f(x)g(x)\big|_{x=0}^1 - \int_0^1 g(x)\,\mathrm{d}f(x).$$

Exercise 1.10. *Compute $\int_0^1 x^3\,\mathrm{d}x^2$.*

Exercise 1.11. *Compute $\iint_{[0,1]^2} x^2 y\,\mathrm{d}^2\,y^2 x$.*

Exercise 1.12 (Riemann–Stieltjes integrals in two dimensions)**.** *Verify that*

[9] Peter Jost Huber, 1934

(i) $\iint_{[0,1]^2} g(x,y)\,\mathrm{d}^2(x\cdot y) = \int_0^1\int_0^1 g(x,y)\,\mathrm{d}x\,\mathrm{d}y$,

(ii) $\iint_{[0,1]^2} g(x,y)\,\mathrm{d}^2\min(x,y) = \int_0^1 g(x,x)\,\mathrm{d}x$ *(comonotonicity) and*

(iii) $\iint_{[0,1]^2} g(x,y)\,\mathrm{d}^2\max(0,x+y-1) = \int_0^1 g(x,1-x)\,\mathrm{d}x$ *(antimonotonicity).*

Exercise 1.13 (Conditional density)**.** *Show that $f(\cdot|y)$ defined in (1.10) and (1.11) are densities.*

Exercise 1.14. *Show that* $\operatorname{var}\begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}\operatorname{var}X & \operatorname{cov}(X,Y)\\ \operatorname{cov}(Y,X) & \operatorname{var}Y\end{pmatrix}$, *where* $\operatorname{cov}(Y,X)=\operatorname{cov}(X,Y)^\top$.

Multinomial distribution

Exercise 1.15 (Multinomial distribution)**.** *The probability mass function (pmf) of the* multinomial distribution *(multivariate binomial distribution)* $\operatorname{bin}(n;p_1,\dots,p_k)$ *(where $\sum_{i=1}^k p_i = 1$) with support*

$$\left\{(i_1,\dots,i_k)\in\mathbb{N}_0^k\colon i_1+\dots+i_k=n\right\}$$

is

$$P\big((X_1,\dots,X_k)=(i_1,\dots,i_k)\big) = \frac{n!}{i_1!\cdot\ldots\cdot i_k!}p_1^{i_1}\cdot\ldots\cdot p_k^{i_k}.$$

Show that

$$\operatorname{bin}(m;p_1,\dots,p_k)+\operatorname{bin}(n;p_1,\dots,p_k)\sim\operatorname{bin}(m+n;p_1,\dots,p_k)$$

for independent random variables.

Exercise 1.16. *Show the marginal distributions $P(X_\ell=i)=\binom{n}{i}p_\ell^i(1-p_\ell)^{n-i}$, $i=0,\dots,n$ (the usual binomial distribution* $\operatorname{bin}(n,p_\ell)$*) for $\ell=1,\dots,k$, and*

$$P(X_\ell=i,X_m=j)=\frac{n!}{i!\,j!\,(n-i-j)!}p_\ell^i p_m^j(1-p_\ell-p_m)^{n-i-j}$$

with support $\left\{(i,j)\in\mathbb{N}_0^2\colon i+j\le n\right\}$.

Exercise 1.17. *Verify the moments*

$$\mathbb{E}\begin{pmatrix}X_1\\X_2\\\vdots\\X_k\end{pmatrix}=n\begin{pmatrix}p_1\\p_2\\\vdots\\p_k\end{pmatrix}\text{ and }\operatorname{var}\begin{pmatrix}X_1\\X_2\\\vdots\\X_k\end{pmatrix}=n\begin{pmatrix}p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k\\ -p_2p_1 & p_2(1-p_2) & \ddots & \vdots\\ \vdots & \ddots & \ddots & -p_{k-1}p_k\\ -p_kp_1 & \dots & -p_kp_{k-1} & p_k(1-p_k)\end{pmatrix}.$$

Show that the covariance matrix is singular *(cf. Remark 1.12). Argue, why X_i and X_j are correlated, why the correlations are negative and why the matrix is singular.*

2 Sample mean and sample variance

> If your experiment needs statistics, you ought to have done a better experiment.
>
> *attributed to* Erneset Rutherford, 1871–1937

Definition 2.1. For random variables X_i, $i = 1, \dots n$, the *sample mean*[1] statistics is

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

(a measure of location) and the (uncorrected) *sample variance* – a measure of spread – is[2]

$$V_n := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2. \tag{2.2}$$

The (Bessel corrected) sample variance is[3]

$$s_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}V_n. \tag{2.3}$$

Remark 2.2. The distinction between s_n^2 and V_n is a common source of confusion. Take care when consulting the literature to determine which convention is used, especially since the uninformative notation s is commonly used for both.

Proposition 2.3. *It holds that (compare with (1.7))*

$$s_n^2 = \frac{1}{2\,n(n-1)}\sum_{i,j=1}^{n}\left(X_i - X_j\right)^2. \tag{2.4}$$

Remark 2.4. Note that there are $n(n-1)$ non-zero summands in (2.4).

[1] Stichprobenmittel

[2] Stichprobenvarianz

[3] korrigierte Stichprobenvarianz

Proof. Indeed,

$$\begin{aligned}\frac{1}{2n^2}\sum_{i,j=1}^n (X_i-X_j)^2 &= \frac{1}{2n^2}\sum_{i,j=1}^n \left(X_i-\overline{X}_n-(X_j-\overline{X}_n)\right)^2\\
&= \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (X_i-\overline{X}_n)^2-2(X_i-\overline{X}_n)\cdot(X_j-\overline{X}_n)+(X_j-\overline{X}_n)^2\\
&= \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (X_i-\overline{X}_n)^2-\frac{1}{n^2}\sum_{i=1}^n (X_i-\overline{X}_n)\cdot\sum_{j=1}^n (X_j-\overline{X}_n)+\frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (X_j-\overline{X}_n)^2\\
&= \frac{n}{2n^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2+0+\frac{n}{2n^2}\sum_{j=1}^n (\overline{X}_n-X_j)^2=\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2=V_n.\end{aligned}$$

Multiply with $\frac{n}{n-1}$ to get the assertion. □

Remark 2.5. We have that $s_n^2=0$ iff $X_i=X_j$ for all $i,\ j=1,\dots,n$.

Proposition 2.6 (Algebraic formula for the variance,[4] cf. Exercise 2.6)**.** *For every (sic!) $\xi\in\mathbb{R}$ (thus for $\xi=\mu=\mathbb{E}\,X$) it holds that*

$$V_n=\frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2=\frac{1}{n}\sum_{i=1}^n (X_i-\xi)^2-\left(\overline{X}_n-\xi\right)^2. \tag{2.5}$$

Particularly ($\xi=0$) it holds that $V_n=\frac{1}{n}\sum_{i=1}^n X_i^2-\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$, i.e., $V_n=\mathbb{E}_{P_n}X^2-\left(\mathbb{E}_{P_n}X\right)^2$ for the empirical measure $P_n(\cdot)=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}(\cdot)$.

Proof. Write

$$\begin{aligned}V_n &= \frac{1}{n}\sum_{i=1}^n\left(X_i-\xi-(\overline{X}_n-\xi)\right)^2=\frac{1}{n}\sum_{i=1}^n (X_i-\xi)^2-2(X_i-\xi)(\overline{X}_n-\xi)+(\overline{X}_n-\xi)^2\\
&= \frac{1}{n}\sum_{i=1}^n (X_i-\xi)^2-(\overline{X}_n-\xi)^2,\end{aligned}$$

and hence the assertion. □

Proposition 2.7. *By involving the sample $X=(X_1,\dots,X_n)$ explicitly in the notation it holds that*

(i) *$\overline{\alpha X+\beta}=\alpha\overline{X}+\beta$ (the sample mean is linear),*

(ii) *$s_n^2(\alpha X)=\alpha^2\cdot s_n^2(X)$ and*

(iii) *$s_n^2(X+c)=s_n^2(X)$ (the sample variance is shift-invariant).*

2.1 PROPERTIES OF THE SAMPLE MEAN AND SAMPLE VARIANCE ESTIMATORS

Some content follows http://www.randomservices.org/random/sample.

Proposition 2.8. *Let X_i be uncorrelated with $\mathbb{E}\,X_i=:\mu$ and $\sigma^2:=\operatorname{var}X_i<\infty$, $i=1,\dots,n$. Then*

[4] Steinerscher Verschiebungssatz

(i) $\mathbb{E}\,\overline{X}_n = \mu$ *and*

(ii) $\operatorname{var}\overline{X}_n = \frac{\sigma^2}{n} \quad (= \mathcal{O}(1/n))$.

Proof. By linearity, $\mathbb{E}\,\overline{X}_n = \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,X_i = \mu$. Further, $\operatorname{cov}(X_i, X_j) = \begin{cases}\sigma^2 & \text{if } i = j,\\ 0 & \text{else}\end{cases}$ and thus

$$\begin{aligned}\operatorname{var}\overline{X}_n &= \mathbb{E}\left(\overline{X}_n - \mathbb{E}\,\overline{X}_n\right)^2 = \mathbb{E}\left(\overline{X}_n - \mu\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right)^2\\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}(X_i - \mu)(X_j - \mu) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \operatorname{cov}(X_i, X_j) = \frac{n}{n^2}\operatorname{var} X = \frac{\sigma^2}{n}.\end{aligned}$$

Hence the assertion. □

Proposition 2.9. *Let $m \le n$ and X_i, $i = 1, \dots, n$ uncorrelated. Then*

(i) $\operatorname{cov}\left(\overline{X}_m, \overline{X}_n\right) = \frac{\sigma^2}{n}$,

(ii) $\operatorname{corr}\left(\overline{X}_m, \overline{X}_n\right) = \sqrt{\frac{m}{n}}$ *(≤ 1) and*

(iii) $\mathbb{E}\,\overline{X}_m \cdot \overline{X}_n = \mu^2 + \frac{\sigma^2}{n}$.

Proof. As X_i, $i > m$, are independent from $\overline{X}_m$ it follows with (1.6) that

$$\begin{aligned}\operatorname{cov}\left(m\,\overline{X}_m, n\,\overline{X}_n\right) &= \operatorname{cov}\left(m\,\overline{X}_m, m\,\overline{X}_m\right) + \operatorname{cov}\left(m\,\overline{X}_m, n\,\overline{X}_n - m\,\overline{X}_m\right)\\ &= \operatorname{var}\left(m\,\overline{X}_m\right) + 0 = m^2\frac{\sigma^2}{m} = m\,\sigma^2,\end{aligned}$$

hence $\operatorname{cov}\left(\overline{X}_m, \overline{X}_n\right) = \frac{\sigma^2}{n}$.

Further, $\operatorname{corr}\left(\overline{X}_m, \overline{X}_n\right) = \frac{\frac{\sigma^2}{n}}{\sqrt{\frac{\sigma^2}{m}\frac{\sigma^2}{n}}} = \sqrt{\frac{m}{n}}$ and $\mathbb{E}\,\overline{X}_m \cdot \overline{X}_n = \operatorname{cov}\left(\overline{X}_m, \overline{X}_n\right) + \mathbb{E}\,\overline{X}_m \cdot \mathbb{E}\,\overline{X}_m = \frac{\sigma^2}{n} + \mu^2$. □

Proposition 2.10. *The moment generating function of $\overline{X}_n$ is $m_{\overline{X}_n}(t) = \mathbb{E}\,e^{t\,\overline{X}_n} = \left(\mathbb{E}\,e^{tX_i}\right)^n = m_{X_i}(\frac{t}{n})^n$.*

Proof. Indeed, $m_{\overline{X}_n}(t) = \mathbb{E}\,e^{t\,\overline{X}_n} = \mathbb{E}\,e^{\frac{t}{n}\sum_{i=1}^n X_i} = \mathbb{E}\prod_{i=1}^n e^{\frac{t}{n}X_i} = \prod_{i=1}^n \mathbb{E}\,e^{\frac{t}{n}X_i} = m_{X_i}\left(\frac{t}{n}\right)^n$. □

Proposition 2.11. *Let X_i be independent and identically distributed (iid) with central 4th moment $\mu_4 := \mathbb{E}\,(X - \mathbb{E}\,X)^4 < \infty$.*[5] *Then*

(i) $\mathbb{E}\,V_n = \frac{n-1}{n}\sigma^2$ *and*

(ii) $\operatorname{var} V_n = \frac{(n-1)^2}{n^3}\mu_4 - \frac{(n-1)(n-3)}{n^3}\sigma^4 \quad (= \mathcal{O}(1/n))$.

Corollary 2.12. *It holds that*

(i) $\mathbb{E}\,s_n^2 = \sigma^2$,

[5] The dimensionless quantity $K := \mu_4/\sigma^4$ is called *kurtosis* (Wölbung, Kurtosis, dt.).

(ii) $\operatorname{var} s_n^2 = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$, *but*

(iii) $\mathbb{E}\sqrt{V_n} < \mathbb{E}\, s_n \le \sigma$, *where* $s_n := \sqrt{s_n^2}$ *is the (uncorrected) sample standard deviation.*

By (iii), V_n and s_n are *negatively biased estimators* that tend to *underestimate* σ.

Proof of the corollary. The first assertion is immediate. Note then that $V_n = \frac{n-1}{n} s_n^2 < s_n^2$ and by monotonicity of $\sqrt{\cdot}$ thus $\sqrt{V_n} < s_n$, (almost) everywhere. From Jensen's inequality for the convex function $\varphi\colon x \mapsto x^2$ we finally get $(\mathbb{E}\, s_n)^2 = \varphi(\mathbb{E}\, s_n) \le \mathbb{E}\,\varphi(s_n) = \mathbb{E}\, s_n^2 = \sigma^2$. □

Proof. Use Steiner's theorem (2.5) with $\mu = \mathbb{E}\, X_i$. Then the expectation is given by

$$\begin{aligned}\mathbb{E}\, V_n &= \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)\right)^2\\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,(X_i-\mu)^2 - \mathbb{E}\,\frac{1}{n^2}\sum_{i,j=1}^n (X_i-\mu)(X_j-\mu) = \sigma^2 - \frac{n}{n^2}\sigma^2 = \frac{n-1}{n}\sigma^2.\end{aligned}$$

From (2.4) we deduce that

$$\operatorname{var} s_n^2 = \operatorname{cov}\left(s_n^2, s_n^2\right) = \frac{1}{4n^2(n-1)^2}\sum_{i,j,k,\ell=1}^n \operatorname{cov}\left((X_i-X_j)^2, (X_k-X_\ell)^2\right).$$

In view of (2.5) we may assume $\mu = \mathbb{E}\, X_i = 0$. Now we have that

(i) $\operatorname{cov}\left((X_i-X_j)^2, (X_k-X_\ell)^2\right) = 0$ if $i = j$ or $k = \ell$; there are $2n^3 - n^2$ such terms.

(ii) $\operatorname{cov}\left((X_i-X_j)^2, (X_k-X_\ell)^2\right) = 0$ if i, j, k and ℓ are all distinct; there are $n(n-1)(n-2)(n-3)$ such terms.

(iii) $\operatorname{cov}\left((X_i-X_j)^2, (X_k-X_\ell)^2\right) = 2\mu_4 + 2\sigma^4$ if $i \ne j$ and $\{k,\ell\} = \{i,j\}$. Indeed

$$\begin{aligned}\operatorname{var}\left((X_i-X_j)^2\right) &= \mathbb{E}(X_i-X_j)^4 - \left(\mathbb{E}(X_i-X_j)^2\right)^2\\ &= \mathbb{E}\left(X_i^4 - 4X_i^3X_j + 6X_i^2X_j^2 - 4X_iX_j^3 + X_j^4\right) - \left(\mathbb{E}\, X_i^2 - 2X_iX_j + X_j^2\right)^2\\ &= 2\mu_4 + 6\sigma^4 - (2\sigma^2)^2 = 2\mu_4 + 2\sigma^4;\end{aligned}$$

there are $2n(n-1)$ such terms.

(iv) $\operatorname{cov}\left((X_i-X_j)^2, (X_k-X_\ell)^2\right) = \mu_4 - \sigma^4$ if $i \ne j$, $k \ne \ell$ and $\#\left(\{i,j\} \cap \{k,\ell\}\right) = 1$. Indeed,

$$\begin{aligned}\operatorname{cov}&\left((X_i-X_j)^2, (X_i-X_\ell)^2\right) = \mathbb{E}(X_i-X_j)^2(X_i-X_\ell)^2 - \mathbb{E}(X_i-X_j)^2\,\mathbb{E}(X_i-X_\ell)^2\\ &= \mathbb{E}\left(X_i^4 - 2X_i^3X_j + X_i^2X_j^2 - 2X_i^3X_\ell + 4X_i^2X_jX_\ell - 2X_iX_j^2X_\ell + X_i^2X_\ell^2 - 2X_iX_jX_\ell^2 + X_j^2X_\ell^2\right)\\ &\quad - \left(\mathbb{E}\, X_i^2 - 2X_iX_j + X_j^2\right)^2\\ &= \mu_4 + 3\sigma^4 - (2\sigma^2)^2 = \mu_4 - \sigma^4;\end{aligned}$$

and there are $4n(n-1)(n-2)$ such terms. By collecting terms we get that

$$\operatorname{var} s_n^2 = \frac{4n(n-1)(1+n-2)\mu_4 + 4n(n-1)(1-n+2)\sigma^4}{4n^2(n-1)^2} = \frac{1}{n}\Big(\mu_4 - \frac{n-3}{n-1}\sigma^4\Big).$$

□

Proposition 2.13. *It holds that*

(i) $\operatorname{cov}\left(\overline{X}_n, s_n^2\right) = \frac{\mu_3}{n}$ *with central third moment*[6] $\mu_3 := \mathbb{E}(X - \mathbb{E}\,X)^3$*, or*

(ii) $\operatorname{corr}\left(\overline{X}_n, s_n^2\right) = \frac{\mu_3}{\sigma\sqrt{\mu_4 - \frac{n-3}{n-1}\sigma^4}}$.

Proof. With (2.4),

$$\operatorname{cov}\bigl(\overline{X}_n, s_n^2\bigr) = \frac{1}{2n^2(n-1)} \sum_{i=1}^n \sum_{j,k=1}^n \operatorname{cov}\bigl(X_i, (X_j - X_k)^2\bigr).$$

Now we have that

(i) $\operatorname{cov}\left(X_i, (X_j - X_k)^2\right) = 0$ if $j = k$; there are n^2 such terms;

(ii) $\operatorname{cov}\left(X_i, (X_j - X_k)^2\right) = 0$ if i, j, k are all distinct; there are $n(n-1)(n-2)$ such terms.

Finally

(iii) $\operatorname{cov}\left(X_i, (X_j - X_k)^2\right) = \mu_3$ if $j \neq k$ and $i \in \{j, k\}$: indeed

$$\begin{aligned}\operatorname{cov}\Bigl(X_j, (X_j - X_k)^2\Bigr) &= \mathbb{E}\,X_j(X_j - X_k)^2 - \mathbb{E}\,X_j \cdot \mathbb{E}(X_j - X_k)^2 \\ &= \mathbb{E}\,X_j^3 - 2\,\mathbb{E}\,X_j^2 X_k + \mathbb{E}\,X_j X_k^2 - \mathbb{E}\,X_j\,\mathbb{E}\,X_j^2 + 2\,\mathbb{E}\,X_j\,\mathbb{E}\,X_j X_k - 2\,\mathbb{E}\,X_j\,\mathbb{E}\,X_k^2 \\ &= \mathbb{E}\,X_j^3 = \mu_3;\end{aligned}$$

there are $2n(n-1)$ such terms.

□

2.2 COVARIANCE AND CORRELATION

Definition 2.14. The (corrected) *sample covariance* of random variables (X_i, Y_i) is (cf. (1.7))

$$q_n := \frac{1}{2n(n-1)} \sum_{i,j=1}^n \bigl(X_i - X_j\bigr)\bigl(Y_i - Y_j\bigr)^\top. \tag{2.6}$$

Remark 2.15. Exercise 2.7 addresses a generalization for a weighted sample.

Proposition 2.16. *It holds that (cf. (2.4) and Exercise 2.8)*

$$q_n = \frac{1}{n-1} \sum_{i=1}^n \bigl(X_i - \overline{X}_n\bigr) \cdot \bigl(Y_i - \overline{Y}_n\bigr)^\top. \tag{2.7}$$

In what follows we discuss univariate random variables only, i.e., $X, Y \in \mathbb{R}$ as the generalizations are obvious.

[6] The dimensionless quantity $S := \mu_3/\sigma^3$ is called *skewness* (Schiefe, dt).

Proposition 2.17 (Parallel computation of the mean and the sample variance)**.** *Let A, B be a partition of the sample, i.e., sets with $A\dot\cup B = \{1, 2, \dots, n\}$ and $n_A + n_B = n$. With an obvious generalization in notation it holds that*[7]

$$\begin{aligned}\overline{X}_n &= \frac{n_A\,\overline{X}_A + n_B\,\overline{X}_B}{n_A + n_B}, \\ s_n^2 &= \frac{n_A - 1}{n-1}s_A^2 + \frac{n_B - 1}{n-1}s_B^2 + \frac{n_A \cdot n_B}{n\,(n-1)}\big(\overline{X}_A - \overline{X}_B\big)^2 \text{ and} \\ q_n &= \frac{n_A - 1}{n-1}q_A + \frac{n_B - 1}{n-1}q_B + \frac{n_A \cdot n_B}{n\,(n-1)}\big(\overline{X}_A - \overline{X}_B\big)\big(\overline{Y}_A - \overline{Y}_B\big).\end{aligned} \tag{2.8}$$

Proof. We have with (2.8) that

$$\begin{aligned}V_n &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{n_A}{n}\overline{X}_A - \frac{n_B}{n}\overline{X}_B\right)^2 \\ &= \frac{n_A}{n}\frac{1}{n_A}\sum_{i\in A}\left(X_i - \overline{X}_A + \frac{n_B}{n}\big(\overline{X}_A - \overline{X}_B\big)\right)^2 + \frac{n_B}{n}\frac{1}{n_B}\sum_{i\in B}\left(X_i - \overline{X}_B - \frac{n_A}{n}\big(\overline{X}_A - \overline{X}_B\big)\right)^2.\end{aligned}$$

It follows from (2.5) with $\xi = \overline{X}_A - \frac{n_B}{n}\big(\overline{X}_A - \overline{X}_B\big)$ (and $\xi = \overline{X}_B + \frac{n_A}{n}\big(\overline{X}_A - \overline{X}_B\big)$ for the second sum, resp.) that

$$\begin{aligned}V_n &= \frac{n_A}{n}\left(V_A + \left(\frac{n_B}{n}\big(\overline{X}_A - \overline{X}_B\big)\right)^2\right) + \frac{n_B}{n}\left(V_B + \left(\frac{n_A}{n}\big(\overline{X}_A - \overline{X}_B\big)\right)^2\right) \\ &= \frac{n_A}{n}V_A + \frac{n_B}{n}V_B + \left(\frac{n_A\,n_B^2}{n^3} + \frac{n_B\,n_A^2}{n^3}\right)\big(\overline{X}_A - \overline{X}_B\big)^2 \\ &= \frac{n_A}{n}V_A + \frac{n_B}{n}V_B + \frac{n_A\,n_B}{n^2}\big(\overline{X}_A - \overline{X}_B\big)^2\end{aligned}$$

and thus the second assertion. The remaining assertion follows by an obvious modification of the preceding proof using Exercise 2.6. □

Corollary 2.18 (Update formulae, cf. Exercise 2.4)**.** *It holds that*

$$\begin{aligned}\overline{X}_{n+1} &= \overline{X}_n + \frac{X_{n+1} - \overline{X}_n}{n+1}, \\ \overline{Y}_{n+1} &= \overline{Y}_n + \frac{Y_{n+1} - \overline{Y}_n}{n+1}, \\ s_{n+1}^2 &= \frac{n-1}{n}s_n^2 + \frac{n+1}{n^2}\big(X_{n+1} - \overline{X}_{n+1}\big)^2 \text{ for } n \ge 1 \text{ and} \\ q_{n+1} &= \frac{n-1}{n}q_n + \frac{n+1}{n^2}\big(X_{n+1} - \overline{X}_{n+1}\big)\big(Y_{n+1} - \overline{Y}_{n+1}\big) \text{ for } n \ge 1,\end{aligned}$$

where s_1^2 and $q_1 \in \mathbb{R}$ are arbitrary.

Proposition 2.19 (Cf. Proposition 2.12)**.** *Let (X_i, Y_i) be independent and identically distributed (iid), then*[8]

$$\mathbb{E}\, q_n = \operatorname{cov}(X, Y).$$

[7] Similar formulae can be shown for higher moments.

[8] Note, that (X_i, Y_i) is independent from (X_j, Y_j) for $i \neq j$, but X_i and Y_i are correlated.

Proof. Use (2.6) or Exercise 2.6 below. □

Definition 2.20 (Pearson)**.** Pearson's correlation coefficient is

$$\rho = \rho_{X,Y} := \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)}{\sqrt{\mathbb{E}(X-\mathbb{E}Y)^2}\cdot\sqrt{\mathbb{E}(Y-\mathbb{E}Y)^2}}.$$

The sample correlation is

$$r_n = r_{X,Y;n} := \frac{q_n}{s_{X;n}\cdot s_{Y;n}} = \frac{\sum_{i=1}^n \left(X_i-\overline{X}_n\right)\left(Y_i-\overline{Y}_n\right)}{\sqrt{\sum_{i=1}^n \left(X_i-\overline{X}_n\right)^2}\cdot\sqrt{\sum_{i=1}^n \left(Y_i-\overline{Y}_n\right)^2}}. \tag{2.9}$$

Equivalent expressions for the sample correlation include

$$r_{X,Y} = \frac{1}{n}\sum_{i=1}^n \frac{X_i-\overline{X}_n}{V_{X;n}}\frac{Y_i-\overline{Y}_n}{V_{Y;n}} = \frac{1}{n-1}\sum_{i=1}^n \frac{X_i-\overline{X}_n}{s_{X;n}}\frac{Y_i-\overline{Y}_n}{s_{Y;n}}.$$

Proposition 2.21. *If X_i and Y_i are independent, then $\mathbb{E}\, r_n = 0$ and $\operatorname{var} r_n = \frac{1}{n-1}$.*

Definition 2.22 (Kendall's τ)**.** Kendall's rank correlation coefficient[9] is

$$\begin{aligned}\rho_\tau &:= P\big(\underbrace{(X-\tilde{X})(Y-\tilde{Y})>0}_{\text{concordant}}\big) - P\big(\underbrace{(X-\tilde{X})(Y-\tilde{Y})<0}_{\text{discordant}}\big) \\ &= \mathbb{E}\operatorname{sign}\left((X-\tilde{X})(Y-\tilde{Y})\right),\end{aligned} \tag{2.10}$$

where $(\tilde{X},\tilde{Y})$ is an independent copy of (X,Y).

Remark 2.23. Figure 2.1 displays various ranks of programming languages.

Definition 2.24. The estimator for Kendall's rank correlation coefficient for independent pairs (X_i,Y_i) is

$$r_\tau := \frac{1}{n(n-1)}\sum_{i\neq j}^n \operatorname{sign}(X_i-X_j)\cdot\operatorname{sign}(Y_i-Y_j),$$

cf. (2.7) and Figure 2.1.

Remark 2.25. Kendall's rank correlation coefficient ρ_τ as well as the estimator r_τ satisfy $-1\le\rho_\tau,\ r_\tau\le 1$.

Definition 2.26 (Spearman's rank correlation coefficient ρ_s)**.** Spearman's[10] rho is

$$\rho_s := \operatorname{corr}\big(F_X(X), F_Y(Y)\big).$$

Definition 2.27. The rank is

$$\operatorname{rg}(X_{(i)}) := i, \tag{2.11}$$

i.e., $X_{\operatorname{rg}(X_i)} = X_{(i)}$, where $X_{(1)}\le X_{(2)}\le\dots\le X_{(n)}$. Spearman's correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables (cf. Exercise 2.9),

$$r_s := \operatorname{corr}\big(\operatorname{rg}(X_i),\operatorname{rg}(Y_i)\big) = \frac{\operatorname{cov}\big(\operatorname{rg}(X_i),\operatorname{rg}(Y_i)\big)}{\sqrt{\operatorname{var}\operatorname{rg}(X_i)}\cdot\sqrt{\operatorname{var}\operatorname{rg}(Y_i)}}.$$

[9] Maurice Kendall, 1907–1983, British statistician

[10] Charles Spearman, 1863–1945, English psychologist

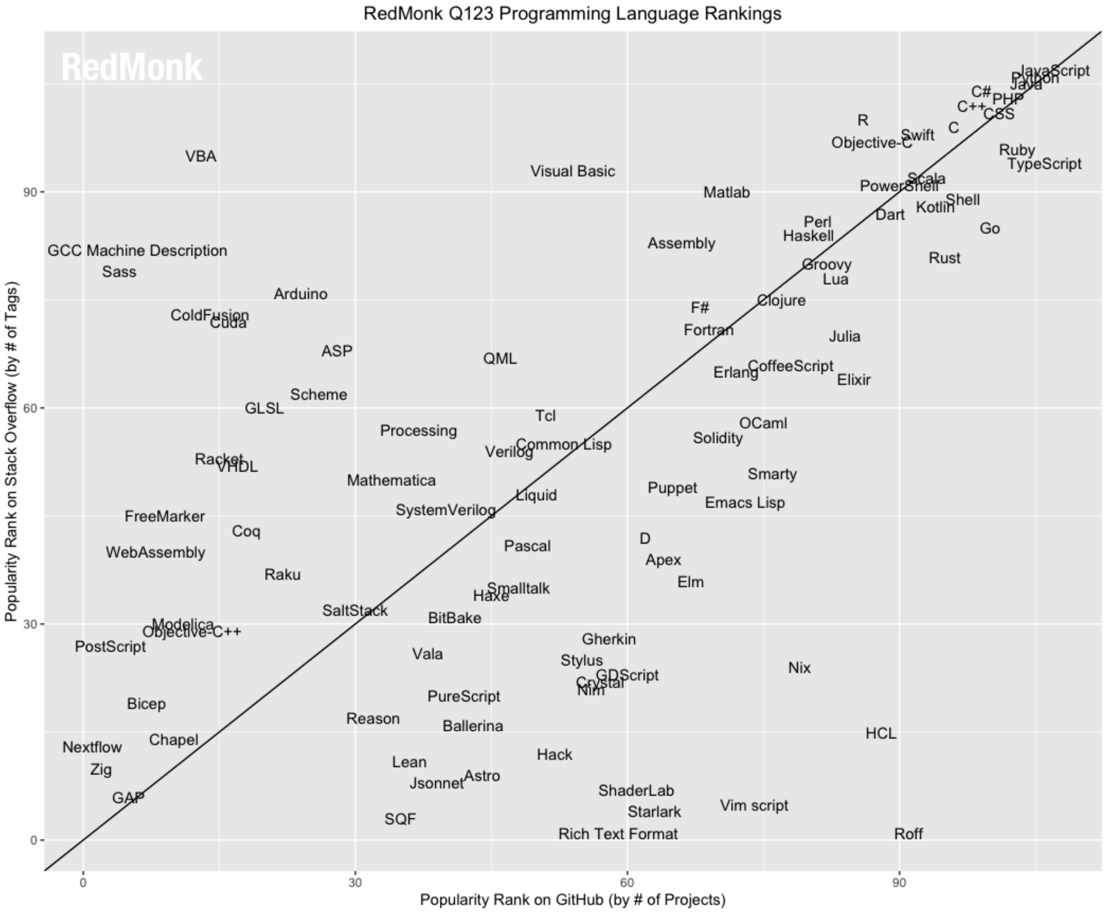


Figure 2.1: Ranking of important programming languages, cf. https://redmonk.com

Remark 2.28. Spearman's rank coefficient is more robust than Pearson's correlation coefficient.

Proposition 2.29. *Let $C(x,y)$ be the copula of (X,Y), i.e., $P(X \le x,\, Y \le y) = C\big(F_X(x), F_Y(y)\big)$, then*

(i) $\rho_\tau = 4\int_0^1 \int_0^1 C(u,v)\,\mathrm{d}C(u,v) - 1$ *and*

(ii) $\rho_s = 12\int_0^1 \int_0^1 \big(C(u,v) - u\cdot v\big)\,\mathrm{d}u\mathrm{d}v = 12\int_0^1 \int_0^1 C(u,v)\,\mathrm{d}u\mathrm{d}v - 3.$

Proof. It follows from (2.10) that $\rho_\tau = 2P\left((X-\tilde{X})(Y-\tilde{Y}) > 0\right) - 1$ and by interchanging the pairs $(\tilde{X},\tilde{Y})$ and (X,Y) that

$$\begin{aligned}\rho_\tau &= 4P\left(X \le \tilde{X},\, Y \le \tilde{Y}\right) - 1\\ &= 4\,\mathbb{E}\,P\left(X \le \tilde{X},\, Y \le \tilde{Y}\,\middle|\,\tilde{X},\tilde{Y}\right) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X \le x, Y \le y)\,\mathrm{d}F(x,y) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C\left(F_X(x),F_Y(y)\right)\,\mathrm{d}C\left(F_X(x),F_Y(y)\right) - 1\\ &= 4\int_0^1\int_0^1 C(u,v)\,\mathrm{d}C(u,v) - 1\end{aligned}$$

and thus (i).

As for Spearman's rho recall that $F_X(X) \sim U[0,1]$ which has variance $\operatorname{var} U = \frac{1}{12}$. Formula (ii) follows from Hoeffding's formula, Proposition 1.14. □

2.3 IF μ WERE KNOWN...

Definition 2.30. The *mean squared error function* and the *mean absolute error function* are

$$\operatorname{mse}(X;\,\xi) \coloneqq \frac{1}{n}\sum_{i=1}^n (X_i - \xi)^2, \qquad \operatorname{mae}(X;\,\xi) \coloneqq \frac{1}{n}\sum_{i=1}^n |X_i - \xi|\,.$$

For convenience, we set $W_n^2 \coloneqq \operatorname{mse}(\mu) = \sum_{i=1}^n (X_i-\mu)^2$ with $\mu = \mathbb{E}\,X$.

Proposition 2.31. *It holds that*

(i) $\mathbb{E}\,W_n^2 = \sigma^2$,

(ii) $\operatorname{var} W_n^2 = \frac{1}{n}\left(\mu_4 - \sigma^4\right)$,

(iii) $W_n^2 \longrightarrow \sigma^2$ *as* $n \to \infty$ *with probability* 1,

(iv) $\sqrt{n}\frac{W_n^2 - \sigma^2}{\mu_4 - \sigma^4} \longrightarrow \mathcal{N}(0,1)$ *in distribution;*

(v) $\mathbb{E}\,W_n \le \sigma$, *i.e.,* W_n *is negatively biased and tends to underestimate* σ.

Proof. $\operatorname{var}(X_i-\mu)^2 = \mathbb{E}(X_i-\mu)^4 - \left(\mathbb{E}(X_i-\mu)^2\right)^2 = \mu_4 - \sigma^4.$ □

Proposition 2.32. *Covariance and correlation are* $\operatorname{cov}\left(\overline{X}_n, W_n^2\right) = \frac{\mu_3}{n}$ *and* $\operatorname{corr}\left(\overline{X}_n, W_n^2\right) = \frac{\mu_3}{\sqrt{\sigma^2(\mu_4-\sigma^4)}}$.

Proof. By independence, $\operatorname{cov}\left(\overline{X}_n, W_n^2\right) = \sum_i^n \frac{1}{n^2}\operatorname{cov}\left(X_i, (X_i-\mu)^2\right)$. But

$$\operatorname{cov}\left(X_i, (X_i-\mu)^2\right) = \operatorname{cov}\left(X_i-\mu, (X_i-\mu)^2\right) = \mathbb{E}(X_i-\mu)^3 - \mathbb{E}(X_i-\mu)\,\mathbb{E}(X_i-\mu)^2 = \mu_3.$$

□

Lemma 2.33. *Additional knowledge is helpful: it holds that* $\operatorname{var} W_n < \operatorname{var} s_n^2$.

Proof. Indeed, $\operatorname{var} s_n^2 - \operatorname{var} W_n = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) - \frac{1}{n}\left(\mu_4-\sigma^4\right) = \frac{2\sigma^4}{n(n-1)} > 0$. □

2.4 PROBLEMS

Exercise 2.1 (Empirical distribution). *For $x_1,\dots,x_N$ given, let X have the discrete distribution $P(X=x_i)=\frac{1}{N}$. Show that $\mathbb{E}\,X = \overline{x}_N$ and $\operatorname{var} X = V_N$. Does the result contradict (i) in Proposition 2.11?*

Exercise 2.2 (Cf. Rüschendorf [17, Beispiel 2.1.3]). *Let $x_1,\dots,x_N$ be given and draw a sample $X_1,\dots,X_n$ ($n\le N$) successively,* without replacement *(i.e., the vector $(X_1,\dots,X_n)$ is iid and uniformly distributed on $x_1,\dots,x_N$; note, however, that $X_1,X_2,\dots,X_n$ are* not *independent). Verify that the estimator $\hat\mu_n := \frac{1}{n}\sum_{i=1}^n X_i$ satisfies*

(i) $\mathbb{E}\,\hat\mu_n = \frac{1}{N}\sum_{k=1}^N x_k =: \mu$ *and*

(ii) $\operatorname{var}\hat\mu_n = \frac{N-n}{N-1}\frac{\tau^2}{n}$, *where* $\tau^2 := \frac{1}{N}\sum_{k=1}^N (x_k-\mu)^2$.

Discuss the result for n small ($n \ll N$) and large ($n=N$, $n=N-1$, etc.).

Hint: observe that $\operatorname{var} X_i = \tau^2$, $\operatorname{cov}(X_i,X_j) = -\frac{\tau^2}{N-1}$, *$i\ne j$, and follow the proof of Proposition 2.8.*

Exercise 2.3. *Show that $V_1 = 0$ and $\sqrt{V_2} = \frac{1}{2}\left|X_2-X_1\right|$.*

Exercise 2.4. *Verify the update formulae in Corollary 2.18.*

Exercise 2.5. *Show that* $\frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)\cdot\left(Y_i-\overline{Y}_n\right) = \frac{1}{n}\sum_{i=1}^n X_iY_i - \overline{X}_n\overline{Y}_n$.

Exercise 2.6 (Generalization of Theorem 2.5 (Steiner)). *Each sample (X_i,Y_i) is assigned a weight w_i. Verify that for arbitrary $\xi,\eta\in\mathbb{R}$ and normalized weights w_i (i.e., $\sum_{i=1}^n w_i = 1$)*

$$\sum_{i=1}^n w_i\left(X_i-\overline{X}_n^w\right)\left(Y_i-\overline{Y}_n^w\right) = \sum_{i=1}^n w_i\,(X_i-\xi)\,(Y_i-\eta) - \left(\overline{X}_n^w-\xi\right)\left(\overline{Y}_n^w-\eta\right), \tag{2.12}$$

where $\overline{X}_n^w := \sum_{i=1}^n w_i\,X_i$ and $\overline{Y}_n^w := \sum_{i=1}^n w_i\,Y_i$ are the weighted means.

Exercise 2.7. *Show that the estimator*

$$q_n^w := \frac{1}{1-\sum_{i=1}^n w_i^2}\sum_{i=1}^n w_i\big(X_i-\overline{X}_n^w\big)\big(Y_i-\overline{Y}_n^w\big)$$

is an unbiased estimator for $\operatorname{cov}(X,Y)$ *for a weighted iid sample $(X_i,Y_i)\sim(X,Y)$. (Hint: use Exercise 2.6 with weights and follow the proof of Proposition 2.11(i).)*

Compare with (2.7).

Exercise 2.8. *Verify (2.7).*

Exercise 2.9. *Show that* $\overline{\operatorname{rg}(X_i)} = \frac{n+1}{2}$ *and* $\operatorname{var}\operatorname{rg}(X_i) = \frac{n^2-1}{12}$, *if all* $\operatorname{rg}(X_i)$ *are distinct.*[11]

Exercise 2.10. *Show that Spearman's coefficient is*

$$r_s = \frac{6}{n^2\left(n^2-1\right)} \sum_{i,j=1}^{n} \left(\operatorname{rg}(X_i) - \operatorname{rg}(X_j)\right) \left(\operatorname{rg}(Y_i) - \operatorname{rg}(Y_j)\right) = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} d_i^2, \tag{2.13}$$

where $d_i \coloneqq \operatorname{rg}(Y_i) - \operatorname{rg}(X_i)$ *is the difference between the two ranks of each observation and all* $\operatorname{rg}(X_i)$*, as well as all* $\operatorname{rg}(Y_i)$ *are distinct.*

Exercise 2.11. *Show that Spearman's coefficient is*

(i) $r_s = 1$, *iff* $\operatorname{rg}(X_i) = \operatorname{rg}(Y_i)$ *for* $i = 1, \dots, n$ *and*

(ii) $r_s = -1$, *iff* $\operatorname{rg}(X_i) + \operatorname{rg}(Y_i) = n + 1$, $i = 1, \dots, n$.

Exercise 2.12. *Derive from Markov's inequality that*

$$P\left(\left|\overline{X}_n - \mu\right| \ge \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2 n} = O\left(\frac{1}{n}\right)$$

and

$$P\left(\left|s_n^2 - \sigma^2\right| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2 n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) = O\left(\frac{1}{n}\right).$$

[11] Recall that $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$.

Normal distribution 3

The law would have been personified by the Greeks and deified, if they had known of it.

Francis Galton, 1822–1911, on the
Central Limit Theorem

3.1 UNIVARIATE NORMAL DISTRIBUTION

An $\mathbb{R}$-valued random variable X with parameters μ and σ^2 is normally distributed (Gaussian,[1] $\mathcal{N}\left(\mu,\sigma^2\right)$) if its density is $\varphi_{\mu,\sigma}(\cdot) \coloneqq \frac{1}{\sigma}\varphi\left(\frac{\cdot-\mu}{\sigma}\right)$, where

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{3.1}$$

is the bell curve. The standard normal distribution is $Z \sim \mathcal{N}(0,1)$, its cdf. is $\Phi(x) \coloneqq \int_{-\infty}^{x} \varphi(z)\,\mathrm{d}z$ (cf. Table 3.1).

Remark 3.1 (Random variable generation). See Algorithm 1 and 2 below (page 49) to efficiently generate Gaussian variables.

Remark 3.2 (φ is a density). For completeness we include a proof that $\varphi(\cdot)$ is a density.[2] Indeed, by Fubini's law,

$$\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z\right)^2 = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,\mathrm{d}x \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,\mathrm{d}y = \iint_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y. \tag{3.2}$$

Employ polar coordinates, i.e., $\begin{pmatrix} x(r,\varphi) \\ y(r,\varphi) \end{pmatrix} \coloneqq \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}$ with Jacobian $\det\begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} = r$. By changing the variables (integration by substitution),

$$(3.2) = \int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2}\cdot r\,\mathrm{d}r\,\mathrm{d}\varphi = \int_0^{\infty} r\,e^{-\frac{1}{2}r^2}\,\mathrm{d}r \cdot \int_0^{2\pi} 1\,\mathrm{d}\varphi = -e^{-\frac{1}{2}r^2}\Big|_{r=0}^{\infty} \cdot \varphi\big|_{\varphi=0}^{2\pi} = 1\cdot 2\pi, \tag{3.3}$$

thus the result.

The variance of a normal distribution is (use[3] $\varphi'(z) = -z\cdot\varphi(z)$ and the product rule)

$$\begin{aligned} \operatorname{var} X &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x \underset{x\leftarrow\mu+\sigma z}{=} \int_{-\infty}^{\infty} \sigma^2 z^2 \varphi(z)\,\mathrm{d}z \\ &= -\sigma^2\int_{-\infty}^{\infty} z\,\varphi'(z)\,\mathrm{d}z = -\,\sigma^2 z\,\varphi(z)\big|_{y=-\infty}^{\infty} + \sigma^2\int_{-\infty}^{\infty} 1\cdot\varphi(z)\,\mathrm{d}z = \sigma^2 \end{aligned} \tag{3.4}$$

by (3.3).

[1] Johann Carl Friedrich Gauß, 1777–1855

[2] Gauß attributes this result to Pierre-Simon Laplace, 1782

[3] This is actually the differential equation which lead Gauß to (3.1).

$\alpha = \Phi(-z_\alpha)$	$z_\alpha = \Phi^{-1}(1-\alpha)$
15.87 %	1.000
10 %	1.282
5 %	1.645
2.5 %	**1.960**
2.28 %	2.000
1.0 %	2.326
0.5 %	2.576
0.13 %	3.000
0.10 %	3.090
0.05 %	3.291
0.01 %	3.719

(a) upper-tailed scores

$\alpha = 2\Phi(-z_\alpha)$	$z_\alpha = \Phi^{-1}(1-\frac{\alpha}{2})$
31.73 %	1.000
10 %	1.645
5 %	**1.960**
4.55 %	2.000
2.5 %	2.241
1.0 %	2.576
0.5 %	2.807
0.27 %	3.000
0.10 %	3.291
0.05 %	3.481
0.01 %	3.891

(b) two-tailed scores

Table 3.1: Z-scores of the normal distribution

Lemma 3.3 (Stein's Lemma[4]). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}(X-\mu)g(X) = \sigma^2\,\mathbb{E}\,g'(X)$, provided that both integrals exist and g decays fast enough at $\pm\infty$.*

Proof. As $z \cdot \varphi(z) = -\varphi'(z)$ and by integration by parts we have that

$$\begin{aligned}\mathbb{E}(X-\mu)g(X) &= \int_{-\infty}^{\infty} g(x)\cdot\frac{x-\mu}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x = -\int_{-\infty}^{\infty} g(x)\cdot\varphi'\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x\\ &= \int_{-\infty}^{\infty} g'(x)\cdot\sigma\,\varphi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x = \sigma^2\int_{-\infty}^{\infty} g'(x)\,\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x = \sigma^2\,\mathbb{E}\,g'(X)\end{aligned}$$

and thus the assertion. □

Example 3.4. From $g(x) = (x-\mu)$ we deduce that $\operatorname{var}(X-\mu) = \sigma^2$; with $g(x) = (x-\mu)^3$ it holds that $\mathbb{E}(X-\mu)^4 = \sigma^2\,\mathbb{E}\,3(X-\mu)^2 = 3\sigma^4$; and $\mathbb{E}(X-\mu)^6 = \sigma^2\,\mathbb{E}\,5(X-\mu)^4 = 15\sigma^6$, etc.

Remark 3.5. For $X \sim \mathcal{N}(\mu, \sigma^2)$ it holds that

$$\mathbb{E}\left[X|a \le X \le b\right] = \mu + \sigma^2\frac{\varphi_{\mu,\sigma}(b) - \varphi_{\mu,\sigma}(a)}{\Phi_{\mu,\sigma}(b) - \Phi_{\mu,\sigma}(a)}.$$

Indeed, the density is $f(x) = \begin{cases}\frac{\varphi_{\mu,\sigma}(x)}{\Phi_{\mu,\sigma}(b)-\Phi_{\mu,\sigma}(a)} & x \in [a,b],\\ 0 & \text{else,}\end{cases}$ where $\varphi_{\mu,\sigma} = \Phi'_{\mu,\sigma}$. The expectation thus is

$$\mathbb{E}\left[X|a \le X \le b\right] = \mu + \int_a^b \frac{(x-\mu)\,\varphi_{\mu,\sigma}(x)}{\Phi_{\mu,\sigma}(b) - \Phi_{\mu,\sigma}(a)}\mathrm{d}x = \mu + \int_a^b \frac{(x-\mu)\,\varphi_{\mu,\sigma}(x)}{\Phi_{\mu,\sigma}(b) - \Phi_{\mu,\sigma}(a)}\mathrm{d}x,$$

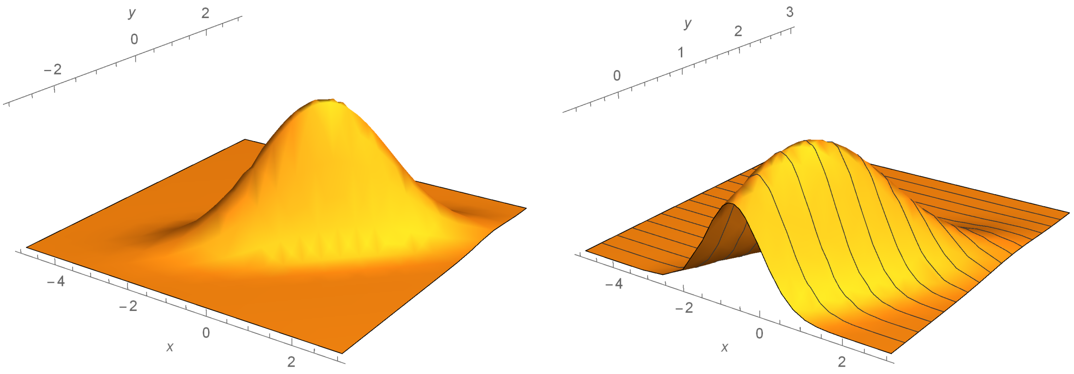
and now the identity $z \cdot \varphi(z) = -\varphi'(z)$ from above applies.

Remark 3.6. A useful series expansion is

$$\Phi(z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\left(z + \frac{z^3}{3} + \frac{z^5}{3\cdot 5} + \frac{z^7}{3\cdot 5\cdot 7} + \dots + \frac{z^{2n+1}}{(2n+1)!!} + \dots\right), \tag{3.5}$$

as follows from $\varphi(z) = \Phi'(z)$ and $\Phi(0) = 1/2$.

[4] Named after Charles Max Stein, 1920–2016, American mathematical statistician



(a) Probability density function

(b) Conditional Gaussians are Gaussian

Figure 3.1: Multivariate normal distribution

3.2 MULTIVARIATE NORMAL DISTRIBUTION

Definition 3.7. The random variable X follows a multivariate normal distribution, $X \sim \mathcal{N}(\mu, \Sigma)$, if

$$X = \mu + L\,Z,$$

where $Z = (Z_1, \dots, Z_n)$ is a vector of independent standard Gaussians ($Z_i \sim \mathcal{N}(0,1)$) and $\Sigma = LL^\top$.

Remark 3.8. The matrix L is a Cholesky factor or $L = \Sigma^{1/2}$ or any other matrix with $LL^\top = \Sigma$.

Remark 3.9 (Σ is symmetric). It follows from $LL^\top = \Sigma$ that $\Sigma = \Sigma^\top$.

Proposition 3.10. *The mean of $X \sim \mathcal{N}(\mu, \Sigma)$ is $\mathbb{E}\, X = \mu$ and the covariance matrix is* $\operatorname{cov} X = \Sigma$.

Proof. By linearity of the expectation,

$$\mathbb{E}\, X = \mathbb{E}\,(\mu + L\,Z) = \mu + L\;\mathbb{E}\, Z = \mu + 0 = \mu$$

and with Corollary 1.9

$$\operatorname{var} X = \operatorname{var}(\mu + L\,Z) = L \cdot \operatorname{var} Z \cdot L^\top = L \cdot \underbrace{\mathbb{E}\left(Z\,Z^\top\right)}_{\mathbb{1}_n \text{ by } (3.4)} \cdot L^\top = \Sigma,$$

the assertion. □

Example 3.11. Cf. Exercise 3.1.

Proposition 3.12 (Multivariate normal distributions are closed under linear transformations)**.** *If $X \sim \mathcal{N}(\mu, \Sigma)$, then*

$$b + A\,X \sim \mathcal{N}\left(b + A\,\mu,\; A\,\Sigma\,A^\top\right). \tag{3.6}$$

Proof. Indeed, $b + A\,X = b + A(\mu + L\,Z) = b + A\mu + ALZ$ and $AL(AL)^\top = A\Sigma A^\top$, the assertion. □

Corollary 3.13. *The marginal distribution of $X \sim \mathcal{N}(\mu, \Sigma)$ are $X_i \sim \mathcal{N}(\mu_i, \sigma_{ii})$.*

Proof. It holds that $X_i = e_i^\top X$, thus $e_i^\top \mu = \mu_i$ and $e_i^\top \Sigma e_i = \sigma_{ii}$ and the result. □

Proposition 3.14. *The density of the* multivariate *normal distribution $\mathcal{N}(\mu,\Sigma)$ with mean $\mu\in\mathbb{R}^n$ and invertible, strictly positive definite covariance matrix $\Sigma\in\mathbb{R}^{n\times n}$ is*

$$f(x)=\frac{1}{\sqrt{(2\pi)^n\,\det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu)^\top\,\Sigma^{-1}\,(x-\mu)\right). \tag{3.7}$$

Proof. The density of the random vector $(Z_1,\dots,Z_n)$ of independent normals ($Z_i\sim\mathcal{N}(0,1)$) is

$$f_{(Z_1,\dots Z_n)}(z_1,\dots,z_n)=\varphi(z_1)\cdot\ldots\cdot\varphi(z_n)=\frac{1}{\sqrt{2\pi}^n}\exp\Big(-\frac{1}{2}z^\top z\Big).$$

The affine linear function $g(z):=\mu+L\,z$ has inverse $g^{-1}(x)=L^{-1}\,(x-\mu)$. By Proposition 1.15 the random vector $X:=g(Z)=\mu+L\,Z$ has density function

$$f_X(x)=f_Z\left(L^{-1}\,(x-\mu)\right)\cdot\left|\det L^{-1}\right|=\frac{1}{\sqrt{(2\pi)^n\det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu)^\top\,\underbrace{L^{-\top}L^{-1}}_{\Sigma^{-1}}\,(x-\mu)\right)$$

and thus the assertion. □

Corollary 3.15. *Suppose that $X\sim\mathcal{N}(\mu,\Sigma)$, where Σ has eigenvectors $\Sigma\,u_i=\sigma_i^2\,u_i$ with $u_i^\top u_i=1$. Then the random variables $u_i^\top X\sim\mathcal{N}\big(u_i^\top\mu,\sigma_i^2\big)$, $i=1,\dots,n$, are* independent *and the density of X is $f_X(x)=\prod_{i=1}^n\frac{1}{\sigma_i}\varphi\left(\frac{u_i^\top x-u_i^\top\mu}{\sigma_i}\right)$.*

Proof. The matrix $U:=(u_1\mid\cdots\mid u_n)$ is unitary and

$$\Sigma=U\begin{pmatrix}\sigma_1^2&0&0\\0&\ddots&0\\0&0&\sigma_n^2\end{pmatrix}U^\top.$$

With (3.7) and (1.9) (or Exercise 1.3) we deduce the density

$$\begin{aligned}f_{U^\top X}(z)&=\frac{1}{\sqrt{(2\pi)^n\det\Sigma}}\exp\left(-\frac{1}{2}(z-U^\top\mu)^\top\begin{pmatrix}\sigma_1^{-2}&0&0\\0&\ddots&0\\0&0&\sigma_n^{-2}\end{pmatrix}(z-U^\top\mu)\right)\\&=\prod_{i=1}^n\frac{1}{\sqrt{2\pi\,\sigma_i^2}}\exp\left(-\frac{1}{2}\left(\frac{z_i-u_i^\top\mu}{\sigma_i}\right)^2\right)=\prod_{i=1}^n\frac{1}{\sigma_i}\varphi\left(\frac{z_i-u_i^\top\mu}{\sigma_i}\right).\end{aligned}$$

This is a product and hence the components $u_i^\top X$ are independent.

The remaining density of X is apparent with $f_X(x)=f_{UU^\top X}(x)$ and (1.9) again. □

Theorem 3.16 (Uncorrelated normals are independent)**.** *Suppose that $(X,Y)\sim\mathcal{N}(\mu,\Sigma)$ follow a multivariate normal distribution. If the components X and Y are uncorrelated, then X and Y are independent.*

Proof. As X and Y are not correlated, it follows that $\Sigma=\begin{pmatrix}\Sigma_{XX}&0\\0&\Sigma_{YY}\end{pmatrix}$. Hence $\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}+\begin{pmatrix}L_{XX}&0\\0&L_{YY}\end{pmatrix}\begin{pmatrix}Z_X\\Z_Y\end{pmatrix}=\begin{pmatrix}\mu_X+L_{XX}\cdot Z_X\\\mu_Y+L_{YY}\cdot Z_Y\end{pmatrix}$. It follows that X and Y are independent, as Z_X and Z_Y are. □

Proposition 3.17. *The moment generating function of the multivariate normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$ is*

$$m_X(t) := \mathbb{E}\, e^{t^\top X} = e^{\mu^\top t + \frac{1}{2} t^\top \Sigma t}, \qquad t \in \mathbb{C}^n. \tag{3.8}$$

Proof. The moment generation function of the *univariate* normal distribution is

$$\begin{aligned} m_X(t) := \mathbb{E}\, e^{t\,X} &= \int_{-\infty}^{\infty} e^{t\,x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x \\ &= e^{\mu\,t + \frac{1}{2}t^2\sigma^2} \cdot \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x-\mu-t\,\sigma^2\right)^2}}_{\text{pdf of } \mathcal{N}(\mu+t\,\sigma^2,\,\sigma^2)} \mathrm{d}x = e^{\mu\,t+\frac{1}{2}t^2\sigma^2}, \end{aligned} \tag{3.9}$$

where we have used the algebraic identity $t\,x - \frac{1}{2\sigma^2}(x-\mu)^2 = \mu\,t + \frac{1}{2}t^2\sigma^2 - \frac{1}{2\sigma^2}(x-\mu-t\,\sigma^2)^2$ and $t \in \mathbb{R}$.

To verify the assertion for $t \in \mathbb{C}$, consider the function $f(s) := \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-t+is)^2}\mathrm{d}s$ in (3.9) and note that

$$f'(s) = -\int_{-\infty}^{\infty} i\frac{x-t+is}{\sigma^2} e^{-\frac{1}{2\sigma^2}(x-t+is)^2}\mathrm{d}s = i \cdot e^{-\frac{1}{2\sigma^2}(x-t+is)^2}\Big|_{x=-\infty}^{\infty} = i \cdot e^{-\frac{x^2}{2\sigma^2}} e^{-i\frac{(x-t)s}{\sigma^2}+\frac{s^2}{2\sigma^2}}\Big|_{x=-\infty}^{\infty} = 0,$$

that is, $f(s) = f(0)$ for all $s \in \mathbb{R}$. Hence the assertion for $t \in \mathbb{C}$.

For the multivariate case define $Y := L^{-1}(X-\mu)$ and note that $Y \sim \mathcal{N}(0, \mathbb{1}_n)$ by Proposition 3.12; in particular the variables Y_i are independent by Theorem 3.16. Then

$$m_X(t) = \mathbb{E}\, e^{t^\top X} = \mathbb{E}\, e^{t^\top(\mu+LY)} = e^{t^\top\mu}\, \mathbb{E} \prod_{i=1}^n e^{(t^\top L)_i \cdot Y_i} = e^{\mu^\top t} \prod_{i=1}^n \mathbb{E}\, e^{(t^\top L)_i \cdot Y_i},$$

as Y_i are independent. Employing the moment generating function for the univariate normal distribution (3.9) we have further

$$\begin{aligned} m_X(t) = e^{\mu^\top t} \prod_{i=1}^n m_{Y_i}\left(\left(t^\top L\right)_i\right) &= e^{\mu^\top t} \prod_{i=1}^n e^{\frac{1}{2}\left(t^\top L\right)_i^2} = e^{\mu^\top t} e^{\sum_{i=1}^n \frac{1}{2}\left(t^\top L\right)_i^2} \\ &= e^{\mu^\top t} e^{\frac{1}{2}t^\top L\left(t^\top L\right)^\top} = e^{\mu^\top t} e^{\frac{1}{2}t^\top L L^\top t} = e^{\mu^\top t + \frac{1}{2}t^\top \Sigma t}, \end{aligned}$$

which concludes the proof. □

Corollary 3.18. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then the centralized k-th ($k = 0, 1, 2, \dots$) moment is*

$$\mathbb{E}(X-\mu)^{2k} = \frac{(2k)!}{k!\,2^k}\sigma^{2k} \text{ and } \mathbb{E}(X-\mu)^{2k+1} = 0;$$

in particular, $\mathbb{E}(X-\mu)^4 = 3\sigma^4$ and $\mathbb{E}(X-\mu)^6 = 15\sigma^6$.

Proof. The coefficients of t^{2k} in the equation $\mathbb{E}\, e^{t(X-\mu)} = e^{\frac{1}{2}\sigma^2 t^2}$ are $\frac{1}{(2k)!}\mathbb{E}(X-\mu)^{2k} = \frac{1}{k!}\left(\frac{\sigma^2}{2}\right)^k$ and hence the assertion. □

Proposition 3.19 (Multivariate normal distributions are closed under convolution, cf. Exercise 3.3)**.** *If $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ are independent (multivariate) normals, then*

$$\alpha X_1 + \beta X_2 \sim \mathcal{N}\left(\alpha\mu_1 + \beta\mu_2,\ \alpha^2\Sigma_1 + \beta^2\Sigma_2\right). \tag{3.10}$$

In other words: normal distributions are closed under addition, "+".

Proof. The moment generating function of $\alpha X_1 + \beta X_2$ is (cf. (3.8))

$$\begin{aligned} m_{\alpha X_1+\beta X_2}(t) &= \mathbb{E}\, e^{t^\top(\alpha X_1+\beta X_2)} = \mathbb{E}\, e^{\alpha t^\top X_1} e^{\beta t^\top X_2} = \mathbb{E}\, e^{\alpha t^\top X_1}\cdot \mathbb{E}\, e^{\beta t^\top X_2} \\ &= e^{\alpha\mu_1^\top t+\frac{1}{2}\alpha^2 t^\top\Sigma_1 t}\cdot e^{\beta\mu_2^\top t+\frac{1}{2}\beta^2 t^\top\Sigma_2 t} = e^{(\alpha\mu_1+\beta\mu_2)^\top t+\frac{1}{2}t^\top(\alpha^2\Sigma_1+\beta^2\Sigma_2)t}. \end{aligned}$$

The latter is the mgf of a $\mathcal{N}\left(\alpha\mu_1+\beta\mu_2, \alpha^2\Sigma_1+\beta^2\Sigma_2\right)$ random variable. □

3.3 CONDITIONAL GAUSSIANS

Conditionals of Gaussians are Gaussian. Cf. Figure 3.1b for an illustration.

Theorem 3.20 (Cf. Liptser and Shiryaev [11, Theorem 13.1])**.** *Suppose that*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right),$$

then the conditional distribution is Gaussian as well; more specifically,

$$Y|\,X \sim \mathcal{N}\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}\,(X-\mu_X)\,,\ \Sigma_{YY} - \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY}\right), \tag{3.11}$$

or

$$X|\,Y \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}\,(Y-\mu_Y)\,,\ \Sigma_{XX} - \Sigma_{XY}\,\Sigma_{YY}^{-1}\,\Sigma_{YX}\right).$$

Proof. Define the Schur complement $S := \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ and observe that

$$\begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbb{1} & 0 \\ -\Sigma_{YX}\Sigma_{XX}^{-1} & \mathbb{1} \end{pmatrix}^\top \cdot \begin{pmatrix} \Sigma_{XX}^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \cdot \begin{pmatrix} \mathbb{1} & 0 \\ -\Sigma_{YX}\Sigma_{XX}^{-1} & \mathbb{1} \end{pmatrix}. \tag{3.12}$$

Set $\tilde{\mu} := \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}\,(x-\mu_X)$. Then we have that

$$\begin{aligned} &\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^\top \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\ &\quad = \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^\top \begin{pmatrix} \mathbb{1} & 0 \\ -\Sigma_{YX}\Sigma_{XX}^{-1} & \mathbb{1} \end{pmatrix}^\top \cdot \begin{pmatrix} \Sigma_{XX}^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \cdot \begin{pmatrix} \mathbb{1} & 0 \\ -\Sigma_{YX}\Sigma_{XX}^{-1} & \mathbb{1} \end{pmatrix} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \\ &\quad = \begin{pmatrix} x-\mu_X \\ y-\tilde{\mu} \end{pmatrix}^\top \cdot \begin{pmatrix} \Sigma_{XX}^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \cdot \begin{pmatrix} x-\mu_X \\ y-\tilde{\mu} \end{pmatrix} \\ &\quad = (x-\mu_X)^\top\Sigma_{XX}^{-1}(x-\mu_X) + (y-\tilde{\mu})^\top S^{-1}(y-\tilde{\mu}). \end{aligned} \tag{3.13}$$

Now note that $X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$ so that the conditional density, with (1.11), is

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{\exp\left(-\frac{1}{2}(3.13)\right)}{\sqrt{(2\pi)^{n_X+n_Y}\det\left(S\,\Sigma_{XX}\right)}} \Bigg/ \frac{\exp\left(-\frac{1}{2}(x-\mu_X)^\top\Sigma_{XX}^{-1}(x-\mu_X)\right)}{\sqrt{(2\pi)^{n_X}\det\Sigma_{XX}}} \\ &= \frac{\exp\left(-\frac{1}{2}(y-\tilde{\mu})^\top S^{-1}(y-\tilde{\mu})\right)}{\sqrt{(2\pi)^{n_Y}\det S}}. \end{aligned}$$

Hence the Gaussian distribution (3.11). □

Corollary 3.21 (Cf. Bishop [1]). *Suppose that*

$$X \sim \mathcal{N}(\mu,\, \Sigma_X) \text{ and} \tag{3.14}$$

$$Y|\, X \sim \mathcal{N}(AX + b,\, \Sigma_Y), \tag{3.15}$$

then

$$Y \sim \mathcal{N}\left(A\mu + b,\, \Sigma_Y + A\Sigma_X A^\top\right) \text{ and} \tag{3.16}$$

$$X|\, Y \sim \mathcal{N}\left(\Sigma\big(A^\top \Sigma_Y^{-1}(Y - b) + \Sigma_X^{-1}\mu\big),\, \Sigma\right), \tag{3.17}$$

where $\Sigma \coloneqq \left(\Sigma_X^{-1} + A^\top \Sigma_Y^{-1} A\right)^{-1}$.

Proof. We derive the distributions from the general Gaussian random variable

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_Y + A\Sigma_X A^\top \end{pmatrix}\right).$$

The marginals (3.14) and (3.16) are apparent. From (3.11) we infer that

$$\begin{aligned} Y|\, X &\sim \mathcal{N}\left(A\mu + b + A\Sigma_X\Sigma_X^{-1}(X - \mu),\, \Sigma_Y + A\Sigma_X A^\top - A\Sigma_X\Sigma_X^{-1}\Sigma_X A^\top\right) \\ &= \mathcal{N}\left(AX + b,\, \Sigma_Y\right) \end{aligned}$$

and thus (3.15). Again from (3.11) we derive that

$$\begin{aligned} X|\, Y &\sim \mathcal{N}\left(\mu + \Sigma_X A^\top \left(\Sigma_Y + A\Sigma_X A^\top\right)^{-1}(Y - A\mu - b),\, \Sigma_X - \Sigma_X A^\top \left(\Sigma_Y + A\Sigma_X A^\top\right)^{-1} A\Sigma_X\right) \\ &= \mathcal{N}\left(\Sigma\big(A^\top\Sigma_Y^{-1}(Y - b) + \Sigma_X^{-1}\mu\big),\, \Sigma\right), \end{aligned} \tag{3.18}$$

where we have employed $(\Sigma_Y + A\Sigma_X A^\top)^{-1} = \Sigma_Y^{-1} - \Sigma_Y^{-1} A\,\Sigma\, A^\top \Sigma_Y^{-1}$ (Woodbury matrix identity, cf. Exercise 3.13) and thus (3.17). □

Theorem 3.22. *For* $X \sim \mathcal{N}(\mu, \Sigma)$ *and a surjective matrix* A *it holds that*

$$\mathbb{E}\left[X|\, AX = y\right] = \mu + \Sigma A^\top \left(A\Sigma A^\top\right)^{-1}(y - A\mu) \tag{3.19}$$

and

$$\operatorname{var}(X|\, AX = y) = (1 - P)\,\Sigma, \tag{3.20}$$

where

$$P \coloneqq \Sigma A^\top \left(A\Sigma A^\top\right)^{-1} A \tag{3.21}$$

is a projection.[5]

Proof. It holds that $P^2 = P$ and $\Sigma\, P^\top = P\,\Sigma$. With

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} \coloneqq \begin{pmatrix} P(X - \mu) \\ (1 - P)\,(X - \mu) \end{pmatrix}$$

and Proposition 3.10 it holds that

$$\operatorname{cov}(Y', X') = \mathbb{E}\, Y'\, X'^\top = \mathbb{E}(1 - P)(X - \mu)(X - \mu)^\top P^\top = (1 - P)\,\Sigma\, P^\top = (1 - P)P\,\Sigma = 0$$

[5] It is sufficient to set $P \coloneqq \Sigma A^\top \left(A\Sigma A^\top\right)^{+} A$, where $\left(A\Sigma A^\top\right)^{+}$ is the Moore–Penrose inverese; cf. (13.6) below.

so that X' and Y' are independent by Theorem 3.16; similarly we have that

$$\operatorname{var} Y' = \mathbb{E}(1-P)(X-\mu)(X-\mu)^\top(1-P) = (1-P)\,\Sigma\,(1-P)^\top = (1-P)\,\Sigma.$$

Note, that $AX = y$ is equivalent to $PX = \Sigma A^\top \left(A\Sigma A^\top\right)^{-1} y =: y'$. Thus

$$\begin{aligned}\mathbb{E}\left[X-\mu \,|\, AX = y\right] &= \mathbb{E}\left[X' + Y' \,|\, PX = y'\right]\\ &= \mathbb{E}\left[X' + Y' \,|\, X' = y' - P\mu\right]\\ &= y' - P\mu\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[(X-\mu)(X-\mu)^\top \middle| AX = y\right] &= \mathbb{E}\left[(X'+Y')\,(X'+Y')^\top \middle| PX = y'\right]\\ &= \mathbb{E}\left[X'X'^\top + X'Y'^\top + Y'X'^\top + Y'Y'^\top \middle| X' = y' - P\mu\right]\\ &= (y'-P\mu)(y'-P\mu)^\top + 0 + 0 + \mathbb{E}\,Y'Y'^\top\\ &= (y'-P\mu)(y'-P\mu)^\top + (1-P)\Sigma\end{aligned}$$

so that $\operatorname{var}(X|\,AX = y) = (1-P)\Sigma$. The assertion now follows. □

Note that the random variable X' has deficient rank, as $\operatorname{var} X' = (1-P)\,\Sigma$.

Remark 3.23 (Cf. Exercise 3.11). The subspaces $B := \{(1-P)x \colon x \in \mathbb{R}^n\}$ and $B^\perp := \{Px \colon x \in \mathbb{R}^n\}$ (with P defined in (3.21)) are orthogonal with respect to the inner product

$$\langle x, y\rangle_\Sigma := x^\top\,\Sigma^{-1}\,y.$$

Remark 3.24 (Caveat). By formal computation (in line with the proofs in this section) and (3.6) we have that

$$\begin{aligned}P\left(X \in \mathrm{d}x \mid AX = y\right) &= P\left(X \in \mathrm{d}x \mid PX = y'\right)\\ &= P\left(\mu + X' + Y' \in \mathrm{d}x \mid Y' = y' - P\mu\right)\\ &= P\left(X' + \mu + y' - P\mu \in \mathrm{d}x \mid Y' = y' - P\mu\right)\\ &= P\left(X' + y' + (1-P)\mu \in \mathrm{d}x\right)\\ &= P\left(y' + (1-P)X \in \mathrm{d}x\right)\end{aligned}$$

so that

$$\begin{aligned}X|\,(Y = y) &\sim \mathcal{N}\left(y' + (1-P)\mu,\ (1-P)\Sigma\right)\\ &= \mathcal{N}\left(\Sigma A^\top\left(A\Sigma A^\top\right)^{-1} y + \left(1 - \Sigma A^\top\left(A\Sigma A^\top\right)^{-1} A\right)\mu,\ (1-P)\Sigma\right)\\ &= \mathcal{N}\left(\mu + \Sigma A^\top\left(A\Sigma A^\top\right)^{-1}(y - A\mu),\ (1-P)\Sigma\right).\end{aligned}$$

However, the matrix $(1-P)\Sigma$ is singular (in general) so that the distribution (cf. (3.7)) does not have a density.

3.4 PROBLEMS

Exercise 3.1. *Give the pdf of a bivariate normal distribution with correlation $\rho \in (-1,1)$ explicitly. Hint:* $\Sigma = \begin{pmatrix}\sigma_1^2 & \rho\,\sigma_1\sigma_2\\ \rho\,\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$.

Exercise 3.2. *Give two random variables which are uncorrelated, but not independent. (Cf. Theorem 3.16)*

Exercise 3.3 (Cf. Proposition 3.19)**.** *Let* $X_1 \sim N(\mu_1, \sigma_1^2)$ *and* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ *be independent. Show that*

$$\alpha X_1 + \beta X_2 \sim \mathcal{N}(\alpha\mu_1 + \beta\mu_2,\, \alpha^2\sigma_1^2 + \beta^2\sigma_2^2).$$

Exercise 3.4. *Assume that* $X_i \sim \mathcal{N}(\mu, \sigma^2)$ *are uncorrelated normals. Show that the Z-transform* $Z_i := \frac{X_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$ *and*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

Exercise 3.5 (Cholesky decomposition)**.** *Let* X, Y *be* independent *standard normals. Define*

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} := \begin{pmatrix} X \\ \rho \cdot X + \sqrt{1-\rho^2} \cdot Y \end{pmatrix}, \text{ i.e., } \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

and show that $X' \sim X$ *and* $Y' \sim Y$*, but* $\operatorname{corr}(X', Y') = \rho$*.*

Exercise 3.6. *For* X, Y independent *standard normals define*

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} := \begin{pmatrix} c & s \\ s & c \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} c\,X + s\,Y \\ s\,X + c\,Y \end{pmatrix},$$

where $\delta := \sqrt{1-\rho^2}$, $c := \sqrt{\frac{1+\delta}{2}}$ *and* $s := \operatorname{sign}(\rho) \cdot \sqrt{\frac{1-\delta}{2}}$*. Show that* $X' \sim X$ *and* $Y' \sim Y$*, but* $\operatorname{corr}(X', Y') = \rho$*.*

Exercise 3.7. *Give the density for the multivariate variable* (X', Y') *in the previous example.*

Exercise 3.8. *Suppose that* $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & C \\ C^\top & \Sigma_Y \end{pmatrix}\right)$*. Show that*

$$\alpha X + \beta Y \sim \mathcal{N}\left(\alpha\,\mu_1 + \beta\,\mu_2,\, \alpha^2\,\Sigma_X + \alpha\,\beta\,\left(C + C^\top\right) + \beta^2 \Sigma_Y\right)$$

(cf. (3.10)).

Exercise 3.9. *For* X, $Y \sim \mathcal{N}(\mu, \Sigma)$ independent *normals define* $X' := \frac{X+Y}{\sqrt{2}}$ *and* $Y' := \frac{X-Y}{\sqrt{2}}$*. Give the distribution of* X' *and* Y' *and show that they are also independent.*

Exercise 3.10. *Suppose that* X *and* $Z \sim \mathcal{N}(0,1)$ *are independent. Define* $Y := X \cdot \operatorname{sign}(Z)$ *and show that*

(i) $Y \sim \mathcal{N}(0,1)$,

(ii) X *and* Y *are not correlated,*

(iii) $X + Y$ *is* not *normal. Is this a contradiction to Proposition 3.19?*

Exercise 3.11. *Verify Remark 3.23 above.*

Exercise 3.12. *Verify (3.12).*

Exercise 3.13. *Verify (3.18).*

Exercise 3.14 (Cauchy distribution)**.** *For* Y, $Z \sim \mathcal{N}(0,1)$*, show that* ${}^{Y}\!/_{Z}$ *has density*

$$f_{Y/Z}(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \qquad x \in \mathbb{R}.$$

Exercise 3.15. *Use integration by parts to verify (3.5).*

Limit theorems 4

> Million-to-one odds happen eight times a day in New York.[a]
>
> [a] population about 8 millions
>
> Penn Jillette, 1955

4.1 LAW OF LARGE NUMBER

Theorem 4.1 (Weak law of large numbers)**.** *Let X_i be iid. with finite second moment. Then*

$$\overline{X}_n \longrightarrow \mu \text{ in probability},$$

i.e., $P\left(\left|\overline{X}_n-\mu\right|>\varepsilon\right) \xrightarrow[n\to\infty]{} 0$ *for every* $\varepsilon>0$.

Proof. By Markov's inequality we have that

$$P\left(\left|\overline{X}_n-\mu\right|>\varepsilon\right)=P\left((\overline{X}_n-\mu)^2>\varepsilon^2\right)\le\frac{1}{\varepsilon^2}\,\mathbb{E}(\overline{X}_n-\mu)^2=\frac{\sigma^2}{\varepsilon^2\,n}\xrightarrow[n\to\infty]{}0$$

and hence the assertion. □

4.2 CENTRAL LIMIT THEOREM

The following elementary analysis and proof follow Kersting and Wakolbinger [7]. Here is an illustrative gif: https://en.wikipedia.org/wiki/Convergence_of_random_variables.

Lemma 4.2. *Let X_i be independent with $\mathbb{E}\,X_i=:\mu$, $\operatorname{var}X_i=:\sigma^2<\infty$, $h\colon\mathbb{R}\to\mathbb{R}$ with $\|h''\|_\infty<\infty$ and h'' Lipschitz continuous, then*

$$\mathbb{E}\,h\left(\frac{X_1+\dots+X_n-n\,\mu}{\sqrt{n}}\right)\to\mathbb{E}\,h(Z)$$

as $n\to\infty$, where $Z\sim\mathcal{N}(0,\sigma^2)$.

Proof. We may assume $\mu=0$. Let Z_i be independent copies of Z, which are independent of all X_i as well. Define $U_i:=\frac{Z_1+\dots+Z_{i-1}+X_{i+1}+\dots+X_n}{\sqrt{n}}$ and note the telescoping series $h\left(\frac{X_1+\dots+X_n}{\sqrt{n}}\right)-h\left(\frac{Z_1+\dots+Z_n}{\sqrt{n}}\right)=\sum_{i=1}^n h\left(U_i+\frac{X_i}{\sqrt{n}}\right)-h\left(U_i+\frac{Z_i}{\sqrt{n}}\right)$. The Taylor series expansion with Peano remainder[1] at U_i is

$$h\left(U_i+\frac{X_i}{\sqrt{n}}\right)-h\left(U_i+\frac{Z_i}{\sqrt{n}}\right)=h'(U_i)\frac{X_i-Z_i}{\sqrt{n}}+h''(U_i)\frac{X_i^2-Z_i^2}{2n}+R_{i,n},$$

[1] Recall that $h(x+\Delta x)=h(x)+h'(x)\,\Delta x+h''(x)\frac{\Delta x^2}{2}+(h''(\xi)-h''(x))\,\frac{\Delta x^2}{2}$ for $\xi\in(x,x+\Delta x)$. Choose $x=U_i$ and $\Delta x=\frac{X_i}{\sqrt{n}}$ ($\Delta x=\frac{Z_i}{\sqrt{n}}$, resp.).

where

$$R_{i,n} := \frac{h''(V_i) - h''(U_i)}{2} \cdot \frac{X_i^2}{n} - \frac{h''(W_i) - h''(U_i)}{2} \cdot \frac{Z_i^2}{n}$$

with $|V_i - U_i| \le \frac{|X_i|}{\sqrt{n}}$ and $|W_i - U_i| \le \frac{|Z_i|}{\sqrt{n}}$.

Define $c_2 := \|h''\|_\infty = \sup_{x\in\mathbb{R}} |h''(x)|$, and assume that there is a finite Lipschitz constant c_3 of h''. With $|\ h''(V_i) - h''(U_i)\ | \le c_3\,|V_i - U_i| \le c_3\frac{|X_i|}{\sqrt{n}}$, the Peano remainder thus satisfies

$$\left|R_{i,n}\right| \le \frac{c_3\,k^3}{2n^{3/2}}\,\mathbb{1}_{\{X_i \le k\}} + \frac{2c_2}{2}\frac{X_i^2}{n}\,\mathbb{1}_{\{X_i > k\}} + \frac{c_3}{2}\frac{|Z_i|^3}{n^{3/2}},$$

where $k > 0$ is arbitrary.

By independence, as $\mathbb{E}\,X_i = \mathbb{E}\,Z_i = \mu$ and $\operatorname{var} X_i = \operatorname{var} Z_i = \sigma^2$,

$$\mathbb{E}\,h'(U_i)\frac{X_i - Z_i}{\sqrt{n}} = \mathbb{E}\,h'(U_i) \cdot \mathbb{E}\,\frac{X_i - Z_i}{\sqrt{n}} = 0 \text{ and}$$

$$\mathbb{E}\,h''(U_i)\frac{X_i^2 - Z_i^2}{2n} = \mathbb{E}\,h''(U_i) \cdot \mathbb{E}\,\frac{X_i^2 - Z_i^2}{2n} = 0.$$

Recall from (3.10) that $\frac{Z_1+\dots+Z_n}{\sqrt{n}} \sim Z$, thus

$$\left|\mathbb{E}\,h\left(\frac{X_1 + \dots + X_n}{\sqrt{n}}\right) - \mathbb{E}\,h(Z)\right| \le \frac{c_3}{2}n\frac{k^3 + \mathbb{E}\,|Z_i|^3}{n^{3/2}} + 2c_2 n\frac{\mathbb{E}\,X_i^2 \cdot \mathbb{1}_{\{X_i > k\}}}{n}$$
$$\xrightarrow[n\to\infty]{} \mathbb{E}\left(X_i^2 \cdot \mathbb{1}_{\{X_i > k\}}\right).$$

Finally note that $\mathbb{E}\left(X_i^2 \cdot \mathbb{1}_{\{X_i > k\}}\right) \xrightarrow[k\to\infty]{} 0$, as $\mathbb{E}\,X_i^2 < \infty$ by assumption. □

Theorem 4.3 (Central limit theorem, CLT)**.** *Let X_i be independent with $\mathbb{E}\,X_i =: \mu$ and $\operatorname{var} X_i =: \sigma^2 < \infty$. Then*

$$P\left(\frac{X_1 + \dots + X_n - n\cdot\mu}{\sqrt{n}\,\sigma} \le x\right) \xrightarrow[n\to\infty]{} \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t;$$

it is said that $\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ in distribution (occasionally also denoted by $\rightsquigarrow \mathcal{N}(0,1)$).

Proof. Let h_1 and h_2 be functions with $\mathbb{1}_{(-\infty,x-\varepsilon]}(\cdot) \le h_1(\cdot) \le \mathbb{1}_{(-\infty,x]}(\cdot) \le h_2(\cdot) \le \mathbb{1}_{(-\infty,x+\varepsilon]}(\cdot)$ and $\left\|h_i'''\right\|_\infty < \infty$. With the preceding lemma, $Z \sim \mathcal{N}(0,1)$ and monotonicity of the expectation it follows that

$$\begin{aligned}
P(Z \le x - \varepsilon) &\le \mathbb{E}\,h_1(Z)\\
&= \lim_{n\to\infty} \mathbb{E}\,h_1\left(\frac{X_1 + \dots + X_n - n\,\mu}{\sqrt{n}\,\sigma}\right)\\
&\le \liminf_{n\to\infty} P\left(\frac{X_1 + \dots + X_n - n\,\mu}{\sqrt{n}\,\sigma} \le x\right)\\
&\le \limsup_{n\to\infty} P\left(\frac{X_1 + \dots + X_n - n\,\mu}{\sqrt{n}\,\sigma} \le x\right)\\
&\le \lim_{n\to\infty} \mathbb{E}\,h_2\left(\frac{X_1 + \dots + X_n - n\,\mu}{\sqrt{n}\,\sigma}\right)\\
&= \mathbb{E}\,h_2(Z)\\
&\le P(Z \le x + \varepsilon).
\end{aligned}$$

The assertion follows, as Z does not give mass to atoms. □

4.3 BOREL–CANTELLI LEMMAS

Lemma 4.4 (First Borel–Cantelli Lemma). *Let E_n be a sequence of events such that $\sum_{n=1} P(E_n) < \infty$. Then*

$$P\left(\limsup_{n\to\infty} E_n\right) = P\left(E_n,\ \textit{infinitely often}\right) = 0,$$

where

$$\limsup E_n := \bigcap_{m\ge 1} \bigcup_{n\ge m} E_n = \{\omega\colon \omega \in E_n \textit{ for infinitely many } n\}\,.$$

Proof. Define $G_m := \bigcup_{n\ge m} E_n$. Then $G_m \supset G := \limsup E_n$ and consequently

$$P(G) \le P(G_m) \le \sum_{n\ge m} P(E_n) \xrightarrow[m\to\infty]{} 0.$$

This is the assertion. □

Lemma 4.5 (Second Borel–Cantelli Lemma). *If E_n is a sequence of* independent *events and $\sum_{n=1} P(E_n) = \infty$, then*

$$P\left(\limsup_{n\to\infty} E_n\right) = 1.$$

Proof. Note first that $G_m \supset G_{m+1}$, and thus $\bigcap_{n\ge m} E_n^c = G_m^c$ is *in*creasing, as m increases. It follows for $m' > m$ that

$$P\left(\bigcap_{n\ge m} E_n^c\right) \le P\left(\bigcap_{n\ge m'} E_n^c\right) = \prod_{n\ge m'} \left(1 - P(E_n)\right) \le \exp\left(-\sum_{n\ge m'} P(E_n)\right) \to 0.$$

by independence and as $1 - x \le e^{-x}$ whenever $x \ge 0$. Hence

$$P\big((\limsup E_n)^c\big) = P\left(\bigcup_m \bigcap_{n\ge m} E_n^c\right) \le \sum_{m=0} P\left(\bigcap_{n\ge m} E_n^c\right) = 0,$$

from which the assertion follows. □

4.4 LAW OF THE ITERATED LOGARITHM

Let X_i be iid. random variables with $P(X_i = \pm 1) = \frac{1}{2}$ and set $S_n := X_1 + X_2 + \dots + X_n$.

Lemma 4.6. *For every $a > 0$ and $u \ge 0$ it holds that*

$$P\left(\sup_{k\le n} S_k \ge a\right) \le \frac{\mathbb{E}\, e^{u\, S_n}}{e^{u\, a}}. \tag{4.1}$$

Proof. Set

$$\begin{aligned} E_0 &:= \{S_1 < a, \dots, S_n < a\}\,,\\ E_1 &:= \{S_1 \ge a\} \text{ and}\\ E_k &:= \{S_1 < a, \dots, S_{k-1} < a, S_k \ge a\} \text{ for } k = 2, \dots, n, \end{aligned}$$

so that $\Omega = \dot{\bigcup}_{k=0}^{n} E_k$. Note further that $\sum_{k=1}^{n} \mathbb{1}_{E_k} = \left\{\sup_{k\le n} S_k \ge a\right\}$ and hence

$$\mathbb{E}\, e^{u\,S_n} \ge \sum_{k=1}^{n} \int_{E_k} e^{u\,S_n} \mathrm{d}P = \sum_{k=1}^{n} \int_{\Omega} e^{u\,S_k}\, \mathbb{1}_{E_k} \cdot e^{u(X_{k+1}+\dots+X_n)} \mathrm{d}P.$$

Now recall that S_k and $X_{k+1}, \dots, X_n$ are independent, hence

$$\int_{\Omega} e^{u\,S_k}\, \mathbb{1}_{E_k} \cdot e^{u(X_{k+1}+\dots+X_n)} \mathrm{d}P \ge e^{u\,a}\, P(E_k) \cdot \left(\mathbb{E}\, e^{u\,X_1}\right)^{n-k} \ge e^{u\,a}\, P(E_k),$$

where we have employed Jensen's inequality $1 = e^{u\,\mathbb{E}X} \le \mathbb{E}\, e^{u\,X}$. It follows that

$$\mathbb{E}\, e^{u\,S_n} \ge \sum_{k=1}^{n} e^{u\,a} P(E_k) = \sum_{k=1}^{n} e^{u\,a} P\left(\cup_{k=1}^{n} E_k\right) = e^{u\,a} \cdot P\left(\sup_{k\le n} S_k \ge a\right)$$

and hence the assertion. □

Lemma 4.7. *For every $a > 0$ we have that*

$$P\left(\sup_{k\le n} S_k \ge a\right) \le e^{-a^2/(2n)} \textit{ and } P\left(\sup_{k\le n} |S_k| \ge a\right) \le 2e^{-a^2/(2n)}. \tag{4.2}$$

Proof. The moment generating function is (cf. Exercise 4.1)

$$\mathbb{E}\, e^{u\,S_n} = \frac{1}{2}\left(e^u + e^{-u}\right) \le e^{u^2/2}. \tag{4.3}$$

It follows with (4.1) that $P\left(\sup_{k\le n} S_k \ge a\right) \le e^{u^2/2 - u\,a}$. To obtain the assertion replace both, u and a, by $\frac{a}{\sqrt{n}}$. □

Theorem 4.8 (Law of the iterated logarithm; Khintchin, 1924)**.** *It holds that*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = 1 \textit{ almost surely and}$$
$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = -1 \textit{ almost surely.}$$

Proof. Assume that $c > 1$ and choose γ with $1 < \gamma < c$. Set $n_r := \lceil \gamma^r \rceil$ (the ceiling function) and consider the events

$$B_r := \left\{ \sup_{n_r < n \le n_{r+1}} |S_n| > \sqrt{2n_r \log\log n_r} \right\}.$$

Applying (4.2) with $a = c\sqrt{2n_r \log\log n_r}$ we have that

$$P(B_r) \le 2e^{-c^2 (n_r \log\log n_r)/n_{r+1}} = 2\left(\frac{1}{\log n_r}\right)^{\frac{c^2 n_r}{n_{r+1}}} \sim 2\left(\frac{1}{r\log\gamma}\right)^{\frac{c^2 n_r}{n_{r+1}}}.$$

But $\frac{n_r}{n_{r+1}} \sim \frac{1}{\gamma} > \frac{1}{c}$. It follows that $\sum_{r=1}^{\infty} P(B_r) \lesssim \sum_{r=1}^{\infty} 2\left(\frac{1}{r\log\gamma}\right)^c < \infty$. We conclude with Borel–Cantelli that B_r can happen with finitely many indices only, and it follows that $A_n := \left\{|S_n| > c\sqrt{2n_r \log\log n_r}\right\}$ happens for finitely many indices as well. Hence, $\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \le 1$, as $c > 1$ was arbitrary.

Next let $c < 1$. Choose $\eta \in (c, 1)$ so that $1 - \eta < \left(\frac{\eta - c}{2}\right)^2$, $\gamma \geq 2$ so that $\eta < \frac{\gamma - 1}{\gamma} < 1$ and set $n_r := \lceil \gamma^r \rceil$. The random variables $S_{n_{r-1}}$ and $D_r := S_{n_r} - S_{n_{r-1}}$ are independent and further, the variables D_r, $r = 1, 2, \dots,$ are independent. Set $B_r := \left\{ D_r > \eta \sqrt{2 n_r \log\log n_r} \right\}$ and $C_r := \left\{ S_{n_{r-1}} > -(\eta - c)\sqrt{2 n_r \log\log n_r} \right\}$ so that $B_r \cap C_r \subset A_{n_r}$. Note, that $4 n_{r-1} \approx 4 \frac{n_r}{\gamma} < 4 n_r (1 - \eta) < n_r (\eta - c)^2$ and thus

$$\begin{aligned} E_r &:= \left\{ |S_{n_{r-1}}| < 2\sqrt{2 n_{r-1} \log\log n_{r-1}} \right\} \\ &\subset \left\{ |S_{n_{r-1}}| < (\eta - c)\sqrt{2 n_{r-1} \log\log n_r} \right\} \\ &\subset \left\{ S_{n_{r-1}} > -(\eta - c)\sqrt{2 n_{r-1} \log\log n_r} \right\} = C_r. \end{aligned}$$

Recall from above that E_r finitely often, hence C_r only finitely often.

With $D_r^* := \frac{\mathbb{1}_{D_r}}{\sqrt{n_r - n_{r-1}}}$ it holds that $B_r = \left\{ D_r^* > \eta \sqrt{2 \frac{n_r}{n_r - n_{r-1}} \log\log n_r} \right\}$ and, as $\frac{n_r}{n_r - n_{r-1}} \approx \frac{\gamma}{\gamma - 1} < \frac{1}{\eta}$ that

$$B_r \supset \left\{ D_r^* > \sqrt{\eta}\sqrt{2 \log\log n_r} \right\} = \left\{ D_r^* > \sqrt{\eta}\sqrt{2 \log(r \log \gamma)} \right\}.$$

But the CLT, $D_r^* \to \mathcal{N}(0, 1)$. It follows that $\sum_{r=1}^\infty P(B_r) \geq \sum_{r=1}^\infty P\left(D_r^* > \sqrt{\eta}\sqrt{2\log(r \log\gamma)} \right) = \infty$. As the events B_r are independent by construction, it follows, again with Borel–Cantelli, that B_r infinitely often. The assertion follows, as $B_r \cap C_r \subset A_r$. □

4.5 PROBLEMS

Exercise 4.1. *Verify the inequality in (4.3).*

5 Important Distributions in Statistics

> Es existiert keine objektive Wahrscheinlichkeit.
>
> Bruno de Finetti, 1906–1985

5.1 χ^2-distribution

Definition 5.1 (χ^2-distribution)**.** The χ^2-distribution (chi-squared) with n degrees of freedom has density

$$f_{\chi_n^2}(x) = \frac{x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)} \qquad (x > 0). \tag{5.1}$$

Remark 5.2 (Gamma function). The function

$$\Gamma(s) := \int_0^\infty x^{s-1} e^{-x}\, \mathrm{d}x \tag{5.2}$$

is Euler's *Gamma function*, aka. *Euler integral of the second kind*. It holds that $\Gamma(s+1) = s \cdot \Gamma(s)$ (thus $\Gamma(n) = (n-1)!$, $n \in \{1, 2, \dots\}$) and, using Remark 3.2,

$$\Gamma\left({}^1\!/\!{}_2\right) = \int_0^\infty x^{-1/2}\, e^{-x}\, \mathrm{d}x \underset{x \leftarrow \frac{x^2}{2}}{=} \int_0^\infty \sqrt{\frac{2}{x^2}} e^{-x^2/2} \frac{2x}{2} \mathrm{d}x = \sqrt{2} \int_0^\infty e^{-x^2/2}\, \mathrm{d}x = \sqrt{2}\frac{\sqrt{2\pi}}{2} = \sqrt{\pi}. \tag{5.3}$$

Definition 5.3 (Gamma distribution)**.** The pdf of the Gamma distribution with paramters $\alpha > 0$ (shape) and $\beta > 0$ (rate) is

$$f_{\Gamma_{\alpha,\beta}}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad (x > 0). \tag{5.4}$$

Definition 5.4 (Erlang distribution)**.** The pdf of the Erlang[1] distribution with parameters $k \in \{1, 2 \dots\}$ (*shape*) and $\lambda > 0$ (*rate*, or *inverse scale*) is

$$f_{E_{k,\lambda}}(x) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} \qquad (x > 0).$$

Remark 5.5. The distribution $E_{1,\lambda} =: E_\lambda$ is the exponential distribution.

Remark 5.6 (Relation between Gamma, Erlang and χ^2). Erlang's distribution is a special case of the Gamma distribution (cf. (5.4)) and it holds that

$$\chi_n^2 \sim \Gamma_{n/2, 1/2}, \quad \chi_{2n}^2 \sim E_{n, 1/2}, \quad \sigma \cdot \Gamma_{\alpha,\beta} \sim \Gamma_{\alpha, \frac{\beta}{\sigma}} \quad \text{and} \quad E_{n,\beta} \sim \Gamma_{n,\beta}. \tag{5.5}$$

In particular we have that

$$2\lambda\, E_{k,\lambda} \sim \chi_{2k}^2 \quad \text{and} \quad \Gamma_{k,\beta} \sim \frac{1}{2\beta} \chi_{2k}^2 \tag{5.6}$$

(cf. Exercise 5.5).

[1] Agner Karrup Erlang, 1878–1929, Danish

Proposition 5.7. *The moment generating function of the Gamma distribution is*

$$m_{\Gamma_{\alpha,\beta}}(t) = \mathbb{E}\, e^{t\cdot\Gamma_{\alpha,\beta}} = \int_0^\infty e^{-(\beta-t)x}\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}\,\mathrm{d}x \underset{x\leftarrow\frac{x}{\beta-t}}{=} \left(\frac{\beta}{\beta-t}\right)^\alpha, \qquad t<\beta. \tag{5.7}$$

The moments of $X\sim\Gamma_{\alpha,\beta}$ are (cf. Exercise 5.1)

$$\mathbb{E}\, X^\gamma = \frac{\Gamma(\alpha+\gamma)}{\beta^\gamma\,\Gamma(\alpha)},\ \textit{in particular}\ \mathbb{E}\, X = \frac{\alpha}{\beta}\ \textit{and}\ \operatorname{var} X = \frac{\alpha}{\beta^2}. \tag{5.8}$$

Proposition 5.8. *If $X\sim\Gamma_{k,\beta}$ and $Y\sim\Gamma_{\ell,\beta}$ are independent, then $X+Y\sim\Gamma_{k+\ell,\beta}$ (cf. Exercise 5.12).*

Proof. As X and Y are independent we have with (5.7) that

$$\begin{aligned}\mathbb{E}\, e^{t(X+Y)} &= \mathbb{E}\, e^{tX}\cdot e^{tY} = \mathbb{E}\, e^{tX}\cdot\mathbb{E}\, e^{tY} = m_{\Gamma_{k,\beta}}(t)\cdot m_{\Gamma_{\ell,\beta}}(t)\\ &= \left(\frac{\beta}{\beta-t}\right)^k\cdot\left(\frac{\beta}{\beta-t}\right)^\ell = \left(\frac{\beta}{\beta-t}\right)^{k+\ell} = m_{\Gamma_{k+\ell,\beta}}(t)\end{aligned}$$

for all $t\in(-\infty,\beta)$ and thus the result. □

2nd, more explicit proof. The convolution is

$$\begin{aligned}f_{X+Y}(x) &= \int_0^x f_k(y)f_\ell(x-y)\,\mathrm{d}y = \int_0^x \frac{\beta^k y^{k-1}}{\Gamma(k)}e^{-\beta y}\frac{\beta^\ell(x-y)^{\ell-1}}{\Gamma(\ell)}e^{-\beta(x-y)}\,\mathrm{d}y\\ &\underset{y\leftarrow xu}{=} \frac{\beta^{k+\ell}x^{k+\ell-1}}{\Gamma(k+\ell)}e^{-\beta x}\cdot\frac{\Gamma(k+\ell)}{\Gamma(k)\,\Gamma(\ell)}\underbrace{\int_0^1 u^{k-1}(1-u)^{\ell-1}\,\mathrm{d}u}_{B(k,\ell)}\end{aligned} \tag{5.9}$$

and thus the result. □

Remark 5.9. The function

$$B(s,t) := \int_0^1 u^{s-1}(1-u)^{t-1}\,\mathrm{d}u = \frac{\Gamma(s)\cdot\Gamma(t)}{\Gamma(s+t)} \tag{5.10}$$

is the Beta function, also called Euler integral of the first kind (cf. Exercise 5.10). Note that $f_{X+Y}(\cdot)$ is a density and thus the identity (5.10) follows from (5.9).

Alternative proof of (5.10). Indeed, using the transformation $\begin{pmatrix}x(r,\varphi)\\ y(r,\varphi)\end{pmatrix} = r\begin{pmatrix}\cos^2\varphi\\ \sin^2\varphi\end{pmatrix}$ with Jacobian

$$\det\begin{pmatrix}\cos^2\varphi & 2r\cos\varphi\sin\varphi\\ \sin^2\varphi & 2r\cos\varphi\sin\varphi\end{pmatrix} = -2r\cos\varphi\sin\varphi,$$

it follows that

$$\begin{aligned}\Gamma(s)\cdot\Gamma(t) &= \int_0^\infty x^{s-1}e^{-x}\,\mathrm{d}x\cdot\int_0^\infty y^{t-1}e^{-y}\,\mathrm{d}y\\ &= 2\int_0^\infty\int_0^{\frac{\pi}{2}}\cos^{2s-1}\varphi\sin^{2t-1}\varphi\cdot r^{s+t-1}e^{-r}\,\mathrm{d}\varphi\,\mathrm{d}r = \Gamma(s+t)\cdot B(s,t)\end{aligned}$$

with $B(s,t) = 2\int_0^{\pi/2}\cos^{2s-1}\varphi\sin^{2t-1}\varphi\,\mathrm{d}\varphi \underset{\varphi\leftarrow\arccos\sqrt{u}}{=} \int_0^1 u^{s-1}(1-u)^{t-1}\,\mathrm{d}u$, the assertion. □

Remark 5.10. Euler's Beta function generalizes the binomial coefficient: for n, $k \in \mathbb{R}$, it holds that $\binom{n}{k} = \frac{1}{(n-1)B(n-k+1,k+1)}$.

Corollary 5.11. *If $X \sim E_{k,\lambda}$ and $Y \sim E_{\ell,\lambda}$ are independent, then $X + Y \sim E_{k+\ell,\lambda}$.*

Proposition 5.12. *Let X_i, $i = 1, \dots n$, be independent and standard normally distributed. Then*

$$X_1^2 + \dots + X_n^2 \sim \chi_n^2$$

follows a χ_n^2-distribution with n degrees of freedom.

Proof. We demonstrate first that $X^2 \sim \chi_1^2$ for $X \sim \mathcal{N}(0,1)$. Indeed,

$$\begin{aligned} F_{X^2}(x) &= P(X^2 \le x) = P\left(-\sqrt{x} \le X \le \sqrt{x}\right) = \Phi\left(\sqrt{x}\right) - \Phi\left(-\sqrt{x}\right) \\ &= \Phi\left(\sqrt{x}\right) - \left(1 - \Phi\left(\sqrt{x}\right)\right) = 2\,\Phi\left(\sqrt{x}\right) - 1 \end{aligned}$$

and hence

$$f_{X^2}(x) = F'_{X^2}(x) = \frac{2\,\varphi\left(\sqrt{x}\right)}{2\,\sqrt{x}} = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} = f_{\chi_1^2}(x)$$

and thus the assertion for $n = 1$, cf. (5.3).

It follows from (5.5) and Proposition 5.8 that

$$\chi_m^2 + \chi_n^2 \sim \Gamma_{\frac{m}{2},\frac{1}{2}} + \Gamma_{\frac{n}{2},\frac{1}{2}} \sim \Gamma_{\frac{m+n}{2},\frac{1}{2}} \sim \chi_{m+n}^2$$

for independent χ_m^2 and χ_n^2 random variables, thus the result. □

5.2 BOX–MULLER TRANSFORM AND THE POLAR METHOD

For two independent, normally distributed X, $Y \sim \mathcal{N}(0,1)$ it follows from Proposition 5.12 and (5.5) that

$$R^2 := X^2 + Y^2 \sim \chi_2^2 \sim E_{1,1/2} \sim E_{1/2} \sim 2E_1 \tag{5.11}$$

is *exponentially distributed* with rate $\lambda = 1/2$. This is the basis for the Box[2]–Muller[3] transform

$$U\left([0,1]^2\right) \sim \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \mapsto \underbrace{\sqrt{-2\log U_1}}_{R} \begin{pmatrix} \cos\left(2\pi U_2\right) \\ \sin\left(2\pi U_2\right) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

which Algorithm 1 exploits to generate normally distributed random variables.

```
Result: Realization of two independent, normally distributed random variables $X$ and $Y$
generate independent uniforms $U_1 \in [0,1]$ and $U_2 \in [0,1]$;
set $R := \sqrt{-2\log U_1}$;                  random radius $R$ with $R^2 \sim E_{1/2} \sim \chi_2^2$ by (5.11)
set $X := R \cdot \cos(2\pi U_2)$ and $Y := R \cdot \sin(2\pi U_2)$;           random angle $U_2$
return $(X, Y)$
```

Algorithm 1: Box–Muller transform, 1958

[2] George Box, 1919–2013

[3] Mervin E. Muller

Marsaglia's[4] polar method (Algorithm 2) is a variant of Box–Muller transform which avoids evaluating sin and cos, as these evaluations are numerically expensive. To this end note that $C^2 + S^2 \sim U([0,1])$ is uniformly distributed, if $(C, S) \sim U\left([-1,1]^2\right)$ are independent uniforms and $C^2 + S^2 \leq 1$. Indeed (cf. Exercise 5.13),

$$P\left(C^2 + S^2 \leq u \,\middle|\, C^2 + S^2 \leq 1\right) = P\left(\sqrt{C^2 + S^2} \leq \sqrt{u} \,\middle|\, \sqrt{C^2 + S^2} \leq 1\right) = \frac{\sqrt{u}^2\,\pi}{1^2\,\pi} = u. \tag{5.12}$$

```
Result: Realization of two independent, normally distributed random variables $X$ and $Y$
repeat
    generate independent uniforms $C \in [-1,1]$ and $S \in [-1,1]$;
until $U := C^2 + S^2 \leq 1$;
set $p := \sqrt{\frac{-2\log U}{U}}$;          note that $p^2\,U = -2\log U \sim E_{1,1/2} \sim \chi^2_2$ by (5.12) and (5.11)
return $(X = p \cdot C,\ Y = p \cdot S)$          $X = \sqrt{p^2\,U} \cdot \frac{C}{\sqrt{U}},\ Y = \sqrt{p^2\,U} \cdot \frac{S}{\sqrt{U}}$
```

Algorithm 2: Marsaglia polar method, 1964

The *Ziggurat algorithm* intends to reduce expensive evaluations as $\log(\cdot)$ and $\sqrt{\cdot}$ to a minimum.

5.3 STUDENT'S T-DISTRIBUTION

The statistics $\overline{X}_n$ and s_n^2 are dependent in general (well, s_n^2 explicitly involves $\overline{X}_n$, cf. (2.3)), and correlated (see Proposition 2.13). The next theorem discusses the situation for Gaussians.

Theorem 5.13 (Cochran's theorem; Gosset[5])**.** *Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$, be independent normals. Then the statistics $\overline{X}_n$ and s_n^2 are independent (sic!) and they follow the distributions*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{n-1}{\sigma^2} s_n^2 \sim \chi^2_{n-1}. \tag{5.13}$$

Proof. We shall assume first $\mu = 0$ and $\sigma = 1$. The matrix

$$U_n := \begin{pmatrix} 1/\sqrt{n} & \dots & \dots & \dots & 1/\sqrt{n} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & \dots & 0 \\ 1/\sqrt{2\cdot 3} & 1/\sqrt{2\cdot 3} & -2/\sqrt{2\cdot 3} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1/\sqrt{(n-1)n} & 1/\sqrt{(n-1)n} & \dots & 1/\sqrt{(n-1)n} & -\frac{n-1}{\sqrt{(n-1)n}} \end{pmatrix} \tag{5.14}$$

is unitary, i.e., $U_n^\top U_n = \mathbb{1}_n$ (Exercise 5.6). Define the linear transform (rotation) $Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} := U_n \cdot \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$.

By (3.6) the distribution is $\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim \mathcal{N}\left(U_n 0,\, U_n^\top\, \mathbb{1}_n\, U_n\right) = \mathcal{N}\left(0, \mathbb{1}_n\right)$ and by Theorem 3.16 the components Y_i

[4] George Marsaglia, 1924–2011

[5] William Sealy Gosset, 1876–1937, was an employee of the brewery Guinness (Dublin) and published as *Student*.

are *independent* normals. Note, that

$$Y_1 = \sqrt{n} \cdot \overline{X}_n$$

and

$$Y_2^2 + Y_3^2 + \dots + Y_n^2 = \left(\sum_{i=1}^n Y_i^2\right) - Y_1^2 = \left(\sum_{i=1}^n X_i^2\right) - n\,\overline{X}_n^2 \underset{(2.5)}{=} \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 = (n-1)\, s_n^2,$$

as U_n is unitary (and thus $\sum_{i=1}^n Y_i^2 = \|Y\|^2 = \|U_n X\|^2 = \|X\|^2 = \sum_{i=1}^n X_i^2$) and by (2.5). The statistics (2.1) and (2.2) are independent as $\overline{X}_n = f_1(Y_1)$ and $s_n^2 = f_2(Y_2, \dots, Y_n)$. Their distributions (5.13) are immediate as well, as Y_i are independent normals.

The assertion for general $\mu \in \mathbb{R}$ and $\sigma > 0$ follows by employing the transformation $\frac{X-\mu}{\sigma}$, the Z-transform. □

Remark 5.14. The statistics s_n *depends explicitly* on $\overline{X}_n$ by (2.2). However, for Gaussians, these quantities are stochastically *in*dependent. This is *not* true for non-Gaussians and even more, independence of $\overline{X}_n$ and s_n actually characterizes Gaussian random variables.

Definition 5.15. The density of Student's t-distribution with n degrees of freedom is

$$f_{t_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\,\pi}\,\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \qquad t \in \mathbb{R}.$$

Proposition 5.16 (Pointwise convergence)**.** *For every $t \in \mathbb{R}$ it holds that $f_{t_n}(t) \longrightarrow \varphi(t)$ (the density of the normal distribution) as $n \longrightarrow \infty$.*

Proof. Use that $\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{1+\frac{t^2}{n}}} \cdot \frac{1}{\sqrt{\left(1+\frac{t^2}{n}\right)^n}} \xrightarrow[n\to\infty]{} 1 \cdot \frac{1}{\sqrt{e^{t^2}}} = e^{-\frac{1}{2}t^2}$. □

For $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ iid normally distributed random variables we have that $\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$; even more, for X_i iid not necessarily normal we have from CLT (Theorem 4.3) that

$$\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The following theorem describes the corresponding distribution when σ^2 is not known explicitly but replaced by its estimate s_n^2.

Theorem 5.17. *Let X_i be iid $\mathcal{N}\left(\mu, \sigma^2\right)$ and $n > 1$, then the statistic*

$$t(X_1, \dots, X_n) := \sqrt{n}\,\frac{\overline{X}_n - \mu}{s_n} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{s_n^2}} \sim t_{n-1} \tag{5.15}$$

follows a Student t distribution with $n-1$ degrees of freedom.

Proof. The joint density of a random variable $Z \sim \mathcal{N}(0,1)$ and an independent variable $Y \sim \chi_{n-1}^2$ is

$$f_{Z,Y}(z,y) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \cdot \frac{y^{\frac{n-1}{2}-1}\,e^{-\frac{y}{2}}}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}.$$

The transformation $\begin{pmatrix} t \\ v \end{pmatrix} = g\begin{pmatrix} z \\ y \end{pmatrix} := \begin{pmatrix} z/\sqrt{y/(n-1)} \\ y \end{pmatrix}$ has inverse $\begin{pmatrix} z \\ y \end{pmatrix} := g^{-1}\begin{pmatrix} t \\ v \end{pmatrix} = \begin{pmatrix} t \cdot \sqrt{v/(n-1)} \\ v \end{pmatrix}$ and Jacobian $\det(g^{-1})'\begin{pmatrix} t \\ v \end{pmatrix} = \det\begin{pmatrix} \sqrt{\frac{v}{n-1}} & \cdots \\ 0 & 1 \end{pmatrix} = \sqrt{\frac{v}{n-1}}$. By (1.9), the density of the transformation $g\begin{pmatrix} Z \\ Y \end{pmatrix} = \begin{pmatrix} T \\ Y \end{pmatrix}$ with $T := \frac{Z}{\sqrt{Y/(n-1)}}$ is

$$f_{T,Y}(t,v) = \frac{e^{-\frac{1}{2}\frac{t^2 v}{n-1}}}{\sqrt{2\pi}} \cdot \frac{v^{\frac{n-1}{2}-1}\, e^{-\frac{v}{2}}}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)} \cdot \sqrt{\frac{v}{n-1}} = \frac{v^{\frac{n}{2}-1}\, e^{-\frac{v}{2}\left(1+t^2/(n-1)\right)}}{\sqrt{(n-1)\pi}\, 2^{\frac{n}{2}}\,\Gamma\left(\frac{n-1}{2}\right)},$$

its marginal density is (recall that $f_T(t)\,\mathrm{d}t = P(T \in \mathrm{d}t) = \int_0^\infty \underbrace{P(T \in \mathrm{d}t, Y \in \mathrm{d}v)}_{f(t,v)\,\mathrm{d}t\mathrm{d}v})$

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,Y}(t,v)\,\mathrm{d}v = \frac{1}{\sqrt{(n-1)\,\pi}\, 2^{\frac{n}{2}}\,\Gamma\left(\frac{n-1}{2}\right)} \int_0^\infty v^{\frac{n}{2}-1}\, e^{-v\frac{1}{2}\left(1+t^2/(n-1)\right)}\,\mathrm{d}v \\ &= \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} = f_{t_{n-1}}(t), \end{aligned}$$

i.e., $T \sim t_{n-1}$.

Recall finally from Theorem 5.13 that $Z := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$ and $Y := \frac{n-1}{\sigma^2} s_n^2 \sim \chi^2_{n-1}$ are independent. It follows that

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{s_n^2}} = \frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{s_n^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}} = T \sim t_{n-1},$$

the assertion. □

Proposition 5.18 (Cf. Proposition 2.21 and WolframMathWorld). *If*

(i) $\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho \cdot \sigma_X \sigma_Y \\ \rho \cdot \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$ *are bivariate normal, then the density of the correlation coefficient* r_n *(cf. (2.9)) is*

$$f_{r_n}(r) = \frac{(n-2)(n-2)!\left(1-\rho^2\right)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}}{\sqrt{2\pi}\Gamma\left(n-\frac{1}{2}\right)(1-\rho r)^{n-\frac{3}{2}}}\; {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; n - \frac{1}{2}; \frac{\rho r + 1}{2}\right),$$

where $-1 \le r \le 1$ *and* ${}_2F_1(a,b;c;z) = \sum_{k=1}^\infty \frac{a\cdot(a+1)\cdots(a+k-1)\cdot b\cdot(b+1)\cdots(b+k-1)}{c\cdot(c+1)\cdots(c+k-1)} \frac{z^k}{k!}$ *is the Gaussian hypergeometric function. Particularly,* $\mathbb{E}\, r_n = \rho - \rho\frac{1-\rho^2}{n-1}$ *and* $\operatorname{var} r_n = \frac{(1-\rho^2)^2}{n-1} + \mathcal{O}\left({}^1\!/n\right)$.

(ii) If X_i *and* Y_i *are bivariate normal and independent* ($\rho = 0$)*, then*

$$\sqrt{\frac{n-2}{1-r_n^2}}\, r_n \sim t_{n-2}. \tag{5.16}$$

The critical value for r_n according (5.16) is $r = \frac{t}{\sqrt{n-2+t^2}}$.

n	$\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sim 1+\frac{1}{4n}+\frac{9}{32n^2}$
1	—
2	1.25331...
3	1.12838...
10	1.02811...
100	1.00253...

Table 5.1: The factor to correct the standard deviation (5.17)

Theorem 5.19. *Let $X_i \sim \mathcal{N}\big(\mu, \sigma^2\big)$ be iid normals. Then (cf. Corollary 2.12 (iii) and Table 5.1)*

$$\mathbb{E}\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sqrt{s_n^2} = \sigma. \tag{5.17}$$

Proof. The distribution $\chi_n := \sqrt{\chi_n^2}$ ($g(x) = \sqrt{x}$ and $g^{-1}(x) = x^2$) has the density (cf. (5.1) and (1.9))

$$f_{\chi_n}(x) := f_{\chi_n^2}(x^2)\cdot 2x = \frac{x^{n-1}e^{-\frac{x^2}{2}}}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)}.$$

It follows from (5.13) that $\sqrt{(n-1)s_n^2} \sim \sigma\cdot\chi_{n-1}$. Using (5.1) we have that

$$\mathbb{E}\chi_{n-1} = \int_0^\infty x\cdot\frac{x^{n-2}e^{-\frac{x^2}{2}}}{2^{\frac{n-1}{2}-1}\Gamma\left(\frac{n-1}{2}\right)}\,\mathrm{d}x \underset{x\leftarrow\sqrt{2x}}{=} \int_0^\infty 2^{\frac{n-2}{2}}\frac{x^{\frac{n-2}{2}}e^{-x}}{2^{\frac{n-1}{2}-1}\Gamma\left(\frac{n-1}{2}\right)}\,\mathrm{d}x = \sqrt{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)},$$

from which the rest is immediate. □

5.4 FISHER’S F-DISTRIBUTION

Definition 5.20. Fisher’s F-distribution with m and n degrees of freedom has the density

$$f_{m,n}(x) = \frac{\left(\frac{m}{n}\right)^{m/2}}{B\left(m/2, n/2\right)}\cdot\frac{x^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}}, \qquad x \geq 0,$$

where B is Euler’s Beta function, cf. (5.10). We shall write $X \sim F_{m,n}$ for such random variables.

Proposition 5.21. *Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ be independent, then*

$$\frac{X/m}{Y/n} \sim F_{m,n}.$$

Proof. Define the transformation $g\begin{pmatrix}x\\y\end{pmatrix} := \begin{pmatrix}\frac{x/m}{y/n}\\ y\end{pmatrix}$ and note that $g^{-1}\begin{pmatrix}z\\v\end{pmatrix} = \begin{pmatrix}m\cdot z\cdot v/n\\ v\end{pmatrix}$ and $\det(g^{-1})' =$

$\det\begin{pmatrix} {}^{mv}\!/_{n} & \cdots \\ 0 & 1 \end{pmatrix} = \frac{mv}{n}$. Then, for independent $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ we find

$$\begin{aligned} f_{g(X,Y)}(z,v) &= f_X\left(\frac{m\,z\,v}{n}\right) f_Y(v) \cdot \frac{m\,v}{n} = \frac{\left(\frac{m\,z\,v}{n}\right)^{\frac{m}{2}-1} e^{-\frac{mzv}{2n}}}{2^{m/2}\Gamma\left({}^{m}\!/_{2}\right)} \cdot \frac{v^{\frac{n}{2}-1} e^{-\frac{v}{2}}}{2^{n/2}\Gamma\left({}^{n}\!/_{2}\right)} \cdot \frac{m\,v}{n} \\ &= \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{2^{\frac{m+n}{2}}\Gamma\left({}^{m}\!/_{2}\right)\Gamma\left({}^{n}\!/_{2}\right)} v^{\frac{m+n}{2}-1} e^{-\frac{v}{2}\left(1+\frac{m}{n}z\right)} \end{aligned}$$

with marginal distribution

$$f_{\frac{X/m}{Y/n}}(z) = \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{2^{\frac{m+n}{2}}\Gamma\left({}^{m}\!/_{2}\right)\Gamma\left({}^{n}\!/_{2}\right)} \int_0^\infty v^{\frac{m+n}{2}-1} e^{-\frac{v}{2}\left(1+\frac{m}{n}z\right)}\,\mathrm{d}v.$$

Now substitute $v \longleftarrow v\frac{2}{1+\frac{m}{n}z}$ to get

$$\begin{aligned} f_{\frac{X/m}{Y/n}}(z) &= \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{2^{\frac{m+n}{2}}\Gamma\left({}^{m}\!/_{2}\right)\Gamma\left({}^{n}\!/_{2}\right)} \cdot \left(\frac{2}{1+\frac{m}{n}z}\right)^{\frac{m+n}{2}} \cdot \int_0^\infty v^{\frac{m+n}{2}-1} e^{-v}\,\mathrm{d}v \\ &= \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} \Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left({}^{m}\!/_{2}\right)\Gamma\left({}^{n}\!/_{2}\right)} \cdot \frac{z^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}z\right)^{\frac{m+n}{2}}}, \end{aligned}$$

which is the assertion. □

5.5 PROBLEMS

Exercise 5.1. *Verify the moments of the Gamma distribution, Eq. (5.8).*

Exercise 5.2. *Show that*

$$\mathbb{E}\, X = n \textit{ and } \operatorname{var} X = 2n$$

for $X \sim \chi^2_n$.

Exercise 5.3. *For every $x \in \mathbb{R}$ it holds that $\sqrt{2n} \cdot f_{\chi^2_n}\left(n + x\sqrt{2n}\right) \xrightarrow[n\to\infty]{} \varphi(x)$.*

Exercise 5.4. *Show that $\overline{X}_n \sim \Gamma_{n\alpha, n\beta}$ for independent $X_i \sim \Gamma_{\alpha,\beta}$.*

Exercise 5.5. *Use $g(x) = 2\lambda x$ and (1.9) to verify Remark 5.6.*

Exercise 5.6. *Show that (5.14) is unitary.*

Exercise 5.7. *If $X \sim F_{n,m}$, then ${}^{1}\!/_{X} \sim F_{m,n}$.*

Exercise 5.8. *If $X \sim t_n$ is Student, then $X^2 \sim F_{1,n}$ and $X^{-2} \sim F_{n,1}$.*

Exercise 5.9. *Show that $\operatorname{var} X = \frac{n}{n-2}$ for $X \sim t_n$ and $n > 2$.*

Beta distribution

Exercise 5.10 (Beta distribution). *The density of the Beta distribution* $\mathrm{B}_{\alpha,\beta}$ *with parameters* $\alpha > 0$ *and* $\beta > 0$ *is*

$$f_{B_{\alpha,\beta}}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad x \in (0,1),$$

where $B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,\mathrm{d}u$ *is the Beta function cf. (5.10).*

Show that $\mathbb{E}\, X^\gamma = \frac{B(\alpha+\gamma,\beta)}{B(\alpha,\beta)}$ *for* $X \sim \mathrm{B}_{\alpha,\beta}$. *In particular,*

$$\mathbb{E}\, X = \frac{\alpha}{\alpha+\beta} \text{ and } \operatorname{var} X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Exercise 5.11. *If* $X \sim \mathrm{B}_{\frac{m}{2},\frac{n}{2}}$, *then* $\frac{nX}{m(1-X)} \sim F_{m,n}$ *or equivalently: if* $Y \sim F_{m,n}$, *then* $\frac{mY/n}{1+mY/n} \sim \mathrm{B}_{\frac{m}{2},\frac{n}{2}}$.

Exercise 5.12. *If* $X \sim \Gamma_{k,\beta}$ *and* $Y \sim \Gamma_{\ell,\beta}$ *are independent, then*

$$\frac{X}{X+Y} \sim \mathrm{B}_{k,\ell} \text{ and } X+Y \sim \Gamma_{k+\ell,\beta}$$

and they are independent as well.

Hint: the transform $g\begin{pmatrix}x\\y\end{pmatrix} := \begin{pmatrix}x+y\\ \frac{x}{x+y}\end{pmatrix}$ *has inverse* $g^{-1}\begin{pmatrix}u\\v\end{pmatrix} = \begin{pmatrix}u\,v\\u(1-v)\end{pmatrix}$ *and Jacobian* $\left|\det\left(g^{-1}\right)'\begin{pmatrix}u\\v\end{pmatrix}\right| = u$.

Exercise 5.13 (Cf. (5.12)). *Let* $U_1, U_2 \sim U[0,1]$ *be uniformly distributed and independent. Show that*

$$P\big(U_1^2+U_2^2 \in \mathrm{d}u\big) = \begin{cases} \frac{\pi}{4}\,\mathrm{d}u & \text{if } u \in [0,1],\\ \left(\frac{\pi}{4} - \arctan\sqrt{u-1}\right)\mathrm{d}u & \text{if } u \in [1,2]. \end{cases}$$

Note particularly that $P\left(U_1^2+U_2^2 \le u \mid U_1^2+U_2^2 \le 1\right) = u$, *i.e.,* $R^2 := U_1^2+U_2^2$ *is uniformly distributed provided that* $R \le 1$.

Statistical Hypothesis Testing

6

Let the data speak for themselves.

attributed to John W. Tukey, 1915–2000

6.1 MATHEMATICAL SETTING AND DEFINITIONS

Definition 6.1. A *statistical model*[1] is

$$\mathcal{E} := (\mathcal{X}, \Sigma, (P_\vartheta)_{\vartheta\in\Theta}),$$

where
(i) $\mathcal{X}$ is the sample space (Stichprobenraum),
(ii) Σ is the sigma algebra,
(iii) $\vartheta \in \Theta$ is a parameter in a set of parameters and
(iv) $\mathcal{P} := \{P_\vartheta \colon \vartheta \in \Theta\}$ is a family of probability measures, $P_\vartheta \colon \Sigma \to [0, 1]$.

A parametrization is said to be *identifiable*, if $\vartheta_1 \neq \vartheta_2 \implies P_{\vartheta_1} \neq P_{\vartheta_2}$ (i.e., the mapping $\vartheta \mapsto P_\vartheta$ is injective).

The model is said to be
- *discrete*, if $\mathcal{X}$ is finite or countably finite;
- *parametric*, if $\Theta \subset \mathbb{R}^d$ for some $d \in \{1, 2, 3, \dots\}$ and *nonparametric* else;
- *binary*, if $\{P_\vartheta \colon \vartheta \in \Theta\} = \{P_0,\ P_1\}$ with $P_0 \neq P_1$.

Remark 6.2. We shall typically write $\mathbb{E}_\vartheta$ ($\operatorname{var}_\vartheta$, resp.) for the expectation (variance, resp.) with respect to the probability measure P_ϑ; e.g., $\mathbb{E}_\vartheta\, X = \int_{\mathcal{X}} X\, \mathrm{d}P_\vartheta$, etc.

The typical problem in statistics is to decide which distribution a sample (data) $x \in \mathcal{X}$ is from.

Definition 6.3. A random variable $t \colon \mathcal{X} \to [0, 1]$ is a *statistical test*. The test is
(i) *non-randomized*, if $\mathcal{X} \xrightarrow{t} \{0, 1\}$,
(ii) a general test $\mathcal{X} \xrightarrow{t} [0, 1]$ is called *randomized*.

For a (non-randomized) test $t(\cdot)$ we shall associate the sets

- $\{x \in \mathcal{X} \colon t(x) = 0\}$ *acceptance region*,[2] and
- $\{x \in \mathcal{X} \colon t(x) = 1\} = C_t$ is the *rejection region* or *critical region*.[3]

Remark 6.4 (Conservative test). Suppose that $t(\cdot)$ is a randomized test, set $p := t(x)$. Then reject H_0 with probability p, i.e., the final decision depends on a further experiment which is independent from the data x.

[1] statistisches Experiment, statistisches Modell

[2] *Annahmebereich*

[3] *Ablehnbereich*, auch *kritischer Bereich*

In practice, the (non-randomized) more conservative test

$$\lfloor t \rfloor (x) := \lfloor t(x) \rfloor = \begin{cases} 1 & \text{if } t(x) = 1, \\ 0 & \text{if } t(x) < 1 \end{cases}$$

is employed instead of the randomized test $t(\cdot)$ (i.e., $\lfloor t \rfloor \le t$).

Remark 6.5. The set of randomized tests is convex: if t_0 and t_1 are (randomized) tests, then $(1-\lambda)\, t_0 + \lambda\, t_1$ is a randomized test as well.

6.2 HYPOTHESIS TESTING: TYPES OF ERRORS

A hypothesis is proposed, the *null hypothesis* H_0 as opposed to the *alternative hypothesis* H_1. To investigate this setting we assume that

$$\Theta = \Theta_0 \,\dot\cup\, \Theta_1 \quad (\text{i.e., } \Theta_0 \cap \Theta_1 = \emptyset \text{ and } \Theta = \Theta_0 \cup \Theta_1)$$

and for an identifiable parametrization thus

$$\mathcal{P} := \underbrace{\{P_\vartheta : \vartheta \in \Theta_0\}}_{=:\mathcal{P}_0} \,\dot\cup\, \underbrace{\{P_\vartheta : \vartheta \in \Theta_1\}}_{=:\mathcal{P}_1}.$$

We shall consider *binary tests* with $\mathcal{P} = \{P := P_0,\ Q := P_1\}$ first and address *composite hypotheses* later.

The test setting. The test problem is usually formulated in terms of an hypothesis versus an alternative. The

- ▹ null hypothesis, H_0: the sample x originates from $P(\cdot) = P_0(\cdot) = P(\cdot \mid H_0)$

is tested against the

- ▹ Alternative, H_1: the sample x originates from $Q(\cdot) = P_1(\cdot) = P(\cdot \mid H_1)$;

here we have implicitly introduced different notations in frequent use.

We shall attribute a sample $x \in \mathcal{X}$ to P_0 if $t(x) = 0$ and to P_1 if $t(x) = 1$.

Definition 6.6. Power and types of errors, cf. Table 6.1.

- ▹ *Type I error*:[4] decision for $Q = P_1$, i.e., $t(x) = 1$, although the sample x is drawn from $P = P_0$. The probability of a type I error is

$$\alpha := \mathbb{E}_P\, t = \int_{\mathcal{X}} t \,\mathrm{d}P = \int_{\mathcal{X}} t(x)\, P(\mathrm{d}x)$$

(i.e., $\alpha = \mathbb{E}_P\, \mathbb{1}_{\{t=1\}} = P(t=1) = P(t>0)$ if t is non-randomized). α is also called the *statistical significance* of the test.

- ▹ *Type II error*:[5] decision for $P = P_0$, i.e., $t(x) = 0$, although the sample x is drawn from $P_1 = Q$. The probability of this misclassification is

$$\beta := \mathbb{E}_Q(1-t) = \int_{\mathcal{X}} \big(1 - t(x)\big)\, Q(\mathrm{d}x)$$

(i.e., $\beta = \mathbb{E}_Q\, \mathbb{1}_{\{t=0\}} = Q(t=0)$ if t is non-randomized).

[4] Fehler erster Art, Signifikanz des Testes, Niveau, Signifikanzniveau

[5] Fehler zweiter Art

	decision about null hypothesis	
null hypothesis H_0 is	accept the null hypothesis H_0 $t = 0$	reject the null hypothesis H_0 $t = 1$
true	correct inference true negative $P(t = 0 \mid H_0) \ge 1 - \alpha$	wrong decision false positive type I error, α $P(t = 1 \mid H_0) \le \alpha$
false (i.e., H_1)	wrong decision false negative type II error, β $P(t = 0 \mid H_1) \le \beta$	correct inference true positive $P(t = 1 \mid H_1) \ge 1 - \beta$

Table 6.1: Error types for binary tests

- ▹ The *power*[6] of a statistical test $t(\cdot)$ is

$$\pi_t(Q) \coloneqq \mathbb{E}_Q\, t = \int_{\mathcal{X}} t(x)\, Q(\mathrm{d}x) = 1 - \beta. \tag{6.1}$$

Desirably, the test $t(\cdot)$ should be chosen so that

$$\int_{\mathcal{X}} t(x)\, P(\mathrm{d}x) \text{ is small and } \int_{\mathcal{X}} t(x)\, Q(\mathrm{d}x) \text{ is large.} \tag{6.2}$$

Memory hook for (6.2):

- ▹ The *price* $\int_{\mathcal{X}} t\, \mathrm{d}P$ should be *small*, but
- ▹ the *quality* $\int_{\mathcal{X}} t\, \mathrm{d}Q$ should be *high*.

There are two major paradigms to construct a statistical test:

6.2.1 Neyman–Pearson

The Neyman[7]–Pearson[8] test specified as follows involves the statistical significance $\alpha \in (0, 1)$, which is chosen and fixed.

Problem: find the test statistics $t\colon \mathcal{X} \to [0, 1]$ so that

$$\begin{aligned} &\underset{t\colon \mathcal{X}\to[0,1]}{\text{maximize}} \int_{\mathcal{X}} t(x)\, Q(\mathrm{d}x), \\ &\text{subject to } \int_{\mathcal{X}} t(x)\, P(\mathrm{d}x) \le \alpha, \end{aligned} \tag{6.3}$$

where the maximum is among all feasible test statistics $t(\cdot)$.

Remark 6.7. Typical α-values often used in practice include $\alpha = 10\,\%, 5\,\%, 1\,\%, 0.1\,\%$. They are small, as α describes the type I error.

[6] Güte, Schärfe, Trennschärfe, Teststärke, Operationscharakteristik

[7] Jerzy Neyman, 1894–1981, Polish mathematisian

[8] Egon Pearson, 1895–1980, British statistician

observation	H	d	decision
$X<Y<Z$	1	1	correct
$X<Z<Y$	1	1	correct
$Y<X<Z$	0	1	wrong
$Y<Z<X$	0	0	correct
$Z<X<Y$	1	0	wrong
$Z<Y<X$	0	0	correct

Table 6.2: Randomized decision rule

6.2.2 Bayes

To select a test $t(\cdot)$ according a Bayes[9] paradigm let $\lambda \in (0,1)$ be chosen. Find a test $t(\cdot)$ which minimizes

$$(1-\lambda)\int_{\mathcal{X}} t(x)\,P(\mathrm{d}x) + \lambda \int_{\mathcal{X}} \big(1-t(x)\big)\,Q(\mathrm{d}x) \to \min! \tag{6.4}$$

6.3 RANDOMIZED DECISION RULE

Consider a random variable (X,Y) with $P(X<Y) = P(Y<X) = \frac{1}{2}$ and the randomized decision rule

$$d(x) := \begin{cases} 1 & \text{if } x \le Z \\ 0 & \text{else} \end{cases}$$

for an independent random variable Z to decide on

$$\begin{aligned} H_0 &\colon Y < X \text{ versus} \\ H_1 &\colon X \le Y. \end{aligned}$$

The decision rule $d(X)$ is successful (cf. Table 6.2) with probability

$$P\,(X<Y<Z) + P(X<Z<Y) + P(Y<Z<X) + P(Z<Y<X); \tag{6.5}$$

it is not successful in the remaining cases $\{Y<X<Z\}$ or $\{Z<X<Y\}$ for which we have that

$$\begin{aligned} P\,(X<Y<Z) = P\,(Y<X<Z) &= \frac{1}{2}P(X,Y<Z) \text{ and} \\ P\,(Z<Y<X) = P\,(Z<X<Y) &= \frac{1}{2}P(Z<X,Y), \end{aligned}$$

as Z is independent. It follows that

$$\begin{aligned} (6.5) &= \frac{1}{2}\big(P\,(X<Y<Z) + P\,(Y<X<Z)\big) \\ &\quad + P(X<Z<Y) + P(Y<Z<X) \\ &\quad + \frac{1}{2}\big(P\,(Z<X<Y) + P\,(Z<Y<X)\big) \\ &= \frac{1}{2} + \frac{1}{2}\big(P(X<Z<Y) + P(Y<Z<X)\big) \\ &> \frac{1}{2}, \end{aligned}$$

[9] Thomas Bayes, 1701–1761, English statistician and philosopher

if Z has strictly positive density on $\mathbb{R}$.

6.4 PROBLEMS

Exercise 6.1 (Coronavirus)**.** *A test for COVID-19 is designed along the hypothesis*

$$H_0\colon\ \textit{the person has corona}.$$

Let t be a test. The person with $t = 0$ is tested positive (probably has corona) while a person with result $t = 1$ is tested negative (i.e., the person presumably does to not have corona). What does the false positive *decision describe?*

Exercise 6.2 (Coronavirus, cntd.)**.** *The RT-PCR-test t for COVID-19 has the reliable and convincing properties (estimated by the British Medical Journal in 2020)*

(i) $P(t = 0 \mid H_0) = 80\,\%$ *(the* sensitivity[10] *of the test) and*

(ii) $P(t = 1 \mid H_1) = 98\,\%$ *(the* specificity[11]*of the test).*

Give the probability of the type I error *(false positive) and the probability of the* type II error *(false negative).*

Suppose further that the prevalence[12] *of the population is $P(H_0) = 1$ ‰, i.e., one out of 1000 randomly chosen persons has corona. Prove Bayes' formula $P(B \mid A) = \frac{P(A|B)\,P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$ and verify that*

$$P\left(H_1 \mid t = 1\right) \approx 99.98\,\%, \quad \textit{but} \quad P\left(H_0 \mid t = 0\right) \approx 3.85\,\%.$$

What are your conclusions given this surprising, probably shocking result? How do the results change, if the prevalence is 5% (as in an old people's home, say)?

Exercise 6.3 (Coronavirus, cntd.)**.** *The test is apparently useless, unless*

$$P\left(H_0 \mid t = 0\right) > P(H_0) \textit{ and } P\left(H_1 \mid t = 1\right) > P(H_1). \tag{6.6}$$

Show that (6.6), iff

$$P(t = 0 \mid H_0) + (t = 1 \mid H_1) > 1.$$

Show that t and H are independent, iff $P(t = 0 \mid H_0) + (t = 1 \mid H_1) = 1$.

Exercise 6.4. *John visits the doctor claiming some discomfort. The doctor is led to believe that he may have some disease A. He then takes some standard procedures for this case: he examines John, carefully observes the symptoms and runs routine laboratory examinations.*

The doctor assumes that $P\left(A|\, H\right) = 0.7$, where H (history) contains the information John provided and all other relevant knowledge he has learned from former patients. To improve the evidence about the illness, the doctor asks John to undertake an independent *examination.*

Examination t provides an uncertain result of the positive negative type with probabilities

$$\begin{cases} P\left(t = 1|\, A^c\right) = & 0.40 \textit{ (test positive, provided non-disease A) and} \\ P\left(t = 1|\, A\right) = & 0.95 \textit{ (test positive, provided disease A).} \end{cases}$$

John goes through the examination with result $t = 1$.

[10] Sensitivität, dt.

[11] Spezifität, dt.

[12] Prävalenz, dt.

(i) What should the doctor infer about John's disease?

(ii) The doctor decides to ask John to undertake a second, more efficient, but also more expensive test $\tilde{t}$ with probabilities

$$\begin{cases} P\left(\tilde{t} = 1 \middle| A^c\right) = & 0.04 \textit{ and} \\ P\left(\tilde{t} = 1 \middle| A\right) = & 0.99. \end{cases}$$

Which result can the doctor predict for the second test $\tilde{t}$?

(iii) The result of the second test is $\tilde{t} = 0$. What should the doctor infer about John's disease?

7 The Neyman–Pearson test

Statistics is the grammar of science.

Karl Pearson, 1857–1936

Let P (Q, resp.) have the density f (g, resp.), i.e., $P(C) = \int_C f(x)\,\mathrm{d}x$ ($Q(C) = \int_C g(x)\,\mathrm{d}x$, resp.) or

$$P(\mathrm{d}x) = f(x)\,\mathrm{d}x, \ (Q(\mathrm{d}x) = g(x)\,\mathrm{d}x, \ \text{resp.}).$$

Note that we have

$$Q(\mathrm{d}x) = g(x)\,\mathrm{d}x = \frac{g(x)}{f(x)} f(x)\,\mathrm{d}x = \frac{g(x)}{f(x)} P(\mathrm{d}x) \tag{7.1}$$

to change the measure.

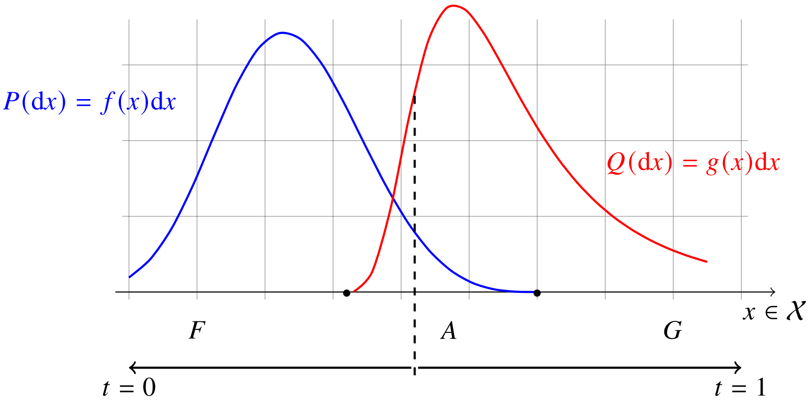
Definition 7.1 (Likelihood ratio)**.** The *likelihood ratio*[1] is the statistic $R(x) := \frac{g(x)}{f(x)}$.

Remark 7.2. With (7.1) we have that $R(\cdot)$ is the Radon–Nikodym derivative, $\mathrm{d}Q = R\,\mathrm{d}P$. We thus have $Q(C) = \int_{\mathcal{X}} \mathbb{1}_C \,\mathrm{d}Q = \int_{\mathcal{X}} \mathbb{1}_C \cdot R\,\mathrm{d}P$ or, by taking linear combinations, $\mathbb{E}_Q\, Y = \mathbb{E}_P\,(Y \cdot R)$.

7.1 DEFINITION

Definition 7.3. A test $t\colon \mathcal{X} \to [0,1]$ is a Neyman–Pearson test if

$$t(x) = \begin{cases} 1 & \text{iff } \frac{g(x)}{f(x)} > c \text{ or } x \in G, \\ 0 & \text{iff } \frac{g(x)}{f(x)} < c \text{ or } x \in F, \end{cases}$$



$P(\mathrm{d}x) = f(x)\mathrm{d}x$
$Q(\mathrm{d}x) = g(x)\mathrm{d}x$
$x \in \mathcal{X}$
F
A
G
$t = 0$
$t = 1$

Figure 7.1: Neyman–Pearson test

[1] Dichtequotient, in German

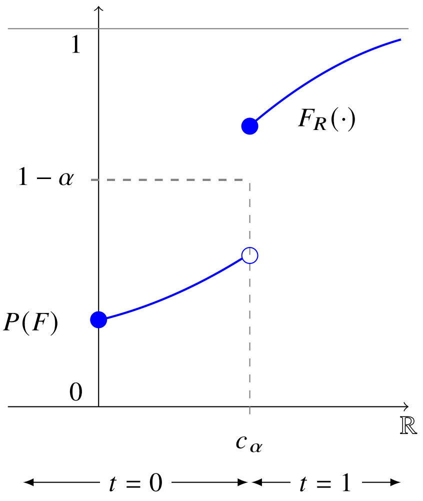


Figure 7.2: cdf of the likelihood ratio

where $c \geq 0$ and (cf. Figure 7.1)

$$\begin{aligned} F &:= \{x\colon f(x) > 0,\ g(x) = 0\}, \\ A &:= \{x\colon f(x) > 0,\ g(x) > 0\}, \\ G &:= \{x\colon f(x) = 0,\ g(x) > 0\}. \end{aligned}$$

We shall write $t \in \mathcal{N}(P, Q)$ for a Neyman–Pearson test of

$$\begin{aligned} &H_0\colon X \sim P = P_0 \text{ versus} \\ &H_1\colon X \sim Q = P_1. \end{aligned}$$

7.2 EXISTENCE

The Neyman Pearson test satisfies

$$0 \leq \int_{\mathcal{X}} t \,\mathrm{d}P \leq 1 - P(F) \text{ and } Q(G) \leq \int_{\mathcal{X}} t \,\mathrm{d}Q \leq 1.$$

Indeed, $0 \leq \int_{\mathcal{X}} t \,\mathrm{d}P \leq \int_{F^\mathrm{c}} 1 \,\mathrm{d}P = 1 - P(F)$ and $1 - \int_{\mathcal{X}} t \,\mathrm{d}Q = \int_{\mathcal{X}} 1 - t \,\mathrm{d}Q \leq \int_{G^\mathrm{c}} 1 \,\mathrm{d}Q = 1 - Q(G)$.

Lemma 7.4. *For every $\alpha < P(F^\mathrm{c})$ there is a (possibly randomized) Neyman–Pearson test $t_\alpha(\cdot)$ with type I error α, i.e., $\int_{\mathcal{X}} t_\alpha(x) P(\mathrm{d}x) = \alpha$.*

Proof. Set

$$F(c) := F_R(c) = P\left(\left\{x \in \mathcal{X}\colon \frac{g(x)}{f(x)} \leq c\right\}\right), \tag{7.2}$$

i.e., $F(\cdot)$ is the cdf. of the likelihood ratio $R(x) = \frac{g(x)}{f(x)}$ under P. Define the quantile function (inverse cdf) $F^{-1}(p) := \inf\{c\colon F(c) \geq p\}$ and set

$$c_\alpha := c(\alpha) := F^{-1}(1 - \alpha). \tag{7.3}$$

(i) If $F\big(c_\alpha\big) > 1-\alpha$, then c_α is a point of *dis*continuity of $F(\cdot)$ (cf. Figure 7.2) and the left limit satisfies $F\big(c_\alpha - 0\big) \le 1-\alpha$ and $P(R = c_\alpha) = F(c_\alpha) - F(c_\alpha - 0) > 0$; set

$$q_\alpha \coloneqq \frac{F(c_\alpha) - (1-\alpha)}{F(c_\alpha) - F(c_\alpha - 0)}.$$

Define the Neyman–Pearson test

$$t_\alpha(x) \coloneqq \begin{cases} 1 & \text{if } x \in G \text{ or } \frac{g(x)}{f(x)} > c_\alpha, \\ q_\alpha & \text{if } \frac{g(x)}{f(x)} = c_\alpha, \\ 0 & \text{if } x \in F \text{ or } \frac{g(x)}{f(x)} < c_\alpha. \end{cases} \tag{7.4}$$

It holds that

$$\begin{aligned} \int_X t_\alpha(x) P(\mathrm{d}x) &= \int_{\left\{\frac{g(x)}{f(x)} > c_\alpha\right\}} \mathrm{d}P + q_\alpha \cdot \Big(F\big(c_\alpha\big) - F\big(c_\alpha - 0\big)\Big) \\ &= 1 - F(c_\alpha) + F(c_\alpha) - (1-\alpha) = \alpha, \end{aligned}$$

which is the desired level.

(ii) Otherwise, $F(c_\alpha) = 1-\alpha$ (this is certainly the case if c_α is a point of continuity of $F(\cdot)$).

It follows that the test (7.4) is Neyman–Pearson with the desired level. □

Remark 7.5. The Neyman–Pearson test t_α is randomized, iff $P(R = c_\alpha) > 0$ and then rejects with probability q_α. The quantities are related by

$$P(R > c_\alpha) + q_\alpha \cdot P(R = c_\alpha) = \alpha.$$

Remark 7.6. The decision of the Neyman–Pearson test is based on the likelihood ratio

$$R(x) = \frac{g(x)}{f(x)} \tag{7.5}$$

and the more conservative test is

$$\lfloor t_\alpha \rfloor (x) = \begin{cases} 1 & \text{if } R(x) > c_\alpha, \\ 0 & \text{if } R(x) \le c_\alpha. \end{cases}$$

Rejection. The rejection of H_0 can be formulated in the following two ways:

(i) Randomized: reject H_0, if

$$R(x) > \underbrace{c_\alpha}_{\text{critical value}} \tag{7.6}$$

and randomize if $R(x) = c_\alpha$.

(ii) Non-randomized: reject H_0, if

$$\underbrace{1 - F_R\big(R(x)\big)}_{p\text{-value}} < \alpha, \tag{7.7}$$

where F_R is the cdf of the test statistics $R(\cdot)$ under P.
This is how statistical program packages operate and $1 - F_R(R(x))$ is the *p-value*. Note that the test (7.7) for given α is automatically conservative in the sense of Remark 6.4 and randomization is not necessary any longer.

The *critical region* or *rejection region* of the test is

$$C_t \coloneqq \{x \colon R(x) > c_\alpha\}$$

so that the test can be formulated as

$$t(x) = \begin{cases} 1 & \text{if } x \in C_t, \\ 0 & \text{else} \end{cases} = \mathbb{1}_{C_t}(x) = \begin{cases} 1 & \text{if } R(x) > c_\alpha, \\ 0 & \text{else.} \end{cases}$$

The

- ▹ *acceptance region* is the set of values of the test statistic for which the null hypothesis is not rejected, i.e., C_t^c.
- ▹ The *critical values* of a statistical test $R(\cdot)$ are the boundary points of the acceptance region of the test, i.e., ∂C_t, cf. (7.6). Depending on the shape of the acceptance region, there can be one or more than one critical value.

Remark 7.7. Under P, the p-value (7.7) is uniformly distributed.

7.3 THE MOST POWERFUL TEST

Proposition 7.8. *Let $t(\cdot)$ be a Neyman–Pearson test with critical value c_t for P versus Q and $\psi(\cdot)$ be any test. Then*

$$c_t \cdot \left(\int_X t \,\mathrm{d}P - \int_X \psi \,\mathrm{d}P\right) \le \int_X t \,\mathrm{d}Q - \int_X \psi \,\mathrm{d}Q. \tag{7.8}$$

Proof. By definition of the Neyman–Pearson test we have that,

$$\begin{aligned} t(x) = 1 \ge \psi(x) &\text{ iff } \frac{g(x)}{f(x)} > c_t \text{ and} \\ t(x) = 0 \le \psi(x) &\text{ iff } \frac{g(x)}{f(x)} < c_t \end{aligned}$$

and thus

$$0 \le \big(t(x) - \psi(x)\big) \cdot \left(\frac{g(x)}{f(x)} - c_t\right) \tag{7.9}$$

whenever $f(x) > 0$, i.e., $x \in A \cup F$. By taking P-expectations of (7.9) with (7.1),

$$c_t \cdot \int_{A\cup F} t - \psi \,\mathrm{d}P \le \int_{A\cup F} (t - \psi)\,\frac{g}{f}\,\mathrm{d}P = \int_{A\cup F} t - \psi \,\mathrm{d}Q.$$

Note that $P(G) = 0$ and $t|_G = 1$, thus

$$c_t \cdot \int_G t - \psi \,\mathrm{d}P = 0 \le \int_G \underbrace{t - \psi}_{\ge 0} \,\mathrm{d}Q.$$

Adding the latter displays gives

$$c_t \cdot \int_X t - \psi \,\mathrm{d}P \le \int_X t - \psi \,\mathrm{d}Q,$$

which is the assertion. □

α	$\Phi^{-1}(\alpha)$
90 %	1.282
95 %	1.645
99 %	2.326
99.5 %	2.576
99.9 %	3.090
99.95 %	3.291
99.99 %	3.719

(a) normal distribution

α	K_α, cf. (9.4)	$c(\alpha)=\sqrt{-\frac{1}{2}\log\frac{1-\alpha}{2}}$
90 %	1.07	1.22
95 %	1.22	1.36
99 %	1.52	1.63
99.5 %	1.63	1.73
99.9 %	1.86	1.95

(b) Kolmogorov Smirnov

Table 7.1: Quantiles

Lemma 7.9 (Neyman–Pearson lemma)**.** *Let $t(\cdot)$ be a Neyman–Pearson test and $\psi(\cdot)$ any other test.*

(i) If $\int_{\mathcal{X}} t\,\mathrm{d}P \ge \int_{\mathcal{X}} \psi\,\mathrm{d}P$, then $\int_{\mathcal{X}} t\,\mathrm{d}Q \ge \int_{\mathcal{X}} \psi\,\mathrm{d}Q$ and

(ii) if $\int_{\mathcal{X}} t\,\mathrm{d}Q \le \int_{\mathcal{X}} \psi\,\mathrm{d}Q$, then $\int_{\mathcal{X}} t\,\mathrm{d}P \le \int_{\mathcal{X}} \psi\,\mathrm{d}P$.

(iii) The Neyman–Pearson test $t_\alpha(\cdot)$ (cf. Lemma 7.4) solves problem (6.3) with type I error α, i.e.,

$$\int_{\mathcal{X}} t_\alpha\,\mathrm{d}Q = \sup\left\{\int_{\mathcal{X}} \psi\,\mathrm{d}Q \colon \psi \text{ is a test with } \int_{\mathcal{X}} \psi\,\mathrm{d}P \le \alpha\right\}. \tag{7.10}$$

Proof. (i) and (ii) are direct consequences of (7.8).

As for (iii) assume that $\int_{\mathcal{X}} \psi\,\mathrm{d}P \le \alpha$. By Lemma 7.4 there is a test t_α with $\int_{\mathcal{X}} t_\alpha\,\mathrm{d}P = \alpha$ so that $\int_{\mathcal{X}} t_\alpha\,\mathrm{d}P = \alpha \ge \int_{\mathcal{X}} \psi\,\mathrm{d}P$. We conclude from (i) that $\int_{\mathcal{X}} t_\alpha\,\mathrm{d}Q \ge \int_{\mathcal{X}} \psi\,\mathrm{d}Q$ and hence "$\ge$" in (7.10). Equality is obtained for $\psi = t_\alpha$. □

Example 7.10. For $\mathcal{X} = \mathbb{R}$ consider the test problem

$$\begin{aligned} &H_0\colon P \sim E_{\lambda=1} \text{ (exponential with rate parameter 1, cf. Definition 5.4)},\\ &H_1\colon Q \sim E_{\lambda=2}. \end{aligned}$$

Note first that $f(x) = e^{-x}$ and $g(x) = 2e^{-2x}$, the likelihood ratio is

$$R(x) = \frac{g(x)}{f(x)} = 2e^{-x}. \tag{7.11}$$

The cdf of the likelihood ratio (7.2) is

$$F_R(c) = P_{E_1}\left(2e^{-x} \le c\right) = \int_{\{x\colon 2e^{-x}\le c\}} e^{-x}\,\mathrm{d}x = \int_{\left\{x \ge -\log\frac{c}{2}\right\}} e^{-x}\,\mathrm{d}x = -e^{-x}\big|_{x=-\log\frac{c}{2}}^{\infty} = \frac{c}{2}.$$

From (7.3) it follows that $c_\alpha = F_R^{-1}(1-\alpha) = 2 - 2\alpha$. The Neyman–Pearson test (7.4) finally reads

$$t_\alpha(x) = \begin{cases} 1 & \text{if } 2e^{-x} > 2 - 2\alpha,\\ 0 & \text{if } 2e^{-x} \le 2 - 2\alpha \end{cases} = \begin{cases} 1 & \text{if } x < -\log(1-\alpha),\\ 0 & \text{if } x \ge -\log(1-\alpha). \end{cases}$$

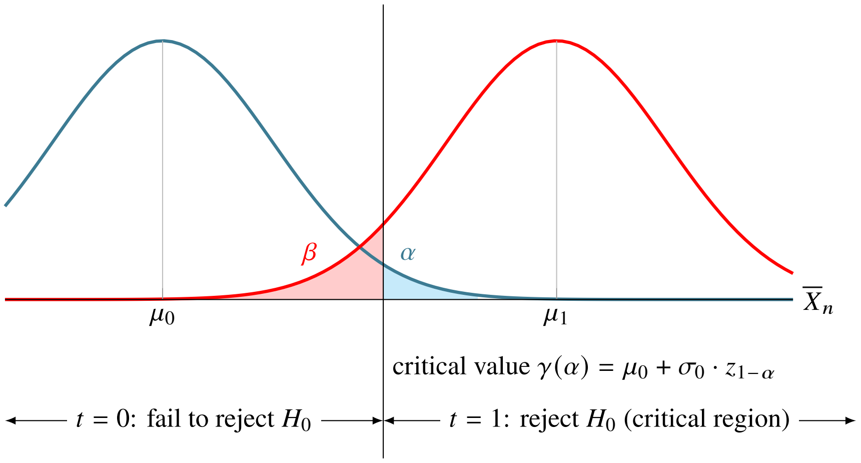


Figure 7.3: Visualization of the upper tailed Z-test (7.13), cf. Example 7.11: type I (α) and type II error (β)

7.4 UPPER TAILED Z-TEST, GAUSS TEST

The most prominent statistical test is probably the Z-test (Gauss test; see Figure 7.3 for interpretation); here $\mathcal{X} = \mathbb{R}^n$.

Theorem 7.11 (Upper tailed Z-test[2]). *Let μ_0 and σ_0 be fixed. For independent, normally distributed X_i consider the test problem*

$$\begin{aligned} H_0 &\colon (X_1, \dots, X_n) \sim \mathcal{N}(\mu_0, \sigma_0^2)^{(n)} \textit{ versus} \\ H_1 &\colon (X_1, \dots, X_n) \sim \mathcal{N}(\mu_1, \sigma_0^2)^{(n)}, \end{aligned} \tag{7.12}$$

where $\mu_0 < \mu_1$. The Neyman–Pearson test is (cf. Table 7.1a)

$$t_\alpha(X) = \begin{cases} 1 & \textit{if } \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sigma_0} > z_{1-\alpha}, \\ 0 & \textit{else} \end{cases} = \begin{cases} 1 & \textit{if } 1 - \Phi\left(\sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sigma_0}\right) < \alpha, \\ 0 & \textit{else}, \end{cases} \tag{7.13}$$

where $P_{\mathcal{N}(0,1)}\left([z_{1-\alpha}, \infty)\right) = \alpha$, i.e., $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the normal distribution. The quantity $Z := \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$ is called Z-score. Note the p-value $\Phi(-Z)$, cf. (7.7).

Proof. Indeed, the likelihood ratio is

$$\begin{aligned} R(x_1, \dots, x_n) = \frac{g(x_1, \dots, x_n)}{f(x_1, \dots, x_n)} &= \frac{\frac{1}{\sqrt{2\pi\sigma_0^2}^n} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_1)^2\right)}{\frac{1}{\sqrt{2\pi\sigma_0^2}^n} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2\right)} \\ &= \exp\left(\frac{2(\mu_1 - \mu_0)}{2\sigma_0^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma_0^2}\left(\mu_1^2 - \mu_0^2\right)\right). \end{aligned} \tag{7.14}$$

[2] Gauß-Test

The distribution of the likelihood ratio $R(X_1, \dots, X_n)$ under $P \sim \mathcal{N}(\mu_0 \mathbb{1}, \sigma_0^2 \mathbb{1})$ is

$$\begin{aligned} F(c) &= P\left(\exp\left(\frac{\mu_1 - \mu_0}{\sigma_0^2} \sum_{i=1}^n X_i - \frac{n}{2\sigma_0^2}\left(\mu_1^2 - \mu_0^2\right)\right) \le c\right) \\ &= P\left(\frac{\mu_1 - \mu_0}{\sigma_0^2} \sum_{i=1}^n X_i \le \log c + \frac{n}{2\sigma_0^2}\left(\mu_1^2 - \mu_0^2\right)\right) \\ &= P\left(\frac{1}{\sigma_0}\overline{X}_n \le \frac{1}{n}\frac{\sigma_0 \log c}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2\sigma_0}\right) \\ &= P\left(\frac{\overline{X}_n - \mu_0}{\sigma_0} \le \frac{1}{n}\frac{\sigma_0 \log c}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2\sigma_0}\right) \\ &= P\left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0} \le \frac{\sigma_0 \log c}{\sqrt{n}(\mu_1 - \mu_0)} + \frac{\sqrt{n}}{2\sigma_0}(\mu_1 - \mu_0)\right) \\ &= \Phi\left(\frac{\sigma_0 \log c}{\sqrt{n}(\mu_1 - \mu_0)} + \frac{\sqrt{n}}{2\sigma_0}(\mu_1 - \mu_0)\right), \end{aligned} \tag{7.15}$$

where we have used that the distribution of the Z-score is $\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu_0}{\sigma_0} \sim \mathcal{N}(0, 1)$ (cf. Exercise 3.4). From (7.3) we get $1 - \alpha = F(c_\alpha)$, i.e.,

$$c_\alpha = \exp\left(\sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0}\Phi^{-1}(1 - \alpha) - \frac{n}{2\sigma_0^2}(\mu_1 - \mu_0)^2\right). \tag{7.16}$$

The test (7.21) thus reads $t_\alpha(x) = \begin{cases} 1 & \text{if (7.14)} > c_\alpha, \\ 0 & \text{else.} \end{cases}$ The conditions simplify further to

$$\frac{2(\mu_1 - \mu_0)}{2\sigma_0^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma_0^2}\left(\mu_1^2 - \mu_0^2\right) > \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma_0}\Phi^{-1}(1 - \alpha) - \frac{n}{2\sigma_0^2}(\mu_1 - \mu_0)^2$$

or equivalently (divide by $\mu_1 - \mu_0$, etc.)

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma_0^2}(\mu_1 + \mu_0) > \sqrt{n}\frac{1}{\sigma_0}\Phi^{-1}(1 - \alpha) - \frac{n}{2\sigma_0^2}(\mu_1 - \mu_0),$$

or

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{x_i - \mu_0}{\sigma_0} > \Phi^{-1}(1 - \alpha).$$

Hence the most powerful test t_α, i.e., (7.13). □

7.5 COMPOSITE HYPOTHESES AND CLASSIFICATION OF TESTS

The Neyman–Pearson lemma (Lemma 7.9) gives rise for the following definition.

Definition 7.12. A binary test $t(\cdot)$ is *most powerful* [3]if there is no other test ψ with

$$\int_X t\,\mathrm{d}P \geq \int_X \psi\,\mathrm{d}P \text{ and}$$
$$\int_X t\,\mathrm{d}Q < \int_X \psi\,\mathrm{d}Q.$$

Remark 7.13. By the Neyman–Pearson lemma (Lemma 7.9) the Neyman–Pearson test t_α (cf. Lemma 7.4) is most powerful.

In what follows we shall consider *composite hypotheses*, that is, Θ consists of more than two points.

Definition 7.14. For a test $t(\cdot)$, the function $\pi_t \colon \vartheta \mapsto \mathbb{E}_\vartheta\, t$ is the *power function* (cf. (6.1))

Remark 7.15. If the test is non-randomized, then $\pi_t(\vartheta) = P_\vartheta(t > 0)$.

Definition 7.16 (Unbiased test)**.** The test $t(\cdot)$ is *unbiased*,[4] if its power $\pi_t(\vartheta) = \mathbb{E}_\vartheta\, t$ satisfies

$$\mathbb{E}_{\vartheta_0}\, t \leq \alpha \text{ for all } \vartheta_0 \in \Theta_0 \text{ and } \alpha \leq \mathbb{E}_{\vartheta_1}\, t \text{ for all } \vartheta_1 \in \Theta_1.$$

Recall that the Gauß-test (Theorem 7.11) does not depend on μ_1 and hence the test is independent of another $\mu_1 > \mu_0$.

Definition 7.17 (UMP and UMPU tests)**.** A test $t(\cdot)$ is a *uniformly most powerful* test (UMP) at significance level α if it is most powerful and in addition, for any other test $\psi(\cdot)$,

$$\mathbb{E}_Q\, \psi \leq \mathbb{E}_Q\, t, \qquad \text{for all } Q \in \mathcal{P}_{\Theta_1}.$$

An unbiased, uniformly most powerful test is UMPU.

7.6 MONOTONE LIKELIHOOD RATIOS

In this subsection we assume that $\Theta \subset \mathbb{R}$.

Definition 7.18 (Monotone likelihood ratio)**.** The class $f_\vartheta(\cdot)$, $\vartheta \in \Theta \subset \mathbb{R}$, is said to possess a *monotone likelihood ratio in the statistic $T(\cdot)$* if

$$\frac{f_{\vartheta_1}(x)}{f_{\vartheta_0}(x)} = R(x) = g_{\vartheta_0,\vartheta_1}\big(T(x)\big),$$

for some function $g_{\vartheta_0,\vartheta_1}(\cdot)$, where all functions $t \mapsto g_{\vartheta_0,\vartheta_1}(t)$ $(\vartheta_0, \vartheta_1 \in \Theta)$ are monotone increasing (decreasing, resp.).

Remark 7.19. The likelihood ratios (7.11) and (7.14) are monotone. Table 7.2 collects further examples.

Lemma 7.20. *Suppose that $f_\vartheta(\cdot)$ has an increasing monotone likelihood ratio in the statistic $T(\cdot)$ and let $h(\cdot)$ be nondecreasing, then $\pi \colon \vartheta \mapsto \mathbb{E}_\vartheta\, h\big(T(X)\big)$ is nondecreasing.*

Proof. Without loss of generality we may assume $h(\cdot) \geq 0$. Let $\vartheta_0 < \vartheta_1$ be chosen. Define

$$A := \left\{x \colon f_{\vartheta_0}(x) > f_{\vartheta_1}(x)\right\}, \quad a := \sup_{x \in A} h\big(T(x)\big)$$

[3] trennschärfster, bester oder mächtigster Test, dt.

[4] unverfälscht

[5] Here, ~ means *up to a constant* not depending on t.

distribution	statistic T	$g(t)$[5]
binomial $bin(n,p)$, $0<p<1$	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$	$\sim \left(\frac{\vartheta_1(1-\vartheta_0)}{\vartheta_0(1-\vartheta_1)}\right)^t$
negative binomial $NB_{n_0,p}$, $0<p<1$	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$	$\sim \left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)^t$
Poisson P_α, $\alpha>0$	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$	$\sim e^{n(\vartheta_1-\vartheta_0)}\left(\frac{\vartheta_1}{\vartheta_0}\right)^t$
exponential E_λ, $\lambda>0$	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$	$\sim e^{t(\vartheta_1-\vartheta_0)}$
normal $\mathcal{N}(\mu,\sigma_0^2)$, $-\infty<\mu<\infty$	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$	$\sim e^{\frac{\vartheta_1-\vartheta_0}{\sigma_0^2}t}$, cf. (7.14)
normal $\mathcal{N}(\mu_0,\sigma^2)$, $\sigma>0$	$\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2$	$\sim e^{\frac{t}{2}\left(\frac{1}{\vartheta_0^2}-\frac{1}{\vartheta_1^2}\right)}$
uniform $U[0,\vartheta)$, $\vartheta>0$	$\max(X_1,\dots,X_n)$	$\sim \mathbb{1}_{[-\infty,\vartheta_0]}(t)$

Table 7.2: Distributions with monotone likelihood ratio

and

$$B := \left\{x\colon f_{\vartheta_0}(x) < f_{\vartheta_1}(x)\right\}, \quad b := \inf_{x\in B} h\big(T(x)\big).$$

For $x\in A$ and $y\in B$ we have that $g_{\vartheta_0,\vartheta_1}\big(T(x)\big) = \frac{f_{\vartheta_1}(x)}{f_{\vartheta_0}(x)} < 1 < \frac{f_{\vartheta_1}(y)}{f_{\vartheta_0}(y)} = g_{\vartheta_0,\vartheta_1}\big(T(y)\big)$ and, as $g_{\vartheta_0,\vartheta_1}(\cdot)$ is increasing thus $T(x)<T(y)$ and $h\left(T(x)\right)\le h\left(T(y)\right)$; it follows that $a\le b$.

Now note that $0 = \int_{\mathcal{X}} f_{\vartheta_1}(x) - f_{\vartheta_0}(x)\,\mathrm{d}x = \left(\int_A + \int_B\right) f_{\vartheta_1}(x) - f_{\vartheta_0}(x)\,\mathrm{d}x$ and thus $\int_B f_{\vartheta_1}(x) - f_{\vartheta_0}(x)\,\mathrm{d}x = -\int_A f_{\vartheta_1}(x) - f_{\vartheta_0}(x)\,\mathrm{d}x$. Hence

$$\begin{aligned}
\pi(\vartheta_1)-\pi(\vartheta_0) &= \int_{\mathcal{X}} h\big(T(x)\big)\left(f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\right)\,\mathrm{d}x\\
&= \int_B h\big(T(x)\big)\underbrace{\left(f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\right)}_{>0 \text{ on } B}\,\mathrm{d}x + \int_A h\big(T(x)\big)\underbrace{\left(f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\right)}_{<0 \text{ on } A}\,\mathrm{d}x\\
&\ge b\cdot\int_B f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\,\mathrm{d}x + a\cdot\int_A f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\,\mathrm{d}x\\
&= (b-a)\cdot\int_B f_{\vartheta_1}(x)-f_{\vartheta_0}(x)\,\mathrm{d}x \ge 0,
\end{aligned}$$

the assertion □

Theorem 7.21 (Karlin–Rubin theorem). *Suppose that $f_\vartheta(\cdot)$ has an increasing monotone likelihood ratio for the statistic T. Let α and c_α be chosen so that the test t has level α at ϑ_0, i.e., $\pi_t(\vartheta_0) = P_{\vartheta_0}\left(T(\cdot)\ge c_\alpha\right) = \alpha$. Then $C := \{x\colon T(x)\ge c_\alpha\}$ is the critical region for a uniformly most powerful test at level α for the one-sided, composite hypotheses*

$$\begin{aligned}
H_0&\colon \vartheta\le\vartheta_0 \text{ versus}\\
H_1&\colon \vartheta>\vartheta_0.
\end{aligned}$$

Further, the power $\pi_t\colon \vartheta\mapsto \mathbb{E}_\vartheta\, t$ is nondecreasing.

Proof. The case $\vartheta<\vartheta_0$: Set $h(\cdot) := \mathbb{1}_{(c_\alpha,\infty)}(\cdot)$. The power function $\pi_t\colon \vartheta\mapsto\mathbb{E}_\vartheta\, t = P_\vartheta\left(T>c_\alpha\right) = \mathbb{E}\,h(T)$ is nondecreasing by the previous lemma, as $h(\cdot)$ is nondecreasing. Thus $\alpha = \pi_t(\vartheta_0) = \sup\left\{\pi_t(\vartheta)\colon \vartheta\le\vartheta_0\right\}$ and $t(\cdot)$ is a test at level α.

Conversely, pick $\vartheta_1 > \vartheta_0$ and consider the simple hypothesis $H_0\colon \vartheta = \vartheta_0$ versus $H_1 : \vartheta = \vartheta_1$. By the Neyman–Pearson lemma (Lemma 7.9), the best choice for the critical region is to choose k_α so that $C = \left\{x\colon \frac{f_{\vartheta_1}(x)}{f_{\vartheta_0}(x)} \geq k_\alpha\right\}$. Because $f_\vartheta(\cdot)$ has in increasing likelihood ratio for the statistics $T(\cdot)$, this is equivalent to $C = \{x\colon T(x) \geq c_\alpha\}$ for some c_α. □

7.7 LIKELIHOOD RATIO TEST

Definition 7.22. The *likelihood ratio test statistic* is $\Lambda(x) := {}^1\!/_{R(x)}$, where $R(x) := \frac{\sup_{\vartheta\in\Theta_0\cup\Theta_1} f_\vartheta(x)}{\sup_{\vartheta\in\Theta_0} f_\vartheta(x)}$. The test is

$$t(x) = \begin{cases} 0 & \text{if } \Lambda(x) > c,\\ 1 & \text{if } \Lambda(x) < c,\end{cases}$$

where c is appropriate.

Remark 7.23. It holds that $0 \leq \Lambda(x) \leq 1$.

Remark 7.24. It holds that the functions $\frac{f_{\vartheta_0}(x)}{\max\{f_\vartheta(x)\colon \vartheta\in\{\vartheta_0\cup\vartheta_1\}\}} = \min\left\{1, \frac{f_{\vartheta_0}(x)}{f_{\vartheta_1}(x)}\right\}$ are monotone functions of each other and thus equivalent for present purposes.

Example 7.25 (Student's t-test)**.** Consider the family $\mathcal{N}(\mu, \sigma^2)$ with $(\mu, \sigma^2) \in \mathbb{R}\times\mathbb{R}_{>0}$. The UMPU test for the problem

$$\begin{aligned} &H_0\colon \mu = \mu_0 \text{ versus}\\ &H_1 : \mu \neq \mu_0\end{aligned}$$

(μ_0 known) is

$$t(X) := \begin{cases} 1 & \text{if } \sqrt{n}\left|\frac{\overline{X}_n-\mu_0}{s_n}\right| > t_{n-1,1-\frac{\alpha}{2}},\\ 0 & \text{else,}\end{cases} \tag{7.17}$$

where $t_{n-1,1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of the Student t_{n-1} distribution with $n-1$ degrees of freedom, $P_{t_{n-1}}\left(\left[-t_{n-1,1-\frac{\alpha}{2}},\ t_{n-1,1-\frac{\alpha}{2}}\right]\right) = \alpha$.

Remark 7.26. Compare the Gauß test (7.13) and the Student test (7.17).

Proof. The regions for (μ, σ^2) are $\Theta_0 = \{\mu_0\}\times\mathbb{R}_{>0}$ and $\Theta := \mathbb{R}\times\mathbb{R}_{>0}$. With $\vartheta = (\mu, \sigma^2) \in \Theta$ we have that $f_\vartheta(x) = \frac{1}{\sqrt{2\pi\sigma^2}^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)$ and thus

$$\sup_{\vartheta\in\Theta_0} f_\vartheta(x) = \frac{1}{\sqrt{2\pi}^n}\left(\frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2\right)^{-n/2}\cdot e^{-n/2}$$

(the maximum is attained at $\sigma^{*2} = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2$) and

$$\sup_{\vartheta\in\Theta} f_\vartheta(x) = \frac{1}{\sqrt{2\pi}^n}\left(\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x}_n)^2\right)^{-n/2}\cdot e^{-n/2},$$

(the maximum at $\mu^* = \overline{x}_n$ and $\sigma^{*2} = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x}_n)^2$. Hence

$$R(x) = \left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2}{\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x}_n)^2}\right)^{n/2}.$$

Now recall from Remark 2.5 (Steiner) that $\frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x}_n)^2 + (\overline{x}_n-\mu_0)^2$ and thus

$$R(x) = \left(1+\frac{1}{n-1}\cdot n\frac{(\overline{x}_n-\mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x}_n)^2}\right)^{n/2},$$

a monotone function in $\sqrt{n}\left|\frac{\overline{x}_n-\mu_0}{s_n}\right|$. By the Karlin–Rubin theorem (Theorem 7.21), the optimal test is a Neyman–Pearson test. Its critical value follows with Theorem 5.17. □

Example 7.27 (Student's t-test)**.** Consider the family $\mathcal{N}(\mu,\sigma^2)$ with $(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_{>0}$. The UMPU test for the problem

$$\begin{aligned} H_0\colon\ &\mu\le\mu_0 \text{ versus}\\ H_1\colon\ &\mu>\mu_0\end{aligned}$$

(μ_0 known) is

$$t(X) := \begin{cases}1 & \text{if } \sqrt{n}\frac{\overline{X}_n-\mu_0}{s_n} > t_{n-1,1-\alpha},\\ 0 & \text{if } \sqrt{n}\frac{\overline{X}_n-\mu_0}{s_n}\le t_{n-1,1-\alpha}\end{cases} = \begin{cases}1 & \text{if } \overline{X}_n > \mu_0+\frac{s_n}{\sqrt{n}}t_{n-1,1-\alpha},\\ 0 & \text{if } \overline{X}_n\le \mu_0+\frac{s_n}{\sqrt{n}}t_{n-1,1-\alpha},\end{cases} \tag{7.18}$$

where $t_{n-1,1-\alpha}$ is the $1-\alpha$-quantile of the Student t_{n-1} distribution with $n-1$ degrees of freedom, $P_{t_{n-1}}\big([t_{n-1,1-\alpha},\infty)\big)=\alpha$.

7.8 THE LIKELIHOOD RATIO FOR THE ALTERNATIVE

In what follows we shall assume that $P(F)=0$ and $Q(G)=0$. The likelihood ratio $R(x)=\frac{g(x)}{f(x)}$ under P is F_R (cf. (7.2)), but here we investigate the ratio under Q,

$$G(u) := Q\,(R\le u) = \int_{\left\{\frac{g(\cdot)}{f(\cdot)}\le u\right\}} g(x)\,\mathrm{d}x = \int_{\left\{\frac{g(\cdot)}{f(\cdot)}\le u\right\}} \frac{g(x)}{f(x)} f(x)\,\mathrm{d}x. \tag{7.19}$$

Lemma 7.28. *It holds that*

$$Q(R\le u)\le P(R\le u),\qquad u\in\mathbb{R},$$

i.e., R is smaller under *P than under Q: first–order stochastic dominance (FSD).*

Proof. Note that $P(R\in \mathrm{d}c)=\mathrm{d}F_R(c)$ by definition of $F_R(\cdot)$, thus

$$G(u) := \int_{\{R\le u\}} R(x)f(x)\,\mathrm{d}x = \mathbb{E}_P\,\mathbb{1}_{\{R\le u\}}\,R = \int_0^u c\,P(R\in\mathrm{d}c) = \int_0^u c\,\mathrm{d}F_R(c).$$

Now note that

$$u\mapsto \mathbb{E}\,(R\mid R\le u) = \frac{\int_0^u c\,\mathrm{d}F(c)}{\int_0^u \mathrm{d}F(c)} \tag{7.20}$$

is nondecreasing in u (Exercise 7.7). It follows that

$$\frac{G(u)}{F_R(u)} = \frac{\int_0^u c\,\mathrm{d}F_R(c)}{\int_0^u \mathrm{d}F_R(c)} \le \frac{\int_0^\infty c\,\mathrm{d}F_R(c)}{\int_0^\infty \mathrm{d}F_R(c)} = \frac{G(\infty)}{F_R(\infty)} = \frac{1}{1} = 1$$

and thus

$$Q(R\le u) = G(u)\le F_R(u) = P(R\le u),$$

i.e, the statistics R is *stochastically smaller* under P than under Q. □

Definition 7.29. The probability of a type II error of the Neyman–Pearson test (cf. (7.19)) is

$$Q\big(R \le c_\alpha\big) = G\big(c_\alpha\big), \tag{7.21}$$

its power is $1 - G\big(c_\alpha\big)$, cf. (6.1).

Definition 7.30. The function

$$h(\alpha) := \sup\left\{\int_X \psi \,\mathrm{d}Q \colon \int_X \psi \,\mathrm{d}P \le \alpha\right\}$$

is the *type II error function*.

Remark 7.31. The most powerful test is t_α. If $P \approx Q$, then it holds that

$$h(\alpha) = 1 - \int_X 1 - t_\alpha \,\mathrm{d}Q = 1 - G(c_\alpha) = 1 - G\left(F_R^{-1}(1-\alpha)\right).$$

Proposition 7.32. *The type ii error function $h(\alpha)$ for the Gauß test (7.13) is $h(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{n}\frac{\mu_1-\mu_0}{\sigma}\right)$.*

Proof. Observe from (7.15) that

$$\begin{aligned} G(u) &= Q\left(\frac{1}{\sigma_0}\sum_{i=1}^n X_i \le \frac{\sigma_0 \log u}{\mu_1-\mu_0} + \frac{n}{2\sigma_0}\left(\mu_1+\mu_0\right)\right) \\ &= Q\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i-\mu_1}{\sigma_0} \le \frac{1}{\sqrt{n}}\frac{\sigma_0 \log u}{\mu_1-\mu_0} - \frac{\sqrt{n}}{2\sigma_0}\left(\mu_1-\mu_0\right)\right) \\ &= \Phi\left(\frac{1}{\sqrt{n}}\frac{\sigma_0 \log u}{\mu_1-\mu_0} - \frac{\sqrt{n}}{2\sigma_0}\left(\mu_1-\mu_0\right)\right), \end{aligned}$$

thus, cf. (7.16),

$$\begin{aligned} h(\alpha) &= 1 - G\left(F^{-1}(1-\alpha)\right) = 1 - G\left(c_\alpha\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{n}\frac{\mu_1-\mu_0}{\sigma_0}\right), \end{aligned}$$

the assertion. □

7.9 PROBLEMS

Exercise 7.1 (Continuation of Example 7.10)**.** *The likelihood under Q is $G(u) = \int_0^u c\,\mathrm{d}F_R(c) = \int_0^u c\,\mathrm{d}\frac{c}{2} = \frac{c^2}{4}\Big|_{c=0}^u = \frac{u^2}{4}$, thus*

$$h(\alpha) = 1 - G\left(F_R^{-1}(1-\alpha)\right) = 1 - G\left(c_\alpha\right) = 1 - G\left(2-2\alpha\right) = 2\alpha - \alpha^2.$$

Exercise 7.2 (Cf. Example 7.10)**.** *Give the Neyman–Pearson test t_α for the exponential distributions*

$$\begin{aligned} &H_0\colon P \sim E_1^{(n)} \text{ versus} \\ &H_1\colon Q \sim E_\lambda^{(n)}, \end{aligned}$$

where $\lambda < 1$. Express the result in terms of the incomplete Gamma function.

Exercise 7.3. *Find the best test for the problem*

$$\begin{aligned} H_0&\colon (X_1,\dots,X_n)\sim\mathcal{N}(\mu_0,\sigma_0^2)^{(n)} \text{ versus}\\ H_1&\colon (X_1,\dots,X_n)\sim\mathcal{N}(\mu_0,\sigma_1^2)^{(n)}, \end{aligned}$$

where $\mu_0\in\mathbb{R}$ is known and $\sigma_0<\sigma_1$.

Exercise 7.4. *Show that the most powerful test for the problem*

$$\begin{aligned} H_0&\colon \sigma^2\le\sigma_0^2 \text{ versus}\\ H_1&\colon \sigma^2>\sigma_0^2 \end{aligned}$$

for the family $\mathcal{N}\left(\mu_0,\sigma^2\right)$ (μ_0 known) has the critical region $C=\left\{x\colon \sum_{i=1}^n(\frac{x_i-\mu_0}{\sigma_0})^2\ge c_\alpha\right\}$, where c_α is the α-quantile of the χ_n^2 distribution, i.e., $P_{\chi^2_{n,\alpha}}([c_\alpha,\infty))=\alpha$ (cf. Table 7.2).

Exercise 7.5. *Verify the functions $g(\cdot)$ for the distributions in Table 7.2.*

Exercise 7.6. *Show that the family of Cauchy random variables with density $f_\vartheta(x)=\frac{1}{\pi}\frac{1}{1+(x-\vartheta)^2}$ does* not *possess a monotone likelihood ratio.*

Exercise 7.7. *Show that (7.20) is nondecreasing.*

8 Bayes' tests

Non-Bayesians have more fun.

Thomas Mikosch' maxim

Recall from (6.4) that we are interested in the test $t(\cdot)$ minimizing

$$(1-\lambda)\cdot\int_{\mathcal{X}} t(x)\,P(\mathrm{d}x)+\lambda\cdot\int_{\mathcal{X}} 1-t(x)\,Q(\mathrm{d}x)\longrightarrow \min! \tag{8.1}$$

Proposition 8.1. *The Neyman–Pearson test*

$$t(x):=\begin{cases}1 & \text{if } \frac{g(x)}{f(x)}>\frac{1-\lambda}{\lambda},\\ 0 & \text{if } \frac{g(x)}{f(x)}\le\frac{1-\lambda}{\lambda}\end{cases} \tag{8.2}$$

is the Bayes' test minimizing (8.1).

Proof. It follows from Lemma 7.9 (Neyman–Pearson lemma) that it is enough to consider Neyman–Pearson tests. Let

$$\alpha:=1-F\left(\frac{1-\lambda}{\lambda}\right) \tag{8.3}$$

so that $c_\alpha=\frac{1-\lambda}{\lambda}$ and t_α be the associated Neyman–Pearson test. By (7.8),

$$\frac{1-\lambda}{\lambda}\left(\int_{\mathcal{X}} t_\alpha\,\mathrm{d}P-\int_{\mathcal{X}}\psi\,\mathrm{d}P\right)\le\int_{\mathcal{X}} t_\alpha\,\mathrm{d}Q-\int_{\mathcal{X}}\psi\,\mathrm{d}Q,$$

where ψ is any other test. Hence

$$(1-\lambda)\int_{\mathcal{X}} t_\alpha\,\mathrm{d}P-\lambda\int_{\mathcal{X}} t_\alpha\,\mathrm{d}Q\le(1-\lambda)\int_{\mathcal{X}}\psi\,\mathrm{d}P-\lambda\int_{\mathcal{X}}\psi\,\mathrm{d}Q$$

and thus

$$(1-\lambda)\int_{\mathcal{X}} t_\alpha\,\mathrm{d}P+\lambda\int_{\mathcal{X}} 1-t_\alpha\,\mathrm{d}Q\le(1-\lambda)\int_{\mathcal{X}}\psi\,\mathrm{d}P+\lambda\int_{\mathcal{X}} 1-\psi\,\mathrm{d}Q,$$

the assertion. □

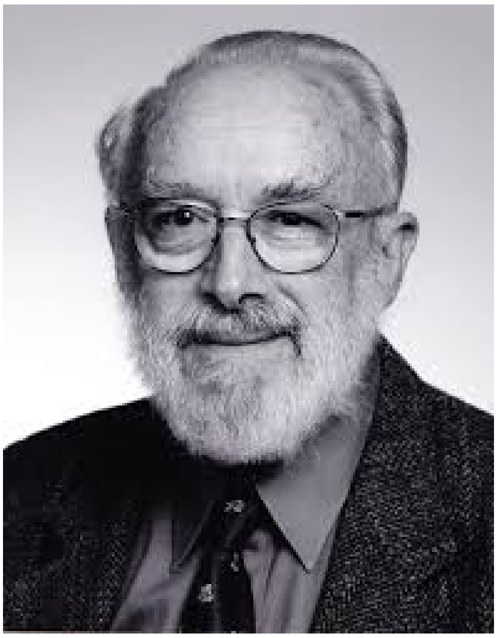
Remark 8.2. It follows from (8.3) that $c_\alpha=F^{-1}(1-\alpha)\ge\frac{1-\lambda}{\lambda}$ and thus Bayes tests are always non-randomized and conservative.

Definition 8.3. The *average error probability* of a Bayes test is

$$k(\lambda):=\inf_{t(\cdot)\in[0,1]}(1-\lambda)\int_{\mathcal{X}} t(x)P(\mathrm{d}x)+\lambda\int_{\mathcal{X}}(1-t(x))Q(\mathrm{d}x).$$

In view of Proposition 8.1 we have $k(\lambda)=(1-\lambda)\left(1-F\left(\frac{1-\lambda}{\lambda}\right)\right)+\lambda\,G\left(\frac{1-\lambda}{\lambda}\right)$.

(a) Ronald Aylmer Fisher, 1890–1962



(b) Dennis Lindley, 1923–2013; interesting obituary

Figure 8.1: Frequentist (Ronald Fisher, Egon Pearson, Jerzy Neyman) versus Bayesian (Jimmie Savage, Bruno de Finetti ('probability does not exist'), Jack Good, Harold Jeffreys, Robert Schlaifer) approaches in statistics

Remark 8.4. The average error probability of problem (7.12) (Example 7.11) is

$$\begin{aligned}k(\lambda) =&(1-\lambda)\left(1-\Phi\left(\frac{1}{\sqrt{n}}\frac{\sigma}{\mu_1-\mu_0}\log\frac{1-\lambda}{\lambda}+\sqrt{n}\frac{\mu_1-\mu_0}{\sigma}\right)\right)\\ &+\lambda\Phi\left(\frac{1}{\sqrt{n}}\frac{\sigma}{\mu_1-\mu_0}\log\frac{1-\lambda}{\lambda}-\sqrt{n}\frac{\mu_1-\mu_0}{\sigma}\right).\end{aligned}$$

MAXIMUM A POSTERIORI INTERPRETATION

The equation (8.1) suggests to consider the combined measure

$$\pi := (1-\lambda)P + \lambda Q$$

for $\lambda \in [0,1]$ fixed. The distributions $H_0\colon P \sim f$ and $H_1\colon Q \sim g$ are called *sampling distributions* or *likelihood*.

The Bayesian setting assumes/ pretends to know a *prior distribution*[1]

$$\pi(H_1) = \lambda, \text{ i.e., } \pi(H_0) = 1-\lambda,$$

selecting the model itself, independently of the data. The density of the measure π is

$$h(x) = (1-\lambda)\cdot \underbrace{f(x)}_{h(x|H_0)} + \lambda\cdot \underbrace{g(x)}_{h(x|H_1)},$$

called *marginal likelihood*, sometimes also termed the *model evidence*. In this setting we have that $P(A) = \int_A h(x \mid H_0)\,\mathrm{d}x$ and $Q(B) = \int_B h(x \mid H_1)\,\mathrm{d}x$.

Let X have density $h(\cdot)$. From Bayes' theorem, the *posterior probabilities*[2] are

$$\pi(H_0 \mid X = x) = \frac{h(x \mid H_0)\cdot\pi(H_0)}{h(x)} = \frac{f(x)\cdot(1-\lambda)}{h(x)}$$

and

$$\pi(H_1 \mid X = x) = \frac{h(x \mid H_1)\cdot\pi(H_1)}{h(x)} = \frac{g(x)\cdot\lambda}{h(x)}.$$

We accept H_0 (i.e., $t(x) = 0$) by comparing the likelihood of the *posterior distribution* (i.e., after observing the data x), iff

$$\pi(H_0 \mid X = x) \ge \pi(H_1 \mid X = x),$$

i.e.,

$$\frac{f(x)\cdot(1-\lambda)}{h(x)} \ge \frac{g(x)\cdot\lambda}{h(x)}.$$

This is precisely the test (8.2),

$$t(x) := \begin{cases} 1 & \text{if } \frac{g(x)}{f(x)} > \frac{1-\lambda}{\lambda},\\ 0 & \text{if } \frac{g(x)}{f(x)} \le \frac{1-\lambda}{\lambda},\end{cases}$$

which is also called the *maximum a posteriori (MAP) test*.

8.1 PROBLEMS

Exercise 8.1. *Verify Remark 8.2.*

[1] a priori (lat.): from the earlier, i.e., *before* knowing the data

[2] a posteriori (lat.): from the later, i.e., *after* having seen the data

Selected tests 9

> If I have seen further it is by standing on the shoulders of giants.
>
> Isaac Newton, 1642–1726

9.1 Fisher's exact test of independence

The test is useful for categorical data that result from classifying objects in two different ways. Fisher's exact test is *exact* because it guarantees an α rate regardless of the sample size. Fisher devised the test in the *lady tasting tea* experiment.

Problem 9.1. Based on Table 9.1a, is it fair to say that it is equally probable for men and woman to study? The test setting addresses the hypothesis $H_0\colon p_{\text{men}} = p_{\text{women}}$ versus $H_1\colon p_{\text{men}} \neq p_{\text{women}}$.

	men	woman	row total
studying	4	6	10
not studying	9	5	14
column total	13	11	24

(a) Data

	A	not A	row total
B	x	y	$z = x + y$
not B	$n_1 - x$	$n_2 - y$	$n - z$
column total	n_1	n_2	$n := n_1 + n_2$

(b) Data, schematic

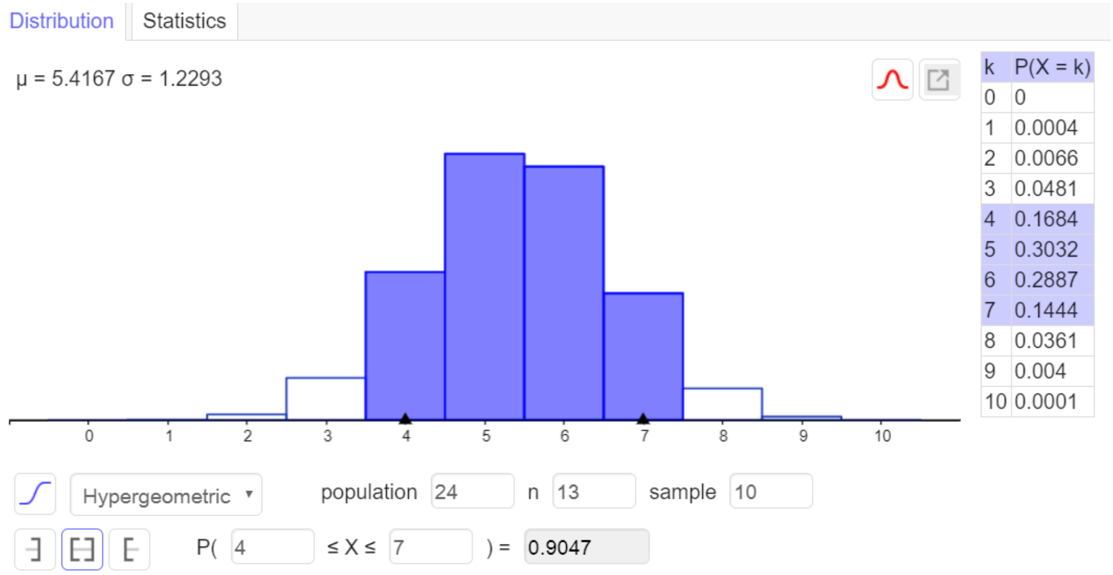
Table 9.1: Contingency table

Remark 9.2. Note that the contingency table is fully determined by the marginals n_1, n_2, z of the table, and a single entry of the table, for example x.

Lemma 9.3. *Suppose that $X \sim bin(n_1, p)$ and $Y \sim bin(n_2, p)$ are independent, then*

$$P(X = x \mid X + Y = z) = \frac{\binom{n_1}{x} \cdot \binom{n_2}{z-x}}{\binom{n_1+n_2}{z}}$$

follows a hypergeometric distribution, which is not dependent *on p(!). The parameters are the population size $n_1 + n_2$, the number of success states in the population n_1 and the draws z.*



μ = 5.4167 σ = 1.2293

k	P(X = k)
0	0
1	0.0004
2	0.0066
3	0.0481
4	0.1684
5	0.3032
6	0.2887
7	0.1444
8	0.0361
9	0.004
10	0.0001

Hypergeometric population 24 n 13 sample 10

P(4 ≤ X ≤ 7) = 0.9047

Figure 9.1: Screenshot from the freeware GeoGebra: the claim H_0 in Problem 9.1 can*not* be rejected at $\alpha = 10\,\%$.

Proof. By independence, $X + Y \sim \operatorname{bin}(n_1 + n_2, p)$. It holds that

$$\begin{aligned}
P(X = x \mid X + Y = z) &= \frac{P(X = x,\ X + Y = z)}{P(X + Y = z)} \\
&= \frac{P(X = x,\ Y = z - x)}{P(X + Y = z)} \\
&= \frac{P(X = x) \cdot P(Y = z - x)}{P(X + Y = z)} \\
&= \frac{\binom{n_1}{x} p^x (1-p)^{n_1 - x} \cdot \binom{n_2}{z-x} p^{z-x} (1-p)^{n_2 - z + x}}{\binom{n_1+n_2}{z} p^z (1-p)^{n_1+n_2-z}} \\
&= \frac{\binom{n_1}{x} \cdot \binom{n_2}{z-x}}{\binom{n_1+n_2}{z}},
\end{aligned}$$

the assertion. □

Figure 9.1 clarifies that the Hypothesis H_0 in Problem 9.1 (perhaps surprisingly) can*not* be rejected at $\alpha = 10\,\%$. That is, it is equally likely for men and women to study. Note further, that the test does *not* involve any estimate for p_{men} or p_{women}.

9.2 GOODNESS OF FIT

9.2.1 Pearson's chi-squared test

Example 9.4. During a visit in a casino, the pattern displayed in Table 9.2 has been observed on a roulette table. Is the table biased? (See the table https://en.wikipedia.org/wiki/Chi-squared_distribution)

	observations O_i	p_i	E_i	Pearson T
red	18	18/37	26.76	2.87
black	35	18/37	26.76	2.54
zero	2	1/37	1.49	0.18
total	55	1	55	$T = 5.58$

Table 9.2: Is this roulette table biased or fair?

Proposition 9.5. *Suppose that the observed counts $(O_1, \dots, O_k)$ follow a multinomial distribution correspond to the expected counts $(E_1, \dots, E_k)$, where $n \coloneqq \sum_{i=1}^k O_i = \sum_{i=1}^k E_i$. Then the Pearson statistics*

$$T \coloneqq \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} \tag{9.1}$$

follows asymptotically, for $n \to \infty$, a χ^2 distribution with $k-1$ degrees of freedom.

Remark 9.6 (Bayesian method). In many situations, $E_i = p_i \cdot n$, with $\sum_{i=1}^k p_i = 1$.

Proof. We illustrate the proof for $k = 2$ first. In this case, $O \coloneqq O_1 \sim \operatorname{bin}(n, p)$ for some $p \in (0,1)$, $O_2 = n - O_1$, $E_1 = \mathbb{E}\, O_1 = np$, $E_2 = \mathbb{E}\, O_2 = n - np$ and $\operatorname{var} O = np(1-p)$. The test statistics (9.1) is

$$\begin{aligned} T &= \frac{(O-np)^2}{np} + \frac{\big(n - O - (n-np)\big)^2}{n-np} \\ &= \frac{(O-np)^2}{np} + \frac{(O-np)^2}{n(1-p)} \\ &= \frac{(O-np)^2}{np(1-p)}. \end{aligned}$$

Now note that $Z \coloneqq \frac{O-np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1)$ by the central limit theorem (Theorem 4.3) so the assertion follows with $T = Z^2 \sim \chi^2_1$ by Proposition 5.12.

The proof for $k > 2$ is technically more involved, but we can proceed as above. From Exercise 1.15 recall the covariance matrix

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & & \vdots \\ \vdots & & \ddots & -p_{k-1}p_k \\ -p_1p_k & \dots & -p_{k-1}p_k & p_k(1-p_k) \end{pmatrix} = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & p_k \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix}^\top.$$

The matrix Σ is singular, as every column (row, resp.) sums to 0. The truncated matrix

$$\Sigma^* = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ddots & 0 & p_{k-1} \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{k-1} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{k-1} \end{pmatrix}^\top$$

is regular with explicit inverse $\Sigma^{*-1} = \begin{pmatrix} \frac{1}{p_1} & 0 & 0 \\ 0 & \ddots & \ddots \\ 0 & \ddots & \frac{1}{p_{k-1}} \end{pmatrix} + \frac{1}{p_k} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$. With $\sum_{i=1}^k O_i = n = n \cdot \sum_{i=1}^k p_i$, the Pearson statistics (9.1) is

$$\begin{aligned} T &= \sum_{i=1}^{k} \frac{(O_i - n\,p_i)^2}{n\,p_i} \\ &= \sum_{i=1}^{k-1} \frac{(O_i - n\,p_i)^2}{n\,p_i} + \frac{(O_k - n\,p_k)^2}{n\,p_k} \\ &= \sum_{i=1}^{k-1} \frac{(O_i - n\,p_i)^2}{n\,p_i} + \frac{\left(\sum_{i=1}^{k-1}(O_i - np_i)\right)^2}{n\,p_k} \\ &= \frac{1}{n}(O^* - np^*)\Sigma^{*-1}(O^* - np^*), \end{aligned}$$

where $O^* := (O_1, \dots, O_{k-1})$ and $p^* := (p_1, \dots, p_{k-1})$. By the central limit theorem, $Z := \frac{1}{\sqrt{n}}\Sigma^{*-1/2}(O^* - np^*) \xrightarrow{d} \mathcal{N}(0, I_{k-1})$. Hence $T = Z^\top Z$ converges to the sum of $k-1$ independent squared normals, that is $T \sim \chi^2_{k-1}$, the assertion. □

9.2.2 G-test

The G-test employs the statistics $G := 2\sum_{i=1}^k O_i \ln \frac{O_i}{E_i}$ instead of T (cf. (9.1)), but is considered to be more robust. We show that

$$G \sim \chi^2_{k-1}$$

asymptotically, for $n = \sum_{i=1}^k O_i = \sum_{i=1}^k E_i \to \infty$.

Indeed, for $n \to \infty$ we have that $\left|\frac{O_i - E_i}{E_i}\right| \ll 1$ (small) for $i = 1, \dots, k$. With $\ln(1+x) = x - \frac{1}{2}x^2 + \mathcal{O}(x^3)$,

$$\begin{aligned} G = 2\sum_{i=1}^{k} O_i \ln \frac{O_i}{E_i} &= 2\sum_{i=1}^{k} (E_i + O_i - E_i) \ln\left(1 + \frac{O_i - E_i}{E_i}\right) \\ &\approx 2\sum_{i=1}^{k} \left(E_i + (O_i - E_i)\right)\left(\frac{O_i - E_i}{E_i} - \frac{1}{2}\left(\frac{O_i - E_i}{E_i}\right)^2\right) \\ &\approx 2\sum_{i=1}^{k} O_i - E_i + \frac{(O_i - E_i)^2}{E_i} - \frac{1}{2}\frac{(O_i - E_i)^2}{E_i} \\ &= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \end{aligned}$$

which is Pearson's chi-squared test statistics T. It follows that

$$G \approx T \sim \chi^2_{k-1}$$

with (9.1) above.

9.3 CONFIDENCE INTERVALS

Definition 9.7. Let $t\colon \mathcal{X} \to [0,1]$ be a test. The confidence[1] interval $C(X)$ satisfies

$$P_\vartheta\left(\{x \in \mathcal{X}\colon \vartheta \in C(X)\}\right) \geq 1-\alpha \qquad (\vartheta \in \Theta).$$

Remark 9.8. For some fixed $\vartheta_0 \in \Theta$ it holds that

$$P_{\vartheta_0}\left(\{x \in \mathcal{X}\colon \vartheta_0 \notin C(X)\}\right) \leq \alpha$$

and

$$P_\vartheta\left(\{x \in \mathcal{X}\colon \vartheta \in C(X)\}\right) \geq 1-\alpha \qquad (\vartheta \in \Theta\backslash\{\vartheta_0\}).$$

So

$$t_C(X) \coloneqq \mathbb{1}_{\{x\in\mathcal{X}\colon\, \vartheta_0\notin C(X)\}}(X)$$

is a test for the problem

$$\begin{aligned} &H_0\colon \vartheta = \vartheta_0,\\ &H_1\colon \vartheta \in \Theta\backslash\{\vartheta_0\}. \end{aligned}$$

Example 9.9. The confidence interval in Example 7.11 (cf. Figure 7.3) is $C(X) = \left[\overline{X}_n - \frac{\sigma_0}{\sqrt{n}} \cdot z_{1-\alpha}, \infty\right)$. Indeed, by (7.13),

$$\mu_0 \notin C(X) \iff \mu_0 < \overline{X}_n - \frac{\sigma_0}{\sqrt{n}} \cdot z_{1-\alpha} \iff \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0} > z_{1-\alpha};$$

here, z_α is the α-quantile of the standard normal distribution,

$$z_\alpha = \Phi^{-1}(\alpha),\ \text{or}\ \Phi(z_\alpha) = \alpha,\ \text{i.e.,}\ \alpha = \int_\infty^{z_\alpha} \varphi(u)\,\mathrm{d}u.$$

Example 9.10 (Cf. Weiß [21])**.** A company produces balls which are supposed to have diameter μ_0. The variance σ_0 of their diameter is known, due to observations over years. During a working shift, the diameter of randomly picked balls is $X_1, \dots, X_n$. Do they deviate significantly from μ_0?

We choose $\mathcal{P} \coloneqq \left\{\mathcal{N}(\mu, \sigma_0^2)^{(n)}\colon \mu \in \mathbb{R}\right\}$ and $H_0\colon P \sim \mathcal{N}(\mu_0, \sigma_0^2)^{(n)}$. Apparently,

$$t(X_1, \dots X_n) \coloneqq \begin{cases} 1 & \text{if } \left|\overline{X}_n - \mu_0\right| \geq c,\\ 0 & \text{else} \end{cases}$$

is a reasonable test, where we still need to determine c. But

$$\begin{aligned} \alpha = P(t=1) &= P\left(\left|\overline{X}_n - \mu_0\right| \geq c \mid \mathcal{N}(\mu_0, \sigma_0^2)\right) = P\left(\left|\overline{X}_n - \mu_0\right| \geq c \mid \mathcal{N}(\mu_0, \sigma_0^2)\right)\\ &= P\left(\left[-\frac{c}{\sigma_0^2}\sqrt{n}, \frac{c}{\sigma_0^2}\sqrt{n}\right] \mid \mathcal{N}(0,1)\right), \end{aligned}$$

so that $c = \frac{\sigma_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$.

[1] confiteor, lat., deutlich zeigen, offenbaren, beichten

Example 9.11 (Cf. Weiß [21]). A person is said to have supernatural skills. So, she is asked to predict a sequence of randomly chosen numbers from $\{1,2,3,4\}$. The person correctly predicts 42 out of 115 numbers. Does the person have supernatural skills, provided the statistician does not want to make a fool of himself with probability 99%?

We choose Bernoulli random variables and the model $\mathcal{P} := \left\{B(1,p)^{(n)} : p \ge \frac{1}{4}\right\}$ and $H_0 : B(1, 1/4)^{(n)}$. For this we consider the test

$$t(X_1, \dots X_n) := \begin{cases} 1 & \text{if } \overline{X}_n - \frac{1}{4} \ge c, \\ 0 & \text{else.} \end{cases}$$

To achieve significance α,

$$\begin{aligned} \alpha = P(t=1) &= P\left(\overline{X}_n - \frac{1}{4} \ge c \mid B(1, 1/4)^{(n)}\right) = P\left(X_1 + \dots X_n \ge n\left(c + \frac{1}{4}\right) \mid B(1, 1/4)^{(n)}\right) \\ &= P\left(\left[n\left(c+\frac{1}{4}\right), \infty\right) \mid B(n, 1/4)\right) \sim P\left(\left[n\left(c+\frac{1}{4}\right), \infty\right) \mid \mathcal{N}(\frac{n}{4}, \frac{3n}{16})\right) \\ &= P\left(\left[\frac{4c}{\sqrt{3}}\sqrt{n}, \infty\right) \mid \mathcal{N}(0,1)\right), \end{aligned}$$

so that $c = \frac{\sqrt{3}}{4\sqrt{n}} z_{1-\alpha}$. But $\overline{X}_n - \frac{1}{4} = \frac{42}{115} - \frac{1}{4} = 0.115 \ge 0.0939 = \frac{\sqrt{3}}{4\sqrt{115}} z_{99\%}$, so that we have to reject H_0 and the person has supernatural skills, indeed.

Example 9.12 (Cf. Weiß [21]). A worker produces 600 items in a working shift. On average, 2.8 are faulty. To supervise the quality the number of faulty items is recorded for every worker. How can one check if a particular worker produces significantly more ($\alpha = 0.05$) faulty items than $\lambda_0 = 2.8$?

To model the situation we choose Poisson random variables (cf. (12.3) below) and the model $\mathcal{P} := \left\{P_\lambda^{(n)} : \lambda \ge \lambda_0\right\}$ the hypothesis $H_0 : P(1, \lambda_0)^{(n)}$ and the test

$$t(X_1, \dots X_n) := \begin{cases} 1 & \text{if } \overline{X}_n - \lambda_0 \ge c, \\ 0 & \text{else.} \end{cases}$$

We have

$$\begin{aligned} \alpha = P(t=1) &= P\left(\overline{X}_n - \lambda_0 \ge c \mid P_{\lambda_0}^{(n)}\right) = P\left(\frac{X_1 + \dots X_n - n\lambda_0}{\sqrt{n\lambda_0}} \ge \sqrt{\frac{n}{\lambda_0}} c \mid P_{\lambda_0}^{(n)}\right) \\ &\sim P\left(\frac{X_1 + \dots X_n - n\lambda_0}{\sqrt{n\lambda_0}} \ge \sqrt{\frac{n}{\lambda_0}} c \mid \mathcal{N}(0,1)\right), \end{aligned}$$

thus $c = \sqrt{\frac{\lambda_0}{n}} z_{1-\alpha}$.

Example 9.13 (Cf. Weiß [21, Bsp. 7.7]). A company produces items which are known to have diameter μ_0. The client expects the items to be very similar. The diameters of a sample are $X_1, \dots, X_n$. Can we ensure the client that the items deviate less than σ_0^2?

We choose $\mathcal{P} := \left\{\mathcal{N}(\mu_0, \sigma^2)^{(n)} : \sigma \ge \sigma_0\right\}$ and $H_0\colon P \sim \mathcal{N}(\mu_0, \sigma_0^2)^{(n)}$. Apparently,

$$t(X_1, \dots X_n) := \begin{cases} 1 & \text{if } \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \ge c, \\ 0 & \text{else} \end{cases}$$

is a reasonable test. For this test,

$$\alpha = P(t=1) = P\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma_0^2}\right)^2 \geq nc \mid \mathcal{N}(\mu_0,\sigma_0^2)\right) = \chi_n^2\,[nc,\infty)\,,$$

so that $c = \frac{1}{n} q_{\chi_n^2}(1-\alpha)$ is the critical value.

Example 9.14 (Cf. Weiß [21]). The interarrival times of a Poisson process are $X_1,\dots,X_n$. Check, if $\lambda > \lambda_0$?

We choose $\mathcal{P} := \left\{E_\lambda^{(n)} : \lambda > 0\right\}$, the hypothesis $H_0 : E_\lambda^{(n)} : \lambda \leq \lambda_0$ and the alternative $H_1 : E_\lambda^{(n)} : \lambda > \lambda_0$. Based on (5.8) and (5.5) we have that $\mathbb{E}\,X = \frac{n}{\lambda}$ for $X \sim E_{n,\lambda}$, so we choose the test

$$t(X_1,\dots X_n) := \begin{cases} 1 & \text{if } \frac{n}{X_1+\dots+X_n} \geq \lambda_0 c, \\ 0 & \text{else.}\end{cases}$$

We want

$$\begin{aligned}\alpha = P(t=1) &= P\left(\frac{n}{X_1+\dots X_n} \geq \lambda_0 c \mid E_\lambda\right)\\ &= P\left(2\lambda_0(X_1+\dots X_n) \leq 2\frac{n}{c} \mid E_\lambda\right)\\ &\leq P\left(\underbrace{2\lambda(X_1+\dots X_n)}_{\chi_{2n}^2,\text{ cf. (5.6)}} \leq \frac{2n}{c} \mid E_\lambda\right)\\ &= \chi_{2n}^2\left(\left[0,\frac{2n}{c}\right]\right).\end{aligned}$$

It follows that $c = \frac{2n}{z_\alpha(\chi_{2n}^2)}$, where $z_\alpha(\chi_{2n}^2)$ is the α-quantile of the χ_{2n}^2 distribution.

9.4 STUDENT'S T-TEST

9.4.1 One sample location test

For the distribution see page 50 and the footnote 5.

In this section the statistics is

$$T := \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}$$

(cf. (5.15)) and t_{n-1} (t_{n-1}^{-1}, resp.) is the cdf (quantile, resp.) of Student's t-distribution.

(i) One tailed test, upper tailed test:[2] consider the test problem

$$\begin{aligned}H_0&\colon \mu = \mu_0,\\ H_1&\colon \mu > \mu_0.\end{aligned}$$

[2] rechtsseitiger Test

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then (compare with (7.13))

$$t_\alpha(X) = \begin{cases} 1 & \text{if } \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n} = T > t_{n-1}^{-1}(1-\alpha), \\ 0 & \text{else.} \end{cases} = \begin{cases} 1 & \text{if } 1 - t_{n-1}\left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}\right) < \alpha, \\ 0 & \text{else.} \end{cases}$$

Note the p-value $t_{n-1}(-T)$, cf. (7.7).

(ii) One tailed test, lower tailed test: the test for the problem

$$\begin{aligned} H_0 &: \mu = \mu_0, \\ H_1 &: \mu < \mu_0, \end{aligned}$$

reads

$$t_\alpha(X) = \begin{cases} 1 & \text{if } \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n} < t_{n-1}^{-1}(\alpha), \\ 0 & \text{else.} \end{cases} = \begin{cases} 1 & \text{if } t_{n-1}\left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}\right) < \alpha, \\ 0 & \text{else.} \end{cases}$$

The p-value is $t_{n-1}(T)$.

(iii) For two-tailed tests: for the problem

$$\begin{aligned} H_0 &: \mu = \mu_0, \\ H_1 &: \mu \neq \mu_0, \end{aligned}$$

the test

$$\begin{aligned} t_\alpha(X) &= \begin{cases} 1 & \text{else,} \\ 0 & \text{if } \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n} \in \left[t_{n-1}^{-1}\left(\frac{\alpha}{2}\right), t_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right)\right]. \end{cases} \\ &= \begin{cases} 1 & \text{else,} \\ 0 & \text{if } t_{n-1}\left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}\right) \in \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]. \end{cases} \end{aligned}$$

can be considered. Its p-value is $2t_{n-1}(|T|)$.

9.4.2 Two sample location test

Let $X_1, \dots X_n$ and $Y_1, \dots Y_m$ be independent samples with unknown, but *equal* (!) σ and consider the problem

$$\begin{aligned} H_0 &: \mu_X = \mu_Y, \\ H_1 &: \mu_X \neq \mu_Y. \end{aligned}$$

Consider the statistics

$$T := \frac{\overline{X}_n - \overline{Y}_m}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} = \sqrt{\frac{nm}{n+m}}\frac{\overline{X}_n - \overline{Y}_m}{S} \sim t_{n+m-2},$$

where $S^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$ is the pooled variance. A possible test is

$$t_\alpha(X) = \begin{cases} 1 & \text{if } |T| > t_{n+m-2}^{-1}\left(1 - \frac{\alpha}{2}\right), \\ 0 & \text{else.} \end{cases}$$

9.5 WELCH'S T-TEST

Let $X_1, \dots X_n$ and $Y_1, \dots Y_m$ be independent samples with unknown, but possibly *unequal* (!) σ and consider the problem

$$H_0 \colon \mu_X = \mu_Y,$$
$$H_1 \colon \mu_X \neq \mu_Y.$$

Welch's t-test considers the statistics

$$T := \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim t_\nu,$$

where $\nu \approx \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n(n-1)} + \frac{s_Y^4}{m(m-1)}}$ (the Welch–Satterthwaite equation) is an approximation for the degrees of freedom.

9.6 FISHER'S F-TEST

The tests have been developed by Fisher.[3]

Recall from (5.13) that $s_n^2(X) \sim \sigma^2 \frac{\chi_{n-1}^2}{n-1}$ which we compare with $s_m^2(Y) \sim \sigma^2 \frac{\chi_{m-1}^2}{m-1}$.

9.6.1 ANOVA

The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. It follows from Proposition 5.21 that

$$F := \frac{s_m^2(Y)}{s_n^2(X)} \sim \frac{\chi_{m-1}^2/(m-1)}{\chi_{n-1}^2/(n-1)} \sim F_{m-1,n-1}.$$

Fisher hence proposes F for the test

$$H_0 \colon \sigma_2^2 = \sigma_1^2,$$
$$H_1 \colon \sigma_2^2 > \sigma_1^2.$$

9.6.2 Lack-of-fit sum of squares

The hypothesis that a proposed regression model fits the data well.

9.7 KOLMOGOROV–SMIRNOV TEST

Recall the empirical distribution function

$$F_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty,x]}(X_i). \tag{9.2}$$

[3] Ronald A. Fisher, 1890–1962, statistician and geneticist

9.7.1 One-sample Kolmogorov–Smirnov test

Definition 9.15. The Kolmogorov–Smirnov statistic for a given distribution function F is (cf. (12.4))

$$D_n := \|F_n - F\|_\infty := \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)|\,.$$

Theorem 9.16 (Kolmogorov)**.** *Let $X_i \sim X$ be iid with continuous distribution function F. Then (cf. Table 7.1b)*

$$\lim_{n\to\infty} P\Big(\sqrt{n}\,\underbrace{\sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)|}_{D_n} \le z\Big) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2z^2}. \tag{9.3}$$

For the problem

$$\begin{aligned} H_0&\colon X_i \sim F,\\ H_1&\colon X_i \nsim F.\end{aligned}$$

the Kolmogorov–Smirnov test is

$$t_\alpha(X_1,\dots,X_n) = \begin{cases} 1 & \text{if } \sqrt{n}\,D_n > K_\alpha,\\ 0 & \text{else,}\end{cases}$$

where

$$P(K > K_\alpha) = \alpha \tag{9.4}$$

and $P(K \le z) = 1 + 2\sum_{k=1}^\infty (-1)^k e^{-2k^2z^2}$, cf. (9.3). Table 7.1b lists these value.

To wit, consider $K := \sup_{t\in[0,1]} |B_t|$ for a Brownian bridge B_t. Then $\sqrt{n}\,D_n \xrightarrow[n\to\infty]{} \sup_{x\in\mathbb{R}} \left|B_{F(x)}\right|$ in distribution and thus the result.

9.7.2 Two-sample Kolmogorov–Smirnov test

The Kolmogorov[4]–Smirnov[5] test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,m} := \sup_{x\in\mathbb{R}} \left| \underbrace{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,x]}(X_i)}_{F_{X,n}(x)} - \underbrace{\frac{1}{m}\sum_{j=1}^m \mathbb{1}_{(-\infty,x]}(Y_j)}_{F_{Y,m}(x)} \right|$$

Theorem 9.17 (Kolmogorov Smirnov)**.** *For $X_i \sim F$ and $Y_i \sim F$, all iid it holds that*

(i) $P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \le z\right) \xrightarrow[n,m\to\infty]{} 1 - 2e^{-2z^2}$ *and*

(ii) $P\left(\sqrt{\frac{nm}{n+m}} \sup_{x\in\mathbb{R}} \left\{F_{X,n}(x) - F_{Y,m}(x)\right\} \le z\right) \xrightarrow[n,m\to\infty]{} 1 - e^{-2z^2}$.

[4] Andrey Kolmogorov, 1903–1987

[5] Nikolai Smirnov, 1900-1960

For the problem

$$\begin{aligned} H_0 &\colon F_{X_i} \sim F_X, \\ H_1 &\colon X_i \nsim F. \end{aligned}$$

The Kolmogorov–Smirnov test is

$$t_\alpha(X_1, \dots, X_n) = \begin{cases} 1 & \text{if } D_n > c(\alpha)\sqrt{\frac{n+m}{nm}}, \\ 0 & \text{else,} \end{cases}$$

where

$$c(\alpha) = \sqrt{-\frac{1}{2}\log\frac{1-\alpha}{2}}. \tag{9.5}$$

Table 7.1b lists these values.

9.8 KUIPER'S TEST

is closely related to the Kolmogorov–Smirnov test, it uses the test statistics

$$D_+ := \max_{i=1,\dots,n}\left\{\frac{i}{n} - F(X_{(i)})\right\} + \max_{i=1,\dots,n}\left\{F(X_{(i)}) - \frac{i-1}{n}\right\}.$$

9.9 CRAMÉR-VON MISES TEST

The test statistics is $\omega_n^2 := \frac{1}{12n} + \sum_{j=1}^n \left(F\left(X_{(j)}\right) - \frac{2j-1}{2n}\right)^2$ (cf. (12.5)). In the limit, this statistics follows an ω^2 distribution with $\omega^2 \sim \int_0^1 B_t^2 \,\mathrm{d}t \sim \sum_{k=1}^\infty \frac{\xi_k^2}{k^2\pi^2}$, where B_t is a Brownian bridge and ξ_k are independent normals.

9.10 WALD TEST

The maximum likelihood is asymptotically normal, i.e., $\lim_{n\to\infty}\hat{\vartheta} \to \mathcal{N}(\vartheta, \Sigma_{\hat{\vartheta}})$ where $\vartheta = (\vartheta_1 \dots \vartheta_k)$ and $\Sigma_{\hat{\vartheta}}$ is the asymptotic non-singular covariance matrix of the Likelihood estimator.

Wald[6] thus proposes to employ the statistics

$$T_W^2 := (\vartheta - \vartheta_0)^\top \Sigma_{\hat{\vartheta}}^{-1} (\vartheta - \vartheta_0) \sim \chi_k^2$$

to test

$$\begin{aligned} H_0 &\colon \vartheta = \vartheta_0, \\ H_1 &\colon \vartheta \neq \vartheta_0. \end{aligned}$$

9.11 SHAPIRO–WILK TEST

https://math.mit.edu/~rmd/465/shapiro.pdf

[6] Abraham Wald, 1902–1950

α	Reject H_0 if
0.10	$JB > 4.6$
0.05	$JB > 6.0$
0.02	$JB > 7.8$
0.01	$JB > 9.2$

Table 9.3: Jarque–Bera

1	2	3	4	5	6	total
74	107	99	98	103	119	600

Table 9.4: Dice

9.12 JARQUE–BERA-TEST

The skewness[7] is $S = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^3}{\left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2\right)^{3/2}}$ and kurtosis[8] is $K = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^4}{\left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2\right)^2}$. Then

$$JB := \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \sim \chi_2^2$$

and the test is (cf. Table 9.3)

$$H_0\colon X_i \sim \mathcal{N}(\mu, \sigma^2)^{(n)} \text{ are normally distributed,}$$
$$H_1\colon X_i \text{ are not normally distributed.}$$

9.13 PROBLEMS

Exercise 9.1. *Based on Table 9.1a, is it fair to say that it is more likely for women to study than for men? Test the hypothesis $H_0\colon p_{men} \le p_{women}$ versus the alternative $H_1\colon p_{men} > p_{women}$.*

Exercise 9.2 (Coin flipping)**.** *Suppose a coins shows 532 heads and 468 tails. Is the coin fair? Give the p-value of the corresponding χ_1^2 distribution for $\alpha = 5\%$.*

Exercise 9.3. *Suppose a dice shows the counts displayed in Table 9.4 after 600 throws. With, $\alpha = 5\%$, is the dice fair?*

[7] Schiefe, cf. Footnote 6 (page 23)

[8] Wölbung, cf. Footnote 5 (page 21)

Descriptive Statistics

10

Also lautet ein Beschluß,
daß der Mensch was lernen muß.

Wilhelm Busch, 1832–1908

Definition 10.1. The *empirical measure*[1] of a set A is the counting measure[2]

$$P_n(A) := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(X_i).$$

The *empirical distribution function* is

$$F_n(x) := P_n\big((-\infty, x]\big) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \le x},$$

where $(-\infty, x] := (-\infty, x_1] \times \cdots \times (-\infty, x_d] \subset \mathbb{R}^d$ is a hypercube for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $\mathbb{1}_{X\le x} = \begin{cases} 1 & \text{if } X_1 \le x_1, \dots, X_d \le x_d, \\ 0 & \text{else} \end{cases}.$

Proposition 10.2. *Let A be fixed and X_i independent and identically distributed (iid.). Then $(X_1, \dots, X_n) \mapsto n \cdot P_n(A) \in \{0, 1, \dots, n\}$ is a binomial random variable with mean $n \cdot P(A)$ and variance $n\, P(A)\,\big(1 - P(A)\big)$. In particular, $P_n(A)$ is an unbiased estimator for $P(A)$. Further, $n\, P_n(A) \sim \operatorname{bin}\big(n, P(A)\big)$.*

Proof. We have that $\mathbb{E}\, n P_n(A) = \sum_{i=1}^n \mathbb{E}\, \mathbb{1}_A(X_i) = n \cdot P(A)$ and it is evident that $P(X_1 \in A, \dots, X_n \in A \text{ exactly } k \text{ times}) = \binom{n}{k} P(A)^k \big(1 - P(A)\big)^{n-k}$ and hence the distribution.

For the variance observe that

$$\begin{aligned}\mathbb{E}\, n^2 P_n(A)^2 = \mathbb{E} \sum_{i,j=1}^n \mathbb{1}_A(X_i) \cdot \mathbb{1}_A(X_j) &= \sum_{i \ne j} P(X_i \in A,\, X_j \in A) + \sum_{i=1}^n P(X_i \in A) \\ &= (n^2 - n) P(A)^2 + n P(A) = n P(A)\big(1 - P(A)\big) + n^2 P(A)^2,\end{aligned}$$

so that $\operatorname{var}\big(n \cdot P_n(A)\big) = n\, P(A)\big(1 - P(A)\big)$. □

Proposition 10.3. *Let X_i be iid with common cdf F. It holds that*

$$\mathbb{E}\, F_n(x) = F(x)$$

and (with $x \wedge y = \min(x, y)$)

$$\operatorname{cov}\big(F_n(x),\, F_n(y)\big) = \frac{1}{n}\big(F(x \wedge y) - F(x) \cdot F(y)\big).$$

[1] Paul Dirac, 1902–1984, English theoretical physicist

[2] The Dirac measure is $\delta_x(A) := \mathbb{1}_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$

Proof. From the preceding proposition we have that

$$P\left(F_n(y)=\frac{i}{n}\right)=\binom{n}{i}F(y)^i\big(1-F(y)\big)^{n-i}$$

and, for $x\le y$,

$$P\left(F_n(x)=\frac{j}{n}\,\middle|\,F_n(y)=\frac{i}{n}\right)=\binom{i}{j}\left(\frac{F(x)}{F(y)}\right)^j\left(1-\frac{F(x)}{F(y)}\right)^{i-j},\qquad j=0,\dots,i.$$

It follows that

$$\begin{aligned}\mathbb{E}\,F_n(x)F_n(y)&=\mathbb{E}\,\mathbb{E}\left(F_n(x)F_n(y)\mid F_n(y)\right)\\&=\sum_{i=0}^n\binom{n}{i}F(y)^i\big(1-F(y)\big)^{n-i}\cdot\sum_{j=0}^i\binom{i}{j}\left(\frac{F(x)}{F(y)}\right)^j\left(1-\frac{F(x)}{F(y)}\right)^{i-j}\cdot\frac{i}{n}\frac{j}{n}\\&=\sum_{i=0}^n\binom{n}{i}F(y)^i\big(1-F(y)\big)^{n-i}\frac{i}{n^2}\cdot i\frac{F(x)}{F(y)}\qquad(10.1)\\&=\frac{1}{n^2}\frac{F(x)}{F(y)}\left(nF(y)\big(1-F(y)\big)+\big(n\,F(y)\big)^2\right)\qquad(10.2)\\&=\frac{1}{n}F(x)\big(1-F(y)\big)+F(x)F(y),\end{aligned}$$

where we have used that $\mathbb{E}\,X=i\,\frac{F(x)}{F(y)}$ in (10.1) for $X\sim\operatorname{bin}\left(i,\frac{F(x)}{F(y)}\right)$ and $\mathbb{E}\,X^2=n\,F(y)\big(1-F(y)\big)+\big(n\,F(y)\big)^2$ for $X\sim\operatorname{bin}\big(n,F(y)\big)$ in (10.2). Hence the assertion. □

10.1 BOX-AND-WHISKER PLOT

A method for graphically depicting groups of numerical data through their quartiles (cf. Figure 10.1) introduced by Tuckey.[3] Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points. Box plots are non-parametric.

10.2 HISTOGRAM, Q–Q AND P–P PLOTS

For two random variables X (Y, resp.) with cdf. F_X (F_Y, resp.), the Q–Q plot (for Quantile–Quantile plot, cf. Figure 10.2b)

$$[0,1]\ni p\mapsto\big(F_X^{-1}(p),\,F_Y^{-1}(p)\big)\in\mathbb{R}^2\qquad(10.3)$$

plots their quantiles against each other. The P–P plot (for Probability–Probability, or Percentage–Percentage plot) is

$$\mathbb{R}\ni q\mapsto\big(F_X(q),\,F_Y(q)\big)\in[0,1]^2.\qquad(10.4)$$

The Q–Q plot is more widely used, but they are both referred to as *probability plot*, and are potentially confused.

[3] John W. Tukey, 1915–2000, American mathematician

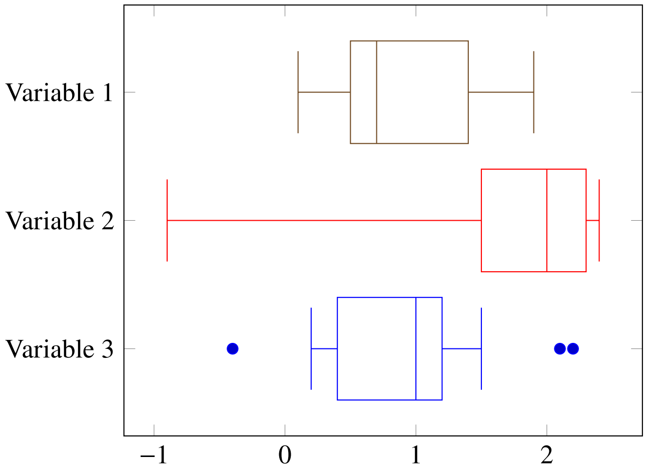


Figure 10.1: Box plot

The usual combinations in probability plots are F_X and the empirical distribution function F_n. Denote the realizations by X_i, $i = 1, \dots, n$, and set

$$F_n^{-1}\left(\frac{i}{n+1}\right) =: X_{(i)}, \qquad i = 1, \dots, n. \tag{10.5}$$

A usual Q–Q plot displays (following Van der Waerden, put $p \in \left\{\frac{i}{n+1} : i = 1, \dots, n\right\}$ in (10.3))

$$i \mapsto \left(F_X^{-1}\left(\frac{i}{n+1}\right), X_{(i)}\right), \qquad i = 1, \dots, n$$

and the P–P plot (choose $q \in \{X_i : i = 1, \dots, n\}$ in (10.4), cf. Figure 10.2c)

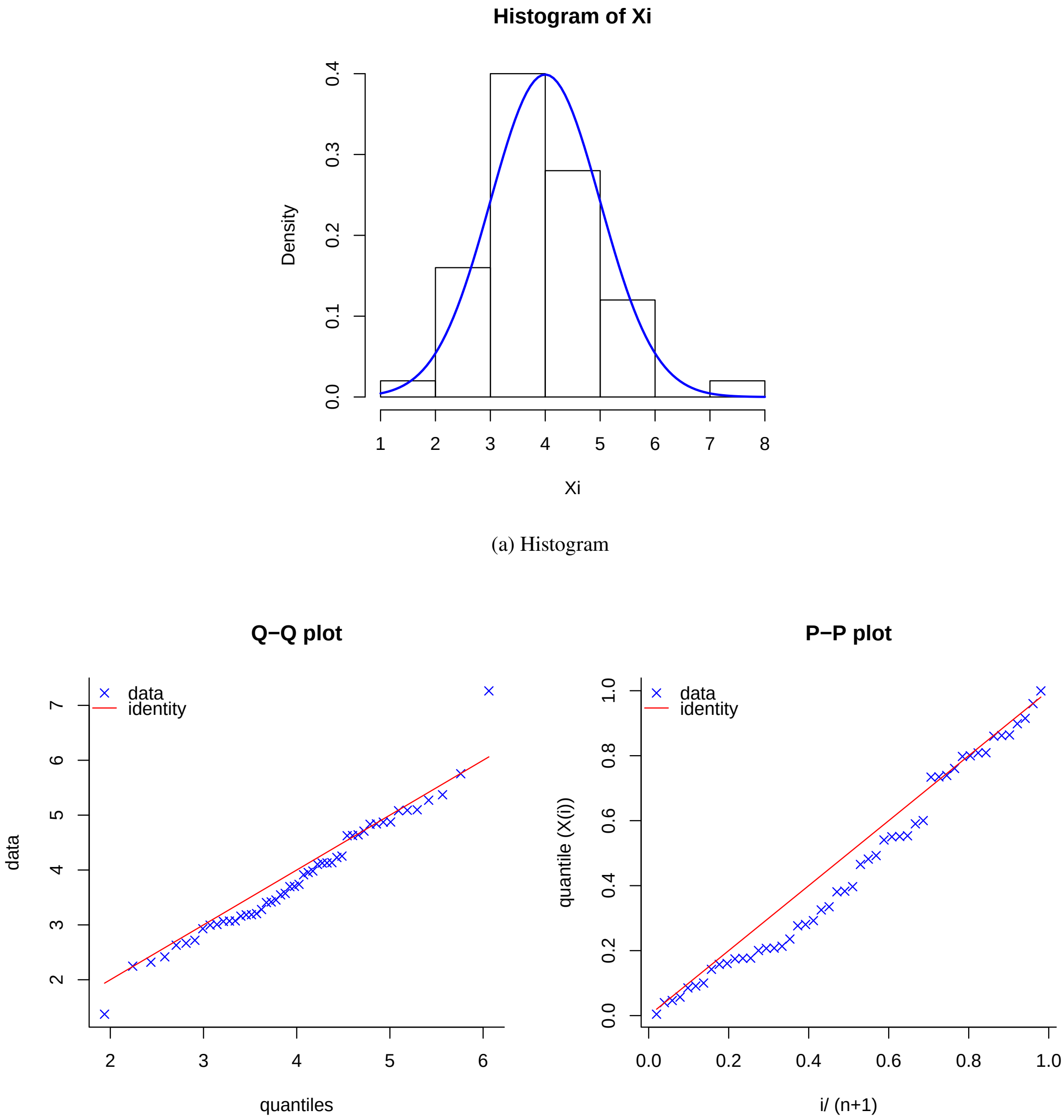
$$i \mapsto \left(\frac{i}{n+1}, F_X\big(X_{(i)}\big)\right), \qquad i = 1, \dots, n. \tag{10.6}$$

Remark 10.4. Q–Q and P–P plots are also used to plot two samples against each other.

10.3 PROBLEMS

Exercise 10.1. *Consider the probability measure (Dirac measure)* $P_{\omega_0}(A) := \begin{cases} 1 & \text{if } \omega_0 \in A, \\ 0 & \text{else} \end{cases}$ *with* $A \subset \mathbb{R}$ *(cf. Footnote 2 on page 93). Give* $\mathbb{E}\,X^2$.

Exercise 10.2. *For* $\lambda \in (0, 1)$ *set* $P := (1-\lambda)P_{\omega_0} + \lambda P_{\omega_1}$. *Compute* $\mathbb{E}\,X$ *using (1.2) and (1.15).*



(a) Histogram

(b) Q–Q plot

(c) P–P plot

Figure 10.2: Histogram, Q–Q and P–P plot of the same 50 samples, distributed according $\mathcal{N}(4, 1)$

Order Statistics

11

Nichts gibt so sehr das Gefühl der
Unendlichkeit als wie die Dummheit.

Ödön von Horváth, 1901–1938

Definition 11.1 (Order statistics, rank, cf. (2.11))**.** We shall write $(\cdot)$ for the permutation on $\{1,2,\dots,n\}$ so that

$$X_{(1)} \le \dots \le X_{(n)}$$

and $\{X_1,\dots,X_n\} = \left\{X_{(1)},\dots,X_{(n)}\right\}$. $X_{(i)}$ is the i^{th} order statistic of the sample, i.e., its rank is (i) (cf. also (2.11)). Occasionally, the notations $X_{1:n} \le \dots \le X_{n:n}$ or $X_{n(1)} \le \dots \le X_{n(n)}$ are used as well to denote the order statistics.

Remark 11.2. Sometimes it is convenient to include the observations $X_{(0)} := -\infty$ and $X_{(n+1)} := +\infty$ in the sample. By convention, $F\big(X_{(0)}\big) = 0$ and $F\big(X_{(n+1)}\big) = 1$.

Remark 11.3. Note that $F_n(x) = \frac{i}{n}$ for $x \in [X_{(i)}, X_{(i+1)})$ and $F_n^{-1}(x) = X_{(i)}$ whenever $x \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$, in particular $X_{(i)} = F_n^{-1}(i/n)$, $i = 0,\dots,n+1$. As well, $X_{(i)} = F_n^{-1}\left(\frac{i}{n+1}\right)$, cf. (9.2).

11.1 DENSITIES

For order statistics, the following hold true.

Proposition 11.4 (Density of order statistics)**.** *Let X_i, $i = 1,\dots,n$, be iid with common density $f(\cdot)$ (cdf $F(\cdot)$, resp.). It holds that*

$$\begin{aligned}
f_{X_{(k)}}(x) &= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} f(x)\big(1-F(x)\big)^{n-k},\\
f_{X_{(j)},X_{(k)}}(x,y) &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!}, \qquad j<k,\ x\le y,\\
&\quad \cdot F(x)^{j-1} f(x)\big(F(y)-F(x)\big)^{k-1-j} f(y)\big(1-F(y)\big)^{n-k},\\
f_{X_{(1)},\dots,X_{(n)}}(x_1,\dots,x_n) &= n!\, f(x_1)\cdots f(x_n), \quad x_1 \le \dots \le x_n.
\end{aligned}$$

Proposition 11.5 (cdf of order statistics)**.** *Let X_i, $i = 1,\dots,n$, be iid. It holds that*

$$\begin{aligned}
P\left(X_{(k)} \in \mathrm{d}x\right) &= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1-F(x)\big)^{n-k}\,\mathrm{d}F(x),\\
P\left(X_{(j)} \in \mathrm{d}x, X_{(k)} \in \mathrm{d}y\right) &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!}\cdot \qquad j<k,\ x\le y,\\
&\quad \cdot F(x)^{j-1}\big(F(y)-F(x)\big)^{k-1-j}\big(1-F(y)\big)^{n-k}\,\mathrm{d}F(x)\,\mathrm{d}F(y),\\
P\left(X_{(1)} \in \mathrm{d}x_1,\dots,X_{(n)} \in \mathrm{d}x_n\right) &= n!\,\mathrm{d}F(x_1)\cdots\mathrm{d}F(x_n), \quad x_1\le\dots\le x_n.
\end{aligned}$$

Corollary 11.6 (cdf and pdf of min and max)**.** *In particular we have that*

$$P\left(\min_{i=1,\dots,n} X_i \in \mathrm{d}x\right) = n\left(1-F(x)\right)^{n-1} P(X \in \mathrm{d}x),$$
$$P\left(\max_{i=1,\dots,n} X_i \in \mathrm{d}x\right) = n\, F_X(x)^{n-1}\, P(X \in \mathrm{d}x)$$

and

$$P\left(\min_{i=1,\dots,n} X_i \le x\right) = 1-\left(1-F(x)\right)^{n},$$
$$P\left(\max_{i=1,\dots,n} X_i \le x\right) = F(x)^n.$$

11.2 FIRST MOMENTS

Recall that $U_k := F(X_k) \sim U[0,1]$ is uniformly distributed and $U_k \in [0,1]$. In what follows we assume that all X_i are independent and we study $U_{(k)} = F(X_{(k)})$. Note, that $U_{(j)}$ and $U_{(k)}$ are dependent, although U_j and U_k are independent.

Corollary 11.7. *It holds that*

(i) $f_{F(X_{(k)})}(p) = \frac{n!}{(k-1)!(n-k)!} p^{k-1}\left(1-p\right)^{n-k}$*, i.e,* $U_{(k)} \sim \mathrm{Beta}(k, n-k+1)$ *and*

(ii) $f_{F(X_{(j)}),F(X_{(k)})}(p,q) = \frac{n!p^{j-1}\left(q-p\right)^{k-1-j}\left(1-q\right)^{n-k}}{(j-1)!(k-j-1)!(n-k)!}$, $0 \le p \le q \le 1$.

Corollary 11.8. *It holds that (cf. Propositions 10.3)*

(i) $\mathbb{E}\, F(X_{(k)}) = \frac{k}{n+1}$,

(ii) $\operatorname{var} F(X_{(k)}) = \frac{1}{n+2} \cdot \frac{k}{n+1}\left(1-\frac{k}{n+1}\right)$ *and*

(iii) $\operatorname{cov}\left(F(X_{(j)}), F(X_{(k)})\right) = \frac{1}{n+2} \cdot \frac{j}{n+1}(1-\frac{k}{n+1})$, $j \le k$.

Remark 11.9. Cf. the P–P plot.

Remark 11.10. Note, that $F(X_{(k)}) \sim \mathrm{Beta}_{k,n+1-k} \approx \mathcal{N}\left(\frac{k}{n+1}, \frac{1}{n+2}\frac{k}{n+1}\left(1-\frac{k}{n+1}\right)\right)$, where Beta is the Beta distribution, cf. Exercise 5.10.

Proof. Use the formula (5.10) for Euler's Beta function to get

$$\begin{aligned}\mathbb{E}\, F(X_{(k)}) &= \int_{-\infty}^{\infty} F(x) f_{X_{(k)}}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{n!}{(k-1)!\,(n-k)!} F(x)^k f(x)(1-F(x))^{n-k}\,\mathrm{d}x \\ &= \int_0^1 \frac{n!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k}\,\mathrm{d}p = \frac{n!}{(k-1)!\,(n-k)!} \cdot \frac{k!(n-k)!}{(n+1)!} = \frac{k}{n+1},\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\, F(X_{(j)}) \cdot F(X_{(k)}) \\
&\quad = \frac{n!}{(j-1)!\,(k-j-1)!\,(n-k)!} \int_0^1 \int_p^1 pq \cdot p^{j-1}(q-p)^{k-1-j}(1-q)^{n-k}\, \mathrm{d}q\mathrm{d}p \\
&\quad = \cdots \int_0^1 \int_0^q p^j (q-p)^{k-1-j} q(1-q)^{n-k}\, \mathrm{d}p\mathrm{d}q \\
&\quad \underset{p\leftarrow qp}{=} \cdots \int_0^1 \int_0^1 (qp)^j q^{k-1-j}(1-p)^{k-1-j} q^2 (1-q)^{n-k}\, \mathrm{d}p\mathrm{d}q \\
&\quad = \cdots \int_0^1 p^j (1-p)^{k-1-j}\, \mathrm{d}p \cdot \int_0^1 q^{k+1}(1-q)^{n-k}\, \mathrm{d}q \\
&\quad = \frac{n!}{(j-1)!\,(k-j-1)!\,(n-k)!} \cdot \frac{j!(k-j-1)!}{k!} \cdot \frac{(k+1)!(n-k)!}{(n+2)!} \\
&\quad = \frac{j(k+1)}{(n+1)(n+2)}.
\end{aligned}$$

For the variance,

$$\begin{aligned}
\mathbb{E}\, F(X_{(k)})^2 &= \int_{-\infty}^{\infty} F(x)^2 f_{X_{(k)}}(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{n!}{(k-1)!\,(n-k)!} F(x)^{k+1} f(x)(1-F(x))^{n-k}\, \mathrm{d}x \\
&= \int_0^1 \frac{n!}{(k-1)!\,(n-k)!} p^{k+1}(1-p)^{n-k}\, \mathrm{d}p \\
&= \frac{n!}{(k-1)!\,(n-k)!} \cdot \frac{(k+1)!(n-k)!}{(n+2)!} = \frac{k(k+1)}{(n+1)(n+2)};
\end{aligned}$$

together with (i) it follows that $\operatorname{var} F(X_{(k)}) = \frac{k(n+1-k)}{(n+1)^2(n+2)}$ and hence the assertion. □

11.3 DERIVED ORDER STATISTICS

Definition 11.11 (Range)**.** We define the following derived statistics.

(i) The *sample median* is $m_X := \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even;} \end{cases}$

(ii) the *range*[1] of a set of data is the difference between the largest and smallest values,

$$\Delta_n := \max_{i=1,\dots n} X_i - \min_{i=1,\dots n} X_i = X_{(n)} - X_{(1)};$$

(iii) the *midrange* is $\frac{1}{2}\left(\max_{i=1,\dots n} X_i - \min_{i=1,\dots n} X_i\right) = \frac{1}{2}\left(X_{(n)} - X_{(1)}\right)$;

(iv) the L-estimator $\frac{1}{2}\left(F^{-1}(1/4) + F^{-1}(3/4)\right)$ is called *midhinge* (the interquartile range is $F^{-1}(3/4) - F^{-1}(1/4)$);

(v) the *pseudomedian* or *Hodges-Lehmann*[2] *estimator* of a population is the median of all $\frac{1}{2}\left(X_i + X_j\right)$, $1 \le i \le j \le n$, i.e, the median of the averages of all $\frac{n(n+1)}{2}$ pairs.

[1] Spannweite, dt.

[2] *Cf. Footnote 7 on page 12.*

Proposition 11.12. *Let $X_1, \dots, X_n$ be iid. Then the range has pdf*

$$f_{\Delta_n}(t) = n(n-1)\int_{-\infty}^{\infty} f_X(x)\big(F_X(x+t) - F_X(x)\big)^{n-2} f_X(x+t)\,\mathrm{d}x, \qquad t \ge 0,$$

and cdf

$$P(\Delta_n \le t) = F_{\Delta_n}(t) = n\int_{-\infty}^{\infty} f_X(x)\big(F_X(x+t) - F_X(x)\big)^{n-1}\,\mathrm{d}x, \qquad t \ge 0.$$

Remark 11.13. Note that $F_{\Delta_n}(t)$ is a cdf, as $F_{\Delta_n}(0) = 0$ and, for $t \to \infty$,

$$F_{\Delta_n}(t) \xrightarrow[t\to\infty]{} n\int_{-\infty}^{\infty} f_X(x)\big(1 - F_X(x)\big)^{n-1}\,\mathrm{d}x = -\left(1 - F_X(x)\right)^n\Big|_{x=-\infty}^{\infty} = 1.$$

Theorem 11.14. *For $p \in (0,1)$ and $n \to \infty$ we have that*

(i) $F(X_{([np])}) \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ *and*

(ii) $X_{([np])} \sim \mathcal{N}\left(F^{-1}(p), \frac{1}{n}\frac{p(1-p)}{f\left(F^{-1}(p)\right)^2}\right)$.

Remark 11.15 (Brownian Bridge). For $t \in [0,1]$ define the process

$$B_t^n := \sqrt{n+2}\cdot\left(F(X_{(t(n+1))}) - t\right), \qquad t \in [0,1],$$

with linear interpolation of adjacent samples and the convention mentioned in Remark 11.2 (i.e., $F(X_{(0)}) = 0$ and $F(X_{(n+1)}) = 1$). Then $B_t^n \approx \mathcal{N}\big(0, t(1-t)\big)$ and

$$\operatorname{cov}\left(B_s^n, B_t^n\right) = st - s \wedge t, \qquad s, t \in \left\{\frac{k}{n+1} : k = 0, \dots n+1\right\},$$

which is the covariance structure of a Brownian bridge. Hence, $B_t^n \xrightarrow[n\to\infty]{} B_t$ in distribution, where B_t is a Brownian bridge.

Remark 11.16. Note as well that (cf. (5.10))

$$\begin{aligned}\mathbb{E}\, X_{(k)} &= \int_{-\infty}^{\infty} x \cdot f_{X_{(k)}}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x \cdot \frac{n!}{(k-1)!\,(n-k)!} F(x)^{k-1} f(x)(1-F(x))^{n-k}\,\mathrm{d}x \\ &= \int_0^1 F^{-1}(p)\frac{n!}{(k-1)!\,(n-k)!} p^{k-1}(1-p)^{n-k}\,\mathrm{d}p \to F^{-1}\left(\frac{k}{n+1}\right).\end{aligned}$$

11.4 PROBLEMS

Exercise 11.1. *Let U_1 and $U_2 \sim U([0,1])$ be uniformly distributed in* $[0,1]$. *Show that* $\max(U_1, U_2)$ *and $\sqrt{U_1}$ have the same distribution. More generally, let U_i and U be independent uniforms. Show that* $\max(U_1, \dots, U_n)$ *and $\sqrt[n]{U}$ have the same distribution.*

Exercise 11.2. *Show that the sample median of the sample $X = (2,2,3,5)$ is $m_X = 2.5$, the pseudomedian is $m_X^* = 2.75$.*

Exercise 11.3. *Show that for independent $X_i \sim E_\lambda$ we have that $X_{(1)} = \min_{i=1,\dots,n} X_i \sim E_{n\lambda}$. Give the expectation and variance of $X_{(1)}$.*

Exercise 11.4. *Let $X_i \sim U[0,1]$, $i=1,\dots,n$, be independent.*

(i) *Show that the range Δ_n has the density $P(\Delta_n \in \mathrm{d}t) = n(n-1)t^{n-2}(1-t)\,\mathrm{d}t$;*

(ii) *show that the moments are given by $\mathbb{E}\,\Delta_n^k = \frac{n(n-1)}{(n+k)(n+k-1)}$ and deduce that $\mathbb{E}\,\Delta_n = 1-\frac{2}{n+1}$ and $\operatorname{var}\Delta_n = \frac{2(n-1)}{(n+1)^2(n+2)} = \mathcal{O}\left(1/n^2\right)$.*

Exercise 11.5. *Let $0 \le j < k \le n$ and u_0 be fixed. Compute the density of $f_{U_{(j)}}\big(\cdot \mid U_{(k)} = u_0\big)$ and give the distribution of $U_{(j)} \mid U_{(k)} = u_0$, i.e., the distribution of $U_{(j)}$ provided that $U_{(k)} = u_0$. Hint: see (1.11).*

Estimating the location parameter

Exercise 11.6. *Let X_i, $i=1,\dots,n$ (n odd, $n=2k-1$), be iid. The sample median is $m = X_{(k)}$, cf. Definition 11.11(i).*

(i) *Show that the sample median has density $P\left(m \in \mathrm{d}v\right) = \frac{(2k-1)!}{(k-1)!^2}\big(F(v)(1-F(v))\big)^{k-1}\,\mathrm{d}F(v)$;*

(ii) *show for $X_i \sim U[0,1]$ that $\mathbb{E}\,m = \frac{1}{2}$ and $\operatorname{var} m = \frac{1}{8+4n} = \mathcal{O}\left(1/n\right)$.*

Exercise 11.7. *Let $U_i \sim U[0,1]$, $i=1,\dots,n$. Show that*

$$\mathbb{E}\,\frac{1}{2}\left(U_{(1)}+U_{(n)}\right) = \frac{1}{2} \text{ and } \operatorname{var}\left(\frac{1}{2}\big(U_{(1)}+U_{(n)}\big)\right) = \frac{1}{2(n+1)(n+2)} = \mathcal{O}\left(1/n^2\right).$$

Exercise 11.8. *Define $F_g^{-1}(p) \coloneqq \sum_{i=0}^n \binom{n}{k}(1-p)^{n-k}p^k \cdot g\big(X_{(k)}\big)$. Verify that $\int_0^1 F_g^{-1}(p)\,\mathrm{d}p$ is an estimator for $\mathbb{E}\,g(X)$ and $F_g^{-1}(p)$ an estimator for $F_{g(X)}^{-1}(p)$. Show that $p \mapsto F_g^{-1}(p)$ is nondecreasing, provided that $g(\cdot)$ is nondecreasing. What does that mean for $g(x)=x$?*

Exercise 11.9. *Let U_1 and U_2 be uniformly distributed on $[0,1]$ and set $X \coloneqq \min(U_1,U_2)$ and $Y \coloneqq \max(U_1,U_2)$. Show that $P\left(X > \alpha/2 \text{ and } Y > \alpha\right) = 1-\alpha$.*

12 Theory of Estimation

All models are wrong, but some are useful.

attributed to George Box, 1919–2013

Given a statistical model

$$\mathcal{E} = (\mathcal{X}, \Sigma, (P_\vartheta)_{\vartheta\in\Theta})$$

(in its original formulation due to Blackwell [2], cf. also Le Cam [8]), the theory of estimation is interested in *estimators* (a *decision rule*)

$$\hat\vartheta\colon \mathcal{X} \to \Theta$$

based on observed data $X \in \mathcal{X}$. Note, that $\hat\vartheta(X)$ is a random variable.

12.1 Loss and expected loss

Definition 12.1. The *risk function* (aka. *frequentist expected loss*) of an estimator $\hat\vartheta\colon \mathcal{X} \to \Theta$ with respect to the loss function (regret function) $\ell\colon \Theta\times\Theta \to \mathbb{R}$ is

$$r_\ell\big(\hat\vartheta(\cdot),\vartheta\big) := \mathbb{E}_\vartheta\, \ell\Big(\hat\vartheta(\cdot),\vartheta\Big) = \int_{\mathcal{X}} \ell\Big(\hat\vartheta(x),\vartheta\Big) P_\vartheta(\mathrm{d}x). \tag{12.1}$$

Examples of frequently used loss functions include the *Minkowski loss*

$$\ell_p(\vartheta',\vartheta) = \|\vartheta-\vartheta'\|_p^p \text{ or } \ell_\epsilon(\vartheta,\vartheta') = \begin{cases} 0 & \text{if } \|\vartheta-\vartheta'\| \le \epsilon,\\ 1 & \text{else.}\end{cases} \tag{12.2}$$

Here, ℓ_1 is called *Laplace* or *modular loss* ($p = 1$), ℓ_2 Gauß loss ($p = 2$) and ℓ_ε is the $0-1$ loss ($\epsilon = 0$ is only a useful idea for discrete distributions). See Definition 1.30 for the Huber loss function.

Example 12.2. The Poisson loss function

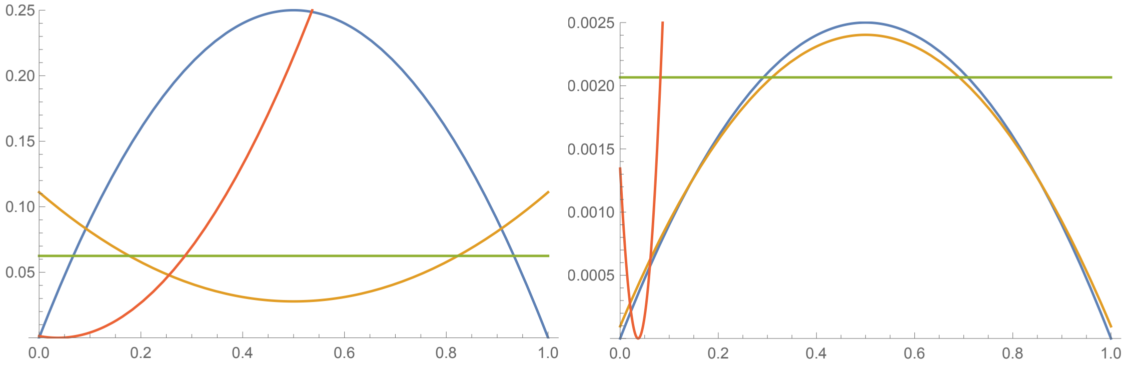
$$\ell(x,\pi) := y\log\frac{y}{\pi} - (y-\pi)$$

is not symmetric.

Example 12.3. Consider the Gauß loss function $\ell(\vartheta',\vartheta) = (\vartheta'-\vartheta)^2$. For the binomial model $S_n \sim bin(n,p)$ built of independent Bernoulli observations $X_1,\dots,X_n$ consider the following estimators for the parameter $p = \vartheta \in \Theta = [0,1]$ ($S_n := X_1 + \dots + X_n$):

(i) $\hat p_1 := \frac{S_n}{n}$, its risk is $r_\ell(\hat p_1,p) = \mathbb{E}_p(\hat p_1-p)^2 = \frac{1}{n^2}\,\mathbb{E}_p\,(S_n - n\,p)^2 = \frac{1}{n^2}\operatorname{var} S_n = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$;

(ii) $\hat p_2 := \frac{S_n+1}{n+2}$ has the risk $r_\ell(\hat p_2,p) = \mathbb{E}_p(\hat p_2-p)^2 = \frac{1}{(n+2)^2}\,\mathbb{E}_p\,\big(\underbrace{S_n+1-(n+2)p}_{S_n-np+1-2p}\big)^2 = \frac{np(1-p)+(1-2p)^2}{(n+2)^2}$;



(a) $n = 1$ (b) $n = 100$

Figure 12.1: The risk in Example 12.3

(iii) The risk for the estimator $\hat{p}_3 := 2.20/65$ is $r_\ell(\hat{p}_3, p) = (p - 2.20/65)^2$;

(iv) The risk for the estimator $\hat{p}_4 := \frac{S_n + \sqrt{n}/2}{n + \sqrt{n}}$ is

$$r_\ell(\hat{p}_4, p) = \mathbb{E}_p\left(\frac{S_n + \sqrt{n}/2}{n + \sqrt{n}} - p\right)^2 = \mathbb{E}_p\left(\frac{S_n - np + \sqrt{n}(1/2 - p)}{n + \sqrt{n}}\right)^2$$
$$= \frac{np(1-p) + \frac{n}{4} - np + np^2}{(n + \sqrt{n})^2} = \frac{n}{4(n + \sqrt{n})^2},$$

which does *not* depend on p.[1]

Figure 12.1 visualizes the risk.

Note, that $r_\ell \xrightarrow[n\to\infty]{} 0$ for all estimators above except for $\hat{p}_3$ (indeed, $r_\ell = \mathcal{O}\left(1/n\right)$).

12.2 METHOD OF MOMENTS

Suppose that $\Theta \subset \mathbb{R}^d$. The jth-generalized moment is $\mathbb{E}_\vartheta\, \varphi_j(X)$, where φ_j are measurable functions; the classical method of moments involves the monomials $\varphi_j(x) := x^j$, $j = 1, \dots, d$. Further, let $h \colon \mathbb{R}^d \to \mathbb{R}^d$ be an (invertible) function (the identity, e.g.). The *moment estimator* $\hat{\vartheta} = \hat{\vartheta}(X) \in \Theta$ is chosen so that

$$h\left(\frac{1}{n}\sum_{i=1}^n \varphi_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^n \varphi_d(X_i)\right) = h\Big(\mathbb{E}_{\hat{\vartheta}}\, \varphi_1(X), \dots, \mathbb{E}_{\hat{\vartheta}}\, \varphi_d(X)\Big).$$

Example 12.4 (Binomial)**.** Choose $\varphi(x) = x$, $h(x) = x$ and $X_i \sim B(p)$ independent Bernoulli variables. Then the moment estimator $\hat{p}$ is $\frac{1}{n}\sum_{i=1}^n X_i = \mathbb{E}_{\hat{p}}\, X_i = \hat{p}$.

Example 12.5 (Cf. Exercise 12.3)**.** To estimate the parameters α and β of the Gamma distribution $\Gamma_{\alpha,\beta}$ one may use (5.8), the resulting moment estimators are

$$\hat{\alpha}(X) = \frac{\overline{X}_n^2}{s_n^2} \text{ and } \hat{\beta}(X) = \frac{\overline{X}_n}{s_n^2}.$$

[1] A decision rule with constant risk is called an *equalizer rule*.

Example 12.6 (Poisson)**.** The Poisson distribution P_λ with parameter $\lambda > 0$ (rate) has the probability mass function (pmf)

$$X \sim P_\lambda \colon P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{12.3}$$

For $X \sim P_\lambda$ it holds that $\mathbb{E}_\lambda X = \operatorname{var}_\lambda X = \lambda$ (cf. Exercise 12.1). Apparently, $\hat{\lambda} \coloneqq \overline{X}_n$ is the moment estimator (for the first moment), but $\hat{\lambda} \coloneqq s_n^2$ is an estimator (a different estimator) for the parameter λ too.

Example 12.7. Consider the Gauß loss function $\ell(x, y) = (x - y)^2$. Then the expected loss of $\hat{\vartheta}(X) = \hat{\lambda} = \overline{X}_n$ for the Poisson distribution (Example 12.6) is $r_\ell(\hat{\vartheta}, \vartheta) = \mathbb{E}_\vartheta \left(\overline{X}_n - \lambda\right)^2 = \operatorname{var} \overline{X}_n = \frac{\lambda}{n}$ (cf. Proposition 2.8). The risk for the estimator $\hat{\vartheta} \coloneqq s_n^2$ satisfies $\mathbb{E}\, s_n^2 = \lambda$, but $\operatorname{var} s_n^2 = \frac{1}{n}\left(\lambda + \left(3 - \frac{n-3}{n-1}\right)\lambda^2\right)$ (cf. Corollary 2.12). Note, that $\operatorname{var} \overline{X}_n < \operatorname{var} s_n^2$ and for this reason one would likely prefer the estimator $\overline{X}_n$ over s_n^2.

12.3 MINIMUM DISTANCE ESTIMATION

Let X_i be iid random sample from the population with cdf $F_\vartheta(\cdot)$, $\vartheta \in \Theta$. Recall the empirical distribution function of the observations $X_1, \dots, X_n$, $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i)$ (cf. (9.2)).

Definition 12.8. The minimum distance estimator $\hat{\vartheta}$ is

$$\hat{\vartheta} \in \operatorname*{arg\,min}_{\vartheta \in \Theta} d\big(F_n(\cdot),\, F_\vartheta(\cdot)\big).$$

Distances $d(\cdot, \cdot)$ for distribution functions include:

(i) Kolmogorov–Smirnov:

$$d\big(F(\cdot), F^*(\cdot)\big) \coloneqq \sup_{x \in \mathbb{R}^d} |F(x) - F^*(x)|\,; \tag{12.4}$$

(ii) Cramér[2]–von Mises[3]:

$$d\big(F(\cdot), F^*(\cdot)\big) \coloneqq \omega^2 \coloneqq \int_{\mathbb{R}^d} \big(F(x) - F^*(x)\big)^2 \,\mathrm{d}F^*(x); \tag{12.5}$$

(iii) Anderson[4]–Darling[5]:

$$d\big(F(\cdot), F^*(\cdot)\big) \coloneqq \int_{\mathbb{R}^d} \frac{\big(F(x) - F^*(x)\big)^2}{F^*(x)\big(1 - F^*(x)\big)} \,\mathrm{d}F^*(x). \tag{12.6}$$

12.4 MAXIMUM LIKELIHOOD

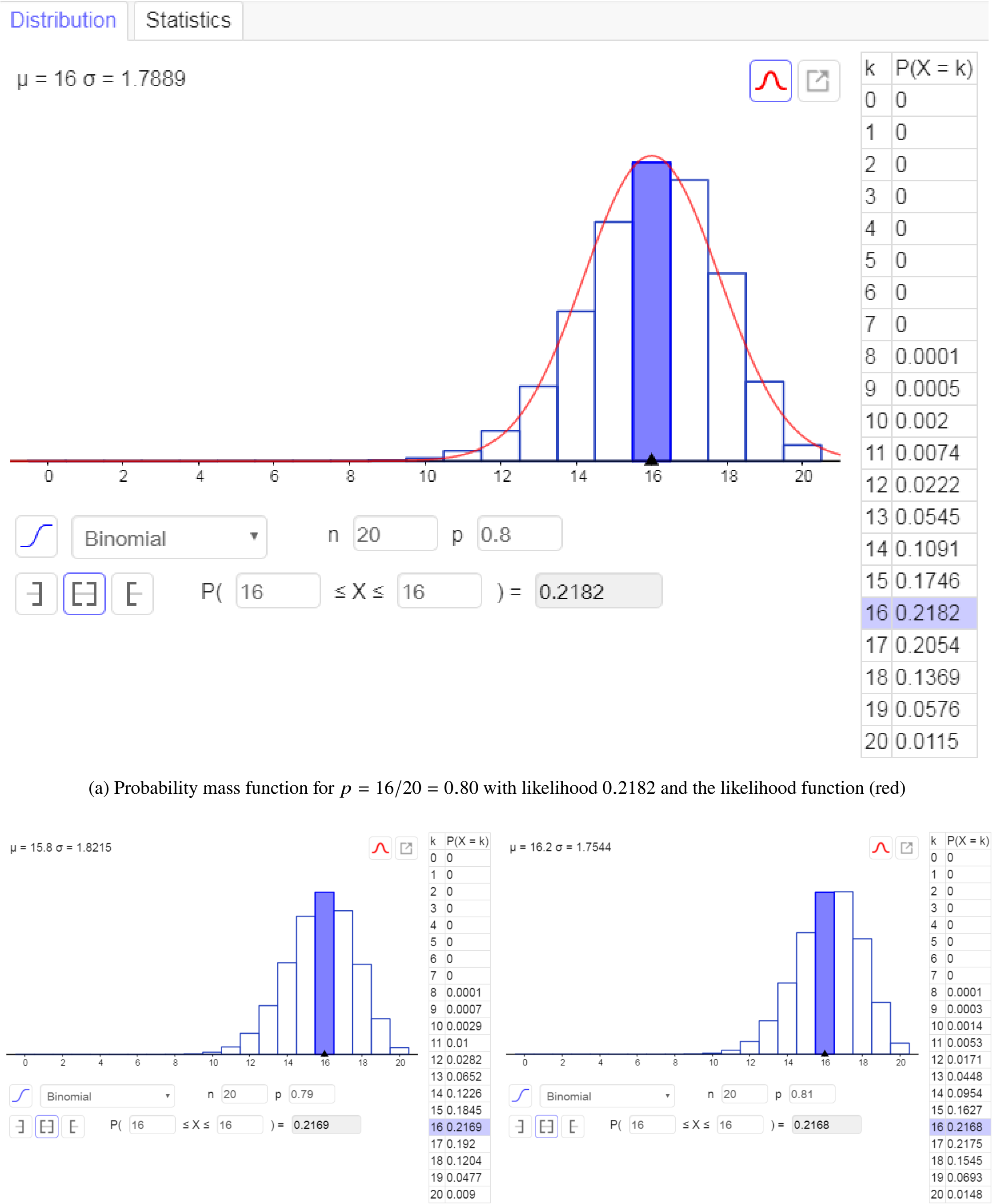
For a parametrization $\vartheta \in \Theta$ let $f_\vartheta(\cdot)$ be the pmf (if the distributions is discrete, $P_\vartheta(\{x\}) = f_\vartheta(x)$) or the density of P_ϑ (if the distribution is continuous, i.e., $P_\vartheta(\mathrm{d}x) = f_\vartheta(x)\,\mathrm{d}x$).

[2] Harald Cramér, 1893–1985, Swedish mathematician and actuary

[3] Richard von Mises, 1881–1973, Austrian-American mathematician

[4] Theodore Wilbur Anderson, 1918–2016, American statistician

[5] Donald Allan Darling, 1915–2014, American statistician



μ = 16 σ = 1.7889

Binomial n 20 p 0.8

P(16 ≤ X ≤ 16) = 0.2182

k	P(X = k)
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0.0001
9	0.0005
10	0.002
11	0.0074
12	0.0222
13	0.0545
14	0.1091
15	0.1746
16	0.2182
17	0.2054
18	0.1369
19	0.0576
20	0.0115

(a) Probability mass function for $p = 16/20 = 0.80$ with likelihood 0.2182 and the likelihood function (red)

μ = 15.8 σ = 1.8215

Binomial n 20 p 0.79

P(16 ≤ X ≤ 16) = 0.2169

k	P(X = k)
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0.0001
9	0.0007
10	0.0029
11	0.01
12	0.0282
13	0.0652
14	0.1226
15	0.1845
16	0.2169
17	0.192
18	0.1204
19	0.0477
20	0.009

μ = 16.2 σ = 1.7544

Binomial n 20 p 0.81

P(16 ≤ X ≤ 16) = 0.2168

k	P(X = k)
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0.0001
9	0.0003
10	0.0014
11	0.0053
12	0.0171
13	0.0448
14	0.0954
15	0.1627
16	0.2168
17	0.2175
18	0.1545
19	0.0693
20	0.0148

(b) pmf for $p = 0.79$ with likelihood 0.2169 at $X = 16$

(c) pmf for $p = 0.81$ with likelihood 0.2168 at $X = 16$

Figure 12.2: Screenshots from the freeware GeoGebra.
The likelihood attains its maximum at $\hat{p} = 80\,\%$ with $P_{\hat{p}}(S_{20} = 16) = 0.2182$

Definition 12.9. The *likelihood function* (cf. Ferguson [4, 5]) is

$$\Theta \ni \vartheta \mapsto L(\vartheta \mid x) \coloneqq f_\vartheta(x) \in \mathbb{R},$$

the function

$$\ell(\vartheta \mid x) \coloneqq \log f_\vartheta(x) = \log L(\vartheta \mid x)$$

is the *log likelihood function*.[6] Here, $f_\vartheta(\cdot)$ is the pdf for continuous, the pmf for discrete distributions.

Remark 12.10. Many notational variants are used in the literature for these functions, for example $f(x,\vartheta)$, $f(x \mid \vartheta)$, $L(\vartheta \mid x)$, $\ell(x \mid \vartheta)$ or $\mathcal{L}(\vartheta; x)$, etc., and the parameters are often interchanged as well.

Definition 12.11 (Maximum likelihood)**.** The *maximum likelihood estimator (MLE)* $\hat\vartheta(\cdot)$ satisfies

$$\hat\vartheta(X) \in \operatorname*{arg\,max}_{\vartheta\in\Theta} f_\vartheta(X) = \operatorname*{arg\,max}_{\vartheta\in\Theta} L(\vartheta \mid X) = \operatorname*{arg\,max}_{\vartheta\in\Theta} \ell(\vartheta \mid X).$$

Remark 12.12. The maximum likelikhood method was popularized by Ronald A. Fisher (1912, 1929), but it has been employed by Gauss, Laplace and Edgeworth earlier.

Definition 12.13 (Score)**.** The gradient of the log-likelihood function

$$V \coloneqq V(\vartheta, x) \coloneqq \nabla_\vartheta\, \ell(\vartheta \mid x) = \nabla_\vartheta\, \log f_\vartheta(x) = \frac{1}{f_\vartheta(x)} \cdot \frac{\partial}{\partial\vartheta} f_\vartheta(x) \tag{12.7}$$

is called *score* or *informant*. The score V indicates the sensitivity of $\vartheta \mapsto \ell(\vartheta \mid x)$.

Remark 12.14 (Relation to maximum likelihood). For independent data $X = (X_1, \dots, X_n)$, the maximum likelihood estimator $\hat\vartheta$ maximizes $\vartheta \mapsto \log \prod_{i=1}^n f_\vartheta(X_i) = \sum_{i=1}^n \log f_\vartheta(X_i)$. Provided sufficient smoothness, the maximum likelihood estimator $\hat\vartheta$ satisfies the first order condition

$$0 = \frac{1}{n} \sum_{i=1}^n V(\hat\vartheta, X_i). \tag{12.8}$$

Remark 12.15 (M-estimator). An estimator of the form

$$\hat\vartheta \in \operatorname*{arg\,min}_{\vartheta\in\Theta} \frac{1}{n} \sum_{i=1}^n \rho(X_i; \vartheta)$$

is called M-estimator (“M” for “maximum likelihood” type). For independent data, the maximum likelihood estimator is an M estimator. This classification was introduced by Peter J. Huber (cf. Footnote 9 on page 16).

Remark 12.16 (ψ-type estmator). An estimator is ψ type, if

(i) $\mathbb{E}\,\psi(X; \vartheta) = 0$ and

(ii) $\frac{1}{n} \sum_{i=1}^n \psi(X_i; \vartheta) = 0$.

Note the relation $\psi(x; \vartheta) = \frac{\partial}{\partial\vartheta} \rho(x; \vartheta)$ to M-estimators.

[6] Not to be confused with the loss function (regret function) in (12.1).

12.4.1 Discrete distributions

Example 12.17 (Binomial distribution, cf. Example 12.4 and Figure 12.2)**.** $X_1, \dots, X_n \sim B(p)$ are independent Bernoulli observations. What is p ($\vartheta = p$)?

The likelihood functions for $S_n := X_1 + \dots + X_n \sim \operatorname{bin}(n, p)$ is $L(p \mid X_1, \dots, X_n) = \binom{n}{S_n} p^{S_n}(1-p)^{n-S_n}$. To find $\hat{p}$ maximizing the likelihood differentiate with respect to p and set $0 = \frac{\partial}{\partial p} L(\hat{p} \mid X_1, \dots, X_n)$, i.e.,

$$0 = \binom{n}{S_n} \hat{p}^{S_n - 1}(1-\hat{p})^{n-S_n-1} \cdot \underbrace{\left(S_n(1-\hat{p}) - (n - S_n)\hat{p}\right)}_{=S_n - n\hat{p}}$$

and thus $\hat{p} = \frac{S_n}{n} = \overline{X}_n$.

Example 12.18 (Poisson distribution)**.** The values $X_1, \dots, X_n$ from independent observations of a Poisson distribution P_λ have been observed. What is a qualified guess for λ?

The likelihood function for the parameter $\vartheta = \lambda$ is $L(\lambda \mid X) = \frac{\lambda^{X_1} e^{-\lambda}}{X_1!} \cdot \ldots \frac{\lambda^{X_n} e^{-\lambda}}{X_n!} = \frac{\lambda^{X_1 + \dots + X_n} e^{-n\lambda}}{X_1! \cdot \ldots X_n!}$ and

$$\ell(\lambda \mid X) = (X_1 + \dots + X_n) \log \lambda - n\lambda - \log\left(X_1! \cdots X_n!\right)$$

The maximum likelihood function satisfies (differentiating with respect to λ)

$$(X_1 + \dots + X_n) \frac{1}{\hat{\lambda}} - n = 0, \text{ i.e., } \hat{\lambda} = \overline{X}_n.$$

12.4.2 Continuous distributions

Example 12.19. We are interested in estimating the parameter θ for the distribution with density $f_\theta(x) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x)$. The maximum likelihood estimator $\hat{\theta}$ for independent observations $X_1, \dots, X_n$ maximizes

$$\theta \mapsto \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i);$$

the maximum is attained at $\hat{\theta}(X) = X_{(n)} := \max_{i=1,\dots n} X_i$. Apparently it holds that $\hat{\theta} \le \theta$ (θ being the true parameter).

Example 12.20 (Normal distribution, variance known)**.** The likelihood function of n independent normal observations $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$ for the unknown μ is (cf. (2.5), Steiner)

$$L(\mu \mid X) = \frac{1}{\sqrt{2\pi\sigma_0^2}^{\,n}} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \mu)^2} = \frac{1}{\sqrt{2\pi\sigma_0^2}^{\,n}} e^{-\frac{n}{2\sigma_0^2}\left(V_n + (\overline{X}_n - \mu)^2\right)}$$

(see Figure 12.3). The likelihood $\mu \to L(\mu \mid X)$ attains its maximum at $\hat{\mu} = \overline{X}_n$.

Example 12.21 (Normal distribution)**.** Given some independent observations with common distribution $X_i \sim \mathcal{N}(\mu, \sigma^2)$, what are useful estimators for μ and σ^2?

The likelihood is

$$L(\mu, \sigma^2 \mid X) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right)$$

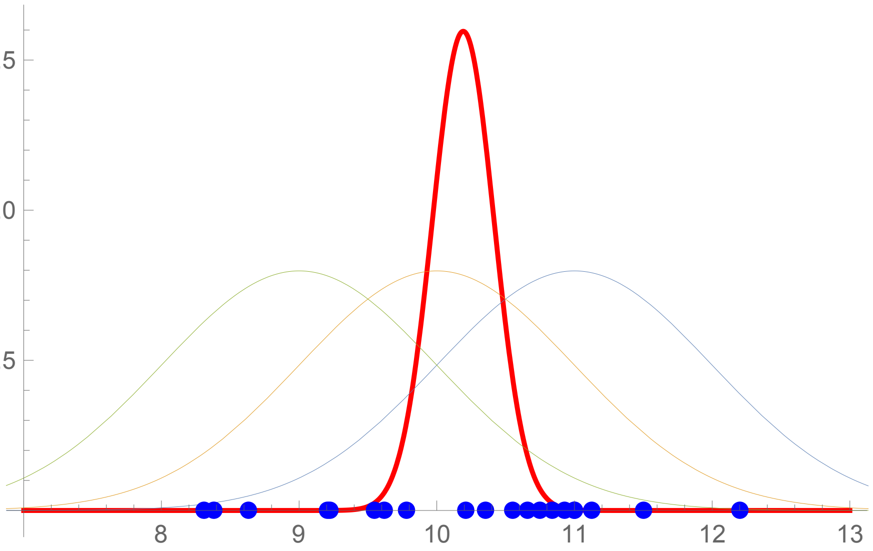


Figure 12.3: Samples (blue), three candidate pdfs ($x \mapsto f_{\mu_i}(x)$ for $\mu \in \{9, 10, 11\}$) and the likelihood function $\mu \mapsto f_\mu(x_1, \dots, x_n)$ (bold, red; cf. Example 12.20)

and thus

$$\begin{aligned}\ell(\mu, \sigma^2 \mid X) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2\right).\end{aligned}$$

Apparently, the maximum with respect to μ is attained at $\hat{\mu} = \overline{X}_n$. Differentiating with respect to the parameter σ^2 gives

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \overline{X}_n)^2, \text{ or } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = V_n = \frac{n-1}{n}s_n^2.$$

Example 12.22 (Censored data)**.** Independent observations of an exponential distribution E_λ are X_i, $i = 1, \dots n$, but only the censored data $\tilde{X}_i := \min(T, X_i)$ are accessible. We are interested in the average lifespan $\mathbb{E}\, X_i = {}^1\!/\!_\lambda$.

The maximum likelihood estimator considers the likelihood function

$$L(\lambda \mid X) = \prod_{i \in n_T} \lambda e^{-\lambda X_i} \cdot \prod_{i \notin n_T} e^{-\lambda T} \text{ or } \ell(\lambda \mid X) = \sum_{i \in n_T} (\log\lambda - \lambda X_i) - \sum_{i \notin n_T} \lambda T,$$

where $n_T := \left\{i \colon \tilde{X}_i < T\right\}$. Differentiating with respect to the parameter λ gives $0 = \sum_{i \in n_T}\left(\frac{1}{\lambda} - X_i\right) - \sum_{i \notin n_T} T$ and thus

$$\frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^n \min(X_i, T)}{|n_T|} = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{X}_i}{{}^{|n_T|}\!/\!_n} = \frac{n}{|n_T|} \cdot \overline{\tilde{X}}_n.$$

Note, that the estimator which results from removing all observations $\{i \colon X_i \geq T\}$ satisfies $\frac{\sum_{i \in n_T} X_i}{|n_T|} \leq T$ and is therefore rather useless.

12.5 GENERAL PROPERTIES OF ESTIMATORS

12.5.1 Bias

Definition 12.23. Let $\hat{\vartheta}(\cdot)$ be an estimator for ϑ and $\gamma\colon \Theta \to \mathbb{R}^d$ a function. We consider the estimator $\hat{\gamma}(\cdot)$ for $\gamma(\vartheta)$.

(i) The *bias*[7] of $\hat{\gamma}$ at ϑ is

$$\operatorname{bias}_\vartheta \hat{\gamma} := \mathbb{E}_\vartheta \hat{\gamma} - \gamma(\vartheta).$$

The estimator $\hat{\vartheta}(\cdot)$ ($\hat{\gamma}(\cdot)$, resp.) is an *unbiased estimator*[8] for ϑ ($\gamma(\vartheta)$, resp.), if

$$\vartheta = \mathbb{E}_\vartheta \hat{\vartheta} \qquad (\mathbb{E}_\vartheta \hat{\gamma} = \gamma(\vartheta),\ \text{resp.}).$$

(ii) The estimator $\hat{\gamma}(\cdot)$ is the *minimum-variance unbiased estimator (MVUE)* if

$$\operatorname{var}_\vartheta \hat{\gamma} \le \operatorname{var}_\vartheta \hat{\gamma}'$$

for every unbiased estimator $\hat{\gamma}'(\cdot)$.

(iii) The mean squared error (MSE)[9] is (cf. (12.1) and Exercise 12.11)

$$\operatorname{mse}_\vartheta \hat{\gamma} := \mathbb{E}_\vartheta \big(\hat{\gamma} - \gamma(\vartheta)\big)^2 = \operatorname{var}_\vartheta \hat{\gamma} + (\operatorname{bias}_\vartheta \hat{\gamma})^2. \tag{12.9}$$

Remark 12.24 (Bias–variance tradeoff). The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

12.5.2 Comparison of estimators

Definition 12.25 (Cf. Rüschendorf [17, Section 2.2], Pflug [13])**.** Let r be a risk function and $\hat{\vartheta}(\cdot)$ an estimator for ϑ.

(i) $\hat{\vartheta}_1$ is *at least as good* as $\hat{\vartheta}_2$ ($\hat{\vartheta}_1 \preceq \hat{\vartheta}_2$), if $r\left(\hat{\vartheta}_1, \vartheta\right) \le r\left(\hat{\vartheta}_2, \vartheta\right)$ for all $\vartheta \in \Theta$.

(ii) $\hat{\vartheta}_1$ is *better* than $\hat{\vartheta}_2$ ($\hat{\vartheta}_1 \prec \hat{\vartheta}_2$), if $\hat{\vartheta}_1 \preceq \hat{\vartheta}_2$ and there exists at least one $\vartheta_0 \in \Theta$ so that $r\left(\hat{\vartheta}_1, \vartheta_0\right) < r\left(\hat{\vartheta}_2, \vartheta_0\right)$.

(iii) The estimator $\hat{\vartheta}^*$ is *admissible*,[10] if there is no better estimator.

(iv) The estimator $\hat{\vartheta}^*$ is *optimal* with respect to the class $\mathcal{C}$, if $\hat{\vartheta}^* \preceq \hat{\vartheta}$ for all $\hat{\vartheta} \in \mathcal{C}$.

(v) The estimator $\hat{\vartheta}^+$ is *minimax*, if $\sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}, \vartheta\right) \ge \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}^+, \vartheta\right)$ for every other estimator $\hat{\vartheta}$.

Remark 12.26. The minimax estimator $\hat{\vartheta}^+$ satisfies $\sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}^+(\cdot), \vartheta\right) = \inf_{\hat{\vartheta}} \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}(\cdot), \vartheta\right)$ and thus takes precaution against the worst.

[7] Verzerrung

[8] erwartungstreu, unverzerrt

[9] mittlere quadratische Abweichung

[10] zulässig

Lemma 12.27. *If $\hat{\vartheta}^+$ is admissible with constant risk, then $\hat{\vartheta}^+$ is minimax.*

Proof. Denote the constant risk by $c := r\big(\hat{\vartheta}^+, \vartheta\big)$. If $\hat{\vartheta}^+$ were not minimax, then, by Remark 12.26, there is an estimator $\hat{\vartheta}$ so that $\sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}, \vartheta\right) < \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}^+, \vartheta\right) = c$. Hence, $\hat{\vartheta}^+$ is not admissible, a contradiction. □

12.6 BAYES ESTIMATOR

Bayesian statistics involves the following steps:

(i) Define the *prior distribution* (measure $\pi(\cdot)$) that incorporates your *subjective beliefs* about a parameter ϑ. The prior represents the initial believe about ϑ. The prior can be informative or uninformative.[11] The prior π is a measure, not necessarily a probability measure, though.

(ii) Gather data.

(iii) Update your *prior distribution* with the data using Bayes' theorem to obtain a posterior distribution. The posterior distribution is a probability distribution that represents your *updated beliefs* about the parameter after having seen the data.

Prior distribution The prior distribution is a measure on the parameters, $\pi(\mathrm{d}\vartheta)$.

Definition 12.28 (Bayes estimator)**.** Let ϑ follow a prior distribution with measure $\pi(\mathrm{d}\vartheta)$ and $\hat{\vartheta}\colon \mathcal{X} \to \Theta$ be a decision rule. The *Bayes risk* is

$$r\big(\hat{\vartheta}(\cdot), \pi\big) := \int_\Theta r_\ell\big(\hat{\vartheta}(\cdot), \vartheta\big)\, \pi(\mathrm{d}\vartheta) = \int_\Theta \underbrace{\int_{\mathcal{X}} \ell\left(\hat{\vartheta}(x), \vartheta\right) P_\vartheta(\mathrm{d}x)}_{\mathbb{E}_\vartheta\, \ell\left(\hat{\vartheta}(\cdot), \vartheta\right)} \pi(\mathrm{d}\vartheta). \tag{12.10}$$

Remark 12.29. Note that $r\big(\hat{\vartheta}(\cdot), \pi\big) = r_\ell\big(\hat{\vartheta}(\cdot), \vartheta\big)$ for the prior measure $\pi = \delta_\vartheta$.

Definition 12.30. The *Bayes rule* or *Bayes estimator* with respect to the prior π is the decision rule $\hat{\vartheta}^\pi(\cdot)$ that minimizes the Bayes risk (12.10), i.e.,

$$r\big(\hat{\vartheta}^\pi(\cdot),\, \pi\big) \le r\big(\hat{\vartheta}(\cdot),\, \pi\big) \tag{12.11}$$

for every decision rule $\hat{\vartheta}(\cdot)$.

Remark 12.31. Different notations for the Bayes estimator in frequent use include $\hat{\vartheta}(\cdot) = \hat{\vartheta}^\pi(\cdot) = \delta(\cdot)$; the second relates to the prior π.

Lemma 12.32. *Suppose that $\hat{\vartheta}^\pi(\cdot)$ is a Bayes decision rule with respect to the prior distribution π. If the risk function of $\hat{\vartheta}^\pi(\cdot)$ satisfies*

$$r\big(\hat{\vartheta}^\pi(\cdot),\, \vartheta\big) \le r\big(\hat{\vartheta}^\pi(\cdot),\, \pi\big) \qquad \textit{for all } \vartheta \in \Theta, \tag{12.12}$$

then $\hat{\vartheta}^\pi(\cdot)$ is a minimax decision rule.

Proof. Suppose that $\hat{\vartheta}^\pi(\cdot)$ is not minimax. Then there is a decision rule $\hat{\vartheta}(\cdot)$ such that $\sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}, \vartheta\right) < \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}^\pi, \vartheta\right)$. It follows that

$$r\left(\hat{\vartheta}(\cdot),\, \pi\right) \le \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta},\, \vartheta\right) < \sup_{\vartheta\in\Theta} r\left(\hat{\vartheta}^\pi,\, \vartheta\right) \underset{(12.12)}{\le} r\left(\hat{\vartheta}^\pi(\cdot),\, \pi\right),$$

contradicting the statement that $\hat{\vartheta}^\pi$ is Bayes with respect to π, cf. (12.11). Hence $\hat{\vartheta}^\pi$ is minimax. □

[11] Vorbewertung

Posterior distributions and disintegration. Assume that the prior π has a density (which we denote again as π), i.e., $\pi(\mathrm{d}\vartheta) = \pi(\vartheta)\,\mathrm{d}\vartheta$. Then the Bayes risk is

$$r\left(\hat{\vartheta}(\cdot), \pi\right) = \iint_{\Theta\times\mathcal{X}} \ell\Big(\hat{\vartheta}(x), \vartheta\Big) P_\vartheta(\mathrm{d}x)\, \pi(\mathrm{d}\vartheta) = \iint_{\Theta\times\mathcal{X}} \ell\Big(\hat{\vartheta}(x), \vartheta\Big)\, \underbrace{f_\vartheta(x)\, \pi(\vartheta)}_{\pi(x,\vartheta)}\, \mathrm{d}x\, \mathrm{d}\vartheta.$$

Note, that $\pi(x, \vartheta) \coloneqq f_\vartheta(x)\,\pi(\vartheta)$ is a density, as $\iint_{\Theta\times\mathcal{X}} \pi(x, \vartheta)\,\mathrm{d}x\,\mathrm{d}\vartheta = 1$. By Fubini's theorem we thus have

$$r\left(\hat{\vartheta}(\cdot), \pi\right) = \int_{\mathcal{X}} \left(\int_\Theta \ell\left(\hat{\vartheta}(x), \vartheta\right)\, \pi(\vartheta \mid x)\,\mathrm{d}\vartheta\right) f(x)\,\mathrm{d}x \tag{12.13}$$

where $\pi(\vartheta \mid x) = \pi(x, \vartheta)/f(x)$ and the marginal density $f_\pi(x) \coloneqq \int_\Theta \pi(x, \vartheta')\,\mathrm{d}\vartheta'$, cf. (1.11).

Definition 12.33. The distribution with density

$$\pi(\vartheta \mid x) \coloneqq \frac{\pi(x, \vartheta)}{\int_\Theta \pi(x, \vartheta')\,\mathrm{d}\vartheta'} = \frac{f_\vartheta(x)\,\pi(\vartheta)}{\int_\Theta f_{\vartheta'}(x)\,\pi(\vartheta')\,\mathrm{d}\vartheta'} \tag{12.14}$$

is the *posterior distribution*.

$\pi(\vartheta \mid x)$ represents the most up-to-date belief in ϑ and is all that is needed for inference. Note, that (12.14) is

$$\pi(\vartheta \mid x) \propto f_\vartheta(x) \times \pi(\vartheta),$$

often stated as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Remark 12.34 (Disintegration). In view of (12.13) we have that

$$\begin{aligned}\min_{\hat{\vartheta}(\cdot)} \int_{\mathcal{X}} \int_\Theta \ell\left(\hat{\vartheta}(x), \vartheta'\right)\, \pi(\mathrm{d}\vartheta' \mid x)\, f_\pi(x)\,\mathrm{d}x \\ = \int_{\mathcal{X}} \left(\min_\vartheta \int_\Theta \ell\left(\vartheta, \vartheta'\right)\, \pi(\mathrm{d}\vartheta' \mid x)\right) f_\pi(x)\,\mathrm{d}x,\end{aligned}$$

i.e., the Bayes estimator $\hat{\vartheta}(x)$ (cf. Definition 12.28) is a (the) minimizer

$$\begin{aligned}\hat{\vartheta}(x) &\in \operatorname*{arg\,min}_{\vartheta\in\Theta} \int_\Theta \ell(\vartheta, \vartheta')\, \pi(\mathrm{d}\vartheta' \mid x) \\ &= \operatorname*{arg\,min}_{\vartheta\in\Theta} \int_\Theta \ell(\vartheta, \vartheta')\, f_{\vartheta'}(x)\, \pi(\mathrm{d}\vartheta').\end{aligned} \tag{12.15}$$

Note, that in (12.15) the density $\pi(\cdot \mid x)$ is replaced by $f_\vartheta(x)\,\pi(\cdot)$, the denominator $f(x)$ in (12.14) does not depend on ϑ. The replacement does *not* integrate to 1, but this does not change the minimizer in (12.10). This measure, which does not integrate to 1, is called an *improper prior*.

Definition 12.35. The quantity

$$\int_\Theta \ell(\vartheta, \vartheta')\, \pi(\mathrm{d}\vartheta' \mid x)$$

is the *Bayesian expected loss* or *posterior expected loss*.

Example 12.36 (Posterior mean, cf. the Laplace approximation, Section 12.7 below)**.** Suppose that $\ell(\vartheta,\vartheta')=(\vartheta-\vartheta')^2$. Following (12.15) we minimize $\vartheta\mapsto\int_\Theta(\vartheta-\vartheta')^2\,\pi(\mathrm{d}\vartheta'\mid X)$. Differentiating with respect to ϑ gives $2\int_\Theta(\vartheta-\vartheta')\,\pi(\mathrm{d}\vartheta'\mid X)=0$, the Bayes estimator thus is the posterior mean

$$\hat{\vartheta}(X)=\int_\Theta\vartheta'\,\pi(\mathrm{d}\vartheta'\mid X).\tag{12.16}$$

Example 12.37 (Posterior median)**.** The median of $\pi(\cdot\mid X)$ is the Bayes estimate with respect to the modular loss (i.e., absolute value loss), cf. Proposition 1.25.

Example 12.38 (Posterior mode)**.** The mode of $\pi(\cdot\mid X)$ is the Bayes estimate with respect to the zero-one loss, cf. (12.2); this is, however, only a useful idea for a discrete distribution. Indeed, the loss function is $\ell_{\epsilon=0}\big(\hat{\vartheta}(X),\vartheta\big)=\begin{cases}1 & \text{if } \hat{\vartheta}(X)\neq\vartheta,\\ 0 & \text{if } \hat{\vartheta}(X)=\vartheta.\end{cases}$ Assume that $\hat{\vartheta}(X)=\vartheta^*$. Then

$$\sum_{\vartheta\in\Theta}\ell_{\epsilon=0}\big(\hat{\vartheta}(X),\vartheta\big)\,\pi(\vartheta\mid X)=\sum_{\vartheta\neq\vartheta^*}\pi(\vartheta\mid X)=1-\pi(\vartheta^*\mid X).$$

Minimizing means maximizing $\pi(\vartheta^*\mid X)$, so $\hat{\vartheta}(X)=\vartheta^*$ is taken to be the most likely value, the mode of $\pi(\cdot\mid X)$.

Example 12.39. We want to estimate p of a distribution $\operatorname{bin}(n,p)$. Assume the prior is $\vartheta=p\sim\operatorname{Beta}(\alpha,\beta)$ (Beta distribution; cf. Exercise 5.10 for the distribution and (5.10) for the Beta function). The Bayes estimator $\hat{p}(\cdot)$ for the Binomial $\operatorname{bin}(n,p)$ distribution minimizes

$$\begin{aligned}&\int_0^1\underbrace{\sum_{k=0}^n\binom{n}{k}p^k(1-p)^{n-k}\big(\hat{p}(k)-p\big)^2}_{\mathbb{E}_p(\hat{p}-p)^2}\cdot\underbrace{\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p}_{\text{prior }\pi(\mathrm{d}p)}\\
&=\sum_{k=0}^n\frac{1}{B(\alpha,\beta)}\binom{n}{k}\int_0^1p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\Big(\hat{p}(k)^2-2p\cdot\hat{p}(k)+p^2\Big)\,\mathrm{d}p &(12.17)\\
&=\sum_{k=0}^n\frac{1}{B(\alpha,\beta)}\binom{n}{k}\begin{pmatrix}\hat{p}(k)^2\cdot B(k+\alpha,n-k+\beta)\\ -2\hat{p}(k)\cdot B(k+\alpha+1,n-k+\beta)\\ +B(k+\alpha+2,n-k+\beta)\end{pmatrix}. &(12.18)\end{aligned}$$

Computing $\hat{p}(k)$ for every k individually in (12.18) and (12.19) gives rise to disintegrating (12.10). To identify the optimal estimator $\hat{p}(\cdot)$ we take the derivative and get the equations (for every k individually!)

$$0=2\hat{p}(k)\cdot B(k+\alpha,n-k+\beta)-2B(k+\alpha+1,n-k+\beta),\tag{12.19}$$

i.e. (use (5.10) again),

$$\hat{p}(k)=\frac{B(k+\alpha+1,n-k+\beta)}{B(k+\alpha,n-k+\beta)}=\frac{\frac{\Gamma(k+\alpha+1)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta+1)}}{\frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)}}=\frac{k+\alpha}{n+\alpha+\beta}.$$

Note, that the estimator (ii) (estimator (iv), resp.) of Example 12.3 is a special cases with particular prior $\alpha=\beta=1$ (i.e., a uniform prior) ($\alpha=\beta=\sqrt{n}/2$, resp.).

Example 12.40 (Continuation of Example 12.39)**.** The posterior distribution in Example 12.39 for $S_n \sim \operatorname{bin}(n,p)$ with prior $p \sim \operatorname{Beta}(\alpha,\beta)$ is

$$\begin{aligned}\pi(p \mid k) &= \frac{f_p(k)\pi(p)}{\int f_{p'}(k)\pi(p')\,\mathrm{d}p'} = \frac{p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p'^k(1-p')^{n-k}p'^{\alpha-1}(1-p')^{\beta-1}\,\mathrm{d}p'}\\ &= \frac{p^{k+\alpha-1}(1-p)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)} \sim B(\alpha+k, \beta+n-k).\end{aligned}$$

Now the objective (12.15) with improper prior is $\int_0^1 (\hat{p}(k)-p)^2 \cdot p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\,\mathrm{d}p$. Compare with (12.17) and (12.19) to see that the Bayes estimator is $\hat{p}(k) = \frac{k+\alpha}{n+\alpha+\beta}$.

12.7 LAPLACE APPROXIMATION

Let $h(\vartheta) := \log p(\vartheta)$ with maximum at ϑ_0, then $p'(\vartheta_0) = 0$, $h'(\vartheta_0) = \frac{p'(\vartheta_0)}{p(\vartheta_0)} = 0$ and $h''(\vartheta_0) < 0$ and the Taylor series expansion is $h(\vartheta) \approx h(\vartheta_0) + \frac{1}{2}(\vartheta-\vartheta_0)^\top h''(\vartheta_0)(\vartheta-\vartheta_0)$. It follows that $p(\vartheta) = \exp\big(h(\vartheta)\big) \approx \exp\big(h(\vartheta_0)\big) \cdot \exp\Big(-\frac{1}{2}(\vartheta-\vartheta_0)^\top\big(-h''(\vartheta_0)\big)(\vartheta-\vartheta_0)\Big)$ so that a random variable with density p is locally well approximated by $Y \sim \mathcal{N}\big(\vartheta_0, -h''(\vartheta_0)^{-1}\big)$.

Now let ϑ be the posterior mode of the density $\vartheta \mapsto p(\vartheta \mid X)$ and consider the estimator (12.16). Then $\hat{\vartheta} = \int_\Theta \vartheta' \cdot p(\vartheta' \mid X)\,\mathrm{d}\vartheta' \approx \mathbb{E}\,Y = \vartheta_0$ so that the posterior mode is a good approximation of the posterior mean.

12.8 FISHER INFORMATION

Recall the notational convenience discussed in Definition 12.9. In this section we assume that every f_ϑ is regular enough, for example differentiable with respect to the parameters ϑ, etc.

Recall the definition of the score function $V(\vartheta,x) = \nabla_\vartheta \log f_\vartheta(x)$ (Definition 12.13).

Proposition 12.41 (Properties of the score)**.** *It holds that*

$$\mathbb{E}_\vartheta\, V(\vartheta,\cdot) = 0. \tag{12.20}$$

Proof. Indeed,

$$\begin{aligned}\mathbb{E}_\vartheta\, V &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(x)\right) f_\vartheta(x)\,\mathrm{d}x = \int_{\mathcal{X}} \frac{1}{f_\vartheta(x)}\frac{\partial f_\vartheta(x)}{\partial\vartheta} \cdot f_\vartheta(x)\,\mathrm{d}x\\ &= \int_{\mathcal{X}} \frac{\partial f_\vartheta(x)}{\partial\vartheta}\,\mathrm{d}x = \frac{\partial}{\partial\vartheta}\int_{\mathcal{X}} f_\vartheta(x)\,\mathrm{d}x = \frac{\partial}{\partial\vartheta}1 = 0.\end{aligned}$$

□

Remark 12.42 (Relation to maximum likelihood). Note, that (12.8) is the empirical version of the property (12.20). The maximum likelihood estimator thus is a moment estimator for the score function V.

Definition 12.43. The *Fisher information* is the variance of the score,

$$I(\vartheta) := \operatorname{var}_\vartheta V. \tag{12.21}$$

Proposition 12.44 (Properties of the Fisher information)**.** *It holds that*

$$I(\vartheta) = \mathbb{E}_\vartheta \left(\frac{\partial}{\partial\vartheta} \log f_\vartheta(X)\right)^2 = -\mathbb{E}_\vartheta \frac{\partial^2}{\partial\vartheta^2} \log f_\vartheta(X). \tag{12.22}$$

Proof. It follows from (12.20) and (12.7) that

$$I(\vartheta) = \operatorname{var}_\vartheta V = \mathbb{E}_\vartheta V^2 = \mathbb{E}_\vartheta \left(\frac{\partial}{\partial\vartheta} \log f_\vartheta(X)\right)^2. \tag{12.23}$$

By the quotient rule we have

$$\frac{\partial^2}{\partial\vartheta^2} \log f_\vartheta(x) = \frac{\partial}{\partial\vartheta} \frac{\frac{\partial}{\partial\vartheta} f_\vartheta(x)}{f_\vartheta(x)} = \frac{\frac{\partial^2}{\partial\vartheta^2} f_\vartheta(x)}{f_\vartheta(x)} - \left(\frac{\frac{\partial}{\partial\vartheta} f_\vartheta(x)}{f_\vartheta(x)}\right)^2 = \frac{\frac{\partial^2}{\partial\vartheta^2} f_\vartheta(x)}{f_\vartheta(x)} - \left(\frac{\partial}{\partial\vartheta} \log f_\vartheta(x)\right)^2.$$

Now note that $\mathbb{E}_\vartheta \frac{\frac{\partial^2}{\partial\vartheta^2} f_\vartheta(X)}{f_\vartheta(X)} = \int_{\mathcal{X}} \frac{\partial^2}{\partial\vartheta^2} f_\vartheta(x)\,\mathrm{d}x = \frac{\partial^2}{\partial\vartheta^2} \int_{\mathcal{X}} f_\vartheta(x)\,\mathrm{d}x = \frac{\partial^2}{\partial\vartheta^2} 1 = 0$ as above, so the result follows by taking expectations with (12.23). □

Example 12.45. The Fisher information of a binomial $\operatorname{bin}(n, p)$ trial for the unknown $p \in (0, 1)$ is $I(p) = \frac{n}{p(1-p)}$.

Indeed, by (12.22),

$$\begin{aligned} I(p) &= -\mathbb{E}_p \frac{\partial^2}{\partial p^2} \log\left(\binom{n}{X} p^X (1-p)^{n-X}\right) = -\mathbb{E}_p \frac{\partial^2}{\partial p^2}\left(\log\binom{n}{X} + X\log p + (n-X)\log(1-p)\right) \\ &= \mathbb{E}_p \frac{X}{p^2} + \frac{n-X}{(1-p)^2} = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}. \end{aligned} \tag{12.24}$$

Remark 12.46. It follows from (12.22) that the Fisher information scales with the number of independent observations (cf. (12.24)).

12.9 INFORMATION INEQUALITIES

Theorem 12.47 (Cramér–Rao[12] inequality)**.** *Let $\hat{\gamma}(\cdot)$ be an unbiased estimator for $\gamma(\vartheta)$, i.e., $\mathbb{E}_\vartheta \hat{\gamma} = \gamma(\vartheta)$, then*

$$\operatorname{var}_\vartheta \hat{\gamma} \geq \frac{\gamma'(\vartheta)^2}{I(\vartheta)}.$$

Proof. As $\hat{\gamma}(\cdot)$ is unbiased it holds that

$$0 = \mathbb{E}_\vartheta \hat{\gamma} - \gamma(\vartheta) = \int_{\mathcal{X}} \big(\hat{\gamma}(x) - \gamma(\vartheta)\big) f_\vartheta(x)\,\mathrm{d}x.$$

Differentiate the latter to get

$$\begin{aligned} 0 = \frac{\partial}{\partial\vartheta} \int_{\mathcal{X}} \big(\hat{\gamma}(x) - \gamma(\vartheta)\big) f_\vartheta(x)\,\mathrm{d}x &= -\int_{\mathcal{X}} \gamma'(\vartheta) f_\vartheta(x)\,\mathrm{d}x + \int_{\mathcal{X}} \big(\hat{\gamma}(x) - \gamma(\vartheta)\big) \frac{\partial}{\partial\vartheta} f_\vartheta(x)\,\mathrm{d}x \\ &= -\gamma'(\vartheta) + \int_{\mathcal{X}} \big(\hat{\gamma}(x) - \gamma(\vartheta)\big) f_\vartheta(x) \frac{\partial}{\partial\vartheta} \log f_\vartheta(x)\,\mathrm{d}x. \end{aligned}$$

[12] Calyampudi Radhakrishna Rao, 1920, Indian-American mathematician and statistician

By the Cauchy–Schwarz inequality, thus

$$\begin{aligned}\gamma'(\vartheta) &= \int_X \big(\hat\gamma(x)-\gamma(\vartheta)\big)\sqrt{f_\vartheta(x)}\cdot\sqrt{f_\vartheta(x)}\frac{\partial}{\partial\vartheta}\log f_\vartheta(x)\,\mathrm{d}x\\ &\le \sqrt{\int_X \big(\hat\gamma(x)-\gamma(\vartheta)\big)^2 f_\vartheta(x)\,\mathrm{d}x}\cdot\sqrt{\int_X f_\vartheta(x)\left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(x)\right)^2\mathrm{d}x}\\ &= \sqrt{\operatorname{var}_\vartheta\hat\gamma}\cdot\sqrt{I(\vartheta)},\end{aligned}$$

where we have used that $\gamma(\vartheta)=\mathbb{E}_\vartheta\,\hat\gamma$ and (12.22). This completes the proof. □

The Cramér–Rao inequality is of particular importance to estimate the parameter itself, $\gamma(\vartheta) := \vartheta$.

Corollary 12.48 (Cramér–Rao bound)**.** *Let $\hat\vartheta(\cdot)$ be an unbiased estimator for ϑ, i.e. $\mathbb{E}_\vartheta\,\hat\vartheta=\vartheta$, then*

$$\operatorname{var}_\vartheta\hat\vartheta \ge \frac{1}{I(\vartheta)}.$$

Corollary 12.49. *Let $\hat\gamma(\cdot)$ be a statistic and set $\gamma(\vartheta) := \mathbb{E}_\vartheta\,\hat\gamma$. Then*

$$\operatorname{cov}_\vartheta\hat\gamma \ge \nabla_\vartheta\gamma(\vartheta)\cdot I(\vartheta)^{-1}\cdot\big(\nabla_\vartheta\gamma(\vartheta)\big)^\top,$$

where $I(\vartheta)_{j,k}=\mathbb{E}_\vartheta\,\frac{\partial}{\partial\vartheta_j}\log f_\vartheta(X)\cdot\frac{\partial}{\partial\vartheta_k}\log f_\vartheta(X) = -\,\mathbb{E}_\vartheta\,\frac{\partial^2}{\partial\vartheta_j\,\partial\vartheta_k}\log f_\vartheta(X)$ is the Fisher information matrix.

Here, $A\ge B$ is the Loewner order and understood to mean that $A-B$ is positive semidefinite.

12.10 SEQUENCES OF ESTIMATORS

Definition 12.50. Let n denote the sample size and let $\hat\vartheta_n(\cdot)$ ($\hat\gamma_n(\cdot)$, resp.) be a sequence of estimators for ϑ ($\gamma(\vartheta)$, resp.).

(i) The sequence of estimators $\hat\vartheta_n$ is a *(weakly) consistent estimator* for ϑ if $\hat\vartheta_n\to\vartheta$ in probability, i.e., $P_\vartheta\left(\left|\hat\vartheta_n-\vartheta\right|>\varepsilon\right)\xrightarrow[n\to\infty]{}0$ for every $\varepsilon>0$;

(ii) $\hat\gamma_n\xrightarrow{\text{a.s.}}\gamma(\vartheta)$, if $P_\vartheta\left(\lim_{n\to\infty}\left|\hat\gamma_n-\gamma(\vartheta)\right|=0\right)=1$;

(iii) The estimator $\hat\vartheta_n$ is *asymptotically normal*, if $\sqrt{n}\left(\hat\vartheta_n-\vartheta\right)\xrightarrow{d}\mathcal{N}(0,V)$ for some variance V.

Remark 12.51. Consistency can often be insured by employing the Markov inequality $P_\vartheta\left(h(\vartheta_n-\vartheta)\ge\varepsilon\right)\le\frac{\mathbb{E}_\vartheta\,h(\vartheta_n-\vartheta)}{\varepsilon}$, for example $h(\cdot)=|\cdot|$ or $h(\cdot)=|\cdot|^2$.

12.11 ASYMPTOTIC NORMALITY AND OPTIMALITY OF THE MAXIMUM LIKELIHOOD ESTIMATOR

The following theorem demonstrates that the Cramér–Rao bound is (asymptotically) sharp for the maximum likelihood estimator.

Theorem 12.52 (Asymptotic normality of the maximum likelihood estimator)**.** *Suppose that*

(i) $\vartheta\in$ int Θ,

(ii) $f_\vartheta(x) > 0$ *and is twice continuously differentiable in* ϑ *in a neighborhood* $N \ni \vartheta$*,*

(iii) $\int \sup_{\vartheta\in N} \|\nabla_\vartheta f_\vartheta(x)\|\,\mathrm{d}x < \infty$ *and* $\int \sup_{\vartheta\in N} \|\nabla_{\vartheta\vartheta} f_\vartheta(x)\|\,\mathrm{d}x < \infty$*,*

(iv) The Fisher information matrix $I(\vartheta) := \mathbb{E}\,(\nabla_\vartheta \log f_\vartheta(x))\,(\nabla_\vartheta \log f_\vartheta(x))^\top$ *exists and is nonsingular,*

(v) $\mathbb{E}\sup_{\vartheta\in N} \|\nabla_\vartheta \log f_\vartheta(x)\| < \infty$.

Then the maximum likelihood estimator $\hat\vartheta_n$ *satisfies*

$$\sqrt{n}\left(\hat\vartheta_n - \vartheta\right) \xrightarrow[n\to\infty]{} \mathcal{N}\left(0, I(\vartheta)^{-1}\right),$$

where $I(\vartheta)$ *is the Fisher information.*

Sketch of the proof. The first order conditions for the maximum likelihood estimator for $\vartheta \in \operatorname{int}\Theta$ (the interior of Θ) reads

$$\nabla_\vartheta \ell(\hat\vartheta_n \mid X) = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\vartheta}\log f_\vartheta(X_i)\Big|_{\vartheta=\hat\vartheta_n} = 0.$$

The Taylor series expansion of the (sufficiently smooth) score function $\vartheta \mapsto V(\vartheta; X)$ around the true parameter ϑ for the maximum likelihood estimator $\hat\vartheta$ is $V(\hat\vartheta) = V(\vartheta) + V'(\tilde\vartheta)(\hat\vartheta - \vartheta)$, i.e.,

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\vartheta}\log f_\vartheta(X_i) + \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\vartheta^2}\log f_\vartheta(X_i)\Big|_{\vartheta=\tilde\vartheta_n} \cdot \left(\hat\vartheta_n - \vartheta\right),$$

where $\tilde\vartheta_n$ is a point intermediate between ϑ and $\hat\vartheta_n$ and thus

$$\sqrt{n}\left(\hat\vartheta_n - \vartheta\right) = \underbrace{\left(-\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\vartheta^2}\log f_\vartheta(X_i)\Big|_{\vartheta=\tilde\vartheta_n}\right)^{-1}}_{\to I(\vartheta)^{-1}} \cdot \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\vartheta}\log f_\vartheta(X_i)}_{\sim\mathcal{N}(0,I(\vartheta))}.$$

By the law of large numbers (LLN) we have that $-\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\vartheta^2}\log f_\vartheta(X_i)\Big|_{\vartheta=\tilde\vartheta_n} \to I(\vartheta)$, cf. (12.22); further, by the CLT (Theorem 4.3), (12.7) and (12.21), the second sum converges in distribution to $\mathcal{N}\left(0, I(\vartheta)\right)$. Hence, by (3.6),

$$\sqrt{n}\left(\hat\vartheta_n - \vartheta\right) \xrightarrow{d} \mathcal{N}\left(0, I(\vartheta)^{-1}\right),$$

the assertion. □

12.12 PROBLEMS

Exercise 12.1 (Poisson distribution, cf. (12.3) in Example 12.6)**.** *Verify that the moment generating function of a Poisson random variable* $X \sim P_\lambda$ *is*

$$\begin{aligned}\mathbb{E}\,e^{tX} &= \exp\left(\lambda(e^t - 1)\right)\\ &= 1 + \lambda t + \frac{t^2}{2}\left(\lambda + \lambda^2\right) + \frac{t^3}{6}\left(\lambda + 3\lambda^2 + \lambda^3\right) + \frac{t^3}{6}\left(\lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4\right) + O(t^5).\end{aligned}$$

Derive that $\mathbb{E}\,X = \lambda = \operatorname{var} X = \mu_3$ *(skewness) and* $\mu_4 = \lambda + 3\lambda^2$ *(kurtosis).*

Exercise 12.2. *Verify that* $X + Y \sim P_{\lambda+\lambda'}$*, if* $X \sim P_\lambda$ *and* $Y \sim P_{\lambda'}$ *are independent.*

Exercise 12.3. *Verify the moment estimator for the Gamma distribution given in Example 12.5.*

Moments

Exercise 12.4. *The density of a distribution is* $f_\theta(x) = \begin{cases} c(\theta)(\theta - x) & \text{if } x \in (0,\theta) \\ 0 & \text{else}. \end{cases}$ *What is* $c(\theta)$*? Give a moment estimator* $\hat{\theta}$ *for* θ*. Find a sample, so that the estimator* $\hat{\theta}$ *is not reasonable.*

Exponential Distribution

Exercise 12.5. *Verify that the moment estimator (cf. Section 12.2) for the parameter* λ *of an exponential distribution* E_λ *with* $\varphi(x) := x$ *is* $\hat{\lambda}(X) = 1/\overline{X}_n$*.*

Exercise 12.6. *For* $X_i \sim E_\lambda$*, show that* $\mathbb{E}\, X_i^\alpha = \frac{\Gamma(1+\alpha)}{\lambda^\alpha}$*. Show that*

$$\hat{\lambda}_\alpha(X) := \left(\frac{\Gamma(1+\alpha)}{\overline{X_n^\alpha}}\right)^{1/\alpha}$$

is a moment estimator for λ*, where* $\overline{X_n^\alpha} = \frac{1}{n}\sum_{i=1}^n X_i^\alpha$*. Compare with Exercise 12.5.*

Exercise 12.7 (Exponential random variables, cf. Exercise 12.5 and Exercise 5.4)**.** *Show that the maximum likelihood estimator to estimate the rate* λ *given independent observations* $X_i \sim E_\lambda$ *is* $\hat{\lambda}(X_1,\dots,X_n) = 1/\overline{X}_n$*. This estimator* $\hat{\lambda}$ *is biased, but*

$$\lambda^*(X_1,\dots,X_n) := \frac{1}{\frac{n}{n-1}\cdot \overline{X}_n}$$

is unbiased (cf. Exercise 5.4).

Exercise 12.8. *Recall from Exercise 11.3 that* $\min_{i=1,\dots,n} X_i \sim E_{n\lambda}$ *for independent* $X_i \sim E_\lambda$*. It follows that* $\mathbb{E}\min_{i=1,\dots n} X_i = \frac{1}{n\lambda}$ *and* $\check{\lambda} := \frac{1}{n \min_{i=1,\dots n} X_i}$ *seem to be a useful estimator to estimate* λ*. Show that* $\mathbb{E}\,\check{\lambda} = \infty$.

Exercise 12.9. *Consider the estimator* $\acute{\mu}(X) := \overline{X}_n$ *and* $\grave{\mu}(X) := n \cdot \min_{i=1,\dots n} X_i$ *of independent, exponentially distributed random variables.*

Show that $\mathbb{E}\,\acute{\mu} = \mathbb{E}\,\grave{\mu} = \frac{1}{\lambda}$ *and compare their variance.*

Exercise 12.10 (Beta distribution)**.** *Verify the moment estimator*

$$\hat{\alpha}(X) := \overline{X}_n\left(\frac{\overline{X}_n(1-\overline{X}_n)}{s_n^2} - 1\right) \text{ and } \hat{\beta}(X) := \left(1-\overline{X}_n\right)\left(\frac{\overline{X}_n(1-\overline{X}_n)}{s_n^2} - 1\right)$$

for the parameters (α,β) *of the Beta distribution (see Exercise 5.10).*

Exercise 12.11. *Demonstrate the bias/variance decomposition of MSE, Eq. (12.9).*

Exercise 12.12. *Let* X_i *be independent* $\mathcal{N}(\mu_0,\sigma^2)$ *observations (*μ_0 *known). Give the MLE for the variance* σ^2*.*

13 Linear models

> Uncertainty is not the same as probability.
>
> A. Shapiro, 1943

13.1 GENERAL LINEAR MODEL

The *general linear model* or *multivariate regression model* is

$$Y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \dots, n. \tag{13.1}$$

Definition 13.1 (Nomenclature)**.**

- The variables $x_i^\top = (x_{i,1}, \dots x_{i,K})$ are observable and called *explanatory variables*;
- The vector $\beta = (\beta_1, \dots \beta_K)^\top$ is the unobservable *regression parameter*;
- The error ε is called *disturbance*, *noise* or simply *error*;
- The vector Y is the *dependent variable*;
- The matrix
$$X := \begin{pmatrix} x_{1,1} & \dots & x_{1,K} \\ & \dots & \\ x_{n,1} & \dots & x_{n,K} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ \dots \\ x_n^\top \end{pmatrix}$$
is the *design matrix*.

The matrix form of (13.1) is $Y = X\beta + \varepsilon$, its expanded form is $Y_i = \sum_{j=1}^K x_{ij}\,\beta_j + \varepsilon_i$.

It is evident that the problem setting (13.1) includes the more general linear problem (affine linear problem)

$$Y_i = \beta_0 + x_i^\top \beta + \varepsilon_i \tag{13.2}$$

by replacing β by (β_0, β) and $x_i^\top$ by $(1, x_{i,1}, \dots, x_{i,K})$, resp.

13.2 GENERALIZED LEAST SQUARES ESTIMATOR

The generalized least squares estimator (LS) for β is found by minimizing the mean squared error

$$\|Y - X\beta\|_\Sigma \to \min_\beta \tag{13.3}$$

with respect to β; here, we have incorporated a positive definite matrix Σ as additional parameter, where $\|x\|_\Sigma^2 := x^\top \Sigma^{-1} x$ is the norm corresponding to the inner product $\langle x, y\rangle_\Sigma := x^\top \Sigma^{-1} y$. The matrix Σ is the variogram, for independent data $\Sigma \sim \mathbb{1}$. (Cf. Remark 3.23 and the Mahalanobis[1] distance, which is unitless

[1] Prasanta Chandra Mahalanobis, 1893–1972, Indian applied statistician

and scale invariant). We arrive at

$$\begin{aligned}\hat\beta \in \operatorname*{arg\,min}_\beta \|Y - X\beta\|_\Sigma^2 &= \operatorname*{arg\,min}_\beta\ (Y - X\beta)^\top \Sigma^{-1} (Y - X\beta) \\ &= \operatorname*{arg\,min}_\beta\ Y^\top\Sigma^{-1}Y - \beta^\top X^\top\Sigma^{-1}Y - Y^\top\Sigma^{-1}X\beta + \beta^\top X^\top\Sigma^{-1}X\beta.\end{aligned} \tag{13.4}$$

Differentiating the latter with respect to β gives the *normal equations*[2] $X^\top\Sigma^{-1}X\beta = X^\top\Sigma^{-1}Y$ and the generalized *least squares estimator* is the *linear* estimator

$$\hat\beta := \left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}Y \in \mathbb{R}^K. \tag{13.5}$$

Remark 13.2. The operator

$$\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1} \in \mathbb{R}^{K\times n}$$

is a generalized Moore–Penrose pseudoinverse for linearly independent columns of X. So if $X^\top\Sigma^{-1}X$ is not invertible, then $\hat\beta$ is still well-defined by (13.4) via the generalized pseudoinverse.

Lemma 13.3. *The matrix*[3]

$$P := X\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1} \in \mathbb{R}^{n\times n} \tag{13.6}$$

has the following properties:

(i) P is a projection onto X, i.e., $P^2 = P$ and $PX = X$;

(ii) $P^\top = \Sigma^{-1}P\Sigma$ is a projection onto $\Sigma^{-1}X$, i.e., $P^\top\Sigma^{-1}X = \Sigma^{-1}X$.

The matrix $\Sigma^{-1/2}P\Sigma^{1/2} = \Sigma^{-1/2}X\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1/2} \in \mathbb{R}^{n\times n}$ is an orthonormal projection.

Remark 13.4. It is convenient to set $\hat Y := X\hat\beta$, cf. (13.3). Note, that

$$\hat Y = X\hat\beta = PY$$

and

$$\varepsilon := Y - X\hat\beta = Y - \hat Y = (1 - P)Y \tag{13.7}$$

is the residual vector, cf. (13.1). The error ε and $\hat Y$ are orthogonal (uncorrelated), as

$$\left\langle \hat Y \mid \varepsilon\right\rangle_\Sigma = Y^\top P^\top\Sigma^{-1}\varepsilon = Y^\top\cdot\Sigma^{-1}\underbrace{P\Sigma\cdot\Sigma^{-1}(1-P)}_{=0}Y = 0;$$

in addition, the error ε is orthogonal to X (use (i)),

$$\langle X \mid \varepsilon\rangle_\Sigma = \underbrace{(PX)^\top}_{X=PX}\Sigma^{-1}(1-P)Y = X^\top\cdot\Sigma^{-1}\underbrace{P\Sigma\cdot\Sigma^{-1}(1-P)}_{=0}Y = 0.$$

[2] Normalgleichungen, dt.

[3] Note that P is the *formal* inverse of (3.21) provided that $X = A^{-1}$.

13.3 GAUSS–MARKOV THEOREM

The reasoning in the preceding section (Section 13.2) does *not* involve stochastic. One may restate the problem (13.1) as $Y(\omega) = x^\top\beta + \varepsilon(\omega)$ to emphasize that x and β are deterministic while ε (and thus Y) are random. Now note that $\operatorname{var} Y_i = \operatorname{var}\varepsilon_i$; note as well that $\mathbb{E}(Y_i \mid x_i) = x_i^\top\beta$, assuming that $\mathbb{E}\,\varepsilon_i = 0$. A typical assumption for the error in (13.1) is $Y_i \sim \mathcal{N}(x_i^\top\beta,\, \sigma^2)$ or

$$Y(x) \sim \mathcal{N}\left(x^\top\beta,\, \Sigma\right),$$

for example.

For dependent errors, the quantity $2\gamma(x_i, x_j) = \operatorname{cov}\left(Y_i, Y_j\right)$ is the variogram. The semivariogram is based on the distance only, $\gamma(\|x_i - x_j\|) = \gamma(x_i, x_j)$.

Theorem 13.5 (Gauß–Markov[4] theorem)**.** *Suppose that Σ is a positive semi-definite matrix and $\varepsilon \sim (0, \Sigma)$, i.e.,*

(i) $\mathbb{E}\,\varepsilon_i = 0$ *($i = 1, \dots, n$) and*

(ii) $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \Sigma_{ij}$ *for $i, j = 1, \dots n$.*

Then the ordinary least squares estimator (13.5) is unbiased, has variance $\operatorname{var}\hat\beta = \left(X^\top\Sigma^{-1}X\right)^{-1}$ *and is the best linear unbiased estimator (BLUE), i.e., has smallest variance.*

Remark 13.6. Note, that the Gauß–Markov theorem (in the form presented above) does not make any assumption on the distribution, nor on independence of the observations; it just involves the first two moments.

Proof of Theorem 13.5. $\hat\beta$ is unbiased, as

$$\begin{aligned}\mathbb{E}\,\hat\beta &= \mathbb{E}\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\cdot Y = \mathbb{E}\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\left(X\beta + \varepsilon\right)\\ &= \beta + \left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\cdot \underbrace{\mathbb{E}\,\varepsilon}_{=0 \text{ by (i)}} = \beta;\end{aligned}$$

the variance is

$$\begin{aligned}\operatorname{var}\hat\beta &= \operatorname{var}\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\left(X\beta + \varepsilon\right) = \operatorname{var}\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\varepsilon\\ &= \left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}\cdot \underbrace{\operatorname{var}\varepsilon}_{=\Sigma \text{ by (ii)}} \cdot\Sigma^{-1}X\left(X^\top\Sigma^{-1}X\right)^{-1} = \left(X^\top\Sigma^{-1}X\right)^{-1}.\end{aligned}$$

Now suppose that $\tilde\beta = C\,Y$ is another linear estimator, then $C = \left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1} + D$ for some matrix $D \in \mathbb{R}^{K\times n}$. It holds that

$$\begin{aligned}\mathbb{E}\,\tilde\beta &= \mathbb{E}\,CY = \mathbb{E}\left(\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1} + D\right)\left(X\beta + \varepsilon\right)\\ &= \left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1}X\beta + DX\beta + \left(\left(X^\top\Sigma^{-1}X\right)^{-1} X^\top\Sigma^{-1} + D\right)\underbrace{\mathbb{E}\,\varepsilon}_{=0 \text{ by (i)}}\\ &= \beta + DX\beta.\end{aligned}$$

[4] Andrey Andreyewich Markov, 1856–1922

Therefore, $\tilde{\beta}$ is unbiased iff $DX = 0$. Now

$$\begin{aligned}
\operatorname{var}\tilde{\beta} &= \operatorname{var}(CY) = C\operatorname{var}(Y)C^\top = C\,\Sigma\,C^\top\\
&= \left(\left(X^\top\Sigma^{-1}X\right)^{-1}X^\top\Sigma^{-1}+D\right)\Sigma\left(\left(X^\top\Sigma^{-1}X\right)^{-1}X^\top\Sigma^{-1}+D\right)^\top\\
&= \left(X^\top\Sigma^{-1}X\right)^{-1}X^\top\Sigma^{-1}\Sigma\Sigma^{-1}X\left(X^\top\Sigma^{-1}X\right)^{-1}\\
&\qquad+\underbrace{D\Sigma\Sigma^{-1}X}_{DX=0}\left(X^\top\Sigma^{-1}X\right)^{-1}+\left(X^\top\Sigma^{-1}X\right)^{-1}\underbrace{X^\top\Sigma^{-1}\Sigma D^\top}_{DX=0}+D\,\Sigma\,D^\top\\
&= \left(X^\top\Sigma^{-1}X\right)^{-1}+D\,\Sigma\,D^\top = \operatorname{var}\hat{\beta}+D\,\Sigma\,D^\top.
\end{aligned}$$

Hence, $\hat{\beta}$ has smaller variance than $\tilde{\beta}$, as $D\,\Sigma\,D^\top$ is positive semidefinite and thus the assertion. □

In the typical and usual setting the observations in (13.10) are independent (and thus $\Sigma_{ij} = \mathbb{E}\,\varepsilon_i\varepsilon_j = 0$ if $i \neq j$). In this case the minimization (13.10) reads

$$\|Y - X\beta\|_\Sigma^2 = \sum_{i=1}^n \left(\frac{y_i - x_i^\top\beta}{\sigma_i}\right)^2 = \sum_{i=1}^n w_i\left(y_i - x_i^\top\beta\right)^2 \to \min,$$

where $\varepsilon_i \sim (0,\sigma_i^2)$ and the weights $w_i = \frac{1}{\sigma_i^2}$. Recall that $\frac{1}{\sigma^2}$ is the precision (and Σ^{-1} the precision matrix). For independent, identically distributed random variables $\varepsilon \sim (0, \sigma^2\,\mathbb{1})$ thus particularly

$$\sigma^2\,\|Y - X\beta\|_\Sigma^2 = \|Y - X\beta\|_{\ell_n^2}^2 = \sum_{i=1}^n \left(y_i - x_i^\top\beta\right)^2 \to \min.$$

Corollary 13.7 (Gauß–Markov theorem, the classical formulation for homoscedastic data)**.** *Suppose that*

- $\mathbb{E}\,\varepsilon_i = 0$,
- $\operatorname{var}\varepsilon_i = \sigma^2 < \infty$*, i.e., the errors are homoscedastic,*[5]
- $\operatorname{cov}(\varepsilon_i,\varepsilon_j) = 0$ *($i \neq j$), i.e., distinct error terms are uncorrelated.*

Then the ordinary least squares (OLS) estimator

$$\hat{\beta} = \left(X^\top X\right)^{-1}X^\top Y$$

(cf. (13.5)) is the best linear unbiased estimator (BLUE) (i.e., with smallest variance), its variance is $\operatorname{var}\hat{\beta} = \sigma^2\cdot(X^\top X)^{-1}$.

13.4 NONLINEAR REGRESSION

For the general, nonlinear problem

$$Y_i = g(x_i;\beta) + \varepsilon_i, \quad i = 1,\dots,n, \tag{13.8}$$

it is occasionally written that

$$\mathbb{E}(Y \mid x) = g(x;\beta). \tag{13.9}$$

[5] Scedasticity (dispersion, Greek). Heteroscedasticity (Varianzheterogenität, dt.) is the absence of homoscedasticity (Varianzhomogenität, dt.).

As in (13.3) one may consider the least squares estimator

$$\hat{\beta} \in \arg\min_{\beta} \left\|(Y_i - g(x_i;\beta))_{i=1}^n\right\|_{\Sigma}, \text{ or } \hat{\beta} \in \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^n \big(Y_i - g(x_i;\beta)\big)^2$$

to extract an estimator $\hat{\beta}$ for β from observations (x_i, Y_i).

A typical example of (13.8) is the linear model $g(x;\beta) = \beta^\top g(x)$, i.e.,

$$Y_i = \sum_{j=1}^J \beta_j g_j(x_i) + \varepsilon_i = \beta^\top g(x_i) + \varepsilon_i \tag{13.10}$$

(the model is linear, as the regression coefficients β_j appear linearly in the relationship (13.10) although $g_j(\cdot)$ are possibly nonlinear). Note, that including the constant function $g_0(\cdot) := 1$ again (cf. (13.2)) gives the problem

$$Y_i = \beta_0 + \sum_{j=1}^J \beta_j\, g_j(x_i) + \varepsilon_i.$$

Remark 13.8. Exercise 13.6 below states the normal equations and the BLUE estimator for (13.10) explicitly.

Example 13.9. Recall the angle addition theorem $A\sin(\omega\cdot x + \varphi) = \beta_c \sin(\omega\cdot x) + \beta_s \cos(\omega\cdot x)$ with $\beta_c = A\cos\varphi$ and $\beta_s = A\sin\varphi$; as well, $A^2 = \beta_s^2 + \beta_c^2$ and $\varphi = \arctan\frac{\beta_s}{\beta_c}$.

13.5 COEFFICIENT OF DETERMINATION

Proposition 13.10. *The best prediction of order* 0 *(i.e., without involving a model) is the weighted mean,*

$$\hat{\beta}_0 = \overline{Y} := \frac{\mathbb{1}^\top\,\Sigma^{-1} Y}{\mathbb{1}\,\Sigma^{-1}\,\mathbb{1}}. \tag{13.11}$$

Proof. Consider the objective $\beta_0 \mapsto \|Y - \beta_0\cdot\mathbb{1}\|_\Sigma^2 = (Y - \beta_0\cdot\mathbb{1})^\top\,\Sigma^{-1}\,(Y - \beta_0\cdot\mathbb{1})$. The first order conditions for the minimum are $(Y - \beta_0\cdot\mathbb{1})^\top\,\Sigma^{-1}\,\mathbb{1} = 0$ and thus the assertion. □

Definition 13.11. The coefficient of determination is a quantity to quantify the quality of a regression. The following terms are involved:

(i) The *total sum of squares* (proportional to the variance of the data) is $\mathsf{TSS} := \left\|\left(Y_i - \overline{Y}\right)_{i=1}^n\right\|_\Sigma^2$, where $\overline{Y} \in \operatorname{range} g(X)$.

(ii) The *explained sum of squares* (also regression sum of squares, $\mathsf{SS}_{\text{model}}$) is $\mathsf{ESS} := \left\|\left(g(x_i;\hat{\beta}) - \overline{Y}\right)_{i=1}^n\right\|_\Sigma^2$.

(iii) The *residual sum of squares* (also sum of squares of residuals, $\mathsf{SS}_{\text{error}}$) is $\mathsf{RSS} := \left\|(Y_i - g(x_i;\hat{\beta}))_{i=1}^n\right\|_\Sigma^2$.

(iv) The *coefficient of determination*[6] is $R^2 := 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}$. The minuend $\mathsf{FVU} := \frac{\mathsf{RSS}}{\mathsf{TSS}}$ is also called *fraction of variance unexplained.*

If the model explains everything, then $R^2 = 1$, while $R^2 = 0$ identifies a useless model.

[6] Bestimmtheitsmaß, dt.

Lemma 13.12. *It holds that* $\mathsf{TSS} = \mathsf{ESS} + \mathsf{RSS}$ *and*

$$R^2 = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}} = \frac{\mathsf{ESS}}{\mathsf{TSS}} \in [0,1].$$

Proof. Note, that $\overline{Y} \cdot \mathbb{1} \in \operatorname{range}(X)$ provided that $X = (\mathbb{1}, \dots)$, cf. (13.2). The statement thus is a consequence of the following more general proposition. □

Proposition 13.13. *Let* $x \in \operatorname{range}(X)$. *Then* $\|Y - x\|_\Sigma^2 = \|Y - \hat{Y}\|_\Sigma^2 + \|\hat{Y} - x\|_\Sigma^2$.

Proof. It holds that $\|Y - x\|_\Sigma^2 = \|Y - \hat{Y} + \hat{Y} - x\|_\Sigma^2 = \|Y - \hat{Y}\|_\Sigma^2 + \|\hat{Y} - x\|_\Sigma^2 + 2\left\langle Y - \hat{Y} \mid \hat{Y} - x\right\rangle_\Sigma$. But (cf. Remark 13.4)

$$\begin{aligned}
\left\langle Y - \hat{Y} \mid \hat{Y} - x\right\rangle_\Sigma &= \langle (1-P)Y \mid PY - x\rangle_\Sigma \\
&= Y^\top (1-P)^\top \Sigma^{-1}(PY - x) \\
&= Y^\top \Sigma^{-1}(1-P)\Sigma\Sigma^{-1}(PY - x) \\
&= Y^\top \Sigma^{-1} \underbrace{(1-P)P}_{=0} Y - Y^\top \Sigma^{-1} \underbrace{(1-P)x}_{=0} = 0,
\end{aligned}$$

as $PX = X$ and thus $Px = x$ for $x \in \operatorname{range}(X)$. □

13.6 NUMERICAL SOLUTION

Computing the inverse explicitly in (13.5) is numerically not stable. For numerically stable computations one may apply the QR algorithm, one of the top 10 algorithms from the 20th century. Here, we outline the *rank-deficient complete orthogonal decomposition*.

13.6.1 QR

Consider the inner product $\langle x \mid y\rangle_\Sigma := x^\top \Sigma^{-1} y$ and set $H_v := \mathbb{1} - 2\frac{v\,v^\top \Sigma^{-1}}{v^\top \Sigma^{-1} v}$. Then

$$\begin{aligned}
\|H_v x\|_\Sigma^2 &= x^\top H_v^\top \Sigma^{-1} H_v x \\
&= x^\top \left(\mathbb{1} - 2\frac{\Sigma^{-1} v\, v^\top}{v^\top \Sigma^{-1} v}\right) \Sigma^{-1} \left(\mathbb{1} - 2\frac{v\, v^\top \Sigma^{-1}}{v^\top \Sigma^{-1} v}\right) x \\
&= x^\top \left(\Sigma^{-1} - 2\frac{\Sigma^{-1} v\, v^\top \Sigma^{-1}}{v^\top \Sigma^{-1} v} - 2\frac{\Sigma^{-1} v\, v^\top \Sigma^{-1}}{v^\top \Sigma^{-1} v} + 4\frac{\Sigma^{-1} v\,\left(v^\top \Sigma^{-1} v\right)}{v^\top \Sigma^{-1} v}\frac{v^\top \Sigma^{-1}}{v^\top \Sigma^{-1} v}\right) x \\
&= x^\top \Sigma^{-1} x = \|x\|_\Sigma^2,
\end{aligned}$$

i.e., $x \longrightarrow H_v x$ is an isometry.

Given a and e with $\|e\|_\Sigma = 1$, set $\lambda := \pm\|a\|_\Sigma$ and choose $v \propto a - \lambda e$, then

$$\begin{aligned}
H_v a &= a - 2\frac{(a-\lambda e)(a-\lambda e)^\top \Sigma^{-1}}{(a-\lambda e)^\top \Sigma^{-1}(a-\lambda e)} a \\
&= a - 2(a - \lambda e)\frac{a^\top \Sigma^{-1} a - \lambda e^\top \Sigma^{-1} a}{a^\top \Sigma^{-1} a - 2\lambda e^\top \Sigma^{-1} a + \lambda^2 e^\top \Sigma^{-1} e} \\
&= a - 2(a - \lambda e)\frac{1}{2} = \lambda\, e.
\end{aligned}$$

For $a = \begin{pmatrix} a_1 \\ a_{2:} \end{pmatrix}$, we may particularly choose $v \propto \begin{pmatrix} a_1 \\ a_{2:} \end{pmatrix} - \lambda\, e_1$, where $\operatorname{sign} \lambda = -\operatorname{sign} a_1$ to avoid loss of significance in the first component. While $v_{2:} = a_{2:}$, v's first component v_1 is

$$v_1 = a_1 + \|a\|_\Sigma \operatorname{sign} a_1 = (|a_1| + \|a\|_\Sigma) \operatorname{sign} a_1$$

(alternatively, set $v_{2:} \coloneqq (\operatorname{sign} a_1)\, a_{2:}$ and $v_1 = |a_1| + \|a\|_\Sigma$). Further we find $\|v\|_\Sigma^2 = \|a\|_\Sigma^2 - 2\lambda\, a^\top \Sigma^{-1} e + \lambda^2 = 2\,\|a\|_\Sigma^2 - 2\lambda\, a \Sigma^{-1} e$. With $\Sigma = \mathbb{1}$,

$$\|v\|^2 = 2\,\|a\|^2 - 2\lambda\, a_1 = 2\,\|a\|\,(|a_1| + \|a\|)\,.$$

13.6.2 Rank-revealing/ rank-deficient complete orthogonal decomposition

Proposition 13.14. *Let $A = Q_1 \begin{pmatrix} 0 & R \\ 0 & 0 \end{pmatrix} Q_2^\top$ with Q_1, Q_2 unitary, R regular and $A^+ \coloneqq Q_2 \begin{pmatrix} 0 & 0 \\ R^{-1} & 0 \end{pmatrix} Q_1^\top$. Then $x = A^+ b$ has smallest norm among all minimizers of $x \mapsto \|b - A\,x\|$.*

Proof. It holds that $(0\ R)\, Q_2^\top A^+ b = (\mathbb{1}\ \ 0)\, Q_1^\top b$ and, as Q_1 is unitary,

$$\begin{aligned} \|b - A\,x\| &= \left\| b - Q_1 \begin{pmatrix} 0 & R \\ 0 & 0 \end{pmatrix} Q_2^\top\, x \right\| \\ &= \left\| \begin{pmatrix} (\mathbb{1}\ \ 0)\, Q_1^\top b \\ (0\ \ \mathbb{1})\, Q_1^\top b \end{pmatrix} - \begin{pmatrix} (0\ R)\, Q_2^\top x \\ 0 \end{pmatrix} \right\| \\ &\geq \left\| \begin{pmatrix} 0 \\ (0\ \ \mathbb{1})\, Q_1^\top b \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} (\mathbb{1}\ \ 0)\, Q_1^\top b \\ (0\ \ \mathbb{1})\, Q_1^\top b \end{pmatrix} - \begin{pmatrix} (0\ R)\, Q_2^\top A^+ b \\ 0 \end{pmatrix} \right\| \\ &= \left\| Q_1^\top b - Q_1^\top\, A\, A^+ b \right\| \\ &= \left\| b - A\, A^+ b \right\|. \end{aligned} \tag{13.12}$$

It follows that $A^+ b$ minimizes $\|b - A\,\cdot\|$.

Finally suppose that x is another minimizer, then equality holds in (13.12), i.e., $(\mathbb{1}\ \ 0)\, Q_1^\top b = (0\, R)\, Q_2^\top x$. Multiply with $Q_2 \begin{pmatrix} 0 \\ R^{-1} \end{pmatrix}$ to get $Q_2 \begin{pmatrix} 0 & 0 \\ R^{-1} & 0 \end{pmatrix} Q_1^\top b = Q_2 \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{1} \end{pmatrix} Q_2^\top x$, i.e., $A^+ b = Q_2 \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{1} \end{pmatrix} Q_2^\top x$. As Q_2 is unitary it follows that $\|A^+ b\| = \left\| \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{1} \end{pmatrix} Q_2^\top x \right\| \leq \left\| Q_2^\top x \right\| = \|x\|$ and thus the assertion. □

Remark 13.15. The norm is strictly convex, thus the generalized inverse A^+ is uniquely defined by the characterization in Proposition 13.14 (i.e., A^+ is independent of the particular choice of Q_1, Q_2 and R).

In what follows we construct the generalized inverse explicitly for the standard inner product with $\Sigma = \mathbb{1}$. For $A \in \mathbb{R}^{m \times n}$, suppose that $A = P \begin{pmatrix} L & 0 \end{pmatrix} Q_1$ (i.e., $A^\top = Q_1 \begin{pmatrix} L^\top \\ 0 \end{pmatrix} P$) for an orthogonal projection Q_1 and a permutation P, where

$$\begin{pmatrix} L & 0 \end{pmatrix} = \begin{pmatrix} \ell_{11} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ell_{22} & \ddots & \vdots & & \\ & \ddots & \ddots & 0 & \vdots & \\ \ell_{k1} & & & \ell_{kk} & 0 & \cdots \\ \vdots & & & \vdots & \vdots & \\ \ell_{m1} & \cdots & & \ell_{mk} & 0 & \cdots \end{pmatrix}$$

is an extended lower triangular rank k matrix; the permutation P is found by pivoting the rows. Define the exchange matrix $P_m := \begin{pmatrix} 0 & 0 & 1 \\ 0 & \iddots & 0 \\ 1 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}$, then

$$P_m \begin{pmatrix} L & 0 \end{pmatrix} P_n = \begin{pmatrix} \cdots & 0 & \ell_{mk} & \cdots & \ell_{m1} \\ & \vdots & \vdots & & \vdots \\ \cdots & 0 & \ell_{kk} & \cdots & \ell_{k1} \\ & \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \ell_{11} \end{pmatrix} = Q_2 \begin{pmatrix} \cdots & 0 & \ell'_{mk} & \cdots & \ell'_{m1} \\ & & 0 & \ddots & \\ & \vdots & \vdots & \ddots & \ell'_{m-k+1,1} \\ & & & \ddots & 0 \\ 0 & \cdots & 0 & \cdots & \vdots \end{pmatrix} = Q_2 \begin{pmatrix} 0 & L' \\ 0 & 0 \end{pmatrix} \tag{13.13}$$

by employing a usual QR decomposition again, where L' is regular upper triangular matrix; note that the exchange matrices allow exploiting the sparse structure in (13.13). It follows that

$$A = P \begin{pmatrix} L & 0 \end{pmatrix} Q_1 = P \cdot P_m Q_2 \begin{pmatrix} 0 & L' \\ 0 & 0 \end{pmatrix} P_n \cdot Q_1$$

and the generalized inverse (Moore–Penrose inverse) A^+ of the rank-k matrix A is

$$A^+ = Q_1 P_n \begin{pmatrix} 0 & 0 \\ L'^{-1} & 0 \end{pmatrix} Q_2 P_m P.$$

Finally $A^+ b = Q_1 \begin{pmatrix} \tilde{x} \\ 0 \end{pmatrix}$, where

$$\begin{pmatrix} \vdots & 0 & \dots & & 0 \\ 0 & 0 & 0 & & \vdots \\ \ell'_{m-k+1,1} & \ddots & & \vdots & \\ \vdots & \ddots & \ddots & & \\ \ell'_{m1} & \cdots & \ell'_{mk} & 0 & \cdots \end{pmatrix} \cdot \begin{pmatrix} \tilde{x} \\ 0 \end{pmatrix} = P_m \begin{pmatrix} 0 & L' \\ 0 & 0 \end{pmatrix} P_n \cdot \begin{pmatrix} \tilde{x} \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{b} \\ \dots \end{pmatrix} := P_m Q_2 P_m P\, b.$$

For an efficient implementation of a rank revealing generalized inverse see https://github.com/aloispichler/Matrix-Class/.

13.7 PROBLEMS

Exercise 13.1. *Verify Footnote 3 on page 120.*

Exercise 13.2 (Linear regression)**.** *Consider the problem* $Y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$ *and show that the slope is*

$$\hat{\beta} = \frac{\frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_i (x_i - \overline{x})^2}$$

with intercept $\hat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x}$. *The regression line passes the point* $(\overline{x}, \overline{y})$ *and it follows that* $\hat{\alpha} + \hat{\beta} x = \overline{y} + \hat{\beta}(x - \overline{x})$.

Hint: Exercise 2.5.

Exercise 13.3. *Show that* $\hat\beta = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = r_{x,y}\frac{s_y}{s_x}$ *(cf. (2.9)).*

Exercise 13.4. *Verify the famous parameters for the linear regression* $Y_i = \alpha + \beta x$ *(with independent* $\varepsilon_i \sim (0, \sigma_i^2)$*)*

$$\begin{aligned}\hat\beta &= \frac{\overline{x\cdot y} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sum_i w_i\,(x_i - \overline{x})\,(y_i - \overline{y})}{\overline{x^2} - \overline{x}^2} = \frac{\left(\sum_i w_i x_i y_i\right) - \left(\sum_i w_i x_i\right)\left(\sum_i w_i y_i\right)}{\left(\sum_i w_i x_i^2\right) - \left(\sum_i w_i x_i\right)^2},\\ \hat\alpha &= \overline{y} - \hat\beta\overline{x},\end{aligned}$$

where the weights are $w_i = \frac{1}{\sigma_i^2}\Big/\sum_j \frac{1}{\sigma_j^2}$ *(cf. (13.11)) and*

$$\overline{x} := \sum_i w_i x_i,\ \overline{x^2} := \sum_i w_i x_i^2,\ \overline{y} := \sum_i w_i y_i,\ \overline{y^2} := \sum_i w_i y_i^2 \text{ and } \overline{x\cdot y} := \sum_i w_i x_i y_i$$

the weighted means.

Exercise 13.5. *Show that the residual* $\hat\varepsilon_i = y_i - x_i^\top\hat\beta$ *(cf. (13.7)) satisfy* $\overline{\hat\varepsilon} = \sum_{i=1}^n w_i\hat\varepsilon_i = 0$ *and further, the residuals* $\hat\varepsilon_i$ *and* x_i *are uncorrelated. Hint: show first that* $\overline{\hat\varepsilon}\cdot\overline{x} = \overline{\hat\varepsilon\cdot x}$*.*

Exercise 13.6. *Given the same conditions as in the Gauß–Markov theorem, Theorem 13.5. Show that the optimal parameter for the linear regression (13.10) with respect to general functions* $g_j(x_1, \dots x_K)$*,* $j = 1, \dots J$*, is*

$$\hat\beta = \left(g(X)^\top\Sigma^{-1}g(X)\right)^{-1} g(X)^\top\Sigma^{-1}Y \in \mathbb{R}^J,$$

where

$$g(X) := \begin{pmatrix} g_1(x_{1,1}, \dots x_{1,K}) & \dots & g_J(x_{1,1}, \dots x_{1,K})\\ & \dots & \\ g_1(x_{n,1}, \dots x_{n,K}) & \dots & g_J(x_{n,1}, \dots x_{n,K})\end{pmatrix} \in \mathbb{R}^{n\times J}$$

is the design matrix. Note, that K and J may differ here (although there are usually not more variables than functions, $J \le K$*).*

Exercise 13.7. *Consider the problem of approximating the data* Y_i *by a simple constant, i.e.,* $g(x) = 1$ *in (13.3). Show that* $\hat\beta_0 = \overline{Y}_n^\Sigma := \sum_{j=1}^n \frac{\sum_{i=1}^n \Sigma_{ij}^{-1}}{\sum_{k,\ell=1}^n \Sigma_{k\ell}^{-1}} Y_j = w^\top Y$ *with weights* $w_j = \sum_{i=1}^n \frac{\Sigma_{ij}^{-1}}{\sum_{k,\ell=1}^n \Sigma_{k\ell}^{-1}}$*,* $w^\top = \frac{\mathbb{1}^\top\Sigma^{-1}}{\mathbb{1}^\top\Sigma^{-1}\mathbb{1}}$*.*

Show in particular that the best approximation is $\hat\beta = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ *whenever* $\Sigma = \sigma^2\,\mathbb{1}$*.*

Exercise 13.8. *Show that* $AA^+A = A$*,* $A^+AA^+ = A^+$*,* $(AA^+)^\top = AA^+$ *and* $(A^+A)^\top = A^+A$*.*

14 Logistic Regression

> Ah! C'est une belle hypothèse; ça explique beaucoup de choses.
>
> Pierre-Simon Laplace, 1749–1827

14.1 THE LOGIT AND LOGISTIC FUNCTIONS

Definition 14.1 (Sigmoid[1])**.** The standard logistic (also sigmoid) function is[2]

$$S(t)=\frac{1}{1+e^{-t}}=\frac{e^t}{e^t+1}=\frac{1}{2}+\frac{1}{2}\tanh\frac{t}{2},$$

cf. Figure 14.1.

$S(\cdot)$ is strictly monotonically increasing with $S(x)\xrightarrow[x\to-\infty]{}0$ and $S(x)\xrightarrow[x\to\infty]{}1$. It holds that $S(-t)=1-S(t)$ and thus $S'(t)=S'(-t)$. The derivative

$$S'(t)=\frac{e^{-t}}{\left(1+e^{-t}\right)^2}=\frac{1}{e^t+2+e^{-t}}=\frac{1}{\left(e^{t/2}+e^{-t/2}\right)^2}$$

is the *logistic kernel*. All derivatives of S can be expressed by S, as

$$S'(t)=S(t)\big(1-S(t)\big) \tag{14.1}$$

(or $S''(t)=S(t)\big(1-S(t)\big)\big(1-2S(t)\big)$, etc.). The function

$$\operatorname{logit}(p):=S^{-1}(p)=\log\frac{p}{1-p},\qquad p\in(0,1), \tag{14.2}$$

(i.e., $t=\log\frac{S(t)}{1-S(t)}$) gives the log-odds, or the logarithm of the odds $\frac{p}{1-p}$ and is called *logit function*.

The antiderivative of the logit function is $\int_{-\infty}^t S(u)\,\mathrm{d}u=\log\left(1+e^t\right)=-\log\big(1-S(t)\big)$. Note as well that $\lim_{\substack{\beta\to\infty,\\ \beta>0}} S(\beta\, t)=\begin{cases}0 & \text{if } t<0,\\ 1/2 & \text{if } t=0,\\ 1 & \text{if } t>0.\end{cases}$

14.2 THE LOGISTIC DISTRIBUTION

Definition 14.2 (Logistic distribution)**.** The logistic distribution has cdf $S\left(\frac{\cdot-m}{s}\right)$ for some $m\in\mathbb{R}$ and $s>0$. We shall write $S_{m,s}$ for a logistic distribution with parameters m and s.

[1] S-shaped; the function smashes $\mathbb{R}$ to $[0,1]$.

[2] Recall that $\tanh t=\frac{e^t-e^{-t}}{e^t+e^{-t}}$.

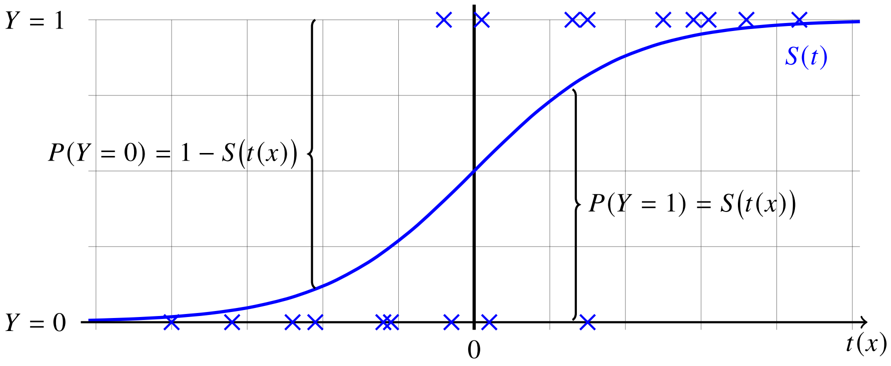


Figure 14.1: Logistic regression and the sigmoid function

Remark 14.3. If $U \in [0, 1]$ is uniformly distributed, then, by (14.2), $m + s \log \frac{1-U}{U}$ follows a logistic distribution.

Remark 14.4. For a $X \sim S_{m,s}$ random variable we have $\mathbb{E}\, X = m$ (by symmetry) and $\operatorname{var} X = \frac{\pi^2}{3} s^2$. Indeed, the variance of $X \sim S_{0,1}$ is

$$\begin{aligned}\operatorname{var} X = \mathbb{E}\, X^2 &= \int_{-\infty}^{\infty} t^2 \frac{e^t}{(e^t + 1)^2}\, \mathrm{d}t = 2 \int_0^{\infty} t^2 \sum_{k=1}^{\infty} (-1)^{k-1} k\, e^{-kt}\, \mathrm{d}t \\ &\underset{t \leftarrow t/k}{=} 2 \sum_{k=1}^{\infty} (-1)^{k-1} \frac{k}{k^2\, k} \int_0^{\infty} t^2 e^{-t}\, \mathrm{d}t = 4 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} = 4 \frac{\pi^2}{12} = \frac{\pi^2}{3} \approx 1.8138^2.\end{aligned}$$

14.3 REGRESSION

Assume that the dependent variable $t(\cdot)$ is a function of the *explanatory variables* $x = (x_1, \dots, x_d)$, $t\colon \mathbb{R}^d \to \mathbb{R}$. Define the random variable (cf. Figure 14.1)

$$Y(x) := \begin{cases} 1 & \text{if } t(x) \ge \varepsilon, \\ 0 & \text{else,} \end{cases} \tag{14.3}$$

where $\varepsilon \in \mathbb{R}$ is random with cdf F_ε, i.e.,

$$P(\varepsilon \le t') = F_\varepsilon(t'). \tag{14.4}$$

Note, that $Y(x)$ is a latent variable, as ε is not observed. We have

$$P\big(Y(x) = 1\big) = P\big(\varepsilon \le t(x)\big) = F_\varepsilon\big(t(x)\big), \tag{14.5}$$

which is occasionally also stated as

$$\mathbb{E}\,\big(Y \mid x\big) = F_\varepsilon\big(t(x)\big)$$

(cf. linear regression and (13.1)).

Logistic regression finally assumes that $\varepsilon \sim S$ follows a logistic distribution, i.e., $F_\varepsilon = S$.

Remark 14.5. Compare with linear regression and (13.9).

observation i	variable x_1	$\dots$	x_d	categorical variable (result) $Y(x_1, \dots x_d)$
1	$\dots$		$\dots$	0
$\dots$	$\dots$		$\dots$	$\dots$
i	$x_{i,1}$		$x_{i,d}$	$Y_i \in \{0,1\}$
n	$\dots$		$\dots$	1

Table 14.1: Problem description logistic regression

14.4 maximum likelihood method

To specify a model, the function $t(\cdot)$ is often assumed to depend on some parameters β and we shall write $t_\beta(\cdot)$. The likelihood function of the observations (x_i, Y_i), $i = 1, \dots, n$, is

$$L(\beta \mid x) \coloneqq \prod_{\{i\colon Y_i=0\}} \big(1 - F_\varepsilon(t_\beta(x_i))\big) \cdot \prod_{\{i\colon Y_i=1\}} F_\varepsilon\big(t_\beta(x_i)\big),$$

the log-likelihood function is

$$\ell(\beta \mid x) = \log L(\beta \mid x) = \sum_{\{i\colon Y_i=0\}} \log\big(1 - F_\varepsilon(t_\beta(x_i))\big) + \sum_{\{i\colon Y_i=1\}} \log F_\varepsilon(t_\beta(x_i)) \tag{14.6}$$

with observations $Y_i \coloneqq Y(x_i) + \varepsilon_i$ corresponding to the explanatory variables $x_i = (x_{i,1}, \dots x_{i,d})$, as Table 14.1 indicates. The best fitting parameters can be determined by the maximum likelihood method,

$$\hat\beta \in \operatorname*{arg\,max}_\beta L(\beta \mid x) = \operatorname*{arg\,max}_\beta \ell(\beta \mid x).$$

The first-order conditions to be solved are the nonlinear equations

$$\sum_{\{i\colon Y_i=0\}} \frac{F'_\varepsilon(t_\beta(x_i))}{1 - F_\varepsilon(t_\beta(x_i))} \frac{\partial}{\partial\beta_j} t_\beta(x_i) = \sum_{\{i\colon Y_i=1\}} \frac{F'_\varepsilon(t_\beta(x_i))}{F_\varepsilon(t_\beta(x_i))} \frac{\partial}{\partial\beta_j} t_\beta(x_i), \quad j = 0, \dots, d. \tag{14.7}$$

These equations can be solved by employing Newton's method, e.g.

14.5 logistic regression

For a logistic regression with $F_\varepsilon = S$ we have with (14.1) that $F'_\varepsilon(t) = S'(t) = S(t)\big(1 - S(t)\big)$ and thus

$$\sum_{\{i\colon Y_i=0\}} S(t_\beta(x_i)) \frac{\partial}{\partial\beta_j} t_\beta(x_i) = \sum_{\{i\colon Y_i=1\}} \big(1 - S(t_\beta(x_i))\big) \frac{\partial}{\partial\beta_j} t_\beta(x_i), \quad j = 0, \dots, d. \tag{14.8}$$

A particular choice in practice for $t_\beta(\cdot)$ is the *linear ansatz*

$$t_\beta(x) \coloneqq \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d. \tag{14.9}$$

For the linear model (14.9) we have $\frac{\partial}{\partial\beta_0} t_\beta(x_i) = 1$ and $\frac{\partial}{\partial\beta_j} t_\beta(x_i) = x_{ij}$, $j = 1, \dots d$ in (14.7).

The logistic regression can be understood in finding the parameters $\hat\beta = (\hat\beta_0, \dots, \hat\beta_d)$ that best fit the problem (14.3), i.e.,

$$Y(x) = \begin{cases} 1 & \text{if } \tilde\beta_0 + \tilde\beta_1 x_1 + \dots + \tilde\beta_d x_d \ge \varepsilon, \\ 0 & \text{else.} \end{cases}$$

14.6 THE PROBIT MODEL AND PROBIT REGRESSION

The probit function is the quantile function of the standard normal distribution, i.e.,

$$\operatorname{probit}(p) := \Phi^{-1}(p), \qquad p \in (0,1),$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2} z^2\right) \mathrm{d}z$$

is the cdf of the standard normal distribution.

If $\varepsilon \sim \mathcal{N}(0,1)$ follows a standard normal distribution in (14.3) and (14.4) (instead of $\varepsilon \sim S_{0,1}$), then all formulae above modify with Φ instead of S. However, the equations corresponding to (14.7) are not so pleasant any longer. Further, the logit tails are heavier than probit tails and for this logistic regression is often more robust compared to probit.

14.7 PROBLEMS

Exercise 14.1 (The binomial model)**.** *Suppose that x_i are irrelevant parameters and the problem specification is $t(x) = \beta_0$. The observations are $k := |\{i \colon Y_i = 1\}|$ and $n - k := |\{i \colon Y_i = 0\}|$. Derive from (14.6) that $S(\hat{\beta}) = \frac{k}{n}$ and $\hat{\beta}_0 = S^{-1}\left({}^{k}/_{n}\right)$ and thus $P(Y = 1) = \frac{k}{n}$.*

15 Distances

> To be deep in history is to cease to be a Protestant.
>
> John Henry Newman, Introduction p. 5

Consider the measure space $(\mathcal{X}, \mathcal{F})$.

Definition 15.1. Let P and Q be probability measures.

(i) The total variation is

$$TV(P, Q) \coloneqq \|P - Q\|_\infty \coloneqq \sup_{A \in \mathcal{F}} |P(A) - Q(A)|\,.$$

(ii) Suppose that $Q \ll \mu$ and $P \ll \mu$ with Radon–Nikodym derivatives $\mathrm{d}P = f\,\mathrm{d}\mu$ and Q $= g\,\mathrm{d}\mu$. For $r \ge 1$, the metric d_r is

$$d_r(P, Q) \coloneqq \left(\int_{\mathcal{X}} \left|f^{1/r} - g^{1/r}\right|^r \mathrm{d}\mu\right)^{1/r} = \left\|f^{1/r} - g^{1/r}\right\|_r.$$

(iii) The Hellinger[1] distance is

$$H(P, Q) \coloneqq \frac{1}{\sqrt{2}} d_2(P, Q).$$

Lemma 15.2. *It holds that*

$$TV(P, Q) = \sup_{\varphi \in \Phi} \left|\int_{\mathcal{X}} \varphi\,\mathrm{d}P - \int_{\mathcal{X}} \varphi\,\mathrm{d}Q\right| = \frac{1}{2} d_1(P, Q), \tag{15.1}$$

where $\varphi \in \Phi$ is a random variable with $\varphi \in [0, 1]$ (i.e., a statistical test).

Proof. Note first that $P(A) - Q(A) = Q\left(A^{\mathrm{c}}\right) - P\left(A^{\mathrm{c}}\right)$ and thus $\|P - Q\| = \sup_{A \in \mathcal{F}} P(A) - Q(A) = \|Q - P\|$. Further $P(A) - Q(A) = \int_A f - g\,\mathrm{d}\mu \le \int_{\{f > g\}} f - g\,\mathrm{d}\mu = \sup_{\varphi \in \Phi} \mathbb{E}_P\,\varphi - \mathbb{E}_Q\,\varphi$, the supremum is thus attained for $A^* = \{f > g\}$ and $\varphi = \mathbb{1}_{A^*}$ and hence $\|P - Q\| = P(\{f > g\}) - Q(\{f > g\})$. Further we have that

$$\begin{aligned} d_1(P, Q) &= \int_{\mathcal{X}} |f - g|\,\mathrm{d}\mu = \int_{\{f > g\}} f - g\,\mathrm{d}\mu + \int_{\{g > f\}} g - f\,\mathrm{d}\mu \\ &= \big(P(\{f > g\}) - Q(\{f > g\})\big) + \big(Q(\{g > f\}) - P(\{g > f\})\big) \\ &= \|P - Q\| + \|Q - P\| = 2\,\|Q - P\| \end{aligned}$$

and thus the result. □

[1] Ernst Hellinger, 1883–1950, German mathematician

Corollary 15.3. *It holds that*

$$2TV(P,Q)=\sup_{\varphi\in[-1,1]}\left|\int_X\varphi\,\mathrm{d}P-\int_X\varphi\,\mathrm{d}Q\right|=d_1(P,Q),$$

where the supremum is among random variables φ with range $\varphi\in[-1,1]$: the supremum is attained for $\varphi=\mathbb{1}_{\{f>g\}}-\mathbb{1}_{\{f<g\}}$.

Lemma 15.4. *It holds that $d_r(P,Q)\le 2^{1/r}$ and thus $0\le TV(P,Q)\le 1$ and $0\le H(P,Q)\le 1$.*

Proof. Indeed, $\left|a-b\right|^r\le a^r+b^r$ and thus $d_r(P,Q)=\left(\int_X\left|f^{1/r}-g^{1/r}\right|^r\,\mathrm{d}\mu\right)^{1/r}\le\left(\int_X f+g\,\mathrm{d}\mu\right)^{1/r}=2^{1/r}$. □

The distances TV and H are topologically equivalent.

Proposition 15.5. *It holds that $H(P,Q)^2\le\left\|P-Q\right\|\le\sqrt{2}H(P,Q)$.*

Proof. By the inequality of the arithmetic and geometric means we have that $\sqrt{fg}\le\frac{1}{2}(f+g)$. Hence, by Cauchy–Schwarz,

$$\begin{aligned}\left\|P-Q\right\|&=\frac{1}{2}\int_X|f-g|\,\mathrm{d}\mu\\&=\frac{1}{2}\int_X\left(\sqrt{f}+\sqrt{g}\right)\cdot\left|\sqrt{f}-\sqrt{g}\right|\,\mathrm{d}\mu\\&\le\frac{1}{2}\left(\int_X\left(\sqrt{f}+\sqrt{g}\right)^2\,\mathrm{d}\mu\right)^{1/2}\cdot\left(\int_X\left(\sqrt{f}-\sqrt{g}\right)^2\,\mathrm{d}\mu\right)^{1/2}\\&=\frac{\sqrt{2}}{2}\left(\int_X f+g+2\sqrt{fg}\,\mathrm{d}\mu\right)^{1/2}\cdot\frac{1}{\sqrt{2}}\left(\int_X\left(\sqrt{f}-\sqrt{g}\right)^2\,\mathrm{d}\mu\right)^{1/2}\\&\le\sqrt{2}H(P,Q).\end{aligned}$$

Further note that $f\wedge g\le\sqrt{fg}$ and thus $f+g-2\sqrt{fg}\le f+g-2f\wedge g=|f-g|$ and

$$\begin{aligned}H(P,Q)^2&=\frac{1}{2}\int_X\left(\sqrt{f}-\sqrt{g}\right)^2\,\mathrm{d}\mu\\&=\frac{1}{2}\int_X f+g-2\sqrt{fg}\,\mathrm{d}\mu\\&\le\frac{1}{2}\int_X|f-g|\,\mathrm{d}\mu\\&=\|P-Q\|,\end{aligned}$$

by (15.1), thus the assertion. □

Definition 15.6. The Kullback[2]–Leibler[3] divergence between P and Q is

$$D_{KL}(P,Q):=\begin{cases}\int_X\left(\log\frac{f}{g}\right)f\,\mathrm{d}\mu & \text{if } P\ll Q,\\ +\infty & \text{else.}\end{cases}$$

[2] Solomon Kullback, 1907–1994, US

[3] Richard Leibler, 1914–2003, US

Lemma 15.7 (First Pinsker's inequality). *It holds that* $\|P-Q\|^2 \le \frac{1}{2} D_{KL}(P,Q)$.

Proof. (The proof follows Tsybakov [19, Lemma 2.5]). Consider the function $\psi(x) = x\log x - x + 1$ and observe that $\psi(1) = 0$, $\psi'(1) = 0$ and $\psi''(x) = \frac{1}{x} \ge 0$, hence $\psi(x) \ge 0$ for all $x \ge 0$ by Taylor's theorem.

Define $g(x) := (x-1)^2 - \left(\frac{4}{3} + \frac{2}{3}x\right)\psi(x)$ and observe that $g(1) = 0$, $g'(1) = 0$ and $g''(x) = -\frac{4\psi(x)}{3x} \le 0$ and thus (again by Taylor's theorem) $g(x) \le 0$, i.e., $(x-1)^2 \le \left(\frac{4}{3} + \frac{2}{3}x\right)\psi(x)$. It follows that

$$\begin{aligned}
\|P-Q\| &= \frac{1}{2}\int_{\mathcal{X}} |f-g|\,\mathrm{d}\mu = \frac{1}{2}\int_{\{g>0\}} \left|\frac{f}{g} - 1\right| g\,\mathrm{d}\mu \\
&\le \frac{1}{2}\int_{\{g>0\}} g\sqrt{\left(\frac{4}{3} + \frac{2}{3}\frac{f}{g}\right)\psi\left(\frac{f}{g}\right)}\,\mathrm{d}\mu = \frac{1}{2}\int_{\{g>0\}} \sqrt{\frac{4}{3}g + \frac{2}{3}f}\cdot\sqrt{g\,\psi\left(\frac{f}{g}\right)}\,\mathrm{d}\mu \\
&\le \frac{1}{2}\sqrt{\int_{\mathcal{X}} \frac{4}{3}g + \frac{2}{3}f\,\mathrm{d}\mu}\cdot\sqrt{\int_{\{g>0\}} g\,\psi\left(\frac{f}{g}\right)\mathrm{d}\mu} \qquad (15.2) \\
&= \frac{1}{2}\sqrt{2}\sqrt{\int_{\mathcal{X}} f\log\frac{f}{g}\,\mathrm{d}\mu} = \sqrt{\frac{1}{2}D_{KL}(P,Q)},
\end{aligned}$$

where we have used Cauchy–Schwarz in (15.2). Thus the assertion. □

Theorem 15.8 (Villani [20]). *It holds that* $\|P-Q\| = \inf_\pi \mathbb{E}_\pi \mathbb{1}_{x\ne y}$, *where* π *is a bivariate probability measure on* $\mathcal{X}\times\mathcal{X}$ *with marginals* P *and* Q, *i.e.,* $\pi(A\times\mathcal{X}) = P(A)$ *and* $\pi(\mathcal{X}\times B) = Q(B)$.

Lemma 15.9 (Scheffé's lemma[4]). *Suppose that* $f_n \to f$ μ*-a.e., then*

$$\int_{\mathcal{X}} |f_n - f|\,\mathrm{d}\mu \xrightarrow[n\to\infty]{} 0 \text{ iff } \int_{\mathcal{X}} |f_n|\,\mathrm{d}\mu \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} |f_n|\,\mathrm{d}\mu.$$

Proof. Indeed, by Fatou's lemma,

$$\begin{aligned}
2\int_{\mathcal{X}} f\,\mathrm{d}\mu &= \int_{\mathcal{X}} \liminf_{n\to\infty} f_n + f - |f_n - f|\,\mathrm{d}\mu \\
&\le \liminf_{n\to\infty} \int_{\mathcal{X}} f_n + f - |f_n - f|\,\mathrm{d}\mu \\
&= 2\int_{\mathcal{X}} f\,\mathrm{d}\mu - \limsup_{n\to\infty} \int_{\mathcal{X}} |f_n - f|\,\mathrm{d}\mu;
\end{aligned}$$

it follows that $0 \le \limsup_{n\to\infty} \int_{\mathcal{X}} |f_n - f|\,\mathrm{d}\mu \le 0$ and thus the result. □

15.1 PROBLEMS

Exercise 15.1. *Show that* d_r *and* H *are distances.*

Exercise 15.2. *Verify Corollary 15.3.*

[4] Henry Scheffé, 1907–1977. The result, however, is based on a result by Frigyes Riesz, 1880–1956

16 Families

> Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.
>
> Pierre-Simon Laplace, 1749–1827

16.1 MONOTONE LIKELIHOOD RATIOS

16.2 EXPONENTIAL FAMILIES

Definition 16.1. The density of an exponential family is

$$f(x|\vartheta) = c(\vartheta)h(x)e^{\vartheta^\top t(x)}.$$

Equivalent forms include $f(x|\vartheta) = h(x)\exp\left(\eta(\vartheta)\cdot T(x) - A(\vartheta)\right)$ or $f(x|\vartheta) = \exp\left(\eta(\vartheta)\cdot T(x) - A(\vartheta) + B(x)\right)$.

Example 16.2 (Poisson Distribution)**.** The pmf is $f(x|\alpha) = e^{-\alpha}\cdot\frac{1}{x!}\cdot e^{\log\alpha\cdot x}$, thus $t(x) = x$.

Example 16.3 (Exponential distribution)**.** The pmf is $f(x|\alpha) = e^{-\alpha}\cdot\frac{1}{x!}\cdot e^{\log\alpha\cdot x}$, thus $t(x) = x$.

Example 16.4 (Normal distribution: unknown mean, known variance)**.** The density is $f_\sigma(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$: it is easily seen that $T_\sigma(x) := \frac{x}{\sigma}$, $h_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}$, $A(\mu) = \frac{\mu^2}{2\sigma^2}$ and $\eta_\sigma(\mu) = \frac{\mu}{\sigma^2}$.

For a vector of independent variables, $f_\sigma(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}^n}e^{-\sum_{i=1}^n\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}^n}e^{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)}$ thus $t(x) = \frac{1}{\sigma}\sum_{i=1}^n x_i$.

16.3 SUFFICIENT FAMILIES

» Sufficiency.

Definition 16.5. A satsitic $t = T(x)$ is sufficient for the underlying parameter ϑ, if

$$P(x \mid t, \vartheta) = P(x \mid t),$$

i.e., the probability of x given $T(x)$ does not independent on ϑ.

Theorem 16.6 (Fisher–Neyman factorization theorem)**.** *The statistics t is sufficient for ϑ if and only if nonnegative functions g and h can be found such that (importantly, h does not depend on ϑ)*

$$f(x|\vartheta) = h(x)\cdot g_\vartheta\big(t(x)\big).$$

Remark 16.7. If f is an exponential family, then t is sufficient.

distribution	minimal sufficient statistics
Binomial, $\text{bin}(n, p)$, n fixed	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$
Poisson, P_α	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$
$L_{a,b}$, $-\infty < a < b < \infty$	$X_{(1)}, X_{(n)}$
Exponential E_λ	$\sum_{i=1}^n X_i$ and thus $\overline{X}_n$
Erlang, $E_{n,\lambda}$	$\sum_{i=1}^n X_i$ and $\sum_{i=1}^n \log X_i$
Normal $\mathcal{N}(\mu, \sigma^2)$	$\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$, and thus $\overline{X}_n$ and V_n

Table 16.1: Minimal sufficient statistics

Nonparametric Statistics — Density Estimation

17

> Wer nichts weiß muss alles glauben.
>
> Marie von Ebner-Eschenbach, 1830–1916

We address essential convergence theorems from multivariate kernel density estimation first. The general assumption for a kernel $k\colon \mathbb{R}^d \to \mathbb{R}$ is that

(i) $k(\cdot) \geq 0$,

(ii) $\int_{\mathbb{R}^d} k(u)\,\mathrm{d}u = 1$ and

(iii)
$$\int_{\mathbb{R}^d} u_i \cdot k(u)\, du = 0 \tag{17.1}$$
for all $i = 1, \dots, d$.

Rescaling: note that $k_h(u) := \frac{1}{h^d} k\left(\frac{u}{h}\right)$, $h > 0$ (more generally, $\frac{1}{h_1 \cdots h_d} k\left(\frac{u_1}{h_1}, \dots, \frac{u_d}{h_d}\right)$) is a kernel, provided that $k(\cdot)$ is a kernel.

Definition 17.1 (Kernel density estimaton (KDE))**.** The kernel density estimator for data $X_i \in \mathbb{R}^d$, $i = 1, \dots, n$, is
$$\hat{f}_n(x) := \frac{1}{n} \sum_{i=1}^n k_h\left(x - X_i\right),$$
where $h > 0$ is the *bandwidth*. In some fields such as signal processing and econometrics it is also termed the *Parzen–Rosenblatt window* method.[1]

Remark 17.2. Note, that $x \mapsto \hat{f}_n(x)$ is a density on $\mathbb{R}^d$ for every $n \in \{1, 2, \dots\}$.

Compare with the histogram.

17.1 THE BIAS TERM

The bias of the density estimator $\hat{f}_n(\cdot)$ can be expressed as
$$\mathbb{E}\, \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} k_h(x-y)\, f(y)\, \mathrm{d}y = (f * k_h)(x), \tag{17.2}$$
where $*$ denotes the usual convolution of densities. It follows from (17.2) that $\hat{f}_n(x)$ is biased in general. The bias can be stated as
$$\begin{aligned} \operatorname{bias} \hat{f}_n(x) := \mathbb{E}\, \hat{f}_n(x) - f(x) &= \frac{1}{n\, h^d} \sum_{i=1}^n \int_{\mathbb{R}^d} k\left(\frac{x-y}{h}\right) \big(f(y) - f(x)\big)\, \mathrm{d}y \\ &\underset{y \leftarrow x - h\, u}{=} \int_{\mathbb{R}^d} k(u) \big(f(x - h \,.\, u) - f(x)\big)\, \mathrm{d}u, \end{aligned} \tag{17.3}$$

[1] Emanuel Parzen, 1929–2016, Murray Rosenblatt, 1926

where we denote the entrywise product (Hadamard product) by $h\,\boldsymbol{.}\,u = (h \cdot u_i)_{i=1}^d$ (similarly for matrices).

It is evident that $\mathbb{E}\,\hat{f}_n(x) \to f(x)$ whenever $h_n \to 0$ and if x is a point of continuity of f. Indeed, by assuming that f is smooth and employing a Taylor series expansion (17.3) reduces to

$$\begin{aligned}\operatorname{bias} \hat{f}_n(x) &= \int_{\mathbb{R}^d} k\,(u) \left(f(x) - f'(x)^\top h\,\boldsymbol{.}\,u + \frac{1}{2}(h\,\boldsymbol{.}\,u)^\top f''(x)(h\,\boldsymbol{.}\,u) - f(x) + o(h^2) \right) \mathrm{d}u \\ &= \frac{1}{2} h^\top \big(f''(x)\,\boldsymbol{.}\,\kappa \big) h + o(h_{max}^2), \qquad (17.4)\end{aligned}$$

whenever (17.1) holds and where κ is the matrix with entries $\kappa_{ij} = \iint u_i u_j k(u)\,\mathrm{d}u$. Note that the expression (17.3), as well as the approximation (17.4) are deterministic quantities, they do not involve any random component. Instead, the bias depends on the density function f and its smoothness, or (local) differentiability. Moreover it should be noted that the bias tends asymptotically to 0 in (17.3) and (17.4), provided that $h_{max} = \max\{h_1, \dots h_d\} \to 0$.

17.2 MEAN SQUARED ERROR

The variance of the multivariate kernel statistics is

$$\begin{aligned}\operatorname{var} \hat{f}_n(x) &= \operatorname{var} \frac{1}{n \cdot h_1 \dots h_d} \sum_{i=1}^n k\left(\frac{x - X_i}{h}\right) = \frac{n}{n^2} \operatorname{var} \frac{1}{h_1 \dots h_d} k\left(\frac{x - X_1}{h}\right) \\ &= \frac{1}{n} \int_{\mathbb{R}^d} \frac{1}{h_1^2 \dots h_d^2} k\left(\frac{x-y}{h}\right)^2 f(y)\,\mathrm{d}y - \frac{1}{n}\left(\frac{1}{h_1 \dots h_d} \int k\left(\frac{x-y}{h}\right) f(y)\,\mathrm{d}y\right)^2 \\ &= \frac{1}{n \cdot h_1 \dots h_d} \int_{\mathbb{R}^d} k\,(u)^2 f(x - h\,\boldsymbol{.}\,u)\,\mathrm{d}u - \frac{1}{n}\left(\int k\,(u)\, f(x - hu)\,\mathrm{d}u\right)^2 \\ &= \frac{f(x)}{n \cdot h_1 \dots h_d} \int_{\mathbb{R}^d} k(u)^2\,\mathrm{d}u + \frac{f(x)^2}{n} + o\left(\frac{1}{n \cdot h_1 \dots h_d}\right), \qquad (17.5)\end{aligned}$$

and the *mean squared error* is given by (cf. (12.9)) is

$$\begin{aligned}\operatorname{mse} \hat{f}_n(x) &:= \mathbb{E}\left(\hat{f}_n(x) - f(x)\right)^2 = \left(\operatorname{bias} \hat{f}_n(x)\right)^2 + \operatorname{var} \hat{f}_n(x). \qquad (17.6) \\ &= \left(\frac{1}{2} h^\top \big(f''(x)\,\boldsymbol{.}\,\kappa\big) h\right)^2 + \frac{f(x)}{n \cdot h_1 \dots h_d} \int_{\mathbb{R}^d} k(u)^2\,\mathrm{d}u + o(\cdots) \qquad (17.7)\end{aligned}$$

To minimize the mean squared error with respect to a particular direction h_i it is advantageous to get rid of the mixed terms $h_i h_j$ $(i \neq j)$ in (17.4) for the bias. This can be accomplished by assuming that

$$\kappa_{ij} = \int_{\mathbb{R}^d} u_i\, u_j\, k(u)\,\mathrm{d}u = 0 \text{ whenever } i \neq j. \qquad (17.8)$$

minimized for

$$h_n^{d+4} \simeq \frac{d}{n} \cdot \frac{f(x) \cdot \int_{\mathbb{R}^d} k(u)^2\,\mathrm{d}u}{\left(\sum_{i=1}^d \kappa_{i,i} \cdot \frac{\partial^2}{\partial x_i^2} f\right)^2}, \text{ i.e., } h_n \simeq \left(\frac{d}{n} \cdot \frac{f(x) \cdot \int_{\mathbb{R}^d} k(u)^2\,\mathrm{d}u}{\left(\sum_{i=1}^d \kappa_{i,i} \cdot \frac{\partial^2}{\partial x_i^2} f\right)^2}\right)^{\frac{1}{d+4}}, \qquad (17.9)$$

which is

$$h_n = n^{-\frac{1}{5}} \cdot \left(\frac{f(x)}{f''(x)^2}\right)^{1/5} \cdot \left(\frac{\int_{-\infty}^{\infty} k(u)^2\,\mathrm{d}u}{\left(\int_{-\infty}^{\infty} u^2\,k(u)\,\mathrm{d}u\right)^2}\right)^{1/5}$$

in dimension $d = 1$.

Minimizing the mean squared error. Minimizing the mean squared error includes minimizing (17.5) over all potential candidates of kernels. The following result provides an answer.

Proposition 17.3 (Epanechnikov)**.** *On $\mathbb{R}^1$, consider kernels with the properties*

(i) $k(\cdot) \geq 0$,
(ii) $\int_{-\infty}^{\infty} k(u)\,\mathrm{d}u = 1$ *and*
(iii) $\int_{-\infty}^{\infty} u^2\,k(u)\,\mathrm{d}u = 1$ *(scaling).*

The Epanechnikov kernel[2] $k_E(u) \coloneqq \frac{3\sqrt{5}}{20}\left(1 - \frac{u^2}{5}\right) \cdot \mathbb{1}_{[-\sqrt{5},\sqrt{5}]}(u)$ *satisfies these properties and minimizes*

$$\int_{-\infty}^{\infty} k(u)^2\,\mathrm{d}u$$

among all kernels with the above properties (i)–(iii). We have $\int_{-\infty}^{\infty} k_E(u)^2\,\mathrm{d}u = \frac{3\sqrt{5}}{25} = 0.268\ldots$*.*

Proof. Note first that $1 - \frac{u^2}{5} < 0$ and $k(u) \geq 0 = k_E(u)$ whenever $|u| > \sqrt{5}$. We thus have

$$\begin{aligned}\int_{-\infty}^{\infty} \big(k(u) - k_E(u)\big) \cdot k_E(u)\,\mathrm{d}u &\geq \frac{3\sqrt{5}}{20}\int_{-\infty}^{\infty} \big(k(u) - k_E(u)\big) \cdot \left(1 - \frac{u^2}{5}\right)\mathrm{d}u \\ &= \frac{3\sqrt{5}}{20}\int_{-\infty}^{\infty} k(u) - k_E(u)\,\mathrm{d}u - \frac{3\sqrt{5}}{100}\int_{-\infty}^{\infty} u^2 k(u) - u^2 k_E(u)\,\mathrm{d}u = 0\end{aligned}$$

by (ii) and (iii). Hence

$$\begin{aligned}\int_{-\infty}^{\infty} k(u)^2\,\mathrm{d}u &= \int_{-\infty}^{\infty} \big(k(u) - k_E(u) + k_E(u)\big)^2\,\mathrm{d}u \\ &= \int_{-\infty}^{\infty} \big(k(u) - k_E(u)\big)^2\,\mathrm{d}u + 2\cdot\int_{-\infty}^{\infty} \big(k(u) - k_E(u)\big)\,k_E(u)\,\mathrm{d}u + \int_{-\infty}^{\infty} k_E(u)^2\,\mathrm{d}u \\ &\geq 0 + 2\cdot 0 + \int_{-\infty}^{\infty} k_E(u)^2\,\mathrm{d}u = \frac{3\sqrt{5}}{25}\end{aligned}$$

and thus the assertion. □

Fact 17.4 (Silverman's rule of thumb)**.** *In practice, Silverman's rule of thumb is often used, i.e.,*

$$h_n = \frac{1.06}{n^{1/5}} s_n.$$

[2] Occasionally, with a different scaling than (iii), the kernel is defined on the support $|u| \leq 1$ by $k_E(u) = \frac{3}{4}\left(1 - u^2\right)$.

17.3 INTEGRATED MEAN SQUARED ERROR

If, instead of the mean squared error at a specific point x in (17.6), the *mean integrated square error* (compare with (12.5) for the distribution function)

$$\operatorname{mise} \hat{f}_n := \int_{\mathbb{R}^d} \operatorname{mse}\left(\hat{f}_n(x)\right) f(x)\,\mathrm{d}x = \mathbb{E}\int_{\mathbb{R}^d}\left(\hat{f}_n(x) - f(x)\right)^2 \mathrm{d}x$$

is to be minimized, then the optimal bandwidth is

$$h_n^{d+4} \simeq \frac{d}{n} \cdot \frac{\int_{\mathbb{R}^d} k(u)^2\,\mathrm{d}u}{\left(\sum_{i=1}^m \kappa_{i,i} \int_{\mathbb{R}^d} f_{x_i x_i}\,\mathrm{d}x\right)^2}, \tag{17.10}$$

which is the same order as in (17.9).[3]

Remark 17.5. Assumption (17.8) is an assumption on the kernel k. Any kernel exhibiting the product form

$$k(u) = k_1(u_1) \cdot k_2(u_2) \cdot \ldots k_d(u_d)$$

satisfies this assumption. The bias (17.4) of a product kernel of the particular form $k(u) = k(u_1) \cdot k(u_2) \cdot \ldots k(u_d)$ reduces to

$$\operatorname{bias} \hat{f}_n(x) \quad = \quad \frac{\kappa_2}{2}\sum_{s=1}^d h_s^2 f_{x_s x_s}(x) + o\big(\max_{s=1,\ldots d} h_s^2\big),$$

where

$$\kappa^{(2)} := \int_{\mathbb{R}^d} u^2\, k(u)\,\mathrm{d}u \tag{17.11}$$

is the second moment (or variance) of the distribution associated with the kernel.

Remark 17.6. Both formulae ((17.9) and (17.10)) for the asymptotic optimal bandwidth involve f'', the Hessian of the density function f. As the function f is unknown (this is what kernel density estimation intends to estimate) the formulae provide the correct asymptotic order, but the optimal constant remains an oracle (cf. Tsybakov [19]). Different methods to obtain an optimal bandwidth as cross-validation are designed to overcome this difficulty and outlined in Racine et al. [16], e.g., or plug-in rules of Sheather [18].

Asymptotic normality. The kernel density estimator (2.9) is a sum of independent, identically distributed random variables. Evoking the central limit theorem (CLT, Theorem 4.3) for independent identically distributed random variables, it is expected that after correcting the bias (17.4), the estimator $\hat{f}_n(x)$ satisfies the CLT

$$\sqrt{n\,h_1 \ldots h_d}\left(\hat{f}_n(x) - f(x) - \frac{\kappa_{(2)}}{2}\sum_{s=1}^d h_s^2 \frac{\partial^2}{\partial x_i^2} f\right) \xrightarrow{d} \mathcal{N}\left(0,\, f(x)\cdot \kappa_{(2)}^d\right), \tag{17.12}$$

where

$$\kappa_{(2)} := \int k(u)^2\,\mathrm{d}u$$

(notice the difference to 17.11). This is indeed the case, as is shown in Li and Racine [10, Theorem 1.3] under mild regularity conditions by employing Liapunov's central limit theorem for triangular arrays.

[3] Note, that $\sum_{i=1}^m \kappa_{i,i} f_{x_i x_i} = \operatorname{div}(\kappa\,.\,\nabla f)$, and $\sum_{i=1}^m \kappa_{i,i} f_{x_i x_i} = \kappa\,\Delta f$ (the Laplace operator) for constant $\kappa_{i,i} = \kappa$.

Remark 17.7 (Over- and undersmoothing). Notice that the bias term in (17.12) cannot be dropped if the bandwidth is chosen as proposed in (17.9) or (17.10), because $\sqrt{n\,h_1\dots h_d}\cdot h_n^2 \sim 1$ whenever $h_n \sim n^{-1/(d+4)}$. By choosing $h_n \sim n^{-\alpha}$ for some $\alpha > {}^1\!/\!_{(d+4)}$, the bias is asymptotically negligible relative to $\hat{f}_n - f$. This is known as undersmoothing.

In case of oversmoothing (for example if $h_n \sim n^{-\alpha}$ and $\alpha < {}^1\!/\!_{(d+4)}$) the normalized term $\sqrt{n\,h_1\dots h_d}\cdot \left(\hat{f}_n - f\right)$ in (17.12) diverges, but $\hat{f}_n - f$ still converges.

18 Analysis of variance — ANOVA

I have nothing to say,
I am saying it,
and that is poetry
as I need it.

John Cage, 1912–1992

By Dana. See https://www.tu-chemnitz.de/mathematik/fima/public/anova.pdf

19 Principal Component Analysis

> Erfunden. Alles erfunden.
>
> Johann Wolfgang von Goethe, 1749–1832, in *Faust II*

19.1 LINEAR MODEL WITH NORMAL ERRORS

Consider the linear model with observations $X_i = \mu + L\,\varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \mathbb{1})$ are independent multivariate normals. Recall, that

$$X \sim \mathcal{N}\left(\mu, \Sigma\right), \tag{19.1}$$

where $\Sigma = L\,L^\top = \operatorname{cov} X$.

Let W be the orthonormal matrix ($WW^\top = \mathbb{1}_d$) providing the factorization (eigendecomposition) of the covariance matrix, $\operatorname{cov}(X) = W\,\Lambda\,W^\top$, or

$$\operatorname{cov}(X) \cdot W = W \cdot \Lambda \text{ and } \Lambda := \begin{pmatrix} \lambda_1 & 0 & \ddots \\ 0 & \ddots & 0 \\ \ddots & 0 & \lambda_d \end{pmatrix}.$$

The matrix $W = (w_1 \mid \dots \mid w_d)$, with columns w_j, does not change lengths.

Remark 19.1. Note, that $L' := W\Lambda^{1/2}$ and $L'' := W\Lambda^{1/2}W^\top$ satisfy $L'\,L'^\top = L''\,L''^\top = L\,L^\top = \Sigma$.

19.2 REDUCING THE DIMENSIONALITY OF THE PROBLEM

Definition 19.2. The transformed variable $\tilde{X} := W^\top X$ is called *principal component*. Note, that $X = W\,\tilde{X}$.

Without loss of generality we assume that $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_d$. The eigenvalue λ_j is the variance of the j-th principal component (the first principal component has the largest possible variance).

Remark 19.3 (Truncated transform). For $d' \le d$, the first d' principal components are

$$(w_1 \mid \dots \mid w_{d'})^\top \cdot X = (\tilde{X}_1, \dots, \tilde{X}_{d'}) \in \mathbb{R}^{d'}.$$

The fraction of the total variance *explained* by the first d' principal components is $\frac{\sum_{j=1}^{d'} \lambda_j}{\sum_{j=1} \lambda_j}$.

Proposition 19.4. *The principal component follows the distribution*

$$\tilde{X} \sim \mathcal{N}\left(W^\top b + W^\top A\,\mu,\ \Lambda\right) \tag{19.2}$$

and its coordinates are independent *(cf. Theorem 3.16).*

Further, it holds that

$$X \sim \mu + W \cdot \Lambda^{1/2} \cdot \mathcal{N}\left(0, \mathbb{1}_d\right). \tag{19.3}$$

Proof. Using (3.6) it holds that

$$\tilde{X} = W^\top X \sim \mathcal{N}\left(W^\top \mu,\, W^\top \Sigma W\right) \sim \mathcal{N}\left(W^\top \mu,\, \Lambda\right).$$

The coordinates or $\tilde{X}$ are independent, as Λ in (19.2) is a diagonal matrix.

Eq. (19.3) follows from (19.2) by shifting the mean and the fact that $\mathcal{N}(0,\lambda) \sim \sqrt{\lambda}\,\mathcal{N}(0,1)$ for univariate normals, i.e.,

$$X \sim W\tilde{X} \sim WW^\top \mu + W \cdot \Lambda^{1/2} \cdot \mathcal{N}\left(0,\, \mathbb{1}_d\right).$$

□

Remark 19.5. Useful approximations including only the first, most important principal components are

$$X_i' \approx \overline{X}_n + \sum_{j=1}^{d'} \xi_j^i \cdot \sqrt{\lambda_j} \cdot w_j$$

where $d' < d$ and w_j are the columns of $W = (w_1 \mid \cdots \mid w_d)$. Note, that $\overline{X}_n$ is an estimator for $b + A\mu$ and $\operatorname{cov}(X)$ has to be estimated as well from the empirical observations.

19.3 KARHUNEN–LOÈVE

Theorem 19.6 (Karhunen–Loève). *The linear model (19.1) has the expansion*

$$X_i \sim \mathbb{E}\,X + \sum_{j=1}^{d} \sqrt{\lambda_j} \cdot \xi_j^i \cdot w_j, \tag{19.4}$$

where ξ_j^i are all independent standard normals, $\xi_j^i \sim \mathcal{N}(0,1)$ and w_j are the columns of $W = (w_1, \dots, w_d)$.

Definition 19.7. The (Fourier series) expansion (19.4) is called Karhunen–Loève expansion of X.

Proposition 19.8. *The models (19.1) and (19.4) cannot be distinguished in distribution.*

Proof. Evident from (19.3). □

Examples

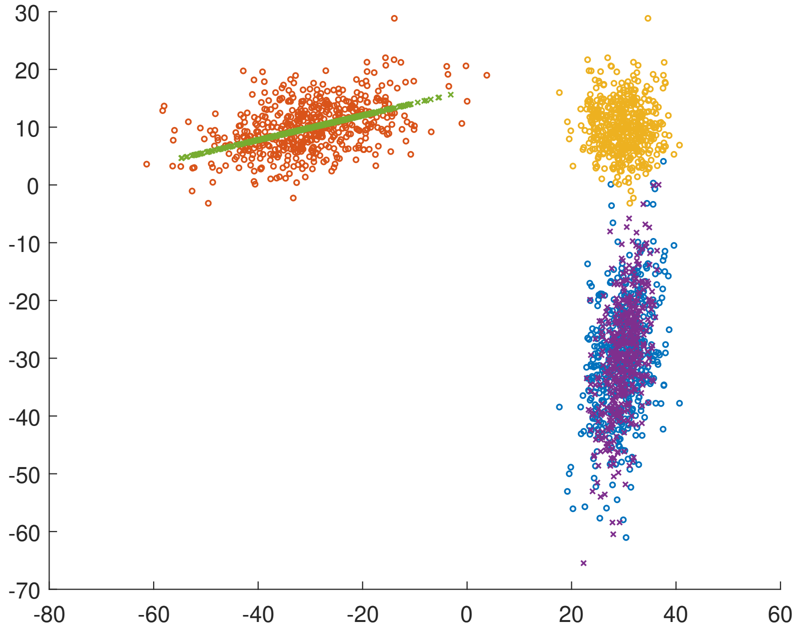
Figure 19.1a visualizes a 3-dimensional stochastic model, $X \in \mathbb{R}^3$; the scatter plot displays the pairs $(X_{i,1}, X_{i,2})$, $(X_{i,1}, X_{i,3})$ and $(X_{i,2}, X_{i,3})$. Included in the display as well are realizations of approximations using only the first (green) and the first two (purple) components.

Figure 19.1b displays realizations of non-normally distributed data $X \in \mathbb{R}^3$. However, the principal components are notably useful linear approximations of the non-normal and non-linear model: indeed, the data and PCA3 (includes all 3 components) are almost indistinguishable, but also PCA2 (includes only the first two components) explains a lot.

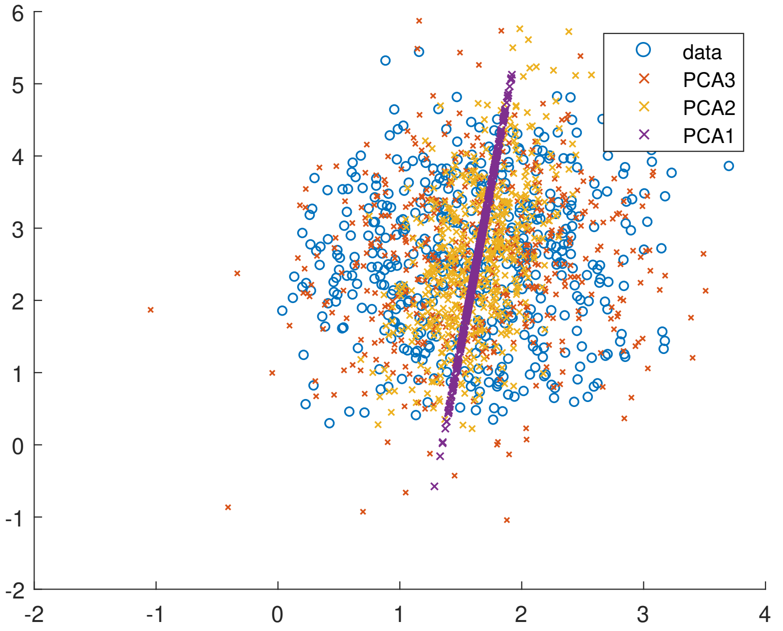
Lemma 19.9. *For the Euclidean norm it holds that* $\operatorname{var}\|X\| \le \mathbb{E}\,\|X - \mathbb{E}\,X\|^2 = \sum_{j=1}^d \lambda_j = \operatorname{trace}\Sigma$.

Proof. From Jensen's inequality we have that $\|\mathbb{E}\,X\| \le \mathbb{E}\,\|X\|$, hence

$$\operatorname{var}\|X\| = \mathbb{E}\,\|X\|^2 - (\mathbb{E}\,\|X\|)^2 \le \mathbb{E}\,\|X\|^2 - \|\mathbb{E}\,X\|^2 = \mathbb{E}\,\|X - \mathbb{E}\,X\|^2.$$



(a) Principal component analysis, 3 dimensional data



(b) PCA of non-normal data

Figure 19.1: Principal component analysis

Further,

$$\begin{aligned}\mathbb{E}\,\|X-\mathbb{E}\,X\|^2 &= \mathbb{E}\,\big\|\tilde{X}-\mathbb{E}\,\tilde{X}\big\|^2 = \mathbb{E}\sum_{j=1}^{d}\big(\tilde{X}_j-\mathbb{E}\,\tilde{X}_j\big)^2\\ &= \sum_{j=1}^{d}\mathbb{E}\,\big(\tilde{X}_j-\mathbb{E}\,\tilde{X}_j\big)^2 = \sum_{j=1}\operatorname{var}\tilde{X}_j = \sum_{j=1}\lambda_j = \operatorname{trace}\Sigma,\end{aligned}$$

as W is unitary and the coordinates of $\tilde{X}$ are independent. □

19.4 RELATION TO SINGULAR VALUE DECOMPOSITION AND THE SCORE

Assume that the model has 0 mean ($b=\mu=0$) and let $\mathbf{X}\in\mathbb{R}^{N\times d}$ denote the matrix collecting all N observations (repetitions). Then the estimated covariance is $\frac{1}{N}\mathbf{X}^\top\mathbf{X}$ and we have the singular value decomposition of the data by $\mathbf{X}=U\Lambda^{1/2}W^\top$ (here, $U\in\mathbb{R}^{N\times N}$ and $W\in\mathbb{R}^{d\times d}$ are orthonormal and

$$\hat{\Lambda}=\begin{pmatrix}\hat{\lambda}_1 & 0 & \ddots\\ 0 & \ddots & 0\\ \vdots & \ddots & \hat{\lambda}_d\\ 0 & \dots & 0\end{pmatrix}\in\mathbb{R}^{N\times d}$$

is a rectangular diagonal matrix carrying the squared singular values).

Definition 19.10. The score matrix is $T:=XW$.

Remark 19.11. The score T_{ij} provides the relative importance of the j-th principal component of the datum X_i. It holds that $T=U\Lambda^{1/2}$.

20 Extreme Value Statistics

Dimidium facti, qui coepit, habet:
sapere aude, incipe.

Quintus Horatius Flaccus, 65 – 8 a. d.

http://www.math.nus.edu.sg/~matsr/ProbI/Lecture12.pdf

Bibliography

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., 2006. ISBN 0387310738. URL https://www.springer.com/de/book/9780387310732. 37

[2] D. Blackwell. Comparison of experiments. 1951. 103

[3] C. Czado and T. Schmidt. *Mathematische Statistik*. Springer Berlin Heidelberg, 2011. doi:10.1007/978-3-642-17261-8. 3

[4] T. S. Ferguson. *Mathematical Statistics. A Decision Theoretic Approach.* Academic Press, New York, 1967. 107

[5] T. S. Ferguson. *A course in large sample theory*. Chapman & Hall, 1996. ISBN 9780412043710. 107

[6] H.-O. Georgii. *Stochastik*. de Gruyter, Berlin, 2002. doi:10.1515/9783110206777. 3

[7] G. Kersting and A. Wakolbinger. *Elementare Stochastik*. Springer Basel, 2010. doi:10.1007/978-3-0346-0414-7. 41

[8] L. Le Cam. Comparison of experiments: A short review. *Lecture Notes-Monograph Series*, 30: 127–138, 1996. URL http://www.jstor.org/stable/4355942. 103

[9] E. L. Lehmann. Some concepts of dependence. In *Selected Works of E. L. Lehmann*, chapter 64, pages 811–827. Springer, 2012. doi:10.1007/978-1-4614-1412-4. 12

[10] Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2006. URL http://books.google.com.au/books?id=Zsa7ofamTIUC. 142

[11] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes I*. Springer, 2nd edition, 2001. doi:10.1007/978-3-662-13043-8. 36

[12] W. K. Newey and J. L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55 (4):819–847, 1987. doi:10.2307/1911031. 15

[13] G. Ch. Pflug. Mathematische Statistik. Lecture notes, 2016. 110

[14] H. Pruscha. *Vorlesungen über Mathematische Statistik*. Springer, 2000. doi:10.1007/978-3-322-82966-5. 3

[15] H. Pruscha. *Statistisches Methodenbuch*. Springer, 2006. doi:10.1007/3-540-29305-1. 3

[16] J. S. Racine, Q. Li, and X. Zhu. Kernel estimation of multivariate conditional distributions. *Annals of Economics and Finance*, 5:211–235, 2004. 142

[17] L. Rüschendorf. *Mathematische Statistik*. Springer Berlin Heidelberg, 2014. doi:10.1007/978-3-642-41997-3. German. 3, 28, 110

[18] S. J. Sheather. An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics & Data Analysis*, 4(1):61–65, 1986. ISSN 0167-9473. doi:10.1016/0167-9473(86)90026-5. 142

[19] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2008. doi:10.1007/b13794. 135, 142

[20] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-821-83312-X. doi:10.1090/gsm/058. URL `http://books.google.com/books?id=GqRXYFxe0l0C`. 135

[21] P. Weiß. *Anwendungsorientierte Stochastik II*. 85, 86, 87

[22] H. Witting and U. Müller-Funk. *Mathematische Statistik III*. Vieweg+Teubner Verlag, 1995. doi:10.1007/978-3-322-90152-1. 3

Index