Mathematical Statistics

Lutz Dümbgen

University of Bern
Fall Semester 2025

December 18, 2025

Bibliography

[1] G. AILAM (1968). On probability properties of random sets and the asymptotic behavior of empirical distribution functions. *Journal of Applied Probability* ***5****(1)*, pp. 196-202.

[2] D. BLACKWELL (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics* ***18***, pp. 105-110.

[3] C. CZADO and T. SCHMIDT (2011). *Mathematische Statistik*. Springer.

[4] DONOHO, D.L. and JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* ***90****(432)*, pp. 1200-1225.

[5] L. DÜMBGEN (2021). *Optimization Methods*. Lecture notes, University of Bern.

[6] L. DÜMBGEN (2025). *Advanced Measure Theory*. Lecture notes, University of Bern.

[7] EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Assocociation* ***68****(341)*, pp. 117-130.

[8] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* ***1***, pp. 361-379.

[9] JOHNSTONE, I.M. (2017). *Gaussian estimation: Sequence and wavelet models*. Manuscript, Stanford University.

[10] P.R. HALMOS (1946). The theory of unbiased estimation. *Annals of Mathematical Statistics* ***17***, pp. 34-43.

[11] W. HOEFFING (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* ***19****(3)*, pp. 293-325.

[12] M.G. KENDALL (1938). A new measure of rank correlation. *Biometrika* ***30****(1-2)*, pp. 81-93.

[13] A.N. KOLMOGOROV (1950). Unbiased estimates. *Izvestiya Akademii Nauk SSSR. Seriya Matematicheskaya* ***14***, pp. 303-326.

[14] E.L. LEHMANN and J.P. ROMANO (2021). *Testing Statistical hhypotheses (4th edition)*. Springer, Cham.

[15] E.L. LEHMANN and H. SCHEFFÉ (1950). Completeness, similar regions, and unbiased estimation. I. *Sankhyā* ***10***, pp. 305-340.

[16] E.L. LEHMANN and H. SCHEFFÉ (1950). Completeness, similar regions, and unbiased estimation. II. *Sankhyā* ***15***, pp. 219-236.

[17] D.W. MÜLLER (1986). *Mathematische Statistik*. Lecture notes, University of Heidelberg.

[18] C.M. STEIN (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* ***1***, pp. 197-206.

[19] C.M. STEIN (1981). Estimating the mean of a multivariate normal distribution. *Annals of Statistics* ***9****(6)*, 1135-1151.

Acknowledgements. The major part of these lecture notes is an updated and translated version of my personal notes from a lecture "Mathematische Statistik" by my former PhD advisor Prof. Dietrich Werner Müller at the University of Heidelberg in the summer of 1986. His intriguing lectures stimulated my own interest in statistics as a mathematical as well as applied discipline. Constructive comments by Stefan Binder, Alessandro Corrent, Claire Descombes, Alexander Henzi, Xinwei Li, Jin M. Löffler, Christof Mahnig, Alexandre Mösching, Géraldine Oppliger, Levi Ryffel, Sara Salvador, Nadine Seitlinger, Philip Stange, Christof Strähl, Nikita van Gils and Raphael Waller on various versions of these lecture notes are gratefully acknowledged.

Contents

Chapter 0

Introduction

Statistics is the art of analysing data and dealing with non-avoidable errors and uncertainties in a concise way. In introductory and many advanced Statistics courses, various procedures such as point estimators, statistical tests and confidence regions are introduced for different settings, but often they seem a bit ad hoc. The purpose of Mathematical Statistics is to present these procedures in a coherent framework and to clarify which procedures are optimal for a given task. This includes the question of how to quantify the quality of a statistical procedure.

An indispensable tool for mathematical statistics is measure theory, including Radon-Nikodym derivaties, conditional expectations, conditional distributions and Markov kernels. These topics are covered in the lecture "Advanced Measure Theory".

From raw data to statistical experiments

Consider some experiment or study which yields raw data $\omega \in \Omega$. We assume that these raw data are random, so we imagine a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Preprocessing the raw data ω leads to data $X(\omega)$ in some measurable space $(\mathcal{X}, \mathcal{B})$, the *sample space*. We assume that the mapping X is measurable.

Typically the underlying space $(\Omega, \mathcal{A}, \mathbb{P})$ is not further specified, and the raw data ω are rarely mentioned. Instead we focus on the distribution of the random variable X and assume that it belongs to a given family $(P_\theta)_{\theta\in\Theta}$ of probability distributions on the sample space $(\mathcal{X}, \mathcal{B})$, indexed by *parameters* θ in a *parameter space* Θ. Thus we assume that there is an unknown true parameter $\theta \in \Theta$ such that X has distribution P_θ. In what follows, the symbol θ may denote this particular *true* parameter or a *potential* parameter. It should become clear from the context in which sense θ is meant. The dependency of probabilities, expectations, variances etc. on θ will sometimes be denoted by a corresponding subscript, leading to $\mathbb{P}_\theta(\cdot)$, $\mathbb{E}_\theta(\cdot)$, $\mathrm{Var}_\theta(\cdot)$ etc. Integrals of measurable functions on $(\mathcal{X}, \mathcal{B})$ with respect to P_θ are denoted with $E_\theta(\cdot)$.

The resulting model for our observed data X is the *statistical experiment*

$$\mathcal{E} \;=\; \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big).$$

At first glance there seems to be some redundancy in this definition of a statistical experiment $\mathcal{E}$, because specifying the family $(P_\theta)_{\theta\in\Theta}$ implies the specification of the measurable space $(\mathcal{X},\mathcal{B})$. But sometimes we shall replace $\mathcal{B}$ with certain sub-σ-fields, so the current definition is useful.

Often the data consist of a tuple of n independent, identically distributed random variables with values in some measurable space $(\mathcal{X},\mathcal{B})$. Then we sometimes write the data as $\boldsymbol{X} = (X_i)_{i=1}^n$ with values in the sample space $(\mathcal{X}^n,\mathcal{B}^{\otimes n})$, and the statistical experiment has the form $\mathcal{E} = \big(\mathcal{X}^n,\mathcal{B}^{\otimes n},(P_\theta^{\otimes n})_{\theta\in\Theta}\big)$ with a given family $(P_\theta)_{\theta\in\Theta}$ of distributions on $(\mathcal{X},\mathcal{B})$. The assumption that $\boldsymbol{X}\sim P_\theta^{\otimes n}$ is often phrased as "$\boldsymbol{X}$ is a sample from P_θ".

Example (Simple location family). Suppose that we observe independent random variables X_1, X_2, …, X_n, where

$$X_i \;=\; \theta+\epsilon_i.$$

Here $\theta\in\mathbb{R}$ is an unknown parameter, and $\epsilon_1,\epsilon_2,\ldots,\epsilon_n$ are independent (unobserved) random errors with given distribution P_0.

One could think about n repeated measurements with a measurement device (e.g. a scale), and P_0 describes the distribution of an unavoidable measurement error, which does not depend on the value of θ.

Thus the tuple $\boldsymbol{X}=(X_i)_{i=1}^n\in\mathbb{R}^n$ has distribution $P_\theta^{\otimes n}$ with

$$P_\theta(B) \;:=\; \mathbb{P}(\theta+\epsilon\in B),\ \epsilon\sim P_0.$$

If the error distribution P_0 has distribution function F_0, then the distribution function F_θ of P_θ is given by $F_\theta(x):=F_0(x-\theta)$, and if P_0 has a density f_0 with respect to Lebesgue measure on $\mathbb{R}$, then P_θ has density f_θ given by $f_\theta(x)=f_0(x-\theta)$. This leads to the statistical experiment

$$\mathcal{E} \;=\; \big(\mathbb{R}^n,\mathrm{Borel}(\mathbb{R})^{\otimes n},(P_\theta^{\otimes n})_{\theta\in\mathbb{R}}\big).$$

Example (Bernoulli experiment). Suppose we observe independent random variables X_1, X_2, …, X_n with values in $\{0,1\}$, where

$$\mathbb{P}(X_i=1) \;=\; \theta \;=\; 1-\mathbb{P}(X_i=0)$$

for an unknown parameter $\theta\in[0,1]$.

As a first specific example, one could think about n repeated tosses of a (not necessarily fair) coin.

Here the tuple $\boldsymbol{X}=(X_i)_{i=1}^n\in\{0,1\}^n$ has the following discrete distribution: For any tuple $\boldsymbol{x}\in\{0,1\}^n$,

$$\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}) \;=\; \prod_{i=1}^n \mathbb{P}(X_i=x_i) \;=\; \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \;=\; \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})}$$

with $T(\boldsymbol{x})=\sum_{i=1}^n x_i\in\{0,1,\ldots,n\}$. This leads to the statistical experiment

$$\mathcal{E} \;=\; \big(\{0,1\}^n,\mathcal{P}(\{0,1\}^n),(P_\theta)_{\theta\in[0,1]}\big),$$

where P_θ is given by

$$P_\theta(\{\boldsymbol{x}\}) \;=\; \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})} \quad \text{for } \boldsymbol{x} \in \{0,1\}^n.$$

Instead of reporting the n outcomes $X_1, X_2, \ldots, X_n$, one could summarize the experiment by the total number $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ with distribution $\mathrm{Bin}(n,\theta)$. This leads to the experiment

$$\mathcal{E} \;=\; \big(\{0,1,\ldots,n\}, \mathcal{P}(\{0,1,\ldots,n\}), (\mathrm{Bin}(n,\theta))_{\theta\in[0,1]}\big).$$

Here is another specific example leading to the same statistical experiment(s) with an explicit description of $(\Omega, \mathcal{A}, \mathbb{P})$: Imagine a population $\mathcal{M}$ and a subpopulation $\mathcal{M}_o$ of individuals with a special property. Suppose that the fraction $\theta = \#\mathcal{M}_o/\#\mathcal{M}$ is unknown. To find out something about θ, we draw a random sample $\omega = (\omega_1, \ldots, \omega_n)$ from this population and report

$$X_i(\omega) \;=\; \begin{cases} 1 & \text{if } \omega_i \in \mathcal{M}_o, \\ 0 & \text{else.} \end{cases}$$

If $\omega_1, \ldots, \omega_n$ are drawn completely at random with replacement, then ω belongs to the finite set $\Omega = \mathcal{M}^n$, $\mathbb{P}$ is the Laplace distribution on $(\Omega, \mathcal{P}(\Omega))$, and the random variables $X_1, \ldots, X_n$ have the properties described before.

If $\omega_1, \ldots, \omega_n$ are drawn completely at random without replacement, then ω belongs to the finite set Ω of all $\omega \in \mathcal{M}^n$ with pairwise different components. Again, $\mathbb{P}$ is the Laplace distribution on $(\Omega, \mathcal{P}(\Omega))$, and the random variables $X_1, \ldots, X_n$ still have the property that $\mathbb{P}(X_i = 1) = \theta = 1 - \mathbb{P}(X_i = 0)$. But now they are stochastically dependent, and $T(\boldsymbol{X})$ follows the hypergeometric distribution $\mathrm{Hyp}(\#\mathcal{M}, \#\mathcal{M}_o, n)$. If we focus on the potential distributions of $T(\boldsymbol{X})$, and the population size $\#\mathcal{M}$ is unknown too, this leads to the statistical experiment

$$\big(\{0,1,\ldots,n\}, \mathcal{P}(\{0,1,\ldots,n\}), (\mathrm{Hyp}(N,m,n))_{(N,m)\in\Theta}\big),$$

where Θ consists of all pairs (N,m) of integers such that $N \geq n$ and $0 \leq m \leq N$.

If $n \ll \#\mathcal{M}$, however, the difference between sampling with replacement and without replacement becomes negligible, and the simpler Bernoulli or binomial experiments are appropriate.

Chapter 1

Measurement Series and Estimators of Location

1.1 Point Estimators

We start from a statistical experiment $\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$. Sometimes one is interested in a function $g(\theta)$ of the true parameter θ with values in some metric space $(\mathbb{G}, d)$, where $g : \Theta \to \mathbb{G}$ is given. In the simplest case, one would like to deduce from the observed data X a simple guess $\widehat{g}(X) \in \mathbb{G}$ of $g(\theta)$.

Definition 1.1 (Point estimator). A *point estimator of* $g(\theta)$ (short: an *estimator*) is a measurable[1] function

$$\widehat{g} : \mathcal{X} \to \mathbb{G}.$$

To compare different estimators $\widehat{g}$, one can quantify their inaccuracy for instance by their mean squared error,

$$E_\theta\big(d(\widehat{g}, g(\theta))^2\big) \;=\; \int_{\mathcal{X}} d(\widehat{g}(x), g(\theta))^2 \, P_\theta(\mathrm{d}x).$$

Note that this quantity depends on the parameter θ. In general, it may happen that one estimator $\widehat{g}_1$ is strictly better than another estimators $\widehat{g}_2$ in a certain region of the parameter space Θ but strictly worse somewhere else.

1.2 Estimators of Location

We consider a simple location family as in the introduction, i.e.

$$\mathcal{E} \;=\; \big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R})^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\mathbb{R}}\big)$$

with

$$P_\theta \;=\; P_0 \star \delta_\theta.$$

[1] $\mathcal{B}$-Borel$(\mathbb{G}, d)$-measurable

Here '$\star$' denotes convolution[2], and δ_θ is the Dirac measure at the point θ.

To ease notation, we identify $(\Omega, \mathcal{A})$ with $(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R})^{\otimes n})$ and set $X_i(\boldsymbol{x}) := x_i$, so $\boldsymbol{X}$ is just the identity mapping. Moreover, we write $\mathbb{P}_\theta := P_\theta^{\otimes n}$.

Equivariant estimators. For a vector $\boldsymbol{x} = (x_i)_{i=1}^n \in \mathbb{R}^n$ and any number $a \in \mathbb{R}$ let

$$a + \boldsymbol{x} \;:=\; \boldsymbol{x} + a \;:=\; (x_i + a)_{i=1}^n$$

The simple location family $\mathcal{E}$ has the property that for arbitrary $\theta, a \in \mathbb{R}$,

$$\boldsymbol{X} \;\sim\; \mathbb{P}_\theta \quad \text{if and only if} \quad \boldsymbol{X} + a \;\sim\; \mathbb{P}_{\theta+a}\,.$$

This motivates the following property of an estimator $\widehat{\theta}$ of θ:

Definition 1.2 (Equivariance). An estimator $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$ is called *equivariant*, if

$$\widehat{\theta}(\boldsymbol{x} + a) \;=\; \widehat{\theta}(\boldsymbol{x}) + a \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \text{ and } a \in \mathbb{R}.$$

Note that the sample mean,

$$\bar{X} \;:=\; \frac{1}{n}\sum_{i=1}^n X_i,$$

as well as the sample median are equivariant estimators. Concerning the sample mean, it follows from the weak law of large numbers that for a fixed distribution P_0 with mean $\int x\, P_0(\mathrm{d}x) = 0$,

$$\lim_{n\to\infty} \mathbb{E}_\theta\, |\bar{X} - \theta| \;=\; 0.$$

If in addition, P_0 has finite variance σ^2, then even

$$\mathbb{E}_\theta\big((\bar{X} - \theta)^2\big) \;=\; \frac{\sigma^2}{n}.$$

Risk functions. For an arbitrary estimator $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$, we consider its risk function $R(\widehat{\theta}, \cdot) : \mathbb{R} \to [0, \infty]$ with

$$R(\widehat{\theta}, \theta) \;:=\; \mathbb{E}_\theta\big((\widehat{\theta} - \theta)^2\big) \;=\; \int_{\mathbb{R}^n} (\widehat{\theta} - \theta)^2 \,\mathrm{d}\,\mathbb{P}_\theta,$$

the mean squared error of $\widehat{\theta}$ in case of the true parameter being θ.

In case of an equivariant estimator $\widehat{\theta}$, its risk function is constant: For arbitrary $\theta \in \mathbb{R}$,

$$R(\widehat{\theta}, \theta) \;=\; R(\widehat{\theta}) := R(\widehat{\theta}, 0) \;=\; \mathbb{E}_0(\widehat{\theta}^2).$$

More generally, if $\widehat{\theta}$ is equivariant, then for any measurable function $h : \mathbb{R} \to \mathbb{R}$ and arbitrary $\theta \in \mathbb{R}$,

$$\mathbb{E}_\theta\, h(\widehat{\theta} - \theta) \;=\; \mathbb{E}_0\, h(\widehat{\theta}),$$

provided that $\mathbb{E}_0\, h(\widehat{\theta})$ is well-defined.

[2]For distributions P and Q on the real line, $P \star Q$ denotes the distribution of $X + Y$ with independent random variables $X \sim P$ and $Y \sim Q$. In particular, for $\theta \in \mathbb{R}$, $P \star \delta_\theta$ is the distribution of $X + \theta$ with $X \sim P$.

1.3 Constructing an Optimal Equivariant Estimator

An equivariant estimator $\widehat{\theta}_*$ is called *optimal (among all equivariant estimators)* if

$$R(\widehat{\theta}_*) \;\leq\; R(\widehat{\theta}) \quad \text{for any equivariant estimator } \widehat{\theta}.$$

If $\boldsymbol{X} \sim \mathbb{P}_\theta$, then $\boldsymbol{X} = \theta + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathbb{P}_0$, and for any equivariant estimator $\widehat{\theta}$,

$$\widehat{\theta}(\boldsymbol{X}) \;=\; \theta + \widehat{\theta}(\boldsymbol{\epsilon}).$$

Of course, we don't know $\boldsymbol{\epsilon}$, but at least we know $\boldsymbol{T}(\boldsymbol{\epsilon})$ for the particular function $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\boldsymbol{T}(\boldsymbol{x}) \;:=\; \boldsymbol{x} - x_1 \;=\; (0,\, x_2 - x_1, \ldots, x_n - x_1)^\top.$$

Indeed, this function $\boldsymbol{T}$ is *invariant* in the sense that

$$\boldsymbol{T}(\boldsymbol{x} + a) \;=\; \boldsymbol{T}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \text{ and } a \in \mathbb{R}.$$

Hence, if we observe $\boldsymbol{X} = \theta + \boldsymbol{\epsilon}$, then

$$\boldsymbol{T}(\boldsymbol{X}) \;=\; \boldsymbol{T}(\boldsymbol{\epsilon}).$$

That means, we know at least $\boldsymbol{T}(\boldsymbol{\epsilon})$. So we could try to improve the estimator $\widehat{\theta}(\boldsymbol{X}) = \theta + \widehat{\theta}(\boldsymbol{\epsilon})$ by subtracting the conditional expectation of $\widehat{\theta}(\boldsymbol{\epsilon})$, given that $\boldsymbol{T}(\boldsymbol{\epsilon})$ is equal to the observed $\boldsymbol{T}(\boldsymbol{X})$. That means, we subtract a reasonable guess of $\widehat{\theta}(\boldsymbol{\epsilon})$ from $\widehat{\theta}(\boldsymbol{X})$. This idea leads to an optimal equivariant estimator indeed.

Theorem 1.3 (Pitman's improvement). *Let $\widehat{\theta}$ be an equivariant estimator with finite risk $R(\widehat{\theta})$. Then*

$$\widehat{\theta}_* \;:=\; \widehat{\theta} - \mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T}))$$

defines an optimal equivariant estimator. It is unique in the sense that for any equivariant estimator $\tilde{\theta}$ and arbitrary $\theta \in \mathbb{R}$,

$$R(\tilde{\theta}) \;=\; R(\widehat{\theta}_*) \quad \textit{implies that} \quad \tilde{\theta} \;=\; \widehat{\theta}_* \;\; \mathbb{P}_\theta\textit{-almost surely}.$$

Remark 1.4 (Invariance and the choice of $\boldsymbol{T}(\cdot)$). Our particular choice of $\boldsymbol{T}(\cdot)$ is somewhat arbitrary. In principle one could take any equivariant estimator $\tilde{\theta} : \mathbb{R}^n \to \mathbb{R}$ and define $\boldsymbol{T}(\boldsymbol{x}) := \boldsymbol{x} - \tilde{\theta}(\boldsymbol{x})$. Inspecting the proof of Theorem 1.3 carefully reveals that Theorem 1.3 remains valid with this definition of $\boldsymbol{T}$. Our particular version corresponds to $\tilde{\theta} = X_1$ and is convenient for explicit calculations.

Exercise 1.5. Consider an estimator $\tilde{\theta} : \mathbb{R}^n \to \mathbb{R}$. Show that $\tilde{\theta}$ is equivariant if and only if $\boldsymbol{T}(\boldsymbol{x}) := \boldsymbol{x} - \tilde{\theta}(\boldsymbol{x})$ is invariant, i.e. $\boldsymbol{T}(\boldsymbol{x} + a) = \boldsymbol{T}(\boldsymbol{x})$ for arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}$.

Remark 1.6 (Characterization of optimality). The particular construction in Theorem 1.3 implies that an equivariant estimator $\widehat{\theta}$ with finite risk $R(\widehat{\theta})$ is optimal if and only if

$$\mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) \;=\; 0 \quad \mathbb{P}_0\text{-almost surely.}$$

Exercise 1.7. Let $\widehat{\theta}_* : \mathbb{R}^n \to \mathbb{R}$ be an optimal equivariant estimator for θ with finite risk $R(\widehat{\theta}_*)$. Show that $\widehat{\theta}_*$ is unbiased, that means,

$$\mathbb{E}_\theta(\widehat{\theta}_*) \;=\; \theta \quad \text{for all } \theta \in \mathbb{R}.$$

Exercise 1.8. Suppose that P_0 is the Laplace distribution on $\{0, 1\}$.

(a) Before starting to apply the general theory, how would you estimate θ?

(b) Determine the conditional distribution of $\boldsymbol{X}$, given that $\boldsymbol{T} = \boldsymbol{y}$, in case of $\theta = 0$. (Which vectors $\boldsymbol{y} \in \mathbb{R}^n$ are relevant?)

(c) Determine the optimal (in terms of mean squared error) equivariant estimator of θ.

Proof of Theorem 1.3. The general theory of conditional expectations shows that we may write $\mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) = g_*(\boldsymbol{T})$ with a measurable function $g_* : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathbb{E}_0\big((\widehat{\theta} - g(\boldsymbol{T}))^2\big) \;=\; \mathbb{E}_0\big((\widehat{\theta} - g_*(\boldsymbol{T}))^2\big) + \mathbb{E}_0\big((g(\boldsymbol{T}) - g_*(\boldsymbol{T}))^2\big)$$

for any measurable function $g : \mathbb{R}^n \to \mathbb{R}$. In particular, for $g \equiv 0$ we obtain the formula

$$\mathbb{E}_0(\widehat{\theta}^2) \;=\; \mathbb{E}_0\big((\widehat{\theta} - g_*(\boldsymbol{T}))^2\big) + \mathbb{E}_0(g_*(\boldsymbol{T})^2). \tag{1.1}$$

Since $\widehat{\theta}$ is equivariant and $\boldsymbol{T}$ is invariant, $\widehat{\theta}_* = \widehat{\theta} - g_*(\boldsymbol{T})$ is an equivariant estimator, too: For arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}$,

$$\widehat{\theta}_*(\boldsymbol{x} + a) \;=\; \underbrace{\widehat{\theta}(\boldsymbol{x} + a)}_{=\widehat{\theta}(\boldsymbol{x})+a} - \underbrace{g_*(\boldsymbol{T}(\boldsymbol{x} + a))}_{=g_*(\boldsymbol{T}(\boldsymbol{x}))} \;=\; \widehat{\theta}(\boldsymbol{x}) + a - g_*(\boldsymbol{T}(\boldsymbol{x})) \;=\; \widehat{\theta}_*(\boldsymbol{x}) + a.$$

Hence, we may rewrite (1.1) as

$$R(\widehat{\theta}) \;=\; R(\widehat{\theta}_*) + \mathbb{E}_0(g_*(\boldsymbol{T})^2).$$

Consequently, $R(\widehat{\theta}) \geq R(\widehat{\theta}_*)$ with equality if and only if $\mathbb{E}_0(g_*(\boldsymbol{T})^2) = 0$, and this is equivalent to

$$\mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) \;=\; 0 \quad \mathbb{P}_0\text{-almost surely.}$$

Finally, let $\tilde{\theta} : \mathbb{R}^n \to \mathbb{R}$ be another equivariant estimator with finite risk $R(\tilde{\theta})$. Then $h := \tilde{\theta} - \widehat{\theta}$ is invariant in the sense that $h(\boldsymbol{x}) = h(\boldsymbol{T}(\boldsymbol{x}))$ for arbitrary $\boldsymbol{x} \in \mathbb{R}^n$. Consequently, if we apply Pitman's recipe to $\tilde{\theta}$ instead of $\widehat{\theta}$, we obtain the estimator

$$\begin{aligned}
\tilde{\theta}_* \;&:=\; \tilde{\theta} - \mathbb{E}_0(\tilde{\theta} \,|\, \sigma(\boldsymbol{T})) \\
&=\; \widehat{\theta} + h(\boldsymbol{T}) - \mathbb{E}_0\big(\widehat{\theta} + h(\boldsymbol{T}) \,\big|\, \sigma(\boldsymbol{T})\big) \\
&=\; \widehat{\theta} + h(\boldsymbol{T}) - \mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) - \underbrace{\mathbb{E}_0(h(\boldsymbol{T}) \,|\, \sigma(\boldsymbol{T}))}_{=h(\boldsymbol{T}) \text{ a.s.}} \\
&=\; \widehat{\theta} - \mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) \;=\; \widehat{\theta}_* \quad \mathbb{P}_0\text{-almost surely.}
\end{aligned}$$

Moreover, since $\tilde{\theta}_* - \widehat{\theta}_*$ is invariant, $\mathbb{P}_\theta(\tilde{\theta}_* \neq \widehat{\theta}_*) = \mathbb{P}_0(\tilde{\theta}_* \neq \widehat{\theta}_*) = 0$ for arbitrary $\theta \in \mathbb{R}$. $\square$

In case of P_0 having a density with respect to Lebesgue measure, there is an explicit formula for the optimal equivariant estimator $\widehat{\theta}_*$:

Corollary 1.9. *Suppose that P_0 has a density f_0 with respect to Lebesgue measure on $\mathbb{R}$, and suppose that there exists an equivariant estimator with finite risk. Then there exists a Borel set $B_* \subset \mathbb{R}^n$ such that $\mathbb{P}_0(\boldsymbol{T}^{-1}(B_*)) = 1$,[3] and for each $\boldsymbol{x} \in \boldsymbol{T}^{-1}(B_*)$, the optimal equivariant estimator $\widehat{\theta}_*$ is given by*

$$\widehat{\theta}_*(\boldsymbol{x}) \;=\; \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x})\,\mathrm{d}\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\,\mathrm{d}\theta,$$

where $f_\theta(\boldsymbol{x}) := \prod_{i=1}^n f_\theta(x_i)$ and $f_\theta(x) := f_0(x-\theta)$ for real numbers x. In other words, $\widehat{\theta}_*(\boldsymbol{x})$ is the mean of the probability distribution $Q_{\boldsymbol{x}}$ on $\mathbb{R}$ with density

$$\theta \;\mapsto\; f_\theta(\boldsymbol{x}) \Big/ \int_{\mathbb{R}} f_\eta(\boldsymbol{x})\,\mathrm{d}\eta.$$

Example 1.10 (Gaussian distributions). Suppose that $P_0 = \mathrm{N}(0,\sigma^2)$ for some $\sigma > 0$. Then

$$\widehat{\theta}_* \;=\; \bar{X}.$$

This follows from Corollary 1.9 and the following calculations: $f_0(x) = C\exp\bigl(-x^2/(2\sigma^2)\bigr)$ with $C = (2\pi\sigma^2)^{-1/2}$, whence

$$f_\theta(\boldsymbol{x}) \;=\; C^n \exp\Bigl(-\frac{\|\boldsymbol{x}-\theta\|^2}{2\sigma^2}\Bigr).$$

But

$$\|\boldsymbol{x}-\theta\|^2 \;=\; \|\boldsymbol{x}-\bar{x}\|^2 + n(\theta-\bar{x})^2$$

with $\bar{x} := n^{-1}\sum_{i=1}^n x_i$, so

$$f_\theta(\boldsymbol{x}) \;=\; f_{\bar{x}}(\boldsymbol{x}) \exp\Bigl(-\frac{(\theta-\bar{x})^2}{2\sigma^2/n}\Bigr),$$

and this implies that the distribution $Q_{\boldsymbol{x}}$ in Corollary 1.9 is equal to $\mathrm{N}(\bar{x},\sigma^2/n)$. Hence

$$\widehat{\theta}_*(\boldsymbol{x}) \;=\; \mathrm{mean}\bigl(\mathrm{N}(\bar{x},\sigma^2/n)\bigr) \;=\; \bar{x}.$$

Example 1.11 (Uniform distributions). Suppose that $P_0 = \mathrm{Unif}([-\sigma,\sigma])$ for some $\sigma > 0$. Then

$$\widehat{\theta}_*(\boldsymbol{x}) \;=\; \bigl(\min(\boldsymbol{x}) + \max(\boldsymbol{x})\bigr)/2$$

with $\min(\boldsymbol{x})$ and $\max(\boldsymbol{x})$ denoting the minimum and maximum of $\{x_1,\ldots,x_n\}$, respectively. This follows from the following considerations: Since $f_0(x) = (2\sigma)^{-1}1_{[-\sigma\le x\le\sigma]}$,

$$\begin{aligned} f_\theta(\boldsymbol{x}) \;&=\; (2\sigma)^{-n}\prod_{i=1}^n 1_{[-\sigma\le x_i-\theta\le\sigma]} \\ &=\; (2\sigma)^{-n}\prod_{i=1}^n 1_{[x_i-\sigma\le\theta\le x_i+\sigma]} \\ &=\; (2\sigma)^{-n} 1_{[\max(\boldsymbol{x})-\sigma\le\theta\le\min(\boldsymbol{x})+\sigma]}. \end{aligned}$$

[3] $\mathbb{P}_\theta(\boldsymbol{T}^{-1}(B_*)) = 1$ for all $\theta\in\mathbb{R}$ by invariance of $\boldsymbol{T}$.

Hence the distribution $Q_{\boldsymbol{x}}$ in Corollary 1.9 is the uniform distribution on the interval with endpoints $\max(\boldsymbol{x}) - \sigma$ and $\min(\boldsymbol{x}) + \sigma$, unless $\max(\boldsymbol{x}) - \min(\boldsymbol{x}) > 2\sigma$. (Note that $\max(\boldsymbol{X}) - \min(\boldsymbol{X}) < 2\sigma$ almost surely.) Consequently,

$$\widehat{\theta}_*(\boldsymbol{x}) \;=\; \text{midpoint of } [\max(\boldsymbol{x}) - \sigma, \min(\boldsymbol{x}) + \sigma] \;=\; \big(\min(\boldsymbol{x}) + \max(\boldsymbol{x})\big)/2.$$

(This definition makes sense no matter how large the difference $\max(\boldsymbol{x}) - \min(\boldsymbol{x})$ is.)

Exercise 1.12. Suppose that $P_0 = \mathrm{Unif}[-\sigma, \sigma]$. As shown before, the optimal equivariant estimator of θ is given by $\widehat{\theta}_*(\boldsymbol{x}) = \big(\min(\boldsymbol{x}) + \max(\boldsymbol{x})\big)/2$.

(a) Determine the risk of $\bar{X}$.

(b) Show that the risk of $\widehat{\theta}_*$ is of order $O(n^{-2})$.

Bonus question: Show that

$$R(\widehat{\theta}_*) \;=\; \frac{2\sigma^2}{(n+1)(n+2)}.$$

Remark 1.13 (Maximum-likelihood estimation). In case of P_0 having density f_0, the function

$$\theta \;\mapsto\; f_\theta(\boldsymbol{X}) = \prod_{i=1}^{n} f_\theta(X_i)$$

is the so-called likelihood function. For any number θ one may interpret $f_\theta(\boldsymbol{X})$ as a measure of plausibility of θ being equal to the true parameter. Indeed a standard estimator of the true parameter θ would be the maximum-likelihood estimator

$$\widehat{\theta}_{\mathrm{ML}} \;:=\; \operatorname*{arg\,max}_{\theta \in \mathbb{R}} \; f_\theta(\boldsymbol{X}),$$

provided the latter is uniquely defined. Our previous calculations show that $\widehat{\theta}_* = \widehat{\theta}_{\mathrm{ML}} = \bar{X}$ in case of P_0 being a centered Gaussian distribution.

The higher popularity of $\widehat{\theta}_{\mathrm{ML}}$ in comparison with $\widehat{\theta}_*$ is due to the fact that the latter estimator is rather difficult to compute explicitly in non-Gaussian models. Moreover, in many settings one can show that

$$f_\theta(\boldsymbol{X}) \;\approx\; f_{\widehat{\theta}_{\mathrm{ML}}}(\boldsymbol{X}) \exp\Big(-\frac{(\theta - \widehat{\theta}_{\mathrm{ML}})^2}{2\widehat{\gamma}^2}\Big)$$

for some random variable $\widehat{\gamma} > 0$ such that $\widehat{\gamma} \to_p 0$ as $n \to \infty$. Hence $Q_{\boldsymbol{X}} \approx \mathrm{N}(\widehat{\theta}_{\mathrm{ML}}, \widehat{\gamma}^2)$ for large sample sizes n, and $\widehat{\theta}_{\mathrm{ML}}$ seems to be a good surrogate for $\widehat{\theta}_*$.

Exercise 1.14. Suppose that the error distribution P_0 is the standard exponential distribution. That means, its density is given by

$$f_0(x) \;=\; \begin{cases} 0 & \text{if } x < 0, \\ \exp(-x) & \text{if } x \geq 0. \end{cases}$$

(a) Determine $f_\theta(\boldsymbol{x})$ for $\theta \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$ in terms of $\min(\boldsymbol{x})$ and $x_+ := \sum_{i=1}^n x_i$.

(b) Determine the maximum likelihood estimator $\widehat{\theta}_{\mathrm{ML}}$.

(c) Determine the optimal equivariant estimator $\widehat{\theta}_*$.

Proof of Corollary 1.9. Consider the linear transformation given by

$$\boldsymbol{x} \;\mapsto\; (x_1,\, x_2 - x_1, \ldots, x_n - x_1)^\top \;=\; \boldsymbol{A}\boldsymbol{x}$$

with the lower triangular matrix

$$\boldsymbol{A} \;=\; \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

Its inverse is given by

$$\boldsymbol{y} \;\mapsto\; (y_1,\, y_2 + y_1, \ldots, y_n + y_1)^\top \;=\; \boldsymbol{A}^{-1}\boldsymbol{y},$$

and $\det(\boldsymbol{A}) = 1 = \det(\boldsymbol{A}^{-1})$. Hence, by the transformation formula for Lebesgue integrals,

$$\begin{aligned} \mathbb{P}_0(\boldsymbol{A}\boldsymbol{X} \in B) \;&=\; \int_{\mathbb{R}^n} 1_{[\boldsymbol{A}\boldsymbol{x} \in B]} f_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &=\; \int_{\mathbb{R}^n} 1_{[\boldsymbol{y} \in B]} f_0(\boldsymbol{A}^{-1}\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\ &=\; \int_B f_0(\boldsymbol{A}^{-1}\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \end{aligned}$$

for any Borel set $B \subset \mathbb{R}^n$. This shows that the distribution of $\boldsymbol{A}\boldsymbol{X}$ under $\mathbb{P}_0$ has density[4]

$$\boldsymbol{y} \;\mapsto\; f_0(\boldsymbol{A}^{-1}\boldsymbol{y}) \;=\; f_0(y_1, y_2 + y_1, \ldots, y_n + y_1).$$

Note that $\boldsymbol{A}\boldsymbol{X} = (X_1, T_2, \ldots, T_n)^\top$ while $T_1 \equiv 0$. Thus the distribution of $(T_2, \ldots, T_n)$ is given by the density

$$(t_2, \ldots, t_n) \;\mapsto\; g(t_2, \ldots, t_n) := \int_{\mathbb{R}} f_0(u, t_2 + u, \ldots, t_n + u)\, du.$$

In particular, there exists a Borel set $B_o \subset \mathbb{R}^n$ with $\mathbb{P}_0(\boldsymbol{T} \in B_o) = 1$ such that for any $\boldsymbol{t} \in B_o$,

$$0 \;<\; g(t_2, \ldots, t_n) := \int_{\mathbb{R}} f_0(u, t_2 + u, \ldots, t_n + u)\, du \;<\; \infty,$$

and the conditional distribution of X_1, given that $\boldsymbol{T} = \boldsymbol{t}$, has density

$$u \;\mapsto\; g(t_2, \ldots, t_n)^{-1} f_0(u, t_2 + u, \ldots, t_n + u).$$

Coming back to our estimator $\widehat{\theta}$ with finite risk $\mathbb{E}_0(\widehat{\theta}^2)$, note that $\widehat{\theta}(\boldsymbol{x}) = x_1 + h(\boldsymbol{T}(\boldsymbol{x}))$ for some measurable function h on $\mathbb{R}^n$, and

$$\infty \;>\; \mathbb{E}_0\, |\widehat{\theta}| \;=\; \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} |u + h(0, t_2, \ldots, t_n)| f_0(u, t_2 + u, \ldots, t_n + u)\, du\, d(t_2, \ldots, t_n).$$

[4] In this proof, densities are with respect to the corresponding Lebesgue measure.

Hence for a Borel set $B_* \subset B_o$ with $\mathbb{P}_0(\boldsymbol{T} \in B_*) = 1$ and arbitrary $\boldsymbol{t} \in B_*$,

$$\int_{\mathbb{R}} |u| f_0(u, t_2 + u, \ldots, t_n + u)\, du \;<\; \infty.$$

In particular, if we plug in $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{x})$ for some $\boldsymbol{x} \in \boldsymbol{T}^{-1}(B_*)$, then

$$f_0(u, t_2 + u, \ldots, t_n + u) \;=\; f_0(x_1 - x_1 + u, x_2 - x_1 + u, \ldots, x_n - x_1 + u) \;=\; f_{x_1 - u}(\boldsymbol{x}),$$

so

$$g(t_2, \ldots, t_n) \;=\; \int_{\mathbb{R}} f_{x_1-u}(\boldsymbol{x})\, du \;=\; \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta,$$

and

$$\begin{aligned}
\mathbb{E}_0\big(\widehat{\theta} \,\big|\, \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{x})\big) \;&=\; \int_{\mathbb{R}} (u + h(\boldsymbol{T}(\boldsymbol{x}))) f_{x_1-u}(\boldsymbol{x})\, du \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta \\
&=\; x_1 + h(\boldsymbol{T}(\boldsymbol{x})) - \int_{\mathbb{R}} (x_1 - u) f_{x_1-u}(\boldsymbol{x})\, du \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta \\
&=\; \widehat{\theta}(\boldsymbol{x}) - \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x})\, d\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta.
\end{aligned}$$

Consequently,

$$\widehat{\theta}_*(\boldsymbol{x}) \;=\; \widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_0\big(\widehat{\theta} \,\big|\, \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{x})\big) \;=\; \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x})\, d\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta$$

for all $\boldsymbol{x} \in \boldsymbol{T}^{-1}(B_*)$, as claimed. □

We end this section with the interesting result that in case of an error distribution P_0 with finite second moment and Lebesgue density f_0, the optimal equivariant estimator is the sample mean if and only if P_0 is a centered Gaussian distribution.

Theorem 1.15 (Kagan–Linnik–Rao). *Let $n \geq 3$, and let P_0 have finite second moment. Then $\widehat{\theta}_* = \bar{X}$ almost surely if and only if $P_0 = \mathrm{N}(0, \sigma^2)$ for some $\sigma \geq 0$.*

Proof of Theorem 1.15. We have verified already that $\widehat{\theta}_* = \bar{X}$ in case of P_0 being a centered Gaussian distribution. Hence it suffices to prove the reverse statement.

In what follows all probabilities and expectations refer to the distribution $\mathbb{P} = \mathbb{P}_0$. The proof of Theorem 1.3 shows that optimality of $\bar{X}$ is equivalent to

$$\mathbb{E}(\bar{X} \,|\, \sigma(\boldsymbol{T})) \;=\; 0$$

almost surely, where $\boldsymbol{T} = \boldsymbol{X} - X_1 = (0, X_2 - X_1, \ldots, X_n - X_1)^\top$. In other words,

$$\mathbb{E}\Big(\sum_{i=1}^n X_i g(\boldsymbol{T})\Big) \;=\; 0 \quad \text{whenever } g(\boldsymbol{T}) \in L^2(\mathbb{P}). \tag{1.2}$$

With $g \equiv 1$ this implies that $X := X_1$ has mean

$$\mathbb{E}(X) \;=\; \int x f_0(x)\, dx \;=\; 0.$$

If $n > 3$, we may take $g(\boldsymbol{T}) = h(T_2, T_3) = h(X_2 - X_1, X_3 - X_1)$ and deduce from independence of $X_1, X_2, \ldots, X_n$ and $\mathbb{E}(X_i) = 0$ that

$$\mathbb{E}\big((X_1 + X_2 + X_3)h(T_2, T_3)\big) \;=\; 0 \quad \text{whenever } h(T_2, T_3) \in L^2(\mathbb{P}). \tag{1.3}$$

Now let ϕ be the characteristic function of P_0, i.e.

$$\phi(t) \;=\; \mathbb{E}(e^{\boldsymbol{i}tX}) \;=\; \int e^{\boldsymbol{i}tx}\, P_0(dx)$$

with the imaginary unit $\boldsymbol{i} \in \mathbb{C}$. We know that $\phi : \mathbb{R} \to \mathbb{C}$ is bounded and continuous with $\phi(0) = 1$. Moreover, since P_0 has finite second moment, ϕ is twice continuously differentiable with derivative $\phi^{(k)}(t) = \boldsymbol{i}^k\, \mathbb{E}(X^k e^{\boldsymbol{i}tX})$ for $k = 1, 2$, so

$$\phi'(0) \;=\; \boldsymbol{i}\, \mathbb{E}(X) \;=\; 0 \quad \text{and} \quad \phi''(0) \;=\; -\,\mathbb{E}(X^2).$$

We may apply (1.3) to the (real and imaginary part of the) complex-valued and bounded function $h(z_1, z_2) = e^{\boldsymbol{i}sz_1 + \boldsymbol{i}tz_2}$ with arbitrary real numbers s, t. This leads to

$$\begin{aligned}
0 \;&=\; \mathbb{E}\big((X_1 + X_2 + X_3)e^{\boldsymbol{i}s(X_2 - X_1) + \boldsymbol{i}t(X_3 - X_1)}\big) \\
&=\; \mathbb{E}\big(X_1 e^{-\boldsymbol{i}(s+t)X_1} e^{\boldsymbol{i}sX_2} e^{\boldsymbol{i}tX_3}\big) \\
&\quad +\; \mathbb{E}\big(e^{-\boldsymbol{i}(s+t)X_1} X_2 e^{\boldsymbol{i}sX_2} e^{\boldsymbol{i}tX_3}\big) + \mathbb{E}\big(e^{-\boldsymbol{i}(s+t)X_1} e^{\boldsymbol{i}sX_2} X_3 e^{\boldsymbol{i}tX_3}\big) \\
&=\; \mathbb{E}(X_1 e^{-\boldsymbol{i}(s+t)X_1})\, \mathbb{E}(e^{\boldsymbol{i}sX_2})\, \mathbb{E}(e^{\boldsymbol{i}tX_3}) \\
&\quad +\; \mathbb{E}(e^{-\boldsymbol{i}(s+t)X_1})\, \mathbb{E}(X_2 e^{\boldsymbol{i}sX_2})\, \mathbb{E}(e^{\boldsymbol{i}tX_3}) + \mathbb{E}(e^{-\boldsymbol{i}(s+t)X_1})\, \mathbb{E}(e^{\boldsymbol{i}sX_2})\, \mathbb{E}(X_3 e^{\boldsymbol{i}tX_3}) \\
&= \phi'(-(s+t))\phi(s)\phi(t) + \phi(-(s+t))\phi'(s)\phi(t) + \phi(-(s+t))\phi(s)\phi'(t),
\end{aligned}$$

where the latter two equalities follow from X_1, X_2, X_3 being independent and identically distributed. Consequently,

$$\phi'(-(s+t))\phi(s)\phi(t) + \phi(-(s+t))\big(\phi'(s)\phi(t) + \phi(s)\phi'(t)\big) \;=\; 0 \quad \text{for arbitrary } s, t \in \mathbb{R}.$$

Now let

$$c \;:=\; \max\big\{t \in (0, \infty] : \phi \neq 0 \text{ on } (-t, t)\big\}.$$

Then

$$\psi(t) \;:=\; \frac{\phi'(t)}{\phi(t)}$$

defines a continuous function $\psi : (-c, c) \to \mathbb{C}$ with $\psi(0) = 0$ and

$$\psi(-(s+t)) + \psi(s) + \psi(t) \;=\; 0 \quad \text{whenever } |s|, |t|, |s+t| < c.$$

But this implies that for some $\alpha \in \mathbb{C}$,

$$\psi(t) \;=\; \alpha t \quad \text{for all } t \in (-c, c),$$

see Exercise 1.16. In other words,

$$\phi'(t) \;=\; \alpha t \phi(t) \quad \text{for } t \in (-c, c).$$

Together with $\phi(0) = 1$, standard results for differential equations imply that

$$\phi(t) \;=\; e^{\alpha t^2/2} \quad \text{for } t \in (-c, c).$$

But continuity of ϕ and the definition of c imply that $c = \infty$. For otherwise, continuity of ϕ and the definition of c would imply that $0 = \phi(\pm c) = e^{\alpha c^2/2}$. Consequently,

$$\phi(t) \;=\; e^{\alpha t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

Since $\phi''(t) = (\alpha + \alpha^2 t^2)\phi(t)$, we may conclude from $\phi''(0) = \alpha = -\mathbb{E}(X^2)$ that α is a negative real number. It is well-known from probability theory that for any $\sigma \geq 0$, the characteristic function of $\mathrm{N}(\mu, \sigma^2)$ is given by $t \mapsto \exp(\dot{\imath} t\mu - \sigma^2 t^2/2)$. Hence the characteristic function of P_0 coincides with the characteristic function of $\mathrm{N}(0, \sigma^2)$, where $\sigma := \sqrt{-\alpha}$. Since any probability distribution is uniquely determined by its characteristic distribution, this shows that $P_0 = \mathrm{N}(0, \sigma^2)$. □

Exercise 1.16. For some $c \in (0, \infty]$, let $\psi : (-c, c) \to \mathbb{C}$ be a continuous function such that

$$\psi(-(s+t)) + \psi(s) + \psi(t) \;=\; 0 \quad \text{whenever } |s|, |t|, |s+t| < c.$$

Show that there exists a constant $\alpha \in \mathbb{C}$ such that

$$\psi(t) \;=\; \alpha t \quad \text{for all } t \in (-c, c).$$

The contents of the next two exercises are probably known from other courses in Statistics.

Exercise 1.17 (Binomial distribution functions). For $n \in \mathbb{N}$ and $p \in [0, 1]$, let $F_{n,p}$ be the distribution function of $\mathrm{Bin}(n, p)$, that is,

$$F_{n,p}(c) \;=\; \sum_{j=0}^{c} \binom{n}{j} p^j (1-p)^{n-j} \quad \text{for } c \in \{0, 1, \ldots, n\}.$$

Show that for any fixed $c \in \{0, 1, \ldots, n-1\}$, $F_{n,p}(c)$ is a continuous function of $p \in [0, 1]$ with $F_{n,0}(c) = 1$, $F_{n,1}(c) = 0$ and

$$F_{n,p}(c) \;=\; n\binom{n-1}{c} \int_p^1 u^c (1-u)^{n-1-c} \, \mathrm{d}u.$$

Exercise 1.18 (Distribution of order statistics). Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics of independent random variables $X_1, \ldots, X_n$ with distribution function F on $\mathbb{R}$.

Show that for $k \in \{1, 2, \ldots, n\}$,

$$\mathbb{P}(X_{(k)} \leq x) \;=\; 1 - F_{n,F(x)}(k-1),$$

with $F_{n,p}$ as in Exercise 1.17. Now deduce from Exercise 1.17 that

$$\mathbb{P}(X_{(k)} \leq x) \;=\; n\binom{n-1}{k-1} \int_0^{F(x)} u^{k-1} (1-u)^{n-k} \, \mathrm{d}u.$$

Exercise 1.19 (Distribution of the sample median). Let $X_1, \ldots, X_n$ be independent random variables with density f and differentiable distribution function F on $\mathbb{R}$. Suppose that $n = 2m + 1$ for some integer $m \geq 1$.

(a) Show that the sample median $M_n := \mathrm{median}(X_1, \ldots, X_n)$ has density

$$f_n(x) \;=\; n\binom{2m}{m} F(x)^m (1 - F(x))^m f(x).$$

(b) Suppose that f is the standard Cauchy density, $f(x) = \pi^{-1}(1 + x^2)^{-1}$. For which values of m is $\mathbb{E}(M_n^2) < \infty$?

Remark 1: One can answer (b) without computing $\mathbb{E}(M_n^2)$ explicitly, utilizing rough bounds for $F(x)$.

Remark 2: One can show here that $n\,\mathbb{E}(M_n^2) \to \pi^2/4$ as $n \to \infty$.

1.4 Beyond Equivariance: Admissibility

Although equivariance is a rather natural requirement, it is not obvious that it isn't too restrictive. Let us first consider a different estimation paradigm.

Bayesian estimation of θ. Suppose that P_0 is given by a density f_0 on $\mathbb{R}$, so the distribution of $\boldsymbol{X}$ is given by the density $f_\theta(\boldsymbol{x}) = \prod_{i=1}^n f_0(x_i - \theta)$, $\boldsymbol{x} \in \mathbb{R}^n$. Now imagine that θ itself is a random variable which is chosen by "mother nature" according to a so-called prior distribution with probability density π on $\mathbb{R}$. That means, for arbitrary Borel sets $C \subset \mathbb{R}$ and $D \subset \mathbb{R}^n$,

$$\mathbb{P}^{\mathrm{B}}(\theta \in C, \boldsymbol{X} \in D) \;=\; \int_C \mathbb{P}_\theta(D)\, \pi(\theta) d\theta \;=\; \int_C \int_D f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\, \pi(\theta) d\theta.$$

Here and thoughout the sequel, the superscript 'B' stands for 'Bayesian' and means that θ is considered as a random variable. (We do not distinguish notationally between the random variable θ and an explicit value θ.)

The latter display and Fubini's theorem show that the joint distribution of $(\theta, \boldsymbol{X})$ is given by the density $(\theta, \boldsymbol{x}) \mapsto g(\theta, \boldsymbol{x}) := f_\theta(\boldsymbol{x})\pi(\theta)$. That means, for any Borel set $B \subset \mathbb{R} \times \mathbb{R}^n$,

$$\mathbb{P}^{\mathrm{B}}((\theta, \boldsymbol{X}) \in B) \;=\; \int_B g(\theta, \boldsymbol{x})\, d(\theta, \boldsymbol{x}).$$

Moreover, by Fubini's theorem,

$$\mathbb{P}^{\mathrm{B}}(\boldsymbol{X} \in D) \;=\; \int_D f^{\mathrm{B}}(\boldsymbol{x})\, d\boldsymbol{x}$$

with

$$f^{\mathrm{B}}(\boldsymbol{x}) \;:=\; \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, \pi(\theta) d\theta.$$

Hence f^{B} describes the marginal distribution of $\boldsymbol{X}$ in the Bayesian framework.

More generally, for any measurable function $h : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}^{\mathrm{B}}\, h(\theta, \boldsymbol{X}) \;=\; \int_{\mathbb{R}\times\mathbb{R}^n} h(\theta, \boldsymbol{x}) g(\theta, \boldsymbol{x})\, d(\theta, \boldsymbol{x}),$$

provided that the latter integral is well-defined. By Fubini's theorem this may be rewritten in two ways:

$$\mathbb{E}^{\mathrm{B}}\, h(\theta, \boldsymbol{X}) \;=\; \int_{\mathbb{R}} \int_{\mathbb{R}^n} h(\theta, \boldsymbol{x}) f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\; \pi(\theta) d\theta \;=\; \int_{\mathbb{R}} \mathbb{E}_\theta\, h(\theta, \boldsymbol{X})\, \pi(\theta) d\theta,$$

and

$$\mathbb{E}^{\mathrm{B}}\, h(\theta, \boldsymbol{X}) \;=\; \int_{\mathbb{R}^n} \int_{\mathbb{R}} h(\theta, \boldsymbol{x})\, Q^{\mathrm{B}}_{\boldsymbol{x}}(d\theta)\; f^{\mathrm{B}}(\boldsymbol{x})\, d\boldsymbol{x},$$

where $Q^{\mathrm{B}}_{\boldsymbol{x}}$ is conditional distribution of θ, given $\boldsymbol{X} = \boldsymbol{x}$, with density

$$\theta \;\mapsto\; \pi(\theta \,|\, \boldsymbol{x}) \;:=\; \begin{cases} f_\theta(\boldsymbol{x})\pi(\theta)/f^{\mathrm{B}}(\boldsymbol{x}) & \text{if } 0 < f^{\mathrm{B}}(\boldsymbol{x}) < \infty, \\ \pi(\theta) & \text{else.} \end{cases}$$

Within the Bayesian framework, $Q^{\mathrm{B}}_{\boldsymbol{x}}$ and $\pi(\cdot \,|\, \boldsymbol{x})$ are called the *posterior distribution* and *posterior density*, respectively, of θ, given $\boldsymbol{X} = \boldsymbol{x}$.

The *Bayes risk* of any estimator $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$ in this framework is defined as

$$\begin{aligned} R^{\mathrm{B}}(\widehat{\theta}) \;&:=\; \mathbb{E}^{\mathrm{B}}\big((\widehat{\theta}(\boldsymbol{X}) - \theta)^2\big) \\ &=\; \int_{\mathbb{R}} \int_{\mathbb{R}^n} (\widehat{\theta}(\boldsymbol{x}) - \theta)^2 f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\; \pi(\theta) d\theta \\ &=\; \int_{\mathbb{R}} R(\widehat{\theta}, \theta)\, \pi(\theta) d\theta. \end{aligned}$$

The general theory of conditional expectations implies that the (essentially) unique minimizer of the Bayes risk is given by

$$\widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;=\; \mathbb{E}^{\mathrm{B}}(\theta \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; \int_{\mathbb{R}} \theta \pi(\theta \,|\, \boldsymbol{x})\, d\theta \;=\; \mathrm{mean}(Q^{\mathrm{B}}_{\boldsymbol{x}}).$$

Furthermore,

$$\begin{aligned} R^{\mathrm{B}}(\widehat{\theta}^{\mathrm{B}}) \;&=\; \int_{\mathbb{R}\times\mathbb{R}^n} (\theta - \widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}))^2 f_\theta(\boldsymbol{x}) \pi(\theta)\, d(\theta, \boldsymbol{x}) \\ &=\; \int_{\mathbb{R}^n} \int_{\mathbb{R}} (\theta - \widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}))^2\, Q^{\mathrm{B}}_{\boldsymbol{x}}(d\theta)\; f^{\mathrm{B}}(\boldsymbol{x})\, d\boldsymbol{x} \\ &=\; \int \mathrm{Var}(Q^{\mathrm{B}}_{\boldsymbol{x}})\, f^{\mathrm{B}}(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned}$$

Note the similiarity between the Bayes-optimal estimator $\widehat{\theta}^{\mathrm{B}}$ and the optimal equivariant estimator $\widehat{\theta}_*$:

$$\begin{aligned} \widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;&=\; \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x})\, \pi(\theta) d\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, \pi(\theta) d\theta, \\ \widehat{\theta}_*(\boldsymbol{x}) \;&=\; \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x})\, d\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x})\, d\theta. \end{aligned}$$

Hence $\widehat{\theta}_*$ may be interpreted as a Bayesian estimator with prior distribution Lebesgue measure, corresponding to $\pi \equiv 1$. Moreover, suppose that π is the density of $\mathrm{N}(\nu, \tau^2)$ for some $\nu \in \mathbb{R}$ and $\tau > 0$. Then $\pi(\theta)$ is proportional to $e^{-(\theta-\nu)^2/(2\tau^2)}$, whence

$$\widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;=\; \int_{\mathbb{R}} \theta f_\theta(\boldsymbol{x}) e^{-(\theta-\nu)^2/(2\tau^2)}\, d\theta \Big/ \int_{\mathbb{R}} f_\theta(\boldsymbol{x}) e^{-(\theta-\nu)^2/(2\tau^2)}\, d\theta.$$

Since $(0,1] \ni e^{-(\theta-\nu)^2/(2\tau^2)} \to 1$ as $\tau \to \infty$, it follows from dominated convergence that

$$\widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;\to\; \widehat{\theta}_*(\boldsymbol{x}) \quad \text{as } \tau \to \infty,$$

provided that the integrals in the enumerator and denominator of $\widehat{\theta}_*(\boldsymbol{x})$ are well-defined.

Example 1.20 (Gaussian model and prior). Suppose that π is the density of $\mathrm{N}(0, \tau^2)$ for some $\tau > 0$, and let $P_0 = \mathrm{N}(0, \sigma^2)$ for given $\sigma > 0$. Then

$$\widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;=\; \frac{n}{n + \sigma^2/\tau^2}\, \bar{x} \quad \text{and} \quad R^{\mathrm{B}}(\widehat{\theta}^{\mathrm{B}}) \;=\; \frac{\sigma^2}{n + \sigma^2/\tau^2}.$$

To verify this, recall that

$$f_\theta(\boldsymbol{x}) \;=\; f_{\bar{x}}(\boldsymbol{x}) \exp\Big(-\frac{n(\theta-\bar{x})^2}{2\sigma^2}\Big).$$

Since $\pi(\theta)$ is proportional to $\exp(-\theta^2/(2\tau^2))$,

$$\begin{aligned} f_\theta(\boldsymbol{x})\pi(\theta) \;&=\; C_1(\boldsymbol{x}) \exp\Big(-\frac{\theta^2}{2\tau^2} - \frac{n(\theta-\bar{x})^2}{2\sigma^2}\Big) \\ &=\; C_2(\boldsymbol{x}) \exp\Big(-\frac{\theta^2}{2}\,\frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2} + \frac{n\bar{x}}{\sigma^2}\,\theta\Big) \\ &=\; C_3(\boldsymbol{x}) \exp\Big(-\frac{(\theta-\beta\bar{x})^2}{2\gamma^2}\Big) \end{aligned}$$

with certain terms $C_1(\boldsymbol{x}), C_2(\boldsymbol{x}), C_3(\boldsymbol{x}) > 0$ and

$$\begin{aligned} \gamma^2 \;&:=\; \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \;=\; \frac{\sigma^2}{n + \sigma^2/\tau^2}, \\ \beta \;&:=\; \frac{n\gamma^2}{\sigma^2} \;=\; \frac{n}{n + \sigma^2/\tau^2}. \end{aligned}$$

In particular, $Q^{\mathrm{B}}_{\boldsymbol{x}} = \mathrm{N}(\beta\bar{x}, \gamma^2)$, so

$$\widehat{\theta}^{\mathrm{B}}(\boldsymbol{x}) \;=\; \beta\bar{x} \quad \text{and} \quad \mathrm{Var}(Q^{\mathrm{B}}_{\boldsymbol{x}}) \;=\; \gamma^2 \;=\; R^{\mathrm{B}}(\widehat{\theta}^{\mathrm{B}}).$$

Admissibility. The use of any estimator $\widehat{\theta}$ is justified if it is admissible in the following sense:

Definition 1.21 (Admissibility). An estimator $\widehat{\theta}$ of θ is called *admissible* if there exists no other estimator $\tilde{\theta}$ such that

$$\begin{aligned} R(\tilde{\theta}, \theta) \;&\le\; R(\widehat{\theta}, \theta) \quad \text{for all } \theta \in \mathbb{R}, \\ R(\tilde{\theta}, \theta_o) \;&<\; R(\widehat{\theta}, \theta_o) \quad \text{for some } \theta_o \in \mathbb{R}. \end{aligned}$$

Example 1.22. A rather trivial example of an admissible, but non-equivariant estimator is given by $\widehat{\theta} \equiv \theta_o$ with some fixed value $\theta_o \in \mathbb{R}$, provided that $f_0 > 0$. Here, $R(\widehat{\theta}, \theta) = (\theta_o - \theta)^2$. If $\tilde{\theta}$ would be another estimator with $R(\tilde{\theta}, \cdot) \leq R(\widehat{\theta}, \cdot)$, then $R(\tilde{\theta}, \theta_o) = 0$ is equivalent to $\mathbb{P}_{\theta_o}(\tilde{\theta} \neq \theta_o) = 0$. But for each $\theta \in \mathbb{R}$, the distribution $\mathbb{P}_\theta$ is absolutely continuous with respect to $\mathbb{P}_{\theta_o}$, so $\mathbb{P}_\theta(\tilde{\theta} \neq \theta_o) = 0$, whence $R(\tilde{\theta}, \theta) = (\theta_o - \theta)^2$.

Exercise 1.23. Suppose that $P_0([-1, 1]) = 1$. Show that the trivial estimator $\widehat{\theta} \equiv 0$ is not admissible.
Proposal: Show that if $\boldsymbol{X} \sim \mathbb{P}_\theta$, then $\max(\boldsymbol{X}) - 1 \leq \theta \leq \min(\boldsymbol{X}) + 1$ almost surely. Now deduce that $\tilde{\theta}(\boldsymbol{X}) := (\max(\boldsymbol{X}) - 1)^+ - (\min(\boldsymbol{X}) + 1)^-$ outperforms $\widehat{\theta}(\boldsymbol{X}) = 0$.

The following theorem shows that in case of a centered Gaussian error distribution P_0, the estimator $\bar{X}$ is indeed admissible.

Theorem 1.24. *If $P_0 = \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$, then $\bar{X}$ is an admissible estimator of θ.*

Proof of Theorem 1.24. The risk function of $\bar{X}$ is constant σ^2/n. Suppose that $\widehat{\theta}$ is an arbitrary estimator such that $R(\widehat{\theta}, \cdot) \leq \sigma^2/n$ on the whole real line. As shown in Exercises 1.25 and 1.26, it follows from $R(\widehat{\theta}, \cdot) < \infty$ on $\mathbb{R}$ that the risk function $R(\widehat{\theta}, \cdot)$ is continuous. Hence if $R(\widehat{\theta}, \theta_o) < \sigma^2/n$ for some $\theta_o \in \mathbb{R}$, then there exist real numbers $\delta > 0$ and $a < b$ such that $R(\widehat{\theta}, \theta) \leq \sigma^2/n - \delta$ for $\theta \in [a, b]$. Now we evaluate the performance of $\widehat{\theta}$ in a Bayesian framework with $\theta \sim \mathrm{N}(0, \tau^2)$ for some $\tau > 0$. Here

$$\begin{aligned} R^{\mathrm{B}}(\widehat{\theta}) &= \mathbb{E}^{\mathrm{B}} R(\widehat{\theta}, \theta) \\ &\leq \mathbb{P}^{\mathrm{B}}(\theta \notin [a, b]) \frac{\sigma^2}{n} + \mathbb{P}^{\mathrm{B}}(\theta \in [a, b]) \Big(\frac{\sigma^2}{n} - \delta \Big) \\ &= \frac{\sigma^2}{n} - \mathbb{P}^{\mathrm{B}}(\theta \in [a, b]) \delta \\ &= \frac{\sigma^2}{n} - \Big(\Phi\Big(\frac{b}{\tau}\Big) - \Phi\Big(\frac{a}{\tau}\Big) \Big) \delta \\ &= \frac{\sigma^2}{n} - \frac{\Phi'(\xi(\tau))(b - a)\delta}{\tau} \end{aligned}$$

for some number $\xi(\tau) \in [a/\tau, b/\tau]$. On the other hand,

$$R^{\mathrm{B}}(\widehat{\theta}) \geq R^{\mathrm{B}}(\widehat{\theta}^{\mathrm{B}}) = \frac{\sigma^2}{n + \sigma^2/\tau^2} \geq \frac{\sigma^2}{n} - \frac{\sigma^4}{n^2 \tau^2}$$

by the elementary inequality $1/(1 + y) \geq 1 - y$ for $y > -1$. These inequalities for $R^{\mathrm{B}}(\widehat{\theta})$ imply that

$$\Phi'(\xi(\tau))(b - a)\delta \leq \frac{\sigma^4}{n^2 \tau}$$

for arbitrary $\tau > 0$. But as $\tau \to \infty$, the left hand side converges to $\Phi'(0)(b - a)\delta > 0$, whereas the right hand side converges to 0. This contradiction shows that $R(\widehat{\theta}, \cdot) \leq \sigma^2/n$ implies that $R(\widehat{\theta}, \cdot) \equiv \sigma^2/n$. □

Exercise 1.25 (Some basic considerations). Let M be a measure on a measurable space $(\Omega, \mathcal{A})$, and let $g, h : \Omega \to \mathbb{R}$ be $\mathcal{A}$-measurable functions such that for real numbers $a < b$,

$$\int (e^{ag} + e^{bg}) |h| \, dM < \infty.$$

(a) Show that

$$L(t) \;:=\; \int e^{tg} h \, dM$$

defines a continuous function $L : [a, b] \to \mathbb{R}$.

(b) Show that L is continuously differentiable on (a, b) with derivative

$$L'(t) \;=\; \int g e^{tg} h \, dM.$$

(c) Show that L is infinitely often differentiable on (a, b) with k-th derivative

$$L^{(k)}(t) \;=\; \int g^k e^{tg} h \, dM.$$

Exercise 1.26 (Continuity of risk functions in simple Gaussian location families). Consider the simple location family with $P_0 = \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$. Let $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$ be an estimator of θ such that the risk

$$R(S, \theta) \;=\; \mathbb{E}_\theta\big((\widehat{\theta} - \theta)^2\big)$$

is finite for any $\theta \in \mathbb{R}$. Show that $R(\widehat{\theta}, \cdot)$ is continuous on $\mathbb{R}$.

Exercise 1.27. Let $Z \sim \mathrm{N}(0, 1)$ and $a \in \mathbb{R}$. Show that $\mathbb{E}(1_{[Z>a]} Z) = \phi(a)$ and $\mathbb{E}(1_{[Z>a]} Z^2) = \Phi(-a) + a\phi(a)$, where ϕ and Φ arethe density and distribution function of Z, respectively.

Exercise 1.28. Suppose that the error distribution equals $P_0 = \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$. If one assumes that $\theta \geq 0$, a possible estimator would be

$$\bar{X}^+.$$

(a) Determine $R(\bar{X}^+, \theta)$ for arbitrary $\theta \in \mathbb{R}$. (Hint: Exercise 1.27.)

(b) Compare $R(\widehat{X}^+, \cdot)$ with $R(\bar{X})$.

Remark: This exercise shows that the estimator $\bar{X}$ of θ is inadmissible in the statistical experiment $\Big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), \big(\mathrm{N}(\theta, \sigma^2)^{\otimes n}\big)_{\theta \geq 0}\Big)$, because $\bar{X}^+$ has strictly smaller risk than $\bar{X}$. Whether or not $\bar{X}^+$ is admissible itself is a different question.

1.5 Location Functionals and Gross Error Models

Estimators as functionals of (empirical) distributions. Consider a random vector $\boldsymbol{X} \in \mathbb{R}^n$ with independent components X_i having distribution P. Most estimators $\widehat{\theta}(\boldsymbol{X})$ may be viewed as a functional $S(\widehat{P})$ of the empirical distribution

$$\widehat{P} \;:=\; \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

i.e.

$$\widehat{P}(B) \;=\; \frac{\#\{i \leq n : X_i \in B\}}{n} \quad \text{and} \quad \int h \, d\widehat{P} \;=\; \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

for $B \subset \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$. For instance

$$\begin{aligned} \bar{X} &= \text{mean}(\widehat{P}), \\ \text{median}(X_1, \ldots, X_n) &= \text{median}(\widehat{P}), \end{aligned}$$

where for arbitrary distributions Q on $\mathbb{R}$,

$$\begin{aligned} \text{mean}(Q) &:= \int x\, Q(dx) \quad \text{provided that } \int |x|\, Q(dx) < \infty, \\ \text{median}(Q) &:= \frac{\min\{x : Q((-\infty, x]) \geq 0.5\} + \max\{x : Q([x, \infty)) \geq 0.5\}}{2}. \end{aligned}$$

It is well-known that the empirical distribution $\widehat{P}$ is a consistent estimator for the underlying distribution P. Precisely,

$$\mathbb{E}\Big(\sup_{\text{intervals } B \subset \mathbb{R}} \big|\widehat{P}(B) - P(B)\big| \Big) = O(n^{-1/2})$$

uniformly in P, and for arbitrary measurable functions $h : \mathbb{R} \to \mathbb{R}$ with $\int |h|\, dP < \infty$,

$$\mathbb{E}\Big| \int h\, d\widehat{P} - \int h\, dP \Big| \to 0 \quad \text{as } n \to \infty.$$

Hence 'reasonable' functionals $S(\cdot)$ should satisfy $S(\widehat{P}) \to_p S(P)$ as $n \to \infty$, at least if P itself is 'reasonable'.

In what follows we consider the families

$$\begin{aligned} \mathcal{P} &:= \big\{\text{probability distributions on } \mathbb{R}\big\}, \\ \mathcal{P}^r &:= \Big\{ P \in \mathcal{P} : \int |x|^r\, P(dx) < \infty \Big\}, \quad r \geq 0, \end{aligned}$$

i.e. $\mathcal{P}^0 = \mathcal{P}$.

Definition 1.29 (Equivariant location functional). An equivariant location functional on $\mathcal{P}^r$ is a function $S : \mathcal{P}^r \to \mathbb{R}$ such that

$$S(P \star \delta_a) = S(P) + a$$

for arbitrary $P \in \mathcal{P}^r$ and $a \in \mathbb{R}$.

Indeed, $\text{mean}(\cdot)$ is an equivariant location functional on $\mathcal{P}^1$, and $\text{median}(\cdot)$ is an equivariant location functional on $\mathcal{P}^0 = \mathcal{P}$.

Gross error models. For a given exponent $r \geq 0$ we consider a simple location family

$$\big(\mathbb{R}^n, \text{Borel}(\mathbb{R}^n), (P_\theta^{\otimes n})_{\theta \in \mathbb{R}}\big) \quad \text{with } P_\theta = P_0 \star \delta_\theta,$$

generated by a given distribution $P_0 \in \mathcal{P}^r$. Now suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with distribution P in a "contamination neighborhood" of some distribution

in $\{P_\theta : \theta \in \mathbb{R}\}$. Precisely, we assume that for some unknown parameter $\theta \in \mathbb{R}$ and some $\epsilon \in (0, 0.5)$,

$$\begin{aligned} P \in \mathcal{U}_\epsilon^r(\theta) &:= \big\{(1-\epsilon)P_\theta + \epsilon Q : Q \in \mathcal{P}^r\big\} \\ &= \big\{Q \in \mathcal{P}^r : Q(B) \geq (1-\epsilon)P_\theta(B) \text{ for any } B \in \mathrm{Borel}(\mathbb{R})\big\}. \end{aligned}$$

The idea behind this "gross error model" is that each observation X_i stems from P_θ with probability $1-\epsilon$, but with a (small) probability ϵ it could follow any other distribution $Q \in \mathcal{P}^r$.

For instance, a well-known problem in sociology is that a certain percentage of people give nonsensical answers on questionnaires. In the natural sciences, it may happen that a measurement device fails completely with small probability or that the measured value is recorded with a wrong or missing decimal point which may result in extreme outliers.

If such a model is realistic, for large sample sizes n we should not worry too much about the sampling error $S(\widehat{P}) - S(P)$ but rather about the systematic error $S(P) - \theta$. Note that in case of an equivariant location functional $S : \mathcal{P}^r \to \mathbb{R}$,

$$\sup_{P \in \mathcal{U}_\epsilon^r(\theta)} \big|S(P) - \theta\big| \;=\; \sup_{P \in \mathcal{U}_\epsilon^r(0)} \big|S(P)\big|$$

for any $\theta \in \mathbb{R}$. For instance, for any $r \geq 1$ and any 'generator' $P_0 \in \mathcal{P}^r$ with $\mathrm{mean}(P_0) = 0$,

$$\sup_{P \in \mathcal{U}_\epsilon(0)} \big|\mathrm{mean}(P)\big| \;\geq\; \sup_{a \in \mathbb{R}} \big|\underbrace{\mathrm{mean}((1-\epsilon)P_0 + \epsilon\delta_a)}_{=\epsilon a}\big| \;=\; \infty.$$

Hence the mean is a problematic functional in the presence of gross errors.

The following theorem of Peter J. Huber, a Swiss mathematician and co-founder of the field of "robust statistics", shows that the median is an optimal equivariant location functional for a broad class of generators P_0.

Theorem 1.30 (Huber). *For any fixed $r \geq 0$ let $P_0 \in \mathcal{P}^r$ with density f_0 such that f_0 is even on $\mathbb{R}$ and non-increasing on $[0, \infty)$. Then for any equivariant location functional $S : \mathcal{P}^r \to \mathbb{R}$ and arbitrary $\epsilon \in (0, 0.5)$,*

$$\sup_{P \in \mathcal{U}_\epsilon^r(0)} \big|S(P)\big| \;\geq\; \sup_{P \in \mathcal{U}_\epsilon^r(0)} \big|\mathrm{median}(P)\big| \;=\; F_0^{-1}\Big(\frac{0.5}{1-\epsilon}\Big),$$

where F_0 and F_0^{-1} are the distribution and quantile function, respectively, of P_0.

Proof of Theorem 1.30. We first show that indeed

$$\sup_{P \in \mathcal{U}_\epsilon(0)^r} \big|\mathrm{median}(P)\big| \;=\; x_\epsilon \;:=\; F_0^{-1}\Big(\frac{0.5}{1-\epsilon}\Big).$$

The assumptions on f_0 imply that with $x_* := \sup\{x > 0 : f_0(x) > 0\} \in (0, \infty]$, the interval $(-x_*, x_*)$ coincides with $\{x \in \mathbb{R} : 0 < F_0(x) < 1\}$, and F_0 is continuous and strictly increasing on $(-x_*, x_*)$. Moreover, $F_0(-x) = 1 - F_0(x)$ for all $x \in \mathbb{R}$. Since $0 < \epsilon < 0.5$, the number x_ϵ

lies in $(0, x_*)$. The distribution function F of $P = (1-\epsilon)P_0 + \epsilon Q \in \mathcal{U}_\epsilon^r(P_0)$ is strictly increasing on $(-x_*, x_*)$ as well and satisfies

$$F(-x_\epsilon) \;\le\; (1-\epsilon)F_0(-x_\epsilon) + \epsilon \;=\; (1-\epsilon)\Big(1 - \frac{0.5}{1-\epsilon}\Big) + \epsilon \;=\; 0.5$$

with equality if, and only if, $Q((-\infty, -x_\epsilon]) = 1$. On the other hand,

$$F(x_\epsilon) \;\ge\; (1-\epsilon)F_0(x_\epsilon) \;=\; (1-\epsilon)0.5/(1-\epsilon) \;=\; 0.5$$

with equality if, and only if, $Q((-\infty, x_\epsilon]) = 0$. These considerations show that the maximum of $\big|\mathrm{median}(P)\big|$ over all $P \in \mathcal{U}_\epsilon^r(0)$ equals x_ϵ.

Now we construct two particular distributions $P^{(1)}, P^{(2)} \in \mathcal{U}_\epsilon^r(0)$ such that

$$P^{(2)} \;=\; P^{(1)} \star \delta_{2x_\epsilon}.$$

If this is possible, then for any equivariant location functional $S : \mathcal{R}^r \to \mathbb{R}$,

$$S(P^{(2)}) - S(P^{(1)}) \;=\; 2x_\epsilon.$$

This implies that $S(P^{(1)}) \le -x_\epsilon$ or $S(P^{(2)}) \ge x_\epsilon$, whence

$$\sup_{P \in \mathcal{U}_\epsilon^r(0)} \big|S(P)\big| \;\ge\; x_\epsilon.$$

The construction starts from the function $(1-\epsilon)f_0$ and noting that

$$\int_{-\infty}^{x_\epsilon} (1-\epsilon)f_0(x)\,dx \;=\; (1-\epsilon)F_0(x_\epsilon) \;=\; 0.5.$$

Shifting this function to the right by $2x_\epsilon$ yields the function $(1-\epsilon)f_0(x-2x_\epsilon)$, and the assumptions about f_0 imply that

$$(1-\epsilon)f_0(x) \;\left\{\begin{matrix}\ge\\ \le\end{matrix}\right\}\; (1-\epsilon)f_0(x-2x_\epsilon) \quad \text{if} \quad x \;\left\{\begin{matrix}\le\\ \ge\end{matrix}\right\}\; x_\epsilon,$$

and

$$\int_{x_\epsilon}^{\infty} (1-\epsilon)f_0(x-2x_\epsilon)\,dx \;=\; \int_{-x_\epsilon}^{\infty} (1-\epsilon)f_0(x)\,dx \;=\; 1 - F_0(-x_\epsilon) \;=\; 0.5.$$

This shows that

$$f^{(2)} \;:=\; (1-\epsilon)\max\big\{f_0(x), f_0(x-2x_\epsilon)\big\}$$

defines a probability density such that the corresponding distribution $P^{(2)}$ belongs to $\mathcal{U}_\epsilon^r(0)$. Instead of shifting $(1-\epsilon)f_0$ to the right, we could shift it by $2x_\epsilon$ to the left and would obtain the density

$$f^{(1)} \;:=\; (1-\epsilon)\max\big\{f_0(x), f_0(x+2x_\epsilon)\big\}$$

of a distribution $P^{(1)} \in \mathcal{U}_\epsilon^r(0)$. But $f^{(2)} = f^{(1)}(\cdot - 2x_\epsilon)$, so $P^{(2)} = P^{(1)} \star \delta_{2x_\epsilon}$, as desired. Figure 1.1 illustrates the construction of $f^{(1)}, f^{(2)}$. □

Remark. There seems to be no simple location family such that $\mathrm{Median}(\boldsymbol{x})$ is the corresponding Pitman estimator. On the other hand, if P_0 is the centered Laplace distribution with density $f_0(x) = (2\sigma)^{-1}\exp(-|x|/\sigma)$, then $\mathrm{Median}(\boldsymbol{x}) = \widehat{\theta}_{\mathrm{ML}}(\boldsymbol{x})$, and one can show that this estimator is approximately optimal as $n \to \infty$.

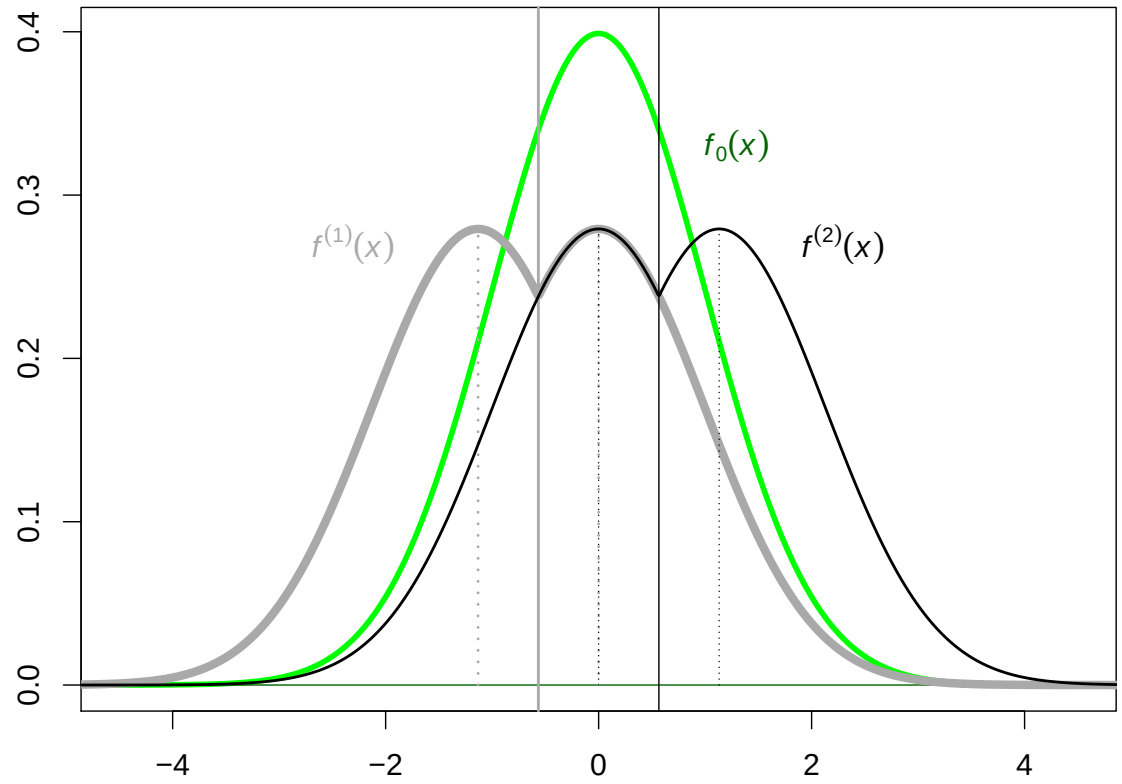

Figure 1.1: Construction of two particular distributions $P^{(1)}, P^{(2)} \in \mathcal{U}_{0.3}(0)$ with densities $f^{(1)}, f^{(2)}$ in case of $P_0 = \mathrm{N}(0, 1)$.

Chapter 2

Statistical Tests

In this chapter we consider a general statistical experiment $\big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$. Recall that $(P_\theta)_{\theta\in\Theta}$ describes potential distributions of observed data $X \in \mathcal{X}$. Suppose for the moment that the observed data are indeed a realization of a random variable with distribution P_θ for an unknown *true parameter* θ. Sometimes we conjecture that θ does not belong to a given set $\Theta_o \subset \Theta$. That means, our *working hypothesis* is that $\theta \in \Theta \setminus \Theta_o$, and we would like to falsify the *null hypothesis* that $\theta \in \Theta_o$ based on the observed data. This can be formalized by a measurable function

$$\varphi : \mathcal{X} \to \{0, 1\}.$$

If $\varphi(X) = 1$, then we *claim that* $\theta \notin \Theta_o$. In other words we *reject the null hypothesis*. In case of $\varphi(X) = 0$ we make *no assertion about* θ.

For theoretical and other reasons it is useful to consider *randomized tests* φ, i.e. measurable mappings

$$\varphi : \mathcal{X} \to [0, 1].$$

The idea is that after observing the data $X \in \mathcal{X}$, we reject the null hypothesis with (conditional) probability $\varphi(X)$. For instance, we could generate an additional random variable $U \sim \mathrm{Unif}[0, 1]$, independently from X, and reject the null hypothesis if $U \le \varphi(X)$. All in all, by Fubini's theorem, the probability of rejecting the null hypothesis equals

$$E_\theta(\varphi) \;=\; \int \varphi \,\mathrm{d}P_\theta.$$

This is equal to $P_\theta(\varphi = 1)$ in case of a $\{0, 1\}$-valued mapping φ.

Definition 2.1 (Statistical test, power function). A *(statistical) test* is a measurable mapping $\varphi : \mathcal{X} \to [0, 1]$. If φ takes only values in $\{0, 1\}$, this can be indicated by saying that φ is a *non-randomized test*. The power function of a test φ is the function

$$\Theta \ni \theta \;\mapsto\; E_\theta(\varphi) = \int \varphi \,\mathrm{d}P_\theta,$$

and $E_\theta(\varphi)$ is the *power of* φ *for parameter* θ.

Note that this definition does not involve any null or working hypothesis. Let us come back to observed data X with distribution P_θ for an unknown true parameter $\theta \in \Theta$. If we use a test φ to check the null hypothesis that $\theta \in \Theta_o$, there are two possible types of error:

Error of the first kind: The true parameter θ belongs to Θ_o, but we reject the null hypothesis.
Error of the second kind: The true parameter θ does not belong to Θ_o, but we do not reject the null hypothesis.

An error of the first kind happens with probability

$$\begin{cases} E_\theta(\varphi) & \text{if } \theta \in \Theta_o, \\ 0 & \text{if } \theta \notin \Theta_o, \end{cases}$$

whereas an error of the second kind occurs with probability

$$\begin{cases} 0 & \text{if } \theta \in \Theta_o, \\ 1 - E_\theta(\varphi) & \text{if } \theta \notin \Theta_o. \end{cases}$$

Traditionally one tries to control the probability of an error of the 1st kind.

Definition 2.2 (Test level). Let $\emptyset \subsetneq \Theta_o \subsetneq \Theta$, and let $\alpha \in (0,1)$. Suppose that φ is a test such that

$$E_\theta(\varphi) \;\leq\; \alpha \quad \text{for all } \theta \in \Theta_o.$$

Then φ is called a *test of the null hypothesis Θ_o at (test) level α*. A shorter formulation: φ is a *level-α test of Θ_o*.

Instead of fixing the test level α and searching for a level-α test φ of the null hypotheses Θ_o, one could also design a test φ and then determine its exact test level

$$\sup_{\theta \in \Theta_o} E_\theta(\varphi).$$

Example 2.3 (Quality control). The producer of a certain gadget wants to learn something about the unknown probability θ that such a device fails in a standardized test of endurance. To this end, he runs an experiment in which n such gadgets are exposed to that endurance test. The outcome of this experiment could be described by a tuple $\boldsymbol{X} = (X_i)_{i=1}^n$ in $\{0,1\}^n$, where X_i specifies whether the i-th gadget fails ($X_i = 1$) or not ($X_i = 0$). Assuming that the n gadgets perform independently, this leads to the Bernoulli experiment described in the introduction, i.e.

$$\big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (P_\theta)_{\theta \in [0,1]}\big),$$

where $P_\theta(\{\boldsymbol{x}\}) = \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})}$, $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$. Note that $T(\boldsymbol{X}) \sim \mathrm{Bin}(n,\theta)$.

Suppose the producer wants to verify that the unknown probability θ is smaller than a given (small) number θ_o. Then he should test the null hypothesis $\Theta_o = [\theta_o, 1]$. If he performs a statistical test of Θ_o at level α, and if that test rejects the null hypothesis, he may claim with confidence $1-\alpha$ that the unknown parameter θ is smaller than θ_o. As we shall justify later, a reasonable non-randomized test for this task is given by $\varphi(\boldsymbol{X}) = 1_{[T(\boldsymbol{X}) \leq k]}$ for some suitable integer $k \in \{0, \ldots, n-1\}$.

With $F_{n,\theta}$ denoting the distribution function of $\mathrm{Bin}(n,\theta)$, the power function of this test is given by

$$E_\theta(\varphi) \;=\; F_{n,\theta}(k).$$

Elementary considerations show that this power function is continuous and strictly decreasing with $E_0(\varphi) = 1$ and $E_1(\varphi) = 0$, see Exercise 1.17. Thus the exact test level is given by

$$\sup_{\theta\in[\theta_o,1]} F_{n,\theta}(k) \;=\; F_{n,\theta_o}(k).$$

Consequently, if the test level $\alpha \in (0,1)$ is given, then one chooses the largest k such that $F_{n,\theta_o}(k) \le \alpha$. If the threshold k is given, one computes the exakt test level $\alpha = F_{n,\theta_o}(k)$.

General goal. Typically we specify a nonempty subset Θ_A of $\Theta \setminus \Theta_o$ and focus on testing the null hypothesis Θ_o *versus the alternative hypothesis* Θ_A. The goal is to construct a level-α test φ of Θ_o with maximal power $E_\theta(\varphi)$ for $\theta \in \Theta_A$.

Exercise 2.4 (De-randomisation). Let $\varphi : \mathcal{X} \to [0,1]$ be a statistical test. Show that for any fixed $\beta \in (0,1)$,

$$\tilde\varphi \;:=\; 1_{[\varphi\ge\beta]}$$

is a non-randomized test satisfying

$$\frac{E_\theta(\varphi)}{\beta} \wedge 1 \;\ge\; E_\theta(\tilde\varphi) \;\ge\; \frac{(E_\theta(\varphi)-\beta)^+}{1-\beta} \quad \text{for all } \theta\in\Theta.$$

Exercise 2.5 (Quality control). The producer of a certain gadget wants to learn something about the unknown probability θ that such a device fails in a standardized test of endurance. To this end, he runs a two-stage experiment:

Stage 1: n_1 such gadgets are exposed to that endurance test, and X_1 is the number of gadgets that fail.

Stage 2: If $X_1 = 0$, another n_2 gadgets are tested, and X_2 is the number of failures in this round. If $X_1 > 0$, we just set $X_2 = 0$.

(a) Describe a statistical experiment $\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$ for the distribution of $X = (X_1, X_2)$. (Use the notation $f_{n_2,\theta}$ for the probability mass function of $\mathrm{Bin}(n_2,\theta)$.)

(b) Let $\varphi(x_1,x_2) := 1_{[x_1=0, x_2\le k]}$ for some $k \in \{0,\dots,n_2-1\}$. Determine the power function $\theta \mapsto E_\theta(\varphi)$ of this test. (Use the notation $F_{n_2,\theta}$ for the distribution function of $\mathrm{Bin}(n_2,\theta)$.)

Plot this function in the case of $n_1 = n_2 = 10$ and $k = 1$.

(c) Show that the power function in part (b) is strictly decreasing with $E_0(\varphi) = 1$ and $E_1(\varphi) = 0$. (Here one should recall or prove the fact that $F_{n_2,\theta}(k)$ is continuous and strictly decreasing in θ with $F_{n_2,0}(k) = 1$ and $F_{n_2,1}(k) = 0$.)

What is the exact level α of this test if the working hypothesis is that θ is smaller than a given value $\theta_o \in (0,1)$? Which value α do you get if $n_1 = n_2 = 10$, $k = 2$, and the working hypothesis is that $\theta < 0.2$?

(d) Let N be the total number of gadgets which are tested. Determine $\mathbb{E}_\theta(N)$.

Plot the function $\theta \mapsto \mathbb{E}_\theta(N)$ in the case of $n_1 = n_2 = 10$.

2.1 The Neyman–Pearson Lemma

We start with the very simple setting of $\Theta = \{0, 1\}$ and $\Theta_o = \{0\}$. Thus we have only two possible distributions P_0, P_1 for the observed data X, and we want to test the null hypothesis $\{0\}$ (that $X \sim P_0$) versus the alternative hypotheses $\{1\}$ (that $X \sim P_1$) at level α.

Theorem 2.6 (Neyman–Pearson). *Suppose that P_0 and P_1 have densities f_0 and f_1, respectively, with respect to some measure M on $(\mathcal{X}, \mathcal{B})$. For any $\alpha \in (0, 1)$ there exist constants $k_\alpha \geq 0$ and $\gamma_\alpha \in [0, 1]$ such that*

$$\varphi_\alpha \;:=\; \begin{cases} 1 & \text{if } f_1 > k_\alpha f_0 \\ \gamma_\alpha & \text{if } f_1 = k_\alpha f_0 \\ 0 & \text{if } f_1 < k_\alpha f_0 \end{cases}$$

defines a test φ_α of $\{0\}$ with the following properties:

(i) The test φ_α has exact level α in the sense that

$$E_0(\varphi_\alpha) \;=\; \alpha.$$

(ii) For any level-α test φ of $\{0\}$,

$$E_1(\varphi) \;\leq\; E_1(\varphi_\alpha).$$

(iii) If φ is a level-α test of $\{0\}$ with $E_1(\varphi) = E_1(\varphi_\alpha)$, then

$$M(f_1 > k_\alpha f_0 \text{ and } \varphi < 1) \;=\; 0 \;=\; M(f_1 < k_\alpha f_0 \text{ and } \varphi > 0).$$

If in addition $k_\alpha > 0$, then $E_0(\varphi) = \alpha$.

(iv) If $P_0 \neq P_1$, then

$$E_1(\varphi_\alpha) \;>\; \alpha.$$

Note that the optimal level-α test φ_α could also be defined in terms of the *likelihood ratio* f_1/f_0 with the conventions that $a/0 := \infty$ for $a > 0$ and $0/0 := 0$:

$$\varphi_\alpha \;:=\; \begin{cases} 1 & \text{if } f_1/f_0 > k_\alpha, \\ \gamma_\alpha & \text{if } f_1/f_0 = k_\alpha, \\ 0 & \text{if } f_1/f_0 < k_\alpha. \end{cases}$$

Remark 2.7 (Existence and choice of M). The assumption that P_0 and P_1 have densities with respect to some measure M on $(\mathcal{X}, \mathcal{B})$ is not a real restriction. If we take $M^o := P_0 + P_1$, then it follows from the theorem of Radon–Nikodym that there exist densities $f_\theta^o = \mathrm{d}P_\theta/\mathrm{d}M^o$ for $\theta = 0, 1$. If M is an arbitrary measure such that a density $f_\theta = \mathrm{d}P_\theta/\mathrm{d}M$ exists for $\theta = 0, 1$, then one can easily verify that $f_\theta^o = f_\theta/(f_1 + f_2)$ on the set $\{f_1 + f_2 > 0\}$, and $M^o(f_1 + f_2 = 0) = 0$. Hence, the resulting optimal test φ_α would be essentially the same, no matter which measure M we start from.

Proof of Theorem 2.6. For the construction of our special test φ_α we consider the auxiliary function $H : [0, \infty) \to [0, 1]$ given by

$$H(r) \;=\; P_0(f_1 \le r f_0).$$

Since $P_0(f_0 = 0) = 0$, we may rewrite this as $H(r) = P_0(f_1/f_0 \le r)$, where $a/0 := \infty$ for $a > 0$ and $0/0 := 0$. Since $P_0(f_1/f_0 = \infty) \le P_0(f_0 = 0) = 0$, H is a distribution function on $[0, \infty)$. That is, H is nonnegative, nondecreasing and right-continuous with limit $H(\infty) = 1$. Consequently, the number

$$k_\alpha \;:=\; \min\bigl\{r \ge 0 : H(r) \ge 1 - \alpha\bigr\}$$

is well-defined. It has the property that

$$P_0(f_1 > k_\alpha f_0) \;=\; 1 - H(k_\alpha) \;\le\; \alpha \;\le\; 1 - H(k_\alpha -) \;=\; P_0(f_1 \ge k_\alpha f_0).$$

If $P_0(f_1 = k_\alpha f_0) = 0$, we set $\gamma_\alpha := 1$. Otherwise we define

$$\gamma_\alpha \;:=\; \frac{\alpha - P_0(f_1 > k_\alpha f_0)}{P_0(f_1 \ge k_\alpha f_0) - P_0(f_1 > k_\alpha f_0)} \;=\; \frac{\alpha - P_0(f_1 > k_\alpha f_0)}{P_0(f_1 = k_\alpha f_0)} \;\in\; [0, 1].$$

In both cases, the test $\varphi_\alpha := 1_{[f_1 = k_\alpha f_0]}\gamma_\alpha + 1_{[f_1 > k_\alpha f_0]}$ satisfies

$$E_0(\varphi_\alpha) \;=\; P_0(f_1 = k_\alpha f_0)\gamma_\alpha + P_0(f_1 > k_\alpha f_0) \;=\; \alpha.$$

This proves property (i).

As to properties (ii-iv), note that for any test φ,

$$(\varphi_\alpha - \varphi)(f_1 - k_\alpha f_0) \;\ge\; 0$$

with strict inequality on the disjoint sets

$$\{f_1 > k_\alpha f_0 \text{ and } \varphi < 1\} \quad \text{and} \quad \{f_1 < k_\alpha f_0 \text{ and } \varphi > 0\},$$

because by construction of φ_α,

$$\varphi_\alpha - \varphi \;=\; \begin{cases} 1 - \varphi \;\ge\; 0 & \text{on } \{f_1 - k_\alpha f_0 > 0\}, \\ \quad -\varphi \;\le\; 0 & \text{on } \{f_1 - k_\alpha f_0 < 0\}. \end{cases}$$

Consequently,

$$\begin{aligned} 0 \;&\le\; \int (\varphi_\alpha - \varphi)(f_1 - k_\alpha f_0)\,\mathrm{d}M \\ &=\; E_1(\varphi_\alpha) - E_1(\varphi) - k_\alpha\bigl(E_0(\varphi_\alpha) - E_0(\varphi)\bigr) \\ &=\; E_1(\varphi_\alpha) - E_1(\varphi) - k_\alpha(\alpha - E_0(\varphi)), \end{aligned}$$

which is equivalent to

$$E_1(\varphi_\alpha) - E_1(\varphi) \;\ge\; k_\alpha(\alpha - E_0(\varphi)). \tag{2.1}$$

Equality holds if and only if

$$M(f_1 > k_\alpha f_0 \text{ and } \varphi < 1) \;=\; 0 \;=\; M(f_1 < k_\alpha f_0 \text{ and } \varphi > 0). \tag{2.2}$$

If φ is a level-α test of $\{0\}$, then the right hand side of (2.1) is non-negative, so $E_1(\varphi) \le E_1(\varphi_\alpha)$. This proves property (ii).

If φ is a level-α test of $\{0\}$ with $E_1(\varphi) = E_1(\varphi_\alpha)$, then the right hand side of (2.1) has to be zero, and the first half of property (iii) is just (2.2). Moreover, if $k_\alpha > 0$, then the right hand side of (2.1) being zero means that $E_0(\varphi) = \alpha$, which proves the second half of property (iii).

Finally, we may compare φ_α with the trivial test $\varphi \equiv \alpha$, so $E_0(\varphi) = E_1(\varphi) = \alpha$. Then (2.1) and (2.2) show that $E_1(\varphi_\alpha) \ge \alpha$ with equality if and only if

$$M(f_1 \ne k_\alpha f_0) \;=\; 0.$$

That means, P_1 has density $k_\alpha f_0$ with respect to M. But then $1 = P_1(\mathcal{X}) = k_\alpha P_0(\mathcal{X}) = k_\alpha$, so $P_1 = P_0$. This proves property (iv). □

Example 2.8. Let $\mathcal{X} = (0, \infty)$ and $P_\theta := \mathrm{Gamma}(a_\theta, b_\theta)$ with shape parameters $a_1 \ge a_0 > 0$ and scale parameters $b_1 \ge b_0 > 0$, where $(a_0, b_0) \ne (a_1, b_1)$. Then the density f_θ of P_θ with respect to Lebesgue measure on $\mathcal{X}$ equals $f_\theta(x) = \Gamma(a_\theta)^{-1} b_\theta^{-a_\theta} x^{a_\theta - 1} e^{-x/b_\theta}$, so

$$\frac{f_1}{f_0}(x) \;=\; \frac{\Gamma(a_0) b_0^{a_0}}{\Gamma(a_1) b_1^{a_1}}\, x^{a_1 - a_0} \exp\bigl((1/b_0 - 1/b_1)x\bigr)$$

is strictly increasing in $x > 0$. Hence the optimal level-α test of $\{0\}$ versus $\{1\}$ is given by

$$\varphi_\alpha(x) \;=\; 1_{[x \ge k_\alpha]},$$

where k_α is the $(1-\alpha)$-quantile of $\mathrm{Gamma}(a_0, b_0)$.

Exercise 2.9. Let $P_0 = \mathrm{N}(0, \sigma^2)$ with $\sigma \le \sqrt{2}$. The corresponding distribution function is $F_0(x) = \Phi(x/\sigma)$. Further let P_1 be the standard logistic distribution with distribution function

$$F_1(x) \;=\; \frac{e^x}{1 + e^x} \;=\; (1 + e^{-x})^{-1}.$$

Show that the Neyman–Pearson test of P_0 versus P_1 (i.e. of $\{0\}$ versus $\{1\}$) at level $\alpha \in (0, 1)$ is given by

$$\varphi_\alpha(x) \;=\; 1_{\left[|x| \,\ge\, \sigma \Phi^{-1}(1 - \alpha/2)\right]}.$$

Hint: Show first that

$$\log \frac{f_1(x)}{f_0(x)}$$

is strictly convex and even, where f_0 and f_1 are the density functions of P_0 and P_1, respectively.

Exercise 2.10. Let be $P_0 = \mathrm{N}(0, 1)$ and $P_1 = \mathrm{N}(\mu, \sigma^2)$ with $\sigma > 1$.

(a) Show that the Neyman–Pearson test of P_0 versus P_1 at level α has the form

$$\varphi_\alpha(x) \;=\; 1_{\left[|x + \mu/(\sigma^2 - 1)| \,\ge\, k_\alpha\right]}$$

for some $k_\alpha = k_\alpha(\mu, \sigma) > 0$.

(b) Determine φ_α in the special case of $\mu = 0$.

(c) Show that the special test φ_α in part (b) has the following property:

$$\int \varphi_\alpha \, \mathrm{dN}(\mu, \sigma^2) \;\geq\; \int \varphi_\alpha \, \mathrm{dN}(0, \sigma^2) \quad \text{for arbitrary } \mu \in \mathbb{R}.$$

Determine the latter power.

2.2 Monotone Density Ratios

In this section we consider a parameter space $\Theta \subset \mathbb{R}$, and we assume that each distribution P_θ has a density $f_\theta > 0$ with respect to a measure M on $(\mathcal{X}, \mathcal{B})$. Moreover, we assume that there exists a measurable function

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

with the following property: For arbitrary $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$, there exists a *non-decreasing function* $g_{\theta_1,\theta_2} : \mathbb{R} \to [0, \infty]$ such that

$$\frac{f_{\theta_2}}{f_{\theta_1}} \;=\; g_{\theta_1,\theta_2}(T).$$

In particular, $0 < g_{\theta_1,\theta_2} < \infty$ on $T(\mathcal{X})$.

Example 2.11 (Bernoulli experiments). Let $\mathcal{X} = \{0,1\}^n$, $\Theta = (0,1)$, and let P_θ describe the joint distribution of n independent random variables with values in $\{0,1\}$ and expectation θ. That means, with M denoting counting measure on $\mathcal{X}$, P_θ has density f_θ given by

$$f_\theta(\boldsymbol{x}) \;=\; \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})}$$

with $T(\boldsymbol{x}) := \sum_{i=1}^n x_i$. Then for $0 < \theta_1 < \theta_2 < 1$ and $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned} \frac{f_{\theta_2}}{f_{\theta_1}}(\boldsymbol{x}) \;&=\; \frac{\theta_2^{T(\boldsymbol{x})}(1-\theta_2)^{n-T(\boldsymbol{x})}}{\theta_1^{T(\boldsymbol{x})}(1-\theta_1)^{n-T(\boldsymbol{x})}} \\ &=\; \frac{(1-\theta_2)^n}{(1-\theta_1)^n}\Big(\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}\Big)^{T(\boldsymbol{x})} \\ &=\; g_{\theta_1,\theta_2}(T(\boldsymbol{x})) \end{aligned}$$

with

$$g_{\theta_1,\theta_2}(t) \;:=\; \frac{(1-\theta_2)^n}{(1-\theta_1)^n}\Big(\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}\Big)^{t}.$$

Note that $g_{\theta_1,\theta_2}(t)$ is strictly increasing in $t \in \mathbb{R}$, because

$$\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1} \;=\; \frac{\theta_2}{1-\theta_2}\Big/\frac{\theta_1}{1-\theta_1} \;>\; 1.$$

Example 2.12 (Gaussian location family). Let $\mathcal{X} = \mathbb{R}^n$, $\Theta = \mathbb{R}$ and $P_\theta = \mathrm{N}(\theta, \sigma^2)^{\otimes n}$ for a fixed standard deviation $\sigma > 0$. Recall that the density f_θ of P_θ with respect to Lebesgue measure on $\mathbb{R}^n$ is given by

$$f_\theta(\boldsymbol{x}) \;=\; (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{\|\boldsymbol{x}-\theta\|^2}{2\sigma^2}\Big) \;=\; (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{\|\boldsymbol{x}-\bar{x}\|^2 + n(\bar{x}-\theta)^2}{2\sigma^2}\Big).$$

Thus for $\theta_1 < \theta_2$,

$$\begin{aligned}\frac{f_{\theta_2}}{f_{\theta_1}}(\boldsymbol{x}) \;&=\; \exp\Big(\frac{n(\bar{x}-\theta_1)^2 - n(\bar{x}-\theta_2)^2}{2\sigma^2}\Big)\\ &=\; \exp\Big(\frac{n(\theta_2-\theta_1)}{\sigma^2}\,\bar{x} + \frac{n(\theta_1^2-\theta_2^2)}{2\sigma^2}\Big)\\ &=\; g_{\theta_1,\theta_2}(T(\boldsymbol{x})),\end{aligned}$$

where $T(\boldsymbol{x}) := \bar{x}$, and

$$g_{\theta_1,\theta_2}(t) \;:=\; \exp\Big(\frac{n(\theta_2-\theta_1)}{\sigma^2}\,t + \frac{n(\theta_1^2-\theta_2^2)}{2\sigma^2}\Big)$$

is strictly increasing in $t \in \mathbb{R}$.

Example 2.13 (Gamma families). As in Example 2.8, let $\mathcal{X} = (0,\infty)$, and let $\mathrm{Gamma}(a,b)$ be the gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Now consider an arbitrary parameter set $\Theta \subset \mathbb{R}$ and parameter pairs $(a_\theta, b_\theta) \in (0,\infty) \times (0,\infty)$ such $\theta \mapsto a_\theta$ and $\theta \mapsto b_\theta$ are non-decreasing. Specific examples with $\Theta = (0,\infty)$ are given by $(a_\theta, b_\theta) = (\theta, b)$ for some fixed $b > 0$, or $(a_\theta, b_\theta) = (a, \theta)$ for some fixed $a > 0$.

In general, for parameters $\theta(1) < \theta(2)$ in Θ,

$$g_{\theta(1),\theta(2)}(x) \;=\; \frac{f_{\theta(1)}}{f_{\theta(2)}}(x) \;=\; C_{\theta(1),\theta(2)} x^{a_{\theta(2)}-a_{\theta(1)}} \exp\big((1/b_{\theta(1)} - 1/b_{\theta(2)})x\big)$$

for some $C_{\theta(1),\theta(2)} > 0$, and this is strictly increasing in $x > 0$, unless $a_{\theta(1)} = a_{\theta(2)}$ and $b_{\theta(1)} = b_{\theta(2)}$.

In statistical models with monotone density ratios as above, there exist optimal tests of null hypotheses of the form

$$\Theta_o \;=\; \Theta \cap (-\infty, \theta_o] \quad \text{or} \quad \Theta_o \;=\; \Theta \cap [\theta_o, \infty)$$

with arbitrary $\theta_o \in \Theta$. An essential tool for the derivation of these tests is the following lemma.

Lemma 2.14 (Stochastic order). *For any bounded and non-decreasing function $h : \mathbb{R} \to \mathbb{R}$,*

$$E_\theta(h(T)) \;=\; \int h(T)\,\mathrm{d}P_\theta$$

is non-decreasing in $\theta \in \Theta$.

Proof of Lemma 2.14. Let $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 < \theta_2$ and $P_{\theta_1} \not\equiv P_{\theta_2}$. With $g := g_{\theta_1,\theta_2}$,

$$\begin{aligned}\int h(T)\,\mathrm{d}P_{\theta_2} - \int h(T)\,\mathrm{d}P_{\theta_1} \;&=\; \int h(T)g(T)\,\mathrm{d}P_{\theta_1} - \int h(T)\,\mathrm{d}P_{\theta_1}\\ &=\; \int h(T)(g(T)-1)\,\mathrm{d}P_{\theta_1}.\end{aligned}$$

For $h \equiv 1$ we obtain

$$\int (g(T) - 1) \, \mathrm{d}P_{\theta_1} \;=\; 0.$$

Since g is non-decreasing and $P_{\theta_1} \not\equiv P_{\theta_2}$, the latter equation implies that for some $t_o \in \mathbb{R}$,

$$g(t) \begin{cases} \geq\ 1 & \text{whenever } t > t_o, \\ \leq\ 1 & \text{whenever } t < t_o. \end{cases}$$

But then we may conclude that

$$\int h(T)(g(T) - 1) \, \mathrm{d}P_{\theta_1} \;=\; \int (h(T) - h(t_o))(g(T) - 1) \, \mathrm{d}P_{\theta_1} \;\geq\; 0,$$

because the latter integrand is everywhere non-negative. □

Theorem 2.15 (Uniformly most powerful (UMP) right-sided tests). *Let $\theta_o \in \Theta$.*

(i) *For any fixed $\alpha \in (0, 1)$ there exist constants $k_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0, 1]$ such that the test*

$$\varphi_\alpha \;:=\; 1_{[T=k_\alpha]}\gamma_\alpha + 1_{[T>k_\alpha]}$$

satisfies

$$E_{\theta_o}(\varphi_\alpha) \;=\; \alpha.$$

(ii) *A test φ_α as in part (i) has the following properties:*

(ii.1) The power function $\theta \mapsto E_\theta(\varphi_\alpha)$ is non-decreasing on Θ with values in $(0, 1)$. In particular,

$$\sup_{\theta \in \Theta : \theta \leq \theta_o} E_\theta(\varphi_\alpha) \;=\; E_{\theta_o}(\varphi_\alpha) \;=\; \alpha.$$

(ii.2) For any test φ with $E_{\theta_o}(\varphi) \leq \alpha$,

$$E_\theta(\varphi) \;\leq\; E_\theta(\varphi_\alpha) \quad \text{for all } \theta \in \Theta \cap (\theta_o, \infty).$$

(ii.3) For arbitrary parameters $\theta_1 < \theta_2$ in Θ, $E_{\theta_1}(\varphi_\alpha) < E_{\theta_2}(\varphi_\alpha)$, unless $P_{\theta_1} \equiv P_{\theta_2}$.

Remark 2.16 (UMP left-sided tests). The previous theorem carries over with obvious modifications to null hypotheses $\Theta_o = \Theta \cap [\theta_o, \infty)$ for some $\theta_o \in \Theta$. Here the optimal level-α test of Θ_o has the form

$$\varphi_\alpha \;=\; 1_{[T=k_\alpha]}\gamma_\alpha + 1_{[T<k_\alpha]}$$

with suitable constants $k_\alpha \in \mathbb{R}$ and $\gamma_\alpha \in [0, 1]$.

Proof of Theorem 2.15. The existence of $\gamma_\alpha \in [0, 1]$ and $k_\alpha \in \mathbb{R}$ such that $\varphi_\alpha := 1_{[T=k_\alpha]}\gamma_\alpha + 1_{[T>k_\alpha]}$ satisfies $E_{\theta_o}(\varphi_\alpha) = \alpha$ can be verified with the same arguments as in the proof of the Neyman–Pearson lemma: We consider the distribution function $H : \mathbb{R} \to [0, 1]$ with

$$H(r) \;:=\; P_{\theta_o}(T \leq r).$$

Then we define

$$k_\alpha \;:=\; \min\bigl\{r \in \mathbb{R} : H(r) \geq 1 - \alpha\bigr\},$$

so

$$P_{\theta_o}(T > k_\alpha) \;\le\; \alpha \;\le\; P_{\theta_o}(T \ge k_\alpha).$$

In case of $P_{\theta_o}(T = k_\alpha) = 0$ we set $\gamma_\alpha = 1$, otherwise

$$\gamma_\alpha \;:=\; \frac{\alpha - P_{\theta_o}(T > k_\alpha)}{P_{\theta_o}(T = k_\alpha)} \;\in\; [0,1].$$

Then one can easily verify that the resulting test φ_α has power α at θ_o. This proves part (i).

As to part (ii), note first that $\varphi_\alpha = h(T)$ with the nondecreasing function $h : \mathbb{R} \to [0,1]$, $h(t) := 1_{[t=k_\alpha]}\gamma_\alpha + 1_{[t>k_\alpha]}$. Consequently, according to Lemma 2.14, the power function of φ_α is non-decreasing on Θ.

For $\theta \in \Theta \setminus \{\theta_o\}$ let $g_\theta := g_{\theta_o,\theta}$ if $\theta > \theta_o$ and $g_\theta := g_{\theta,\theta_o}^{-1}$ if $\theta < \theta_o$. Then $g_\theta(T) = f_\theta / f_{\theta_o} > 0$, and this implies that $0 < E_\theta(\varphi_\alpha) < 1$. Indeed, $0 = E_\theta(\varphi) = \int \varphi g_\theta(T)\,\mathrm{d}P_{\theta_o}$ would imply that $P_{\theta_o}(\varphi_\alpha > 0) = 0$ and $E_{\theta_o}(\varphi_\alpha) = 0$. And $1 = E_\theta(\varphi_\alpha) = 1 - \int (1-\varphi_\alpha) g_\theta\,\mathrm{d}P_{\theta_o}$ would imply that $P_{\theta_o}(\varphi_\alpha < 1) = 0$ and $\mathbb{E}_{\theta_o}(\varphi_\alpha) = 1$.

These considerations prove property (ii.1).

For arbitrary $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$, the function $g := g_{\theta_1,\theta_2} \in [0,\infty]$ is non-decreasing on $\mathbb{R}$ with $g(T) = f_{\theta_2}/f_{\theta_1} \in (0,\infty)$. Moreover, $0 < g(k_\alpha) < \infty$, because $g(k_\alpha) = 0$ would imply that $\{T \le k_\alpha\} \subset \{g(T) \le 0\} = \emptyset$ and $\varphi_\alpha \equiv 1$, whereas $g(k_\alpha) = \infty$ would imply that $\{T \ge k_\alpha\} \subset \{g(T) \ge \infty\} = \emptyset$ and $\varphi_\alpha \equiv 0$. Hence, any test φ satisfies the inequality

$$(\varphi_\alpha - \varphi)\big(f_{\theta_2} - g(k_\alpha) f_{\theta_1}\big) \;=\; (\varphi_\alpha - \varphi)\big(g(T) - g(k_\alpha)\big) f_{\theta_1} \;\ge\; 0.$$

Consequently

$$\begin{aligned} 0 \;&\le\; \int (\varphi_\alpha - \varphi)\big(f_{\theta_2} - g(k_\alpha) f_{\theta_1}\big)\,\mathrm{d}M \\ &=\; E_{\theta_2}(\varphi_\alpha) - E_{\theta_2}(\varphi) - g(k_\alpha)\big(E_{\theta_1}(\varphi_\alpha) - E_{\theta_1}(\varphi)\big), \end{aligned}$$

so

$$E_{\theta_2}(\varphi_\alpha) - E_{\theta_2}(\varphi) \;\ge\; g_{\theta_1,\theta_2}(k_\alpha)\big(E_{\theta_1}(\varphi_\alpha) - E_{\theta_1}(\varphi)\big). \tag{2.3}$$

In the special case of $\theta_1 = \theta_o$, it follows from (2.3) that $E_{\theta_2}(\varphi_\alpha) \ge E_{\theta_2}(\varphi)$ for arbitrary $\theta_2 > \theta_o$ and any test φ satisfying $E_{\theta_o}(\varphi) \le \alpha = E_{\theta_o}(\varphi_\alpha)$. This proves property (ii.2).

As to property (ii.3), (2.3) shows that for arbitrary parameters $\theta_1 < \theta_2$, the test φ_α is an optimal test of the simple null hypothesis $\{\theta_1\}$ versus the simple alternative hypothesis $\{\theta_2\}$ at level $E_{\theta_1}(\varphi_\alpha) \in (0,1)$. Thus it follows from the Neyman–Pearson lemma that $E_{\theta_2}(\varphi_\alpha) > E_{\theta_1}(\varphi_\alpha)$ unless $P_{\theta_1} \equiv P_{\theta_2}$. □

Example 2.11 (Bernoulli experiments, cont.) Note that the distribution P_θ^T of T equals the binomial distribution $\mathrm{Bin}(n,\theta)$. Let $f_{n,\theta}$ and $F_{n,\theta}$ denote the probability mass function and distri-

bution function of $\mathrm{Bin}(n,\theta)$, respectively, i.e.

$$\begin{aligned} f_{n,\theta}(k) &:= \binom{n}{k}\theta^k(1-\theta)^{n-k}, \\ F_{n,\theta}(x) &:= \sum_{k\le x} f_{n,\theta}(k) \end{aligned}$$

for $k, x \in \{0,1,\dots,n\}$. With the corresponding quantiles

$$F^{-1}_{n,\theta}(u) := \min\{x : F_{n,\theta}(x) \ge u\}, \quad u \in (0,1],$$

for fixed $\theta_o \in (0,1)$ and $\alpha \in (0,1)$, the optimal level-α test of $(0,\theta_o]$ versus $(\theta_o,1)$ is given by

$$\varphi_\alpha = 1_{[T=k_\alpha]}\gamma_\alpha + 1_{[T>k_\alpha]},$$

where

$$k_\alpha = F^{-1}_{n,\theta_o}(1-\alpha) \quad \text{and} \quad \gamma_\alpha = \frac{F_{n,\theta_o}(k_\alpha) - 1 + \alpha}{f_{n,\theta_o}(k_\alpha)}.$$

The power of this test at $\theta \in (0,1)$ equals

$$E_\theta(\varphi_\alpha) = 1 - F_{n,\theta}(k_\alpha) + f_{n,\theta}(k_\alpha)\gamma_\alpha.$$

Example 2.12 (Gaussian location family, cont.) Note that in case of $\boldsymbol{X} \sim \mathrm{N}(\theta,\sigma^2)^{\otimes n}$, the sample mean $\bar{X} = T(\boldsymbol{X})$ has distribution $\mathrm{N}(\theta,\tau^2)$ with $\tau := \sigma/\sqrt{n}$. Hence,

$$P_{\theta_o}(T \ge r) = 1 - \Phi\Big(\frac{r-\theta_o}{\tau}\Big) = \Phi\Big(\frac{\theta_o - r}{\tau}\Big)$$

equals α if and only if $r = \theta_o - \Phi^{-1}(\alpha)\tau$. Consequently an optimal level-α test of $(-\infty,\theta_o]$ versus (θ_o,∞) is given by

$$\varphi_\alpha(\boldsymbol{x}) = 1_{[\bar{x}\ge k_\alpha]} \quad \text{with} \quad k_\alpha = \theta_o - \Phi^{-1}(\alpha)\tau.$$

The power of this test φ_α at an arbitrary parameter θ equals

$$P_\theta(T \ge k_\alpha) = \Phi\Big(\frac{\theta - k_\alpha}{\tau}\Big) = \Phi\big(\Phi^{-1}(\alpha) + (\theta-\theta_o)/\tau\big).$$

Exercise 2.17. Motivated by the Hardy–Weinberg law in genetics, consider the statistical model $(P_\theta)_{\theta\in(0,1)}$ with

$$P_\theta := \mathrm{Mult}\big(n, \boldsymbol{p}(\theta)\big) \quad \text{and} \quad \boldsymbol{p}(\theta) := \big((1-\theta)^2, 2\theta(1-\theta), \theta^2\big).$$

(a) Show that this model has monotone density ratios for a suitable test statistic $T : \mathbb{N}_0^3 \to \mathbb{N}_0$.

(b) Determine the distribution P_θ^T of this test statistic T.

(c) Describe an optimal level-α test of "$\theta \le 0.5$" versus "$\theta > 0.5$".

2.3 The Generalized Neyman–Pearson Lemma

Our goal is to construct optimal tests of null hypotheses Θ_o such that $\#\Theta_o > 1$. In the setting of monotone density ratios, we solved this problem for $\Theta_o = \Theta \cap (-\infty, \theta_o]$ or $\Theta_o = \Theta \cap [\theta_o, \infty)$ with a "least favourable parameter" $\theta_o \in \Theta$. But what about null hypotheses without such a unique least favourable parameter? To deal with such settings, we start with a rather general consideration.

Theorem 2.18 (Generalized Neyman–Pearson lemma). *Let $\mathcal{T}$ be the set of all statistical tests $\varphi : \mathcal{X} \to [0,1]$. Let M be a σ-finite measure on $(\mathcal{X}, \mathcal{B})$, and let $f_1, \ldots, f_m, f_{m+1} \in \mathcal{L}^1(M)$ for some integer $m \geq 1$. Further let $\boldsymbol{\alpha} \in \mathbb{R}^m$ and define*

$$\mathcal{T}(\boldsymbol{\alpha}) \;:=\; \Big\{ \varphi \in \mathcal{T} : \int \varphi \boldsymbol{f} \,\mathrm{d}M = \boldsymbol{\alpha} \Big\}$$

with $\boldsymbol{f} = (f_j)_{j=1}^m : \mathcal{X} \to \mathbb{R}^m$.

(i) *If $\mathcal{T}(\boldsymbol{\alpha}) \neq \emptyset$, then there exists a test $\varphi_{\boldsymbol{\alpha}} \in \mathcal{T}(\boldsymbol{\alpha})$ such that*

$$\int \varphi_{\boldsymbol{\alpha}} f_{m+1} \,\mathrm{d}M \;\geq\; \int \varphi f_{m+1} \,\mathrm{d}M \quad \text{for all } \varphi \in \mathcal{T}(\boldsymbol{\alpha}).$$

(ii) *Suppose that $\varphi_{\boldsymbol{\alpha}}$ is a test in $\mathcal{T}(\boldsymbol{\alpha})$ such that*

$$\varphi_{\boldsymbol{\alpha}}(x) \;=\; \begin{cases} 1 & \text{if } f_{m+1}(x) > \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f}(x) \\ 0 & \text{if } f_{m+1}(x) < \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f}(x) \end{cases}$$

for a certain $\boldsymbol{k}_{\boldsymbol{\alpha}} = (k_{\boldsymbol{\alpha},j})_{j=1}^m \in \mathbb{R}^m$. Then $\varphi_{\boldsymbol{\alpha}}$ has the optimality property in part (i). More generally,

$$\int \varphi_{\boldsymbol{\alpha}} f_{m+1} \,\mathrm{d}M \;\geq\; \int \varphi f_{m+1} \,\mathrm{d}M$$

for arbitrary tests $\varphi \in \mathcal{T}$ such that for $1 \leq j \leq m$,

$$\int \varphi f_j \,\mathrm{d}M \;\begin{cases} \geq\; \alpha_j & \text{if } k_{\boldsymbol{\alpha},j} < 0, \\ \leq\; \alpha_j & \text{if } k_{\boldsymbol{\alpha},j} > 0. \end{cases}$$

(iii) *Suppose that $\boldsymbol{\alpha}$ is an interior point of the set*

$$\Big\{ \int \varphi \boldsymbol{f} \,\mathrm{d}M : \varphi \in \mathcal{T} \Big\} \;\subset\; \mathbb{R}^m.$$

Then there exists a test $\varphi_{\boldsymbol{\alpha}}$ as described in part (ii).

Proof of Theorem 2.18. In what follows let

$$\begin{aligned} \mathcal{K}_m \;&:=\; \Big\{ \int \varphi \boldsymbol{f} \,\mathrm{d}M : \varphi \in \mathcal{T} \Big\} \;\subset\; \mathbb{R}^m, \\ \mathcal{K}_{m+1} \;&:=\; \Big\{ \Big(\int \varphi \boldsymbol{f} \,\mathrm{d}M, \int \varphi f_{m+1} \,\mathrm{d}M \Big) : \varphi \in \mathcal{T} \Big\} \;\subset\; \mathbb{R}^m \times \mathbb{R}. \end{aligned}$$

The set $\mathcal{K}_{m+1}$ is a compact and convex subset of $\mathbb{R}^m \times \mathbb{R}$. This can be verified in two different ways:

With $\mathcal{F} := \mathcal{L}^1(M)$, the set

$$\mathcal{K} \;:=\; \Big\{\Big(\int \varphi f \,\mathrm{d}M\Big)_{f\in\mathcal{F}} : \varphi \in \mathcal{T}\Big\}$$

is a compact and convex subset of $\mathbb{R}^{\mathcal{F}}$, equipped with the product topology, see Theorem A.1 in Appendix A. But $\mathcal{K}_{m+1}$ is the image of $\mathcal{K}$ under the linear and continuous mapping

$$\mathbb{R}^{\mathcal{F}} \ni (x_f)_{f\in\mathcal{F}} \;\mapsto\; \big((x_{f_j})_{j=1}^m, x_{f_{m+1}}\big) \in \mathbb{R}^m \times \mathbb{R},$$

whence it is a compact and convex subset of $\mathbb{R}^m \times \mathbb{R}$.

Alternatively, one can verify directly that $\mathcal{K}_{m+1}$ is a convex and bounded subset of $\mathbb{R}^m \times \mathbb{R}$. But Theorem A.3 implies that it is closed, whence it is compact.

Proof of part (i): Since $\mathcal{K}_{m+1}$ is compact and convex, its intersection with the set

$$\{\boldsymbol{\alpha}\} \times \mathbb{R}$$

is empty or of the form

$$\{\boldsymbol{\alpha}\} \times [a, b]$$

with real numbers $a \le b$. In the latter case there exists a test $\varphi_{\boldsymbol{\alpha}} \in \mathcal{T}(\boldsymbol{\alpha})$ such that

$$\int \varphi_{\boldsymbol{\alpha}} f_{m+1} \,\mathrm{d}M \;=\; b \;=\; \max_{\varphi\in\mathcal{T}(\boldsymbol{\alpha})} \int \varphi f_{m+1} \,\mathrm{d}M.$$

Proof of part (ii): Let $\varphi_{\boldsymbol{\alpha}} \in \mathcal{T}(\boldsymbol{\alpha})$ have the specified special form. Then for any other test φ,

$$(\varphi_{\boldsymbol{\alpha}} - \varphi)(f_{m+1} - \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f}) \;\ge\; 0,$$

whence

$$\begin{aligned}\int \varphi_{\boldsymbol{\alpha}} f_{m+1} \,\mathrm{d}M - \int \varphi f_{m+1} \,\mathrm{d}M \;&=\; \int (\varphi_{\boldsymbol{\alpha}} - \varphi)(f_{m+1} - \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f}) \,\mathrm{d}M \\ &\quad + \sum_{j=1}^m k_{\boldsymbol{\alpha},j}\Big(\alpha_j - \int \varphi f_j \,\mathrm{d}M\Big) \\ &\ge \sum_{j=1}^m k_{\boldsymbol{\alpha},j}\Big(\alpha_j - \int \varphi f_j \,\mathrm{d}M\Big).\end{aligned}$$

The right hand side equals 0, if $\varphi \in \mathcal{T}(\boldsymbol{\alpha})$, so $\varphi_{\boldsymbol{\alpha}}$ maximizes $\int \varphi f_{m+1} \,\mathrm{d}M$ over all tests $\varphi \in \mathcal{T}(\boldsymbol{\alpha})$. More generally, the right hand side is non-negative for all tests φ such that for $1 \le j \le m$,

$$\int \varphi f_j \,\mathrm{d}M \;\begin{cases} \le\; \alpha_j & \text{if } k_{\boldsymbol{\alpha},j} > 0, \\ \ge\; \alpha_j & \text{if } k_{\boldsymbol{\alpha},j} < 0. \end{cases}$$

Proof of part (iii): Let

$$\mathcal{C} \;:=\; \{\boldsymbol{\alpha}\} \times (b, \infty)$$

with

$$b \;:=\; \max_{\varphi\in\mathcal{T}(\boldsymbol{\alpha})} \int \varphi f_{m+1} \,\mathrm{d}M.$$

Then $\mathcal{K}_{m+1}$ and $\mathcal{C}$ are disjoint convex subsets of $\mathbb{R}^m \times \mathbb{R}$. Consequently they may be separated weakly by a hyperplane. That means, there exists a nonzero vector $(\boldsymbol{k}, u) \in \mathbb{R}^m \times \mathbb{R}$ such that

$$\big\langle (\boldsymbol{x}, y), (\boldsymbol{k}, u) \big\rangle \;\le\; \big\langle (\boldsymbol{\alpha}, z), (\boldsymbol{k}, u) \big\rangle \quad \text{for arbitrary } (\boldsymbol{x}, y) \in \mathcal{K}_{m+1} \text{ and } z > b$$

with $\langle \cdot, \cdot \rangle$ denoting the standard inner product on $\mathbb{R}^m \times \mathbb{R}$. Thus

$$\boldsymbol{k}^\top \boldsymbol{x} + uy \;\le\; \boldsymbol{k}^\top \boldsymbol{\alpha} + uz \quad \text{for arbitrary } (\boldsymbol{x}, y) \in \mathcal{K}_{m+1} \text{ and } z > b.$$

Fixing one point $(\boldsymbol{x}, y) \in \mathcal{K}_{m+1}$ and letting $z \to \infty$ shows that $u \ge 0$. In case of $u = 0$, we would have $\boldsymbol{k} \neq \boldsymbol{0}$ and

$$\boldsymbol{k}^\top \boldsymbol{x} \;\le\; \boldsymbol{k}^\top \boldsymbol{\alpha} \quad \text{for arbitrary } \boldsymbol{x} \in \mathcal{K}_m.$$

But then $\boldsymbol{\alpha}$ would be a boundary point of $\mathcal{K}_m$, rather than an interior point. Consequently, $u > 0$, and we may assume without loss of generality that $u = 1$. Consequently, for arbitrary tests $\varphi \in \mathcal{T}$,

$$\boldsymbol{k}^\top \int \varphi \boldsymbol{f} \,\mathrm{d}M + \int \varphi f_{m+1} \,\mathrm{d}M \;\le\; \boldsymbol{k}^\top \boldsymbol{\alpha} + b.$$

If $\varphi_{\boldsymbol{\alpha}} \in \mathcal{T}(\boldsymbol{\alpha})$ with $\int \varphi_{\boldsymbol{\alpha}} f_{m+1} \,\mathrm{d}M = b$, then with $\boldsymbol{k}_{\boldsymbol{\alpha}} := -\boldsymbol{k}$, we may rewrite the previous inequality as

$$\int (\varphi_{\boldsymbol{\alpha}} - \varphi)(f_{m+1} - \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f}) \,\mathrm{d}M \;\ge\; 0$$

for arbitrary tests φ. Applying this inequality to the special test

$$\tilde{\varphi}_{\boldsymbol{\alpha}} \;:=\; \begin{cases} 1 & \text{if } f_{m+1} > \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f} \\ 0 & \text{if } f_{m+1} < \boldsymbol{k}_{\boldsymbol{\alpha}}^\top \boldsymbol{f} \\ \varphi_{\boldsymbol{\alpha}} & \text{else} \end{cases}$$

shows that

$$M(\varphi_{\boldsymbol{\alpha}} \neq \tilde{\varphi}_{\boldsymbol{\alpha}}) \;=\; 0.$$

Hence we may replace $\varphi_{\boldsymbol{\alpha}}$ with $\tilde{\varphi}_{\boldsymbol{\alpha}}$. □

2.4 Tests of Two-Sided Hypotheses

2.4.1 One-parameter exponential families (with natural parametrization)

In what follows we apply the generalized Neyman–Pearson lemma to a particular type of statistical model $\big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta \in \Theta}\big)$: Let Θ be a real interval, and suppose that for some σ-finite measure M on $(\mathcal{X}, \mathcal{B})$,

$$f_\theta(x) = \frac{\mathrm{d}P_\theta}{\mathrm{d}M}(x) \;=\; C(\theta) h(x) \exp(\theta T(x))$$

for given measurable functions $h : \mathcal{X} \to [0, \infty)$, $T : \mathcal{X} \to \mathbb{R}$ and the normalisation constant

$$C(\theta) \;=\; \Big(\int h e^{\theta T} \,\mathrm{d}M \Big)^{-1}.$$

We also assume that $M(h > 0 \text{ and } T \neq c) > 0$ for any real constant c, so $P_{\theta_1} \neq P_{\theta_2}$ for arbitrary different $\theta_1, \theta_2 \in \Theta$.

Example 2.11 (Bernoulli sequences, cont.) This statistical model is an exponential family with $\mathcal{X} = \{0,1\}^n$, M being counting measure on $\mathcal{X}$, and

$$\begin{aligned} \Theta &:= \mathbb{R}, \\ \theta(p) &:= \log(p/(1-p)), \\ h(\boldsymbol{x}) &:= 1, \\ T(\boldsymbol{x}) &:= \sum_{i=1}^{n} x_i, \\ C(\theta) &:= (1+e^{\theta})^{-n}, \qquad \text{i.e. } C(\theta(p)) = (1-p)^n. \end{aligned}$$

Example 2.19 (Poisson distributions). The family of Poisson distributions $\mathrm{Poiss}(\lambda)$, $\lambda > 0$, is a exponential family with $\mathcal{X} = \mathbb{N}_0$, M being counting measure on $\mathcal{X}$, and

$$\begin{aligned} \Theta &:= \mathbb{R}, \\ \theta(\lambda) &:= \log(\lambda), \\ h(x) &:= (x!)^{-1}, \\ T(x) &:= x, \\ C(\theta) &:= \exp(-e^{\theta}), \qquad \text{i.e. } C(\theta(\lambda)) = e^{-\lambda}. \end{aligned}$$

Example 2.12 (Gaussian location family, cont.) The family of distributions $\mathrm{N}(\mu, \sigma^2)^{\otimes n}$, $\mu \in \mathbb{R}$, is an exponential family with $\mathcal{X} = \mathbb{R}^n$, M being Lebesgue measure on $\mathbb{R}^n$, and

$$\begin{aligned} \Theta &:= \mathbb{R}, \\ \theta(\mu) &:= n\mu/\sigma^2, \\ h(\boldsymbol{x}) &:= (2\pi\sigma^2)^{-1/2} \exp\bigl(-\|\boldsymbol{x}\|^2/(2\sigma^2)\bigr), \\ T(\boldsymbol{x}) &:= \bar{x}, \\ C(\theta) &:= \exp\bigl(-\sigma^2\theta^2/(2n)\bigr). \end{aligned}$$

Alternatively, one could choose, for instance,

$$\begin{aligned} \theta(\mu) &:= \mu, \\ T(\boldsymbol{x}) &:= n\bar{x}/\sigma^2, \\ C(\theta) &:= \exp\bigl(-n\theta^2/(2\sigma^2)\bigr), \end{aligned}$$

or

$$\begin{aligned} \theta(\mu) &:= \sqrt{n}\mu/\sigma, \\ T(\boldsymbol{x}) &:= \sqrt{n}\bar{x}/\sigma, \\ C(\theta) &:= \exp(-\theta^2/2). \end{aligned}$$

An advantage of the latter parametrization is that $P_\theta^T = \mathrm{N}(\theta, 1)$.

2.4.2 Two-sided hypotheses, version 1

Now we consider the problem of testing

$$\Theta_o \ := \ \big\{\theta \in \Theta : \theta \le \theta_1 \text{ or } \theta \ge \theta_2\big\} \quad \text{versus} \quad \Theta_A \ := \ (\theta_1, \theta_2)$$

with given parameters $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 < \theta_2$.

Theorem 2.20. ***(i)*** *For any fixed $\alpha \in (0,1)$ there exist real constants $c_1 \le c_2$ and $\gamma_1, \gamma_2 \in [0,1]$ (with $\gamma_2 = 0$ in case of $c_1 = c_2$) such that*

$$\varphi_\alpha \ := \ 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[c_1<T<c_2]}$$

is a test satisfying

$$E_{\theta_1}(\varphi_\alpha) \ = \ E_{\theta_2}(\varphi_\alpha) \ = \ \alpha.$$

(ii) *A test φ_α as in part (i) has the following properties:*

(ii.1) For any level-α test φ of $\{\theta_1, \theta_2\}$,

$$E_\theta(\varphi_\alpha) \ \ge \ E_\theta(\varphi) \quad \text{for all } \theta \in (\theta_1, \theta_2).$$

(ii.2) For arbitrary tests φ such that $E_{\theta_1}(\varphi) = E_{\theta_2}(\varphi) = \alpha$,

$$E_\theta(\varphi_\alpha) \ \le \ E_\theta(\varphi) \quad \text{for all } \theta \in \Theta_o.$$

In particular, φ_α is a level-α test of Θ_o, i.e.

$$E_\theta(\varphi_\alpha) \ \le \ \alpha \quad \text{for all } \theta \in \Theta_o.$$

For the proof of this and later theorems, we need an elementary result about weighted sums of two exponential functions.

Lemma 2.21. *For real numbers $c_1 \le c_2$ and $d_1 < d_2$ there exists a unique vector $\boldsymbol{b} \in \mathbb{R}^2$ such that the function $A : \mathbb{R} \to \mathbb{R}$,*

$$A(t) \ := \ \sum_{j=1}^{2} b_j e^{d_j t}$$

satisfies

$$\begin{cases} A(c_1) = A(c_2) = 1, \\ A'(c_1) = 0 \qquad\qquad \text{if } c_1 = c_2. \end{cases}$$

If $d_1 < 0 < d_2$, then $b_1, b_2 > 0$ and

$$A \begin{cases} < 1 & \text{on } (c_1, c_2), \\ > 1 & \text{on } \mathbb{R} \setminus [c_1, c_2]. \end{cases}$$

If $0 < d_1 < d_2$, then $b_1 > 0 > b_2$, whereas $d_1 < d_2 < 0$ implies that $b_1 < 0 < b_2$. In both cases,

$$A \begin{cases} > 1 & \text{on } (c_1, c_2), \\ < 1 & \text{on } \mathbb{R} \setminus [c_1, c_2]. \end{cases}$$

Proof of Lemma 2.21. We start with existence of a unique $\boldsymbol{b}$ satisfying the stated (in)equalities. Suppose first that $c_1 < c_2$. Then the condition $A(c_1) = A(c_2) = 1$ is equivalent to

$$\boldsymbol{A}\boldsymbol{b} \;=\; (1,1)^\top$$

with the matrix

$$\boldsymbol{A} \;:=\; \begin{bmatrix} e^{d_1c_1} & e^{d_2c_1} \\ e^{d_1c_2} & e^{d_2c_2} \end{bmatrix}.$$

Note that

$$\det(\boldsymbol{A}) \;=\; e^{d_1c_1+d_2c_2} - e^{d_2c_1+d_1c_2} \;=\; e^{d_2c_1+d_1c_2}\big(e^{(d_2-d_1)(c_2-c_1)} - 1\big) \;>\; 0.$$

Hence the equation $\boldsymbol{A}\boldsymbol{b} = (1,1)^\top$ has the unique solution

$$\boldsymbol{b} \;=\; \det(\boldsymbol{A})^{-1} \begin{bmatrix} e^{d_2c_2} & -e^{d_2c_1} \\ -e^{d_1c_2} & e^{d_1c_1} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \;=\; \det(\boldsymbol{A})^{-1} \begin{bmatrix} e^{d_2c_1}\big(e^{d_2(c_2-c_1)} - 1\big) \\ e^{d_1c_2}\big(e^{-d_1(c_2-c_1)} - 1\big) \end{bmatrix},$$

and the stated inequalities for b_1, b_2 are clearly satisfied.

In case of $c_1 = c_2 = c$, the equations $A(c) = 1$ and $A'(c) = 0$ are equivalent to

$$\boldsymbol{A}\boldsymbol{b} \;=\; (1,0)^\top$$

with the matrix

$$\boldsymbol{A} \;:=\; \begin{bmatrix} e^{d_1c} & e^{d_2c} \\ d_1e^{d_1c} & d_2e^{d_2c} \end{bmatrix}.$$

Again, $\det(\boldsymbol{A}) = (d_2 - d_1)e^{(d_1+d_2)c} > 0$, so the equation $\boldsymbol{A}\boldsymbol{b} = (1,0)^\top$ has the unique solution

$$\boldsymbol{b} \;=\; \det(\boldsymbol{A})^{-1} \begin{bmatrix} d_2e^{d_2cc} & -e^{d_2c} \\ -d_1e^{d_1c} & e^{d_1c} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \;=\; \det(\boldsymbol{A})^{-1} \begin{bmatrix} e^{d_2c}d_2 \\ -e^{d_1c}d_1 \end{bmatrix},$$

and the stated inequalities for b_1, b_2 are clearly satisfied.

It remains to verify the additional inequalities for A. In the case of $b_1, b_2 > 0$, the function A is strictly convex, whence $A < 1$ on (c_1, c_2) and $A > 1$ on $\mathbb{R} \setminus [c_1, c_2]$.

In case of $0 < d_1 < d_2$ and $b_1 > 0 > b_2$,

$$A'(t) \;=\; b_1d_1e^{d_1t} + b_2d_2e^{d_2t} \;=\; |b_2|d_2e^{d_1t}\Big(\frac{b_1d_1}{|b_2|d_2} - e^{(d_2-d_1)t}\Big) \;\begin{cases} > 0 & \text{if } t < t_o \\ < 0 & \text{if } t > t_o \end{cases}$$

for some $t_o \in \mathbb{R}$. If $c_1 < c_2$, it follows from $A(c_1) = A(c_2) = 1$ that $t_o \in (c_1, c_2)$, whence $\{A > 1\} = (c_1, c_2)$ and $\{A < 1\} = \mathbb{R} \setminus [c_1, c_2]$. If $c_1 = c_2$, it follows from $A'(c_1) = 0$ that $t_o = c_1$, whence $\{A > 1\} = \emptyset$ and $\{A < 1\} = \mathbb{R} \setminus \{c_1\}$.

Analogous considerations apply in case of $d_1 < d_2 < 0$ and $b_1 < 0 < b_2$. □

Proof of Theorem 2.20. Since $P_\theta(h = 0) = 0$ for all $\theta \in \Theta$, we may replace $\mathcal{X}$ with $\{h > 0\}$ and $M(\mathrm{d}x)$ with $h(x)M(\mathrm{d}x)$, so $h \equiv 1$. Then our statistical experiment has monotone density ratios, and P_θ has density $f_\theta = C(\theta)\exp(\theta T)$ with respect to M.

Proof of part (i). We fix an arbitrary parameter $\theta_* \in (\theta_1, \theta_2)$ and construct an optimal level-α test φ_α of $\{\theta_1, \theta_2\}$ versus $\{\theta_*\}$ by means of the generalized Neyman–Pearson lemma. To this end we consider the point $\boldsymbol{\alpha} := (\alpha, \alpha)^\top$ and the set

$$\mathcal{K}_2 \;=\; \Big\{ \int \varphi \boldsymbol{f} \,\mathrm{d}M : \varphi \in \mathcal{T} \Big\}$$

with $\boldsymbol{f} := (f_{\theta_1}, f_{\theta_2})^\top : \mathcal{X} \to \mathbb{R}^2$. Let $\varphi_0 \equiv 0$, let φ_1 be an optimal level-α test of $\{\theta_1\}$ versus $\{\theta_2\}$, and let φ_2 be an optimal level-α test of $\{\theta_2\}$ versus $\{\theta_1\}$. Then

$$\begin{aligned}
\int \varphi_0 \boldsymbol{f} \,\mathrm{d}M \;&=\; (0,0)^\top, \\
\int \varphi_1 \boldsymbol{f} \,\mathrm{d}M \;&=\; (\alpha, \alpha_1)^\top \quad \text{for some } \alpha_1 > \alpha, \\
\int \varphi_2 \boldsymbol{f} \,\mathrm{d}M \;&=\; (\alpha_2, \alpha)^\top \quad \text{for some } \alpha_2 > \alpha.
\end{aligned}$$

This implies that $\boldsymbol{\alpha}$ is an interior point of the set $\mathcal{K}_2$, see also Exercise 2.22 below. Consequently, there exists a test φ_α such that $E_{\theta_1}(\varphi_\alpha) = E_{\theta_2}(\varphi_\alpha) = \alpha$ with

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } f_{\theta_*} > k_1 f_{\theta_1} + k_2 f_{\theta_2} \\ 0 & \text{if } f_{\theta_*} < k_1 f_{\theta_1} + k_2 f_{\theta_2} \end{cases}$$

for certain constants $k_1, k_2 \in \mathbb{R}$. With $b_j := k_j C(\theta_j)/C(\theta_*)$ we may also write

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } A(T) < 1 \\ 0 & \text{if } A(T) > 1 \end{cases}$$

with

$$A(t) \;:=\; b_1 e^{(\theta_1 - \theta_*)t} + b_2 e^{(\theta_2 - \theta_*)t}.$$

Since $E_{\theta_1}(\varphi_\alpha), E_{\theta_2}(\varphi_\alpha) < 1$, we may conclude that $\max(b_1, b_2) > 0$. But then the function A is strictly monotone or strictly convex. Hence the set $\{A = 1\}$ has at most two elements, and we may replace φ_α with

$$\bar{\varphi}_\alpha \;:=\; \begin{cases} 1 & \text{if } A(T) < 1 \\ M(T=c)^{-1} \displaystyle\int_{\{T=c\}} \varphi_\alpha \,\mathrm{d}M & \text{if } T = c \in \{A = 1\} \\ 0 & \text{else} \end{cases}$$

with the convention that $0/0 := 0$. This does not change the power function of φ_α, because

$$\int_{\{T=c\}} \varphi_\alpha \,\mathrm{d}P_\theta \;=\; C(\theta) e^{\theta c} \int_{\{T=c\}} \varphi_\alpha \,\mathrm{d}M \;=\; C(\theta) e^{\theta c} \int_{\{T=c\}} \bar{\varphi}_\alpha \,\mathrm{d}M \;=\; \int_{\{T=c\}} \bar{\varphi}_\alpha \,\mathrm{d}P_\theta$$

for any $\theta \in \Theta$ and $c \in \{A = 1\}$. Consequently, we may assume that the test φ_α has the form

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } A(T) < 1 \\ 0 & \text{if } A(T) > 1 \\ \gamma(c) & \text{if } T = c \in \{A = 1\} \end{cases}$$

with numbers $\gamma_c \in [0,1]$, $c \in \{A = 1\}$.

Suppose that $b_1 \le 0 < b_2$ or $b_1 > 0 \ge b_2$. In this case $A(\cdot)$ would be strictly monotone, so $E_\theta(\varphi_\alpha)$ would be strictly increasing or strictly decreasing in $\theta \in \Theta$, see Theorem 2.15 (ii). But this would contradict the equation $E_{\theta_1}(\varphi_\alpha) = E_{\theta_2}(\varphi_\alpha)$. Hence $b_1, b_2 > 0$, and $A(\cdot)$ is strictly convex with $A(t) \to \infty$ as $|t| \to \infty$. Consequently, our test φ_α has the asserted form

$$\varphi_\alpha \;=\; 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[c_1<T<c_2]}$$

with real numbers $c_1 \le c_2$ and $\gamma_1, \gamma_2 \in [0,1]$, where $\gamma_2 = 0$ if $c_1 = c_2$.

Proof of part (ii). Let φ_α be a test as in part (i). For arbitrary fixed $\theta \in \Theta \setminus \{\theta_1, \theta_2\}$ let $d_{\theta 1} := \theta_1 - \theta < d_{\theta 2} := \theta_2 - \theta$. Then Lemma 2.21 shows that there exist real numbers $b_{\theta 1}, b_{\theta 2}$ such that

$$A_\theta(t) \;:=\; b_{\theta 1} e^{(\theta_1-\theta)t} + b_{\theta 2} e^{d(\theta_2-\theta)t}$$

satisfies

$$\begin{cases} A_\theta(c_1) = A_\theta(c_2) = 1, \\ A'_\theta(c_1) = 0 \qquad\qquad \text{if } c_1 = c_2. \end{cases}$$

With $k_{\theta j} := b_{\theta j} C(\theta)/C(\theta_j)$, we may write

$$A_\theta(T) \;=\; k_{\theta 1}\frac{f_{\theta_1}}{f_\theta} + k_{\theta 2}\frac{f_{\theta_2}}{f_\theta} \quad \begin{cases} > 1 & \text{if } f_\theta < k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}, \\ < 1 & \text{if } f_\theta > k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}. \end{cases}$$

Suppose first that $\theta_1 < \theta < \theta_2$. Lemma 2.21 shows that both components $b_{\theta j}$ are strictly positive, and

$$A_\theta \quad \begin{cases} > 1 & \text{on } \mathbb{R} \setminus [c_1, c_2], \\ < 1 & \text{on } (c_1, c_2). \end{cases}$$

Hence

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } f_\theta > k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}, \\ 0 & \text{if } f_\theta < k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}. \end{cases}$$

Consequently, property (ii.1) follows from part (ii) of the generalized Neyman–Pearson lemma.

In case of $\theta < \theta_1$ or $\theta > \theta_2$, Lemma 2.21 yields the inequalities

$$A_\theta \quad \begin{cases} < 1 & \text{on } \mathbb{R} \setminus [c_1, c_2], \\ > 1 & \text{on } (c_1, c_2). \end{cases}$$

Hence

$$1 - \varphi_\alpha \;=\; \begin{cases} 1 & \text{if } f_\theta > k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}, \\ 0 & \text{if } f_\theta < k_{\theta 1} f_{\theta_1} + k_{\theta 2} f_{\theta_2}. \end{cases}$$

Consequently we may deduce from part (ii) of the generalized Neyman–Pearson lemma that $E_\theta(\varphi) \le E_\theta(1 - \varphi_\alpha)$ for any test φ such that $E_{\theta_1}(\varphi) = E_{\theta_2}(\varphi) = 1 - \alpha$. In other words, $E_\theta(\varphi) \ge E_\theta(\varphi_\alpha)$ for any test φ such that $E_{\theta_1}(\varphi) = E_{\theta_2}(\varphi) = \alpha$. This is property (ii.2). Considering the special test $\varphi \equiv \alpha$ shows that φ_α is a level-α test of Θ_o. □

Exercise 2.22. Let $K \subset \mathbb{R}^2$ be a convex set containing the three points $(0,0)^\top$, $(\alpha, \alpha_1)^\top$ and $(\alpha_2, \alpha)^\top$ with real numbers $\alpha > 0$ and $\alpha_1, \alpha_2 > \alpha$. Show that $(\alpha, \alpha)^\top$ is an interior point of K.

2.4.3 Two-sided hypotheses, version 2

Version 2a. At first we consider tests of

$$\Theta_o \ := \ [\theta_1, \theta_2] \quad \text{versus } \Theta_A \ := \ \Theta \setminus \Theta_o$$

with given interior points $\theta_1 < \theta_2$ of Θ. This testing problem is essentially the reverse of the testing problem in Theorem 2.20.

Theorem 2.23. ***(i)*** *For any fixed $\alpha \in (0,1)$ there exist real constants $\gamma_1, \gamma_2 \in [0,1]$ and $c_1 \le c_2$ (with $\gamma_2 = 0$ in case of $c_1 = c_2$) such that*

$$\varphi_\alpha \ := \ 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[T<c_1 \text{ or } T>c_2]}$$

is a test satisfying

$$E_{\theta_1}(\varphi_\alpha) \ = \ E_{\theta_2}(\varphi_\alpha) \ = \ \alpha.$$

(ii) *A test φ_α as in part (i) has the following properties:*

(ii.1)

$$E_\theta(\varphi_\alpha) \ \le \ \alpha \quad \text{for all } \theta \in \Theta_o.$$

(ii.2) For any test φ satisfying $E_{\theta_1}(\varphi) = E_{\theta_2}(\varphi) = \alpha$,

$$E_\theta(\varphi_\alpha) \ \ge \ E_\theta(\varphi) \quad \text{for all } \theta \in \Theta_A.$$

Proof of Theorem 2.23. We may apply Theorem 2.20 with $1-\alpha$ in place of α to obtain an optimal level-$(1-\alpha)$ test φ_* of $\Theta \setminus (\theta_1, \theta_2)$ versus (θ_1, θ_2). Then the precise properties of φ_* provided by Theorem 2.20 imply that $\varphi_\alpha := 1 - \varphi_*$ has the properties stated in Theorem 2.23. □

Version 2b. Now we consider tests of

$$\Theta_o \ := \ \{\theta_o\} \quad \text{versus} \quad \Theta_A \ := \ \Theta \setminus \Theta_o$$

for a given interior point θ_o of Θ. Without further constraints on the tests, there exists no globally optimal level-α test of $\{\theta_o\}$. For let

$$\begin{aligned} \varphi_{\alpha,l} \ &= \ 1_{[T<k_{\alpha,l}]} + 1_{[T=k_{\alpha,l}]}\gamma_{\alpha,l}, \\ \varphi_{\alpha,r} \ &= \ 1_{[T>k_{\alpha,r}]} + 1_{[T=k_{\alpha,r}]}\gamma_{\alpha,r} \end{aligned}$$

with real constants $k_{\alpha,l}, k_{\alpha,r}$ and $\gamma_{\alpha,l}, \gamma_{\alpha,r} \in [0,1]$ such that $E_{\theta_o}(\varphi_{\alpha,l}) = E_{\theta_o}(\varphi_{\alpha,r}) = \alpha$. Then by Theorem 2.15 (ii), the power functions of $\varphi_{\alpha,l}$ and $\varphi_{\alpha,r}$ are strictly decreasing and strictly increasing, respectively, and any test φ with $E_{\theta_o}(\varphi) \le \alpha$ satisfies

$$E_\theta(\varphi) \ \le \ \begin{cases} E_\theta(\varphi_{\alpha,l}) & \text{if } \theta \le \theta_o \\ E_\theta(\varphi_{\alpha,r}) & \text{if } \theta \ge \theta_o \end{cases}$$

To obtain a unique optimal test we shall restrict our attention to tests φ such that

$$E_\theta(\varphi) \ \ge \ \alpha \ = \ E_{\theta_o}(\varphi),$$

i.e. the power of φ is nowhere lower than the power of the trivial level-α test $\varphi \equiv \alpha$.

Before going into further detail, let us mention an important property of power functions in the present setting of a one-parameter exponential family: It follows from Exercise 1.25 that for any test φ, the power function

$$\Theta \ni \theta \;\mapsto\; E_\theta(\varphi)$$

is continuous on Θ and continuously differentiable on the interior of Θ with derivative

$$\begin{aligned}
\frac{d}{d\theta} E_\theta(\varphi) \;&=\; \frac{d}{d\theta} \frac{\int \varphi e^{\theta T} h \,\mathrm{d}M}{\int e^{\theta T} h \,\mathrm{d}M} \\
&=\; \frac{\int \varphi T e^{\theta T} h \,\mathrm{d}M}{\int e^{\theta T} h \,\mathrm{d}M} - \frac{\int \varphi e^{\theta T} \,\mathrm{d}M \int T e^{\theta T} \,\mathrm{d}M}{\bigl(\int e^{\theta T} \,\mathrm{d}M\bigr)^2} \\
&=\; \int \varphi T \,\mathrm{d}P_\theta - \int \varphi \,\mathrm{d}P_\theta \int T \,\mathrm{d}P_\theta \\
&=\; \mathrm{Cov}_\theta(\varphi, T).
\end{aligned}$$

Hence, if φ is a test such that

$$E_\theta(\varphi) \;\geq\; \alpha \;\geq\; E_{\theta_o}(\varphi) \quad \text{for all } \theta \in \Theta \setminus \{\theta_o\}, \tag{2.4}$$

then

$$E_{\theta_o}(\varphi) \;=\; \alpha \quad \text{and} \quad \mathrm{Cov}_{\theta_o}(\varphi, T) \;=\; 0.$$

Theorem 2.24. ***(i)*** *For any fixed $\alpha \in (0,1)$, there exist real constants $\gamma_1, \gamma_2 \in [0,1]$ and $c_1 \leq c_2$ (with $\gamma_2 = 0$ in case of $c_1 = c_2$) such that*

$$\varphi_\alpha \;:=\; 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[T<c_1 \text{ or } T>c_2]}$$

is a test satisfying

$$E_{\theta_o}(\varphi_\alpha) \;=\; \alpha \quad \textit{and} \quad \mathrm{Cov}_{\theta_o}(\varphi_\alpha, T) \;=\; 0.$$

(ii) *A test φ_α as in part (i) has the following property: For any test φ such that $E_{\theta_o}(\varphi) = \alpha$ and $\mathrm{Cov}_{\theta_o}(\varphi, T) = 0$,*

$$E_\theta(\varphi_\alpha) \;\geq\; E_\theta(\varphi) \quad \textit{for all } \theta \in \Theta.$$

In particular, φ_α has property (2.4)*, and its power function is pointwise maximal among all tests with that property.*

Proof of Theorem 2.24. Without loss of generality we may assume that $\theta_o = 0$ and $M = P_0$, $h \equiv 1$, because

$$\frac{f_\theta}{f_{\theta_o}} \;=\; \frac{C(\theta)}{C(\theta_o)} \exp((\theta - \theta_o)T) \quad \text{on } \{h > 0\},$$

that means, $C(\theta)C(\theta_o)^{-1}\exp((\theta-\theta_o)T)$ is a density of P_θ with respect to P_{θ_o}. We may further assume without loss of generality that $E_0(T) = 0$, because

$$C(\theta) \cdot \exp(\theta T) \;=\; C(\theta)\exp(\theta E_0(T)) \cdot \exp\bigl(\theta(T - E_0(T))\bigr).$$

Now we construct for a fixed $\theta_1 \in \Theta \setminus \{0\}$ a test φ_α maximizing $E_{\theta_1}(\varphi_\alpha)$ under the constraints that

$$E_0(\varphi_\alpha) \;=\; \int \varphi_\alpha \,\mathrm{d}P_0 \;=\; \alpha \quad \text{and} \quad E_0(\varphi_\alpha T) \;=\; \int \varphi_\alpha T \,\mathrm{d}P_0 \;=\; 0.$$

This may be achieved with the generalized Neyman–Pearson lemma applied to $f_1 := 1$, $f_2 := T$, $f_3 := f_{\theta_1}$ and $M := P_0$. With the four tests $\varphi := 0$, $\varphi := 1$, $\varphi := 1_{[T>0]}$ and $\varphi := 1_{[T\le 0]}$ it follows that

$$\mathcal{K}_2 \;:=\; \Big\{ \int \varphi \boldsymbol{f} \,\mathrm{d}P_0 : \varphi \in \mathcal{T} \Big\}$$

with $\boldsymbol{f} := (f_1, f_2)^\top$ contains the points $(0,0)^\top$, $(1,0)^\top$, $(\beta,\gamma)^\top$ and $(1-\beta,-\gamma)^\top$ with $\beta = P_0(T>0) \in (0,1)$ and $\gamma = \int T^+ \,\mathrm{d}P_0 = \int T^- \,\mathrm{d}P_0 > 0$. Hence $(\alpha,0)^\top$ is an interior point of $\mathcal{K}_2$. Consequently there exist a test φ_α and real constants k_1, k_2 such that

$$\int \varphi_\alpha \,\mathrm{d}P_0 \;=\; \alpha, \quad \int \varphi_\alpha T \,\mathrm{d}P_0 \;=\; 0$$

and

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } f_{\theta_1} - k_1 - k_2 T > 0, \\ 0 & \text{if } f_{\theta_1} - k_1 - k_2 T < 0. \end{cases}$$

Note that $f_{\theta_1} - k_1 - k_2 T = A(T)$, where $A(t) := C(\theta_1) e^{\theta_1 t} - k_1 - k_2 t$ is strictly convex in $t \in \mathbb{R}$. Thus, the set $\{A = 0\}$ consists of at most two different points. We may replace φ_α with

$$\begin{cases} 1 & \text{if } A(T) > 0 \\ \gamma(c) := \displaystyle\int_{\{T=c\}} \tilde{\varphi}_* \,\mathrm{d}P_0 / P_0(T=c) & \text{if } T = c \in \{A=0\} \\ 0 & \text{if } A(T) < 0 \end{cases}$$

with the convention that $0/0 := 0$. This changes neither the power function of φ_α nor the integral $\int \varphi_\alpha T \,\mathrm{d}P_0$. Thus, φ_α can be written as a function of T. If $\{A=0\} = \emptyset$, then strict convexity of A would imply that $A > 0$ on $\mathbb{R}$, that is, $\varphi_\alpha \equiv 1$, a contradiction to $\int \varphi_\alpha \,\mathrm{d}P_0 = \alpha \in (0,1)$. If $\{A=0\} = \{c_1\}$ for some $c_1 \in \mathbb{R}$ but $A'(c_1) \neq 0$, we could deduce from convexity of A that either $\varphi_\alpha = 1_{[T=c_1]}\gamma(c_1) + 1_{[T>c_1]}$ or $\varphi_\alpha = 1_{[T=c_1]}\gamma(c_1) + 1_{[T<c_1]}$, that is, φ_α is a monotone function of T. But then it would follow from Exercise 2.25 below that φ_α is constant P_0-almost surely. Together with $\int \varphi_\alpha \,\mathrm{d}P_0 = \alpha$ and $\int \varphi_\alpha T \,\mathrm{d}P_0 = \int T \,\mathrm{d}P_0 = 0$, this would mean that $\gamma(c_1) = \alpha$ and $T = c_1 = 0$ P_0-almost surely, a contradiction to our assumptions. Hence, we may assume that $\{A=0\} = \{c_1, c_2\}$ with real numbers $c_1 \le c_2$, where $A'(c_1) = 0$ in case of $c_1 = c_2$. Strict convexity of A implies that in both cases, φ_α has the form described in part (i).

As to part (ii), consider an arbitrary fixed parameter $\theta \neq 0$. One can easily verify that there exist real constants $k_{\theta 1}, k_{\theta 2}$ such that $A_\theta(t) := C(\theta) e^{\theta t} - k_{\theta 1} - k_{\theta 2} t$ satisfies

$$\begin{cases} A_\theta(c_1) = A_\theta(c_2) = 0, \\ A'_\theta(c_1) = 0 & \text{if } c_1 = c_2. \end{cases}$$

By strict convexity of A_θ,

$$A_\theta(t) \;\begin{cases} > 0 & \text{if } t \in \mathbb{R} \setminus [c_1, c_2], \\ < 0 & \text{if } t \in (c_1, c_2), \end{cases}$$

whence

$$\varphi_\alpha \;=\; \begin{cases} 1 & \text{if } f_\theta > k_{\theta 1} + k_{\theta 2} T, \\ 0 & \text{if } f_\theta < k_{\theta 1} + k_{\theta 2} T. \end{cases}$$

Consequently, it follows from part (ii) of the generalized Neyman–Pearson lemma that $E_\theta(\varphi_\alpha) \geq E_\theta(\varphi)$ for any test φ such that $\int \varphi \,\mathrm{d}P_0 = \alpha$ and $\int \varphi T \,\mathrm{d}P_0 = 0$. Taking $\varphi \equiv \alpha$ reveals that φ_α has property (2.4) and is optimal among all tests with that property. □

Exercise 2.25. Let Y, T be random variables such that $\mathbb{E}(Y^2), \mathbb{E}(T^2) < \infty$, and suppose that $Y = g(T)$ for some non-decreasing function $g : \mathbb{R} \to \mathbb{R}$. Show that

$$\mathrm{Cov}(Y, T) \;\geq\; 0$$

with equality if and only if $Y = g(\mathbb{E}(T))$ almost surely.

Exercise 2.26. This exercise provides further details about the power function of statistical tests in a one-parameter exponential family with natural parametrization and test statistic $T : \mathcal{X} \to \mathbb{R}$.

(a) Suppose that $f : \mathcal{X} \to \mathbb{R}$ is measurable such that

$$h_f(\theta) \;:=\; E_\theta(f)$$

is well-defined in $\mathbb{R}$ for all $\theta \in \Theta$. Show that h_f is continuous on Θ and differentiable on the interior of Θ with

$$h_f'(\theta) \;=\; E_\theta(fT) - E_\theta(f)E_\theta(T) \;=\; \mathrm{Cov}_\theta(f, T).$$

Hint: Consider Exercise 1.25.

(b) Show that for interior points θ of Θ,

$$h_f''(\theta) \;=\; \mathrm{Cov}_\theta(f, T^2) - 2h_f'(\theta)E_\theta(T).$$

(c) Let θ be an interior point of Θ such that $h_f'(\theta) = 0$. Show that

$$h_f''(\theta) \;=\; E_\theta\big((T - c_1)(T - c_2)(f - E_\theta(f))\big)$$

for arbitrary $c_1, c_2 \in \mathbb{R}$.

(d) Now consider optimal tests of two-sided hypotheses, that is

$$\varphi_\alpha \;:=\; 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + \begin{cases} 1_{[c_1<T<c_2]} & \text{(Type 1)} \\ 1_{[T<c_1 \text{ or } T>c_2]} & \text{(Type 2)} \end{cases}$$

with $c_1 \leq c_2$ and $\gamma_1, \gamma_2 \in [0, 1]$, where $\gamma_2 = 0$ if $c_1 = c_2$. Show that

$$(\varphi_\alpha - E_\theta(\varphi_\alpha))(T - c_1)(T - c_2) \begin{cases} \leq 0 & \text{(Type 1)} \\ \geq 0 & \text{(Type 2)} \end{cases}$$

with strict inequality in case of $T \notin \{c_1, c_2\}$. Deduce from this and part (c), that in case of $h_{\varphi_\alpha}'(\theta) = 0$,

$$h_{\varphi_\alpha}''(\theta) \begin{cases} < 0 & \text{(type 1)} \\ > 0 & \text{(type 2)} \end{cases}$$

unless P_θ^T is concentrated on $\{c_1, c_2\}$.

2.4.4 Summary and some first applications

For notational convenience, we formulated and derived the previous results for one-parameter exponential fanilies with "natural parametrization". That means, the exponential term of the density f_θ contains the product of the test statistic T with the parameter θ. In the examples we looked at in Section 2.4.1, this necessitated a transformation of the original parameters. To get a more complete picture, let us summarize the main results of this section in terms of the original parametrization.

Definition 2.27 (One-parameter exponential family). A one-parameter exponential family is a statistical experiment

$$\big(\mathcal{X}, \mathcal{B}, (P_\lambda)_{\lambda\in\Lambda}\big)$$

of the following form: The parameter space Λ is a real interval. For a σ-finite measure M on $(\mathcal{X}, \mathcal{B})$ and measurable functions $h : \mathcal{X} \to [0, \infty)$, $T : \mathcal{X} \to \mathbb{R}$, each distribution P_λ has density $f_\lambda = \mathrm{d}P_\lambda/\mathrm{d}M$ given by

$$f_\lambda(x) \;=\; C(\lambda)h(x)\exp\big(\theta(\lambda)T(x)\big),$$

where $\theta : \Lambda \to \mathbb{R}$ is a differentiable mapping with $\theta' > 0$ on Λ or $\theta' < 0$ on Λ. Moreover, $M(h > 0 \text{ and } T \neq c) > 0$ for any real constant, so $P_{\lambda_1} \neq P_{\lambda_2}$ whenever $\lambda_1 \neq \lambda_2$.

The parameter $\theta(\lambda)$ in the previous definition is called "natural parameter". The set $\Theta = \theta(\Lambda)$ is a subset of the "natural parameter space"

$$\Theta_{\mathrm{nat}} \;:=\; \Big\{\theta \in \mathbb{R} : \int he^{\theta T}\,\mathrm{d}M < \infty\Big\}.$$

Now we consider tests φ_α of one of the following types:

$$\varphi_\alpha \;=\; 1_{[T=c_1]}\gamma_1 + 1_{[c_1<T<c_2]} + 1_{[T=c_2]}\gamma_2, \tag{2.5}$$

$$\varphi_\alpha \;=\; 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[T<c_1 \text{ or } T>c_2]}, \tag{2.6}$$

where $c_1 \leq c_2$ and $\gamma_1, \gamma_2 \in [0, 1]$ (with $\gamma_2 = 0$ if $c_1 = c_2$).

If M^T is continuous in the sense that $M(T = c) = 0$ for any $c \in \mathbb{R}$, it suffices to consider tests φ_α of the following type:

$$\varphi_\alpha \;=\; 1_{[c_1\leq T\leq c_2]}, \tag{2.7}$$

$$\varphi_\alpha \;=\; 1_{[T\leq c_1 \text{ or } T\geq c_2]}, \tag{2.8}$$

where $c_1 < c_2$.

Two-sided test, version 1. For given points $\lambda_1 < \lambda_2$ in Λ, an optimal level-α test of

$$\Lambda \setminus (\lambda_1, \lambda_2) \quad \text{versus} \quad (\lambda_1, \lambda_2)$$

is given by (2.5) or (2.7), provided that

$$E_{\lambda_1}(\varphi_\alpha) \;=\; \alpha \;=\; E_{\lambda_2}(\varphi_\alpha).$$

Two-sided test, version 2a. For given interior points $\lambda_1 < \lambda_2$ of Λ, an optimal level-α test of

$$[\lambda_1, \lambda_2] \quad \text{versus} \quad \Lambda \setminus [\lambda_1, \lambda_2]$$

with exact power α at λ_1 and λ_2 is given by (2.6) or (2.8), provided that

$$E_{\lambda_1}(\varphi_\alpha) \;=\; \alpha \;=\; E_{\lambda_2}(\varphi_\alpha).$$

Two-sided test, version 2b. For a given interior point λ_o of Λ, an optimal level-α test of

$$\{\lambda_o\} \quad \text{versus} \quad \Lambda \setminus \{\lambda_o\}$$

with power function bounded from below by α is given by (2.6) or (2.8), provided that

$$E_{\lambda_o}(\varphi_\alpha) \;=\; \alpha \quad \text{and} \quad \mathrm{Cov}_{\lambda_o}(\varphi_\alpha, T) \;=\; 0.$$

Example 2.12 (Gaussian location family, cont.) We observe a random vector $\boldsymbol{X} \in \mathbb{R}^n$ with distribution $\mathrm{N}(\mu, \sigma^2)^{\otimes n}$ for a given $\sigma > 0$ and some unknown $\mu \in \mathbb{R}$. As shown before, the model $\big(\mathrm{N}(\mu, \sigma^2)^{\otimes n}\big)_{\mu \in \mathbb{R}}$ is a one-parameter exponential familiy with natural parameter $\theta(\mu) = n\mu/\sigma^2$ and test statistic $T(\boldsymbol{X}) := \bar{X}$. For any fixed μ_o, an optimal level-α test of

$$\{\mu_o\} \quad \text{versus} \quad \mathbb{R} \setminus \{\mu_o\}$$

is given by

$$\varphi_\alpha(\boldsymbol{X}) \;:=\; \begin{cases} 1 & \text{if } |\bar{X} - \mu_o| \geq \tau \Phi^{-1}(1 - \alpha/2), \\ 0 & \text{else,} \end{cases}$$

where $\tau := \sigma/\sqrt{n}$ is the standard deviation of $\bar{X}$. This follows from the fact that φ_α is of type (2.8), and with $Z := (\bar{X} - \mu)/\tau \sim \mathrm{N}(0,1)$,

$$\begin{aligned} E_{\mu_o}(\varphi_\alpha) \;&=\; \mathbb{P}(|Z| \geq \Phi^{-1}(1 - \alpha/2)) \;=\; \alpha, \\ \mathrm{Cov}_{\mu_o}(\varphi_\alpha, T) \;&=\; \tau\, \mathbb{E}\big(1_{[|Z| \leq \Phi^{-1}(1-\alpha/2)]} Z\big) \;=\; 0. \end{aligned}$$

Suppose we want to verify the working hypothesis that

$$|\mu - \mu_o| \;<\; \delta$$

for given numbers $\mu_o \in \mathbb{R}$ and $\delta > 0$. This corresponds to Version 1 of a two-sided test with boundary parameters $\lambda_1 = \mu_o - \delta$ and $\lambda_2 = \mu_o + \delta$. By symmetry reasons, a possible ansatz for φ_α would be

$$\varphi_\alpha(\boldsymbol{X}) \;:=\; \begin{cases} 1 & \text{if } |\bar{X} - \mu_o| \leq \tau c_\alpha \\ 0 & \text{else} \end{cases}$$

for a suitable constant $c_\alpha > 0$. Indeed, this defines a test of type (2.7), and with Z as above we may write

$$\begin{aligned} E_{\mu_o \pm \delta}(\varphi_\alpha) \;=\; \mathbb{P}_{\pm\delta}\big(|\bar{X}| \leq \tau c_\alpha\big) \;&=\; \mathbb{P}\big(|\pm \delta/\tau + Z| \leq c_\alpha\big) \\ &=\; \Phi(c_\alpha \mp \delta/\tau) - \Phi(-c_\alpha \mp \delta/\tau) \\ &=\; \Phi(c_\alpha + \delta/\tau) + \Phi(c_\alpha - \delta/\tau) - 1. \end{aligned}$$

This is obviously a continuous and strictly increasing function of c_α with value 0 for $c_\alpha = 0$ and limit 1 as $c_\alpha \to \infty$. Hence there exists a unique $c_\alpha(\delta/\tau) > 0$ such that

$$E_{\mu_o \pm \delta}(\varphi_\alpha) \;=\; \alpha.$$

The precise value of $c_\alpha(\delta/\tau)$ has to be computed numerically.

If we let $\delta \downarrow 0$, we obtain $c_\alpha(0) = \Phi^{-1}((1+\alpha)/2)$. Indeed, an optimal level-α test of

$$\mathbb{R} \setminus \{\mu_o\} \quad \text{versus} \quad \{\mu_o\}$$

rejects the null hypothesis if

$$\left|\bar{X} - \mu_o\right| \;\leq\; \tau \Phi^{-1}((1+\alpha)/2).$$

In this case, we may claim with confidence $1-\alpha$ that $\mu = \mu_o$. This sounds almost miraculous, but note that the inequality $|\bar{X} - \mu_o| \leq \tau\Phi^{-1}((1+\alpha)/2)$ occurs with probability at most α. Instead of trying to prove that $\mu = \mu_o$, one should rather compute an upper $(1-\alpha)$-confidence bound for $|\mu - \mu_o|$, see later.

Exercise 2.28. Let $X \sim \mathrm{Gamma}(a, b)$ with shape parameter $a > 0$ and scale parameter $b > 0$, i.e. X has density $f_{a,b}$ with respect to Lebesgue measure on $(0, \infty)$, where

$$f_{a,b}(x) \;=\; \Gamma(a)^{-1} b^{-1} (x/b)^{a-1} e^{-x/b}.$$

(a) Suppose that $a > 0$ is given but $b > 0$ is unknown. Verify that $\big(\mathrm{Gamma}(a,b)\big)_{b>0}$ corresponds to a one-parameter exponential family with test statistic $T(X) = X$.

(b) Suppose that $b > 0$ is given but $a > 0$ is unknown. Verify that $\big(\mathrm{Gamma}(a,b)\big)_{a>0}$ corresponds to a one-parameter exponential family with test statistic $T(X) = \log(X)$.

(c) Assuming that $a > 0$ is given but $b > 0$ is an unknown parameter in $(0, \infty)$, determine an optimal level-α test of

$$\{1\} \quad \text{versus} \quad (0,\infty) \setminus \{1\}.$$

(d) Modify your test in part (c) to become an optimal test of

$$\{b_o\} \quad \text{versus} \quad (0,\infty) \setminus \{b_o\}.$$

for arbitrary fixed $b_o > 0$.

Exercise 2.29. Let $X \sim \mathrm{Bin}(n, p)$ with given $n \in \mathbb{N}$ and unknown $p \in [0, 1]$.

(a) Fix small numbers $\alpha \in (0,1)$ and $\delta \in (0, 0.5)$. Construct a statistical procedure to verify with given confidence $1-\alpha$ that $|p - 0.5| < \delta$.

(b) How large should n be such that for a given small $\alpha' \in (0,1)$, this conclusion is drawn with probability at least $1-\alpha'$ in case of $p = 0.5$? Give a numerical answer to this question in case of $\alpha = 0.05$, $\alpha' = 0.3$ and $\delta = 0.1$.

2.5 Tests and Confidence regions

For any statistical model $(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ there is a close relationship between tests and confidence regions. Let uns first clarify what we mean by confidence regions.

Definition 2.30 (Randomized confidence region). A *(randomized) confidence region* is a mapping $C : \mathcal{X} \times \Theta \to [0, 1]$ such that for any fixed $\theta \in \Theta$, the mapping $C(\cdot, \theta)$ is $\mathcal{A}$-measurable.

Suppose that for a given test level $\alpha \in (0, 1)$,

$$\int C(x, \theta)\, P_\theta(\mathrm{d}x) \;\geq\; 1 - \alpha \quad \text{for arbitrary } \theta \in \Theta.$$

Then C is called a *(randomized) confidence region with confidence level* $1 - \alpha$, or shortly: a $(1 - \alpha)$*-confidence region*.

Interpretation, and the meaning of confidence. Suppose that we observe a random data set $X \in \mathcal{X}$ with distribution P_θ, where $\theta \in \Theta$ is unknown. Let U be a random variable with uniform distribution on $[0, 1]$ and independent from X. Then we claim that θ is contained in the set

$$\mathcal{C}(X, U) \;:=\; \big\{\theta \in \Theta : C(X, \theta) \geq U\big\}.$$

In case of C being non-randomized, i.e. C taking only values in $\{0, 1\}$, we do not need the extra random variable U, because the set $\mathcal{C}(X, U)$ equals

$$\mathcal{C}(X) \;:=\; \big\{\theta \in \Theta : C(X, \theta) = 1\big\}$$

almost surely.

If C is a $(1 - \alpha)$-confidence region, then

$$\left.\begin{array}{c} \mathbb{P}_\theta(\mathcal{C}(X) \ni \theta) \\ \mathbb{P}_\theta(\mathcal{C}(X, U) \ni \theta) \end{array}\right\} = \; E_\theta C(\cdot, \theta) \;\geq\; 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

That means, the confidence region covers the unknown true parameter with probability at least $1 - \alpha$. This statement involves the *random variable* X (and U, if needed) and is true *prior* to observing X (and generating U).

Once we have observed X (and generated U, if needed), the claim that $\mathcal{C}(X)$ or $\mathcal{C}(X, U)$ contains the unknown true parameter is simply true or false. Hence it would be a bit ridiculous to say that "with probability $1 - \alpha$, the confidence region $\mathcal{C}(X)$ or $\mathcal{C}(X, U)$ contains θ". Instead, one may claim with *confidence* $1 - \alpha$ that $\theta \in \mathcal{C}(X)$ or $\theta \in \mathcal{C}(X, U)$. This formulation indicates that for a specific observed data set, the claim is simply true or false, but we use a procedure which leads to a correct statement in at least $(1 - \alpha) \cdot 100$ percent of applications in the long run.

Duality between confidence regions and tests of one-point hypotheses. If $C : \mathcal{X} \times \Theta \to [0, 1]$ is a (randomized) confidence region, then for any $\theta_o \in \Theta$, a test of $\{\theta_o\}$ is given by $\varphi(\cdot, \theta_o) := 1 - C(\cdot, \theta_o)$. On the other hand, if for any $\theta_o \in \Theta$ we have defined a test $\varphi(\cdot, \theta_o)$ of $\{\theta_o\}$, then $C(x, \theta) := 1 - \varphi(x, \theta)$ defines a confidence region C. The confidence region C has confidence level $1 - \alpha$ if and only if each test $\varphi(\cdot, \theta_o)$ is a level-α test of $\{\theta_o\}$.

Example 2.12 (Gaussian location family, cont.) We observe $\boldsymbol{X} \sim \mathrm{N}(\mu, \sigma^2)^{\otimes n}$ with unknown mean $\mu \in \mathbb{R}$ and given standard deviation $\sigma > 0$. As shown before, an optimal level-α test of $\{\mu_o\}$ with power function at least α everywhere is given by

$$\varphi(\boldsymbol{X}, \mu_o) \;:=\; 1_{\left[|\bar{X}-\mu_o| \geq \tau\Phi^{-1}(1-\alpha/2)\right]},$$

where $\tau = \sigma/\sqrt{n}$ is the standard deviation of $\bar{X}$. This leads to the $(1-\alpha)$-confidence region C given by

$$C(\boldsymbol{X}, \mu) \;=\; 1_{\left[|\bar{X}-\mu| < \tau\Phi^{-1}(1-\alpha/2)\right]} \;=\; 1_{\left[\mu \in (\bar{X} \pm \tau\Phi^{-1}(1-\alpha/2))\right]},$$

i.e. $\mathcal{C}(\boldsymbol{X}) = \big(\bar{X} \pm \tau\Phi^{-1}(1-\alpha/2)]\big)$.

Exercise 2.31. As in Exercise 2.28, suppose we observe $X \sim \mathrm{Gamma}(a, b)$ with given shape parameter $a > 0$ and unknown scale parameter $b > 0$. An ansatz for a confidence interval for b is

$$\mathcal{C}(X) \;:=\; [X/\kappa_2, X/\kappa_1]$$

with constants $0 < \kappa_1 < \kappa_2$.

(a) Show that for any choice of (κ_1, κ_2), the coverage probability $\mathbb{P}_b(\mathcal{C}(X) \ni b)$ is constant in $b > 0$. Then characterize the set of all pairs (κ_1, κ_2) such that the confidence level is exactly equal to $1 - \alpha$.

(b) A potential measure for the size of $\mathcal{C}(X)$ is the ratio of its upper and lower boundary,

$$\frac{X/\kappa_1}{X/\kappa_2} \;=\; \frac{\kappa_2}{\kappa_1},$$

or the logarithm thereof. Determine the unique pair (κ_1, κ_2) minimizing this size. (The solution is characterized by some equation which could be solved numerically for specific α.) Show also that this pair satisfies $\kappa_1 < a < \kappa_2$.

(c) What is the relation of the solution in part (b) to the optimal test in Exercise 2.28 (b)?

Duality between confidence regions and tests of composite hypotheses. There exist numerous variants of the duality between confidence regions and tests of composite hypotheses. We consider here just one specific version. Suppose we are interested in an upper bound for $g(\theta)$ with a given function

$$g : \Theta \to [0, \infty).$$

For instance, if (Θ, d) is a metric space, we are sometimes interested in the distance between the unknown true parameter and some given point $\theta_o \in \Theta$, so $g(\theta) := d(\theta, \theta_o)$.

An upper confidence bound b_α for $g(\theta)$ corresponds to a measurable function $b_\alpha : \mathcal{X} \to [0, \infty]$ such that

$$P_\theta(b_\alpha \geq g(\theta)) \;\geq\; 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

Such a confidence bound gives rise to the level-α test

$$\varphi_\alpha(x, \delta) \;:=\; 1_{[b_\alpha(x) \leq \delta]}$$

of the null hypothesis

$$\Theta(\delta) \;:=\; \big\{\theta \in \Theta : g(\theta) > \delta\big\}$$

for arbitrary numbers $\delta \geq 0$. Indeed, if $g(\theta) > \delta$, then

$$P_\theta(\varphi_\alpha(\cdot, \delta) = 1) \;=\; P_\theta(b_\alpha \leq \delta) \;\leq\; P_\theta(b_\alpha < g(\theta)) \;\leq\; \alpha.$$

On the other hand, suppose that for any number $\delta \geq 0$, we have constructed a level-α test $\varphi_\alpha(\cdot, \delta) : \mathcal{X} \to \{0, 1\}$ of the null hypothesis $\Theta(\delta)$. Then

$$b_\alpha(x) \;:=\; \sup\big(\{0\} \cup \{\delta \geq 0 : \varphi_\alpha(x, \delta) = 0\}\big)$$

defines an upper $(1 - \alpha)$-confidence bound for $g(\theta)$. Indeed, for any $\theta \in \Theta$ and $\delta := g(\theta)$,

$$P_\theta(g(\theta) \leq b_\alpha) \;\geq\; P_\theta(\varphi_\alpha(\cdot, \delta) = 0) \;=\; 1 - P_\theta(\varphi_\alpha(\cdot, \delta) = 1) \;\geq\; 1 - \alpha.$$

Finding an "optimal" upper confidence bound b_α may be interpreted as finding optimal level-α tests $\varphi_\alpha(\cdot, \delta)$ of the null hypotheses $\Theta(\delta)$, $\delta \geq 0$. Ideally, the function $\varphi_\alpha(\cdot, \cdot)$ is even non-decreasing in its second argument.

Example 2.12 (Gaussian location family, cont.) As shown in the next exercise, for any given $\mu_o \in \mathbb{R}$, a simple upper $(1-\alpha)$-confidence bound for $|\mu - \mu_o|$ is given by $|\bar{X} - \mu_o| + \tau\Phi^{-1}(1-\alpha)$. But if we think about the duality of tests and confidence regions, a good upper $(1 - \alpha)$-confidence bound b_α for $|\mu - \mu_o|$ should satisfy the following condition: For any $\delta \geq 0$,

$$\varphi_\alpha(\boldsymbol{x}, \delta) \;:=\; 1_{[b_\alpha(\boldsymbol{x}) \leq \delta]}$$

defines an optimal level-α test of

$$\mathbb{R} \setminus [\mu_o \pm \delta] \quad \text{versus} \quad [\mu_o \pm \delta].$$

This is essentially Version 1 of our two-sided testing problem, except that the alternative hypothesis is chosen to be a closed rather than an open interval. We know already a solution for this testing problem: An optimal level-α test is given by

$$\varphi_\alpha(\boldsymbol{x}, \delta) \;:=\; 1_{\left[|\bar{x} - \mu_o|/\tau < c_\alpha(\delta/\tau)\right]}$$

where for $\gamma \geq 0$, $c_\alpha(\gamma)$ is the unique number $c > 0$ such that

$$\Phi(c + \gamma) + \Phi(c - \gamma) \;=\; 1 + \alpha.$$

Recall that $c_\alpha(0) = \Phi^{-1}((1 + \alpha)/2)$, but for $\delta > 0$, there is no simple formula for $c_\alpha(\gamma)$. An optimal upper confidence bound for $|\mu - \mu_o|$ is given by

$$\begin{aligned} b_\alpha(\boldsymbol{x}) \;&:=\; \sup\big(\{0\} \cup \big\{\delta \geq 0 : \varphi_\alpha(\boldsymbol{x}, \delta) = 0\big\}\big) \\ &=\; \sup\big(\{0\} \cup \big\{\delta \geq 0 : c(\delta/\tau) \leq |\bar{x} - \mu_o|/\tau\big\}\big) \\ &=\; \sup\big(\{0\} \cup \big\{\tau\gamma : \gamma > 0, \Phi\big(|\bar{x} - \mu_o|/\tau + \gamma\big) + \Phi(|\bar{x} - \mu_o|/\tau - \gamma) \geq 1 + \alpha\big\}\big). \end{aligned}$$

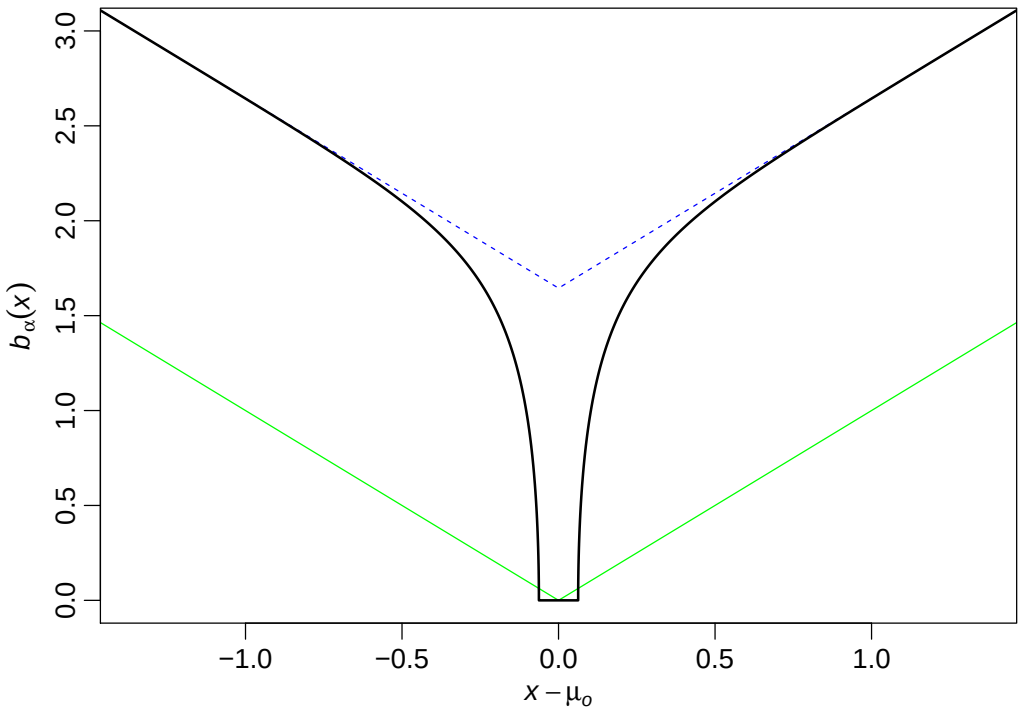

Figure 2.1: Optimal (black) and ad hoc (dashed blue) upper 95%-confidence bound for $|\mu - \mu_o|$ in a simple Gaussian shift model.

If $|\bar{x} - \mu_o|/\tau \leq \Phi^{-1}((1+\alpha)/2)$, then $b_\alpha(\boldsymbol{x}) = 0$. Otherwise, $b_\alpha(\boldsymbol{x})$ is τ times the unique solution $\gamma > 0$ of the equation

$$\Phi\big(|\bar{x} - \mu_o|/\tau + \gamma\big) + \Phi\big(|\bar{x} - \mu_o|/\tau - \gamma\big) \;=\; 1 + \alpha,$$

which can be computed numerically. Note that for any fixed $c > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\big(\Phi(c+\gamma) + \Phi(c-\gamma)\big) \;=\; \phi(c+\gamma) - \phi(c-\gamma) \;=\; \phi(c+\gamma)(1 - \exp(2c\gamma)) \;<\; 0.$$

Figure 2.1 depicts the optimal upper 95%-confidence bound (black curve) for $|\mu - \mu_o|$ and the simple upper bound $|\bar{x} - \mu_o| + \tau\Phi^{-1}(1-\alpha)$ (dashed blue line) in the case of $\sigma = 1$ and $n = 1$.

Exercise 2.32 (An ad hoc uppper bound for $|\mu - \mu_o|$). Suppose we observe $\boldsymbol{X} \sim \mathrm{N}(\mu, \sigma^2)^{\otimes n}$ for some unknown parameter $\mu \in \mathbb{R}$ and given $\sigma > 0$.

(a) Show that for any given $\mu_o \in \mathbb{R}$,

$$\mathcal{C}_\alpha(\boldsymbol{X}) \;:=\; \big[\min(\bar{X} - \tau\Phi^{-1}(1-\alpha), \mu_o), \max(\bar{X} + \tau\Phi^{-1}(1-\alpha), \mu_o)\big]$$

is also a $(1-\alpha)$-confidence interval for μ. Precisely,

$$\mathbb{P}_\mu(\mathcal{C}_\alpha(\boldsymbol{X}) \ni \mu) \;=\; \begin{cases} 1 & \text{if } \mu = \mu_o, \\ 1-\alpha & \text{for any } \mu \neq \mu_o. \end{cases}$$

(b) Show that

$$b_\alpha(\boldsymbol{X}) \;:=\; |\bar{X} - \mu_o| + \tau\Phi^{-1}(1-\alpha)$$

is an upper $(1-\alpha)$-confidence bound for $|\mu-\mu_o|$, i.e.

$$\mathbb{P}_\mu\big(b_\alpha(X) \geq |\mu-\mu_o|\big) \;\geq\; 1-\alpha \quad \text{for all } \mu \in \mathbb{R}.$$

2.6 Stochastic Order, P-Values, Confidence Bounds

Practitioners use mostly non-randomized tests and confidence regions. Nevertheless the results in the previous sections show that certain non-randomized tests are close to optimal. In the present section we describe a more pragmatic approach to non-randomized tests and confidence regions via so-called p-values. These methods require weaker conditions on our statistical model $(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$.

The setting. We consider a statistical experiment $(\mathcal{X}, \mathcal{B}, (P_\theta)_\theta)$ and a given function $g : \Theta \to \mathbb{R}$. Observing $X \sim P_\theta$ for some unknown $\theta \in \Theta$, our goal is to construct a $(1-\alpha)$-confidence region for $g(\theta)$.

We assume that there exists a statistic $T : \mathcal{X} \to \mathbb{R}$ such that its distribution P_θ^T depends only on $g(\theta)$. That is, there exists a family $(Q_\gamma)_{\gamma\in\Gamma}$ of probability distributions on the real line such that $\Gamma = g(\Theta)$, and for any parameter $\theta \in \Theta$,

$$P_\theta(T \in B) \;=\; Q_{g(\theta)}(B), \quad B \in \mathrm{Borel}(\mathbb{R}).$$

In many examples, $\Theta \subset \mathbb{R}$ and $g(\theta) = \theta$.

P-Values. For any hypothetical value γ of $g(\theta)$, we consider the *left-sided p-value* $\pi_\gamma^{\mathrm{left}}(T(X))$, where

$$\pi_\gamma^{\mathrm{left}}(t) \;:=\; Q_\gamma((-\infty, t]),$$

and the *right-sided p-value* $\pi_\gamma^{\mathrm{right}}(T(X))$, where

$$\pi_\gamma^{\mathrm{right}}(t) \;:=\; Q_\gamma([t, \infty)).$$

Both p-values quantify the plausibility of the null hypothesis that $g(\theta) = \gamma$. Small values of $\pi_\gamma^{\mathrm{left}}(T(X))$ or $\pi_\gamma^{\mathrm{right}}(T(X))$ indicate that $T(X)$ is "suspiciously small" or "suspiciously large", respectively, for the hypothetical value γ of $g(\theta)$. One can combine these one-sided p-values to the *two-sided p-value* $\pi_\gamma^{\mathrm{two}}(T(X))$, where

$$\pi_\gamma^{\mathrm{two}} \;:=\; 2\min(\pi_\gamma^{\mathrm{left}}, \pi_\gamma^{\mathrm{right}}).$$

Small values of $\pi_\gamma^{\mathrm{two}}(T(X))$ indicated that the value $T(X)$ is "suspiciously extreme" for the hypothetical value γ of $g(\theta)$.

For any given test level $\alpha \in (0,1)$ and $\theta \in \Theta$,

$$\mathbb{P}_\theta\big(\pi_{g(\theta)}(T(X)) \leq \alpha\big) \;\leq\; \alpha, \tag{2.9}$$

where π_γ stands for π_γ^{left}, $\pi_\gamma^{\text{right}}$ or π_γ^{two}. Equality holds if the distribution $Q_{g(\theta)}$ is continuous. This is well-known from introductory courses on mathematical statistics and follows from elementary considerations. Hence, a level-α test of the null hypothesis that $g(\theta) = \gamma$ is given by

$$\varphi_{\alpha,\gamma}(x) \;:=\; 1_{[\pi_\gamma(T(x)) \le \alpha]},$$

and a $(1-\alpha)$-confidence region for $g(\theta)$ is given by

$$\mathcal{C}_\alpha(x) \;:=\; \big\{\gamma \in \Gamma : \pi_\gamma(T(x)) > \alpha\big\}.$$

Stochastically ordered distributions. There is no guarantee that the particular tests $\varphi_{\alpha,\gamma}$ or confidence regions $\mathcal{C}_\alpha$ above are useful at all. They are useful if, for instance, the family $(Q_\gamma)_{\gamma\in\Gamma}$ satisfies the following conditions which are all equivalent:

(SO.1) For any fixed $t \in \mathbb{R}$, $Q_\gamma((-\infty, t])$ is non-increasing in $\gamma \in \Gamma$.

(SO.2) For any fixed $t \in \mathbb{R}$, $Q_\gamma([t, \infty))$ is non-decreasing in $\gamma \in \Gamma$.

(SO.3) For any fixed $u \in (0,1)$, $\min\big\{t \in \mathbb{R} : Q_\gamma((-\infty, t]) \ge u\big\}$ is non-decreasing in $\gamma \in \Gamma$.

(SO.4) For any non-decreasing function $h : \mathbb{R} \to [0, \infty)$, $\int h \,\mathrm{d}Q_\gamma$ is non-decreasing in $\gamma \in \Gamma$.

If (SO.1-4) are satisfied, the distributions Q_γ are *stochastically increasing* in $\gamma \in \Gamma$.

Exercise 2.33. Show that Conditions (SO.1-4) are equivalent.

Lemma 2.14 shows that any statistical model with monotone density ratios satisfies condition (SO.4) and hence (SO.1-4), where $g(\theta) = \theta$. Note also that in Example 2.11 one may extend the parameter space to $\Theta = [0, 1]$, and the stochastic order constraint remains valid.

Confidence bounds. If the family $(Q_\gamma)_{\gamma\in\Gamma}$ satisfies (SO.1–4), the confidence regions $\mathcal{C}_\alpha$ correspond to confidence bounds or confidence intervals. Precisely, if $\pi_\gamma = \pi_\gamma^{\text{left}}$, then

$$\mathcal{C}_\alpha(x) \;=\; \mathcal{C}_\alpha^{\text{left}}(x) \;=\; \big\{\gamma \in g(\Theta) : Q_\gamma((-\infty, T(x)]) > \alpha\big\}$$

is of the form $\{\gamma \in \Gamma : \gamma < b_\alpha(x)\}$ or $\{\gamma \in \Gamma : \gamma \le b_\alpha(x)\}$ for some $b_\alpha(x) \in \big[\inf(\Gamma), \sup(\Gamma)\big]$. In particular, b_α is an upper $(1-\alpha)$-confidence bound for $g(\theta)$ in the sense that

$$\mathbb{P}_\theta\big(b_\alpha(X) \ge g(\theta)\big) \;\ge\; 1-\alpha \quad \text{for all } \theta \in \Theta.$$

Similarly, if $\pi_\gamma = \pi_\gamma^{\text{right}}$, then

$$\mathcal{C}_\alpha(x) \;=\; \mathcal{C}_\alpha^{\text{right}}(x) \;=\; \big\{\gamma \in g(\Theta) : Q_\gamma([T(x), \infty)) > \alpha\big\}$$

is of the form $\{\gamma \in \Gamma : \gamma > a_\alpha(x)\}$ or $\{\gamma \in \Gamma : \gamma \ge a_\alpha(x)\}$ for some $a_\alpha(x) \in \big[\inf(\Gamma), \sup(\Gamma)\big]$. This yields a lower $(1-\alpha)$-confidence bound a_α for $g(\theta)$ in the sense that

$$\mathbb{P}_\theta\big(a_\alpha(X) \le g(\theta)\big) \;\ge\; 1-\alpha \quad \text{for all } \theta \in \Theta.$$

Finally, if $\pi_\gamma = \pi_\gamma^{\text{two}}$, then $\mathcal{C}_\alpha(x) = \mathcal{C}_{\alpha/2}^{\text{left}}(x) \cap \mathcal{C}_{\alpha/2}^{\text{right}}(x)$ is an interval in Γ with boundaries $a_{\alpha/2}(x)$ and $b_{\alpha/2}(x)$, so

$$\mathbb{P}_\theta\big([a_{\alpha/2}(X), b_{\alpha/2}(X)] \ni g(\theta)\big) \;\ge\; 1-\alpha \quad \text{for all } \theta \in \Theta.$$

Example 2.11 (Bernoulli sequences, cont.) For any $\theta_o \in [0,1]$ the left- and right-sided p-values are given by

$$F_{n,\theta_o}(T(X)) \quad \text{and} \quad 1 - F_{n,\theta_o}(T(X)-1),$$

respectively. The resultung confidence intervals are

$$\mathcal{C}_\alpha^{\text{left}}(x) \;=\; \begin{cases} [0, b_\alpha(x)) & \text{if } T(x) < n, \\ [0,1] & \text{if } T(x) = n, \end{cases}$$

$$\mathcal{C}_\alpha^{\text{right}}(x) \;=\; \begin{cases} [0,1] & \text{if } T(x) = 0, \\ (a_\alpha(x), 1] & \text{if } T(x) > 0, \end{cases}$$

where $b_\alpha(x)$ is the unique $p \in (0,1)$ such that $F_{n,p}(T(x)) = \alpha$ (if $T(x) < n$) while $a_\alpha(x)$ is the unique $p \in (0,1)$ such that $F_{n,p}(T(x)-1) = 1-\alpha$ (if $T(x) > 0$). (Recall that for any integer $t \in \{0,1,\ldots,n-1\}$, the function $p \mapsto F_{n,p}(t)$ is continuous and strictly decreasing with boundary values $F_{n,0}(t) = 1$ and $F_{n,1}(t) = 0$.)

Example 2.34 (Ratio of two standard deviations). Suppose we observe independent random variables $X_{ij} \sim \mathrm{N}(\mu_i, \sigma_i^2)$, where $i = 1,2$ and $j = 1,\ldots,n_i$ with $n_1, n_2 \geq 2$. The parameter $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2) \in \mathbb{R} \times (0,\infty) \times \mathbb{R} \times (0,\infty)$ is unknown. Suppose we are mainly interested in the ratio $g(\theta) = \sigma_1/\sigma_2$ of the two standard deviations σ_1 and σ_2. With the sample means $\bar{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ and sample variances

$$S_i^2 \;:=\; \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

it is well-known that S_1^2 and S_2^2 are stochastically independent, where $(n_i - 1)S_i^2/\sigma_i^2$ follows a chi-squared distribution with $n_i - 1$ degrees of freedom. Denoting the tuple of all $n_1 + n_2$ observations with $\boldsymbol{X}$, a plausible estimator of $g(\theta)$ is $T(\boldsymbol{X}) := S_1/S_2$, and $T(\boldsymbol{X})^2/g(\theta)^2$ follows Fisher's F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The latter distribution has a continuous distribution function $F = F_{n_1-1,n_2-1}$, so for $t \geq 0$,

$$\mathbb{P}_\theta(T(\boldsymbol{X}) \leq t) \;=\; F(t^2/g(\theta)^2) \;=\; 1 - \mathbb{P}_\theta(T(\boldsymbol{X}) \geq t).$$

This leads to the left-sided p-values $\pi_\gamma^{\text{left}}(t) = F(t^2/\gamma^2)$. Solving the equation $F(T(\boldsymbol{X})^2/\gamma^2) = \alpha$ for γ yields the upper $(1-\alpha)$-confidence bound

$$b_\alpha(\boldsymbol{X}) \;=\; \frac{S_1/S_2}{\sqrt{F^{-1}(\alpha)}}$$

for σ_1/σ_2. Analogously, the right-sided p-values are given by $\pi_\gamma^{\text{right}}(t) = 1 - F(t^2/\gamma^2)$ and lead to the lower $(1-\alpha)$-confidence bound

$$a_\alpha(\boldsymbol{X}) \;=\; \frac{S_1/S_2}{\sqrt{F^{-1}(1-\alpha)}}$$

for σ_1/σ_2.

Exercise 2.35. This is a preparation for the next exercise. Let $X \sim \mathrm{Bin}(n,p)$ with unknown $p \in [0,1]$. Show that for arbitrary integers $0 \leq c_1 \leq c_2 \leq n$ with $c_2 - c_1 < n$, the log-probability $\log \mathbb{P}_p(c_1 \leq X \leq c_2) \in [-\infty, 0]$ is continuous and strictly concave in $p \in [0,1]$.

Exercise 2.36 (Confidence bounds for $|p-1/2|$ in binomial model). Let $X \sim \mathrm{Bin}(n,p)$ with unknown $p \in [0,1]$ and given $n \geq 2$. With $m := \lfloor n/2 \rfloor$ let $T(x) := \min(x, n-x) \in \{0,1,\ldots,m\}$.

(a) Show that the distribution of $T(X)$ depends only on $g(p) := \min(p, 1-p) \in \Gamma = [0,1/2]$. With Q_γ denoting this distribution in the case of $g(p) = \gamma$, determine an explicit formula for $Q_\gamma([t,\infty))$, $t \in \{0,1,\ldots,m\}$.

(b) Show by means of Exercise 2.35 that $Q_\gamma([t,\infty))$ is strictly monotone increasing in γ. Use this fact to define an upper $(1-\alpha)$-confidence bound for $|p-1/2|$.

(c) Implement this confidence bound as a computer program and illustrate the results for some $n \geq 20$ and $\alpha = 5\%$.

Example 2.12 (Gaussian location family, cont.) We consider $P_\theta = \mathrm{N}(\theta, \sigma^2)^{\otimes n}$ with unknown $\theta \in \mathbb{R}$ and given $\sigma > 0$. With $T(\boldsymbol{x}) = \bar{x}$ and $\tau = \sigma/\sqrt{n}$, the distribution of T under P_θ is $Q_\theta = \mathrm{N}(\theta, \tau^2)$. It has continuous distribution function $\Phi\big((\cdot - \theta)/\tau\big)$, so the left- and right-sided p-values for any given parameter θ_o are given by

$$\Phi\big((\bar{x} - \theta_o)/\tau\big) \quad \text{and} \quad \Phi\big((\theta_o - \bar{x})/\tau\big),$$

respectively. The resulting confidence intervals are

$$\begin{aligned} \mathcal{C}_\alpha^{(\ell)}(\boldsymbol{x}) &= (-\infty, b_\alpha(\boldsymbol{x})) && \text{with} \quad b_\alpha(\boldsymbol{x}) := \bar{x} + \Phi^{-1}(1-\alpha)\tau, \\ \mathcal{C}_\alpha^{(r)}(\boldsymbol{x}) &= (a_\alpha(\boldsymbol{x}), \infty) && \text{with} \quad a_\alpha(\boldsymbol{x}) := \bar{x} - \Phi^{-1}(1-\alpha)\tau. \end{aligned}$$

Exercise 2.37. Find probability distributions P_0 and P_1 with finite support $\mathcal{X} \subset \mathbb{R}$ and distribution functions F_0 and F_1, respectively, such that $F_0 \leq_{\mathrm{st.}} F_1$ (i.e. $F_0 \geq F_1$) but $g(x) := P_1(\{x\})/P_0(\{x\})$ is not monotone in $x \in \mathcal{X}$.

Example 2.38 (Capture-recapture). The unknown size N of a population of animals is sometimes estimated with a capture-recapture experiment: At first, a random sample of size n_1 is drawn from the population without replacement, and all animals in this catch are marked and then released. After some time a second sample of size n_2 is drawn without replacement, and one determines the number X of marked animals in this second catch. That means, X is the number of animals which were catched twice. Ideally, X is a random variable with distribution $\mathrm{Hyp}(N, n_1, n_2)$. This leads to the statistical experiment

$$\big(\mathcal{X}, \mathcal{P}(\mathcal{X}), \big(\mathrm{Hyp}(N, n_1, n_2)\big)_{N \geq \max(n_1, n_2)}\big),$$

where $\mathcal{X} := \{0,1,\ldots,\min(n_1,n_2)\}$ and

$$\mathrm{Hyp}(N, n_1, n_2)(\{x\}) = \binom{n_1}{x}\binom{N-n_1}{n_2-x}\Big/\binom{N}{n_2} = \binom{n_2}{x}\binom{N-n_2}{n_1-x}\Big/\binom{N}{n_1}$$

for $x \in \mathcal{X}$ with the convention that $\binom{k}{\ell} := 0$ if $\ell > k$. Possible point estimators for N are given by

$$\widehat{N}(x) := \frac{n_1 n_2}{x} \quad \text{or} \quad \widehat{N}(x) := \frac{(n_1+1)(n_2+1)}{x+1}.$$

It follows from Exercise 2.39 below that the distributions $\mathrm{Hyp}(N, n_1, n_2)$ are stochastically *decreasing* in N. That means, if F_N denotes the distribution function of $\mathrm{Hyp}(N, n_1, n_2)$, then

$$F_N(x) \text{ is non-decreasing in } N \geq \max(n_1, n_2)$$

for any $x \in \mathcal{X}$, and

$$\lim_{N\to\infty} F_N(0) \;=\; 1.$$

Consequently, a lower $(1-\alpha)$-confidence bound for N is given by

$$a_\alpha(x) \;:=\; \min\bigl\{N \geq \max(n_1, n_2) : F_N(x) > \alpha\bigr\}$$

while an upper bound is given by

$$b_\alpha(x) \;=\; \begin{cases} \infty & \text{if } x = 0, \\ \max\bigl\{N \geq \max(n_1, n_2) : F_N(x-1) < 1-\alpha\bigr\} & \text{if } x > 0. \end{cases}$$

Exercise 2.39. For integers $n_1, n_2 \geq 1$ and $N \geq \max(n_1, n_2)$, let F_N be the distribution function of the hypergeometric distribution $\mathrm{Hyp}(N, n_1, n_2)$.

(a) Show that $F_N(x) \leq F_{N+1}(x)$ for arbitrary $x \in \{0, 1, \ldots, \min(n_1, n_2)\}$. Proposal: Think of an urn with n_1 black, $N - n_1$ white and one red ball from which you draw $n_2 + 1$ balls one by one without replacement.

(b) Show that

$$\lim_{N\to\infty} F_N(0) \;=\; 1.$$

Exercise 2.40. Let X be a random variable with distribution $\mathrm{Hyp}(N, n_1, n_2)$ with given parameters $n_1, n_2 \in \mathbb{N}$ and an unknown parameter $N \geq \max(n_1, n_2)$. Determine $E_N(\widehat{N})$ for the point estimator $\widehat{N} := (n_1 + 1)(n_2 + 1)/(X + 1)$.

Chapter 3

Decision Problems and Procedures, Sufficiency and Completeness

In the present chapter we introduce some fundamental concepts and results from statistical decision theory. Estimation and testing problems may be viewed as special cases of *decision problems*, while point estimators and statistical tests are corresponding *decision procedures*.

3.1 Decision Problems and Procedures

As in the previous chapters, we consider a statistical experiment

$$\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big).$$

If there is no doubt about the sample space $(\mathcal{X}, \mathcal{B})$, we just write $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$.

Decision spaces. A *decision space* is a measurable space $(\mathcal{V}, \mathcal{C})$ representing the possible conclusions we could draw about the unknown true parameter θ.

Loss functions. A *loss function* is a mapping

$$L : \mathcal{V} \times \Theta \to [0, \infty]$$

such that $L(\cdot, \theta)$ is $\mathcal{C}$-measurable for any fixed $\theta \in \Theta$. Here $L(v, \theta)$ quantifies the loss (e.g. the costs) when drawing the conclusion $v \in \mathcal{V}$ while the true parameter equals $\theta \in \Theta$.

Some people use loss functions which can take negative values, but for our purposes it suffices to consider non-negative functions L.

Example 3.1 (Point estimation). Consider a given mapping $g : \Theta \to \mathbb{R}^q$. For instance, for the statistical experiment $\mathcal{E} = \big(\mathcal{N}(\mu, \sigma^2)^{\otimes n}\big)_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)}$, one could think about $g(\mu, \sigma) := \mu$ or $g(\mu, \sigma) := \sigma$. Our decision could be a guess $v \in \mathbb{R}^q$ for the unknown true value $g(\theta)$. A potential loss function would be given by

$$L(v, \theta) \;:=\; \|v - g(\theta)\|^r$$

for some norm $\|\cdot\|$ on $\mathbb{R}^q$ and some exponent $r > 0$. In case of $r \geq 1$, this loss function is a special case of the more general loss function L given by

$$L(v, \theta) \ := \ \psi(v - g(\theta))$$

for some convex function $\psi : \mathbb{R}^q \to [0, \infty)$ such that $\psi(v) \to \infty$ as $\|v\| \to \infty$.

Example 3.2 (Statistical tests). Suppose we have split Θ into two disjoint sets Θ_0 and Θ_1. The question is whether the unknown true parameter θ belongs to Θ_0 or Θ_1. Hence the two potential decisions would be 0 and 1, respresenting the claims that $\theta \in \Theta_0$ or $\theta \in \Theta_1$, respectively. A potential loss function would be the indicator of a wrong conclusion, i.e.

$$L(v, \theta) \ := \ 1_{[v=1, \theta \in \Theta_0]} + 1_{[v=0, \theta \in \Theta_1]}.$$

More generally, one could specify costs $\lambda_j > 0$ for an error of the j-th kind and set

$$L(v, \theta) \ := \ 1_{[v=1, \theta \in \Theta_0]} \lambda_1 + 1_{[v=0, \theta \in \Theta_1]} \lambda_2. \tag{3.1}$$

Decision problems. A triplet $(\mathcal{E}, (\mathcal{V}, \mathcal{C}), L)$, consisting of a statistical experiment, a decision space and a loss function is called a *decision problem*. If there is a standard σ-field $\mathcal{C}$ on $\mathcal{V}$, we just write $(\mathcal{E}, \mathcal{V}, L)$.

Decision procedures. A *non-randomized decision procedure* is a measurable mapping $\rho : \mathcal{X} \to \mathcal{V}$. That means, if we observe $X \sim P_\theta$ with unknown $\theta \in \Theta$, then we draw the conclusion $\rho(X) \in \mathcal{V}$ about θ.

More generally, a *decision procedure* is a stochastic kernel ρ from $(\mathcal{X}, \mathcal{B})$ to $(\mathcal{V}, \mathcal{C})$. That means,

$$\rho : \mathcal{X} \times \mathcal{C} \to [0, 1]$$

is a mapping such that

$$\begin{aligned} &\text{for any } x \in \mathcal{X}, \quad \rho(x, \cdot) \text{ is a probability measure on } (\mathcal{V}, \mathcal{C}), \\ &\text{for any } C \in \mathcal{C}, \quad \rho(\cdot, C) \text{ is } \mathcal{B}\text{-measurable on } \mathcal{X}. \end{aligned}$$

Now the interpretation is that having observed $X \sim P_\theta$, we draw a random conclusion about θ from the probability measure $\rho(X, \cdot)$.

With slightly ambiguous notation, a non-randomized decision procedure $\rho : \Omega \to \mathcal{V}$ corresponds to the stochastic kernel $\rho(\cdot, \cdot)$ with

$$\rho(\omega, \cdot) \ := \ \delta_{\rho(\omega)}.$$

Risk functions. The performance of a decision procedure ρ is quantified by its risk function $R(\rho, \cdot) : \Theta \to [0, \infty]$,

$$R(\rho, \theta) \ := \ \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \theta) \, \rho(x, \mathrm{d}v) \, P_\theta(\mathrm{d}x).$$

Example 3.1 (Point estimation, continued). A point estimator for $g(\theta)$ is a measurable mapping $\widehat{g} : \mathcal{X} \to \mathbb{R}^q$. This corresponds to the (non-randomized) decision procedure

$$\rho(x, \cdot) \ := \ \delta_{\widehat{g}(x)}.$$

Its risk function is given by

$$R(\widehat{g}, \theta) \ = \ \int_{\mathcal{X}} \|\widehat{g} - g(\theta)\|^r \, \mathrm{d}P_\theta \ = \ E_\theta\big(\|\widehat{g} - g(\theta)\|^r\big)$$

or

$$R(\widehat{g}, \theta) \ = \ \int_{\mathcal{X}} \psi(\widehat{g} - g(\theta)) \, \mathrm{d}P_\theta \ = \ E_\theta \psi(\widehat{g} - g(\theta)).$$

Remark 3.3 (De-randomisation for point estimation). In the context of point estimation, it is often sufficient to consider non-randomized decision procedures, i.e. usual point estimators. Specifically, suppose that $L(\cdot, \theta)$ is convex and coercive[1] on $\mathbb{R}^q$ for any $\theta \in \Theta$. Then for any decision procedure ρ, i.e. a stochastic kernel from $(\mathcal{X}, \mathcal{B})$ to $(\mathbb{R}^q, \mathrm{Borel}(\mathbb{R}^q))$, there exists a point estimator $\widehat{g} : \mathcal{X} \to \mathbb{R}^q$ such that for any $\theta \in \Theta$,

$$R(\widehat{g}, \theta) \ = \ \int_{\mathcal{X}} L(\widehat{g}(x), \theta) \, P_\theta(\mathrm{d}x)$$

is no larger than

$$R(\rho, \theta) \ = \ \int_{\mathcal{X}} \int_{\mathbb{R}^q} L(v, \theta) \, \rho(x, \mathrm{d}v) \, P_\theta(\mathrm{d}x).$$

Indeed, the assumption that $L(\cdot, \theta)$ is coercive implies that for suitable constants $a(\theta) \in \mathbb{R}$ and $b(\theta) > 0$,

$$L(v, \theta) \ \geq \ a(\theta) + b(\theta)\|v\| \quad \text{for all } v \in \mathbb{R}^q.$$

In particular, if $x \in \mathcal{X}$ and $\theta \in \Theta$ satisfy

$$\int_{\mathbb{R}^q} L(v, \theta) \, \rho(x, \mathrm{d}v) \ < \ \infty,$$

then

$$\int_{\mathbb{R}^q} \|v\| \, \rho(x, \mathrm{d}v) \ \leq \ -b(\theta)^{-1} a(\theta) + b(\theta)^{-1} \int_{\mathbb{R}^q} \|v\| \, \rho(x, \mathrm{d}v) \ < \ \infty.$$

Consequently, we may pick an arbitrary fixed point $v_o \in \mathbb{R}^q$ and define

$$\widehat{g}(x) \ := \ \int_{\mathbb{R}^q} v \, \rho(x, \mathrm{d}v) \quad \text{if} \ \int_{\mathbb{R}^q} \|v\| \, \rho(x, \mathrm{d}v) \ < \ \infty,$$

and $\widehat{g}(x) := v_o$ otherwise. Then, Jensen's inequality implies that for arbitrary $x \in \mathcal{X}$ and $\theta \in \Theta$,

$$\int_{\mathbb{R}^q} L(v, \theta) \, \rho(x, \mathrm{d}v) \ \geq \ L(\widehat{g}(x), \theta),$$

whence

$$R(\rho, \theta) \ \geq \ R(\widehat{g}, \theta).$$

[1] $L(v, \theta) \to \infty$ as $\|v\| \to \infty$

Example 3.2 (Statistical tests, continued). With the decision space $\mathcal{V} = \{0, 1\}$, any decision procedure ρ may be written as

$$\rho(x, \cdot) \;=\; (1 - \varphi(x))\delta_0 + \varphi(x)\delta_1$$

for some measurable function $\varphi : \mathcal{X} \to [0, 1]$, i.e. a test on $\mathcal{X}$. With the power function $\theta \mapsto E_\theta(\varphi)$ of φ and the loss function L in (3.1),

$$R(\varphi, \theta) \;=\; 1_{[\theta\in\Theta_0]}\lambda_1 E_\theta(\varphi) + 1_{[\theta\in\Theta_1]}\lambda_2(1 - E_\theta(\varphi)).$$

Bayes risks. Suppose that Θ itself is equipped with a σ-field $\mathcal{D}$, and suppose that the loss function $L : \mathcal{V} \times \Theta \to [0, \infty]$ is $\mathcal{C} \otimes \mathcal{D}$-measurable. Further suppose that $\theta \mapsto P_\theta(B)$ is $\mathcal{D}$-measurable for any fixed $B \in \mathcal{B}$. Then we may view $(P_\theta)_{\theta\in\Theta}$ as a stochastic kernel too and consider the following Bayesian model: Let Π be a probability distribution on $(\Theta, \mathcal{D})$, a so-called *prior (distribution)*. One could imagine that "mother nature" chooses a parameter $\theta \in \Theta$ from that distribution Π. Then, conditional on θ, we observe a random variable $X \sim P_\theta$. The Bayes risk of a decision procedure ρ for the prior Π is defined as

$$R(\rho, \Pi) \;:=\; \int_\Theta R(\rho, \theta)\, \Pi(\mathrm{d}\theta) \;=\; \int_\Theta \int_\mathcal{X} \int_\mathcal{V} L(v, \theta)\, \rho(x, \mathrm{d}v)\, P_\theta(\mathrm{d}x)\, \Pi(\mathrm{d}\theta).$$

With $\mathbb{P}^{\mathrm{B}}$ and $\mathbb{E}^{\mathrm{B}}$ denoting probabilities and expectations in this Bayesian model, one can also write

$$R(\rho, \Pi) \;=\; \mathbb{E}^{\mathrm{B}} \int_\mathcal{V} L(v, \theta)\, \rho(X, \mathrm{d}v),$$

and in case of a non-randomized procedure ρ this becomes

$$R(\rho, \Pi) \;=\; \mathbb{E}^{\mathrm{B}}\, L(\rho(X), \theta).$$

3.2 Some Optimality Concepts and Results

Let $(\mathcal{E}, (\mathcal{V}, \mathcal{C}), L)$ be a given decision problem. Our goal is to devise decision procedures ρ with low risks $R(\rho, \theta)$. Typically there is no "free lunch": If ρ has very small risk $R(\rho, \theta_1)$ for some parameter $\theta_1 \in \Theta$, it will often have rather large risk $R(\rho, \theta_2)$ for some other parameter $\theta_2 \in \Theta$.

Minimax-optimality. A decision procedure ρ_* is called *minimax-optimal*, if

$$\sup_{\theta\in\Theta} R(\rho_*, \theta) \;=\; \min_\rho \sup_{\theta\in\Theta} R(\rho, \theta).$$

Throughout this chapter, "$\min_\rho$" or "$\inf_\rho$" stand for the minimum or infimum over all decision procedures ρ.

Admissibility. A decision procedure ρ_* is called *admissible*, if there exists no decision procedure ρ satisfying

$$R(\rho, \theta) \;\le\; R(\rho_*, \theta) \quad \text{for all } \theta \in \Theta$$

and

$$R(\rho, \theta_o) \;<\; R(\rho_*, \theta_o) \quad \text{for at least one } \theta_o \in \Theta.$$

Bayes-optimality. For a given prior Π on Θ, a decision procedure ρ_* is called *Bayes-optimal* for this prior Π, if

$$R(\rho_*, \Pi) \;=\; \min_{\rho} R(\rho, \Pi).$$

Least favourable priors. A prior Π_* is called least favourable, if

$$\inf_{\rho} R(\rho, \Pi_*) \;=\; \max_{\Pi} \inf_{\rho} R(\rho, \Pi),$$

where "$\max_\Pi$" stands for the maximum over all priors Π on Θ.

Here are three simple results establishing minimaxity, Bayes-optimality and admissibility of decision procedures.

Lemma 3.4. *Let Π_* be a prior on Θ, and let ρ_* be a Bayes-optimal decision procedure for Π_*. Suppose further that*

$$R(\rho_*, \Pi_*) \;=\; \sup_{\theta\in\Theta} R(\rho_*, \theta).$$

Then ρ_* is minimax-optimal, and Π_* is a least favourable prior.

Proof of Lemma 3.4. For any decision procedure ρ,

$$\sup_{\theta\in\Theta} R(\rho, \theta) \;\geq\; R(\rho, \Pi_*),$$

and by our assumptions on ρ_*,

$$R(\rho, \Pi_*) \;\geq\; R(\rho_*, \Pi_*) \;=\; \sup_{\theta\in\Theta} R(\rho_*, \theta).$$

Hence ρ_* is minimax-optimal.

For any prior Π on Θ,

$$\inf_{\rho} R(\rho, \Pi) \;\leq\; R(\rho_*, \Pi) \;\leq\; \sup_{\theta\in\Theta} R(\rho_*, \theta) \;=\; R(\rho_*, \Pi_*) \;=\; \inf_{\rho} R(\rho, \Pi_*)$$

by assumption. Hence Π_* is least favourable. □

Lemma 3.5. *Let ρ_* be a decision procedure such that for a sequence $(\Pi_k)_{k\geq 1}$ of priors,*

$$\sup_{\theta\in\Theta} R(\rho_*, \theta) \;=\; \lim_{k\to\infty} \inf_{\rho} R(\rho, \Pi_k).$$

Then ρ_* is minimax-optimal.

Proof of Lemma 3.5. For any decision procedure ρ_o,

$$\sup_{\theta\in\Theta} R(\rho_o, \theta) \;\geq\; \limsup_{k\to\infty} R(\rho_o, \Pi_k) \;\geq\; \lim_{k\to\infty} \inf_{\rho} R(\rho, \Pi_k) \;=\; \sup_{\theta\in\Theta} R(\rho_*, \theta).$$

□

Lemma 3.6. *Let Θ be a topological space equipped with its Borel-σ-field. Suppose that our decision problem is such that any real-valued risk function is automatically continuous. If Π is a prior on Θ such that $\Pi(U) > 0$ for any non-void open set $U \subset \Theta$, and if ρ_* is a Bayes-optimal decision procedure for Π such that $R(\rho_*, \Pi) < \infty$ and $R(\rho, \cdot)$ is real-valued, then ρ_* is admissible.*

Exercise 3.7. Let Θ be a metric space. Suppose that each distribution P_θ has a density f_θ with respect to some σ-finite measure M on $(\mathcal{X}, \mathcal{B})$ such that $f_\theta(x)$ is continuous in $\theta \in \Theta$ for any fixed $x \in \mathcal{X}$. Further, suppose that the loss function L is bounded, and let $L(v, \theta)$ be continuous in $\theta \in \Theta$ for any fixed $v \in \mathcal{V}$. Show that each decision procedure has bounded and continuous risk function. *Hint:* Use Scheffé's theorem (Theorem A.4) and dominated convergence.

Proof of Lemma 3.6. Suppose that ρ_* is not admissible. That means, there exists a decision procedure ρ such that $R(\rho, \cdot) \leq R(\rho_*, \cdot) < \infty$ and $R(\rho, \theta_*) < R(\rho_*, \theta_o)$ for some $\theta_o \in \Theta$. Since both risk functions $R(\rho, \cdot)$ and $R(\rho_*, \cdot)$ are continuous, there exist an open set $U \subset \Theta$ and an $\epsilon > 0$ such that $R(\rho, \cdot) \leq R(\rho_*, \cdot) - \epsilon$ on U. But then

$$\begin{aligned} R(\rho, \Pi) &= \int_{\Theta \setminus U} R(\rho, \theta)\, \Pi(d\theta) + \int_U R(\rho, \theta)\, \Pi(d\theta) \\ &\leq \int_{\Theta \setminus U} R(\rho_*, \theta)\, \Pi(d\theta) + \int_U (R(\rho_*, \theta) - \epsilon)\, \Pi(d\theta) \\ &= R(\rho_*, \Pi) - \epsilon \Pi(U) \\ &< R(\rho_*, \Pi), \end{aligned}$$

a contradiction to Bayes-optimality of ρ_*. □

Finding Bayes-optimal decision procedures. In the Bayesian model, we have a random pair $(\theta, X) \in \Theta \times \mathcal{X}$, and the key to finding Bayes-optimal decision procedures is to find the *posterior (distribution) of θ, given the observation X*, where "posterior (distribution)" is synonymous to "conditional distribution". Let us explain this in the special setting of distributions P_θ having densities f_θ with respect to some measure M on $(\mathcal{X}, \mathcal{B})$, where $(\theta, x) \mapsto f_\theta(x)$ is measurable on $\Theta \times \mathcal{X}$. Further suppose that Π has a density π with respect to some measure N on $(\Theta, \mathcal{D})$. Then for arbitrary $D \in \mathcal{D}$ and $B \in \mathcal{B}$,

$$\mathbb{P}^{\mathrm{B}}(\theta \in D, X \in B) = \int_D \int_B f_\theta(x)\, M(\mathrm{d}x)\, \pi(\theta)\, N(\mathrm{d}\theta) = \int_B \int_D f_\theta(x) \pi(\theta)\, N(\mathrm{d}\theta)\, M(\mathrm{d}x).$$

Setting $D = \Theta$, we see that

$$\mathbb{P}^{\mathrm{B}}(X \in B) = \int_B f(x)\, M(\mathrm{d}x)$$

with

$$f(x) := \int_\Theta f_\theta(x) \pi(\theta)\, N(\mathrm{d}\theta) \in [0, \infty].$$

In particular, $M(f = \infty) = 0$. If we define

$$\pi(\theta \,|\, x) := \begin{cases} f(x)^{-1} f_\theta(x) \pi(\theta) & \text{if } 0 < f(x) < \infty, \\ \pi(\theta) & \text{else,} \end{cases}$$

then

$$\mathbb{P}^{\mathrm{B}}(\theta \in D, X \in B) = \int_B f(x) \int_D \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)\, M(\mathrm{d}x).$$

With standard tools from measure theory one can deduce that for any measurable function $h : \Theta \times \mathcal{X} \to [0, \infty]$,

$$\mathbb{E}^{\mathrm{B}}\, h(\theta, X) \;=\; \int_\Theta \int_{\mathcal{X}} h(\theta, x)\, P_\theta(\mathrm{d}x)\, \Pi(\mathrm{d}\theta) \;=\; \int_{\mathcal{X}} f(x) \int_\Theta h(\theta, x) \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)\, M(\mathrm{d}x).$$

For any fixed $x \in \mathcal{X}$, the function $\pi(\cdot \,|\, x)$ is a probability density with respect to N. We interpret the corresponding distribution as the *conditional distribution of θ, given that $X = x$*. In that sense,

$$\int_\Theta h(\theta, x) \pi(\theta \,|\, x)\, N(\mathrm{d}\theta) \;=\; \mathbb{E}^{\mathrm{B}}(h(\theta, X) \,|\, X = x).$$

Coming back to a specific decision space $(\mathcal{V}, \mathcal{C})$ and a measurable loss function $L : \mathcal{V} \times \Theta \to [0, \infty]$, let ρ be an arbitrary decision procedure. Then,

$$\begin{aligned} R(\rho, \Pi) \;&=\; \int_\Theta \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \theta)\, \rho(x, \mathrm{d}v)\, P_\theta(\mathrm{d}x)\, \Pi(\mathrm{d}\theta) \\ &=\; \int_{\mathcal{X}} f(x) \int_{\mathcal{V}} \int_\Theta L(v, \theta) \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)\, \rho(x, \mathrm{d}v)\, M(\mathrm{d}x). \end{aligned}$$

If $\rho_* : \mathcal{X} \to \mathcal{V}$ is measurable such that

$$\rho_*(x) \;\in\; \operatorname*{arg\,min}_{v \in \mathcal{V}} \int_\Theta L(v, \theta) \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)$$

for all $x \in \mathcal{X}$, then ρ_* is a Bayes-optimal, non-randomized decision procedure for the prior distribution Π. With conditional expectations one can also write

$$\rho_*(x) \;\in\; \operatorname*{arg\,min}_{v \in \mathcal{V}} \mathbb{E}^{\mathrm{B}}(L(v, \theta) \,|\, X = x).$$

Let us apply this in the context of point estimation with a measurable function $g : \Theta \to \mathbb{R}^q$, where $L(v, \theta) = \|v - g(\theta)\|^2$ with some Euclidean norm $\|\cdot\|$, i.e. $\|y\| = \sqrt{y^\top A y}$ for some symmetric, positive definite matrix $A \in \mathbb{R}^{q \times q}$. According to Remark 3.3, it suffices to consider non-randomized procedures, i.e. point estimators of $g(\theta)$. If we assume that

$$\mathbb{E}^{\mathrm{B}}(\|g(\theta)\|^2) \;=\; \int_\Theta \|g(\theta)\|^2\, \pi(\theta)\, N(\mathrm{d}\theta) \;<\; \infty,$$

then the formula

$$\mathbb{E}^{\mathrm{B}}(\|g(\theta)\|^2) \;=\; \int_{\mathcal{X}} f(x) \int_\Theta \|g(\theta)\|^2 \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)\, M(\mathrm{d}x)$$

shows that we may replace $\pi(\theta \,|\, x)$ with $\pi(\theta)$ whenever $\int_\Theta \|g(\theta)\|^2 \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)$ is infinite. After that modification,

$$\widehat{g}_*(x) \;:=\; \mathbb{E}^{\mathrm{B}}(g(\theta) \,|\, X = x) \;=\; \int_\Theta g(\theta) \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)$$

defines a Bayes-optimal point estimator of $g(\theta)$ for the prior Π. Indeed, for any any point estimator $\widehat{g} : \mathcal{X} \to \mathbb{R}^q$,

$$\begin{aligned} R(\widehat{g}, \Pi) \;&=\; \int_{\mathcal{X}} f(x) \int_\Theta \|\widehat{g}(x) - g(\theta)\|^2 \pi(\theta \,|\, x)\, N(\mathrm{d}\theta)\, M(\mathrm{d}x) \\ &=\; \int_{\mathcal{X}} f(x) \Big(\|\widehat{g}(x) - \widehat{g}_*(x)\|^2 + \int_\Theta \|\widehat{g}_*(x) - g(\theta)\|^2 \pi(\theta \,|\, x)\, N(\mathrm{d}\theta) \Big)\, M(\mathrm{d}x) \\ &=\; \mathbb{E}^{\mathrm{B}}\big(\|\widehat{g}(X) - \widehat{g}_*(X)\|^2\big) + R(\widehat{g}_*, \Pi). \end{aligned}$$

For dimension $q = 1$ and $L(v, \theta) = (v - g(\theta))^2$, we get the useful equation

$$R(\widehat{g}_*, \Pi) \;=\; \int_{\mathcal{X}} f(x)\, \mathrm{Var}^{\mathrm{B}}(g(\theta) \,|\, X = x)\, M(\mathrm{d}x) \;=\; \mathbb{E}^{\mathrm{B}}\big(\mathrm{Var}^{\mathrm{B}}(g(\theta) \,|\, X)\big).$$

Example 3.8 (Gaussian location model). For a given sample size $n \in \mathbb{N}$ and a given standard deviation $\sigma > 0$, let $\mathcal{E} = \big(\mathcal{N}(\mu, \sigma^2)^{\otimes n}\big)_{\mu \in \mathbb{R}}$. The sample mean $\widehat{\mu}_*(\boldsymbol{x}) := \bar{x}$ is a minimax-optimal and admissible point estimator of $g(\mu) = \mu$ for the quadratic loss

$$L(v, \mu) \;:=\; (v - \mu)^2.$$

We have verified admissibility already in Theorem 1.24, but for the convenience of readers who skipped Chapter 1, and to illustrate the use of Lemma 3.5, we provide the main arguments once more. Note first that

$$R(\widehat{\mu}_*, \cdot) \;\equiv\; \frac{\sigma^2}{n}.$$

The density f_μ of $\mathcal{N}(\mu, \sigma^2)^{\otimes n}$ with respect to Lebesgue measure on $\mathbb{R}^n$ can be rewritten as

$$f_\mu(\boldsymbol{x}) \;=\; f_{\bar{x}}(\boldsymbol{x}) \exp\Big(-\frac{n(\mu - \bar{x})^2}{2\sigma^2}\Big).$$

For $k \geq 1$ let $\Pi_k := \mathrm{N}(0, k^2)$. Its density π with respect to Lebesgue measure on $\mathbb{R}$ is given by

$$\pi(\mu) \;=\; C \exp\Big(-\frac{\mu^2}{2k^2}\Big)$$

for some $C > 0$. Thus,

$$\begin{aligned} f_\mu(\boldsymbol{x})\pi(\mu) \;&=\; h_1(\boldsymbol{x}) \exp\Big(-\Big(\frac{1}{k^2} + \frac{n}{\sigma^2}\Big)\frac{\mu^2}{2} + \frac{n\bar{x}}{\sigma^2}\mu\Big) \\ &=\; h_2(\boldsymbol{x}) \exp\Big(-\Big(\mu - \frac{n\bar{x}}{n + \sigma^2/k^2}\Big)^2 \Big/ \Big(2\frac{\sigma^2}{n + \sigma^2/k^2}\Big)\Big) \end{aligned}$$

with certain factors $h_1(\boldsymbol{x}), h_2(\boldsymbol{x}) > 0$. As a function of μ, the right-hand side is the density of

$$\mathrm{N}\Big(\frac{n\bar{x}}{n + \sigma^2/k^2}, \frac{\sigma^2}{n + \sigma^2/k^2}\Big)$$

up to a positive factor depending only on $\boldsymbol{x}$. Hence the posterior distribution of μ, given that $\boldsymbol{X} = \boldsymbol{x}$, is this particular Gaussian distribution. In particular, a Bayes-optimal estimator of μ is given by

$$\widehat{\mu}^{\mathrm{B}}(\boldsymbol{x}) \;=\; \frac{n\bar{x}}{n + \sigma^2/k^2}$$

with Bayes risk

$$R(\widehat{\mu}^{\mathrm{B}}, \Pi) \;=\; \frac{\sigma^2}{n + \sigma^2/k^2}.$$

Since this converges to σ^2/n as $k \to \infty$, Lemma 3.5 shows that $\widehat{\mu}_*$ is minimax-optimal.

To prove admissibility of $\widehat{\mu}_*$, suppose that $\widehat{\mu}$ is another estimator of μ such that $R(\widehat{\mu}, \cdot) \leq \sigma^2/n$. By means of Exercise 1.25 one can show that $R(\widehat{\mu}, \cdot)$ is continuous on $\mathbb{R}$. So $R(\widehat{\mu}, \mu_o) < \sigma^2/n$

for some μ_o would imply that for suitable real numbers $a < b$ and $\epsilon > 0$, $R(\widehat{\mu}, \mu) \le \sigma^2/n - \epsilon$ for all $\mu \in [a, b]$. But then

$$R(\widehat{\mu}, \Pi_k) \;\le\; \frac{\sigma^2}{n} - \epsilon \Pi_k([a,b]) \;=\; \frac{\sigma^2}{n} - \epsilon\big(\Phi(b/k) - \Phi(a/k)\big),$$

whereas

$$R(\widehat{\mu}, \Pi_k) \;\ge\; R(\widehat{\mu}^{\mathrm{B}}, \Pi_k) \;=\; \frac{\sigma^2}{n + \sigma^2/k^2}.$$

Combining these inequalites leads to

$$k\epsilon\Big(\Phi(b/k) - \Phi(a/k)\Big) \;\le\; k\Big(\frac{\sigma^2}{n} - \frac{\sigma^2}{n + \sigma^2/k^2}\Big) \;<\; \frac{\sigma^4}{kn^2}.$$

But as $k \to \infty$, the left-hand side converges to $\epsilon(b-a)\Phi'(0) > 0$ while the right-hand side converges to 0, which yields a contradiction.

Exercise 3.9 (Point estimation of a binomial parameter). For $\theta \in [0,1]$ let $P_\theta := \mathrm{Bin}(n, \theta)$. We consider point estimators of θ with loss function $L : [0,1] \times [0,1] \to [0, \infty)$ given by

$$L(v, \theta) \;:=\; (v - \theta)^2.$$

(a) Consider the Bayesian model with a random parameter $\theta \sim \mathrm{Beta}(a, b)$ with given "hyperparameters" $a, b > 0$, and a random observation X with conditional distribution P_θ, given θ. Here $\mathrm{Beta}(a, b)$ is the distribution on $(0, 1)$ with Lebesgue density

$$\pi_{a,b}(\theta) \;=\; \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \quad B(a,b) \;:=\; \int_0^1 u^{a-1}(1-u)^{b-1}\, du \;=\; \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Show that

$$\mathcal{L}^{\mathrm{B}}(\theta \,|\, X = x) \;=\; \mathrm{Beta}(a + x, b + n - x) \quad \text{and} \quad \mathbb{E}^{\mathrm{B}}(\theta \,|\, X = x) \;=\; \frac{a + x}{a + b + n}.$$

(b) Part (a) implies that

$$\widehat{\theta}_{a,b}(x) \;:=\; \frac{a + x}{a + b + n}$$

minimizes the Bayes risk

$$R(\widehat{\theta}, \mathrm{Beta}(a,b)) \;:=\; \int_0^1 R(\widehat{\theta}, \theta)\, \mathrm{Beta}(a,b)(\mathrm{d}\theta)$$

over all estimators $\widehat{\theta} : \{0, 1, \ldots, n\} \to \mathbb{R}$. Determine the risk function $R(\widehat{\theta}_{a,b}, \cdot)$ of this estimator $\widehat{\theta}_{a,b}$.

(c) Now find parameters a, b such that the risk function in part (b) is constant. What are the consequences for the corresponding prior $\mathrm{Beta}(a, b)$ and the corresponding estimator $\widehat{\theta}_{a,b}$?

Unbiasedness. Sometimes it is difficult to find decision procedures satisfying some optimality criterion. But often the problem gets easier if we impose some additional constraints which are quite natural by themselves. Here is one such constraint: A decision procedure ρ is called *unbiased*, if for arbitrary $\theta, \eta \in \Theta$,

$$R(\rho, \theta) \;=\; \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \theta)\, \rho(x, \mathrm{d}v)\, P_\theta(\mathrm{d}x) \;\le\; \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \eta)\, \rho(x, \mathrm{d}v)\, P_\theta(\mathrm{d}x). \tag{3.2}$$

Example 3.1 (Point estimation, continued). Let $L(v,\theta) := \|v - g(\theta)\|^2$ with some Euclidean norm $\|\cdot\|$ on $\mathbb{R}^q$, that is, $\|y\| = \sqrt{y^\top A y}$ for some symmetric, positive definite matrix $A \in \mathbb{R}^{q\times q}$. For any estimator $\widehat{g} : \mathcal{X} \to \mathbb{R}^q$, the risk $R(\widehat{g},\theta) = E_\theta\big(\|\widehat{g} - g(\theta)\|^2\big)$ is finite if, and only if, $E_\theta\big(\|\widehat{g}\|^2\big) < \infty$. In the latter case,

$$\begin{aligned} E_\theta\big(\|\widehat{g} - g(\eta)\|^2\big) &= E_\theta\big(\|\widehat{g} - E_\theta(\widehat{g})\|^2\big) + \big\|E_\theta(\widehat{g}) - g(\eta)\big\|^2, \\ E_\theta\big(\|\widehat{g} - g(\theta)\|^2\big) &= E_\theta\big(\|\widehat{g} - E_\theta(\widehat{g})\|^2\big) + \big\|E_\theta(\widehat{g}) - g(\theta)\big\|^2, \end{aligned}$$

so $\widehat{g}$ is unbiased if an only if for all $\theta \in \Theta$,

$$\big\|E_\theta(\widehat{g}) - g(\theta)\big\| \;=\; \min_{\eta\in\Theta} \big\|E_\theta(\widehat{g}) - g(\eta)\big\|.$$

Consequently, if $\widehat{g}$ is an estimator such that

$$E_\theta\big(\|\widehat{g}\|^2\big) \;<\; \infty \quad \text{and} \quad E_\theta(\widehat{g}) \;\in\; \text{closure}\{g(\eta) : \eta \in \Theta\}$$

for all $\theta \in \Theta$, then $\widehat{g}$ is unbiased if, and only if,

$$E_\theta(\widehat{g}) \;=\; g(\theta) \quad \text{for all } \theta \in \Theta. \tag{3.3}$$

In what follows, a point estimator $\widehat{g}$ will be called unbiased if it satisfies (3.3).

Example 3.2 (Statistical tests, continued). We consider the general loss function $L(v,\theta) = 1_{[v=1,\theta\in\Theta_0]}\lambda_1 + 1_{[v=0,\theta\in\Theta_1]}\lambda_2$ with $\lambda_1, \lambda_2 > 0$. Then

$$\int_{\mathcal{X}} \int_{\mathcal{V}} L(v,\eta)\, \rho(x,dv)\, P_\theta(\mathrm{d}x) \;\geq\; R(\varphi,\theta) \quad \text{for arbitrary } \theta,\eta \in \Theta$$

is easily shown to be equivalent to

$$\begin{cases} \lambda_2(1 - E_{\theta_0}(\varphi)) \;\geq\; \lambda_1 E_{\theta_0}(\varphi) & \text{for any } \theta_0 \in \Theta_0, \\ \lambda_1 E_{\theta_1}(\varphi) \;\geq\; \lambda_2(1 - E_{\theta_1}(\varphi)) & \text{for any } \theta_1 \in \Theta_1. \end{cases}$$

In other words,

$$E_\theta(\varphi) \begin{cases} \leq\; \alpha & \text{for } \theta \in \Theta_0 \\ \geq\; \alpha & \text{for } \theta \in \Theta_1 \end{cases}$$

with

$$\alpha \;:=\; \frac{\lambda_2}{\lambda_1 + \lambda_2} \in (0,1).$$

Coming from the other end, for a given test level $\alpha \in (0,1)$, a level-α test of Θ_0 with power at least α for each $\theta \in \Theta_1$ is unbiased in the sense of (3.2) if $\lambda_2/\lambda_1 = \alpha/(1-\alpha)$.

Exercise 3.10 (Estimating functions of a binomial parameter). Unbiasedness of point estimators seems often a natural constraint. But it is potentially too restrictive. Consider the statistical experiment $\big(\mathrm{Bin}(n,\theta)\big)_{\theta\in[0,1]}$ and an arbitrary function $g : [0,1] \to \mathbb{R}$.

(a) Suppose that $\widehat{g} : \{0,1,\ldots,n\} \to \mathbb{R}$ is an unbiased estimator of $g(\theta)$, i.e. $E_\theta(\widehat{g}) = g(\theta)$ for all $\theta \in [0,1]$. Show that $g(\theta)$ is a polynomial in θ of order at most n.

(b) Show that an unbiased estimator $\widehat{g}$ as in part (a) is unique.

(c) Determine the estimator $\widehat{g}$ explicitly in case of $g(\theta) = \theta^k$ for some $k \in \{1,\ldots,n\}$. Hint: Consider factorials $[x]_k := \prod_{0\leq i<k}(x-i)$.

3.3 Informativity and Sufficiency

This section is about the comparison of two statistical experiments with one and the same parameter space Θ.

3.3.1 Informativity

Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ and $\tilde{\mathcal{E}} = (\tilde{\mathcal{X}}, \tilde{\mathcal{B}}, (\tilde{P}_\theta)_{\theta\in\Theta})$ be statistical experiments. Experiment $\tilde{\mathcal{E}}$ is called *at least as informative as* experiment $\mathcal{E}$, if the following condition is satisfied: Let $(\mathcal{V}, \mathcal{C})$ be an arbitrary decision space and $L : \mathcal{V} \times \Theta \to (-\infty, \infty]$ be any loss function. Then for any decision procedure $\rho : \mathcal{X} \times \mathcal{C} \to [0, 1]$, there exists a decision procedure $\tilde{\rho} : \tilde{\mathcal{X}} \times \mathcal{C} \to [0, 1]$ such that

$$R(\tilde{\rho}, \cdot) \;\le\; R(\rho, \cdot) \quad \text{on } \Theta.$$

This looks like a very strong condition, because it involves arbitrary decision spaces and loss functions. Nevertheless there is an elegant criterion due to David Blackwell:

Lemma 3.11 (Blackwell's criterion). *Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ and $\tilde{\mathcal{E}} = (\tilde{\mathcal{X}}, \tilde{\mathcal{B}}, (\tilde{P}_\theta)_{\theta\in\Theta})$ be statistical experiments. Suppose there exists a stochastic kernel K from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ such that for any $\theta \in \Theta$ and $B \in \mathcal{B}$,*

$$P_\theta(B) \;=\; \tilde{P}_\theta \otimes K(\tilde{\mathcal{X}} \times A) \;=\; \int_{\tilde{\mathcal{X}}} K(\tilde{x}, B)\, \tilde{P}_\theta(\mathrm{d}\tilde{x}).$$

Then $\tilde{\mathcal{E}}$ is at least as informative as $\mathcal{E}$.

The intuition behind Blackwell's criterion. Suppose that Bob and Alice are planning statistical experiments $\mathcal{E}$ and $\tilde{\mathcal{E}}$, respectively, to do inference about an unknown parameter in Θ.

Alice: "My experiment is at least as good as yours." Bob: "How do you know?"
Alice: "Well, tell me any decision space and loss function." Bob: "$(\mathcal{V}, \mathcal{C})$ and L."
Alice: "Okay, and what is vour favourite decision procedure for that?" Bob: "ρ."
Alice: "All right, this is what I will do: If my experiment yields data $\tilde{X} \sim \tilde{P}_\theta$, where $\theta \in \Theta$ is unknown, I will use Mr. Blackwell's kernel K to generate new data $X \sim K(\tilde{X}, \cdot)$. And then, I'll use your ρ to draw a decision $V \sim \rho(X, \cdot)$ about θ." Bob: "That looks like a swindle!"

Proof of Lemma 3.11. The assumption implies that for any measurable function $h : (\mathcal{X}, \mathcal{B}) \to [0, \infty]$,

$$\int_{\mathcal{X}} h \,\mathrm{d}P_\theta \;=\; \int_{\tilde{\mathcal{X}}} \int_{\mathcal{X}} h(x)\, K(\tilde{x}, \mathrm{d}x)\, \tilde{P}_\theta(\mathrm{d}\tilde{x}).$$

By assumption, this is true for indicator functions $h = 1_B$ with $B \in \mathcal{B}$. By linearity of integration, this is true for measurable functions taking only finitely many different values in $[0, \infty)$. By monotone convergence, the asserted equation is true for arbitrary measurable functions $h : \mathcal{X} \to [0, \infty]$.

For a given decision space $(\mathcal{V}, \mathcal{C})$ and decision procedure $\rho : \mathcal{X} \times \mathcal{C} \to [0,1]$ we define a decision procedure $\tilde{\rho} : \tilde{\mathcal{X}} \times \mathcal{C} \to [0,1]$ as follows: For $C \in \mathcal{C}$ we set

$$\tilde{\rho}(\tilde{x}, C) \;:=\; \int_{\mathcal{X}} \rho(x, C)\, K(\tilde{x}, \mathrm{d}x).$$

This construction of $\tilde{\rho}$ implies that for any measurable function $h : \mathcal{V} \to [0, \infty]$ and $\tilde{x} \in \tilde{\mathcal{X}}$,

$$\int_{\mathcal{V}} h(v)\, \tilde{\rho}(\tilde{x}, \mathrm{d}v) \;=\; \int_{\mathcal{X}} \int_{\mathcal{V}} h(v)\, \rho(x, \mathrm{d}v)\, K(\tilde{x}, \mathrm{d}x).$$

Hence for loss functions $L : \mathcal{V} \times \Theta \to [0, \infty]$ and any $\theta \in \Theta$,

$$\begin{aligned} R(\tilde{\rho}, \theta) \;&=\; \int_{\tilde{\mathcal{X}}} \int_{\mathcal{V}} L(v, \theta)\, \tilde{\rho}(\tilde{x}, \mathrm{d}v)\, \tilde{P}_\theta(\mathrm{d}\tilde{x}) \\ &=\; \int_{\tilde{\mathcal{X}}} \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \theta)\, \rho(x, \mathrm{d}v)\, K(\tilde{x}, \mathrm{d}x)\, \tilde{P}_\theta(\mathrm{d}\tilde{x}) \\ &=\; \int_{\mathcal{X}} \int_{\mathcal{V}} L(v, \theta)\, \rho(x, \mathrm{d}v)\, P_\theta(\mathrm{d}x) \\ &=\; R(\rho, \theta). \end{aligned}$$

□

Example 3.12 (Sampling with and without replacement). Let $\mathcal{M}$ be a population of *known* size $N = \#\mathcal{M}$. Suppose we draw a sample of size $n \geq 2$ from $\mathcal{M}$ *with* replacement. Precisely, we obtain a sample $\omega = (\omega_1, \ldots, \omega_n)$ in the set $\Omega := \mathcal{M}^n$ with N^n different elements completely at random. This corresponds to the uniform distribution $\mathbb{P}$ on Ω.

In case of $N \geq n$, sampling *without* replacement would be an alternative strategy, leading to the set $\tilde{\Omega}$ of all $[N]_n = N(N-1)\cdots(N-n+1)$ tuples $\omega \in \Omega$ with pairwise different elements. The uniform distribution on $\tilde{\Omega}$ is denoted by $\tilde{\mathbb{P}}$.

Suppose that we are interested in a specific property $\theta \in \Theta$ of the population $\mathcal{M}$. Without specifying random variables $X : \Omega \to \mathcal{X}$, $\tilde{X} : \tilde{\Omega} \to \tilde{\mathcal{X}}$ and how their distributions depend on θ, one can say that sampling without replacement is more informative than sampling with replacement. If $n = N$, this is rather obvious, because any sample $\omega \in \tilde{\Omega}$ contains all individuals of $\mathcal{M}$. But what about the more interesting case of $n < N$?

To this end, imagine two gentlemen, Mr. Diligent and Mr. Lazy. Mr. Diligent draws a sample $\tilde{\omega}$ from $\tilde{\Omega}$ completely at random, step by step. Mr. Lazy watches Mr. Diligent and defines ("draws") a sample $\omega \in \Omega$ as follows: He sets $\omega_1 := \tilde{\omega}_1$. For $k = 1, 2, \ldots, n-1$ he imagines drawing ω_{k+1} completely at random from $\mathcal{M}$. Conditional on the outcomes of the first k draws of Mr. Diligent, this would yield $\tilde{\omega}_j$ with probability $1/N$ for $1 \leq j \leq k$, and with probability $1 - k/N$, ω_{k+1} would have the same distribution as $\tilde{\omega}_{k+1}$. Thus Mr. Lazy chooses

$$\omega_{k+1} \;=\; \tilde{\omega}_J$$

with a random index

$$J \;\sim\; \frac{1}{N} \sum_{j=1}^{k} \delta_j + \Big(1 - \frac{k}{N}\Big) \delta_{k+1}.$$

Exercise 3.13 (From sampling without to sampling with replacement). Show that the construction of ω from $\tilde{\omega}$ in Example 3.12 corresponds to a stochastic kernel K from $\tilde{\Omega}$ to Ω such that for any $x \in \tilde{\Omega}$,

$$K\big(x, \{x_1, \ldots, x_n\}^n\big) \;=\; 1,$$

and for any $y \in \mathcal{M}^n$,

$$\frac{1}{[N]_n} \sum_{x \in \tilde{\Omega}} K(x, \{y\}) \;=\; \frac{1}{N^n}.$$

Exercise 3.14 (Estimating the reciprocal of a population size). At first glance one could think that sampling from a population $\mathcal{M}$ without replacement is always more informative than sampling with replacement. But this is not true in general. For instance, suppose that the size N of the population $\mathcal{M}$ is unknown. Then drawing a random sample of size n without replacement from that population reveals nothing about the population size N except that $N \geq n$. But sampling with replacement yields some information:

For a given integer $n \geq 2$, let $\mathbb{P}_{\mathcal{M}}$ be the uniform distribution on $\mathcal{M}^n$. We are interested in constructing an unbiased estimator of $g(\mathcal{M}) := 1/N$ in the sense of (3.3).

(a) Determine the expectation of X_i where

$$X_i(\omega) \;:=\; \frac{1_{[\omega_i \in \{\omega_j : j \neq i\}]}}{\#\{\omega_1, \omega_2, \ldots, \omega_n\}}, \quad \text{for } \omega \in \mathcal{M}^n.$$

for all $\omega \in \mathcal{M}^n$ and $i \in \{1, 2, \ldots, n\}$.

(b) Propose an unbiased estimator of $g(\mathcal{M})$.

(c) For $1 \leq i < j \leq n$, let $X_{ij}(\omega) := 1_{[\omega_i = \omega_j]}$. Determine the expectation of X_{ij} and propose an unbiased estimator of $g(\mathcal{M})$.

(d) Determine the standard deviation of your estimator in part (c).

3.3.2 Sufficiency

The concept of sufficiency is a special instance of Blackwell's criterion. We consider a statistical experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta \in \Theta})$. Now we want to know whether it is sufficient to restrict one's attention to partial information about the experiment's outcome. Partial information could mean that we want to replace an observation $X \sim P_\theta$ with a given function $T(X)$, or we want to restrict our attention to events B_o in a sub-σ-field $\mathcal{B}_o$ of $\mathcal{B}$.

Definition 3.15 (Sufficient statistic). A measurable mapping $T : (\mathcal{X}, \mathcal{B}) \to (\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ is called a *sufficient statistic for* $\mathcal{E}$, if there exists a stochastic kernel K from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ describing the conditional distribution of $X \sim P_\theta$, given $T(X)$, for any $\theta \in \Theta$. In other words, for arbitrary $\theta \in \Theta$, $B \in \mathcal{B}$ and $\tilde{B} \in \tilde{\mathcal{B}}$,

$$P_\theta(\{T \in \tilde{B}\} \cap B) \;=\; \int_{\tilde{B}} K(t, B)\, P_\theta^T(\mathrm{d}t).$$

Instead of $K(t, B)$ one could also use the more intuitive notations $P(B \,|\, T = t)$ or $\mathbb{P}(X \in B \,|\, T(X) = t)$, and $\int_{\mathcal{X}} h(x)\, K(t, \mathrm{d}x)$ could be denoted as $E(h \,|\, T = t)$ or $\mathbb{E}(h(X) \,|\, T(X) = t)$. Due to sufficiency of T, no subscript θ is necessary.

Sufficiency of T implies that the experiment $\mathcal{E}^T := (\tilde{\mathcal{X}}, \tilde{\mathcal{B}}, (P_\theta^T)_{\theta\in\Theta})$ is at least as informative as $\mathcal{E}$. In other words, one may reduce raw data $X \sim \mathbb{P}_\theta$ to $T(X)$ without any loss of information. Indeed, having reduced X to $T(X)$, one could generate an artificial observation $X_o \sim K(T(X), \cdot)$, and $X_o \sim P_\theta$, too, no matter what value the unknown parameter $\theta \in \Theta$ has. Sufficiency can often be verified with the following criterion:

Theorem 3.16 (Halmos–Savage, Neyman). *Suppose that $(\mathcal{X}, d)$ is a separable and complete metric space, and let $\mathcal{B} = \mathrm{Borel}(\mathcal{B}, d)$. Further let M be a σ-finite measure on $(\mathcal{X}, \mathcal{B})$ such that each distribution P_θ has a density f_θ with respect to M. Then the following two conditions are equivalent:*

(i) The statistic T is sufficient for $\mathcal{E}$.

(ii) There exists a measurable function $h : \mathcal{X} \to [0, \infty)$ such that for any $\theta \in \Theta$, the density f_θ can be chosen such that

$$f_\theta(x) \;=\; g_\theta(T(x))h(x)$$

with $g_\theta : \tilde{\mathcal{X}} \to [0, \infty)$ measurable.

The fact that condition (ii) implies sufficiency of T for $\mathcal{E}$ is called Neyman's factorization criterion.

Corollary 3.17. *Let $\mathcal{X}$ and $\tilde{\mathcal{X}}$ be countable sets equipped with $\mathcal{B} = \mathcal{P}(\mathcal{X})$ and $\tilde{\mathcal{X}} = \mathcal{P}(\tilde{\mathcal{X}})$. Then, T is sufficient if and only if there exists a function $h : \mathcal{X} \to [0, \infty)$ such that for any $\theta \in \Theta$ and $x \in \mathcal{X}$,*

$$P_\theta(\{x\}) \;=\; g_\theta(T(x))h(x)$$

with $g_\theta : \tilde{\mathcal{X}} \to [0, \infty)$.

For readers feeling uneasy about measure theory it may be instructive to prove the latter corollary directly. It is a consequence of Theorem 3.16 if we use the metric $d(x, x') := 1_{[x\neq x']}$ on $\mathcal{X}$ and counting measure M on $\mathcal{X}$, that is, $M(\{x\}) = 1$ for all $x \in \mathcal{X}$.

Proof of Theorem 3.16. A measure M on $(\mathcal{X}, \mathcal{B})$ is σ-finite if and only if there exists a measurable function $J : \mathcal{X} \to (0, \infty)$ such that $\int J \,\mathrm{d}M = 1$. But then we could replace M with the measure M_o given by $M_o(B) := \int_B J \,\mathrm{d}M$, the density f_θ with f_θ/J and the function h in condition (ii) with h/J. Hence we may and do assume without loss of generality that M is a probability measure.

Suppose that condition (ii) is satisfied. By our assumption on $(\mathcal{X}, \mathcal{B})$, there exists a stochastic kernel K_o from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ such that for arbitrary $\tilde{B} \in \tilde{\mathcal{B}}$ and $B \in \mathcal{B}$,

$$M(\{T \in \tilde{B}\} \cap B) \;=\; \int_{\tilde{B}} K_o(t, B)\, M^T(\mathrm{d}t).$$

More generally, for arbitrary measurable and non-negative functions f on $\tilde{\mathcal{X}} \times \mathcal{X}$,

$$\int_{\mathcal{X}} f(T(x), x)\, M(\mathrm{d}x) \;=\; \int_{\tilde{\mathcal{X}}} \int_{\mathcal{X}} f(t, x)\, K_o(t, \mathrm{d}x)\, M^T(\mathrm{d}t).$$

This implies that for arbitrary $\theta \in \Theta$, $\tilde{B} \in \tilde{\mathcal{B}}$ and $B \in \mathcal{B}$,

$$\begin{aligned} P_\theta(\{T \in \tilde{B}\} \cap B) &= \int_{\mathcal{X}} 1_{\tilde{B}}(T(x)) 1_B(x) g_\theta(T(x)) h(x)\, M(\mathrm{d}x) \\ &= \int_{\tilde{\mathcal{X}}} \int_{\mathcal{X}} 1_{\tilde{B}}(t) g_\theta(t) 1_B(x) h(x)\, K_o(t, \mathrm{d}x)\, M^T(\mathrm{d}t) \\ &= \int_{\tilde{B}} g_\theta(t) \int_B h(x)\, K_o(t, \mathrm{d}x)\, M^T(\mathrm{d}t). \end{aligned}$$

Taking $B = \mathcal{X}$ shows that

$$P_\theta(T \in \tilde{B}) \;=\; \int_{\tilde{B}} f_\theta^T(t)\, M^T(\mathrm{d}t),$$

where

$$f_\theta^T(t) \;:=\; g_\theta(t) H(t) \quad \text{with} \quad H(t) \;:=\; \int_{\mathcal{X}} h(x)\, K_o(t, \mathrm{d}x).$$

In particular, the set $\tilde{N} := \{t \in \tilde{\mathcal{X}} : H(t) = \infty\}$ satisfies $P_\theta^T(\tilde{N} \cap \{g_\theta > 0\}) = 0$ for any $\theta \in \Theta$. Thus we may replace $f_\theta^T(t)$ with $1_{[H(t)<\infty]} f_\theta^T(t)$ to obtain a density f_θ^T of P_θ^T with respect to M^T. Then,

$$K(t, B) \;:=\; \begin{cases} H(t)^{-1} \displaystyle\int_B h(x)\, K_o(t, \mathrm{d}x) & \text{if } 0 < H(t) < \infty \\ M(B) & \text{else} \end{cases}$$

defines a stochastic kernel K from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ such that

$$P_\theta(\{T \in \tilde{B}\} \cap B) \;=\; \int_{\tilde{B}} f_\theta^T(t) K(t, B)\, M^T(\mathrm{d}t) \;=\; \int_{\tilde{B}} K(t, B)\, P_\theta^T(\mathrm{d}t)$$

for arbitrary $\theta \in \Theta$, $\tilde{B} \in \tilde{\mathcal{B}}$ and $B \in \mathcal{B}$. Hence, condition (i) is satisfied too.

Suppose that condition (i) is satisfied. By our assumption on $(\mathcal{X}, \mathcal{B})$, the space $L^1(M)$ is separable with respect to the corresponding L^1-norm $\|\cdot\|$, i.e. $\|g\| := \int |g|\, \mathrm{d}M$. This implies that the set $\{f_\theta : \theta \in \Theta\} \subset L^1(M)$ is separable too. Thus, for a suitable subset $\{\theta_\ell : \ell \in \mathbb{N}\}$ of Θ, the set $\{f_{\theta_\ell} : \ell \in \mathbb{N}\}$ is dense in $\{f_\theta : \theta \in \Theta\}$. This implies that all distributions P_θ, $\theta \in \Theta$, are absolutely continuous with respect to the probability measure

$$Q \;:=\; \sum_{\ell=1}^{\infty} 2^{-\ell} P_{\theta_\ell}.$$

Indeed, if $P_\theta(B) > 0$ for some $B \in \mathcal{B}$, then for suitable $\ell \in \mathbb{N}$,

$$\big|P_\theta(B) - P_{\theta_\ell}(B)\big| \;\le\; \|f_\theta - f_{\theta_\ell}\| \;<\; P_\theta(B),$$

whence $Q(B) \ge 2^{-\ell} P_{\theta_\ell}(B) > 0$. By assumption, there exists a stochastic kernel K from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ such that for all $\theta \in \Theta$, $\tilde{B} \in \tilde{\mathcal{B}}$ and $B \in \mathcal{B}$,

$$P_\theta(\{T \in \tilde{B}\} \cap B) \;=\; \int_{\tilde{B}} K(t, B)\, P_\theta^T(\mathrm{d}t).$$

The construction of Q implies that

$$Q(\{T \in \tilde{B}\} \cap B) \;=\; \int_{\tilde{B}} K(t, B)\, Q^T(\mathrm{d}t),$$

and via measure-theoretic induction one can show that

$$\int_B g(T(x))Q(\mathrm{d}x) \;=\; \int_{\tilde{\mathcal{X}}} g(t)K(t,B)\,Q^T(\mathrm{d}t)$$

for any $B \in \mathcal{B}$ and any measurable function $g : \tilde{\mathcal{X}} \to [0,\infty)$. Specifically, for any $\theta \in \Theta$, the distribution P_θ^T is absolutely continuous with respect to Q^T, so choosing a density $g_\theta = \mathrm{d}P_\theta^T/\mathrm{d}Q^T$ leads to the equality

$$P_\theta(B) \;=\; \int_{\tilde{\mathcal{X}}} K(t,B)\,P_\theta^T(\mathrm{d}t) \;=\; \int_{\tilde{\mathcal{X}}} g_\theta(t)K(t,B)\,Q^T(\mathrm{d}t) \;=\; \int_B g_\theta(T(x))Q(\mathrm{d}x).$$

Since Q is absolutely continuous with respect to M, there exists a density $h = \mathrm{d}Q/\mathrm{d}M$, and we may write

$$P_\theta(B) \;=\; \int_B g_\theta(T(x))h(x)\,M(\mathrm{d}x).$$

Thus, $g_\theta(T)h = \mathrm{d}P_\theta/\mathrm{d}M$, which proves condition (ii). □

Example 3.18 (Bernoulli sequences). Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with values in $\{0,1\}$ and unknown parameter $\theta = \mathbb{P}(X_i = 1) = \mathbb{E}(X_i) \in [0,1]$. This leads to the statistical experiment $\mathcal{E} = \big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (P_\theta)_{\theta\in[0,1]}\big)$ with $P_\theta := ((1-\theta)\delta_0 + \theta\delta_1)^{\otimes n}$. The density f_θ of P_θ with respect to counting measure on $\{0,1\}^n$ is given by

$$f_\theta(\boldsymbol{x}) \;=\; \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \;=\; \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})},$$

i.e. a function of $T(\boldsymbol{x}) := \sum_{i=1} x_i$ only. Hence T is a sufficient statistic for $\mathcal{E}$. Indeed, T has distribution $\mathrm{Bin}(n,\theta)$, and for any $t \in \{0,1,\ldots,n\}$, the conditional distribution $P_\theta(\cdot\,|\,T=t)$ is the uniform distribution on the set of all $\boldsymbol{x} \in \{0,1\}^n$ with $T(\boldsymbol{x}) = t$.

Example 3.19 (Gaussian samples). Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with distribution $\mathcal{N}(\mu,\sigma^2)$, the mean $\mu \in \mathbb{R}$ and the standard deviation $\sigma > 0$ being unknown. This leads to the statistical experiment $\mathcal{E} = \big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), (P_\theta)_{\theta\in\Theta}\big)$ with $\Theta = \mathbb{R}\times(0,\infty)$ and $P_{\mu,\sigma} := \mathcal{N}(\mu,\sigma^2)^{\otimes n}$. The density $f_{\mu,\sigma}$ of $P_{\mu,\sigma}$ with repect to Lebesgue measure on $\mathbb{R}^n$ times $(2\pi)^{-n/2}$ is given by

$$\begin{aligned} f_{\mu,\sigma}(\boldsymbol{x}) \;&=\; \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma\Big) \\ &=\; \exp\Big(-\sum_{i=1}^n \frac{(x_i-\bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} - n\log\sigma\Big) \\ &=\; g_{\mu,\sigma}(T(\boldsymbol{x})), \end{aligned}$$

where $T(\boldsymbol{x}) = (T_1(\boldsymbol{x}), T_2(\boldsymbol{x}))$ and

$$\begin{aligned} T_1(\boldsymbol{x}) \;&:=\; \bar{x} = n^{-1}\sum_{i=1}^n x_i, \\ T_2(\boldsymbol{x}) \;&:=\; \sum_{i=1}^n (x_i-\bar{x})^2, \\ g_{\mu,\sigma}(t_1,t_2) \;&:=\; \exp\Big(-\frac{n(t_1-\mu)^2}{2\sigma^2} - \frac{t_2}{2\sigma^2} - n\log\sigma\Big). \end{aligned}$$

Hence the statistic $T = (T_1, T_2) : \mathbb{R}^n \to \mathbb{R} \times [0, \infty)$ is sufficient for the experiment $\mathcal{E}$.

It is worthwhile here to verify sufficiency directly, based on standard arguments in connection with student's t distribution: Let $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ be an orthonormal basis of $\mathbb{R}^n$, where $\boldsymbol{b}_1 = n^{-1/2}\mathbf{1}_n$. Then the distribution $P_{\mu,\sigma}$ coincides with the distribution of

$$\boldsymbol{X} \;:=\; \mu \mathbf{1}_n + \sigma \sum_{i=1}^{n} Z_i \boldsymbol{b}_i$$

with stochastically independent random variables $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$. The statistic $T(\boldsymbol{X})$ is equal to

$$\Big(\mu + n^{-1/2}\sigma Z_1, \sigma^2 \sum_{i=2}^{n} Z_i^2\Big),$$

and we may write

$$\boldsymbol{X} \;=\; T_1(\boldsymbol{X})\mathbf{1}_n + \sqrt{T_2(\boldsymbol{X})} \sum_{i=2}^{n} W_i \boldsymbol{b}_i$$

with

$$W_i \;:=\; \Big(\sum_{j=2}^{n} Z_j^2\Big)^{-1/2} Z_i.$$

The random vector $\boldsymbol{W} := (W_i)_{i=2}^n$ is uniformly distributed on the unit sphere in $\mathbb{R}^{n-1}$ and stochastically independent from $\sum_{i=2}^n Z_i^2$. Hence the conditional distribution of $\tilde{\boldsymbol{X}}$, given $\tilde{T}$, does not depend on the parameter (μ, σ).

Exercise 3.20 (Gamma distributions). Let $\boldsymbol{X} = (X_i)_{i=1}^n$ have $n \geq 2$ independent components with

$$X_i \;\sim\; \mathrm{Gamma}(a, b)$$

and unknown parameters $a, b > 0$.

(a) Determine an $\mathbb{R}^2$-valued sufficient statistic $T(\boldsymbol{X})$ for the corresponding statistical experiment $\big(\mathrm{Gamma}(a,b)\big)_{a,b>0}$.

(b) Determine the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X})$ in case of $n = 2$.

Exercise 3.21 (Markov chains with finite state space). Let $\boldsymbol{X} = (X_t)_{t=0}^n$ be a Markov chain with values in $\mathcal{X} = \{1, \ldots, L\}$ for some integer $L \geq 2$ and fixed starting point $X_0 = x_* \in \mathcal{X}$. That means, for $1 \leq k < n$ and $y_0, \ldots, y_n, z \in \mathcal{X}$,

$$\mathbb{P}_\theta\big(X_{k+1} = z \,\big|\, (X_t)_{t=0}^k = (y_t)_{t=0}^k\big) \;=\; \theta_{y_k, z}$$

with an unknown matrix $\theta \in [0,1]^{L \times L}$ such that

$$\sum_{z=1}^{L} \theta_{y,z} \;=\; 1 \quad \text{for all } y \in \mathcal{X}.$$

Let Θ be the set of all such matrices θ, and let P_θ be the resulting distribution of $\boldsymbol{X}$ on the finite set $\mathcal{X}^{n+1}$, where a tuple $\boldsymbol{x} \in \mathcal{X}^{n+1}$ is represented as $(x_t)_{t=0}^n$. Determine a sufficient statistic for the statistical experiment $(P_\theta)_{\theta \in \Theta}$.

Exercise 3.22. Let $\mathcal{M}$ be a population of individuals with identification numbers in $\mathbb{Z}$. We assume that the set of all identification numbers equals $\{a, \dots, b\}$ with unknown integers $a \le b$. We only know that $b - a + 1 \ge n$ for some given integer $n \ge 2$.

Now we draw a random sample of size n without replacement from $\mathcal{M}$ and note the tuple $\boldsymbol{X}(\omega)$ of the observed identification numbers.

Show that $T(\boldsymbol{x}) = \big(\min(\boldsymbol{x}), \max(\boldsymbol{x})\big)$ is a sufficient statistic for this experiment. Describe the conditional distribution of $\boldsymbol{x}$ given $T(\boldsymbol{x}) = (s_1, s_2)$.

Hint: The formal definition of the experiment involves

$$\begin{aligned} \mathcal{X} &= \{\boldsymbol{x} \in \mathbb{Z}^n : x_i \neq x_j \text{ whenever } i \neq j\}, \\ P_{a,b} &= \mathrm{Unif}\big(\mathcal{X} \cap \{a, \dots, b\}^n\big), \\ \Theta &= \big\{(a,b) \in \mathbb{Z}^n : b - a + 1 \ge n\big\}. \end{aligned}$$

Exercise 3.23 (Poisson experiment). Let $\lambda > 0$ be the unknown concentration of certain bacteria in some lake. To estimate λ, some people take probes of volumes $a_1, \dots, a_m > 0$ and determine the numbers $X_1, \dots, X_m$ of bacteria found therein. This leads to independent random variables $X_1, \dots, X_m$ with $X_i \sim \mathrm{Poiss}(a_i\lambda)$.

The corresponding statistical experiment is $\big(\otimes_{i=1}^m \mathrm{Poiss}(a_i\lambda)\big)_{\lambda>0}$ with sample space $\mathbb{N}_0^m$. Show that $T(\boldsymbol{X}) = X_+$ is a sufficient statistic for this experiment. What is the distribution of $T(\boldsymbol{X})$ when $\boldsymbol{X} \sim P_\lambda$, and what is the conditional distribution of $\boldsymbol{X}$, given that $T(\boldsymbol{X}) = N$?

Definition 3.24 (Sufficient sub-σ-fields). Let $\mathcal{B}_o$ be a σ-field over $\mathcal{X}$ such that $\mathcal{B}_o \subset \mathcal{B}$. It is called a *sufficient sub-σ-field for $\mathcal{E}$*, if there exists a stochastic kernel K from $(\mathcal{X}, \mathcal{B}_o)$ to $(\mathcal{X}, \mathcal{B})$ describing the conditional distribution of $X \sim X_\theta$, given $\mathcal{B}_o$, for any $\theta \in \Theta$. In other words, for arbitrary $\theta \in \Theta$, $B \in \mathcal{B}$ and $B_o \in \mathcal{B}_o$,

$$P_\theta(B_o \cap B) \;=\; \int_{B_o} K(x, B)\, P_\theta(\mathrm{d}x).$$

Note that sufficiency of $\mathcal{B}_o$ is equivalent to sufficiency of the statistic

$$T : (\mathcal{X}, \mathcal{B}) \;\to\; (\mathcal{X}, \mathcal{B}_o), \quad T(x) := x.$$

Sufficiency of $\mathcal{B}_o$ implies that the experiment $\mathcal{E}_o := (\mathcal{X}, \mathcal{B}_o, (P_\theta)_{\theta\in\Theta})$ is at least as informative as $\mathcal{E}$. In other words, when analyzing raw data $X \sim P_\theta$, we may restrict our attention to decision procedures which are $\mathcal{B}_o$-measurable rather than $\mathcal{B}$-measurable.

Example 3.25 (Invariant distributions). Let $\mathcal{G}$ be a finite group of measurable bijective mappings $g : \mathcal{X} \to \mathcal{X}$. That means, for arbitrary $g, h \in \mathcal{G}$, both $h \circ g$ and g^{-1} belong to $\mathcal{G}$, too. Now let $\mathcal{B}_\mathcal{G}$ be the set of $\mathcal{G}$-invariant sets $B \in \mathcal{B}$, i.e.

$$g(B) = \{g(x) : x \in B\} \;=\; B \quad \text{for all } g \in \mathcal{G}.$$

This is obviously a sub-σ-field of $\mathcal{B}$.

Now suppose that all distributions P_θ, $\theta \in \Theta$, are $\mathcal{G}$-invariant in the sense that

$$P_\theta^g \;=\; P_\theta \quad \text{for all } g \in \mathcal{G},$$

where P_θ^g is the image measure $P_\theta \circ g^{-1}$. Then $\mathcal{B}_\mathcal{G}$ is sufficient for $\mathcal{E}$, and the conditional distribution $P_\theta(\cdot \,|\, \mathcal{B}_o)$ is given by the stochastic kernel

$$K(x, B) \;:=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} 1_B(g(x))$$

i.e.

$$K(x, \cdot) \;=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} \delta_{g(x)}.$$

Hence the experiment $\mathcal{E}_\mathcal{G} := \big(\mathcal{X}, \mathcal{B}_\mathcal{G}, (P_\theta)_{\theta \in \Theta}\big)$ is at least as informative as $\mathcal{E}$.

Proof: Obviously, $K(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ for any $x \in \mathcal{X}$. For fixed $B \in \mathcal{B}$, the function $x \to K(x, B)$ is certainly $\mathcal{B}$-measurable. To verify $\mathcal{B}_\mathcal{G}$-measurability it suffices to show that $K(h(x), B) = K(x, A)$ for arbitrary $x \in \mathcal{X}$ and $h \in \mathcal{G}$, see Exercise 3.26 below. But

$$K(h(x), B) \;=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} 1_A(g \circ h(x)) \;=\; \frac{1}{\#\mathcal{G}} \sum_{\tilde{g} \in \mathcal{G}} 1_A(\tilde{g}(x)),$$

because for any fixed $h \in \mathcal{G}$, the mapping $\mathcal{G} \ni g \mapsto g \circ h \in \mathcal{G}$ is bijective. It remains to be shown that for arbitrary $\theta \in \Theta$, $B_o \in \mathcal{B}_\mathcal{G}$ and $B \in \mathcal{B}$,

$$P_\theta(B_o \cap B) \;=\; \int_{B_o} K(x, B)\, P_\theta(\mathrm{d}x).$$

But the right hand side equals

$$\frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} \int_{B_o} 1_B(g(x))\, P_\theta(\mathrm{d}x),$$

and each summand equals

$$\begin{aligned}
\int_{B_o} 1_B(g(x))\, P_\theta(\mathrm{d}x) \;&=\; P_\theta(B_o \cap g^{-1}(B)) \\
&=\; P_\theta(g^{-1}(g(B_o) \cap B)) \\
&=\; P_\theta(g^{-1}(B_o \cap B)) \qquad\qquad \text{(by } \mathcal{G}\text{-invariance of } B_o) \\
&=\; P_\theta^g(B_o \cap B) \\
&=\; P_\theta(B_o \cap B) \qquad\qquad \text{(by } \mathcal{G}\text{-invariance of } P_\theta).
\end{aligned}$$

Exercise 3.26. In the setting of Example 3.25, let $(\mathcal{V}, \mathcal{C})$ be another measurable space, and let $\rho : (\mathcal{X}, \mathcal{B}) \to (\mathcal{V}, \mathcal{C})$ be a measurable mapping.

(a) Suppose that ρ is $\mathcal{G}$-invariant in the sense that $\rho \circ g \equiv \rho$ for arbitrary $g \in \mathcal{G}$. Show that ρ is $\mathcal{B}_\mathcal{G}$-$\mathcal{C}$-measurable.

(b) Suppose that ρ is $\mathcal{B}_\mathcal{G}$ measurable, and suppose that $\mathcal{C}$ "separates points in $\mathcal{V}$". That means, for arbitrary different points $v_1, v_2 \in \mathcal{V}$ there exists a set $C \in \mathcal{C}$ such that $v_1 \in C$ but $v_2 \notin C$. Show that ρ is $\mathcal{G}$-invariant.

Example 3.27 (Permutation-invariance). For some integer $n \geq 2$ consider the product space $(\mathcal{X}^n, \mathcal{B}^{\otimes n})$. Further let $\mathcal{S}_n$ be the group of all permutations of $\{1, \ldots, n\}$, that means, bijective mappings $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. Any $\pi \in \mathcal{S}_n$ induces a measurable bijection $g_\pi : \mathcal{X}^n \to \mathcal{X}^n$,

$$\boldsymbol{x} = (x_i)_{i=1}^n \;\mapsto\; g_\pi(\boldsymbol{x}) \;:=\; (x_{\pi(i)})_{i=1}^n.$$

Indeed one can easily show that for permutations $\pi, \sigma \in \mathcal{S}_n$,

$$g_\pi \circ g_\sigma \;=\; g_{\sigma \circ \pi}.$$

(Note also that $\pi \in \mathcal{S}_n$ is uniquely determined by the mapping g_π, unless $\#\mathcal{X} = 1$.) Thus $\mathcal{G} := \{g_\pi : \pi \in \mathcal{S}_n\}$ is a group of measurable bijections of $(\mathcal{X}^n, \mathcal{B}^{\otimes n})$.

A distribution $\mathbb{P}$ on $(\mathcal{X}^n, \mathcal{B}^{\otimes n})$ is called *permutation-invariant* or *exchangeable* if it is $\mathcal{G}$-invariant. In other words, if $X = (X_i)_{i=1}^n$ has distribution $\mathbb{P}$, then for any $\pi \in \mathcal{S}_n$, the random tuple $(X_{\pi(i)})_{i=1}^n$ has distribution $\mathbb{P}$, too.

If $\mathcal{E} = \big(\mathcal{X}^n, \mathcal{B}^{\otimes n}, (\mathbb{P}_\theta)_{\theta \in \Theta}\big)$ involves only permutation-invariant distributions $\mathbb{P}_\theta$, then the sub-σ-field $\mathcal{A}_\mathcal{G}$ of all permutation-invariant sets in $\mathcal{B}^{\otimes n}$ is sufficient for $\mathcal{E}$. The corresponding stochastic kernel K from $(\mathcal{X}^n, \mathcal{A}_\mathcal{G})$ to $(\mathcal{X}^n, \mathcal{B}^{\otimes n})$ is given by

$$K(\boldsymbol{x}, \cdot) \;=\; \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} \delta_{g_\pi(\boldsymbol{x})}.$$

Example 3.28 (Sign invariance). For some integer $n \geq 1$ consider the measurable space $\mathbb{R}^n$. Any sign vector $\xi \in \{-1, 1\}^n$ induces a measurable bijection $g_\xi : \mathbb{R}^n \to \mathbb{R}^n$,

$$\boldsymbol{x} = (x)_{i=1}^n \;\mapsto\; g_\xi(\boldsymbol{x}) \;:=\; (\xi_i x_i)_{i=1}^n.$$

Note that $\{-1, 1\}^n$ with component-wise multiplication is an Abelian group. The corresponding family $\mathcal{G} := \big\{g_\xi : \xi \in \{-1, 1\}^n\big\}$ is an Abelian group, too. Indeed, for arbitrary $\xi, \zeta \in \{-1, 1\}^n$,

$$g_\xi \circ g_\zeta \;=\; g_\gamma \quad \text{with} \quad \gamma = (\xi_i \zeta_i)_{i=1}^n.$$

A distribution $\mathbb{P}$ on $\mathbb{R}^n$ is called *sign-invariant*, if it is $\mathcal{G}$-invariant. In other words, if $\boldsymbol{X} = (X_i)_{i=1}^n$ has distribution $\mathbb{P}$, then for any sign vector $\xi \in \{-1, 1\}^n$, the random vector $(\xi_i X_i)_{i=1}^n$ has distribution $\mathbb{P}$, too.

If $\mathcal{E} = \big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), (\mathbb{P}_\theta)_{\theta \in \Theta}\big)$ consists of sign-invariant distributions, then the sub-σ-field $\mathcal{A}_\mathcal{G}$ of all sign-invariant Borel sets in $\mathbb{R}^n$ is sufficient for $\mathcal{E}$. The corresponding stochastic kernel K from $(\mathbb{R}^n, \mathcal{A}_\mathcal{G})$ to $(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n))$ is given by

$$K(\boldsymbol{x}, \cdot) \;=\; \frac{1}{2^n} \sum_{\xi \in \{-1,1\}^n} \delta_{g_\xi(\boldsymbol{x})}.$$

3.4 Complete Statistical Experiments

Definition 3.29 (Complete statistical experiment). A statistical experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta \in \Theta})$ is called *boundedly complete*, if for any bounded and measurable function $f : \mathcal{X} \to \mathbb{R}$,

$$\int f \,\mathrm{d}P_\theta \;=\; 0 \quad \text{for all } \theta \in \Theta$$

implies that

$$P_\theta(f \neq 0) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

The experiment $\mathcal{E}$ is called *complete* if for any function $f \in \bigcap_{\theta\in\Theta} \mathcal{L}^1(P_\theta)$,

$$\int f \, dP_\theta \;=\; 0 \quad \text{for all } \theta \in \Theta$$

implies that

$$P_\theta(f \neq 0) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

Obviously, completeness of $\mathcal{E}$ implies bounded completeness.

Example 3.30 (Simple d-parameter exponential families). Let M be a σ-finite measure on $\mathbb{R}^d$ and $h : \mathbb{R}^d \to [0, \infty)$ a measurable function such that for all parameters θ in a set $\Theta \subset \mathbb{R}^d$,

$$0 \;<\; \int \exp(\theta^\top x) h(x) \, M(\mathrm{d}x) \;<\; \infty.$$

Then we consider the statistical experiment $\mathcal{E} = \big(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (P_\theta)_{\theta\in\Theta}\big)$ where

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}M}(x) \;=\; C(\theta) h(x) \exp(\theta^\top x)$$

with $C(\theta) := \big(\int h(x) \exp(\theta^\top x) \, M(\mathrm{d}x)\big)^{-1}$. If Θ has nonempty interior, then $\mathcal{E}$ is complete.

This follows immediately from Theorem A.5 in Appendix A.3.

Example 3.31 (Product measures). Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let Θ be a family of probability distributions on $(\mathcal{X}, \mathcal{B})$. For some integer $n \geq 2$ consider $\mathcal{E} = (\mathcal{X}^n, \mathcal{B}^{\otimes n}, (P^{\otimes n})_{P\in\Theta})$. This experiment is *not* boundedly complete. To see this, consider some bounded measurable function $h : \mathcal{X} \to \mathbb{R}$ with $\mathrm{Var}_{P_o}(h) > 0$ for at least one distribution $P_o \in \Theta$. Then $f(\boldsymbol{x}) := h(x_1) - h(x_2)$ defines a bounded and measurable function $f : \mathcal{X}^n \to \mathbb{R}$ such that $\int f \, \mathrm{d}P^{\otimes n} = 0$ for any probability distribution P, but $P_o^{\otimes n}(f \neq 0) > 0$, because $\int f^2 \, \mathrm{d}P_o^{\otimes n} = 2 \, \mathrm{Var}_{P_o}(h) > 0$.

However, if we replace $\mathcal{A} = \mathcal{B}^{\otimes n}$ with the sub-σ-field $\mathcal{A}_o$ of all permutation invariant sets $A \in \mathcal{A}$, then the experiment $\mathcal{E}_o = (\mathcal{X}^n, \mathcal{A}_o, (P^{\otimes n})_{P\in\Theta})$ is often complete, provided that Θ is sufficiently rich.

Special case 1. Suppose that $\mathcal{B} = \mathcal{P}(\mathcal{X})$ and Θ consists of all probability distributions on $\mathcal{X}$ with finite support. Then $\mathcal{E}_o$ is complete. More precisely, if $f : \mathcal{X}^n \to \mathbb{R}$ is permutation-invariant such that $\int f \, \mathrm{d}P^{\otimes n}$ for any probability measure P on $\mathcal{X}$ with finite support, then $f \equiv 0$.

Proof: Consider arbitrary points $x_1, \ldots, x_n \in \mathcal{X}$. For $\lambda \in (0, \infty)^n$, $P_\lambda := \lambda_+^{-1} \sum_{i=1}^n \lambda_i \delta_{x_i}$ is a probability distribution in Θ. Thus $\int f \, \mathrm{d}P^{\otimes n} = 0$ for all $P \in \Theta$ implies that

$$0 \;=\; \lambda_+^n \int f \, \mathrm{d}P_\lambda^{\otimes n} \;=\; \sum_{i(1),\ldots,i(n)=1}^{n} \Big(\prod_{\ell=1}^{n} \lambda_{i(\ell)}\Big) f(x_{i(1)}, \ldots, x_{i(\ell)}) \tag{3.4}$$

for all $\lambda \in (0,\infty)^n$. Note that the right-hand side of (3.4) is an n-variate polynomial of order n in λ. Since

$$\frac{\partial^n}{\partial\lambda_1\partial\lambda_2\cdots\partial\lambda_n}\prod_{\ell=1}^n \lambda_{i(\ell)} \;=\; \begin{cases} 1 & \text{if } \{i(1),\ldots,i(n)\} = \{1,\ldots,n\} \\ 0 & \text{else,} \end{cases} \tag{3.5}$$

it follows from (3.4) and permutation-invariance of f that

$$0 \;=\; \sum_{\pi\in\mathcal{S}_n} f(x_{\pi(1)},\ldots,x_{\pi(n)}) = \; n!f(x_1,\ldots,x_n).$$

□

Special case 2. Suppose that $(\mathcal{X},\mathcal{B},M)$ is a σ-finite measure space. Suppose that Θ consists of distributions which are absolutely continuous with respect to M. If it contains the convex hull of all probability measures $B \mapsto M(A)^{-1}M(A\cap B)$, where $A\in\mathcal{B}$ with $0<M(A)<\infty$, then $\mathcal{E}_o$ is complete.

Proof: Note first that if $f:\mathcal{X}^n\to\mathbb{R}$ belongs to $\bigcap_{P\in\Theta}\mathcal{L}^1(P^{\otimes n})$, then

$$\int_{A^n} |f|\,\mathrm{d}M^{\otimes n} \;<\; \infty$$

for any $A\in\mathcal{B}$ such that $M^{\otimes n}(A)<\infty$.

Now fix arbitrary sets $A_1,\ldots,A_n\in\mathcal{B}$ such that $0<M(A_i)<\infty$. For any tuple $\lambda\in(0,\infty)^n$ consider the probability measure $P_\lambda\in\Theta$ given by

$$P_\lambda(B) \;:=\; \Big(\sum_{i=1}^n \lambda_i M(A_i)\Big)^{-1}\sum_{i=1}^n \lambda_i M(A_i\cap B).$$

If $f:\mathcal{X}^n\to\mathbb{R}$ is measurable and permutation-invariant, it follows from $\int f\,\mathrm{d}P^{\otimes n}=0$ for all $P\in\Theta$ that

$$0 \;=\; \sum_{i(1),\ldots,i(n)=1}^n \prod_{\ell=1}^n \lambda_{i(\ell)} \int_{B_{i(1)}\times\cdots\times B_{i(n)}} f\,\mathrm{d}M^{\otimes n}.$$

By means of this equation, (3.5) and permutation-invariance of f one can deduce that

$$\int_{A_1\times\cdots\times A_n} f\,\mathrm{d}M^{\otimes n} \;=\; 0 \tag{3.6}$$

for arbitrary sets $A_1,\ldots,A_n\in\mathcal{B}$ with finite measure $M(A_i)$. But $\mathcal{X}=\bigcup_{\ell\geq 1}\mathcal{X}_\ell$ with measurable sets $\mathcal{X}_1\subset\mathcal{X}_2\subset\mathcal{X}_3\subset\cdots$ having finite measure under M. For arbitrary sets $A_1,\ldots,A_n$ in $\mathcal{B}_\ell := \{B\in\mathcal{B}: B\subset\mathcal{X}_\ell\}$, we can deduce from (3.6) that the finite measures $Q_\ell^\pm$ on $(\mathcal{X}_\ell^n,\mathcal{B}_\ell^{\otimes n})$, where $Q_\ell^\pm(A) := \int_A f^\pm\,dM^{\otimes n}$ for $A\in\mathcal{B}_\ell^{\otimes n}$, coincide on cartesian products $A_1\times\cdots\times A_n$ of sets $A_i\in\mathcal{B}_\ell$. But the family of these sets is closed under intersections and generates $\mathcal{B}_\ell^{\otimes n}$. Consequently, by Dynkin's theorem, $Q_\ell^+\equiv Q_\ell^-$. Taking $A=\{f^+>0\}\cap\mathcal{X}_\ell^n\subset\{f^-=0\}\cap\mathcal{X}_\ell^n$ and $A=\{f^->0\}\cap\mathcal{X}_\ell^n\subset\{f^+=0\}\cap\mathcal{X}_\ell^n$, we can conclude from $Q_\ell^+(A)=Q_\ell^-(A)$ that $M^{\otimes n}(\{f\neq 0\}\cap\mathcal{X}_\ell^n)=0$ for all ℓ. As $\ell\to\infty$, we conclude that $M^{\otimes n}(f\neq 0)=0$. □

Special case 3. Let $\mathcal{X} = \mathbb{R}$, equipped with its Borel σ-field. Suppose Θ consists of distributions with a density with respect to Lebesgue measure, and for an arbitrary fixed $\sigma > 0$, suppose that Θ contains all finite convex combinations of Gaussian distributions $\mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$. Then $\mathcal{E}_o$ is complete.

Proof: With the same arguments as in the previous special cases, one can show that for any $\mathcal{A}_o$-measurable function $f : \mathbb{R}^n \to \mathbb{R}$, it follows from $\int f \,\mathrm{d}P^{\otimes n} = 0$ for arbitrary $P \in \Theta$ that

$$\int f(\boldsymbol{x}) \prod_{i=1}^{n} \phi_{\mu_i,\sigma}(x_i) \,\mathrm{d}\boldsymbol{x} \;=\; 0 \quad \text{for arbitrary } \boldsymbol{\mu} \in \mathbb{R}^n,$$

where $\phi_{\mu,\sigma}(x) = (2\pi\sigma^2)^{-1/2} \exp\bigl(-(x-\mu)^2/(2\sigma^2)\bigr)$. This is equivalent to

$$\int f(\boldsymbol{x}) \exp\bigl(-\|\boldsymbol{x}\|^2/(2\sigma^2)\bigr) \exp(\boldsymbol{\mu}^\top \boldsymbol{x}) \, d\boldsymbol{x} \;=\; 0 \quad \text{for arbitrary } \boldsymbol{\mu} \in \mathbb{R}^n.$$

But then it follows from Theorem A.5 in the Appendix that the finite measures

$$\mathrm{Borel}(\mathbb{R}^n) \ni A \;\mapsto\; \int_A f^{\pm}(\boldsymbol{x}) \exp\bigl(-\|\boldsymbol{x}\|^2/(2\sigma^2)\bigr) \,\mathrm{d}\boldsymbol{x}$$

are identical, whence $f(\boldsymbol{x}) = 0$ for Lebesgue-almost all $\boldsymbol{x} \in \mathbb{R}^n$. □

Unbiased estimation. With the concepts of sufficiency and completeness one can say something about unbiased point estimators in the sense of (3.3). This was first noted by Rao, P.R. Halmos [10] and later extended by numerous authors including Blackwell [2], Kolmogorov [13], Lehmann and Scheffé [15, 16].

Theorem 3.32. *Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ be a statistical experiment with a sufficient statistic $T : (\mathcal{X}, \mathcal{B}) \to (\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$. Consider some function $g : \Theta \to \mathbb{R}^q$ and a loss function $L : \mathbb{R}^q \times \Theta \to [0, \infty)$ such that $L(\cdot, \theta)$ is convex and coercive[2] for any fixed $\theta \in \Theta$.*

(a) *Suppose that $\widehat{g} : \mathcal{X} \to \mathbb{R}^q$ is an estimator with finite risk function $\theta \mapsto E_\theta(L(\widehat{g}, \theta))$. Then*

$$\breve{g}(x) \;:=\; E(\widehat{g} \,|\, T = T(x))$$

defines another unbiased estimator such that for any $\theta \in \Theta$,

$$E_\theta(\breve{g}) \;=\; E_\theta(\widehat{g}) \quad \text{and} \quad E_\theta L(\breve{g}, \theta) \;\le\; E_\theta L(\widehat{g}, \theta).$$

If $L(\cdot, \theta)$ is strictly convex, then the latter inequality is strict unless $P_\theta(\widehat{g} \neq \breve{g}) = 0$.

(b) *Suppose $\widehat{g}$ is unbiased for $g(\theta)$, that is, $E_\theta(\widehat{g}) = g(\theta)$ for all $\theta \in \Theta$. Then the estimator $\breve{g}$ is unbiased too. If $\mathcal{E}^T = (\tilde{\mathcal{X}}, \tilde{\mathcal{B}}, (P_\theta^T)_{\theta\in\Theta})$ is complete, $\breve{g}$ is essentially the uniqe unbiased estimator which is a function of T. That is, if $\tilde{g} = \tilde{h}(T)$ is another unbiased estimator of $g(\theta)$, then $P_\theta(\tilde{g} \neq \breve{g}) = 0$ for arbitrary $\theta \in \Theta$.*

[2] $L(v, \theta) \to \infty$ as $\|v\| \to \infty$

Proof of Theorem 3.32. By sufficiency of T, there exists a stochastic kernel K from $(\tilde{\mathcal{X}}, \tilde{\mathcal{B}})$ to $(\mathcal{X}, \mathcal{B})$ describing the conditional distribution of $X \sim P_\theta$, given $T(X)$, for any parameter $\theta \in \Theta$. This gives rise to conditional expectations

$$E(h \,|\, T = t) \;:=\; \int h(x)\, K(t, \mathrm{d}x).$$

As to part (a), note that for any $\theta \in \Theta$, $R(\widehat{g}, \theta) < \infty$ implies that

$$\infty \;>\; E_\theta \|\widehat{g}\| \;=\; \int_{\tilde{\mathcal{X}}} E\big(\|\widehat{g}\| \,\big|\, T = t\big)\, P_\theta^T(\mathrm{d}t),$$

see also the proof of Remark 3.3. Hence, the set $\tilde{N} := \big\{t \in \tilde{\mathcal{X}} : E\big(\|\widehat{g}\| \,\big|\, T = t\big) = \infty\big\}$ satisfies $P_\theta^T(\tilde{N}) = 0$. Consequently, we may redefine $K(t, \cdot) := \delta_{x_o}$ for some $x_o \in \mathcal{X}$ whenever $t \in \tilde{N}$. Then,

$$\check{h}(t) \;:=\; E(\widehat{g} \,|\, T = t)$$

is well-defined in $\mathbb{R}^q$ for any $t \in \tilde{\mathcal{X}}$. With this function $\check{h}$ we define $\check{g} = \check{h}(T)$. Note that

$$E_\theta(\widehat{g}) \;=\; \int_{\tilde{\mathcal{X}}} E(\widehat{g} \,|\, T = t)\, P_\theta^T(\mathrm{d}t) \;=\; \int_{\tilde{\mathcal{X}}} \check{h}(t)\, P_\theta^T(\mathrm{d}t) \;=\; E_\theta(\check{h}(T)) \;=\; E_\theta(\check{g}).$$

Moreover, applying Jensen's inequality to the conditional expectations $E(\cdot \,|\, T = t)$ leads to

$$\begin{aligned} E_\theta L(\widehat{g}, \theta) \;&=\; \int_{\tilde{\mathcal{X}}} E\big(L(\widehat{g}, \theta) \,\big|\, T = t\big)\, P_\theta^T(\mathrm{d}t) \\ &\geq\; \int_{\tilde{\mathcal{X}}} L\big(E(\widehat{g} \,|\, T = t), \theta\big)\, P_\theta^T(\mathrm{d}t) \\ &=\; \int_{\tilde{\mathcal{X}}} L\big(\check{h}(t), \theta\big)\, P_\theta^T(\mathrm{d}t) \\ &=\; E_\theta L(\check{g}, \theta). \end{aligned}$$

If $L(\cdot, \theta)$ is strictly convex, the inequality

$$E\big(L(\widehat{g}, \theta) \,\big|\, T = t\big) \;\geq\; L\big(E(\widehat{g} \,|\, T = t), \theta\big)$$

is strict, unless $P(\widehat{g} \neq \check{g} \,|\, T = t) = 0$. Hence $E_\theta L(\widehat{g}, \theta) > E_\theta L(\check{g}, \theta)$, unless $P_\theta(\widehat{g} \neq \check{g}) = 0$.

As to part (b), suppose that $\check{g} = \check{h}(T)$ and $\tilde{g} = \tilde{h}(T)$ are unbiased estimators of $g(\theta)$. Then the difference $\Delta := \tilde{h} - \check{h}$ satisfies

$$\int_{\tilde{\mathcal{X}}} \Delta \,\mathrm{d}P_\theta^T \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

But then completeness of $\mathcal{E}^T$ implies that $0 = P_\theta^T(\Delta \neq 0) = P_\theta(\tilde{g} \neq \check{g})$ for arbitrary $\theta \in \Theta$. $\square$

Exercise 3.33. Consider the statistical experiment in Exercise 3.22.

(a) Show that $(P_{a,b}^T)_{(a,b)\in\Theta}$ is complete.

(b) Show that $E_{a,b}(T_2 - T_1) = (b - a + 2) \cdot (n - 1)/(n + 1)$.

Proposal 1: For $\boldsymbol{X} \sim P_{a,b}$, consider its order statistics $a \leq X_{(1)} < X_{(2)} < \cdots < X_{(n)} \leq b$. Then show that $(Y_j)_{j=1}^{n+1}$ with $Y_j := X_{(j)} - X_{(j-1)}$, $X_{(0)} := a - 1$, $X_{(n+1)} := b + 1$ is uniformly distributed on the set of all $\boldsymbol{y} \in \mathbb{N}^{n+1}$ with $y_+ = b - a + 2$.

Proposal 2: One could also determine the distribution of $T_2 - T_1$ under $P_{a,b}$ and then compute its expectation, using the formula $\sum_{\ell=n}^{N} [\ell]_n = [N+1]_{n+1}/(n+1)$.

(c) Determine an optimal unbiased estimator $\widehat{g}$ of $g(a,b) := \#\{a,\dots,b\} = b - a + 1$.

Exercise 3.34. For $\theta > 0$, let $P_\theta := \mathrm{Unif}[0,\theta]^{\otimes n}$.

(a) Show that $T(\boldsymbol{x}) := \max(x_1,\dots,x_n)$ is a sufficient statistic for $(P_\theta)_{\theta>0}$.

(b) Show that $(P_\theta^T)_{\theta>0}$ is complete.

(c) Determine an optimal unbiased estimator $\widehat{\theta} : [0,\infty)^n \to \mathbb{R}$ of θ.

(d) Show that the estimator $\widehat{\theta}$ from part (c) is not admissible with respect to mean squared error. *Hint:* Consider the estimators $\widehat{\theta}_\lambda := \lambda T$, $\lambda > 0$.

3.5 U-Statistics

The material in this section is based on the famous paper [11] by W. Hoeffding. Let $X_1,\dots,X_n$ be independent random variables with unknown distribution P on a measurable space $(\mathcal{X},\mathcal{B})$. Let Θ be a given family of probability distributions on $(\mathcal{X},\mathcal{B})$. Assuming that the unknown distribution P belongs to Θ, the corresponding statistical experiment is

$$\big(\mathcal{X}^n, \mathcal{B}^{\otimes n}, (P^{\otimes n})_{P\in\Theta}\big).$$

Note that all distributions $P^{\otimes n}$, $P \in \Theta$, are exchangeable (permutation-invariant). Hence for any given function $g : \Theta \to \mathbb{R}^q$, an unbiased point estimator $\widehat{g} : \mathcal{X}^n \to \mathbb{R}^q$ of $g(P)$ can be improved by replacing $\widehat{g}(x_1,x_2,\dots,x_n)$ with

$$\frac{1}{n!}\sum_{\pi\in\mathcal{S}_n} \widehat{g}(x_{\pi(1)}, x_{\pi(2)},\dots,x_{\pi(n)}).$$

And if Θ is sufficiently rich, the latter estimator is essentially unique among all permutation-invariant estimators, see Example 3.31 and Theorem 3.32.

U-Statistics. Now we consider a particular type of parameter $g(P) \in \mathbb{R}$. Let $h_{\mathrm{raw}} : \mathcal{X}^m \to \mathbb{R}$ be a given measurable function such that

$$g(P) \;:=\; \int h_{\mathrm{raw}}\, dP^{\otimes m} \;\in\; \mathbb{R}$$

for all $P \in \Theta$. Obviously, for $n \geq m$, a naive unbiased estimator for $g(P)$ is given by $\widehat{g}(\boldsymbol{x}) := h_{\mathrm{raw}}(x_1,\dots,x_m)$, where $\boldsymbol{x} = (x_i)_{i=1}^n$. Averaging this naive estimator over all permutations of $\boldsymbol{x}$ yields the following unbiased estimator of $g(P)$:

$$\breve{g}_n(\boldsymbol{x}) \;:=\; \frac{1}{[n]_m} \sum_{i_1,\dots,i_m=1}^{n} 1_{[i_1,\dots,i_m \text{ different}]} h_{\mathrm{raw}}(x_{i_1},\dots,x_{i_m}), \tag{3.7}$$

where $[n]_m = \prod_{i=0}^{m-1}(n-i)$. Such an estimator is called a *U-statistic of order m with (general) kernel h_{raw}*.

In case of $m \geq 2$, we may restrict our attention to functions $h : \mathcal{X}^m \to \mathbb{R}$ which are permutation-invariant, i.e. symmetric in its m arguments. Indeed, we could replace $h_{\mathrm{raw}}(x_1, \ldots, x_m)$ with its symmetrization

$$h(x_1, \ldots, x_m) \;:=\; \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} h_{\mathrm{raw}}(x_{\pi(1)}, \ldots, x_{\pi(m)}).$$

This is a permutation-invariant function satisfying $\int h \, dP^{\otimes m} = \int h_{\mathrm{raw}} \, dP^{\otimes m} = g(P)$. Then the estimator $\check{g}_n$ in (3.7) can be written as

$$\check{g}_n(\boldsymbol{x}) \;=\; \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} h(x_{i_1}, \ldots, x_{i_m}). \tag{3.8}$$

Such an estimator is called a *U-statistic of order m with (symmetric) kernel h*.

One could choose Θ to be the family of *all* distributions P such that $\int |h| \, \mathrm{d}P^{\otimes m} < \infty$. Then it contains all discrete distributions with finite support. Hence the results of the previous section show that $\check{g}_n$ is the *unique* permutation-invariant and measurable function on $\mathcal{X}^n$ such that $\mathbb{E}\, \check{g}_n(\boldsymbol{X}) = g(P)$ for all $P \in \Theta$.

Example 3.35 (Mean and variance of a distribution and sample). Let $\mathcal{X} = \mathbb{R}$, and suppose that $\int |x| \, P(dx) < \infty$. Then the mean $\mu(P) := \int x \, P(dx)$ equals $g(P)$ with $h : \mathbb{R} \to \mathbb{R}$, $h(x) = x$, and the sample mean is a U-statistic of order 1:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \;=\; \binom{n}{1}^{-1} \sum_{1 \leq i \leq n} h(x_i).$$

In case of $\int x^2 \, P(dx) < \infty$ one may write

$$\sigma^2(P) := \int (x - \mu(P))^2 \, P(dx) \;=\; \int_{\mathbb{R}^2} h \, dP^{\otimes 2} \quad \text{with} \quad h(x_1, x_2) \;:=\; (x_1 - x_2)^2/2.$$

The corresponding U-statistic is just the usual sample variance:

$$\begin{aligned}
\binom{n}{2}^{-1} \sum_{1 \leq i_1 < i_2 \leq n} (x_{i_1} - x_{i_2})^2/2 \;&=\; \frac{1}{n(n-1)} \sum_{i,j=1}^{n} (x_i - x_j)^2/2 \\
&=\; \frac{1}{n(n-1)} \Big(n \sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i \Big)^2 \Big) \\
&=\; \frac{1}{n-1} \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \Big) \\
&=\; \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\end{aligned}$$

Example 3.36. Let $\mathcal{X} = [0, \infty)$, and suppose that we want to quantify whether P has strong right tails. On possibility to quantify this property would be consider

$$g(P) \;:=\; \mathbb{P}(X_3 > X_1 + X_2).$$

(Maybe this is not such a brilliant proposal; the main point is to illustrate the construction of U-statistics.) This corresponds to $\mathbb{E}\, h_{\mathrm{raw}}(X_1, X_2, X_3)$ with $h_{\mathrm{raw}}(x_1, x_2, x_3) \;:=\; 1_{[x_3 > x_1 + x_2]}$.

Symmetrizing this kernel leads to

$$h(x_1, x_2, x_3) \;=\; 3^{-1} 1_{[2\max(x_1,x_2,x_3) > x_1+x_2+x_3]},$$

and the resulting unbiased estimator $\breve{g}_n$ for $g(P)$ would be the U-statistic of order three with this symmetric kernel h. One could also start from the non-symmetric kernel $h_{\mathrm{raw}}(x_1, x_2, x_3)$. This would lead to the alternative formula

$$\breve{g}_n(\boldsymbol{x}) \;=\; \frac{2}{[n]_3} \sum_{1 \le i < j < k \le n} 1_{[x_{(i)} + x_{(j)} < x_{(k)}]}$$

with the order statistics $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ of $\boldsymbol{x} \in \mathbb{R}^n$.

Here is a first result about the variance of a U-statistic which will suffice for our purposes:

Lemma 3.37 (Hoeffding). *Suppose that $\int h^2 \, dP^{\otimes m} < \infty$. Let*

$$\begin{aligned} h_m(x_1, \ldots, x_m) &:= h(x_1, \ldots, x_m) - g(P), \\ h_k(x_1, \ldots, x_k) &:= \mathbb{E}\, h_m(x_1, \ldots, x_k, X_{k+1}, \ldots, X_m) \quad \text{for } 1 \le k < m, \\ \sigma_k^2 &:= \mathbb{E}\big(h_k(X_1, \ldots, X_k)^2\big) \quad \text{for } 1 \le k \le m. \end{aligned}$$

Then $\sigma_1^2 \le \cdots \le \sigma_m^2$, and

$$\mathrm{Var}(\breve{g}_n(\boldsymbol{X})) \;=\; \sum_{k=1}^{m} \mathbb{P}(Y = k)\, \sigma_k^2$$

with $Y \sim \mathrm{Hyp}(n, m, m)$, i.e. $\mathbb{P}(Y = k) = \binom{m}{k}\binom{n-m}{m-k} \big/ \binom{n}{m}$.

This lemma is only a simplified version of Hoeffding's [11] findings. Stronger statements about the variances σ_k^2 and the distribution of $\breve{g}_n(\boldsymbol{X})$ can be deduced from the so-called Hoeffding decomposition (Section A.4), see Exercise 3.46.

Proof of Lemma 3.37. For any index set $J \subset \{1, \ldots, n\}$ with $1 \le \#J \le m$ let $X_J := (X_i)_{i \in J}$. With $\mathcal{J}_{n,m}$ denoting the set of all index sets $J \subset \{1, 2, \ldots, n\}$ with m elements, we may

$$\breve{g}_n(\boldsymbol{X}) - g(P) \;=:\; U_n \;=\; \binom{n}{m}^{-1} \sum_{J \subset \mathcal{J}_{n,m}} h_m(X_J).$$

It follows from independence of the X_i, Fubini's theorem, symmetry of h and the definition of h_k, $1 \le k \le m$, that

$$\mathbb{E}\big(h_m(X_I) h_m(X_J)\big) \;=\; \sigma^2_{\#(I \cap J)} \quad \text{for } I, J \in \mathcal{J}_{n,m}, \tag{3.9}$$

where $\sigma_0^2 := 0$. Indeed, $\mathbb{E}\, h_m(X_1, \ldots, X_m) = 0$ by definition of h_m, so in case of $I \cap J = \emptyset$, it follows from stochastic independence of X_I and X_J that

$$\mathbb{E}\big(h_m(X_I) h_m(X_J)\big) \;=\; \mathbb{E}(h_m(X_I))\, \mathbb{E}(h_m(X_J)) \;=\; 0.$$

Moreover, by definition of σ_m^2,

$$\mathbb{E}\big(h_m(X_I) h_m(X_J)\big) \;=\; \sigma_m^2 \quad \text{if } I = J.$$

If $1 \le k := \#(I \cap J) < m$, then stochastic independence of $X_{I\cap J}, X_{I\setminus J}, X_{J\setminus I}$ and the definition of h_k imply that

$$\begin{aligned}
\mathbb{E}\big(h_m(X_I)h_m(X_J)\big) &= \mathbb{E}\big(h_m(X_{I\cap J}, X_{I\setminus J})h_m(X_{I\cap J}, X_{J\setminus I})\big) \\
&= \mathbb{E}\,\mathbb{E}\big(h_m(X_{I\cap J}, X_{I\setminus J})h_m(X_{I\cap J}, X_{J\setminus I}) \,\big|\, X_{I\cap J}\big) \\
&= \mathbb{E}\big[\mathbb{E}(h_m(X_{I\cap J}, X_{I\setminus J}) \,|\, X_{I\cap J})\,\mathbb{E}(h_m(X_{I\cap J}, X_{J\setminus I}) \,|\, X_{I\cap J})\big] \\
&= \mathbb{E}\big[h_k(X_{I\cap J})^2\big] \\
&= \sigma_k^2.
\end{aligned}$$

Equation (3.9) yields the specific formula for $\mathrm{Var}(\check{g}_n(\boldsymbol{X}))$, because

$$\begin{aligned}
\mathrm{Var}(\check{g}_n(\boldsymbol{X})) = \mathbb{E}(U_nU_n) &= \binom{n}{m}^{-2} \sum_{I,J\subset\mathcal{J}_{n,m}} \mathbb{E}\big(h_m(X_I)h_m(X_J)\big) \\
&= \binom{n}{m}^{-2} \sum_{I,J\subset\mathcal{J}_{n,m}} \sigma^2_{\#(I\cap J)} \\
&= \binom{n}{m}^{-2} \sum_{k=1}^{m} \#\big\{(I,J) \in \mathcal{J}_{n,m}^2 : \#(I\cap J) = k\big\}\sigma_k^2 \\
&= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\sigma_k^2 \\
&= \sum_{k=1}^{m} \mathbb{P}(Y = k)\sigma_k^2.
\end{aligned}$$

As to the second last step, to choose a pair (I, J) of sets in $\mathcal{J}_{n,m}$ such that $\#(I \cap J) = k$, there are $\binom{n}{m}$ possibilities for I, and for given I there are $\binom{m}{k}\binom{n-m}{m-k}$ possibilities for J.

It remains to prove the inequality $\sigma_k^2 \le \sigma_{k+1}^2$ for $1 \le k < m$. Note first that

$$h_k(x_1, \ldots, x_k) = \mathbb{E}\, h_{k+1}(x_1, \ldots, x_k, X_{k+1}).$$

In case of $k = m-1$ this is just the definition of $h_k(x_1, \ldots, x_k)$, and otherwise it is a consequence of Fubini's theorem, because

$$\begin{aligned}
h_k(x_1, \ldots, x_k) &= \mathbb{E}\, h_m(x_1, \ldots, x_k, X_{k+1}, X_{k+2}, \ldots, X_m) \\
&= \mathbb{E}\,\mathbb{E}\big(h_m(x_1, \ldots, x_k, X_{k+1}, X_{k+2}, \ldots, X_m) \,\big|\, X_{k+1}\big) \\
&= \mathbb{E}\, h_{k+1}(x_1, \ldots, x_k, X_{k+1}).
\end{aligned}$$

But then the Cauchy–Schwarz inequality implies that

$$\begin{aligned}
\sigma_k^2 &= \mathbb{E}\big(\mathbb{E}\big(h_{k+1}(X_1, \ldots, X_k, X_{k+1}) \,\big|\, X_1, \ldots, X_k\big)^2\big) \\
&\le \mathbb{E}\big(\mathbb{E}\big(h_{k+1}(X_1, \ldots, X_k, X_{k+1})^2 \,\big|\, X_1, \ldots, X_k\big)\big) \\
&= \mathbb{E}\big(h_{k+1}(X_1, \ldots, X_k, X_{k+1})^2\big) \\
&= \sigma_{k+1}^2.
\end{aligned}$$

□

Corollary 3.38. *In case of* $\int h^2\, dP^{\otimes m} < \infty$,

$$\mathrm{Var}(\check{g}_n(\boldsymbol{X})) \;=\; \frac{m^2\sigma_1^2}{n} + O(n^{-2}).$$

Proof of Corollary 3.38. This expansion follows immediately from Lemma 3.37 and the fact that for $Y \sim \mathrm{Hyp}(n, m, m)$, the probability $\mathbb{P}(Y = k)$ equals

$$\begin{aligned}
\binom{n}{m}^{-1}\binom{m}{k}\binom{n-m}{m-k} \;&=\; \frac{m![m]_k\,[n-m]_{m-k}}{[n]_m\,k!(m-k)!} \\
&=\; \frac{[m]_k^2\,n^{m-k}(1+O(n^{-1}))}{k!\,n^m(1+O(n^{-1}))} \\
&=\; \frac{[m]_k^2(1+O(n^{-1}))}{k!\,n^k} \\
&=\; \begin{cases} m^2/n + O(n^{-2}) & \text{if } k = 1, \\ O(n^{-2}) & \text{if } k \geq 2. \end{cases}
\end{aligned}$$

□

The following result shows that U-statistics may be approximated by an average of independent, identically distributed random variables and satisfy a Central Limit Theorem:

Theorem 3.39 (Hoeffding). *Under the conditions of Lemma 3.37,*

$$\check{g}_n(\boldsymbol{X}) \;=\; g(P) + \frac{m}{n}\sum_{i=1}^{n} h_1(X_i) + R_n$$

where

$$\mathbb{E}(R_n^2) \;=\; O(n^{-2}).$$

Moreover,

$$\sqrt{n}(\check{g}_n(\boldsymbol{X}) - g(P)) \;\to_{\mathcal{L}}\; \mathcal{N}(0, m^2\sigma_1^2)$$

as $n \to \infty$.

Our proof of Theorem 3.39 utilizes a general approximation result of Jaroslav Hájek:

Lemma 3.40 (Hájek projection). *Let* $X_1, X_2, \ldots, X_n$ *be arbitrary independent random variables with values in* $(\mathcal{X}_1, \mathcal{B}_1), (\mathcal{X}_2, \mathcal{B}_2), \ldots, (\mathcal{X}_n, \mathcal{B}_n)$, *and let* T *be a random variable of the form* $T = f(X_1, X_2, \ldots, X_n)$ *such that* $\mathbb{E}(T) = 0$ *and* $\mathbb{E}(T^2) < \infty$. *Then* $T_i := \mathbb{E}(T \mid X_i)$ *satisfies* $\mathbb{E}(T_i) = 0$ *for* $1 \leq i \leq n$. *Moreover, for arbitrary random variables* $Y_1, Y_2, \ldots, Y_n$ *of type* $Y_i = f_i(X_i)$ *with* $\mathbb{E}(Y_i) = 0$ *and* $\mathbb{E}(Y_i^2) < \infty$,

$$\begin{aligned}
\mathbb{E}\Big(\Big(T - \sum_{i=1}^{n} Y_i\Big)^2\Big) \;&=\; \mathbb{E}(T^2) - \sum_{i=1}^{n}\mathbb{E}(T_i^2) + \sum_{i=1}^{n}\mathbb{E}((Y_i - T_i)^2) \\
&\geq\; \mathbb{E}(T^2) - \sum_{i=1}^{n}\mathbb{E}(T_i^2)
\end{aligned}$$

with equality if and only if $Y_i = T_i$ *almost surely for* $1 \leq i \leq n$.

Proof of Lemma 3.40. By Fubini's theorem, $0 = \mathbb{E}(T) = \mathbb{E}(\mathbb{E}(T \mid X_i)) = \mathbb{E}(T_i)$. Moreover,

$$\begin{aligned}
\mathbb{E}\Big(\Big(T - \sum_{i=1}^n Y_i\Big)^2\Big) &= \mathbb{E}(T^2) - 2\sum_{i=1}^n \mathbb{E}(TY_i) + \sum_{i,j=1}^n \mathbb{E}(Y_iY_j) \\
&= \mathbb{E}(T^2) - 2\sum_{i=1}^n \mathbb{E}(\mathbb{E}(TY_i \mid X_i)) + \sum_{i=1}^n \mathbb{E}(Y_i^2) \\
&= \mathbb{E}(T^2) - 2\sum_{i=1}^n \mathbb{E}\big(\mathbb{E}(T \mid X_i)Y_i\big) + \sum_{i=1}^n \mathbb{E}(Y_i^2) \\
&= \mathbb{E}(T^2) - 2\sum_{i=1}^n \mathbb{E}(T_iY_i) + \sum_{i=1}^n \mathbb{E}(Y_i^2) \\
&= \mathbb{E}(T^2) - \sum_{i=1}^n \mathbb{E}(T_i^2) + \sum_{i=1}^n \mathbb{E}((Y_i - T_i)^2),
\end{aligned}$$

because $\mathbb{E}(Y_iY_j) = \mathbb{E}(Y_i)\,\mathbb{E}(Y_j) = 0$ in case of $i \neq j$. □

Proof of Theorem 3.39. It follows from Lemma 3.40 that

$$\check{g}_n(\boldsymbol{X}) - g(P) \;=\; U_n \;=\; \sum_{i=1}^n \mathbb{E}(U_n \mid X_i) + R_n$$

with

$$\mathbb{E}(R_n^2) \;=\; \mathbb{E}(U_n^2) - \sum_{i=1}^n \mathbb{E}\big(\mathbb{E}(U_n \mid X_i)^2\big).$$

But

$$\begin{aligned}
\mathbb{E}(U_n \mid X_i) &= \binom{n}{m}^{-1} \sum_{J \subset \{1,\ldots,n\}:\#J=m} \mathbb{E}\big(h_m(X_J) \mid X_i\big) \\
&= \binom{n}{m}^{-1} \sum_{J \subset \{1,\ldots,n\}:\#J=m} 1_{[i \in J]} h_1(X_i) \\
&= \binom{n}{m}^{-1} \#\big\{J \subset \{1,\ldots,n\} : \#J = m, i \in J\big\}\, h_1(X_i) \\
&= \binom{n}{m}^{-1} \binom{n-1}{m-1} h_1(X_i) \\
&= \frac{m}{n}\, h_1(X_i).
\end{aligned}$$

Hence

$$U_n \;=\; \frac{m}{n} \sum_{i=1}^n h_1(X_i) + R_n$$

with

$$\mathbb{E}(R_n^2) \;=\; \mathbb{E}(U_n^2) - \frac{m^2\sigma_1^2}{n} \;=\; \mathrm{Var}(\check{g}_n(\boldsymbol{X})) - \frac{m^2\sigma_1^2}{n} \;=\; O(n^{-2})$$

according to Corollary 3.38. Consequently, $R_n = O_p(n^{-1})$, whence

$$\sqrt{n}(\check{g}_n(\boldsymbol{X}) - g(P)) \;=\; \frac{m}{\sqrt{n}} \sum_{i=1}^n h_1(X_i) + O_p(n^{-1/2}).$$

It follows from the Central Limit Theorem that the right hand side converges in distribution to a Gaussian random variable with mean 0 and variance $m^2\,\mathbb{E}(h_1(X_1)^2) = m^2\sigma_1^2$ as $n \to \infty$. $\square$

Remark 3.41 (Hoeffding's decomposition in case of $m = 2$). In case of $m = 2$ one may write

$$\check{g}_n(\boldsymbol{X}) \;=\; g(P) + \frac{2}{n}\sum_{i=1}^{n} h_1(X_i) + \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_2^o(X_i, X_j)$$

with

$$\begin{aligned} h_2^o(x_1,x_2) \;&:= h_2(x_1,x_2) - h_1(x_1) - h_1(x_2) \\ &= h(x_1,x_2) - \mathbb{E}\,h(x_1,X_2) - \mathbb{E}\,h(X_1,x_2) + \mathbb{E}\,h(X_1,X_2). \end{aligned}$$

Moreover, the $n + \binom{n}{2}$ random variables $h_1(X_i)$ $(1 \le i \le n)$ and $h_2^o(X_i, X_j)$ $(1 \le i < j \le n)$ are easily shown to be centered and uncorrelated with

$$\mathbb{E}\big(h_2^o(X_i,X_j)^2\big) \;=\; \sigma_2^2 - 2\sigma_1^2 \;\le\; \sigma_2^2.$$

Hence the remainder

$$R_n \;=\; \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_2^o(X_i, X_j)$$

satisfies the (in)equalities

$$\mathbb{E}(R_n^2) \;=\; \binom{n}{2}^{-1}(\sigma_2^2 - 2\sigma_1^2) \;=\; \frac{2\sigma_2^2 - 4\sigma_1^2}{n(n-1)} \;\le\; \frac{2\sigma_2^2}{n(n-1)}.$$

Example 3.35 (Sample variance, continued). With $h(x_1,x_2) = (x_1 - x_2)^2/2$, the auxiliary function h_1 is given by

$$\begin{aligned} h_1(x) \;&=\; \mathbb{E}\,h(x,X_1) - \sigma^2(P) \\ &=\; x^2/2 - x\mu(P) + \mathbb{E}(X_1^2)/2 - \sigma^2(P) \\ &=\; \big[(x-\mu(P))^2 - \sigma^2(P)\big]/2, \end{aligned}$$

and this leads to the representation

$$\begin{aligned} S_X^2 \;&=\; \binom{n}{2} \sum_{1\le i<j\le n} h(X_i,X_j) \\ &=\; \sigma^2(P) + \frac{2}{n}\sum_{i=1}^{n} h_1(X_i) + R_n \\ &=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu(P))^2 + R_n, \end{aligned}$$

where

$$\mathbb{E}(R_n^2) \;=\; O(n^{-2}) \quad \text{if } \int x^4\,P(dx) < \infty.$$

Moreover, as $n \to \infty$,

$$\sqrt{n}(S_X^2 - \sigma(P)^2) \;\to_{\mathcal{L}}\; \mathcal{N}\big(0, \mathbb{E}[(X_1 - \mu(P))^4] - \sigma(P)^4\big).$$

Exercise 3.42. Extend our considerations about sample variances to the sample covariance

$$(n-1)^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

of independent random pairs $(X_1,Y_1),\dots,(X_n,Y_n)$ with unknown distribution on $\mathbb{R}\times\mathbb{R}$.

Example 3.43 (Kendall's τ). Let (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n) be independent random pairs with distribution P on $\mathbb{R}\times\mathbb{R}$. A nonparametric measure of correlation of X_1 and Y_1, proposed by Maurice Kendall [12], is given by

$$\tau(P) \;:=\; \mathbb{E}\big(\operatorname{sign}(X_2-X_1)\operatorname{sign}(Y_2-Y_1)\big).$$

This is the probability, that the two observation pairs (X_1,Y_1) and (X_2,Y_2) are "concordant", i.e.

$$\operatorname{sign}(X_2-X_1)=\operatorname{sign}(Y_2-Y_1)\neq 0,$$

minus the probability that they are "discordant", i.e.

$$-\operatorname{sign}(X_2-X_1)=\operatorname{sign}(Y_2-Y_1)\neq 0.$$

If X_1 and Y_1 are stochastically independent, then the four random variables X_1X_2, Y_1, Y_2 are stochastically independent, and

$$\tau(P) \;=\; \mathbb{E}\operatorname{sign}(X_2-X_1)\,\mathbb{E}\operatorname{sign}(Y_2-Y_1) \;=\; 0,$$

because the distributions of X_2-X_1 and Y_2-Y_1 are symmetric around 0.

Note that $\tau(P)=\mathbb{E}\,h\big((X_1,Y_1),(X_2,Y_2)\big)$ with the kernel

$$h\big((x_1,y_1),(x_2,y_2)\big) \;:=\; \operatorname{sign}(x_2-x_1)\operatorname{sign}(y_2-y_1).$$

Consequently, an unbiased estimator for $\tau(P)$ is given by Kendall's τ statistic

$$\check{\tau}_n(\boldsymbol{X}) \;:=\; \binom{n}{2}^{-1}\sum_{1\le i<j\le n}\operatorname{sign}(X_j-X_i)\operatorname{sign}(Y_j-Y_i),$$

which is a U-statistic of order 2 with kernel h.

Since $|h|\le 1$, we may apply Theorem 3.39 and conclude that

$$\check{\tau}_n(\boldsymbol{X}) \;=\; \tau(P)+\frac{2}{n}\sum_{i=1}^{n}h_1(X_i,Y_i)+R_n$$

with $h_1(x,y) \;:=\; \mathbb{E}\big(\operatorname{sign}(x-X_1)\operatorname{sign}(y-Y_1)\big)-\tau(P)$ and a remainder term R_n such that $\mathbb{E}(R_n^2)=O(n^{-2})$. Moreover, as $n\to\infty$,

$$\sqrt{n}\big(\check{\tau}_n(\boldsymbol{X})-\tau(P)\big) \;\to_{\mathcal{L}}\; \mathcal{N}\big(0,4\,\mathbb{E}[h_1(X_1,Y_1)^2]\big).$$

Let us consider the following special case: Suppose that X_1 and Y_1 are stochastically independent with continuous distribution function F and G, respectively. Then one can easily verify that

$$h_1(x,y) \;=\; (2F(x)-1)(2G(y)-1).$$

But $2F(X_1) - 1$ and $2G(Y_1) - 1$ are stochastically independent and uniformly distributed on $[-1, 1]$. From this one can easily deduce that

$$\mathbb{E}[h_1(X_1, Y_1)^2] \;=\; 1/9,$$

so

$$\sqrt{n}\,\check{\tau}_n(\boldsymbol{X}) \;\to_{\mathcal{L}}\; \mathcal{N}(0, 4/9).$$

Exercise 3.44. Let $X_1, X_2, \ldots, X_n$ be independent random variables with unknown distribution P on $\mathbb{R}$. For $m \in \mathbb{N}$ let

$$g_m(P) \;:=\; \mathbb{E}\,\mathrm{Med}(X_1, \ldots, X_m)$$

with the sample median function $\mathrm{Med}(\ldots)$. Show that for $n \geq m$, the corresponding U-statistic

$$\check{g}(\boldsymbol{x}) \;=\; \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} \mathrm{Med}(x_{i_1}, \ldots, x_{i_m})$$

is a L-statistic, that means,

$$\check{g}(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} w_i x_{(i)}$$

with suitable weights $w_1, w_2, \ldots, w_n \geq 0$ and the order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ of $\boldsymbol{x} \in \mathbb{R}^n$.

Hint: Distinguish the cases of odd and even m. Eventually you should see that for any integer $k \geq 0$, the estimators for $m = 2k + 1$ and $m = 2k + 2$ are identical!

Exercise 3.45. Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with distribution P on $\mathbb{R}$. Suppose further that $\mathbb{E}\big(|X_1|^3\big) < \infty$.

(a) Determine an optimal unbiased estimator of the centered third moment,

$$g(P) \;:=\; \mathbb{E}\big((X_1 - \mathbb{E}(X_1))^3\big).$$

Hint: Determine a measurable function $h : \mathbb{R}^3 \to \mathbb{R}$ (symmetric or not) such that $g(P) = \mathbb{E}\, h(X_1, X_2, X_3)$, and construct a corresponding U-statistic $\check{g}_n(\boldsymbol{X})$.

(b) A naive estimator for $g(P)$ is given by

$$\widehat{g}_n(\boldsymbol{X}) \;:=\; \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^3.$$

Show that this estimator can be written as a function of the sums $S_\ell := \sum_{i=1}^n X_i^\ell$, $1 \leq \ell \leq 3$. In particular, the computation of $\widehat{g}_n(\boldsymbol{X})$ requires $O(n)$ steps.

(c) Show that

$$\check{g}_n \;=\; \frac{n^2}{(n-1)(n-2)}\, \widehat{g}_n.$$

In particular, $\check{g}_n(\boldsymbol{X})$ can also be computed in $O(n)$ steps, and both estimators remain unchanged when $\boldsymbol{X}$ is replaced with $\boldsymbol{X} + c = (X_i + c)_{i=1}^n$ for any $c \in \mathbb{R}$.

(d) Determine

$$\lim_{n\to\infty} n\big(\mathbb{E}\,\widehat{g}_n(\boldsymbol{X}) - g(P)\big).$$

(e) Determine the asymptotic distribution of $\sqrt{n}\big(\check{g}_n(\boldsymbol{X}) - g(P)\big)$ and $\sqrt{n}\big(\widehat{g}_n(\boldsymbol{X}) - g(P)\big)$, assuming that $\mathbb{E}(X_1^6) < \infty$.

Proposals: In addition to the sums S_ℓ, the double sum $S_{21} := \sum_{i,j=1}^n 1_{[i\neq j]} X_i^2 X_j$ and the triple sum $S_{111} := \sum_{i,j,k=1}^n 1_{[i,j,k \text{ different}]} X_i X_j X_k$ might be useful. In part (e) it suffices to consider the case $\mathbb{E}(X_1) = 0$. This is justified by part (c).

Exercise 3.46 (Refinements via Hoeffding's decomposition). With the notation of Lemma 3.37 and its proof, let

$$h_k^o(x_1,\ldots,x_k) \;:=\; \sum_{\ell=1}^k (-1)^{k-\ell} \sum_{I\subset\{1,\ldots,k\}:\#I=\ell} h_\ell(x_I)$$

for $1 \le k \le m$. The general Hoeffding decomposition, presented in Section A.4, implies that the random variables $h_k^o(X_J)$, where $J \subset \{1,\ldots,n\}$ with $k = \#J \in \{1,\ldots,m\}$, are centered and uncorrelated, where

$$h_k(x_1,\ldots,x_k) \;=\; \sum_{\ell=1}^k \sum_{I\subset\{1,\ldots,k\}:\#I=\ell} h_\ell^o(x_I).$$

(a) Setting $\tau_k^2 := \mathbb{E}\big(h_k^o(X_1,\ldots,X_k)^2\big)$, show that $\sigma_k^2 = \mathbb{E}\big(h_k(X_1,\ldots,X_k)^2\big)$ is equal to

$$\sigma_k^2 \;=\; \sum_{\ell=1}^k \binom{k}{\ell} \tau_\ell^2.$$

Deduce from this representation that σ_k^2/k is non-decreasing in $k \in \{1,\ldots,m\}$.

(b) Show that the variance of $U_n = \check{g}_n(\boldsymbol{X}) - g(P)$ can be written as

$$\mathbb{E}(U_n^2) \;=\; \frac{m^2}{n} \sum_{k=1}^m \mathbb{P}(Z = k-1) \frac{\sigma_k^2}{k},$$

where $Z \sim \mathrm{Hyp}(n-1, m-1, m-1)$. Deduce from this and part (a) that

$$\frac{m^2}{n}\sigma_1^2 \;\le\; \mathbb{E}(U_n^2) \;\le\; \frac{m}{n}\sigma_m^2.$$

(c) Show that U_n equals

$$U_n \;=\; \sum_{k=1}^m \frac{[m]_k}{[n]_k} \sum_{J\subset\{1,\ldots,n\}:\#J=k} h_k^o(X_J),$$

and that

$$\mathbb{E}(U_n^2) \;=\; \sum_{k=1}^m \frac{[m]_k}{[n]_k} \binom{m}{k} \tau_k^2.$$

Show that $n\,\mathbb{E}(U_n^2)$ is non-increasing in $n \ge m$. Furthermore, show that

$$\mathbb{E}(U_n^2) \;\le\; \frac{m^2}{n}\sigma_1^2 + \frac{m^2}{n^2}(\sigma_m^2 - m\sigma_1^2).$$

Chapter 4

Exponential Families

4.1 Definitions and Basic Properties

Definition 4.1 (Exponential families). A statistical experiment $\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$ is called an *exponential family* if there exist a σ-finite measure M on $(\mathcal{X}, \mathcal{B})$, a measurable mapping $h : \mathcal{X} \to [0, \infty)$, a mapping $\alpha : \Theta \to \mathbb{R}^d$ and a measurable mapping $T : \mathcal{X} \to \mathbb{R}^d$ such that for any $\theta \in \Theta$,

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}M}(x) \;=\; h(x)\exp\big(\alpha(\theta)^\top T(x) - \kappa(\theta)\big)$$

with

$$\kappa(\theta) \;:=\; \log\int h\exp(\alpha(\theta)^\top T)\,\mathrm{d}M.$$

(In particular, we assume that $\int h\exp(\alpha(\theta)^\top T)\,\mathrm{d}M < \infty$ for all $\theta \in \Theta$.)

Definition 4.2 (Natural exponential families). For a given measure space $(\mathcal{X}, \mathcal{B}, M)$ and measurable mappings $h : \mathcal{X} \to [0, \infty)$, $T : \mathcal{X} \to \mathbb{R}^d$, the corresponding *natural exponential family* is given by $\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta_{\mathrm{nat}}}\big)$ with the *natural parameter space*

$$\Theta_{\mathrm{nat}} \;:=\; \Big\{\theta \in \mathbb{R}^d : \int h\exp(\theta^\top T)\,\mathrm{d}M < \infty\Big\},$$

and the probability distributions P_θ are given by

$$\begin{aligned}\frac{\mathrm{d}P_\theta}{\mathrm{d}M}(x) \;&:=\; h(x)\exp\big(\theta^\top T(x) - \kappa(\theta)\big),\\ \kappa(\theta) \;&:=\; \log\int h\exp(\theta^\top T)\,\mathrm{d}M.\end{aligned}$$

Example 4.3 (Gaussian samples). Let $\mathcal{X} = \mathbb{R}^n$, equipped with its Borel σ-field, let $\Theta = \mathbb{R} \times (0, \infty)$, and for $\theta = (\mu, \sigma) \in \Theta$, let

$$P_{\mu,\sigma} \;:=\; \mathcal{N}(\mu, \sigma^2)^{\otimes n}.$$

With M denoting Lebesgue measure on $\mathbb{R}^n$ times $(2\pi)^{-n/2}$, the density of $P_{\mu,\sigma}$ with respect to

M is given by

$$\begin{aligned}\frac{\mathrm{d}P_{\mu,\sigma}}{\mathrm{d}M} &= \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma\Big)\\ &= \exp\Big(\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i + \frac{-1}{2\sigma^2}\sum_{i=1}^n x_i^2 - n\log\sigma - \frac{n\mu^2}{2\sigma^2}\Big)\\ &= \exp\big(\alpha(\mu,\sigma)^\top T(\boldsymbol{x}) - \kappa(\mu,\sigma)\big)\end{aligned}$$

with

$$\begin{aligned}\alpha(\mu,\sigma) &:= \Big(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\Big)^\top,\\ T(\boldsymbol{x}) &:= \Big(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\Big)^\top,\\ \kappa(\mu,\sigma) &:= n\log\sigma + \frac{n\mu^2}{2\sigma^2}.\end{aligned}$$

Here one easily verifies that

$$\Theta_{\mathrm{nat}} = \mathbb{R}\times(-\infty,0).$$

Example 4.4 (Gamma distributions). Let $\mathcal{X} = (0,\infty)$, equipped with its Borel σ-field. For $a,b>0$ let $\mathrm{Gamma}(a,b)$ be the gamma distribution on $(0,\infty)$ with shape parameter $a>0$ and scale parameter $b>0$. That means, $\mathrm{Gamma}(a,b)$ has Lebesgue density

$$\begin{aligned}f_{a,b}(x) &= \frac{1}{\Gamma(a)b}\Big(\frac{x}{b}\Big)^{a-1}\exp\Big(-\frac{x}{b}\Big)\\ &= \exp\big(\alpha(a,b)^\top T(x) - \kappa(a,b)\big)\end{aligned}$$

with

$$\begin{aligned}\alpha(a,b) &:= (a-1,-1/b)^\top,\\ T(x) &:= (\log x, x)^\top,\\ \kappa(a,b) &:= a\log b + \log\Gamma(a).\end{aligned}$$

Here one easily verifies that

$$\Theta_{\mathrm{nat}} = (-1,\infty)\times(-\infty,0).$$

Remark 4.5 (Convexity and smoothness). Let $\mathcal{E}$ be a natural exponential family as in Definition 4.2. It follows from convexity of the exponential function that the set Θ_{nat} is a convex subset of $\mathbb{R}^d$. Moreover, if $f\in\bigcap_{\theta\in\Theta_{\mathrm{nat}}} L^1(P_\theta)$, then one can deduce from Exercise 4.7 below that the function $H:\Theta\to\mathbb{R}$,

$$H(\theta) := \int f\,\mathrm{d}P_\theta,$$

is twice continuously differentiable on the interior of Θ_{nat} with gradient

$$\nabla H(\theta) = \mathrm{Cov}_\theta(f,T) = \int fT\,\mathrm{d}P_\theta - H(\theta)\int T\,\mathrm{d}P_\theta.$$

Remark 4.6 (Sufficiency). Suppose that $\mathcal{E}$ is an exponential family as in Definition 4.1, where $(\mathcal{X}, d)$ is a separable and complete metric space and $\mathcal{B} = \mathrm{Borel}(\mathcal{X}, d)$. Then T is a sufficient statistic for $\mathcal{E}$. This follows immediately from Neyman's factorization criterion.

Exercise 4.7 (Basic considerations). This exercise is a multivariate version of Exercise 1.25. Let Θ be an open convex subset of Θ_{nat}, and let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function such that

$$\int |f| \exp(\theta^\top T) h \,\mathrm{d}M \;<\; \infty \quad \text{for all } \theta \in \Theta.$$

Then the function $L : \Theta \to \mathbb{R}$, $L(\theta) := \int f \exp(\theta^\top T) h \,\mathrm{d}M$, is infinitely often differentiable with partial derivatives

$$\frac{\partial^m}{\partial\theta_{\ell_1} \cdots \partial\theta_{\ell_m}} L(\theta) \;=\; \int f \prod_{i=1}^{m} T_{\ell_i} \exp(\theta^\top T) h \,\mathrm{d}M.$$

for $m \in \mathbb{N}$ and $\ell_1, \ldots, \ell_m \in \{1, \ldots, d\}$.

Proposal: Let $\theta^o \in \Theta$ and $\epsilon > 0$ such that $\Gamma(\theta^o, \epsilon) := \times_{i=1}^d [\theta_i^o - \epsilon, \theta_i^o + \epsilon] \subset \Theta$. Show first that for $\delta \in (0, \epsilon]$ and $\theta \in \Gamma(\theta^o, \delta)$,

$$\exp\bigl(\theta^\top T + (\epsilon - \delta)\|T\|_1\bigr) \;\le \sum_{\xi \in \{-1,1\}^d} \exp((\theta^o + \epsilon\xi)^\top T),$$

where $\|T\|_1 := \sum_{i=1}^d |T_i|$.

Theorem 4.8 (Completeness in exponential families). *Let $\mathcal{E}$ be an exponential family as in Definition 4.1. Suppose that the set $\{\alpha(\theta) : \theta \in \Theta\} \subset \mathbb{R}^d$ contains an interior point. Then the statistical model $\bigl(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (P_\theta^T)_{\theta\in\Theta}\bigr)$ is complete.*

Proof of Theorem 4.8. We may assume without loss of generality that $h \equiv 1$. Otherwise we could replace M with $\tilde{M}$, where $\tilde{M}(B) := \int_B h \,\mathrm{d}M$. Note that the image measures[1] M^T and P_θ^T on $\mathbb{R}^d$ satisfy

$$\frac{\mathrm{d}P_\theta^T}{\mathrm{d}M^T}(t) \;=\; \exp\bigl(\alpha(\theta)^\top t - \kappa(\theta)\bigr).$$

Hence for a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, the property

$$\int_{\mathbb{R}^d} f \,\mathrm{d}P_\theta^T \;=\; 0 \quad \text{for all } \theta \in \Theta$$

is equivalent to

$$\int_{\mathbb{R}^d} f(t) \exp(\alpha(\theta)^\top t)\, M^T(\mathrm{d}t) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

Since $\alpha(\Theta)$ contains an interior point, it follows from Theorem A.5 in Section A.3 that $f(t) = 0$ for M^T-almost all $t \in \mathbb{R}^d$. In particular, $P_\theta^T(f \neq 0) = 0$ for all $\theta \in \Theta$. □

[1] $M^T(B) := M(T \in B)$ and $P_\theta^T(B) := P_\theta(T \in B)$

4.2 Nuisance Parameters

In this section we consider statistical experiments $\mathcal{E} = \big(\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$ of the following type: Θ is a convex open subset of $\mathbb{R}^d \times \mathbb{R}$ with projections

$$\begin{aligned} N &:= \big\{\nu \in \mathbb{R}^d : (\nu,\gamma) \in \Theta \text{ for some } \gamma \in \mathbb{R}\big\}, \\ \Gamma &:= \big\{\gamma \in \mathbb{R} : (\nu,\gamma) \in \Theta \text{ for some } \nu \in \mathbb{R}^d\big\}. \end{aligned}$$

Each parameter $\theta = (\nu,\gamma) \in \Theta$ consists of a "nuisance parameter" $\nu \in N$ and a parameter $\gamma \in \Gamma$ of primary interest. The question is how to deal with the nuisance parameter ν if we are only interested in γ.

We assume that $\mathcal{E}$ is an exponential family with natural parametrization: There exist a σ-finite measure M on $(\mathcal{X},\mathcal{X})$ and measurable functions $S : (\mathcal{X},\mathcal{B}) \to \mathbb{R}^d$, $Y : (\mathcal{X},\mathcal{B}) \to \mathbb{R}$ such that for arbitrary $(\nu,\gamma) \in \Theta$,

$$\frac{\mathrm{d}P_{\nu,\gamma}}{\mathrm{d}M} \;=\; \exp\big(\nu^\top S + \gamma Y - \kappa(\nu,\gamma)\big)$$

with

$$\kappa(\nu,\gamma) \;:=\; \log \int \exp\big(\nu^\top S + \gamma Y\big)\,\mathrm{d}M \;<\; \infty.$$

Here the pair (S,Y) is a sufficient statistic, provided that $(\mathcal{X},d)$ is a separable and complete metric space, equipped with its Borel σ-field. Thus we may restrict our attention to decision procedures depending only on (S,Y). The following result shows that under the measure $P_{\nu,\gamma}$, the conditional distribution of Y, given that $S = s$, depends on s and the parameter γ but not on the nuisance parameter ν! Hence we may get rid of the nuisance parameter by conditioning on S.

Proposition 4.9. *Let us fix any parameter $(\nu_o,\gamma_o) \in \Theta$ and choose a stochastic kernel K from $\mathbb{R}^d$ to $\mathbb{R}$ describing the conditional distribution of Y, given S, under the measure P_{ν_o,γ_o}. That means,*

$$P_{\nu_o,\gamma_o}(Y \in B \,|\, S = \cdot) \;=\; K(\cdot,B) \quad \text{for any Borel set } B \subset \mathbb{R}.$$

Then there exist a measurable weight function $w : \mathbb{R} \to (0,\infty)$ and a Borel set $C \subset \mathbb{R}^d$ with the following properties: $P_{\nu_o,\gamma_o}(S \in C) = 0$, and for any $\gamma \in \Gamma$,

$$\tilde\kappa(s,\gamma) \;:=\; \log \int_{\mathbb{R}} \exp(\gamma y)\, w(y)\, K(s,\mathrm{d}y) \;<\; \infty \quad \text{for all } s \in \mathbb{R}^d \setminus C.$$

Moreover,

$$K_\gamma(s,B) \;:=\; \begin{cases} \displaystyle\int_B \exp\big(\gamma y - \tilde\kappa(s,\gamma)\big)\, w(y)\, K(s,\mathrm{d}y) & \text{if } s \in \mathbb{R}^d \setminus C, \\ 1_{[0\in B]} & \text{if } s \in C, \end{cases}$$

defines a stochastic kernel K_γ from $\mathbb{R}^d$ to $\mathbb{R}$ describing the conditional distribution of Y, given S, under any measure $P_{\nu,\gamma}$, $\nu \in N$. That means,

$$P_{\nu,\gamma}(Y \in B \,|\, S = \cdot) \;=\; K_\gamma(\cdot,B) \quad \text{for all } \nu \in N \text{ and any Borel set } B \subset \mathbb{R}.$$

Proof of Proposition 4.9. The assumption on K is equivalent to

$$E_{\nu_o,\gamma_o}\big(g(S)h(Y)\big) \;=\; E_{\nu_o,\gamma_o}\Big(g(S)\int h(y)\,K(S,\mathrm{d}y)\Big)$$

for arbitrary measurable functions $g:\mathbb{R}^d\to[0,\infty]$ and $h:\mathbb{R}\to[0,\infty]$. For any other parameter $(\nu,\gamma)\in\Theta$, this implies that

$$\begin{aligned}
&E_{\nu,\gamma}\big(g(S)h(Y)\big)\\
&\quad= E_{\nu_o,\gamma_o}\Big(g(S)h(Y)\frac{\exp(\nu^\top S+\gamma Y-\kappa(\nu,\gamma))}{\exp(\nu_o^\top S+\gamma_o Y-\kappa(\nu_o,\gamma_o))}\Big) \qquad (4.1)\\
&\quad= E_{\nu_o,\gamma_o}\Big(g(S)\exp\big((\nu-\nu_o)^\top S-\kappa(\nu,\gamma)\big)\int h(y)\exp(\gamma y)\,w(y)\,K(S,dy)\Big) \qquad (4.2)
\end{aligned}$$

with

$$w(y) \;:=\; \exp\big(\kappa(\nu_o,\gamma_o)-\gamma_o y\big).$$

Taking $h\equiv 1$ shows that

$$E_{\nu,\gamma}(g(S)) \;=\; E_{\nu_o,\gamma_o}\big(g(S)f_{\nu,\gamma}(S)\big)$$

with

$$f_{\nu,\gamma}(s) \;:=\; \exp\big((\nu-\nu_o)^\top s-\kappa(\nu,\gamma)\big)\int\exp(\gamma y)\,w(y)\,K(s,\mathrm{d}y).$$

Taking $g\equiv 1$ shows that for any fixed $\gamma\in\Gamma$, the set $C(\gamma)$ of all $s\in\mathbb{R}^d$ such that the integral $\int\exp(\gamma y)\,w(y)\,K(s,dy)$ is infinite satisfies $P_{\nu,\gamma}(S\in C(\gamma))=0$ for all $\nu\in N$. Since $P_{\nu,\gamma}$ has a strictly positive density with respect to P_{ν_o,γ_o}, we may even conclude that

$$P_{\nu_o,\gamma_o}(S\in C(\gamma)) \;=\; 0.$$

But then the set $C:=\bigcup_{\gamma\in\Gamma\cap\mathbb{Q}}C(\gamma)$ satisfies

$$P_{\nu_o,\gamma_o}(S\in C) \;=\; 0,$$

and by convexity of the exponential function,

$$\int\exp(\gamma y)\,w(y)\,K(s,\mathrm{d}y) \;<\; \infty \quad\text{for all } s\in\mathbb{R}^d\setminus C \text{ and } \gamma\in\Gamma.$$

Hence the stochastic kernel K_γ described in the proposition is well-defined, and it follows from (4.2) that for arbitrary measurable functions $g:\mathbb{R}^d\to[0,\infty]$ and $h:\mathbb{R}\to[0,\infty]$,

$$\begin{aligned}
E_{\nu,\gamma}\big(g(S)h(Y)\big) \;&=\; E_{\nu_o,\gamma_o}\Big(g(S)\int h(y)\,K_\gamma(S,\mathrm{d}y)\,f_{\nu,\gamma}(S)\Big)\\
&=\; E_{\nu,\gamma}\Big(g(S)\int h(y)\,K_\gamma(S,\mathrm{d}y)\Big).
\end{aligned}$$

This shows that the kernel K_γ describes the conditional distribution of Y, given S, under the measure $P_{\nu,\gamma}$. □

Neyman's construction of tests. Suppose we want to find a good level-α test of the null hypothesis

$$\Theta_o \;=\; \big\{(\nu,\gamma) : \gamma \le \gamma_o\big\}$$

for a given number $\gamma_o \in \Gamma$ such that

$$\big\{\nu \in \mathbb{R}^d : (\nu,\gamma_o) \in \Theta\big\} \;=\; N.$$

To this end we fix any nuisance parameter ν_o and choose for any $s \in \mathbb{R}^d$ numbers $k_\alpha(s) \in \mathbb{R}$ and $\gamma_\alpha(s) \in [0,1]$ such that the test $\varphi_\alpha : \mathbb{R}^d \times \mathbb{R} \to [0,1]$ with

$$\varphi_\alpha(s,y) \;:=\; \begin{cases} 0 & \text{if } y < k_\alpha(s) \\ \gamma_\alpha(s) & \text{if } y = k_\alpha(s) \\ 1 & \text{if } y > k_\alpha(s) \end{cases}$$

satisfies

$$E_{\nu_o,\gamma_o}\big(\varphi_\alpha(S,Y)\,\big|\,S\big) \;=\; \alpha \quad \text{almost surely.}$$

Then φ_α is a level-α test of Θ_o, and has a certain optimality property:

Theorem 4.10 (UMP unbiased tests). *For given test level $\alpha \in (0,1)$, let φ_α be the special test just described. This test belongs to the class Φ_α of all tests $\varphi : \mathbb{R}^d \times \mathbb{R} \to [0,1]$ such that*

$$E_{\nu,\gamma}\varphi(S,Y) \;\begin{cases} \le \alpha & \text{if } \gamma \le \gamma_o, \\ \ge \alpha & \text{if } \gamma > \gamma_o. \end{cases}$$

For arbitrary $\varphi \in \Phi_\alpha$ and $(\nu,\gamma) \in \Theta \setminus \Theta_o$,

$$E_{\nu,\gamma}\varphi_\alpha(S,Y) \;\ge\; E_{\nu,\gamma}\varphi(S,Y).$$

Proof of Theorem 4.10. With the stochastic kernels K_γ, $\gamma \in \Gamma$, and the Borel set $C \subset \mathbb{R}^d$ as in Proposition 4.9, we may alter the set C if necessary such that

$$\int_{\mathbb{R}} \varphi_\alpha(s,y)\, K_{\gamma_o}(s,\mathrm{d}y) \;=\; \alpha \quad \text{for all } s \in \mathbb{R}^d \setminus C.$$

For any test $\varphi : \mathbb{R}^d \times \mathbb{R} \to [0,1]$, its power

$$E_{\nu,\gamma}\varphi(S,Y) \;=\; \int_{\mathcal{X}} \varphi(S,Y) \exp\big(\nu^\top S + \gamma Y - \kappa(\nu,\gamma)\big)\,\mathrm{d}M$$

is a continuous function of $\gamma \in \Gamma(\nu)$ for any fixed $\nu \in N$; see exercises. Here $\Gamma(\nu)$ denotes the open interval

$$\Gamma(\nu) \;:=\; \{\gamma \in \mathbb{R} : (\nu,\gamma) \in \Theta\} \;\ni\; \gamma_o.$$

If $\varphi \in \Phi_\alpha$, this implies that

$$E_{\nu,\gamma_o}\varphi(S,Y) \;=\; \alpha \quad \text{for arbitrary } \nu \in N.$$

We may also write

$$E_{\nu,\gamma}\varphi(S,Y) \;=\; E_{\nu,\gamma}\Big(\int_{\mathbb{R}} \varphi(S,y)\, K_\gamma(S,\mathrm{d}y)\Big).$$

But the restricted statistical experiment $\mathcal{E}_o := \big(\mathcal{X}, \mathcal{B}, (P_{\nu,\gamma_o})_{\nu\in N}\big)$ is an exponential family with natural parametrization, because

$$\frac{\mathrm{d}P_{\nu,\gamma_o}}{\mathrm{d}M} \;=\; \exp\big(\nu^\top S - \kappa(\nu,\gamma_o)\big)\exp(\gamma_o Y),$$

i.e. with the modified measure

$$M_o(\mathrm{d}x) \;:=\; \exp(\gamma_o Y(x))\, M(\mathrm{d}x)$$

we may write

$$\frac{\mathrm{d}P_{\nu,\gamma_o}}{\mathrm{d}M_o} \;=\; \exp\big(\nu^\top S - \kappa(\nu,\gamma_o)\big).$$

Since N is open, the corresponding family $\big(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (P^S_{\nu,\gamma_o})_{\nu\in N}\big)$ is complete, that means, it follows from

$$\alpha \;=\; E_{\nu,\gamma_o}\Big(\int_{\mathbb{R}} \varphi(S,y)\, K_{\gamma_o}(S,dy)\Big) \quad \text{for all } \nu\in N$$

that

$$\int_{\mathbb{R}} \varphi(S,y)\, K_{\gamma_o}(S,\mathrm{d}y) \;=\; \alpha \quad \text{almost surely}$$

under any measure P_{ν,γ_o}, $\nu\in N$. But our special construction of φ_α and Theorem 2.15 imply that for $\gamma>\gamma_o$ and $s\in\mathbb{R}^d\setminus C$,

$$\int_{\mathbb{R}} \varphi(s,y)\, K_\gamma(s,\mathrm{d}y) \;\le\; \int_{\mathbb{R}} \varphi_\alpha(s,y)\, K_\gamma(s,\mathrm{d}y) \quad \text{whenever } \int_{\mathbb{R}} \varphi(s,y)\, K_{\gamma_o}(s,\mathrm{d}y) \le \alpha.$$

Thus for arbitrary $(\nu,\gamma)\in\Theta$ with $\gamma>\gamma_o$,

$$\begin{aligned} E_{\nu,\gamma}\varphi(S,Y) \;&=\; E_{\nu,\gamma}\Big(\int_{\mathbb{R}} \varphi(S,y)\, K_\gamma(S,\mathrm{d}y)\Big) \\ &\le\; E_{\nu,\gamma}\Big(\int_{\mathbb{R}} \varphi_\alpha(S,y)\, K_\gamma(S,\mathrm{d}y)\Big) \;=\; E_{\nu,\gamma}\varphi_\alpha(S,Y). \end{aligned}$$

□

Example 4.11 (Fisher's exact test and odds ratios). Suppose that $(Y_1,Z_1),\dots,(Y_n,Z_n)$ are independent, identically distributed random variables with values in $\{0,1\}\times\{0,1\}$ such that all four probabilities

$$\theta_{yz} \;:=\; \mathbb{P}(Y=y, Z=z), \quad y,z\in\{0,1\},$$

are strictly positive; here (Y,Z) denotes any of the n pairs (Y_i,Z_i). With the parameter $\theta := (\theta_{00},\theta_{01},\theta_{10},\theta_{11})$, the "correlation" between Y and Z may be quantified in terms of the odds ratio

$$\rho=\rho(\theta) \;:=\; \frac{\theta_{11}\theta_{00}}{\theta_{01}\theta_{10}} \;=\; \frac{\mathrm{odds}(Y=1\,|\,Z=1)}{\mathrm{odds}(Y=1\,|\,Z=0)} \;=\; \frac{\mathrm{odds}(Z=1\,|\,Y=1)}{\mathrm{odds}(Z=1\,|\,Y=0)}.$$

Suppose we are only interested in ρ. To design good tests or confidence regions for γ let us rewrite the model as a suitable natural exponential family:

We write the complete data as $(\boldsymbol{Y},\boldsymbol{Z})\in\mathcal{X}:=\{0,1\}^n\times\{0,1\}^n$, where $\boldsymbol{Y}=(Y_i)_{i=1}^n$ and $\boldsymbol{Z}=(Z_i)_{i=1}^n$. With counting measure M on $\mathcal{X}$, we obtain a statistical experiment with distributions P_θ on $\mathcal{X}$ such that

$$\begin{aligned} \log\frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;&=\; \log\big(\theta_{00}^{H_{00}}\theta_{01}^{H_{01}}\theta_{10}^{H_{10}}\theta_{11}^{H_{11}}\big) \\ &=\; H_{00}\log\theta_{00} + H_{01}\log\theta_{01} + H_{10}\log\theta_{10} + H_{11}\log\theta_{11} \end{aligned}$$

with the absolute frequencies

$$H_{yz}(\boldsymbol{y}, \boldsymbol{z}) \;:=\; \#\{i \le n : y_i = y, z_i = z\}.$$

By means of the marginal frequencies

$$\begin{aligned} H_{+1} \;&:=\; H_{01} + H_{11}, \\ H_{1+} \;&:=\; H_{10} + H_{11}, \end{aligned}$$

we may write $H_{00} = n - H_{+1} - H_{1+} + H_{11}$, $H_{01} = H_{+1} - H_{11}$, $H_{10} = H_{1+} - H_{11}$, and this leads to

$$\log \frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;=\; H_{+1} \log \frac{\theta_{01}}{\theta_{00}} + H_{1+} \log \frac{\theta_{10}}{\theta_{00}} + H_{11} \log \gamma(\theta) + n \log \theta_{00}.$$

With

$$\begin{aligned} \nu = \nu(\theta) \;&:=\; \Big(\log \frac{\theta_{01}}{\theta_{00}}, \log \frac{\theta_{10}}{\theta_{00}} \Big)^\top \;\in\; \mathbb{R}^2, \\ \gamma = \gamma(\theta) \;&:=\; \log \frac{\theta_{11}\theta_{00}}{\theta_{01}\theta_{10}} \;=\; \log \rho \;\in\; \mathbb{R}, \end{aligned}$$

we may write

$$\begin{bmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{bmatrix} \;=\; (1 + e^{\nu_1} + e^{\nu_2} + e^{\nu_1+\nu_2+\gamma})^{-1} \begin{bmatrix} 1 & e^{\nu_1} \\ e^{\nu_2} & e^{\nu_1+\nu_2+\gamma} \end{bmatrix}.$$

In particular, for any choice of $\nu \in \mathbb{R}^2$ and $\gamma \in \mathbb{R}$ there exists a probability parameter θ such that $\nu = \nu(\theta)$ and $\gamma = \gamma(\theta)$. Moreover,

$$\log \frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;=\; \nu^\top S + \gamma Y - \kappa(\nu, \gamma)$$

with

$$\begin{aligned} S \;&:=\; (H_{+1}, H_{1+})^\top, \\ Y \;&:=\; H_{11}, \\ \kappa(\nu, \gamma) \;&:=\; -n \log \theta_{00} \;=\; n \log(1 + e^{\nu_1} + e^{\nu_2} + e^{\nu_1+\nu_2+\gamma}). \end{aligned}$$

Consequently, if we want to construct tests or confidence regions for ρ, we should concentrate on the conditional distribution of H_{11}, given (H_{+1}, H_{1+}). For arbitrary integers $s, z \ge 0$, it follows from $H_{+1} = s$ and $H_{1+} = z$ that H_{11} has some value in

$$\big\{\max(0, s + z - n), \ldots, \min(s, z)\big\}.$$

For any number k in the latter set,

$$\begin{aligned} &P_\theta(H_{11} = k, H_{+1} = s, H_{1+} = z) \\ &\quad= \; P_\theta(H_{11} = k, H_{01} = s - k, H_{10} = z - k, H_{00} = n - s - z + k) \\ &\quad= \; \frac{n!}{k!(s-k)!(z-k)!(n-z-s+k)!}\, \theta_{11}^k \theta_{01}^{s-k} \theta_{10}^{z-k} \theta_{00}^{n-z-s+k} \\ &\quad= \; n!\, \theta_{01}^s \theta_{10}^z \theta_{00}^{n-z-s}\, \frac{\rho^k}{k!(s-k)!(z-k)!(n-z-s+k)!}. \end{aligned}$$

Consequently,

$$\begin{aligned} P_\theta(H_{11} = k \,|\, H_{+1} = s, H_{1+} = z) \;&=\; \frac{P_\theta(H_{11} = k, H_{+1} = s, H_{1+} = z)}{\sum_\ell P_\theta(H_{11} = \ell, H_{+1} = s, H_{1+} = z)} \\ &=\; C(n,s,z,\rho)^{-1} \, \frac{\rho^k}{k!(s-k)!(z-k)!(n-z-s+k)!} \end{aligned}$$

with

$$C(n,s,z,\rho) \;:=\; \sum_{\ell=\max(0,s+z-n)}^{\min(s,z)} \frac{\rho^\ell}{\ell!(s-\ell)!(z-\ell)!(n-z-s+\ell)!}.$$

Alternatively one may write

$$P_\theta(H_{11} = k \,|\, H_{+1} = s, H_{1+} = z) \;=\; \tilde{C}(n,s,z,\rho)^{-1} \, \mathrm{Hyp}_{n,s,z}(\{k\}) \, \rho^k$$

with the hypergeometric distribution $\mathrm{Hyp}_{n,s,z}$ and

$$\tilde{C}(n,s,z,\rho) \;:=\; \sum_{\ell=\max(0,s+z-n)}^{\min(s,z)} \mathrm{Hyp}_{n,s,z}(\{\ell\}) \, \rho^\ell.$$

In particular, if $\rho = 1$, the conditional distribution of H_{11}, given $H_{+1} = s$ and $H_{1+} = z$, equals $\mathrm{Hyp}_{n,s,z}$.

Example 4.12 (McNemar's test). Traditionally, McNemar's test is described in the context of two-by-two tables as in the previous example. But it may be transferred to a more general setting: Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with values in the finite set $\{1, 2, \ldots, K\}$ with $K \geq 3$. The parameter vector $\theta = (\theta_j)_{j=1}^K$ with $\theta_j := \mathbb{P}(X = j) > 0$ is unknown. Suppose we are mainly interested in the ratio

$$\rho = \rho(\theta) \;:=\; \frac{\theta_1}{\theta_2}.$$

Our statistical experiment is given by the sample space $\{1, 2, \ldots, K\}^n$, equipped with counting measure M, and the distributions P_θ given by

$$\log \frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;=\; \log \prod_{j=1}^K \theta_j^{H_j} \;=\; \sum_{j=1}^K H_j \log(\theta_j)$$

with the absolute frequencies

$$H_j(\boldsymbol{x}) \;:=\; \#\{i \leq n : x_i = j\}.$$

Since $H_K = n - \sum_{j<K} H_j$, we may rewrite this as

$$\begin{aligned} \log \frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;&=\; \sum_{j=1}^{K-1} H_j \log(\theta_j/\theta_K) + n \log(\theta_K) \\ &=\; H_1 \log \frac{\theta_1}{\theta_2} + (H_1 + H_2) \log \frac{\theta_2}{\theta_K} + \sum_{2<j<K} H_j \log \frac{\theta_j}{\theta_K} + n \log(\theta_K). \end{aligned}$$

With

$$\nu = \nu(\theta) \;:=\; \Big(\log\frac{\theta_{\ell+1}}{\theta_K}\Big)_{\ell=1}^{K-2} \;\in\; \mathbb{R}^{K-2},$$
$$\gamma = \gamma(\theta) \;:=\; \log\frac{\theta_1}{\theta_2} \;=\; \log\rho$$

we may write

$$\theta \;=\; \frac{(e^{\gamma+\nu_1}, e^{\nu_1}, \ldots, e^{\nu_{K-2}}, 1)^\top}{e^{\gamma+\nu_1} + e^{\nu_1} + \cdots + e^{\nu_{K-2}} + 1}.$$

In particular, for any choice of $\nu \in \mathbb{R}^{K-2}$ and $\gamma \in \mathbb{R}$ there exists a probability vector θ such that $\nu = \nu(\theta)$ and $\gamma = \gamma(\theta)$. Moreover,

$$\log\frac{\mathrm{d}P_\theta}{\mathrm{d}M} \;=\; \nu^\top S + \gamma Y - \kappa(\nu,\gamma)$$

with

$$\begin{aligned} S_1 \;&:=\; H_1 + H_2, \\ S_\ell \;&:=\; H_{\ell+1} \quad \text{for } 2 \le \ell \le K-2, \\ Y \;&:=\; H_1, \\ \kappa(\nu,\gamma) \;&:=\; -n\log(\theta_K) \;=\; n\log(e^{\gamma+\nu_1} + e^{\nu_1} + \cdots + e^{\nu_{K-2}} + 1). \end{aligned}$$

Consequently, if we want to construct tests or confidence regions for ρ, we should concentrate on the conditional distribution of H_1, given $(H_1 + H_2, H_3, \ldots, H_K)$. For arbitrary integers $m, s_3, \ldots, s_K \ge 0$ with $m + \sum_{j=3}^K s_j = n$ and $k \in \{0, \ldots, m\}$,

$$\begin{aligned} &P_\theta(H_1 = k, H_1 + H_2 = m, H_j = s_j \text{ for } j \ge 3) \\ &\quad= P_\theta(H_1 = k, H_2 = m-k, H_j = s_j \text{ for } j \ge 3) \\ &\quad= \frac{n!}{k!(m-k)!\prod_{j\ge3} s_j!}\theta_1^k\theta_2^{m-k}\prod_{j\ge3}\theta_j^{s_j} \\ &\quad= \binom{m}{k}\pi^k(1-\pi)^{m-k}\frac{n!}{m!\prod_{j\ge3} s_j!}(\theta_1+\theta_2)^m\prod_{j\ge3}\theta_j^{s_j} \end{aligned}$$

with

$$\pi = \pi(\theta) \;:=\; \frac{\theta_1}{\theta_1+\theta_2} \;=\; \frac{\rho}{1+\rho} \;\in\; (0,1).$$

Consequently, the conditional distribution of H_1, given $H_1 + H_2, H_3, \ldots, H_K$, equals

$$\mathrm{Bin}(H_1 + H_2, \pi) \;=\; \mathrm{Bin}\Big(H_1 + H_2, \frac{\rho}{1+\rho}\Big).$$

Any test or confidence region for π may be translated into a test or a confidence region for ρ via the inverse transformation $\rho = \pi/(1-\pi)$.

Example 4.13 (Comparing two Poisson parameters). Suppose we observe independent random variables $Y \sim \mathrm{Poiss}(\lambda)$ and $Z \sim \mathrm{Poiss}(\mu)$ with unknown parameters $\lambda, \mu > 0$. Suppose further that we are mainly interested in the ratio $\rho := \lambda/\mu$. With $\mathcal{X} := \mathbb{N}_0 \times \mathbb{N}_0$ and counting measure M on $\mathcal{X}$, this situation corresponds to a statistical model with distributions $P_{\lambda,\mu}$ given by

$$\frac{\mathrm{d}P_{\lambda,\mu}}{\mathrm{d}M}(y,z) \;=\; e^{-\lambda}\frac{\lambda^y}{y!}e^{-\mu}\frac{\mu^z}{z!} \;=\; \frac{\exp(y\log\lambda + z\log\mu - \lambda - \mu)}{y!z!}.$$

Replacing M with the measure M_o given by $M_o(\{(y,z)\}) := (y!z!)^{-1}$ we may write

$$\begin{aligned}\log \frac{\mathrm{d}P_{\lambda,\mu}}{\mathrm{d}M_o}(y,z) &= y\log\lambda + z\log\mu - (\lambda+\mu)\\ &= y\log(\lambda/\mu) + (y+z)\log\mu - (\lambda+\mu)\\ &= \nu(y+z) + \gamma y - \kappa(\nu,\gamma)\end{aligned}$$

with

$$\begin{aligned}\nu &:= \log\mu,\\ \gamma &:= \log(\lambda/\mu) = \log\rho,\\ \kappa(\nu,\gamma) &:= \lambda+\mu = \exp(2\nu+\gamma).\end{aligned}$$

Consequently, for inference about ρ we should analyze the conditional distribution of Y, given $Y+Z$. But for arbitrary integers $m \geq 0$ and $k \in \{0,\ldots,m\}$,

$$\begin{aligned}P_{\lambda,\mu}(Y=k, Y+Z=m) &= P_{\lambda,\mu}(Y=k)P_{\lambda,\mu}(Z=m-k)\\ &= e^{-(\lambda+\mu)}\frac{\lambda^k\mu^{m-k}}{k!(m-k)!}\\ &= \binom{m}{k}\pi^k(1-\pi)^{m-k}\,e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^m}{m!}\\ &= \mathrm{Bin}_{m,\pi}(\{k\})\mathrm{Poiss}_{\lambda+\mu}(\{m\})\end{aligned}$$

with

$$\pi := \frac{\lambda}{\lambda+\mu} = \frac{\rho}{1+\rho} \in (0,1).$$

This shows that the conditional distribution of Y, given $Y+Z$, is equal to

$$\mathrm{Bin}(Y+Z,\pi) = \mathrm{Bin}\Big(Y+Z, \frac{\rho}{1+\rho}\Big).$$

Hence we may construct tests and confidence regions for π, and these translate into tests and confidence regions for ρ via the inverse transformation $\rho = \pi/(1-\pi)$.

Avoiding conditional distributions

In the previous examples we computed the conditional distribution of Y, given S, explicitly. In various settings this step can be avoided by means of the following result:

Lemma 4.14 (Basú). *Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ be a statistical experiment, and let $S : (\mathcal{X},\mathcal{B}) \to (\mathcal{X}',\mathcal{B}')$ be a sufficient statistic for $\mathcal{E}$ such that the experiment $\mathcal{E}^S = (\mathcal{X}',\mathcal{X}',(P_\theta^S)_{\theta\in\Theta})$ is boundedly complete. If $V : (\mathcal{X},\mathcal{B}) \to (\mathcal{X}'',\mathcal{B}'')$ is a measurable mapping such that its distributions P_θ^V, $\theta\in\Theta$, are identical, then S and V are stochastically independent under each measure P_θ, $\theta\in\Theta$.*

Proof of Lemma 4.14. By sufficency of S there exists a stochastic kernel K from $(\mathcal{X}',\mathcal{B}')$ to $(\mathcal{X},\mathcal{B})$ describing the conditional distribution of $X \sim P_\theta$, given S, simultaneously for all $\theta\in\Theta$.

That means, for arbitrary $B' \in \mathcal{B}'$, $B'' \in \mathcal{B}''$ and $\theta \in \Theta$,

$$\begin{aligned} P_\theta(S \in B', V \in B'') &= E_\theta\big(E_\theta(1_{B'}(S)1_{B''}(V) \,|\, S)\big) \\ &= E_\theta\big(1_{B'}(S)P_\theta(V \in B'' \,|\, S)\big) \\ &= E_\theta\big(1_{B'}(S)K(S, \{V \in B''\})\big). \end{aligned}$$

Setting $B' = \mathcal{X}'$, we obtain the equation

$$P_\theta(V \in B'') \;=\; E_\theta K(S, \{V \in A''\}).$$

By our assumption on V, the left hand side does not depend on $\theta \in \Theta$, and we denote this number with $P(V \in B'')$. Hence $f(s) := K(s, \{V \in B''\}) - P(V \in B'')$ defines a bounded measurable function on $(\mathcal{X}', \mathcal{B}')$ such that $\int f \,\mathrm{d}P_\theta^S = 0$ for all $\theta \in \Theta$. By completeness of $\mathcal{E}^S$,

$$P_\theta(f(S) \neq 0) \;=\; P_\theta\big(K(S, \{V \in B''\}) \neq P(V \in B'')\big) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

Hence for arbitrary $B' \in \mathcal{B}'$, $B'' \in \mathcal{B}''$ and $\theta \in \Theta$,

$$\begin{aligned} P_\theta(S \in B', V \in B'') &= E_\theta\big(1_{B'}(S)K(S, \{V \in B''\})\big) \\ &= E_\theta\big(1_{B'}(S)P(V \in B'')\big) \\ &= P_\theta(S \in B')P(V \in B''), \end{aligned}$$

which proves stochastic independence of S and V under P_θ. □

Application to exponential families. Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ be a natural exponential family as described at the beginning of this section with open and convex parameter space $\Theta = N \times \Gamma \subset \mathbb{R}^d \times \mathbb{R}$ and sufficient statistic $(S, Y) \in \mathbb{R}^d \times \mathbb{R}$. Writing $\theta \in \Theta$ as $\theta = (\nu, \gamma)$, we know from Neyman's theory how to construct optimal unbiased level-α tests of the null hypotheses "$\gamma \leq \gamma_o$" or "$\gamma \geq \gamma_o$" or "$\gamma = \gamma_o$" for any given value $\gamma_o \in \Gamma$.

Note that for the restricted experiment $\mathcal{E}_o = (\mathcal{X}, \mathcal{B}, (P_{\nu,\gamma_o})_{\nu\in N})$ the statistic S is sufficient, and the family $\mathcal{E}_o^S = (\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (P^S_{\nu,\gamma_o})_{\nu\in N})$ is complete. Suppose we can identify a real-valued statistic

$$V(\cdot, \gamma_o) \;=\; f(S, Y)$$

with the following two properties:

- The distribution of $V(X, \gamma_o)$, $X \sim P_{\nu,\gamma_o}$, does not depend on $\nu \in N$.
- $V(\cdot, \gamma_o) = f(S, Y)$ is strictly increasing in Y almost everywhere.

Then optimal unbiased level-α tests of the null hypotheses above may be constructed in terms of $V(\cdot, \gamma_o)$ and its unconditional distribution under P_{ν,γ_o}, where $\nu \in N$ is arbitrary.

For instance if $V(X, \gamma_o)$, $X \sim P_{\nu,\gamma_o}$ has continuous distribution function F_{γ_o}, then the right-sided p-value

$$1 - F_{\gamma_o}(V(X, \gamma_o))$$

yields an optimal unbiased level-α test of "$\gamma \leq \gamma_o$", whereas the left-sided p-value

$$F_{\gamma_o}(V(X, \gamma_o))$$

is optimal for the null hypothesis "$\gamma \geq \gamma_o$".

Example 4.15 (Student's t-test for a Gaussian mean). As in Example 4.3 we consider the statistical experiment $\mathcal{E} = \big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), (\mathcal{N}(\mu, \sigma^2)^{\otimes n})_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)}\big)$. Suppose we want to construct optimal unbiased level-α tests of "$\mu \leq \mu_o$" or "$\mu \geq \mu_o$" or "$\mu = \mu_o$", where μ_o is a given fixed number. In all three cases we have to deal with the nuisance parameter σ. With M denoting Lebesgue measure on $\mathbb{R}^n$ times $(2\pi)^{-n/2}$, we may write

$$\begin{aligned}
\log \frac{\mathrm{d}P_{\mu,\sigma}}{\mathrm{d}M}(\boldsymbol{x}) &= -\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma \\
&= -\sum_{i=1}^n \frac{(x_i-\mu_o)^2}{2\sigma^2} + \frac{n(\mu-\mu_o)(\bar{x}-\mu_o)}{\sigma^2} - n\log\sigma - \frac{n(\mu-\mu_o)^2}{2\sigma^2} \\
&= \nu S(\boldsymbol{x}) + \gamma Y(\boldsymbol{x}) - \kappa(\nu,\gamma),
\end{aligned}$$

where

$$\begin{aligned}
\nu &:= \frac{-1}{2\sigma^2} \in (-\infty, 0), \\
\gamma &:= \frac{\sqrt{n}(\mu-\mu_o)}{\sigma^2} \in \mathbb{R}, \\
\kappa(\nu,\gamma) &:= n\log\sigma + \frac{n(\mu-\mu_o)^2}{2\sigma^2}, \\
S(\boldsymbol{x}) &:= \sum_{i=1}^n (x_i-\mu_o)^2, \\
Y(\boldsymbol{x}) &:= \sqrt{n}(\bar{x}-\mu_o).
\end{aligned}$$

Hence an unbiased test of "$\mu \leq \mu_o$" (or of "$\mu = \mu_o$") may be identified with an unbiased test of "$\gamma \leq 0$" (or of "$\gamma = 0$").

Instead of determining the conditional distribution of Y, given S, under $P_{\mu_o,\sigma}$ for some $\sigma > 0$ directly, we apply Basú's lemma and recall student's method from introductory statistics courses: With the sample standard deviation

$$\widehat{\sigma}(\boldsymbol{X}) := \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2}$$

it is well-known that

$$V(\boldsymbol{X}, \mu_o) := \frac{\sqrt{n}(\bar{X}-\mu_o)}{\widehat{\sigma}(\boldsymbol{X})} \sim t_{n-1}$$

whenever $\mu = \mu_o$, irrespective of $\sigma > 0$. But

$$\sum_{i=1}^n (X_i-\bar{X})^2 = \sum_{i=1}^n (X_i-\mu_o)^2 - n(\bar{X}-\mu_o)^2 = S(\boldsymbol{X}) - Y(\boldsymbol{X})^2,$$

so

$$V(\cdot, \mu_o) \;=\; \frac{\sqrt{n-1}\,Y}{\sqrt{S - Y^2}},$$

which is strictly monotone increasing in Y on the set $\{0 < Y^2 < S\}$. (Note that $0 < Y(\boldsymbol{X})^2 < S(\boldsymbol{X})$ almost surely.) Hence with F_{n-1} denoting the distribution function of t_{n-1}, the left-sided p-value

$$F_{n-1}(V(\boldsymbol{X}, \mu_o))$$

yields optimal unbiased tests of "$\mu \geq \mu_o$", the right-sided p-value

$$1 - F_{n-1}(V(\boldsymbol{X}, \mu_o))$$

yields optimal unbiased tests of "$\mu \leq \mu_o$", and the traditional $(1-\alpha)$-confidence interval

$$\Big[\bar{X} \;\pm\; \frac{\widehat{\sigma}(\boldsymbol{X})}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\Big]$$

for μ is based on optimal unbiased level-α tests of one-point hypotheses "$\mu = \mu_o$", $\mu_o \in \mathbb{R}$.

Example 4.16 (Comparing two Gamma scale parameters). Suppose we observe independent random variables

$$X_1 \;\sim\; \mathrm{Gamma}(a_1, \beta_1) \quad \text{and} \quad X_2 \;\sim\; \mathrm{Gamma}(a_2, \beta_2)$$

with given shape parameters $a_1, a_2 > 0$ and unknown scale parameters $\beta_1, \beta_2 > 0$. Suppose we are mainly interested in the ratio

$$\rho = \rho(\boldsymbol{\beta}) \;:=\; \beta_1/\beta_2,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2) \in (0,\infty) \times (0,\infty)$. With $\mathcal{X} := (0,\infty)^2$ and Lebesgue measure M on $\mathcal{X}$, this corresponds to the statistical model with distributions $P_{\boldsymbol{\beta}}$ given by

$$\frac{\mathrm{d}P_{\boldsymbol{\beta}}}{\mathrm{d}M}(\boldsymbol{x}) \;=\; \frac{(x_1/\beta_1)^{a_1-1}\exp(-x_1/\beta_1)}{\Gamma(a_1)\beta_1}\,\frac{(x_2/\beta_2)^{a_2-1}\exp(-x_2/\beta_2)}{\Gamma(a_2)\beta_2}.$$

With the modified measure M_o given by

$$\frac{\mathrm{d}M_o}{\mathrm{d}M}(\boldsymbol{x}) \;:=\; \frac{x_1^{a_1-1}x_2^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)}$$

we may write

$$\log\frac{\mathrm{d}P_{\boldsymbol{\beta}}}{\mathrm{d}M_o}(\boldsymbol{x}) \;=\; \frac{-1}{\beta_1}\,x_1 + \frac{-1}{\beta_2}\,x_2 - a_1\log\beta_1 - a_2\log\beta_2.$$

For a hypothetical value ρ_o of ρ we may rewrite the log-likelihood as

$$\begin{aligned}\log\frac{\mathrm{d}P_{\boldsymbol{\beta}}}{\mathrm{d}M_o}(\boldsymbol{x}) \;&=\; \frac{\rho-\rho_o}{\beta_2\rho}\,x_1/\rho_o + \frac{-1}{\beta_2}\,(x_1/\rho_o + x_2) - a_1\log\beta_1 - a_2\log\beta_2\\ &=\; \nu S(\boldsymbol{x}) + \gamma Y(\boldsymbol{x}) - \kappa(\nu,\gamma)\end{aligned}$$

with

$$\begin{aligned}\nu \;&:=\; \frac{-1}{\beta_2} \;\in\; (-\infty, 0),\\ \gamma \;&:=\; \frac{\rho-\rho_o}{\beta_2\rho} \;\in\; \mathbb{R},\\ S(\boldsymbol{x}) \;&:=\; x_1/\rho_o + x_2,\\ Y(\boldsymbol{x}) \;&:=\; x_1/\rho_o\end{aligned}$$

and the normalization constant $\kappa(\nu,\gamma) = a_1 \log \beta_1 + a_2 \log \beta_2$. Note that (ν,γ) lies in the open convex set

$$\big\{(\nu,\gamma) \in (-\infty,0) \times \mathbb{R} : \gamma < -\nu\big\},$$

which contains $(-\infty,0) \times \{0\}$. Hence optimal unbiased tests of the null hypothesis "$\rho \le \rho_o$" or "$\rho \ge \rho_o$" or "$\rho = \rho_o$" may be viewed as optimal unbiased tests of the null hypothesis "$\gamma \le 0$" or "$\gamma \ge 0$" or "$\gamma = 0$" and could be constructed with the conditional distribution of Y, given S, under any distribution $P_{\beta_2\rho_o,\beta_2}$.

In case of $\rho = \rho_o$, the distribution of $(X_1/\rho_o, X_2)$ coincides with the distribution of $\beta_2(Z_1, Z_2)$ with independent random variables

$$Z_1 \;\sim\; \mathrm{Gamma}(a_1,1) \quad \text{and} \quad Z_2 \;\sim\; \mathrm{Gamma}(a_2,1).$$

Then the test statistic

$$V(\boldsymbol{X},\rho_o) \;:=\; \frac{Y}{S}(\boldsymbol{X}) \;=\; \frac{X_1/\rho_0}{X_1/\rho_o + X_2}$$

has the same distribution as

$$\frac{Z_1}{Z_1+Z_2} \;\sim\; \mathrm{Beta}(a_1,a_2),$$

irrespective of $\beta_2 = -1/\nu$. Here we refer to the well-known fact that $Z_1/(Z_1{+}Z_2)$ and $Z_1{+}Z_2$ are stochastically independent with distributions $\mathrm{Beta}(a_1,a_2)$ and $\mathrm{Gamma}(a_1+a_2,1)$, respectively. The independence would also follow from Basú's lemma.

Since $V(\cdot,\rho_o) = Y/S$ is strictly increasing in Y, optimal tests are obtained by comparing the value $V(\boldsymbol{X},\rho_o)$ with $\mathrm{Beta}(a_1,a_2)$. Specifically, if we denote the u-quantile of $\mathrm{Beta}(a_1,a_2)$ with $q_{a_1,a_2}(u)$, then an optimal unbiased level-α test of "$\rho \ge \rho_o$" rejects this null hypothesis if

$$V(\boldsymbol{X},\rho_o) \;\le\; q_{a_1,a_2}(\alpha).$$

This leads to the $(1-\alpha)$-confidence region

$$\begin{aligned} C_\alpha(\boldsymbol{X}) \;&:=\; \big\{\rho_o > 0 : V(\boldsymbol{X},\rho_o) > q_{a_1,a_2}(\alpha)\big\} \\ &=\; \Big\{\rho_o > 0 : \frac{X_1/X_2}{X_1/X_2 + \rho_o} > q_{a_1,a_2}(\alpha)\Big\} \\ &=\; \Big(0,\; \frac{X_1}{X_2}\big(q_{a_1,a_2}(\alpha)^{-1} - 1\big)\Big). \end{aligned}$$

for ρ. Analogously, an optimal unbiased level-α test of "$\rho \le \rho_o$" rejects this null hypothesis if

$$V(\boldsymbol{X},\rho_o) \;\ge\; q_{a_1,a_2}(1-\alpha),$$

and this leads to the $(1-\alpha)$-confidence region

$$C_\alpha(\boldsymbol{X}) \;:=\; \Big(\frac{X_1}{X_2}\big(q_{a_1,a_2}(1-\alpha)^{-1} - 1\big),\; \infty\Big).$$

for ρ.

Chapter 5

Some Asymptotics

5.1 Testing, Total Variation and Hellinger Distances

Let P, Q be two probability distributions on a measurable space $(\mathcal{X}, \mathcal{B})$. In this section we introduce statistically meaningful measures of distance between P and Q and establish connections between them.

Testing affinity and distance. Let X be a random variable with unknown distribution in $\{P, Q\}$. Now we consider a statistical test $\varphi : \mathcal{X} \to [0, 1]$ and interpret $\varphi(X)$ as the probability of claiming that $X \sim Q$, whereas $1 - \varphi(X)$ is the probability of claiming that $X \sim P$. The quality of φ could be measured by the risk

$$\mathbb{P}(\text{error of 1st kind}) + \mathbb{P}(\text{error of 2nd kind}) \;=\; \int \varphi \,\mathrm{d}P + \int (1 - \varphi) \,\mathrm{d}Q.$$

Hence a natural measure of similarity between P and Q is given by the *testing affinity*

$$\eta_{\mathrm{T}}(P, Q) \;:=\; \inf_{\text{tests } \varphi} \Big(\int \varphi \,\mathrm{d}P + \int (1 - \varphi) \,\mathrm{d}Q \Big),$$

where small values indicate that P and Q easy to distinguish, i.e. dissimilar. As shown in the next lemma, the latter infimum is always a minimum. Moreover, the *testing distance*

$$D_{\mathrm{T}}(P, Q) \;:=\; 1 - \eta_{\mathrm{T}}(P, Q)$$

defines a metric on the space of all probability distributions on $(\mathcal{X}, \mathcal{B})$.

Lemma 5.1 (Testing affinity and distance). *Let M be a measure on $(\mathcal{X}, \mathcal{B})$ such that densities $f = \mathrm{d}P/\mathrm{d}M$ and $g = \mathrm{d}Q/\mathrm{d}M$ exist. Then*

$$\eta_{\mathrm{T}}(P, Q) \;=\; \int \min(f, g) \,\mathrm{d}M \;\in\; [0, 1].$$

The minimum is attained by any test φ such that $\varphi = 0$ on $\{f > g\}$ and $\varphi = 1$ on $\{f < g\}$. There exists always such a measure M. The testing distance can be written as

$$D_{\mathrm{T}}(P, Q) \;=\; \frac{1}{2} \int |f - g| \,\mathrm{d}M.$$

It defines a metric on the space of probability measures on $(\mathcal{X}, \mathcal{B})$ with values in $[0, 1]$.

Proof of Lemma 5.1. Note first that for any test φ,

$$\begin{aligned}\int \varphi \,\mathrm{d}P + \int (1-\varphi)\,\mathrm{d}Q \;&=\; \int \big(\varphi(x)f(x) + (1-\varphi(x))g(x)\big)\, M(\mathrm{d}x)\\ &\geq\; \int \min\big(f(x), g(x)\big)\, M(\mathrm{d}x)\end{aligned}$$

with equality if

$$\varphi(x) \;=\; \begin{cases} 0 & \text{if } f(x) > g(x),\\ 1 & \text{if } f(x) < g(x).\end{cases}$$

This proves already the first part of the lemma. As to the existence of a dominating measure M, let $M := P + Q$. Then $P(A) = Q(A) = 0$ for any $A \in \mathcal{B}$ such that $M(A) = 0$. By the theorem of Radon–Nikodym, there exist densities $f = \mathrm{d}P/\mathrm{d}M$ and $g = \mathrm{d}Q/\mathrm{d}M$.

Since $f, g \geq \min(f, g) \geq 0$, it is clear that $0 \leq \int \min(f, g)\,\mathrm{d}M \leq 1$. Moreover,

$$\begin{aligned}D_{\mathrm{T}}(P,Q) \;=\; 1 - \eta_{\mathrm{T}}(P,Q) \;&=\; \int \Big(\frac{f+g}{2} - \min(f,g)\Big)\,\mathrm{d}M\\ &=\; \frac{1}{2}\int \big(f + g - 2\min(f,g)\big)\,\mathrm{d}M\\ &=\; \frac{1}{2}\int \big(\max(f,g) - \min(f,g)\big)\,\mathrm{d}M\\ &=\; \frac{1}{2}\int |f-g|\,\mathrm{d}M.\end{aligned}$$

Obviously this equals 0 if, and only if, $f = g$ M-almost everywhere, which is equivalent to $P \equiv Q$. Moreover, $D_{\mathrm{T}}(Q,P) = D_{\mathrm{T}}(P,Q)$.

It remains to prove the triangle inequality for D_{T}. Let P, Q, R be probability distributions on $(\mathcal{X}, \mathcal{B})$. Let M be a measure on $(\mathcal{X}, \mathcal{B})$ such that densities $f = \mathrm{d}Q/\mathrm{d}M$, $g = \mathrm{d}Q/\mathrm{d}M$ and $h = \mathrm{d}R/\mathrm{d}M$ exist. A possible choice would be $M := P + Q + R$, again by the theorem of Radon–Nikodym. Then by the the triangle inequality,

$$\begin{aligned}D_{\mathrm{T}}(P,R) \;=\; \frac{1}{2}\int |f-h|\,\mathrm{d}M \;&\leq\; \frac{1}{2}\int |f-g|\,\mathrm{d}M + \frac{1}{2}\int |g-h|\,\mathrm{d}M\\ &=\; D_{\mathrm{T}}(P,Q) + D_{\mathrm{T}}(Q,R).\end{aligned}$$

□

Total variation distance. Another measure of distance is given by the total variation distance

$$D_{\mathrm{TV}}(P,Q) \;:=\; \sup_{B\in\mathcal{B}} \big|P(B) - Q(B)\big|.$$

One could easily verify directly that this defines a metric on the space of probability measures on $(\mathcal{X}, \mathcal{B})$. But the next lemma shows that $D_{\mathrm{TV}} = D_{\mathrm{T}}$.

Lemma 5.2. *For arbitrary probability distributions P and Q on $(\mathcal{X}, \mathcal{B})$,*

$$D_{\mathrm{TV}}(P,Q) \;=\; \max_{B\in\mathcal{B}} \big(P(B) - Q(B)\big) \;=\; \max_{B'\in\mathcal{B}} \big(Q(B') - P(B')\big) \;=\; D_{\mathrm{T}}(P,Q).$$

If P and Q have densities f and g as in Lemma 5.1, then the latter two maxima are attained for arbitrary sets B, B' such that $\{f > g\} \subset B \subset \{f \geq g\}$ and $\{f < g\} \subset B' \subset \{f \leq g\}$.

Proof of Lemma 5.2. Since $P(B) - Q(B) = Q(\mathcal{X} \setminus B) - P(\mathcal{X} \setminus B)$,

$$D_{\mathrm{TV}}(P,Q) \;=\; \sup_{B\in\mathcal{B}}\big(P(B) - Q(B)\big) \;=\; \sup_{B'\in\mathcal{B}}\big(Q(B') - P(B')\big).$$

In case of P and Q having densities f and g, respectively, with respect to some measure M on $(\mathcal{X},\mathcal{B})$,

$$P(B) - Q(B) \;=\; \int 1_B(f-g)\,\mathrm{d}M \;\le\; \int (f-g)^+\,\mathrm{d}M$$

with equality if $\{f > g\} \subset B \subset \{f \ge g\}$. Analogously,

$$Q(B') - P(B') \;\le\; \int (g-f)^+\,\mathrm{d}M \;=\; \int (f-g)^-\,\mathrm{d}M$$

with equality if $\{f < g\} \subset B' \subset \{f \le g\}$. Consequently,

$$D_{\mathrm{TV}}(P,Q) \;=\; \frac{1}{2}\int \big((f-g)^+ + (f-g)^-\big)\,\mathrm{d}M \;=\; \frac{1}{2}\int |f-g|\,\mathrm{d}M \;=\; D_{\mathrm{T}}(P,Q).$$ □

Hellinger affinity and distance. In many situations it turns out that $\eta_{\mathrm{T}}(P,Q)$ and $D_{\mathrm{T}}(P,Q)$ are difficult to compute explicitly. As we shall see later, interesting proxys are given by the *Hellinger affinity*

$$\eta_{\mathrm{H}}(P,Q) \;:=\; \int \sqrt{fg}\,\mathrm{d}M$$

and the *Hellinger distance*

$$D_{\mathrm{H}}(P,Q) \;:=\; \sqrt{\frac{1}{2}\int \big(\sqrt{f}-\sqrt{g}\big)^2\,\mathrm{d}M},$$

where M is some measure on $(\mathcal{X},\mathcal{B})$ such that densities $f = \mathrm{d}P/\mathrm{d}M$ and $g = \mathrm{d}Q/\mathrm{d}M$ exist. Note also that

$$D_{\mathrm{H}}(P,Q)^2 \;=\; \frac{1}{2}\int \big(f+g-2\sqrt{fg}\big)\,\mathrm{d}M \;=\; 1 - \eta_{\mathrm{H}}(P,Q).$$

As in case of $\eta_{\mathrm{T}}(P,Q)$ and $D_{\mathrm{T}}(P,Q)$, the choice of M is irrelevant. Precisely, let $M_o := P+Q$. By the theorem of Radom–Nikodym, there exist densities $f_o = \mathrm{d}P/\mathrm{d}M_o$ and $g_o = \mathrm{d}Q/\mathrm{d}M_o$, and these are M_o-almost everywhere unique. If M is an arbitrary measure such that densities $f = \mathrm{d}Q/\mathrm{d}M$ and $g = \mathrm{d}Q/\mathrm{d}M$ exist, then $h := f+g$ is a density of M_o with respect to M, and one can easily show that f/h and g/h (with $0/0 := 0$) are densities of P and Q, respectively, with respect to M_o. Consequently, $M_o(f_o \ne f/h) = 0 = M_o(g_o \ne g/h)$, whence

$$\int \sqrt{f_o g_o}\,\mathrm{d}M_o \;=\; \int \sqrt{(f/h)(g/h)}\,\mathrm{d}M_o \;=\; \int \sqrt{(f/h)(g/h)}\,h\,\mathrm{d}M \;=\; \int \sqrt{fg}\,\mathrm{d}M.$$

The following lemma shows that testing and Hellinger distance induce the same topology on the space of probability distributions on $(\mathcal{X},\mathcal{B})$.

Lemma 5.3 (Relationships between testing and Hellinger distance).

$$1 - \sqrt{1 - \eta_{\mathrm{H}}^2} \;\le\; \eta_{\mathrm{T}} \;\le\; \eta_{\mathrm{H}}$$

and

$$D_{\mathrm{H}}^2 \;\le\; D_{\mathrm{T}} \;\le\; D_{\mathrm{H}}\sqrt{2 - D_{\mathrm{H}}^2}.$$

Proof of Lemma 5.3. With explicit densities $f := \mathrm{d}P/\mathrm{d}M$ and $g := \mathrm{d}Q/\mathrm{d}M$ it follows from $\min(a,b) \le \sqrt{ab}$ for real numbers $a, b \ge 0$ that

$$\eta_{\mathrm{T}}(P,Q) \;=\; \int \min(f,g)\,\mathrm{d}M \;\le\; \int \sqrt{fg}\,\mathrm{d}M \;=\; \eta_{\mathrm{H}}(P,Q).$$

In particular,

$$D_{\mathrm{T}}(P,Q) \;=\; 1 - \eta_{\mathrm{T}}(P,Q) \;\ge\; 1 - \eta_{\mathrm{H}}(P,Q) \;=\; D_{\mathrm{H}}(P,Q)^2.$$

As to the other bounds, it follows from $a - b = \bigl(\sqrt{a} - \sqrt{b}\bigr)\bigl(\sqrt{a} + \sqrt{b}\bigr)$ for real numbers $a, b \ge 0$ and the Cauchy–Schwarz inequality that

$$\begin{aligned}
1 - \eta_{\mathrm{T}}(P,Q) = D_{\mathrm{T}}(P,Q) \;&=\; \frac{1}{2}\int \bigl|\sqrt{f} - \sqrt{g}\bigr|\bigl|\sqrt{f} + \sqrt{g}\bigr|\,\mathrm{d}M \\
&\le\; \frac{1}{2}\sqrt{\int \bigl(f + g - 2\sqrt{fg}\bigr)\,\mathrm{d}M \int \bigl(f + g + 2\sqrt{fg}\bigr)\,\mathrm{d}M} \\
&=\; \sqrt{(1 - \eta_{\mathrm{H}}(P,Q))(1 + \eta_{\mathrm{H}}(P,Q))} \\
&=\; \begin{cases} \sqrt{1 - \eta_{\mathrm{H}}(P,Q)^2}, \\ D_{\mathrm{H}}(P,Q)\sqrt{2 - D_{\mathrm{H}}(P,Q)^2}. \end{cases}
\end{aligned}$$

This proves that

$$D_{\mathrm{T}} \;\le\; D_{\mathrm{H}}\sqrt{2 - D_{\mathrm{H}}^2} \quad \text{and} \quad 1 - \eta_{\mathrm{T}} \;\le\; \sqrt{1 - \eta_{\mathrm{H}}^2},$$

where the latter inequality is equivalent to $\eta_{\mathrm{T}} \ge 1 - \sqrt{1 - \eta_{\mathrm{H}}^2}$. □

Remark 5.4. As mentioned already, the formulae for η_{T}, D_{T}, η_{H} and D_{H} are independent of the choice of the dominating measure M. Thus some authors write symbolically

$$\begin{aligned}
\eta_{\mathrm{T}}(P,Q) \;&=\; \int \min(\mathrm{d}P, \mathrm{d}Q), \\
D_{\mathrm{T}}(P,Q) \;&=\; \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|, \\
\eta_{\mathrm{H}}(P,Q) \;&=\; \int \sqrt{\mathrm{d}P\,\mathrm{d}Q}, \\
D_{\mathrm{H}}(P,Q) \;&=\; \sqrt{\frac{1}{2}\int \bigl(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\bigr)^2}.
\end{aligned}$$

Exercise 5.5 (More about the relation between testing and Hellinger distance). **(a)** Show that the inequalities

$$D_{\mathrm{H}}^2 \;\le\; D_{\mathrm{T}} \;\le\; D_{\mathrm{H}}\sqrt{2 - D_{\mathrm{H}}^2}$$

are equivalent to

$$\sqrt{1-\sqrt{1-D_{\mathrm{T}}^2}} \;\le\; D_{\mathrm{H}} \;\le\; \sqrt{D_{\mathrm{T}}}.$$

(b) Visualize these two pairs of bounds graphically.

(c) Construct for any $\gamma \in (0,1]$ two distributions P_γ and Q_γ on a suitable sample space $(\mathcal{X},\mathcal{B})$ such that $D_{\mathrm{T}}(P_\gamma,Q_\gamma) = D_{\mathrm{H}}^2(P_\gamma,Q_\gamma) = \gamma$.

Remark 5.6. Let $(\mathcal{X}',\mathcal{B}')$ be a second measurable space, and let $\tau : \mathcal{X} \to \mathcal{X}'$ be a bijection such that both τ and τ^{-1} are measurable. Then

$$\kappa(P,Q) \;=\; \kappa(P^\tau,Q^\tau) \quad \text{for } \kappa = \eta_{\mathrm{T}}, D_{\mathrm{T}}, \eta_{\mathrm{H}}, D_{\mathrm{H}}.$$

The proof of these equalities is left to the reader as an exercise. (Recall that for any measure M on $(\mathcal{X},\mathcal{B})$, its "push-forward" measure M^τ on $(\mathcal{X}',\mathcal{B}')$ is given by $M^\tau(B') := M(\tau^{-1}(B'))$ for $B' \in \mathcal{B}'$.)

Example 5.7 (Testing and Hellinger distance for univariate Gaussian shift). For real numbers μ_1, μ_2 and $\sigma > 0$,

$$\begin{aligned}
\eta_{\mathrm{T}}\big(\mathrm{N}(\mu_1,\sigma^2),\mathrm{N}(\mu_2,\sigma^2)\big) \;&=\; 2\Phi\Big(-\frac{|\mu_1-\mu_2|}{2\sigma}\Big),\\
\eta_{\mathrm{H}}\big(\mathrm{N}(\mu_1,\sigma^2),\mathrm{N}(\mu_2,\sigma^2)\big) \;&=\; \exp\Big(-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}\Big),
\end{aligned}$$

where Φ is the standard Gaussian distribution function. Hence

$$\begin{aligned}
D_{\mathrm{T}}\big(\mathrm{N}(\mu_1,\sigma^2),\mathrm{N}(\mu_2,\sigma^2)\big) \;&=\; 2\Phi\Big(\frac{|\mu_1-\mu_2|}{2\sigma}\Big) - 1,\\
D_{\mathrm{H}}\big(\mathrm{N}(\mu_1,\sigma^2),\mathrm{N}(\mu_2,\sigma^2)\big) \;&=\; \sqrt{1-\exp\Big(-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}\Big)}.
\end{aligned}$$

To verify these formulae, we first apply Remark 5.6 to the bijection $\tau : \mathbb{R} \to \mathbb{R}$ with $\tau(x) := \sigma^{-1}(x - \min(\mu_1,\mu_2))$. Then it suffices to verify the asserted formulae for the testing and Hellinger affinity in case of $P = \mathrm{N}(0,1)$ and $Q = \mathrm{N}(\mu,1)$, where

$$\mu \;:=\; \frac{|\mu_1-\mu_2|}{\sigma}.$$

But with $C = (2\pi)^{-1/2}$ and the standard Gaussian density $\phi(x) := C\exp(-x^2/2)$, the testing affinity $\eta_{\mathrm{T}}\big(\mathrm{N}(0,1),\mathrm{N}(\mu,1)\big)$ equals

$$\begin{aligned}
\int \min\big(\phi(x),\phi(x-\mu)\big)\,\mathrm{d}x \;&=\; C\int \exp\Big(-\frac{\max(x^2,(x-\mu)^2)}{2}\Big)\,\mathrm{d}x\\
&=\; 2C\int_{-\infty}^{\mu/2} \exp\Big(-\frac{(x-\mu)^2}{2}\Big)\,\mathrm{d}x\\
&=\; 2\int_{-\infty}^{\mu/2} \phi(x-\mu)\,\mathrm{d}x\\
&=\; 2\Phi(-\mu/2).
\end{aligned}$$

Moreover, the Hellinger affinity $\eta_{\mathrm{H}}\big(\mathrm{N}(0,1),\mathrm{N}(\mu,1)\big)$ is equal to

$$\begin{aligned}\int \sqrt{\phi(x)\phi(x-\mu)}\,\mathrm{d}x \;&=\; C\int \exp\Big(-\frac{x^2+(x-\mu)^2}{4}\Big)\,\mathrm{d}x\\ &=\; C\int \exp\Big(-\frac{x^2-x\mu+\mu^2/2}{2}\Big)\,\mathrm{d}x\\ &=\; C\int \exp\Big(-\frac{(x-\mu/2)^2+\mu^2/4}{2}\Big)\,\mathrm{d}x\\ &=\; \exp(-\mu^2/8)\int \phi(x-\mu/2)\,\mathrm{d}x\\ &=\; \exp(-\mu^2/8).\end{aligned}$$

Exercise 5.8 (Hellinger distance for multivariate Gaussian shift). For any dimension $d \geq 1$, consider arbitrary vectors $\mu_1, \mu_2 \in \mathbb{R}^d$ and a symmetric, positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. Show that

$$\begin{aligned}\eta_{\mathrm{T}}\big(\mathrm{N}_d(\mu_1,\Sigma),\mathrm{N}_d(\mu_2,\Sigma)\big) \;&=\; 2\Phi\Big(-\sqrt{(\mu_2-\mu_1)^\top\Sigma^{-1}(\mu_2-\mu_1)}/2\Big),\\ \eta_{\mathrm{H}}\big(\mathrm{N}_d(\mu_1,\Sigma),\mathrm{N}_d(\mu_2,\Sigma)\big) \;&=\; \exp\big(-(\mu_2-\mu_1)^\top\Sigma^{-1}(\mu_2-\mu_1)/8\big).\end{aligned}$$

Hint: Consider the transformation $\tau : \mathbb{R}^d \to \mathbb{R}^d$, $\tau(x) := \Sigma^{-1/2}(x-\mu_1)$, and use Remark 5.6.

Remark 5.9 (Product measures). For $j = 1, 2$, let P_j and Q_j be probability measures on a measurable space $(\mathcal{X}_j, \mathcal{B}_j)$. Then

$$\eta_{\mathrm{H}}(P_1\otimes P_2, Q_1\otimes Q_2) \;=\; \eta_{\mathrm{H}}(P_1,Q_1)\eta_{\mathrm{H}}(P_2,Q_2).$$

For if $\mathrm{d}P_j/\mathrm{d}M_j = f_j$ and $\mathrm{d}Q_j/\mathrm{d}M_j = g_j$, then by Fubini's theorem,

$$\begin{aligned}\eta_{\mathrm{H}}(P_1\otimes P_2, Q_1\otimes Q_2) \;&=\; \int_{\mathcal{X}_1\times\mathcal{X}_2}\sqrt{f_1(x_1)f_2(x_2)g_1(x_1)g_2(x_2)}\,M_1\otimes M_2(\mathrm{d}(x_1,x_2))\\ &=\; \int_{\mathcal{X}_1}\sqrt{f_1g_1}\,\mathrm{d}M_1\int_{\mathcal{X}_2}\sqrt{f_2g_2}\,\mathrm{d}M_2\\ &=\; \eta_{\mathrm{H}}(P_1,Q_1)\eta_{\mathrm{H}}(P_2,Q_2).\end{aligned}$$

Inductively this implies that

$$\eta_{\mathrm{H}}(P^{\otimes n}, Q^{\otimes n}) \;=\; \eta_{\mathrm{H}}(P,Q)^n$$

for arbitrary integers $n \geq 1$.

5.2 Asymptotics for Repeated Binary Experiments

Suppose we observe independent random variables $X_1, \ldots, X_n$ with unknown distribution $R \in \{P, Q\}$, where P and Q are two different given probability distributions on $(\mathcal{X}, \mathcal{B})$. Then

$$D_{\mathrm{T}}(P^{\otimes n}, Q^{\otimes n}) \;\geq\; D_{\mathrm{H}}(P^{\otimes n}, Q^{\otimes n})^2 \;=\; 1-\eta_{\mathrm{H}}(P,Q)^n$$

converges to 1 exponentially fast. That means, there exists a sequence of tests $\varphi_n : \mathcal{X}^n \to \{0,1\}$ such that

$$\mathbb{E}_P\,\varphi_n(\boldsymbol{X}_n) + \mathbb{E}_Q(1-\varphi_n(\boldsymbol{X}_n)) \;\to\; 0,$$

where $\boldsymbol{X}_n := (X_i)_{i=1}^n$. Throughout this section, asymptotic statements are meant as $n \to \infty$, unless stated otherwise.

More interesting is the situation when P and Q depend on the sample size n. That means, for each sample size $n \geq 1$ we observe $\boldsymbol{X}_n = (X_{ni})_{i=1}^n$ with independent components $X_{n1}, \ldots, X_{nn}$ having unknown distribution $R_n \in \{P_n, Q_n\}$, where P_n and Q_n are different distributions on $(\mathcal{X}, \mathcal{B})$. The question is, under which conditions on $(P_n)_n$ and $(Q_n)_n$ the potential distributions $P_n^{\otimes n}$ and $Q_n^{\otimes n}$ of $\boldsymbol{X}_n$ satisfy one of the following three conditions:

- The are (asymptotically) indistinguishable, i.e.

$$D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;\to\; 0.$$

- They are (asymptotically) perfectly distinguishable, i.e.

$$D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;\to\; 1.$$

- They are (asymptotically) "interesting" in the sense that

$$\liminf_{n\to\infty} D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;>\; 0 \quad \text{and} \quad \limsup_{n\to\infty} D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;<\; 1.$$

It follows from Lemma 5.3 and Exercise 5.5 (a) that the previous three scenarios are equivalent to the analogous ones with D_{H} or D_{H}^2 in place of D_{T}. But note that

$$\begin{aligned} D_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;&=\; 1 - \eta_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n}) \\ &=\; 1 - \eta_{\mathrm{H}}(P_n, Q_n)^n \\ &=\; 1 - \big(1 - D_{\mathrm{H}}(P_n, Q_n)^2\big)^n, \end{aligned}$$

and the subsequent Lemma 5.11 shows that

$$D_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;=\; 1 - \exp\big(-nD_{\mathrm{H}}(P_n, Q_n)^2\big) + O(n^{-1}).$$

This implies the following results:

Lemma 5.10. ***(a)*** *For $a \in \{0, 1\}$,*

$$\lim_{n\to\infty} D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;=\; a \quad \textit{if and only if} \quad \lim_{n\to\infty} nD_{\mathrm{H}}(P_n, Q_n)^2 \;=\; \begin{cases} \infty & \textit{if } a = 1, \\ 0 & \textit{if } a = 0. \end{cases}$$

(b) *As $n \to \infty$, the distance $D_{\mathrm{T}}(P_n^{\otimes n}, Q_n^{\otimes n})$ stays bounded away from 0 and 1 if and only if $nD_{\mathrm{H}}(P_n, Q_n)^2$ stays bounded away from 0 and ∞.*

Lemma 5.11 (Ailam 1968). *For arbitrary real numbers $x \in [0, 1]$ and $n \geq 1$,*

$$0 \;\leq\; e^{-nx} - (1 - x)^n \;\leq\; e^{-1}n^{-1}.$$

Proof of Lemma 5.11. Since $e^y \geq 1 + y$ for arbitrary $y \in \mathbb{R}$, we know that

$$H_n(x) \;:=\; e^{-nx} - (1 - x)^n$$

satisfies the inequality

$$H_n(x) \;=\; (e^{-x})^n - (1-x)^n \;\geq\; 0$$

for arbitrary $x \in [0,1]$. Note also that $H_n(0) = 0$ and $H_n(1) = e^{-n} > 0$. Moreover,

$$H_n'(x) \;=\; n(1-x)^{n-1} - ne^{-nx} \begin{cases} \geq\; 0 & \text{if } n = 1, \\ =\; 0 & \text{if } n > 1 \text{ and } x = 0, \\ <\; 0 & \text{if } n > 1 \text{ and } x = 1. \end{cases}$$

Hence, $\max_{x\in[0,1]} H_1(x) = H_1(0) = e^{-1}$. For $n > 1$, any maximizer x_n of H_n over $[0,1]$ has to satisfy $0 < x_n < 1$ and $H_n'(x_n) = 0$, i.e.

$$(1-x_n)^{n-1} \;=\; e^{-nx_n}.$$

Consequently, $\max_{x\in[0,1]} H_n(x)$ can be written as

$$H_n(x_n) \;=\; e^{-nx_n} - (1-x_n)e^{-nx_n} \;=\; x_n e^{-nx_n} \;\leq\; n^{-1} \max_{s\geq 0} se^{-s} \;=\; n^{-1}e^{-1},$$

because elementary calculations show that se^{-s} is maximized for $s = 1$. □

Example 5.12 (Gaussian distributions). Let $P_n = \mathrm{N}(\mu_n, \sigma_n)$ and $Q_n = \mathrm{N}(\nu_n, \sigma_n)$ with arbitrary means $\mu_n, \nu_n \in \mathbb{R}$ and the same standard deviation $\sigma_n > 0$. Then,

$$D_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;\to\; a \in [0,1]$$

if and only if

$$\frac{\sqrt{n}|\mu_n - \mu_n|}{\sigma_n} \;\to\; \sqrt{-8\log(1-a^2)} \in [0,\infty].$$

This follows immediately from the explicit formula

$$D_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;=\; 1 - \eta_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n}) \;=\; 1 - \exp\Big(-\frac{n(\mu_n - \nu_n)^2}{8\sigma_n^2}\Big).$$

Expansions of root-densities and log-likelihood ratios. Suppose that for some measure M on $(\mathcal{X}, \mathcal{B})$ and arbitrary $n \geq 1$, the densities

$$f_n \;:=\; \frac{\mathrm{d}P_n}{\mathrm{d}M} \quad \text{and} \quad g_n \;:=\; \frac{\mathrm{d}Q_n}{\mathrm{d}M}$$

exist. Suppose that these densities satisfy the following condition:

(C.1) For some probability measure P on $(\mathcal{X}, \mathcal{B})$ with density $f = \mathrm{d}P/\mathrm{d}M$,

$$f_n \;\to\; f \quad \text{in } L^1(M).$$

Furthermore, for some function $h \in L^2(M)$ with $\|h\|_2 > 0$,

$$h_n := \sqrt{n}\big(\sqrt{g_n} - \sqrt{f_n}\big) \;\to\; h \quad \text{in } L^2(M).$$

Here and throughout the sequel, $L^r(M)$ is the space of all (equivalence classes of) measurable functions functions $h : \mathcal{X} \to \mathbb{R}$ such that $\|h\|_r < \infty$, where

$$\|h\|_r \;:=\; \Big(\int |h|^r \,\mathrm{d}M\Big)^{1/r}$$

for $r \geq 1$. (Two functions $h, \tilde{h}$ are viewed as equivalent if $M(h \neq \tilde{h}) = 0$.)

Some first consequences of this condition:

Lemma 5.13. *Under condition (C.1),*

$$nD_{\mathrm{H}}(P_n, Q_n)^2 \;\to\; \|h\|_2^2/2,$$

and

$$D_{\mathrm{T}}(P, P_n), D_{\mathrm{T}}(P, Q_n), D_{\mathrm{T}}(P_n, Q_n) \;\to\; 0.$$

Proof of Lemma 5.13. The convergence of h_n to h in $L^2(M)$ implies that

$$h_n^2 \;\to\; h^2 \quad \text{in } L^1(M),$$

because $\|h_n^2 - h^2\|_1 \leq 2\|h\|_2\|h_n - h\|_2 + \|h_n - h\|_2^2$, see Exercise 5.14 below. In particular,

$$nD_{\mathrm{H}}(P_n, Q_n)^2 \;=\; \|h_n\|_2^2/2 \;\to\; \|h\|_2^2/2.$$

Note that $f_n \to f$ in $L^1(M)$ is equivalent to $D_{\mathrm{T}}(P, P_n) \to 0$, and by Lemma 5.3, $D_{\mathrm{T}}(P_n, Q_n) \leq \sqrt{2}\, D_{\mathrm{H}}(P_n, Q_n) \to 0$. Hence by the triangle inequality, $D_{\mathrm{T}}(P, Q_n) \to 0$ as well. □

Exercise 5.14. Show that for arbitrary functions $g, h \in L^2(M)$,

$$\|g^2 - h^2\|_1 \;\leq\; 2\|h\|_2\|g - h\|_2 + \|g - h\|_2^2.$$

(One can even show that $\big\|\,|g|^r - |h|^r\big\|_1 \leq r\|h\|_r^{r-1}\|g - h\|_r + \|g - h\|_r^r$ for any $r \in (1, 2]$ and arbitrary $g, h \in L^r(M)$.)

For the next result we have to augment condition (C.1) by an additional one:

(C.2) The functions f and h in (C.1) satisfy

$$M(h \neq 0 = f) \;=\; 0.$$

Lemma 5.15. *Suppose that Conditions (C.1-2) are satisfied. Let $(A_n)_n$ be an arbitrary sequence of events $A_n \in \mathcal{B}$ such that*

$$\min\{P_n(A_n), Q_n(A_n)\} \;=\; O(n^{-1}).$$

Then

$$n\big|Q_n(A_n) - P_n(A_n)\big| \;\to\; 0.$$

Proof of Lemma 5.15. In terms of the densities f_n and g_n we may write

$$\begin{aligned} n\big(Q_n(A_n) - P_n(A_n)\big) \;&=\; n\int_{A_n} (g_n - f_n)\,\mathrm{d}M \\ &=\; n\int_{A_n} \big(\sqrt{g_n} - \sqrt{f_n}\big)\big(\sqrt{g_n} + \sqrt{f_n}\big)\,\mathrm{d}M \\ &=\; \begin{cases} \displaystyle\int_{A_n} h_n\big(2\sqrt{nf_n} + h_n\big)\,\mathrm{d}M, \\ \displaystyle\int_{A_n} h_n\big(2\sqrt{ng_n} - h_n\big)\,\mathrm{d}M. \end{cases} \end{aligned}$$

Consequently, by the Cauchy–Schwarz inequality,

$$n\big|Q_n(A_n) - P_n(A_n)\big| \;\le\; 2\sqrt{n\min\{P_n(A_n), Q_n(A_n)\}\int_{A_n} h_n^2\,\mathrm{d}M} + \int_{A_n} h_n^2\,\mathrm{d}M.$$

Hence it suffices to show that

$$\int_{A_n} h_n^2\,\mathrm{d}M \;\to\; 0.$$

Since $h_n^2 \to h^2$ in $L^1(M)$, this is equivalent to

$$\int_{A_n} h^2\,\mathrm{d}M \;\to\; 0.$$

But it follows from Lemma 5.13 and the assumption that $\min\{P_n(A_n), Q_n(A_n)\} = O(n^{-1})$ that $P(A_n) \to 0$. Consequently, for any fixed $C > 0$,

$$\int_{A_n} h^2\,\mathrm{d}M \;\le\; \int_{\{h^2 > Cf\}} h^2\,\mathrm{d}M + CP(A_n) \;\to\; \int_{\{h^2 > Cf\}} h^2\,\mathrm{d}M,$$

and

$$\lim_{C\to\infty} \int_{\{h^2 > Cf\}} h^2\,\mathrm{d}M \;=\; \int_{\{h^2 > 0 = f\}} h^2\,\mathrm{d}M \;=\; 0$$

by dominated convergence and our assumption (C.2) on f and h. □

Implications for log-likelihood ratios. Now we consider again the random observation tuple $\boldsymbol{X}_n = (X_{ni})_{i=1}^n$ with independent components X_{ni} having distribution $R_n \in \{P_n, Q_n\}$. Optimal tests of "$R_n = P_n$" versus "$R_n = Q_n$" are based on the log-likelihood ratio

$$\Lambda_n \;:=\; \log\Big(\frac{\mathrm{d}Q_n^{\otimes n}}{\mathrm{d}M^{\otimes n}}(\boldsymbol{X}_n)\Big/\frac{\mathrm{d}P_n^{\otimes n}}{\mathrm{d}M^{\otimes n}}(\boldsymbol{X}_n)\Big) \;=\; \sum_{i=1}^n \log\frac{g_n}{f_n}(X_{ni}) \;\in\; [-\infty, \infty]$$

with the conventions that $\log(0) := -\infty$, $\log(\infty) := \infty$, $a/0 := \infty$ for $a > 0$ and $log(0/0) := 0$. Indeed, since $P_n(f_n > 0) = 1 = Q_n(g_n > 0)$, the random variable Λ_n is well-defined almost surely, and

$$\mathbb{P}_{P_n}(\Lambda_n < \infty) \;=\; 1, \quad \mathbb{P}_{Q_n}(\Lambda_n > -\infty) \;=\; 1.$$

It may happen with strictly positive probability that $\Lambda_n \in \{-\infty, \infty\}$, but this probability converges to 0. Here is a precise statement:

Theorem 5.16. *Under conditions (C.1-2),*

$$\Lambda_n \;\to_{\mathcal{L}}\; \begin{cases} \mathrm{N}\big(-2\|h\|_2^2, 4\|h\|_2^2\big) & \text{if } R_n = P_n \text{ for all } n, \\ \mathrm{N}\big(+2\|h\|_2^2, 4\|h\|_2^2\big) & \text{if } R_n = Q_n \text{ for all } n. \end{cases}$$

In this result "$\to_{\mathcal{L}}$" stands for convergence in distribution, meaning that for any fixed continuous function $J : [-\infty, \infty] \to \mathbb{R}$,

$$\mathbb{E}\, J(\Lambda_n) \;\to\; \begin{cases} \mathbb{E}\, J(2\|h\|_2 Z - 2\|h\|_2^2) & \text{if } R_n = P_n \text{ for all } n \\ \mathbb{E}\, J(2\|h\|_2 Z + 2\|h\|_2^2) & \text{if } R_n = Q_n \text{ for all } n \end{cases}$$

with a random variable $Z \sim \mathrm{N}(0,1)$. Since the limiting distributions are continuous, and since

$$\mathbb{P}\big(2\|h\|_2 Z \mp 2\|h\|_2^2 \le k\big) \;=\; \Phi\Big(\frac{k \pm 2\|h\|_2^2}{2\|h\|_2}\Big),$$

Theorem 5.16 may be rephrased as follows: For arbitrary $k \in \mathbb{R}$,

$$\mathbb{P}(\Lambda_n \le k) \;\to\; \begin{cases} \Phi\Big(\dfrac{k + 2\|h\|_2^2}{2\|h\|_2}\Big) & \text{if } R_n = P_n \text{ for all } n, \\ \Phi\Big(\dfrac{k - 2\|h\|_2^2}{2\|h\|_2}\Big) & \text{if } R_n = Q_n \text{ for all } n. \end{cases}$$

This implies the following result about optimal tests of "$R_n = P_n$" versus "$R_n = Q_n$":

Corollary 5.17. *Let $\varphi_{n,\alpha} : \mathcal{X}^n \to [0,1]$ be an optimal level-α test of "$R_n = P_n$" versus "$R_n = Q_n$". Then, under conditions (C.1-2),*

$$\mathbb{E}_{Q_n}\, \varphi_{n,\alpha}(\boldsymbol{X}_n) \;\to\; \Phi\big(\Phi^{-1}(\alpha) + 2\|h\|_2\big).$$

The same result holds true for optimal level-α tests of "$R_n = Q_n$" versus "$R_n = A_n$".

Remark 5.18. Corollary 5.17 shows that under conditions (C.1-2), testing $P_n^{\otimes n}$ versus $Q_n^{\otimes n}$ or vice versa is asymptotically as difficult as testing

$$\mathrm{N}(0,1) \quad \text{versus} \quad \mathrm{N}(\mu, 1),$$

where

$$\mu \;:=\; 2\|h\|_2.$$

This is coherent with the fact that

$$D_{\mathrm{H}}(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;=\; 1 - \exp\big(-n D_{\mathrm{H}}(P_n, Q_n)^2\big) + O(n^{-1})$$

converges to

$$1 - \exp(-\|h\|_2^2/2) \;=\; 1 - \exp\big(-\mu^2/8\big) \;=\; D_{\mathrm{H}}\big(\mathrm{N}(0,1), \mathrm{N}(\mu,1)\big)^2.$$

Remark 5.19 (Contiguity). Corollary 5.17 implies that the two sequences $(P_n^{\otimes n})_n$ and $(Q_n^{\otimes n})_n$ are *contiguous* in the sense that for arbitrary measurable sets $A_n \subset \mathcal{X}^n$,

$$\lim_{n\to\infty} P_n^{\otimes n}(A_n) \;=\; 0 \quad \text{if and only if} \quad \lim_{n\to\infty} Q_n^{\otimes n}(A_n) \;=\; 0.$$

Indeed, consider the test $\varphi_n(\boldsymbol{X}_n) := 1_{A_n}(\boldsymbol{X}_n)$. If $P_n^{\otimes n}(A_n) \le \epsilon \in (0,1)$ for sufficiently large n, then Corollary 5.17 implies that

$$\limsup_{n\to\infty} Q_n^{\otimes n}(A_n) \;\le\; \Phi(\Phi^{-1}(\epsilon) + 2\|h\|_2).$$

Likewise, if $Q_n^{\otimes n}(A_n) \le \epsilon$ for sufficiently large n, then

$$\limsup_{n\to\infty} P_n^{\otimes n}(A_n) \;\le\; \Phi(\Phi^{-1}(\epsilon) + 2\|h\|_2).$$

Contiguity follows from the fact that $\Phi(\Phi^{-1}(\epsilon) + 2\|h\|_2) \to 0$ as $\epsilon \to 0$.

Proof of Corollary 5.17. According to the Neyman–Pearson lemma, we may assume that for some constant $k_{n,\alpha} \in [-\infty, \infty)$,

$$\varphi_{n,\alpha}(\boldsymbol{X}_n) \;=\; \begin{cases} 0 & \text{if } \Lambda_n < k_{n,\alpha}, \\ 1 & \text{if } \Lambda_n > k_{n,\alpha}. \end{cases}$$

But for fixed $k \in \mathbb{R}$,

$$\mathbb{P}_{P_n}(\Lambda_n > k) \;\to\; 1 - \Phi\Big(\frac{k + 2\|h\|_2^2}{2\|h\|_2}\Big).$$

The right hand side is strictly decreasing in k and equals α if, and only if, k is equal to

$$k_\alpha \;:=\; -2\|h\|_2^2 + 2\|h\|_2\Phi^{-1}(1-\alpha) \;= -2\|h\|_2^2 - 2\|h\|_2\Phi^{-1}(\alpha).$$

Hence $k_{n,\alpha} \to k_\alpha$, and $\mathbb{E}_{Q_n}\, \varphi_{n,\alpha}(\boldsymbol{X}_n)$ converges to

$$1 - \Phi\Big(\frac{k_\alpha - 2\|h\|_2^2}{2\|h\|_2}\Big) \;=\; 1 - \Phi\big(-2\|h\|_2 - \Phi^{-1}(\alpha)\big) \;=\; \Phi\big(\Phi^{-1}(\alpha) + 2\|h\|_2\big).$$

When testing "$R_n = Q_n$" versus "$R_n = P_n$", we consider tests $\varphi_{n,\alpha}$ such that $\varphi_{n,\alpha}(\boldsymbol{X}_n) = 1$ if $\Lambda_n < k_{n,\alpha}$ and $\varphi_{n,\alpha}(\boldsymbol{X}_n) = 0$ if $\Lambda_n > k_{n,\alpha}$, where $k_{n,\alpha} \in (-\infty, \infty]$. □

Proof of Theorem 5.16. It suffices to consider the case $(R_n)_n = (P_n)_n$, because interchanging the roles of $(P_n)_n$ and $(Q_n)_n$ would result in replacing Λ_n with $-\Lambda_n$, and conditions (C.1-2) would still be satisfied with $-h$ in place of h.

Since $\sqrt{g_n} = \sqrt{f_n} + h_n/\sqrt{n}$, we may write

$$\Lambda_n \;=\; 2\sum_{i=1}^n \log \frac{\sqrt{g_n}}{\sqrt{f_n}}(X_{ni}) \;=\; 2\sum_{i=1}^n \log\Big(1 + \frac{\tilde{h}_n(X_{ni})}{\sqrt{n}}\Big)$$

with $\tilde{h}_n := h_n/\sqrt{f_n} \in \big[-\sqrt{n}, \infty\big]$. It follows from the well-known Taylor series of $\log(1+y)$ for $y \in (-1,1)$ that

$$\log(1+y) \;=\; y - \frac{y^2}{2} + \mathrm{rem}(y) \quad \text{with} \quad |\mathrm{rem}(y)| \;\le\; \frac{|y|^3}{3(1-|y|)^+}$$

for arbitrary $y \in [-1, \infty)$. Consequently, with

$$D_n \;:=\; \max_{1\le i\le n} \frac{|\tilde{h}_n(X_{ni})|}{\sqrt{n}}$$

we obtain the expansion

$$\Lambda_n \;=\; \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\tilde{h}_n(X_{ni}) - \frac{1}{n}\sum_{i=1}^{n}\tilde{h}_n(X_{ni})^2 + \mathrm{Rem}_n, \tag{5.1}$$

$$\text{where}\quad |\mathrm{Rem}_n| \;\le\; \frac{2D_n}{3(1-D_n)^+}\cdot\frac{1}{n}\sum_{i=1}^{n}\tilde{h}_n(X_{ni})^2.$$

Now we apply the Central Limit Theorem as formulated in Corollary A.14: Suppose we can show that

$$\sqrt{n}\int \tilde{h}_n \,\mathrm{d}P_n \;\to\; \mu, \tag{5.2}$$

$$\int \tilde{h}_n^2 \,\mathrm{d}P_n \;\to\; \sigma^2, \tag{5.3}$$

$$\int \tilde{h}_n^2 1_{[\tilde{h}_n^2 \ge n\epsilon]}) \,\mathrm{d}P_n \;\to\; 0 \quad \text{for any fixed } \epsilon > 0. \tag{5.4}$$

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{h}_n(X_{ni}) \;\to_{\mathcal{L}}\; \mathrm{N}(\mu,\sigma^2), \quad \frac{1}{n}\sum_{i=1}^{n}\tilde{h}_n(X_{ni})^2 \;\to_p\; \sigma^2, \quad D_n \;\to_p\; 0.$$

These facts and (5.1) imply that

$$\Lambda_n \;\to_{\mathcal{L}}\; \mathrm{N}(2\mu-\sigma^2, 4\sigma^2).$$

Consequently, it suffices to verify (5.2) with $\mu = -\|h\|_2^2/2$, (5.3) with $\sigma^2 = \|h\|_2^2$ and (5.4).

As to (5.2),

$$\begin{aligned}
\sqrt{n}\int \tilde{h}_n \,\mathrm{d}P_n \;&=\; n\int \frac{\sqrt{g_n}-\sqrt{f_n}}{\sqrt{f_n}}\, f_n \,\mathrm{d}M \\
&=\; n\int \bigl(\sqrt{f_n g_n} - f_n\bigr)\,\mathrm{d}M \\
&=\; n(\eta_{\mathrm{H}}(P_n,Q_n)-1) \\
&=\; -nD_{\mathrm{H}}(P_n,Q_n)^2 \\
&\to\; -\|h\|_2^2/2,
\end{aligned}$$

see Lemma 5.13. Concerning (5.3),

$$\begin{aligned}
\int \tilde{h}_n^2 \,\mathrm{d}P_n \;&=\; \int \frac{h_n^2}{f_n}\, f_n \,\mathrm{d}M \\
&=\; \int_{\{f_n>0\}} h_n^2 \,\mathrm{d}M \\
&=\; \int_{\{f_n>0\}} h^2 \,\mathrm{d}M + o(1) \\
&=\; \int h^2 \,\mathrm{d}M - \int_{\{f_n=0\}} h^2 \,\mathrm{d}M + o(1) \\
&\to\; \|h\|_2^2.
\end{aligned}$$

This follows from the fact that for $C > 0$,

$$\int_{\{f_n=0\}} h^2 \,\mathrm{d}M \;\le\; \int_{\{h^2>Cf\}} h^2 \,\mathrm{d}M + CP(f_n = 0) \;\to\; \int_{\{h^2>Cf\}} h^2 \,\mathrm{d}M,$$

because $P(f_n = 0) = P_n(f_n = 0)+o(1) = o(1)$, and $\int_{\{h^2>Cf\}} h^2 \,\mathrm{d}M \to \int_{\{h^2>0=f\}} h^2 \,\mathrm{d}M = 0$ as $C \to \infty$ by assumption (C.2).

It remains to verify (5.4), that means, for any fixed $\epsilon > 0$,

$$\int_{\{h_n^2 \ge \epsilon^2 n f_n\}} h_n^2 \,\mathrm{d}M \;\to\; 0.$$

Again, since $h_n^2 \to h^2$ in $L^1(M)$, it suffices to show that

$$\int_{\{h_n^2 \ge \epsilon^2 n f_n\}} h^2 \,\mathrm{d}M \;\to\; 0,$$

and the left hand side is equal to

$$\int_0^\infty M(h^2 > r \text{ and } h_n^2 \ge \epsilon^2 n f_n)\, dr \;\le\; \int_0^\infty M(h^2 > r)\, dr \;=\; \int h^2 \,\mathrm{d}M.$$

Consequently, by dominated convergence, it suffices to show that for any fixed $r > 0$,

$$M(h^2 > r \text{ and } h_n^2 \ge \epsilon^2 n f_n) \;\to\; 0.$$

Indeed, it follows from Markov's inequality that for any fixed $\delta > 0$,

$$\begin{aligned}
&M(h^2 > r \text{ and } h_n^2 \ge \epsilon^2 n f_n) \\
&\quad\le\; M\big(h^2 > r \text{ and } h^2 + \delta \ge \epsilon^2 n (f - \delta)\big) + M(|h_n^2 - h^2| \ge \delta) + M(|f_n - f| \ge \delta) \\
&\quad\le\; M\big(h^2 > r \text{ and } h^2 + \delta \ge \epsilon^2 n (f - \delta)\big) + \delta^{-1}\|h_n^2 - h^2\|_1 + \delta^{-1}\|f_n - f\|_1 \\
&\quad\to\; M(h^2 > r \text{ and } f \le \delta).
\end{aligned}$$

Letting $\delta \downarrow 0$, the right hand side converges to $M(h^2 > r \text{ and } f = 0)$, and this equals 0 by assumption (C.2). □

5.3 Fisher Information

Consider a statistical experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ with Θ being an open subset of $\mathbb{R}^d$. Suppose that each P_θ is given by a density $f_\theta > 0$ with respect to some measure M on $(\mathcal{X}, \mathcal{B})$ such that for any $x \in \mathcal{X}$,

$$\Theta \ni \theta \;\mapsto\; f_\theta(x)$$

is continuously differentiable with gradient

$$\dot{f}_\theta(x) \;:=\; \Big(\frac{\partial f_\theta(x)}{\partial \theta_i}\Big)_{i=1}^d.$$

Assuming that for any $\theta \in \Theta$, each component of $f_\theta^{-1/2}\dot{f}_\theta$ belongs to $L^2(M)$, the matrix

$$J(\theta) \;:=\; \int \frac{\dot{f}_\theta \dot{f}_\theta^\top}{f_\theta} \,\mathrm{d}M$$

is well-defined and called the *Fisher information (matrix) of $\mathcal{E}$ at θ*.

Here is yet another interpretation of this matrix: Let $\lambda_\theta := \log f_\theta$. Then for any $x \in \mathcal{X}$, the mapping

$$\Theta \ni \theta \;\mapsto\; \lambda_\theta(x)$$

is continuously differentiable with gradient

$$\dot\lambda_\theta(x) \;:=\; \Big(\frac{\dot f_\theta}{f_\theta}(x)\Big)_{i=1}^d,$$

and

$$J(\theta) \;=\; \int \dot\lambda_\theta \dot\lambda_\theta^\top \,\mathrm{d}P_\theta.$$

Let $(\theta_n)_n$ be a sequence in Θ such that for fixed $\theta \in \Theta$ and $\delta \in \mathbb{R}^d$,

$$\sqrt{n}(\theta_n - \theta) \;\to\; \delta.$$

Then the densities f_{θ_n} and f_θ satisfy

$$\sqrt{n}\big(\sqrt{f_{\theta_n}} - \sqrt{f_\theta}\big) \;\to\; h_{\theta,\delta} := \frac{\delta^\top \dot f_\theta}{2\sqrt{f_\theta}}$$

almost everywhere. If in addition

$$\limsup_{n\to\infty}\, n \int \big(\sqrt{f_{\theta_n}} - \sqrt{f_\theta}\big)^2 \,\mathrm{d}M \;\le\; \int h_{\theta,\delta}^2 \,\mathrm{d}M, \tag{5.5}$$

then by Scheffé's theorem, conditions (C.1-2) in the previous section are satisfied with P_θ in place of P_n, P_{θ_n} in place of Q_n and limit function $h = h_{\theta,\delta}$. That means, testing $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$ is asymptotically as difficult as testing $\mathrm{N}(0,1)$ versus $\mathrm{N}(\mu,1)$ with

$$\mu \;=\; 2\|h\|_2 \;=\; \sqrt{\delta^\top J(\theta)\delta}.$$

Condition (5.5) is satisfied, if

$$D_{\mathrm{H}}(P_\theta, P_{\theta+\delta})^2 \;=\; 1 - \eta_{\mathrm{H}}(P_\theta, P_{\theta+\delta}) \;\le\; \frac{\delta^\top J(\theta)\delta}{8} + o(\|\delta\|^2) \quad \text{as } \delta \to 0. \tag{5.6}$$

Definition 5.20 (Regular statistical experiment). A statistical experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ satisfying condition (5.6) for all $\theta \in \Theta$ is called a *regular statistical experiment*.

Note that for $\theta \in \Theta$ and $\delta \in \mathbb{R}^d$ such that $\{\theta + t\delta : t \in [0,1]\} \subset \Theta$, the Cauchy–Schwarz or Jensen's inequality implies that

$$D_{\mathrm{H}}(P_\theta, P_{\theta+\delta})^2 \;=\; \frac{1}{2}\int_{\mathcal{X}} \big(\sqrt{f_{\theta+\delta}} - \sqrt{f_\theta}\big)^2 \,\mathrm{d}M \;=\; \frac{1}{2}\int_{\mathcal{X}} \Big(\int_0^1 \frac{\delta^\top \dot f_{\theta+t\delta}}{2\sqrt{f_{\theta+t\delta}}} \,\mathrm{d}t\Big)^2 \,\mathrm{d}M$$

is not larger than

$$\frac{1}{8}\int_{\mathcal{X}} \int_0^1 \frac{(\delta^\top \dot f_{\theta+t\delta})^2}{f_{\theta+t\delta}} \,\mathrm{d}t \,\mathrm{d}M \;=\; \frac{1}{8}\int_0^1 \delta^\top J(\theta+t\delta)\delta \,\mathrm{d}t.$$

Consequently, a sufficient condition for regularity is given by

$$J(\cdot) \text{ is continuous on } \Theta. \tag{5.7}$$

Example 5.21 (Fisher information in location families). Let $\mathcal{X}$ and Θ be the real line, equipped with Lebesgue measure. Further let $f_\theta(x) = f(x-\theta)$ for some continuously differentiable probability density $f > 0$ such that

$$I(f) \;:=\; \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)}\,\mathrm{d}x \;<\; \infty.$$

Then $\mathcal{E}$ is regular with $J(\theta) = I(f)$ for all $\theta \in \mathbb{R}$. Indeed, $\dot{f}(x-\theta) = -f'(x-\theta)$, whence

$$J(\theta) \;=\; \int_{\mathbb{R}} \frac{f'(x-\theta)^2}{f(x-\theta)}\,\mathrm{d}x \;=\; I(f).$$

Since this is constant in $\theta \in \mathbb{R}$, criterion (5.7) is satisfied, whence the location family $\mathcal{E}$ is regular.

Special case 1. Let f be the standard Gaussian density, $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. Here $f'(x) = -xf(x)$, whence

$$I(f) \;=\; \int_{\mathbb{R}} x^2 f(x)\,\mathrm{d}x \;=\; 1.$$

Special case 2. Let f be the standard logistic density, $f(x) = e^x/(1+e^x)^2$. The corresponding distribution function F is given by $F(x) = e^x/(1+e^x)$, and $f = F(1-F)$. Consequently, $f' = (1-2F)f$, whence

$$I(f) \;=\; \int_{-\infty}^{\infty} (1-2F(x))^2 f(x)\,\mathrm{d}x \;=\; \int_0^1 (1-2u)^2\,\mathrm{d}u \;=\; 1/3.$$

Example 5.22 (Fisher information in natural exponential families). Suppose that

$$f_\theta \;=\; \exp\bigl(\theta^\top T - \kappa(\theta)\bigr)$$

for some measurable mapping $T : (\mathcal{X}, \mathcal{B}) \to \mathbb{R}^d$. Then the experiment $\mathcal{E}$ is regular. Obviously, $f_\theta > 0$. As shown in Exercise 4.7, κ is infinitely often differentiable with

$$\kappa(\theta+\delta) \;=\; \kappa(\theta) + \delta^\top E_\theta(T) + \frac{1}{2}\,\delta^\top \operatorname{Var}_\theta(T)\delta + o(\|\delta\|^2)$$

as $\delta \to 0$. In particular $f_\theta(x)$ is a smooth function of θ with

$$\dot{f}_\theta \;=\; (T - E_\theta(T))f_\theta, \quad \dot{\lambda}_\theta \;=\; T - E_\theta(T),$$

so

$$J(\theta) \;=\; \operatorname{Var}_\theta(T).$$

This matrix is positive definite for any $\theta \in \Theta$, unless M^T is concentrated on some hyperplane in $\mathbb{R}^d$. It is also differentiable in $\theta \in \Theta$, so $\mathcal{E}$ is regular by criterion (5.7).

Remark 5.23 (Smooth transformations of parameters). Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\theta})$ be a regular statistical experiment with Fisher information $J(\cdot)$, and let $\tau : \Psi \to \Theta$ be a diffeomorphism from another open set $\Psi \subset \mathbb{R}^d$ onto Θ. That means, τ is bijective and continuously differentiable with nonsingular Jacobian matrix

$$D\tau(\psi) \;=\; \Bigl(\frac{\partial\tau_i(\psi)}{\partial\psi_j}\Bigr)_{i,j=1}^{d}$$

for any $\psi \in \Psi$. Then the experiment $\tilde{\mathcal{E}} := (\mathcal{X}, \mathcal{B}, (\tilde{P}_\psi)_{\psi\in\Psi})$ with $\tilde{P}_\psi := P_{\tau(\psi)}$ is regular, too, and its Fisher information $\tilde{J}(\cdot)$ is given by

$$\tilde{J}(\psi) \;=\; D\tau(\psi)^\top J(\tau(\psi)) D\tau(\psi).$$

This follows from the fact that $\tilde{P}_\psi$ has density $f_{\tau(\psi)}$ with respect to M, and with $\theta = \tau(\psi)$ the chain rule implies that

$$\frac{\partial \tilde{f}_\psi(x)}{\partial \psi_j} \;=\; \sum_{i=1}^d \frac{\partial f_\theta(x)}{\partial \theta_i} \frac{\partial \tau_i(\psi)}{\partial \psi_j} \;=\; \bigl(D\tau(\psi)^\top \dot{f}_\theta\bigr)_j.$$

Moreover, as $\delta \to 0$,

$$\Delta \;:=\; \tau(\psi+\delta) - \tau(\psi) \;=\; D\tau(\psi)\delta + o(\|\delta\|) \;=\; O(\|\delta\|),$$

whence

$$\begin{aligned} D_{\mathrm{H}}(\tilde{P}_{\psi+\delta}, \tilde{P}_\psi)^2 \;=\; D_{\mathrm{H}}(P_{\theta+\Delta}, P_\theta)^2 \;&\le\; \frac{\Delta^\top J(\theta)\Delta}{8} + o(\|\Delta\|^2) \\ &=\; \frac{\delta^\top D\tau(\psi)^\top J(\theta) D\tau(\psi)\delta}{8} + o(\|\delta\|^2). \end{aligned}$$

Thus, $\tilde{\mathcal{E}}$ inherits property (5.6) from $\mathcal{E}$.

Example 5.24 (Binomial distributions). We observe $X \sim \mathrm{Bin}(n,p)$ with an unknown parameter $p \in (0,1)$. The natural parameter for the experiment $\bigl(\mathrm{Bin}(n,p)\bigr)_{p\in(0,1)}$ is given by $\tau(p) := \log(p/(1-p))$ with sufficient statistic $T(x) = x$, and $\tau : (0,1) \to \mathbb{R}$ is a diffeomorphism with $\tau'(p) = (p(1-p))^{-1}$. Since $\mathrm{Var}_p(X) = np(1-p)$, Fisher information at p is given by $\tilde{J}(p) = \tau'(p)^2 \,\mathrm{Var}_p(X)$, i.e.

$$\tilde{J}(p) \;=\; \frac{n}{p(1-p)}.$$

Note that $\tilde{J}(p) = \mathrm{Var}_p(\widehat{p})^{-1}$ with $\widehat{p}(x) := x/n$, which is not a coincidence as explained later.

Example 5.25 (Poisson distributions). We observe $X \sim \mathrm{Poiss}(\lambda)$ with an unknown parameter $\lambda > 0$. The natural parameter for the experiment $\bigl(\mathrm{Poiss}(\lambda)\bigr)_{\lambda>0}$ is given by $\tau(\lambda) := \log\lambda$ with sufficient statistic $T(x) = x$, and $\tau : (0,\infty) \to \mathbb{R}$ is a diffeomrophism with $\tau'(\lambda) = \lambda^{-1}$. Since $\mathrm{Var}_\lambda(X) = \lambda$, Fisher information at λ is given by $\tilde{J}(\lambda) = \tau'(\lambda)^2 \,\mathrm{Var}_\lambda(X)$, i.e.

$$\tilde{J}(\lambda) \;=\; \lambda^{-1}.$$

Again $\tilde{J}(\lambda) = \mathrm{Var}_\lambda(\widehat{\lambda})^{-1}$ with $\widehat{\lambda}(x) := x$.

Implications for point estimation

With our results about testing in Section 5.2 one can prove various precision bounds for point estimators. We present one particular result which can also be viewed as a very simplified version of the Hájek–Le Cam convolution theorem:

Theorem 5.26 (Asymptotic version of the Cramér–Rao bound). *Let $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ be a regular statistical experiment with Fisher information $J(\cdot)$. For each sample size $n \geq 1$ let $\widehat{\theta}_n : \mathcal{X}^n \to \mathbb{R}^d$ be an estimator such that for a fixed $\theta \in \Theta$,*

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{X}_n) - \theta_n) \;\to_{\mathcal{L}}\; \mathrm{N}_d(0, \Sigma(\theta))$$

whenever $\theta_n = \theta + n^{-1/2}\delta$ for an arbitrary fixed $\delta \in \mathbb{R}^d$ and sufficiently large sample sizes n. Then $J(\theta)$ is positive definite, and

$$\Sigma(\theta) \;\geq\; J(\theta)^{-1}$$

in the sense that $\eta^\top\Sigma(\theta)\eta \geq \eta^\top J(\theta)^{-1}\eta$ for arbitrary $\eta \in \mathbb{R}^d$.

Proof of Theorem 5.26. For arbitrary fixed vectors $\delta, \eta \in \mathbb{R}^d \setminus \{0\}$ define

$$\theta_n \;:=\; \theta + n^{-1/2}\delta$$

(or $\theta_n := \theta$ if $\theta + n^{-1/2}\delta \notin \Theta$) and

$$T_n(\eta) \;:=\; \sqrt{n}(\widehat{\theta}_n(\boldsymbol{X}_n) - \theta)^\top\eta.$$

It follows from the assumption about $\widehat{\theta}_n$ that

$$T_n(\eta) \;\to_{\mathcal{L}}\; \begin{cases} \mathrm{N}\big(0, \eta^\top\Sigma(\theta)\eta\big) & \text{if } \boldsymbol{X}_n \sim P_\theta^{\otimes n}, \\ \mathrm{N}\big(\delta^\top\eta, \eta^\top\Sigma(\theta)\eta\big) & \text{if } \boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}, \end{cases}$$

because (for sufficiently large n)

$$T_n(\eta) \;=\; \sqrt{n}(\widehat{\theta}_n(\boldsymbol{X}_n) - \theta_n)^\top\eta + \delta^\top\eta.$$

Suppose first that $\eta^\top\Sigma(\theta)\eta = 0$. If we take $\delta = \eta$, then

$$\varphi_n(\boldsymbol{X}_n) \;:=\; 1_{[T_n(\eta)\geq\|\eta\|^2/2]}$$

defines a statistical test of $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$ such that

$$\mathbb{E}\,\varphi_n(\boldsymbol{X}_n) \;\to\; \begin{cases} 0 & \text{if } \boldsymbol{X}_n \sim P_\theta^{\otimes n}, \\ 1 & \text{if } \boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}. \end{cases}$$

In particular, $D_{\mathrm{T}}(P_\theta^{\otimes n}, P_{\theta_n}^{\otimes n}) \to 1$, which would be equivalent to $D_{\mathrm{H}}(P_\theta^{\otimes n}, P_{\theta_n}^{\otimes n})^2 \to 1$. But this would contradict the fact that $D_{\mathrm{H}}(P_\theta^{\otimes n}, P_{\theta_n}^{\otimes n})^2 \to 1 - \exp(-\eta^\top J(\theta)\eta/8) < 1$. Consequently, the matrix $\Sigma(\theta)$ has to be positive definite.

For general δ, η,

$$\varphi_n(\boldsymbol{X}_n) \;:=\; 1_{[T_n(\eta)\geq 0]}$$

defines a statistical test of $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$ such that

$$\mathbb{E}\,\varphi_n(\boldsymbol{X}_n) \;\to\; \begin{cases} 0.5 & \text{if } \boldsymbol{X}_n \sim P_\theta^{\otimes n}. \\ \Phi\Big(\dfrac{\delta^\top\eta}{\sqrt{\eta^\top\Sigma(\theta)\eta}}\Big) & \text{if } \boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}. \end{cases}$$

If we would replace φ_n with an optimal level-0.5 test of $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$, then the asymptotic power under the alternative hypothesis would be

$$\Phi\big(\sqrt{\delta J(\theta)\delta}\big).$$

Consequently,

$$\frac{\delta^\top \eta}{\sqrt{\eta^\top \Sigma(\theta)\eta}} \;\le\; \sqrt{\delta^\top J(\theta)\delta}.$$

In particular, taking $\eta = \delta$, we see that $\delta^\top J(\theta)\delta > 0$, so $J(\theta)$ is positive definite too. For general $\delta, \eta \in \mathbb{R}^d \setminus \{0\}$ we get the inequality

$$\delta^\top \eta \;\le\; \sqrt{\eta^\top \Sigma(\theta)\eta}\sqrt{\delta^\top J(\theta)\delta}.$$

Setting $\delta = J(\theta)^{-1}\eta$ yields the inequality

$$\eta^\top J(\theta)^{-1}\eta \;\le\; \eta^\top \Sigma(\theta)\eta$$

for arbitrary $\eta \in \mathbb{R}^d \setminus \{0\}$. □

Example 5.27 (Maximum-likelihood estimation in natural exponential families). Let $\mathcal{E}$ be a natural exponential family with sufficient statistic $T : (\mathcal{X}, \mathcal{B}) \to \mathbb{R}^d$ such that M^T is not concentrated on a hyperplane in $\mathbb{R}^d$. Let $(\theta_n)_n$ be a sequence in Θ with limit $\theta \in \Theta$, and let $\boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}$. Then the log-likelihood function

$$L_n = L_n(\cdot, \boldsymbol{X}_n) : \Theta \;\to\; \mathbb{R}, \quad L_n(\theta) \;:=\; \sum_{i=1}^n \log f_\theta(X_{ni})$$

has the following property: With probability tending to 1, there exists a unique maximizer $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X}_n)$ of L_n, and

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to\; \mathrm{N}_d(0, J(\theta)^{-1}) \;=\; \mathrm{N}_d(0, \mathrm{Var}_\theta(T)^{-1}).$$

To verify these claims, note first that

$$L_n(\theta) \;=\; n(\bar{T}_n^\top \theta - \kappa(\theta)\big)$$

with $\bar{T}_n := n^{-1}\sum_{i=1}^n T(X_{ni})$, and thus

$$\begin{aligned} \nabla L_n(\theta) \;&=\; n(\bar{T}_n - E_\theta(T)), \\ D^2 L_n(\theta) \;&=\; -n\,\mathrm{Var}_\theta(T). \end{aligned}$$

This shows that L_n is strictly concave, whence L_n has a unique maximizer or no maximizer at all.

One can deduce from the multivariate version of Lindeberg's Central Limit Theorem that

$$Z_n \;:=\; \sqrt{n}(\bar{T}_n - E_{\theta_n}(T)) \;\to_{\mathcal{L}}\; \mathrm{N}_d(0, \mathrm{Var}_\theta(T)).$$

Now we introduce the localized log-likelihood function $H_n : \mathbb{R}^d \to [-\infty, \infty)$ with

$$H_n(\delta) \;:=\; L_n(\theta_n + n^{-1/2}\delta) - L_n(\theta_n),$$

where $H_n(\delta) = -\infty$ if, and only if, $\kappa(\theta_n + n^{-1/2}\delta) = \infty$. Then elementary calculations reveal that

$$H_n(\delta) \;=\; Z_n^\top \delta - \frac{1}{2}\,\delta^\top \operatorname{Var}_\theta(T)\delta + \operatorname{Rem}_n(\delta)$$

where

$$\sup_{\delta:\|\delta\|\le C} |\operatorname{Rem}_n(\delta)| \;\to_p\; 0$$

for any fixed $C > 0$. From this one may deduce that with probability tending to one, H_n has a unique maximizer given by

$$\operatorname{Var}_\theta(T)^{-1} Z_n + o_p(1).$$

But this is equivalent to saying that with asymptotic probability 1, the unique maximizer $\widehat{\theta}_n$ of L_n exists and satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;=\; \operatorname{Var}_\theta(T)^{-1} Z_n + o_p(1).$$

In particular,

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to_{\mathcal{L}}\; \mathrm{N}_d\big(0, \operatorname{Var}_\theta(T)^{-1}\big)$$

because $Z_n \to_{\mathcal{L}} \mathrm{N}_d(0, \operatorname{Var}_\theta(T))$.

Example 5.28 (Maximum-likelihood estimation in smoothly parametrized exponential families). Let $\tilde{\mathcal{E}} = (\mathcal{X}, \mathcal{B}, (\tilde{P}_\psi)_{\psi\in\Psi})$ be an exponential family with sufficient statistic $T : \mathcal{X} \to \mathbb{R}^d$, i.e. $\tilde{P}_\psi$ has density

$$\tilde{f}_\psi \;=\; \exp\big(\tau(\psi)^\top T - \kappa(\tau(\psi))\big)$$

for some bijective mapping $\tau : \Psi \to \Theta$, where Ψ and Θ are open subsets of $\mathbb{R}^d$, and

$$\kappa(\theta) \;:=\; \log \int \exp(\theta^\top T)\,\mathrm{d}M.$$

Suppose further that τ is a diffeomorphism.

Now let $(\psi_n)_n$ be a sequence in Ψ with limit $\psi \in \Psi$, and let $\boldsymbol{X}_n \sim \tilde{P}_{\psi_n}^{\otimes n}$. Then the log-likelihood function $L_n = L_n(\cdot, \boldsymbol{X}_n) := \sum_{i=1}^n \log \tilde{f}_\psi(X_{ni})$ has the following property: With asymptotic probability 1, there exists a unique maximizer $\widehat{\psi}_n = \widehat{\psi}_n(\boldsymbol{X}_n)$ of $L_n(\cdot)$, and

$$\sqrt{n}(\widehat{\psi}_n - \psi_n) \;\to_{\mathcal{L}}\; \mathrm{N}_d\big(0, \tilde{J}(\psi)^{-1}\big)$$

with the Fisher information $\tilde{J}(\cdot)$ of $\tilde{\mathcal{E}}$, i.e. $\tilde{J}(\psi) = D\tau(\psi)^\top \operatorname{Var}_\psi(T) D\tau(\psi)$.

With $\theta_n := \tau(\psi_n)$ and $\theta := \tau(\psi)$, it follows from Example 5.27 that with asymptotic probability 1 there exists a unique maximimum likelihood estimator $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X}_n)$ for the experiment $\mathcal{E} = (\mathcal{X}, \mathcal{B}, (P_\theta)_{\theta\in\Theta})$ with $P_{\tau(\psi)} := \tilde{P}_\psi$ such that

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to_{\mathcal{L}}\; \mathrm{N}_d(0, \operatorname{Var}_\psi(T)^{-1}).$$

But then $\widehat{\psi}_n := \tau^{-1}(\widehat{\theta}_n)$ is a maximum likelihood estimator for $\tilde{\mathcal{E}}$, and elementary calculus reveals that

$$\begin{aligned}
\sqrt{n}(\widehat{\psi}_n - \psi_n) \;&=\; D\tau(\psi)^{-1}\sqrt{n}(\widehat{\theta}_n - \theta_n) + o_p(1) \\
&\to_{\mathcal{L}}\; \mathrm{N}_d\big(0, D\tau(\psi)^{-1} \operatorname{Var}_\psi(T)(D\tau(\psi)^{-1})^\top\big) \\
&=\; \mathrm{N}_d\big(0, \tilde{J}(\psi)^{-1}\big).
\end{aligned}$$

Chapter 6

Stein's Identity and Shrinkage Estimators

Consider independent random variables $X_1, \ldots, X_n$ with distribution $\mathrm{N}_q(\mu, \sigma^2 I_q)$, where $\mu \in \mathbb{R}^q$ is unknown and $\sigma > 0$ is given. With similar arguments as in Chapter 1, one can show that the optimal equivariant estimator of μ is given by the sample mean $\bar{X}$, where optimality of an estimator $\widehat{\mu}$ refers to the risk

$$R(\widehat{\mu}, \mu) \;:=\; E_\mu[\|\widehat{\mu} - \mu\|^2]$$

with the standard Euclidean norm $\|\cdot\|$ on $\mathbb{R}^q$. Note also that the sample mean is the maximum-likelihood estimator $\widehat{\mu}_{\mathrm{ML}}$ of μ. Surprisingly, this estimator is *not* admissible in dimension $q \geq 3$. Precisely, there exist non-equivariant estimators $\widehat{\mu}$ such that

$$R(\widehat{\mu}, \mu) \;<\; \sigma^2 q/n \;=\; R(\widehat{\mu}_{\mathrm{ML}}, \mu).$$

In the present chapter, this result will be deduced and embedded in a more general framework. By means of sufficiency and rescaling arguments, one can reduce the statement to the case of $n = 1$ and $\sigma = 1$. Thus we consider just one observation $X \sim \mathrm{N}(\mu, I_q)$ with unknown mean $\mu \in \mathbb{R}^q$.

6.1 Stein's identity

The following theorem provides two versions of a very useful identity.

Theorem 6.1 (C. Stein, L. Isserlis). ***(a)*** *Let $X \sim \mathrm{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$. Then for any absolutely continuous function $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}\,|f'(X)| < \infty$,*

$$\mathbb{E}[(X - \mu) f(X)] \;=\; \sigma^2\, \mathbb{E}\, f'(X).$$

(b) *Let $X = \mu + Z\gamma + Y$ with fixed vectors $\mu, \gamma \in \mathbb{R}^q$ and independent random variables $Z \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}_q(0, \Gamma)$. Further let $f, Df(\cdot, \gamma) : \mathbb{R}^q \to \mathbb{R}$ be measurable functions such that with probability one, $f(X + t\gamma)$ is absolutely continuous in $t \in \mathbb{R}$ with derivative $Df(X + t\gamma, \gamma)$. If $\mathbb{E}\,|Df(X, \gamma)| < \infty$, then $\mathbb{E}\,|Zf(X)| < \infty$ if and only if $\mathbb{E}\,|f(\mu + Y)| < \infty$, and in that case,*

$$\mathbb{E}[Zf(X)] \;=\; \mathbb{E}\, Df(X, \gamma).$$

Corollary 6.2. *Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors with a Gaussian joint distribution. If $f : \mathbb{R}^q \to \mathbb{R}$ is continuously differentiable such that $\mathbb{E}\,\|(X - \mathbb{E}\,X)f(Y)\| + \mathbb{E}\,\|\nabla f(Y)\| < \infty$, then*

$$\mathbb{E}[(X - \mathbb{E}\,X)f(Y)] \;=\; \mathrm{Cov}(X, Y)\,\mathbb{E}[\nabla f(Y)].$$

Proof of Theorem 6.1. As to part (a), note that X is distributed like $\mu + \sigma Z$ with $Z \sim \mathrm{N}(0, 1)$. Let ϕ be the standard Gaussian density. It follows from $\phi'(z) = -z\phi(z)$, $\lim_{|z|\to\infty} \phi(z) = 0$ and Fubini's theorem that

$$\begin{aligned}
\sigma^2\,\mathbb{E}\,\big|f'(X)\big| \;&=\; \sigma^2 \int_{\mathbb{R}} |f'(\mu + \sigma z)|\phi(z)\,\mathrm{d}z \\
&=\; \sigma^2 \int_{\mathbb{R}} |f'(\mu + \sigma z)| \int_{\mathbb{R}} (1_{[0<z<t]} + 1_{[t\le z\le 0]})|t|\phi(t)\,\mathrm{d}t\,\mathrm{d}z \\
&=\; \sigma^2 \int_{\mathbb{R}} |t|\phi(t) \int_{\mathbb{R}} (1_{[0<z<t]} + 1_{[t\le z\le 0]})|f'(\mu + \sigma z)|\,\mathrm{d}z\,\mathrm{d}t \\
&\ge\; \sigma \int_{\mathbb{R}} |t|\phi(t)|f(\mu + \sigma t) - f(\mu)|\,\mathrm{d}t \\
&=\; \mathbb{E}\big[|X - \mu||f(X) - f(\mu)|\big] \\
&\ge \mathbb{E}\big[|X - \mu||f(X)|\big] - (2/\pi)^{1/2}\sigma|f(\mu)|.
\end{aligned}$$

Thus, finiteness of $\mathbb{E}\,|f'(X)|$ implies finiteness of $\mathbb{E}\big[|X - \mu||f(X)|\big]$, and we may repeat the previous calculations without absolute values. This leads to

$$\begin{aligned}
\sigma^2\,\mathbb{E}\,f'(X) \;&=\; \sigma^2 \int_{\mathbb{R}} f'(\mu + \sigma z)\phi(z)\,\mathrm{d}z \\
&=\; \sigma^2 \int_{\mathbb{R}} f'(\mu + \sigma z) \int_{\mathbb{R}} (1_{[0<z<t]} + 1_{[t\le z\le 0]})|t|\phi(t)\,\mathrm{d}t\,\mathrm{d}z \\
&=\; \sigma^2 \int_{\mathbb{R}} |t|\phi(t) \int_{\mathbb{R}} (1_{[0<z<t]} + 1_{[t\le z\le 0]})f'(\mu + \sigma z)\,\mathrm{d}z\,\mathrm{d}t \\
&=\; \sigma \int_{\mathbb{R}} |t|\phi(t)\,\mathrm{sign}(t)[f(\mu + \sigma t) - f(\mu)]\,\mathrm{d}t \\
&=\; \sigma \int_{\mathbb{R}} t\phi(t)[f(\mu + \sigma t) - f(\mu)]\,\mathrm{d}t \\
&=\; \mathbb{E}\big[(X - \mu)[f(X) - f(\mu)]\big] \\
&=\; \mathbb{E}[(X - \mu)f(X)],
\end{aligned}$$

because $\mathbb{E}(X - \mu) = 0$.

As to part (b), it follows from the previous calculations, applied to Z in place of X, that with probability one,

$$\begin{aligned}
\mathbb{E}\big[|Df(X, \gamma)|\,\big|\,Y\big] \;&=\; \mathbb{E}\big[|Df(\mu + Y + Z\gamma, \gamma)|\,\big|\,Y\big] \\
&\ge\; \mathbb{E}\Big[\big|Z\big(f(\mu + Y + Z\gamma) - f(\mu + Y)\big)\big|\,\Big|\,Y\Big] \\
&=\; \mathbb{E}\Big[\big|Z\big(f(X) - f(\mu + Y)\big)\big|\,\Big|\,Y\Big] \\
&\ge\; \begin{cases} \mathbb{E}\big[|Zf(X)|\,\big|\,Y\big] - (2/\pi)^{1/2}|f(\mu + Y)|, \\ (2/\pi)^{1/2}|f(\mu + Y)| - \mathbb{E}\big[|Zf(X)|\,|\,Y\big], \end{cases}
\end{aligned}$$

where we used independence of Z and Y. Consequently,

$$\mathbb{E}\,|Zf(X)| \;\le\; \mathbb{E}\,|Df(X,\gamma)| + (2/\pi)^{1/2}\,\mathbb{E}\,|f(\mu+Y)|$$

and

$$(2/\pi)^{1/2}\,\mathbb{E}\,|f(\mu+Y)| \;\le\; \mathbb{E}\,|Df(X,\gamma)| + \mathbb{E}\,|Zf(X)|.$$

This proves the equivalence of the conditions $\mathbb{E}\,|Zf(X)| < \infty$ and $\mathbb{E}\,|f(\mu+Y)| < \infty$, provided that $\mathbb{E}\,|Df(X,\gamma)| < \infty$. Assuming these inequalities, the same calculation without absolute values shows that

$$\mathbb{E}[Df(X,\gamma)\,|\,Y] \;=\; \mathbb{E}[Zf(X)\,|\,Y]$$

almost surely, and taking the expectation of both sides yields the asserted equality. □

Proof of Corollary 6.2. It suffices to consider the case $p = 1$ and $\sigma := \mathrm{Var}(X)^{1/2} > 0$, because $\mathbb{E}[(X - \mathbb{E}\,X)f(Y)]$ and $\mathrm{Cov}(X,Y)\,\mathbb{E}[\nabla f(Y)]$ are vectors in $\mathbb{R}^p$ with i-th component equal to $\mathbb{E}[(X_i - \mathbb{E}\,X_i)f(Y)]$ and $\mathrm{Cov}(X_i,Y)\,\mathbb{E}[\nabla f(Y)]$, respectively, and the latter two would be zero if $\mathrm{Var}(X_i) = 0$.

Let $\tilde{X} := [X, Y^\top]^\top \in \mathbb{R}^{1+q}$. Then $\tilde{X} \sim \mathrm{N}_{1+q}(\tilde{\mu}, \tilde{\Sigma})$ with $\tilde{\mu} = [\mathbb{E}\,X, (\mathbb{E}\,Y)^\top]^\top$ and

$$\tilde{\Sigma} \;=\; \begin{bmatrix} \sigma^2 & \sigma\gamma^\top \\ \sigma\gamma & \mathrm{Var}(Y) \end{bmatrix},$$

where $\gamma := \sigma^{-1}\,\mathrm{Cov}(Y,X)$. Following the recipe in Section A.6, we write $\tilde{X} = \tilde{\mu} + Z\tilde{\gamma} + \tilde{Y}$, where $Z := \sigma^{-1}(X - \mathbb{E}\,X) \sim \mathrm{N}(0,1)$ and $\tilde{\gamma} := \mathrm{Cov}(\tilde{X}, Z) = [\sigma, \gamma^\top]^\top$, and the remainder $\tilde{Y}$ is stochastically independent from Z. With $\tilde{f}(\tilde{X}) := f(Y)$, the function $\tilde{f}(\tilde{X} + t\tilde{\gamma})$ is continuously differentiable in $t \in \mathbb{R}$ with derivative

$$D\tilde{f}(\tilde{X} + t\tilde{\gamma}, \tilde{\gamma}) \;=\; \tilde{\gamma}^\top \nabla \tilde{f}(\tilde{X} + t\tilde{\gamma}) \;=\; \gamma^\top \nabla f(Y + t\gamma).$$

Hence, Theorem 6.1 (b) yields the equation

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}\,X)f(Y)] \;=\; \sigma\,\mathbb{E}[Z\tilde{f}(\tilde{X})] \;=\; \sigma\,\mathbb{E}\,D\tilde{f}(\tilde{X},\tilde{\gamma}) \;&=\; \sigma\,\mathbb{E}[\gamma^\top \nabla f(Y)] \\ &=\; \mathrm{Cov}(X,Y)\,\mathbb{E}[\nabla f(Y)]. \end{aligned}$$

□

6.2 Shrinkage estimators

The setting and some heuristics. Suppose that we observe a random vector

$$X \;\sim\; \mathrm{N}_q\,(\mu, I_q)$$

with unknown mean $\mu \in \mathbb{R}^q$. That is, we observe $X = \mu + Z$ with an unobserved error vector $Z = (Z_i)_{i=1}^q \sim \mathrm{N}_q\,(0, I_q)$. The question is how to estimate μ by some estimator $\widehat{\mu} = \widehat{\mu}(X)$. The imprecision of such an estimator is quantified by its risk

$$R(\widehat{\mu}, \mu) \;:=\; E_\mu\big[\|\widehat{\mu} - \mu\|^2\big].$$

The maximum-likelihood estimator in this setting would be given by $\widehat{\mu}_{\mathrm{ML}}(x) = x$, and

$$R\left(\widehat{\mu}_{\mathrm{ML}}, \mu\right) \;=\; q.$$

But now let us consider a more general class of estimators: For $\lambda \in \mathbb{R}$ we consider

$$\widehat{\mu}_{\lambda}(x) \;:=\; \lambda x,$$

so $\widehat{\mu}_{\mathrm{ML}} = \widehat{\mu}_1$. Its risk is given by

$$\begin{aligned} R(\widehat{\mu}_{\lambda}, \mu) \;&=\; \mathbb{E}_{\mu}\left[\|\lambda X - \mu\|^2\right] \\ &=\; \mathbb{E}\left[\|\lambda Z - (1-\lambda)\mu\|^2\right] \\ &=\; \lambda^2 q + (1-\lambda)^2 \|\mu\|^2 \end{aligned}$$

because $\mu^{\top} Z \sim\; \mathrm{N}\left(0, \|\mu\|^2\right)$. As a function of λ and for fixed μ, this risk is minimal if and only if λ equals

$$\lambda(\mu) \;:=\; \frac{\|\mu\|^2}{q + \|\mu\|^2}$$

and the resulting risk is equal to

$$R\left(\widehat{\mu}_{\lambda(\mu)}, \mu\right) \;=\; \frac{q\|\mu\|^2}{q + \|\mu\|^2} < \min\left\{q, \|\mu\|^2\right\}$$

Unfortunately, the optimal value $\lambda(\mu)$ depends on the unknown parameter μ. But

$$\mathbb{E}\left[\|X\|^2\right] \;=\; \mathbb{E}\left[\|Z\|^2 + 2\mu^{\top} Z + \|\mu\|^2\right] \;=\; q + \|\mu\|^2,$$

so one could try to estimate $\lambda(\mu)$ by $\left(\|X\|^2 - q\right)^+ / \|X\|^2$, where $a^+ := \max(a, 0)$. This leads to the preliminary estimator

$$\widehat{\mu}(x) \;:=\; \frac{\left(\|x\|^2 - q\right)^+}{\|x\|^2}\, x \;=\; \left(1 - \frac{q}{\|x\|^2}\right)^+ x$$

considered by Stein (1956).

Concerning $\|X\|^2$ as an estimator of $\|\mu\|^2 + q$, note that

$$\begin{aligned} \mathbb{E}_{\mu}\left[\left(\frac{\|X\|^2}{q + \|\mu\|^2} - 1\right)^2\right] \;&=\; \mathbb{E}\left[\left(\frac{\|Z\|^2 + 2\mu^{\top} Z + \|\mu\|^2}{q + \|\mu\|^2} - 1\right)^2\right] \\ &=\; \frac{\mathbb{E}\left[\left(\|Z\|^2 - q + 2\mu^{\top} Z\right)^2\right]}{\left(q + \|\mu\|^2\right)^2} \\ &=\; \frac{2q + 4\|\mu\|^2}{\left(q + \|\mu\|^2\right)^2} \;\leq\; \frac{4}{q + \|\mu\|^2} \leq \frac{4}{q}, \end{aligned}$$

where we used the facts that $\mathbb{E}\left[\left(\|Z\|^2 - q\right)\mu^{\top} Z\right] = 0$ by symmetry of the distribution of Z, and that $\mathbb{E}\left[\left(\|Z\|^2 - q\right)^2\right] = \mathrm{Var}\left(\|Z\|^2\right) = q\,\mathrm{Var}\left(Z_1^2\right) = 2q$. Hence the relative error of $\|X\|^2$ as an estimator of $\|\mu\|^2 + q$ converges to 0 as $q \to \infty$, uniformly in μ.

The James–Stein estimator. James and Stein (1961) proposed the estimators

$$\widehat{\mu}_{\mathrm{JS}}(X) \;:=\; \Big(1-\frac{q-2}{\|X\|^2}\Big)X \quad \text{and} \quad \widehat{\mu}_{\mathrm{JS}+}(X) \;:=\; \Big(1-\frac{q-2}{\|X\|^2}\Big)^{+}X$$

and proved that for both versions of $\widehat{\mu}$,

$$R(\widehat{\mu},\mu) \;<\; q$$

for all $q \geq 3$ and $\mu \in \mathbb{R}^q$.

As shown by Efron and Morris (1973), $\widehat{\mu}_{\mathrm{JS}}$ can be viewed as an *empirical Bayes estimator*: Suppose that (μ, X) is a random pair, where $\mu \sim \mathrm{N}_q\,(0, \beta I_q)$ and $X \mid \mu \sim \mathrm{N}_q\,(\mu, I_q)$. In this setting,

$$\begin{bmatrix}\mu\\ X\end{bmatrix} \;\sim \mathrm{N}_{2q}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\beta I_q & \beta I_q\\ \beta I_q & (1+\beta)I_q\end{bmatrix}\right),$$

and the conditional distribution of μ, given X, is equal to

$$\mathrm{N}_q\left(\frac{\beta}{1+\beta}X, \frac{\beta}{1+\beta}I_q\right),$$

see Section A.6. Hence, the Bayes-optimal predictor of μ is given by

$$\widehat{\mu}^{\mathrm{B}}_{\beta}(X) \;:=\; \frac{\beta}{1+\beta}X \;=\; \left(1-\frac{1}{1+\beta}\right)X.$$

If β itself is an unknown parameter, one could try to estimate $1/(1+\beta) = q/\,\mathbb{E}\left[\|X\|^2\right]$ by $q/\|X\|^2$. But this estimator would be biased. Indeed, elementary calculations reveal that

$$\mathbb{E}\left[Y^{-1}\right] = b^{-1}(a-1)^{-1} \quad \text{for } Y \sim \mathrm{Gamma}(a,b),\ a>1,\ b>0, \tag{6.1}$$

where b stands for a scale parameter (not a rate parameter). In the present Bayesian setting, $X \sim \mathrm{N}_q\,(0, (1+\beta)I_q)$, so

$$\|X\|^2 \;\sim\; \mathrm{Gamma}\big(q/2, 2(1+\beta)\big),$$

whence

$$\mathbb{E}\left[\|X\|^{-2}\right] = \frac{1}{(q-2)(1+\beta)}$$

which motivates the proposed estimator $\widehat{\mu}_{\mathrm{JS}}$.

The risk of shrinkage estimators. Coming back to the setting of $X \sim \mathrm{N}_q\,(\mu, I_q)$ with an unknown fixed vector $\mu \in \mathbb{R}^q$, we want to compute and bound the risk $R\,(\widehat{\mu}_{\mathrm{JS}}, \mu)$ in case of $q \geq 3$. To this end we write

$$\widehat{\mu}_{\mathrm{JS}}(X) \;=\; X - g_{\mathrm{JS}}(X)$$

with

$$g_{\mathrm{JS}}(X) \;:=\; (q-2)\|X\|^{-2}X$$

In general, let

$$\widehat{\mu}(X) \;=\; X - g(X)$$

with a measurable function $g : \mathbb{R}^q \to \mathbb{R}^q$ such that $\mathbb{E}\left[\|g(X)\|^2\right] < \infty$. Then,

$$R(\widehat{\mu}, \mu) \;=\; \mathbb{E}\big[\|X - \mu - g(X)\|^2\big] \;=\; q - 2\,\mathbb{E}\big[(X-\mu)^\top g(X)\big] + \mathbb{E}\big[\|g(X)\|^2\big].$$

With the standard basis $e_1, \ldots, e_q$ of $\mathbb{R}^q$, suppose that each component g_i of g satisfies the following conditions: With probability one, $g_i\,(X + te_i)$ is absolutely continuous in $t \in \mathbb{R}$ with derivative $Dg_i(X + te_i, e_i)$, and $\mathbb{E}\,|Dg_i(X, e_i)| < \infty$. Then we may apply Stein's identity componentwise and deduce that

$$\begin{aligned} R(\widehat{\mu}, \mu) \;&=\; q - 2\sum_{i=1}^{q} \mathbb{E}[(X_i - \mu_i)g(X)] + \mathbb{E}[\|g(X)\|^2] \\ &=\; q - 2\sum_{i=1}^{q} \mathbb{E}[Dg_i(X, e_i)] + \mathbb{E}[\|g(X)\|^2] \end{aligned}$$

A remarkable feature of this identity is the conclusion that

$$\widehat{R}(X) \;:=\; q - 2\sum_{i=1}^{q} Dg_i(X, e_i) + \|g(X)\|^2$$

is an unbiased estimator of the risk $R(\widehat{\mu}, \mu)$. It is called *Stein's unbiased risk estimator (SURE).*

Specifically, if $g = g_{\mathrm{JS}}$, then

$$Dg_i(X, e_i) \;=\; (q-2)\left(\frac{1}{\|X\|^2} - \frac{2X_i^2}{\|X\|^4}\right)$$

and

$$\sum_{i=1}^{q} Dg_i(X, e_i) \;=\; (q-2)^2\|X\|^{-2}.$$

This leads to

$$\widehat{R}(X) \;=\; q - (q-2)^2\|X\|^{-2}$$

and yields already the first part of the following result:

Theorem 6.3. *For $X \sim \mathrm{N}_q(\mu, I_q)$ with $q \geq 3$,*

$$R(\widehat{\mu}_{\mathrm{JS}}, \mu) \;=\; q - (q-2)^2\,\mathbb{E}[\|X\|^{-2}] \;\leq\; q - \frac{(q-2)^2}{q - 2 + \|\mu\|^2} \;=\; \frac{(q-2)\|\mu\|^2}{q - 2 + \|\mu\|^2} + 2.$$

This shows clearly that

$$R(\widehat{\mu}_{\mathrm{JS}}, \mu) \;<\; q$$

for any $\mu \in \mathbb{R}^q$. Note also that the upper bound in Theorem 6.3 is not larger than

$$2 + \frac{q\|\mu\|^2}{q + \|\mu\|^2} \;=\; R(\widehat{\mu}_{\lambda(\mu)}, \mu) + 2,$$

so the risk of James-Stein estimator is at most the risk of the oracle estimator $\widehat{\mu}_{\lambda(\mu)}$ plus 2.

Proof of Theorem 6.3. The first equation for $R(\widehat{\mu}_{\mathrm{JS}}, \mu)$ is a consequence of the specific form of Stein's unbiased risk estimator $\widehat{R}(X)$ for $\widehat{\mu}_{\mathrm{JS}}$. Note that $\|X\|^2$ follows a noncentral χ^2 distribution with q degrees of freedom and noncentrality parameter $\|\mu\|^2$, see Section A.7. That is, $\|X\|^2$ has the same distribution as

$$\sum_{k=1}^{q+2N} Z_k^2$$

with independent random variables $Z_1, Z_2, Z_3, \ldots \sim \mathrm{N}(0,1)$ and $N \sim \mathrm{Poiss}(\|\mu\|^2/2)$. Consequently, since $\sum_{k=1}^{q+2m} Z_k^2 \sim \mathrm{Gamma}((q+2m)/2, 2)$ and $\mathbb{E}(N) = \|\mu\|^2/2$, equation (6.1) and Jensen's inequality yield the inequality

$$\mathbb{E}[\|X\|^{-2}] \;=\; \mathbb{E}\,\mathbb{E}\Big(\Big(\sum_{k=1}^{q+2N} Z_k^2\Big)^{-1} \,\Big|\, N\Big) \;=\; \mathbb{E}\Big(\frac{1}{q+2N-2}\Big) \;\geq\; \frac{1}{q-2+\|\mu\|^2}.$$

Consequently, $R(\widehat{\mu}_{\mathrm{JS}}, \mu)$ is not larger than

$$q - \frac{(q-2)^2}{q-2+\|\mu\|^2} \;=\; 2 + \frac{(q-2)(q-2+\|\mu\|^2) - (q-2)^2}{q-2+\|\mu\|^2} \;=\; 2 + \frac{(q-2)\|\mu\|^2}{q-2+\|\mu\|^2}.$$

□

Soft thresholding. As an illustration of SURE, consider the soft-thresholding estimator

$$\widehat{\mu}_\lambda^{\mathrm{DJ}}(X) \;:=\; \big((1-\lambda/|X_i|)^+ X_i\big)_{i=1}^q$$

of Donoho and Johnstone (1995), where $\lambda > 0$. This estimator is designed for situations in which preprocessing of raw data leads to a random vector $X \sim \mathrm{N}_q(\mu, I_q)$ with an unknown signal vector $\mu \in \mathbb{R}^q$ such that presumably a large proportion of its components are nearly 0, a so-called *sparse* vector.

Here

$$g_i(X) \;=\; \big(1 - (1-\lambda/|X_i|)^+\big)\, X_i \;=\; \begin{cases} \lambda & \text{if } X_i \geq \lambda \\ X_i & \text{if } |X_i| \leq \lambda \\ -\lambda & \text{if } X_i \leq -\lambda \end{cases}$$

and

$$Dg_i(X, e_i) \;=\; 1_{[|X_i| \leq \lambda]}.$$

Consequently, an unbiased estimator of $R(\widehat{\mu}_\lambda^{\mathrm{DJ}}, \mu)$ is given by

$$\widehat{R}_\lambda^{\mathrm{DJ}}(X) \;=\; q - 2\#\,\{i \leq q : |X_i| \leq \lambda\} + \sum_{i=1}^q \min(X_i^2, \lambda^2).$$

This is a right-continuous function of $\lambda \geq 0$, with discontinuities at the nonzero values among $|X_1|, \ldots, |X_q|$. Between two adjacent points in $\{0, |X_1|, \ldots, |X_n|\}$, it is monotone increasing, so the minimum of $\lambda \mapsto \widehat{R}_\lambda^{\mathrm{DJ}}(X)$ is attained on the latter set.

Exercise 6.4. To compute the risk function of $\widehat{\mu}_\lambda^{\mathrm{DJ}}$ numerically, it suffices to find an explicit formula for dimension $q = 1$. As before, let Φ and ϕ be the standard Gaussian distribution and density function, respectively. Show that for $X \sim \mathrm{N}(\mu, 1)$ and $\lambda \geq 0$,

$$\begin{aligned} R(\widehat{\mu}_\lambda^{\mathrm{DJ}}, \mu) \;&= 1 + \lambda^2 + (\mu^2 - \lambda^2 - 1)\big(\Phi(\mu+\lambda) - \Phi(\mu-\lambda)\big) \\ &\quad -\; (\mu+\lambda)\phi(\mu-\lambda) + (\mu-\lambda)\phi(\mu+\lambda). \end{aligned}$$

Exercise 6.5. Determine Stein's unbiased risk estimator $\widehat{R}(X)$ for the two estimators

$$\widehat{\mu}_{1,\lambda}(X) \;:=\; \big((1-\lambda^2/X_i^2)^+ X_i\big)_{i=1}^q \quad \text{and} \quad \widehat{\mu}_{2,\lambda}(X) \;:=\; \Big(\frac{X_i^3}{\lambda^2+X_i^2}\Big)_{i=1}^q.$$

Do you see potential advantages or disadvantages of these estimators over $\widehat{\mu}_\lambda^{\mathrm{DJ}}$?

Appendix A

Auxiliary Results

A.1 Two Compactness Properties of Statistical Tests

Let M be a σ-finite measure on a measurable space $(\Omega, \mathcal{A})$. Further, let $\mathcal{F} := \mathcal{L}^1(M)$, the set of all $\mathcal{A}$-measurable functions $f : \Omega \to \mathbb{R}$ with $\int |f| \, dM < \infty$, and let $\mathcal{T}$ be the set of all $\mathcal{A}$-measurable functions $\varphi : \Omega \to [0, 1]$. Then the set $\mathcal{T}$ satisfies the following compactness condition:

Theorem A.1 (Weak compactness of $\mathcal{T}$). *The set*

$$\Big\{ \Big(\int \varphi f \, dM \Big)_{f \in \mathcal{F}} : \varphi \in \mathcal{T} \Big\}$$

is a convex and compact subset of $\mathbb{R}^{\mathcal{F}}$, where the latter set is equipped with the usual product topology.

The set $\mathbb{R}^{\mathcal{F}}$ is the set of all tuples $(x_f)_{f \in \mathcal{F}}$ with components $x_f \in \mathbb{R}$. The product topology on this set is the smallest topology such that the mapping $\mathbb{R}^{\mathcal{F}} \ni x \mapsto x_f \in \mathbb{R}$ is continuous for arbitrary $f \in \mathcal{F}$.

Exercise A.2. The proof of Theorem A.1 relies on Tikhonov's theorem, hence on the axiom of choice. Prove a simpler result without this tool in the special case of a countable set Ω and M being the counting measure on Ω: For arbitrary $m \in \mathbb{N}$ and functions $f_1, \ldots, f_m \in \mathcal{L}^1(M)$, the set

$$\Big\{ \Big(\int \varphi f_j \, dM \Big)_{j=1}^{m} : \varphi \in \mathcal{T} \Big\}$$

is a convex and compact subset of $\mathbb{R}^m$.

Proof of Theorem A.1. Writing $f \in \mathcal{F}$ as $f^+ - f^-$ with $f^{\pm} := \max(\pm f, 0)$, we know that

$$\int \varphi f \, dM \;\in\; K_f \;:=\; \Big[-\int f^- \, dM, \int f^+ \, dM \Big]$$

for any $\varphi \in \mathcal{T}$ and $f \in \mathcal{F}$. Hence,

$$\mathcal{K}_* \;:=\; \Big\{ \Big(\int \varphi f \, dM \Big)_{f \in \mathcal{F}} : \varphi \in \mathcal{T} \Big\}$$

is a subset of

$$\mathcal{K} \;:=\; \big\{x \in \mathbb{R}^{\mathcal{F}} : x_f \in K_f \text{ for all } f \in \mathcal{F}\big\}.$$

Since each K_f, $f \in \mathcal{F}$, is a compact interval, it follows from Tikhonov's theorem that $\mathcal{K}$ is a compact subset of $\mathbb{R}^{\mathcal{F}}$.

Linearity of integrals implies that $\mathcal{K}_*$ is a subset of

$$\mathcal{K}_o \;:=\; \big\{x \in \mathcal{K} : f \mapsto x_f \text{ is linear}\big\}.$$

That means, $\mathcal{K}_o$ consists of all tuples $x \in \mathcal{K}$ such that for arbitrary $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned} x_{\lambda f} \;&=\; \lambda x_f, \\ x_{f+g} \;&=\; x_f + x_g. \end{aligned}$$

Note that each of the previous two constraints defines a closed subset of $\mathbb{R}^{\mathcal{F}}$. Hence the set $\mathcal{K}_o$ is a closed subset of $\mathcal{K}$, i.e. it is compact. Moreover, one can easily verify that $\mathcal{K}_o$ is convex.

Now the assertion of Theorem A.1 is true if we can show that $\mathcal{K}_* = \mathcal{K}_o$. That means, we have to show that for any fixed $x \in \mathcal{K}_o$ there exists a test $\varphi \in \mathcal{T}$ such that $x_f = \int \varphi f \, dM$ for arbitrary $f \in \mathcal{F}$. Indeed, it follows from linearity of $f \mapsto x_f$ and the inclusion $x_f \in K_f$ for all $f \in \mathcal{F}$ that $f \mapsto x_f$ defines a continuous linear functional on $\mathcal{L}^1(M)$, equipped with the seminorm $\|f\| := \int |f| \, dM$. Consequently, by Riesz' representation theorem for the dual space of $L^1(M)$, there exists a bounded measurable function $\varphi : \Omega \to \mathbb{R}$ such that $x_f = \int f\varphi \, dM$ for all $f \in \mathcal{F}$. In particular, since $0 \le x_{1_A} = \int_A \varphi \, dM \le M(A)$ for all $A \in \mathcal{A}$, the function φ satisfies $M(\{\varphi < 0\} \cup \{\varphi > 1\}) = 0$. Hence, we may assume that $\varphi \in \mathcal{T}$. □

The next results establishes a sequential compactness property of $\mathcal{T}$. Its proof is constructive in the sense that it does not use the axiom of choice.

Theorem A.3 (Weak sequential compactness of $\mathcal{T}$). *Let $(\varphi_n)_{n\ge1}$ be a sequence in $\mathcal{T}$. Then there exist a subsequence $(\varphi_{n(k)})_{k\ge1}$ and a test $\varphi \in \mathcal{T}$ such that*

$$\lim_{k\to\infty} \int \varphi_{n(k)} f \, dM \;=\; \int \varphi f \, dM \quad \textit{for any } f \in \mathcal{F}.$$

Proof of Theorem A.3. It suffices to consider the case of a probability measure M. For if M is infinite, but σ-finite, there exists a probability measure M_o on $(\Omega, \mathcal{A})$ and a measurable function $g_o : \Omega \to (0, \infty)$ such that M has density g_o with respect to M_o In particular, $f \mapsto f g_o$ defines a linear bijection from $\mathcal{L}^1(M)$ to $\mathcal{L}^1(M_o)$, so we could replace M with M_o and any $f \in \mathcal{F}$ with $f g_o$.

Let $\mathcal{A}_o$ be the σ-field generated by the tests φ_n, $n \ge 1$. This sub-σ-field of $\mathcal{A}$ has a countable generator, for instance, the family of all sets $\{\varphi_n \le q\}$, $n \ge 1$, $q \in \mathbb{Q}$. Consequently, there exists a countable field $\mathcal{A}_{oo}$ of Ω such that $\mathcal{A}_o = \sigma(\mathcal{A}_{oo})$. By Cantor's diagonalisation trick, there exists a subsequence $(\varphi_{n(k)})_{k\ge1}$ such that

$$L(1_{A_{oo}}) \;:=\; \lim_{k\to\infty} \int \varphi_{n(k)} 1_{A_{oo}} \, dM$$

exists for all $A_{oo} \in \mathcal{A}_{oo}$.

Let $\mathcal{F}_*$ be the set of all $f \in \mathcal{F}$ such that the limit

$$L(f) \;:=\; \lim_{k\to\infty} \int \varphi_{n(k)} f \, dM$$

exists. One can easily verify that $\mathcal{F}_*$ is a linear subspace of $\mathcal{F}$, and L is continuous on $\mathcal{F}_*$ with respect to the seminorm $f \mapsto \|f\| := \int |f| \, dM$. Precisely, for any $f \in \mathcal{F}$ and $n \geq 1$,

$$-\int f^- \, dM \;\leq\; \int \varphi_n f \, dM \;\leq\; \int f^+ \, dM,$$

and thus,

$$-\int f^- \, dM \;\leq\; L(f) \;\leq\; \int f^+ \, dM \quad \text{for all } f \in \mathcal{F}_*. \tag{A.1}$$

The space $\mathcal{F}_*$ is also closed with respect to $\|\cdot\|$. For if $(f_\ell)_{\ell\geq 1}$ is a sequence in $\mathcal{F}_*$ with limit $f \in \mathcal{F}$, then for any fixed $\ell \geq 1$,

$$\begin{aligned}
&\limsup_{k,k'\to\infty} \Big| \int \varphi_{n(k)} f \, dM - \int \varphi_{n(k')} f \, dM \Big| \\
&\qquad \leq\; 2\|f - f_\ell\| + \limsup_{k,k'\to\infty} \Big| \int \varphi_{n(k)} f_\ell \, dM - \int \varphi_{n(k')} f_\ell \, dM \Big| \;=\; 2\|f - f_\ell\|,
\end{aligned}$$

and the right hand side tends to 0 as $\ell \to \infty$. Consequently, $\big(\int \varphi_{n(k)} f \, dM\big)_{k\geq 1}$ is a Cauchy sequence in $\mathbb{R}$.

The space $\mathcal{F}_*$ contains all indicator functions $1_{A_{oo}}$, $A_{oo} \in \mathcal{A}_{oo}$, and the linear span of the latter functions is dense in $\mathcal{L}^1(M|_{\mathcal{A}_o})$ with respect to $\|\cdot\|$. Thus, $\mathcal{F}_*$ contains all functions in $\mathcal{L}^1(M|_{\mathcal{A}_o})$. But for an arbitrary function $f \in \mathcal{F}$ and its conditional expectation $f_o := \mathbb{E}(f \,|\, \mathcal{A}_o)$ with respect to the probability measure M,

$$\int \varphi_n f \, dM \;=\; \int \varphi_n f_o \, dM$$

for all $n \geq 1$, so $\mathcal{F}_* = \mathcal{F}$.

It follows from (A.1) that L is a continuous linear functional on $\mathcal{L}^1(M)$, so by Riesz' representation theorem for the dual space of $L^1(M)$, there exists a bounded measurable function $\varphi : \Omega \to \mathbb{R}$ such that $L(f) = \int f\varphi \, dM$ for all $f \in \mathcal{F}$. In particular, since $0 \leq L(1_A) = \int_A \varphi \, dM \leq M(A)$ for all $A \in \mathcal{A}$, the function φ satisfies $M(\{\varphi < 0\} \cup \{\varphi > 1\}) = 0$. Hence, we may assume that $\varphi \in \mathcal{T}$. □

A.2 Scheffé's Theorem

A standard result in measure theory is the theorem about dominated convergence. There is a more general version which provides necessary and sufficient conditions for convergence in $\mathcal{L}^p(M)$, where M is a σ-finite measure on a measurable space $(\mathcal{X}, \mathcal{B})$, and $p \in [1, \infty)$. The space $\mathcal{L}^p(M)$ is equipped with the seminorm $\|\cdot\|_p$,

$$\|f\|_p \;:=\; \Big(\int |f|^p \, \mathrm{d}M\Big)^{1/p}.$$

To state the result, we need the notion of stochastic convergence with respect to M: If f and f_n, $n \in \mathbb{N}$, are measurable real-valued functions on $\mathcal{X}$, we say that $(f_n)_n$ converges stochastically to f with respect to M if for any $B \in \mathcal{B}$ with $M(B) < \infty$ and arbitrary $\epsilon > 0$,

$$\lim_{n\to\infty} M\big(\{|f_n - f| \geq \epsilon\} \cap B\big) \;=\; 0.$$

A short notation for that is $(f_n)_n \to_M f$.

Theorem A.4 (Scheffé). *Let f and f_n, $n \in \mathbb{N}$, be functions in $\mathcal{L}^p(M)$. The following two conditions are equivalent:*

(i) $\lim_{n\to\infty} \|f_n - f\|_p = 0$.

(ii) $(f_n)_n \to_M f$ *and*

$$\limsup_{n\to\infty} \int |f_n|^p \,\mathrm{d}M \;\leq\; \int |f|^p \,\mathrm{d}M.$$

Proof. Suppose that condition (i) holds true. By the triangle inequality for $\|\cdot\|_p$,

$$\big|\|f_n\|_p - \|f\|_p\big| \;\leq\; \|f_n - f\|_p,$$

so $\int |f_n|^p \,\mathrm{d}M = \|f_n\|_p^p$ converges to $\int |f|^p \,\mathrm{d}M = \|f\|_p^p$ as $n \to \infty$. Moreover, for any $\epsilon > 0$,

$$M(\{|f_n - f| \geq \epsilon\}) \;\leq\; \epsilon^{-p} \int |f_n - f|^p \,\mathrm{d}M \;\to\; 0.$$

Hence, condition (ii) holds true as well.

Now suppose that condition (ii) is satisfied. We define

$$\tilde{f}_n \;:=\; \min(|f_n|, |f|)\,\mathrm{sign}(f_n).$$

Note that $|\tilde{f}_n| \leq |f|$ and $|\tilde{f}_n - f| \leq 2|f|$. Moreover, $|f_n - f| \geq |\tilde{f}_n - f|$, whence $(\tilde{f}_n)_n \to_M f$. Now we show that $\lim_{n\to\infty} \|\tilde{f}_n - f\|_p = 0$. To this end we write

$$\int |\tilde{f}_n - f|^p \,\mathrm{d}M \;=\; \int_0^\infty M\big(|\tilde{f}_n - f|^p \geq t\big)\,\mathrm{d}t \;=\; \int_0^\infty h_n(t)\,\mathrm{d}t,$$

where $h_n(t) := M\big(|\tilde{f}_n - f|^p \geq t\big)$. But $|\tilde{f}_n - f| \leq 2|f|$ implies that

$$h_n(t) \;=\; M\big(\{|\tilde{f}_n - f| \geq t^{1/p}\} \cap \{2^p|f|^p \geq t\}\big) \leq\; g(t) \;:=\; M(2^p|f|^p \geq t),$$

and

$$\int_0^\infty g(t)\,\mathrm{d}t \;=\; 2^p \int_0^\infty M(|f|^p \geq s)\,\mathrm{d}s \;=\; 2^p\|f\|_p^p \;<\; \infty.$$

Since $g(t)$ is nonincreasing in $t > 0$, this implies that for any fixed $t > 0$, $g(t)$ is finite, so $(\tilde{f}_n)_n \to_M f$ implies that $\lim_{n\to\infty} h_n(t) = 0$. Hence, by dominated convergence,

$$\lim_{n\to\infty} \|\tilde{f}_n - f\|_p^p \;=\; \lim_{n\to\infty} \int_0^\infty h_n(t)\,\mathrm{d}t \;=\; 0.$$

Finally,

$$\|f_n - f\|_p \;\leq\; \|f_n - \tilde{f}_n\|_p + \|\tilde{f}_n - f\|_p,$$

so it suffices to show that $\limsup_{n\to\infty} \|f_n - \tilde{f}_n\|^p \le 0$. But $|f_n - \tilde{f}_n| = |f_n| - |\tilde{f}_n|$, whence

$$|f_n - \tilde{f}_n|^p \;=\; \big(|f_n| - |\tilde{f}_n|\big)^p \;=\; \int_0^{|f_n|-|\tilde{f}_n|} ps^{p-1}\,\mathrm{d}s \;\le\; \int_{|\tilde{f}_n|}^{|f_n|} ps^{p-1}\,\mathrm{d}s \;=\; |f_n|^p - |\tilde{f}_n|^p.$$

Consequently, since $\lim_{n\to\infty} \|\tilde{f}_n\|_p^p = \|f\|_p^p$,

$$\begin{aligned}\limsup_{n\to\infty}\, \|f_n - \tilde{f}_n\|_p^p \;&\le\; \limsup_{n\to\infty}\Big(\int |f_n|^p\,\mathrm{d}M - \int |\tilde{f}_n|^p\,\mathrm{d}M\Big)\\ &=\; \limsup_{n\to\infty} \int |f_n|^p\,\mathrm{d}M - \int |f|^p\,\mathrm{d}M\\ &\le\; 0\end{aligned}$$

by the second part of condition (ii). This proves condition (i). □

A.3 Uniqueness of Moment-Generating Functions

In the context of completeness of statistical experiments and exponential families we utilize a classical result from measure theory.

Theorem A.5. *Let M be a measure on $\mathbb{R}^d$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that*

$$\int_{\mathbb{R}^d} \exp(u^\top x) f(x)\, M(dx) \;=\; 0$$

for all u in a nonempty open set $U \subset \mathbb{R}^d$. Then

$$M(f \neq 0) \;=\; 0.$$

Proof of Theorem A.5. Suppose first that $0 \in \mathbb{R}^d$ is an interior point of U. Then for some $\epsilon > 0$,

$$L(u) := \int \exp(u^\top x) f(x)\, M(dx) \;=\; 0 \quad \text{for all } u \in (-\epsilon, \epsilon)^d.$$

One can easily verify that $L(u)$ is well-defined in $\mathbb{C}$ for all complex vectors

$$u \;\in\; U_o^d \quad \text{with} \quad U_o \;:=\; \{z \in \mathbb{C} : -\epsilon < \mathrm{Re}\, z < \epsilon\}.$$

Moreover, for any given index $j \in \{1, \ldots, d\}$, $L(u)$ is a holomorphic (i.e. complex differentiable) function of $u_j \in U_o$ while $(u_k)_{k\neq j}$ is fixed; see Exercise A.6. But it is well-known from complex analysis that a holomorphic function $H : U_o \to \mathbb{C}$ with $H \equiv 0$ on $(-\epsilon, \epsilon)$ satisfies $H \equiv 0$ on U_o. Consequently, we may conclude inductively for $j = 1, 2, \ldots, d$ that

$$\begin{aligned} L \equiv 0 \text{ on } (-\epsilon,\epsilon)^d, \quad &\text{whence} \quad L \equiv 0 \text{ on } U_o \times (-\epsilon,\epsilon)^{d-1},\\ &\text{whence} \quad L \equiv 0 \text{ on } U_o^2 \times (-\epsilon,\epsilon)^{d-2},\\ &\cdots\\ &\text{whence} \quad L \equiv 0 \text{ on } U_o^d. \end{aligned}$$

Since $\{\boldsymbol{i}y : y \in \mathbb{R}^d\} \subset U_o^d$ (with $\boldsymbol{i} = \sqrt{-1}$), the latter equality for L implies that

$$\int \exp(\boldsymbol{i}y^\top x) f^+(x)\, M(dx) \;=\; \int \exp(\boldsymbol{i}y^\top x) f^-(x)\, M(dx) \quad \text{for all } y \in \mathbb{R}^d.$$

That means, the characteristic functions of the finite measures Q^+ and Q^-, where

$$Q^\pm(A) \;:=\; \int_A f^\pm\, dM,$$

are identical. But a finite measure is uniquely determined by its characteristic function, so $Q^+ \equiv Q^-$. In particular, $Q^\pm(\mathbb{R}^d) = Q^\pm(f^\pm > 0) = Q^\mp(f^\pm > 0) = 0$, and this implies that $M(f \neq 0) = 0$.

In case of a general open set U, let u_o be an interior point of U. Then $V := U - u_o$ is an open neighborhood of 0, and with $g(x) := \exp(u_o^\top x) f(x)$ the assumption reads

$$\int \exp(v^\top x) g(x)\, M(dx) \;=\; 0 \quad \text{for all } v \in V.$$

But then the previous considerations show that $M(f \neq 0) = M(g \neq 0) = 0$. □

Exercise A.6. Let M be a measure on a measurable space $(\Omega, \mathcal{A})$, and let $g : \Omega \to \mathbb{R}$, $h : \Omega \to \mathbb{C}$ be $\mathcal{A}$-measurable functions such that for real numbers $a < b$,

$$\int (e^{ag} + e^{bg})|h|\, dM \;<\; \infty.$$

Show that

$$L(z) \;:=\; \int h \exp(zg)\, dM$$

defines a holomorphic function on $\{z \in \mathbb{C} : a < \mathrm{Re}(z) < b\}$.

A.4 Hoeffding's Decomposition

Hoeffding's decomposition is a generalization of Hájek's projection as described in Lemma 3.40. The setting is the same, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with stochastically independent random variables $X_1, \ldots, X_n$ with values in $(\mathcal{X}_1, \mathcal{B}_1), \ldots, (\mathcal{X}_n, \mathcal{B}_n)$, respectively. Now we consider the Hilbert space $\mathbb{H}$ of all random variables $Y \in L^2(\mathbb{P})$ which are a measurable function of the random tuple $X := (X_1, \ldots, X_n)$ with values in $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

For any nonvoid set $K \subset \{1, \ldots, n\}$ let $\mathbb{H}_K$ be the subspace of all random variables $Y \in \mathbb{H}$ which are a measurable function of

$$X_K \;:=\; (X_i)_{i \in K},$$

and let $\mathbb{H}_\emptyset$ be the subspace of all constant random variables. In particular, $\mathbb{H} = \mathbb{H}_{\{1,\ldots,n\}}$. The orthogonal projection of $\mathbb{H}$ onto $\mathbb{H}_K$ is given by Π_K with

$$\Pi_K Y \;:=\; \begin{cases} \mathbb{E}(Y) & \text{if } K = \emptyset \\ \mathbb{E}(Y \mid X_K) & \text{else} \end{cases}$$

for $Y \in \mathbb{H}$. Strictly speaking, we should write $\mathbb{E}(Y \,|\, \sigma(X_K))$, but $\mathbb{E}(Y \,|\, X_K)$ is more convenient and intuitive. One treats X_K temporarily as a fixed tuple, and if $Y = f(X_K, X_L)$ with $L = \{1, \ldots, n\} \setminus K$, then

$$\mathbb{E}(Y \,|\, X_K) \;=\; \int f(X_K, \boldsymbol{z})\, P_L(d\boldsymbol{z}),$$

where P_L denotes the distribution of X_L.

A key property of these projections Π_K is that

$$\Pi_J \Pi_K \;=\; \Pi_{J \cap K} \quad \text{for arbitrary } J, K \subset \{1, \ldots, n\}. \tag{A.2}$$

This can be easily derived from Fubini's theorem. In particular, $\Pi_J \Pi_K = \Pi_K \Pi_J$. Now we define

$$\Pi^o_K \;:=\; \sum_{I \subset K} (-1)^{\#(K \setminus I)} \Pi_I.$$

The next result shows that Π^o_K describes an orthogonal projection, too. And the corresponding subspaces $\Pi^o_K \mathbb{H}$, $K \subset \{1, \ldots, n\}$, comprise a decomposition of $\mathbb{H}$ into pairwise orthogonal subspaces.

Theorem A.7. ***(a)*** *For arbitrary sets $K \subset \{1, \ldots, n\}$.*

$$\Pi_K \;=\; \sum_{J \subset K} \Pi^o_J.$$

In particular, the identity operator I may be written as

$$I \;=\; \sum_{K \subset \{1, \ldots, n\}} \Pi^o_K.$$

(b) *For arbitrary sets $J, K \in \{1, \ldots, n\}$,*

$$\Pi^o_J \Pi_K \;=\; \Pi_K \Pi^o_J \;=\; 1_{[J \subset K]} \Pi^o_J,$$

and

$$\Pi^o_J \Pi^o_K \;=\; 1_{[J = K]} \Pi^o_J.$$

(c) *Each operator Π^o_K describes the orthogonal projection of $\mathbb{H}$ onto the linear space*

$$\mathbb{H}^o_K \;:=\; \mathbb{H}_K \cap \Big(\sum_{J \subsetneq K} \mathbb{H}_J \Big)^{\perp}.$$

These spaces $\mathbb{H}^o_K$, $K \subset \{1, \ldots, n\}$, are pairwise orthogonal.

Corollary A.8 (Hoeffding's decomposition). *Any random variable $Y \in \mathbb{H}$ can be written as*

$$Y \;=\; \sum_{K \subset \{1, \ldots, n\}} \Pi^o_K Y,$$

and the random variables $\Pi^o_K Y$, $K \subset \{1, \ldots, n\}$, are uncorrelated with $\Pi^o_\emptyset Y \equiv \mathbb{E}(Y)$ and $\mathbb{E}(\Pi^o_K Y) = 0$ if $K \neq \emptyset$.

For an additional random variable $Z \in \mathbb{H}$,

$$\mathbb{E}(YZ) \;=\; \sum_{K \subset \{1, \ldots, n\}} \mathbb{E}(\Pi^o_K Y \, \Pi^o_K Z).$$

This corollary follows essentially from Theorem A.7, except for the statements about $\Pi^o_\emptyset Y$ and $\mathbb{E}(\Pi^o_K Y)$. But $\mathbb{H}^o_\emptyset = \mathbb{H}_\emptyset$ is the space of constants, so $\Pi^o_\emptyset Y = \Pi_\emptyset Y \equiv \mathbb{E}(Y)$, and for any nonempty set $K \subset \{1,\dots,n\}$, it follows from $\mathbb{H}^o_K \perp \mathbb{H}_\emptyset$ that $\mathbb{E}(\Pi^o_K Y) = \langle \Pi^o_K Y, 1\rangle = 0$.

Example A.9 (The case $n = 2$). Suppose that $Y = f(X_1, X_2)$. Then the Hoeffding decomposition of Y reads

$$Y - \mathbb{E}(Y) \;=\; f^o_1(X_1) + f^o_2(X_2) + f^o_{12}(X_1, X_2).$$

The three summands on the right hand side are given by

$$f^o_1(x_1) \;:=\; \mathbb{E}\, f(x_1, X_2) - \mathbb{E}(Y), \quad f^o_2(x_2) \;:=\; \mathbb{E}\, f(X_1, x_2) - \mathbb{E}(Y)$$

and

$$f^o_{12}(x_1, x_2) \;:=\; f(x_1, x_2) - \mathbb{E}\, f(x_1, X_2) - \mathbb{E}\, f(X_1, x_2) + \mathbb{E}(Y).$$

Moreover, for arbitrary random variables $g_1(X_1)$ and $g_2(X_2)$ in $L^2(\mathbb{P})$,

$$\mathbb{E}\big(f^o_1(X_1)g_2(X_2)\big) \;=\; 0 \;=\; \mathbb{E}\big(f^o_{12}(X_1, X_2)g_2(X_2)\big)$$

and

$$\mathbb{E}\big(f^o_2(X_2)g_1(X_1)\big) \;=\; 0 \;=\; \mathbb{E}\big(f^o_{12}(X_1, X_2)g_1(X_1)\big).$$

In particular, the random variables $f^o_1(X_1)$, $f^o_2(X_2)$ and $f^o_{12}(X_1, X_2)$ are centered and uncorrelated.

Proof of Theorem A.7. We start with a simple combinatorial fact. For any finite set S,

$$\sum_{L\subset S} (-1)^{\#L} \;=\; 1_{[S=\emptyset]}. \tag{A.3}$$

This follows essentially from the binomial formula, because

$$\begin{aligned}\sum_{L\subset S} (-1)^{\#L} \;&=\; \sum_{\ell=0}^{\#S} \#\{L \subset S : \#L = \ell\}(-1)^\ell \\ &=\; \sum_{\ell=0}^{\#S} \binom{\#S}{\ell}(-1)^\ell(+1)^{\#S-\ell} \;=\; (1-1)^{\#S} \;=\; 1_{[S=\emptyset]}.\end{aligned}$$

As to part (a), by definition of Π^o_K and formula (A.3),

$$\begin{aligned}\sum_{J\subset K} \Pi^o_J \;&=\; \sum_{J\subset K}\sum_{I\subset J} (-1)^{\#(J\setminus I)}\Pi_I \\ &=\; \sum_{I\subset K}\Big(\sum_{J\subset K: I\subset J} (-1)^{\#(J\setminus I)}\Big)\Pi_I \\ &=\; \sum_{I\subset K}\Big(\sum_{L\subset K\setminus I} (-1)^{\#L}\Big)\Pi_I \;=\; \sum_{I\subset K} 1_{[K\setminus I=\emptyset]}\Pi_I \;=\; \Pi_K.\end{aligned}$$

As to part (b), it follows from (A.2) that

$$\left.\begin{matrix}\Pi_J^o\Pi_K\\ \Pi_K\Pi_J^o\end{matrix}\right\} \;=\; \sum_{I\subset J}(-1)^{\#(J\setminus I)}\Pi_{I\cap K}$$
$$= \sum_{\tilde I\subset J\cap K}\;\sum_{L\subset J\setminus K}(-1)^{\#(J\setminus(\tilde I\cup L))}\Pi_{\tilde I}$$
$$= \sum_{\tilde I\subset J\cap K}(-1)^{\#(J\setminus\tilde I)}\Big(\sum_{L\subset J\setminus K}(-1)^{-\#L}\Big)\Pi_{\tilde I}$$
$$= \sum_{\tilde I\subset J\cap K}(-1)^{\#(J\setminus\tilde I)}1_{[J\subset K]}\Pi_{\tilde I} \;=\; 1_{[J\subset K]}\sum_{\tilde I\subset J}(-1)^{\#(J\setminus\tilde I)}\Pi_{\tilde I} \;=\; 1_{[J\subset K]}\Pi_J^o,$$

where the second to last step follows from $(-1)^{-\#L}=(-1)^{\#L}$ and formula (A.3). This proves the first identities of part (b), and the second one follows from

$$\Pi_J^o\Pi_K^o \;=\; \sum_{I\subset K}(-1)^{\#(K\setminus I)}\Pi_J^o\Pi_I$$
$$= \sum_{I\subset K}(-1)^{\#(K\setminus I)}1_{[J\subset I]}\Pi_J^o$$
$$= 1_{[J\subset K]}\Big(\sum_{I\subset K:J\subset I}(-1)^{\#(K\setminus I)}\Big)\Pi_J^o$$
$$= 1_{[J\subset K]}\Big(\sum_{L\subset K\setminus J}(-1)^{\#L}\Big)\Pi_J^o \;=\; 1_{[J\subset K]}1_{[K\subset J]}\Pi_J^o \;=\; 1_{[J=K]}\Pi_J^o.$$

It remains to prove part (c). As shown in Exercise A.10, a linear operator $\Pi:\mathbb{H}\to\mathbb{H}$ describes an orthogonal projection if and only if it satisfies $\Pi^2=\Pi$ and is self-adjoint, that means $\langle\Pi Y,Z\rangle=\langle Y,\Pi Z\rangle$ for all $Y,Z\in\mathbb{H}$. By definition, all operators Π_J, $J\subset\{1,\dots,n\}$, have these properties, so Π_K^o, being a linear combination of self-adjoint operators, is self-adjoint, too. Moreover, it follows from part (b) that $\Pi_K^o\Pi_K^o=\Pi_K^o$, whence Π_K^o describes the orthogonal projection of $\mathbb{H}$ onto some linear subspace $\mathbb{H}_K^o$. The subspaces $\mathbb{H}_K^o$, $K\subset\{1,\dots,n\}$, are pairwise orthogonal, because for different index sets J,K and $Y\in\mathbb{H}_J^o$, $Z\in\mathbb{H}_K^o$,

$$\langle Y,Z\rangle \;=\; \langle\Pi_J^oY,\Pi_K^oZ\rangle \;=\; \langle Y,\Pi_J^o\Pi_K^oZ\rangle \;=\; \langle Y,0\rangle \;=\; 0.$$

Finally, by part (a), for any $K\subset\{1,\dots,n\}$,

$$\mathbb{H}_K \;=\; \sum_{J\subset K}\mathbb{H}_J^o \;=\; \mathbb{H}_K^o+\sum_{J\subsetneq K}\mathbb{H}_J^o,$$

so $\mathbb{H}_K^o$ equals

$$\mathbb{H}_K\cap\Big(\sum_{J\subsetneq K}\mathbb{H}_J^o\Big)^\perp.$$

But $\mathbb{H}_I\subset\mathbb{H}_J$ for $I\subset J$, so $\mathbb{H}_J^o\subset\mathbb{H}_J$, whence

$$\sum_{J\subsetneq K}\mathbb{H}_J^o \;\subset\; \sum_{J\subsetneq K}\mathbb{H}_J.$$

On the other hand, for any fixed $\tilde J\subsetneq K$,

$$\sum_{J\subsetneq K}\mathbb{H}_J^o \;\supset\; \sum_{J\subset\tilde J}\mathbb{H}_J^o \;=\; \mathbb{H}_{\tilde J},$$

so

$$\sum_{J\subsetneq K} \mathbb{H}_J^o \;\supset\; \sum_{J\subsetneq K} \mathbb{H}_J.$$

Consequently,

$$\sum_{J\subsetneq K} \mathbb{H}_J^o \;=\; \sum_{J\subsetneq K} \mathbb{H}_J,$$

and this leads to the asserted representation of $\mathbb{H}_K^o$. □

Exercise A.10 (Projections and orthogonal projections). Let $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ be a real Hilbert space, and let $\Pi : \mathbb{H} \to \mathbb{H}$ be a linear mapping which is idempotent, that means, $\Pi^2 = \Pi$.

(a) Show that there exist linear subspaces $\mathbb{H}_1, \mathbb{H}_2$ of $\mathbb{H}$ such that $\mathbb{H}_1 \cap \mathbb{H}_2 = \{0\}$, $\mathbb{H}_1 + \mathbb{H}_2 = \mathbb{H}$ and

$$\Pi x \;=\; \begin{cases} x & \text{if } x \in \mathbb{H}_1, \\ 0 & \text{if } x \in \mathbb{H}_2. \end{cases}$$

Hint: Write $x \in \mathbb{H}$ as $x = x_1 + x_2$ with $x_1 = \Pi x$ and $x_2 = x - \Pi x$.

(b) Show that $\mathbb{H}_1 \perp \mathbb{H}_2$ if and only if Π is self-adjoint, that means, $\langle \Pi x, y\rangle = \langle x, \Pi y\rangle$ for all $x, y \in \mathbb{H}$. In this case, Π is the orthogonal projection onto $\mathbb{H}_1$

Exercise A.11. Let $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ be a real Hilbert space, and let Π_1, Π_2 be orthogonal projections onto subspaces $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively. Further let Π_0 be the orthogonal projection onto $\mathbb{H}_0 := \mathbb{H}_1 \cap \mathbb{H}_2$. Show that the following three statements are equivalent:

(i) $\mathbb{H}_1 \cap \mathbb{H}_0^\perp \perp \mathbb{H}_2 \cap \mathbb{H}_0^\perp$.

(ii) $\Pi_1\Pi_2 = \Pi_0$.

(iii) $\Pi_1\Pi_2 = \Pi_2\Pi_1$.

A.5 Weak Law of Large Numbers and Central Limit Theorem

In connection with asymptotic considerations, the subsequent versions of the Weak Law of Large Numbers and Lindeberg's Central Limit Theorem are rather useful. Throughout this section asymptotic statements refer to $n \to \infty$, unless specified differently.

Theorem A.12 (WLLN). *For any integer $n \geq 1$ let $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ be independent random variables such that*

$$\sum_{i=1}^n \mathbb{E}\,|Y_{ni}| \;=\; O(1),$$
$$\sum_{i=1}^n \mathbb{E}\big(1_{[|Y_{ni}|>\epsilon]}|Y_{ni}|\big) \;\to\; 0 \quad \textit{for any fixed } \epsilon > 0.$$

Then with $\mu_{ni} := \mathbb{E}(Y_{ni})$,

$$\mathbb{E}\Big|\sum_{i=1}^n (Y_{ni} - \mu_{ni})\Big| \;\to\; 0 \quad \textit{and} \quad \mathbb{E}\max_{1\leq i\leq n} |Y_{ni}| \;\to\; 0.$$

Theorem A.13 (CLT). *For any integer $n \geq 1$ let $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ be independent random variables such that for some real numbers μ and $\sigma > 0$,*

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}(Y_{ni}) \;\to\; \mu \quad \textit{and} \quad \sum_{i=1}^{n} |\,\mathbb{E}(Y_{ni})| \;=\; O(1),\\
&\sum_{i=1}^{n} \mathbb{E}(Y_{ni}^2) \;\to\; \sigma^2,\\
&\sum_{i=1}^{n} \mathbb{E}\big(1_{[Y_{ni}^2 > \epsilon]} Y_{ni}^2\big) \;\to\; 0 \quad \textit{for any fixed } \epsilon > 0.
\end{aligned}$$

Then

$$\sum_{i=1}^{n} Y_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(\mu, \sigma^2)$$

and

$$\mathbb{E}\Big|\sum_{i=1}^{n} Y_{ni}^2 - \sigma^2\Big| \;\to\; 0, \quad \mathbb{E}\Big(\max_{1\leq i\leq n} Y_{ni}^2\Big) \;\to\; 0, \quad \sum_{i=1}^{n} \mathbb{E}(Y_{ni})^2 \;\to\; 0.$$

Corollary A.14. *For any integer $n \geq 1$ let $X_{n1}, X_{n2}, \ldots, X_{nn}$ be independent and identically distributed random variables such that for some real numbers μ and $\sigma > 0$,*

$$\begin{aligned}
&\sqrt{n}\,\mathbb{E}(X_{n1}) \;\to\; \mu,\\
&\mathbb{E}(X_{n1}^2) \;\to\; \sigma^2,\\
&\mathbb{E}\big(1_{[X_{n1}^2 > \epsilon n]} X_{n1}^2\big) \;\to\; 0 \quad \textit{for any fixed } \epsilon > 0.
\end{aligned}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(\mu, \sigma^2)$$

and

$$\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n} X_{ni}^2 - \sigma^2\Big| \;\to\; 0, \quad \mathbb{E}\Big(\max_{1\leq i\leq n} \frac{X_{ni}^2}{n}\Big) \;\to\; 0.$$

Proof of Theorem A.12. Let $M := \limsup_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\,|Y_{ni}|$. For arbitrary fixed $\epsilon > 0$ set

$$Y_{ni1} \;:=\; 1_{[|Y_{ni}| \leq \epsilon]} Y_{ni}, \quad Y_{ni2} \;:=\; 1_{[|Y_{ni}| > \epsilon]} Y_{ni}$$

and $\mu_{nik} := \mathbb{E}(Y_{nik})$. Then $\mathbb{E}\big|\sum_{i=1}^{n}(Y_{ni} - \mu_{ni})\big|$ is bounded from above by

$$\begin{aligned}
\mathbb{E}\Big|\sum_{i=1}^{n}(Y_{ni1} - \mu_{ni1})\Big| + \sum_{i=1}^{n}\big(\mathbb{E}\,|Y_{ni2}| + |\mu_{ni2}|\big) \;&\leq\; \sqrt{\mathrm{Var}\Big(\sum_{i=1}^{n} Y_{ni1}\Big)} + 2\sum_{i=1}^{n} \mathbb{E}\,|Y_{ni2}|\\
&\leq\; \sqrt{\sum_{i=1}^{n} \mathbb{E}(Y_{ni1}^2)} + o(1)\\
&\leq\; \sqrt{\epsilon \sum_{i=1}^{n} \mathbb{E}\,|Y_{ni}|} + o(1)\\
&\leq\; \sqrt{\epsilon M + o(1)} + o(1)\\
&\to\; \sqrt{\epsilon M}.
\end{aligned}$$

Furthermore,

$$\mathbb{E}\Big(\max_{1\le i\le n}|Y_{ni}|\Big) \;\le\; \epsilon + \sum_{i=1}^{n}\mathbb{E}\,|Y_{ni2}| \;\to\; \epsilon.$$

Since $\epsilon > 0$ may be arbitrarily small, these calculations yield the assertions. □

Remarks on the proof of Theorem A.13. One can deduce from Theorem A.12, applied to Y_{ni}^2 in place of Y_{ni}, that

$$\mathbb{E}\Big|\sum_{i=1}^{n}Y_{ni}^2-\sigma^2\Big| \;\to\; 0 \quad\text{and}\quad \mathbb{E}\Big(\max_{1\le i\le n}Y_{ni}^2\Big) \;\to\; 0.$$

In particular, with $\mu_{ni} := \mathbb{E}(Y_{ni})$,

$$\max_{1\le i\le n}\mu_{ni}^2 \;\le\; \max_{1\le i\le n}\mathbb{E}(Y_{ni}^2) \;\le\; \mathbb{E}\Big(\max_{1\le i\le n}Y_{ni}^2\Big) \;\to\; 0,$$

whence

$$\sum_{i=1}^{n}\mu_{ni}^2 \;\le\; \max_{1\le j\le n}|\mu_{nj}|\sum_{i=1}^{n}|\mu_{ni}| \;\to\; 0.$$

But with $\sigma_{ni}^2 := \mathrm{Var}(Y_{ni})$ this implies that

$$\sum_{i=1}^{n}\sigma_{ni}^2 \;=\; \sum_{i=1}^{n}\mathbb{E}(Y_{ni}^2)-\sum_{i=1}^{n}\mu_{ni}^2 \;\to\; \sigma^2.$$

Moreover, for any fixed $\epsilon > 0$, the inequality $\max_{1\le i\le n}|\mu_{ni}| \le \epsilon/2$ is satisfied for sufficiently large n, and in that case, $|Y_{ni}-\mu_{ni}| > \epsilon$ implies that $|Y_{ni}| \ge \epsilon/2$ and $|Y_{ni}-\mu_{ni}| \le 2|Y_{ni}|$. Hence for sufficiently large n,

$$\sum_{i=1}^{n}\mathbb{E}\big(1_{[|Y_{ni}-\mu_{ni}|>\epsilon]}(Y_{ni}-\mu_{ni})^2\big) \;\le\; 4\sum_{i=1}^{n}\mathbb{E}\big(1_{[|Y_{ni}|>\epsilon/2]}Y_{ni}^2\big) \;\to\; 0.$$

Consequently, the centered random variables $Z_{ni} := Y_{ni}-\mu_{ni}$ satisfy the assumptions of the more traditional CLT:

$$\begin{aligned}
&\mathbb{E}(Z_{ni}) \;=\; 0 \quad \text{for all } n\ge 1 \text{ and } 1\le i\le n,\\
&\sum_{i=1}^{n}\mathbb{E}(Z_{ni}^2) \;\to\; \sigma^2,\\
&\sum_{i=1}^{n}\mathbb{E}\big(1_{[|Z_{ni}|>\epsilon]}Z_{ni}^2\big) \;\to\; 0 \quad \text{for any fixed } \epsilon>0.
\end{aligned}$$

These conditions imply that

$$\sum_{i=1}^{n}Z_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(0,\sigma^2),$$

and thus

$$\sum_{i=1}^{n}Y_{ni} \;=\; \mu+o(1)+\sum_{i=1}^{n}Z_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(\mu,\sigma^2).$$

□

A.6 Conditional distributions of Gaussian Random Vectors

Assuming that the reader is familiar with basic properties of multivariate random vectors and Gaussian distributions, let us recall some particular facts: Suppose that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are random vectors on the same probability space such that $\mathbb{E}(\|X\|^2) + \mathbb{E}(\|Y\|^2) < \infty$, where $\Sigma_{XX} := \mathrm{Var}(X)$ is positive definite. With $\mu_X := \mathbb{E}(X)$, $\mu_Y := \mathbb{E}(Y)$, $\Sigma_{XY} := \mathrm{Cov}(X, Y)$, $\Sigma_{YX} := \Sigma_{XY}^\top = \mathrm{Cov}(Y, X)$ and $\Sigma_{YY} = \mathrm{Var}(Y)$,

$$\check{Y}(X) \;:=\; \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) \tag{A.4}$$

is the unique optimal linear predictor of Y, given X. That is, for arbitrary fixed $a \in \mathbb{R}^q$ and $B^{q\times p}$,

$$\mathbb{E}(\|Y - a - BX\|^2) \;\geq\; \mathbb{E}(\|Y - \check{Y}(X)\|^2)$$

with equality if and only if $a = \mu_Y - B\mu_X$ and $B = \Sigma_{YX}\Sigma_{XX}^{-1}$. Moreover, the prediction error $Y - \check{Y}$ satisfies $\mathbb{E}(Y - \check{Y}(X)) = 0$ and

$$\mathrm{Cov}(Y - \check{Y}(X), X) \;=\; 0, \quad \mathrm{Var}(Y - \check{Y}(X)) \;=\; \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{A.5}$$

In particular, if X and Y have a joint Gaussian distribution, then X and $Y - \check{Y}(X)$ are stochastically independent, and the representation

$$Y \;=\; \check{Y}(X) + (Y - \check{Y}(X))$$

shows that

$$\begin{aligned} \mathcal{L}(Y \,|\, X) \;&=\; \mathrm{N}_q\big(\check{Y}(X), \mathrm{Var}(Y - \check{Y}(X))\big) \\ &=\; \mathrm{N}_q\big(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\big). \end{aligned} \tag{A.6}$$

A.7 Gamma and Noncentral Chi-Squared Distributions

It is a consequence of Theorem A.5 that the distribution of a random variable $Y \geq 0$ is uniquely determined by its moment-generating function $t \mapsto \mathbb{E}\exp(tY)$, $t \leq 0$. With this one can derive a couple of well-known results about gamma and chi-squared distributions.

Fact 1. Let $Y \sim \mathrm{Gamma}(a, b)$, the gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Then

$$\mathbb{E}\exp(tY) \;=\; (1 - bt)^{-a} \quad \text{for } t < 1/b.$$

Indeed, Y has the same distribution as bY_o, where $Y_o \sim \mathrm{Gamma}(a, 1)$, so it suffices to verify the formula for $b = 1$. Here,

$$\begin{aligned} \mathbb{E}\exp(tY) \;&=\; \frac{1}{\Gamma(a)}\int_0^\infty \exp(ty) y^{a-1} e^{-y}\,\mathrm{d}y \\ &=\; \frac{1}{\Gamma(a)}\int_0^\infty y^{a-1} e^{-(1-t)y}\,\mathrm{d}y \\ &=\; (1-t)^{-a}\frac{1}{\Gamma(a)}\int_0^\infty z^{a-1} e^{-z}\,\mathrm{d}z \qquad (z = (1-t)y) \\ &=\; (1-t)^{-a}. \end{aligned}$$

Fact 2. Let $Z \sim \mathrm{N}(0,1)$. Then,

$$Z^2 \;\sim\; \chi^2_1 \;=\; \mathrm{Gamma}(1/2, 2).$$

Indeed, for $t < 1/2$,

$$\begin{aligned}
\mathbb{E}\exp(tZ^2) \;&=\; (2\pi)^{-1/2} \int_{\mathbb{R}} \exp(tz^2) e^{-z^2/2}\,\mathrm{d}z \\
&=\; (2\pi)^{-1/2} \int_{\mathbb{R}} e^{-(1-2t)z^2/2}\,\mathrm{d}z \\
&=\; (1-2t)^{-1/2}(2\pi)^{-1/2} \int_{\mathbb{R}} e^{-x^2/2}\,\mathrm{d}x \qquad (x = (1-2t)^{1/2}x) \\
&=\; (1-2t)^{-1/2}.
\end{aligned}$$

Fact 3 (Noncentral chi-squared distributions with one degree of freedom). For $\mu \in \mathbb{R}$ and $X \sim \mathrm{N}(\mu, 1)$, the random variable X^2 has the same distribution as

$$\sum_{k=1}^{1+2N} Z_k^2$$

with independent random variables $Z_1, Z_2, Z_3, \ldots \sim \mathrm{N}(0,1)$ and $N \sim \mathrm{Poiss}(\mu^2/2)$. This distribution is the *noncentral chi-squared distribution with one degree of freedom and noncentrality parameter* μ^2.

Indeed, writing $X = \mu + Z$ with $Z \sim \mathrm{N}(0,1)$, we obtain the formula

$$\begin{aligned}
\mathbb{E}\exp(tX^2) \;&=\; (2\pi)^{-1/2} \int_{\mathbb{R}} \exp(t(\mu+z)^2) e^{-z^2/2}\,\mathrm{d}z \\
&=\; (2\pi)^{-1/2} \int_{\mathbb{R}} \exp\bigl(-(1-2t)z^2/2 + 2t\mu z + t\mu^2\bigr)\,\mathrm{d}z \\
&=\; \exp\Bigl(t\mu^2 + \frac{2t^2\mu^2}{1-2t}\Bigr)(2\pi)^{-1/2} \int_{\mathbb{R}} \exp\Bigl(-(1-2t)\Bigl(z - \frac{2t\mu}{1-2t}\Bigr)^2/2\Bigr)\,\mathrm{d}z \\
&=\; (1-2t)^{-1/2} \exp\Bigl(\frac{t\mu^2}{1-2t}\Bigr)(2\pi)^{-1/2} \int_{\mathbb{R}} e^{-x^2/2}\,\mathrm{d}x \\
&\qquad\qquad \Bigl(x = (1-2t)^{-1/2}\Bigl(z - \frac{2t\mu}{1-2}\Bigr)\Bigr) \\
&=\; (1-2t)^{-1/2} \exp\Bigl(\frac{t\mu^2}{1-2t}\Bigr)
\end{aligned}$$

for $t < 1/2$. But the right-hand side is equal to

$$\begin{aligned}
(1-2t)^{-1/2} e^{-\mu^2/2} \exp\Bigl(\frac{\mu^2/2}{1-2t}\Bigr) \;&=\; e^{-\mu^2/2} \sum_{k=0}^{\infty} \frac{(\mu^2/2)^k}{k!}\,(1-2t)^{-1/2-k} \\
&=\; \sum_{k=0}^{\infty} \mathbb{P}(N=k)(1-2t)^{-(1+2k)/2} \\
&=\; \sum_{k=0}^{\infty} \mathbb{P}(N=k)\,\mathbb{E}\exp\Bigl(t\sum_{\ell=1}^{1+2k} Z_\ell^2\Bigr) \\
&=\; \mathbb{E}\exp\Bigl(t\sum_{\ell=1}^{1+2N} Z_\ell^2\Bigr).
\end{aligned}$$

Fact 4 (Noncentral chi-squared distributions with $q > 1$ degrees of freedom). For $\mu \in \mathbb{R}^q$ and $X \sim \mathrm{N}_q(\mu, 1)$, the distribution of the random variable $\|X\|^2$ is called the *noncentral chi-squared distribution with q degrees of freedom and noncentrality parameter $\|\mu\|^2$*. To see that it really depends only on $\|\mu\|$, let $B \in \mathbb{R}^{q\times q}$ be an orthogonal matrix such that $B\mu = (\|\mu\|, 0, \ldots, 0)^\top$. Writing $X = \mu + Z$ with $Z \sim \mathrm{N}_q(0, I_q)$, we see that $\|X\|^2 = \|BX\|^2 = \|B\mu + BZ\|^2$ has the same distribution as

$$\|B\mu + Z\|^2 \;=\; (Z_1 + \|\mu\|)^2 + \sum_{i=2}^{q} Z_i^2,$$

because $BZ \sim \mathrm{N}_q(0, I_q)$ too. Now one can deduce from Fact 3, applied to $Z_1 + \|\mu\| \sim \mathrm{N}(\|\mu\|, 1)$, that this distribution coincides with the distribution of

$$\sum_{k=1}^{q+2N} Z_k^2$$

with independent random variables $Z_1, Z_2, Z_3, \ldots \sim \mathrm{N}(0, 1)$ and $N \sim \mathrm{Poiss}(\|\mu\|^2/2)$.