Advanced Measure Theory and Mathematical Statistics

Lutz Dümbgen

University of Bern
Spring Semester 2022

October 20, 2022

Bibliography

[1] E.M. ALFSEN (1963). A simplified constructive proof of the existence and uniqueness of Haar measure. *Mathematica Scandinavica* ***12***, pp. 106-116.

[2] 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.

[3] H. BAUER (2001). *Measure and integration theory*. Walter de Gruyter & Co., Berlin.

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

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

[6] A. HAAR (1933). Der Massbegriff in der Theorie der Kontinuierlichen Gruppen. *Annals of Mathematics* ***34****(1)*, pp. 147-169.

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

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

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

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

Acknowledgements. The second 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, Christof Strähl and Nikita van Gils on the current notes are gratefully acknowledged.

Contents

I Advanced Measure Theory **11**

1 Abstract Integrals **13**

1.1 Lattices and Stone Lattices 13

1.2 Abstract and Usual Integrals 15

1.3 Representations of Dual Spaces 20

2 Signed Measures **25**

2.1 The Hahn–Jordan Decomposition 25

2.2 Radon–Nikodym Derivatives 28

2.2.1 Finite measures 30

2.2.2 Absolute continuity and σ-finite measures 32

2.3 Another Riesz Representation 34

3 Conditional Expectations **37**

3.1 Conditional expectations with respect to a sub-σ-field 37

3.2 Conditional expectations as orthogonal projections 40

3.3 Conditional expectations given another random variable 41

4 Stochastic Kernels **47**

4.1 Stochastic Kernels and Fubini's Theorem 47

4.2 Decomposing Measures on Product Spaces 51

4.3 Conditional Expectations and Distributions 55

5 Haar Measure **59**

5.1 Locally Compact Topological Groups 59

5.2 Left- and Right-Invariant Measures 62

5.3 Some Explicit Constructions 69

II Mathematical Statistics **73**

6 Measurement Series and Estimators of Location **75**

6.1 Statistical Experiments and Point Estimators 75

6.2 Estimators of Location 76

6.3 Constructing an Optimal Equivariant Estimator 77

6.4 Beyond Equivariance: Admissibility 85

6.5 Location Functionals and Gross Error Models 90

7 Statistical Tests **95**

7.1 The Neyman–Pearson Lemma 97

7.2 Monotone Density Ratios 100

7.3 Stochastic Order, P-Values, Confidence Bounds 105

7.4 The Generalized Neyman–Pearson Lemma 109

7.5 Tests of Two-Sided Hypotheses 111

7.5.1 One-parameter exponential families (with natural parametrization) 111

7.5.2 Two-sided hypotheses, version 1 113

7.5.3 Two-sided hypotheses, version 2 117

7.5.4 Summary and some first applications 120

7.6 Tests and Confidence regions 124

8 Decision Problems and Procedures, Sufficiency and Completeness **129**

8.1 Decision Problems and Procedures 129

8.2 Some Optimality Concepts and Results 132

8.3 Informativity and Sufficiency 136

8.3.1 Informativity 136

8.3.2 Sufficiency 138

8.4 Complete Statistical Experiments 144

8.5 U-Statistics 148

9 Exponential Families **159**

9.1 Definitions and Basic Properties 159

9.2 Nuisance Parameters 161

10 Some Asymptotics **175**

10.1 Testing, Total Variation and Hellinger Distances 175

10.2 Asymptotics for Repeated Binary Experiments 180

10.3 Fisher Information 188

A Auxiliary Results **195**

A.1 Some Basics from Measure Theory 195

A.2 Two Compactness Properties of Statistical Tests 199

A.3 Uniqueness of Moment-Generating Functions 201

A.4 Hoeffding's Decomposition 202

A.5 Weak Law of Large Numbers and Central Limit Theorem 206

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. Hence, the first part of this course is devoted to these aspects of measure theory.

Part I

Advanced Measure Theory

Chapter 1

Abstract Integrals

We tacitly assume that the reader is familiar with standard measure and integration theory. Some basic notions and results are listed in Section A.1.

In standard measure theory, one starts from a measurable space $(\Omega, \mathcal{A})$, consisting of a set Ω and a σ-field $\mathcal{A}$ over Ω. Then one constructs or defines a measure μ on $\mathcal{A}$, and this leads to the integral $\int f\,d\mu$ of measurable functions $f : \Omega \to [0, \infty]$ or $f : \Omega \to \mathbb{R}$. If we restrict our attention to the set $\mathcal{L}^1(\mu)$ of real-valued, measurable functions f on Ω such that $\int |f|\,d\mu < \infty$, then the integral $\int f\,d\mu$ has the following essential properties:

Linearity: The set $\mathcal{L}^1(\mu)$ is a real vector space[1], and $f \mapsto \int f\,d\mu$ is linear in $f \in \mathcal{L}^1(\mu)$.

Positivity: If $f \in \mathcal{L}^1(\mu)$ is nonnegative, then $\int f\,d\mu \geq 0$.

Montone convergence: If $(f_n)_n$ is a sequence of nonnegative functions in $\mathcal{L}^1(\mu)$ which is pointwise increasing with limit $f \in \mathcal{L}^1(\mu)$, then $\int f_n\,d\mu \to \int f\,d\mu$ as $n \to \infty$.

In the present section we follow the reverse route. Throughout let $\mathcal{F}$ be a real vector space of real-valued functions on a set Ω. Suppose that $J : \mathcal{F} \to \mathbb{R}$ is a linear and positive functional satisfying monotone convergence. Does there exist a σ-field $\mathcal{A}$ over Ω and a measure μ on $\mathcal{A}$ such that $\mathcal{F} \subset \mathcal{L}^1(\mu)$ and $J(f) = \int f\,d\mu$ for all $f \in \mathcal{F}$?

1.1 Lattices and Stone Lattices

For two functions $f, g\colon \Omega \to [-\infty, \infty]$, the inequalities $f \leq g$ or $f \geq g$ are always meant pointwise. Their pointwise maximum and minimum are denoted with $f \vee g$ and $f \wedge g$, respectively, that is,

$$f \vee g(\omega) \;:=\; \max\{f(\omega), g(\omega)\} \quad \text{and} \quad f \wedge g(\omega) \;:=\; \min\{f(\omega), g(\omega)\}.$$

Similarly, the pointwise maximum and minimim of f and a real constant c is denoted with $f \vee c$ and $f \wedge c$, respectively. Specifically, we write

$$f^+ \;:=\; f \vee 0 \qquad \text{and} \quad f^- \;:=\; (-f) \vee 0 = -(f \wedge 0),$$

so $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

[1] with the pointwise addition of functions and the pointwise multiplication of functions with scalars

Definition 1.1 (Lattice and Stone lattice). The linear function space $\mathcal{F}$ is called a *lattice* if $|f| \in \mathcal{F}$ for any $f \in \mathcal{F}$.

A lattice $\mathcal{F}$ is called a *Stone lattice* if $f \wedge 1 \in \mathcal{F}$ for arbitrary $f \in \mathcal{F}$.

Example 1.2. Let (Ω, d) be a metric space. Then the following sets of functions $f : \Omega \to \mathbb{R}$ are Stone lattices:

$$\begin{aligned} \mathcal{C}(\Omega) &= \{f\colon \text{continuous}\}, \\ \mathcal{C}_b(\Omega) &= \{f\colon \text{continuous and bounded}\}, \\ \mathcal{C}_{\mathrm{Lip}}(\Omega) &= \{f\colon \text{Lipschitz-continuous}\}, \\ \mathcal{C}_{\mathrm{Lip},b}(\Omega) &= \{f\colon \text{Lipschitz-continuous and bounded}\}. \end{aligned}$$

Exercise 1.3 (Lattices). Show that the following three properties of a linear function space $\mathcal{F}$ are equivalent:

(i) $|f| \in \mathcal{F}$ for arbitrary $f \in \mathcal{F}$.

(ii) $f^+ \in \mathcal{F}$ for arbitrary $f \in \mathcal{F}$.

(iii) $f \vee g, f \wedge g \in \mathcal{F}$ for arbitrary $f, g \in \mathcal{F}$.

From now on, let $\mathcal{F}$ be a Stone lattice. The set of nonnegative functions in $\mathcal{F}$ is denoted with $\mathcal{F}^+$, that is,

$$\mathcal{F}^+ = \{f \in \mathcal{F}\colon f \geq 0\} = \{f^+\colon f \in \mathcal{F}\}.$$

Exercise 1.4 (Stone lattices). Show that for arbitrary $f \in \mathcal{F}$ and $c \geq 0$,

$$f \wedge c,\ f \vee (-c),\ (f-c)^+,\ (f+c)^- \in \mathcal{F}.$$

Find an example of a Stone lattice $\mathcal{F}$ such that $f \vee 1 \notin \mathcal{F}$ for any $f \in \mathcal{F}$.

The following families of functions and sets play an important role in connection with integrals.

Definition 1.5 (Extension of $\mathcal{F}^+$, $\mathcal{F}$-open sets). The *extension of* $\mathcal{F}^+$ is the family $\mathcal{F}^*$ of all functions $g : \Omega \to [0, \infty]$ such that

$$g = \sup_{n \geq 1} f_n$$

for some sequence $(f_n)_{n \geq 1}$ in $\mathcal{F}^+$.

A set $U \subset \Omega$ is called $\mathcal{F}$-open if

$$U = \{g > 0\}$$

for some $g \in \mathcal{F}^*$. The family of all $\mathcal{F}$-open sets is denoted with $\mathcal{U}(\mathcal{F})$.

These families $\mathcal{F}^*$ and $\mathcal{U}(\mathcal{F})$ have some important properties summarized in the next three lemmas.

Lemma 1.6 (Extension of $\mathcal{F}^+$). *For a function $g : \Omega \to [0, \infty]$, the following three properties are equivalent:*

(i) $g = \sup_{n \geq 1} f_n$ for some sequence $(f_n)_{n \geq 1}$ in $\mathcal{F}^+$.

(ii) $g = \lim_{n\geq 1} f_n$ *for some sequence* $(f_n)_{n\geq 1}$ *in* $\mathcal{F}^+$ *which is pointwise increasing.*

(iii) $g = \sum_{n\geq 1} f_n$ *for some sequence* $(f_n)_{n\geq 1}$ *in* $\mathcal{F}^+$.

Lemma 1.7 (Properties of $\mathcal{F}^*$). *For arbitrary scalars* $c \geq 0$ *and functions* g, h, $g_1, g_2, g_3, \ldots$ *in* $\mathcal{F}^*$, *the following functions belong to* $\mathcal{F}^*$ *too:*

$$cg,\ g \wedge h,\ g \wedge c,\ \sup_{n\geq 1} g_n,\ \sum_{n\geq 1} g_n.$$

Lemma 1.8 ($\mathcal{F}$-open sets). *For a set* $U \subset \Omega$, *the following three properties are equivalent:*

(i) $U = \{g > 0\}$ *for some function* $g \in \mathcal{F}^*$.

(ii) $U = \{g > c\}$ *for some function* $g \in \mathcal{F}^*$ *and* $c \geq 0$.

(iii) $1_U \in \mathcal{F}^*$.

Remark 1.9. Lemmas 1.7 and 1.8 imply that the family $\mathcal{U}(\mathcal{F})$ is closed under intersections and countable unions. Moreover, the σ-fields $\sigma(\mathcal{F})$ and $\sigma(\mathcal{U}(\mathcal{F}))$ coincide, where $\sigma(\mathcal{F})$ is the smallest σ-field over Ω such that all functions in $\mathcal{F}$ are measurable, and $\sigma(\mathcal{U}(\mathcal{F}))$ is the smallest σ-field over Ω containing all sets in $\mathcal{U}(\mathcal{F})$.

To see the latter claim, note that any function in $\mathcal{F}^*$ is a pointwise limit of a sequence in $\mathcal{F}^+ \subset \mathcal{F}$, whence $\mathcal{U}(\mathcal{F}) \subset \sigma(\mathcal{F})$, and this implies that $\sigma(\mathcal{U}(\mathcal{F})) \subset \sigma(\mathcal{F})$. On the other hand, the σ-field $\sigma(\mathcal{F})$ is generated by all sets $\{f > c\}$ and $\{f < -c\}$, $f \in \mathcal{F}$ and $c > 0$. But $\{f > c\} = \{g > 0\}$ with $g := f - f \wedge c \in \mathcal{F}^+$, so $\{f > c\} \in \mathcal{U}(\mathcal{F})$, and $\{f < -c\} = \{-f > c\}$ belongs to $\mathcal{U}(\mathcal{F})$ too. This shows that $\sigma(\mathcal{F}) \in \sigma(\mathcal{U}(\mathcal{F}))$.

The proofs of Lemmas 1.6 and 1.7 are left to the reader as an exercise.

Proof of Lemma 1.8. It is clear that property (i) of U implies property (ii). Suppose that U has property (ii), that is, $U = \{g > c\}$ for some function $g \in \mathcal{F}^*$ and some scalar $c \geq 0$. Let $(f_n)_{n\geq 1}$ be a pointwise increasing sequence in $\mathcal{F}^+$ with limit g. Then

$$\tilde{f}_n \ := \ (n(f_n - c)^+) \wedge 1$$

defines a pointswise increasing sequence $(\tilde{f}_n)_{n\geq 1}$ with limit $1_U = 1_{\{g>c\}}$. Thus U has property (iii). Finally, if U has property (iii), then it has property (i) with $g := 1_U \in \mathcal{F}^*$. □

1.2 Abstract and Usual Integrals

As before, let $\mathcal{F}$ be a Stone lattice of functions $f : \Omega \to \mathbb{R}$ with its subcone $\mathcal{F}^+$ of nonnegative functions $f \in \mathcal{F}$.

Definition 1.10 (Abstract integral). A functional $J : \mathcal{F} \to \mathbb{R}$ is called an *abstact integral (on the Stone lattice* $\mathcal{F}$*)* if it has the following properties:

Linearity: For $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$,

$$J(\lambda f) \ = \ \lambda J(f) \quad \text{and} \quad J(f + g) \ = \ J(f) + J(g).$$

Positivity: $J(f) \geq 0$ for all $f \in \mathcal{F}^+$.

Monotone convergence: Let $(f_n)_n$ be a pointwise increasing sequence of functions $f_n \in \mathcal{F}^+$ with limit $f \in \mathcal{F}$. Then,

$$\lim_{n\to\infty} J(f_n) \;=\; J(f).$$

Linearity and positivity of an abstract integral J on $\mathcal{F}$ imply that

$$J(f) \;\leq\; J(g) \quad \text{whenever } f, g \in \mathcal{F} \text{ with } f \leq g.$$

This follows from $g = f + h$ with $h = g - f \in \mathcal{F}^+$, so $J(g) = J(f) + J(h) \geq J(f)$.

Exercise 1.11. Let $\mathcal{F}$ be the set of all functions (sequences) $f : \mathbb{N} \to \mathbb{R}$ such that $\lim_{\omega\to\infty} f(\omega)$ exists in $\mathbb{R}$. Verify that $\mathcal{F}$ is a Stone lattice. Does $J(f) := \lim_{\omega\to\infty} f(\omega)$ define an abstract integral on $\mathcal{F}$?

From now on, let $J\colon \mathcal{F} \to \mathbb{R}$ be an abstract integral. In view of our goal to achieve, we extend J to a functional on the extension $\mathcal{F}^*$ of $\mathcal{F}^+$. For $g \in \mathcal{F}^*$, let

$$J(g) \;:=\; \sup_{f\in\mathcal{F}^+\colon f\leq g} J(f).$$

Note that if $g \in \mathcal{F}^+$, then $J(f) \leq J(g)$ for all $f \in \mathcal{F}^+$ such that $f \leq g$, so the new definition of $J(g)$ yields the former value of $J(g)$. It is also obvious that for $g, h \in \mathcal{F}^*$,

$$J(g) \;\leq\; J(h) \;\text{ if } g \leq h,$$

because $g \leq h$ implies that $\{f \in \mathcal{F}^+\colon f \leq g\} \subset \{f \in \mathcal{F}^+\colon f \leq h\}$. The next lemma provides an alternative representation of $J(g)$ which is often convenient.

Lemma 1.12. *If $g = \lim_{n\to\infty} f_n$ with a pointwise increasing sequence $(f_n)_n$ in $\mathcal{F}^+$, then $J(g) = \lim_{n\to\infty} J(f_n)$.*

Proof of Lemma 1.12. Since $(f_n)_n$ is pointwise increasing, the sequence $(J(f_n))_n$ is increasing, and $f_n \leq g$ for each n, whence $J(g) \geq \lim_{n\to\infty} J(f_n)$.

On the other hand, if f is any fixed function in $\mathcal{F}^+$ such that $f \leq g$, then $(f_n \wedge f)_n$ is pointwise increasing with limit f. Consequently,

$$\lim_{n\to\infty} J(f_n) \;\geq\; \lim_{n\to\infty} J(f_n \wedge f) \;=\; J(f),$$

and this shows that $J(g) \leq \lim_{n\to\infty} J(f_n)$. □

The next lemma shows that J, as a functional on $\mathcal{F}^*$, has various desirable properties, where we use the convention that $0 \cdot \infty := 0$.

Lemma 1.13. *For arbitrary $c \geq 0$ and functions $g, g_1, g_2, g_3, \ldots \in \mathcal{F}^*$,*

$$J(cg) \;=\; cJ(g) \quad \textit{and} \quad J\Big(\sum_{n\geq 1} g_n\Big) \;=\; \sum_{n\geq 1} J(g_n).$$

If $(g_n)_{n\geq 1}$ is pointwise increasing with limit g, then

$$J(g) \;=\; \lim_{n\to\infty} J(g_n).$$

Proof of Lemma 1.13. Let $(f_n)_{n\geq 1}$ be a pointwise increasing sequence in $\mathcal{F}^+$ with limit g. Then, $cf_n \in \mathcal{F}^+$ and $cf_n \uparrow cg$ as $n \to \infty$, whence

$$J(cg) \;=\; \lim_{n\to\infty} J(cf_n) \;=\; c \lim_{n\to\infty} J(f_n) \;=\; cJ(g).$$

For each $n \geq 1$, let $(f_{n,k})_{k\geq 1}$ be a pointwise increasing sequence in $\mathcal{F}^+$ with limit g_n. Then

$$f_N \;:=\; \sum_{n=1}^{N} f_{n,N}$$

defines a pointwise increasing sequence $(f_N)_{N\geq 1}$ in $\mathcal{F}^+$ with limit $\sum_{n\geq 1} g_n$, because for any fixed integer $N_o \geq 1$ and $N \geq N_o$,

$$\sum_{n\geq 1} g_n \;\geq\; f_N \;\geq\; \sum_{n=1}^{N_o} f_{n,N} \;\to\; \sum_{n=1}^{N_o} g_n \quad \text{as } N \to \infty,$$

and $\sum_{n=1}^{N_o} g_n \uparrow \sum_{n\geq 1} g_n$ as $N_o \to \infty$. Consequently, for any fixed $N_o \geq 1$,

$$J\Big(\sum_{n\geq 1} g_n\Big) \;=\; \lim_{N\to\infty} J(f_N) \;=\; \lim_{N\to\infty} \sum_{n=1}^{N} J(f_{n,N}) \;\begin{cases} \leq \displaystyle\sum_{n\geq 1} J(g_n), \\ \geq \displaystyle\sum_{n=1}^{N_o} J(g_n), \end{cases}$$

and letting $N_o \to \infty$ reveals that $J\big(\sum_{n\geq 1} g_n\big) = \sum_{n\geq 1} J(g_n)$.

Suppose that $(g_n)_{n\geq 1}$ is pointwise increasing with limit g. It is tempting to write $g = \sum_{n\geq 1} \tilde{g}_n$ with $\tilde{g}_n := g_n - g_{n-1}$, $g_0 := 0$, and refer to the previous result about countable sums. But note that in general, $\tilde{g}_n \notin \mathcal{F}^*$. However,

$$f_N \;:=\; \max_{n\leq N} f_{n,N}$$

defines a pointwise increasing sequence $(f_N)_{N\geq 1}$ in $\mathcal{F}^+$ with limit g, because for any fixed integer $N_o \geq 1$ and $N \geq N_o$,

$$g \;\geq\; g_N \;\geq\; f_N \;\geq\; \max_{n\leq N_o} f_{n,N} \;\uparrow\; g_{N_o}$$

as $N \to \infty$, and $g_{N_o} \uparrow g$ as $N_o \to \infty$. Consequently, for any fixed $N_o \geq 1$,

$$J(g) \;=\; \lim_{N\to\infty} J(f_N) \;\begin{cases} \leq \displaystyle\lim_{N\to\infty} J(g_N), \\ \geq J(g_{N_o}), \end{cases}$$

and letting $N_o \to \infty$ reveals that $J(g) = \lim_{N\to\infty} J(g_N)$. □

Now we have some essential ingredients for the proof of the following theorem.

Theorem 1.14 (Daniell–Port–Stone). *There exists a unique measure μ on the σ-field $\sigma(\mathcal{F})$ with the following properties: $\mathcal{F} \subset \mathcal{L}^1(\mu)$,*

$$J(f) \;=\; \int f \, d\mu \quad \textit{for all } f \in \mathcal{F},$$

and for any set $B \in \sigma(\mathcal{F})$,

$$\mu(B) \;=\; \inf\bigl\{\mu(U)\colon U \in \mathcal{U}(\mathcal{F}), U \supset B\bigr\}$$

with $\inf(\emptyset) := \infty$.

Proof of Theorem 1.14. The proof is divided into six steps.

Step 1. For $S \subset \Omega$ let

$$\mu(S) \;:=\; \inf\{J(g)\colon g \in \mathcal{F}^*, g \geq 1_S\}.$$

This defines an outer measure on Ω.

Proof: Since $1_\emptyset \equiv 0 \in \mathcal{F}^*$ and $J(0) = 0$, we see that $\mu(\emptyset) = 0$. To verify that μ is an outer measure, it remains to verify that for arbitrary sets $S, S_1, S_2, S_3, \ldots \subset \Omega$ with $S \subset \bigcup_{n\geq 1} S_n$,

$$\mu(S) \;\leq\; \sum_{n\geq 1} \mu(S_n).$$

This is obvious if $\mu(S_n) = \infty$ for some $n \geq 1$. Otherwise, for an arbitrary fixed (small) number $\epsilon > 0$ and any index $n \geq 1$, there exists a function $g_n \in \mathcal{F}^*$ such that $g_n \geq 1_{S_n}$ and $J(g_n) \leq \mu(S_n) + 2^{-n}\epsilon$. But then the definition of μ and Lemma 1.13 imply that

$$\mu(S) \;\leq\; J\Bigl(\sup_{n\geq 1} g_n\Bigr) \;\leq\; J\Bigl(\sum_{n\geq 1} g_n\Bigr) \;=\; \sum_{n\geq 1} J(g_n) \;\leq\; \sum_{n\geq 1} \mu(S_n) + \epsilon.$$

As $\epsilon \downarrow 0$, we obtain the desired inequality.

Step 2. $\mu(U) = J(1_U)$ for any $U \in \mathcal{U}(\mathcal{F})$.

Proof: By Lemma 1.8, $1_U \in \mathcal{F}^*$, and $J(g) \geq J(1_U)$ for any $g \in \mathcal{F}^*$ such that $g \geq 1_U$.

Step 3. Every set $U \in \mathcal{U}(\mathcal{F})$ is μ-measurable, that is,

$$\mu(S) \;=\; \mu(S \cap U) + \mu(S \setminus U).$$

In particular, by Carathéodory's theory of outer measures, the restriction of μ to $\sigma(\mathcal{F})$ defines a measure on that σ-field.

Proof: Since μ is an outer measure, it suffices to show that $\mu(S) \geq \mu(S \cap U) + \mu(S \setminus U)$ in case of $\mu(S) < \infty$. Let $g \in \mathcal{F}^*$ with $g \geq 1_S$. Let $(g_n)_{n\geq 1}$ and $(f_n)_{n\geq 1}$ be pointwise increasing sequences in $\mathcal{F}^+$ with limits g and $g \wedge 1_U$, respectively. Without loss of generality let $g_n \geq f_n$ for all $n \geq 1$. On the one hand,

$$\mu(S \cap U) \;\leq\; J(g \wedge 1_U) \;=\; \lim_{n\to\infty} J(f_n).$$

On the other hand, for any fixed $n_o \geq 1$,

$$1_{S\setminus U} \;=\; 1_S(1 - 1_U) \;\leq\; g(1 - 1_U) \;\leq\; g - g \wedge 1_U \;\leq\; g - f_{n_o}.$$

But $g - f_{n_o}$ is the limit of the pointwise increasing sequence $(g_n - f_{n_o})_{n \geq n_o}$ in $\mathcal{F}^+$, whence $g - f_{n_o} \in \mathcal{F}^*$ and

$$\begin{aligned} \mu(S \setminus U) \;\leq\; J(g - f_{n_o}) \;&=\; \lim_{n\to\infty} J(g_n - f_{n_o}) \\ &=\; \lim_{n\to\infty} J(g_n) - J(f_{n_o}) \\ &=\; J(g) - J(f_{n_o}). \end{aligned}$$

Consequently, $\mu(S \cap U) + \mu(S \setminus U) \leq J(g) + \lim_{n\to\infty} J(f_n) - J(f_{n_o})$, and as $n_o \to \infty$, we obtain the inequality

$$\mu(S \cap U) + \mu(S \setminus U) \;\leq\; J(g)$$

for any $g \in \mathcal{F}^*$ such that $g \geq 1_S$. This implies the desired inequality $\mu(S \cap U) + \mu(S \setminus U) \leq \mu(S)$.

Step 4. $\mathcal{F} \subset \mathcal{L}^1(\mu)$ and $J(f) = \int f \, d\mu$ for all $f \in \mathcal{F}$.

Proof: It suffices to show that $J(f) = \int f \, d\mu$ for any $f \in \mathcal{F}^+$. For $n \geq 1$ let

$$f_n \;:=\; 2^{-n} \sum_{k \geq 1} 1_{\{f > 2^{-n} k\}}.$$

For any $\omega \in \{f > 0\}$, $f_n(\omega)$ is the largest point on the grid $\{2^{-n} z \colon z \in \mathbb{N}_0\}$ which is strictly smaller than $f(\omega)$. Since $\{2^{-n} z \colon z \in \mathbb{N}_0\} \subset \{2^{-(n+1)} z \colon z \in \mathbb{N}_0\}$ and $f_n = 0$ on $\{f = 0\}$, the function sequence $(f_n)_{n \geq 1}$ is pointwise increasing with limit f. For any constant $c \geq 0$, the set $\{f > c\}$ is $\mathcal{F}$-open, and

$$J(1_{\{f > c\}}) \;=\; \mu(\{f > c\}) \;=\; \int 1_{\{f > c\}} \, d\mu.$$

Consequently, $f_n \in \mathcal{F}^*$, so by Lemma 1.13 and monotone convergence for usual integrals,

$$J(f) \;=\; \lim_{n\to\infty} J(f_n) \quad \text{and} \quad \int f \, d\mu \;=\; \lim_{n\to\infty} \int f_n \, d\mu.$$

But for any fixed n, Lemma 1.13 and standard properties of usual integrals reveal that

$$J(f_n) \;=\; 2^{-n} \sum_{k \geq 1} J\big(1_{\{f > 2^{-n} k\}}\big) \;=\; 2^{-n} \sum_{k \geq 1} \int 1_{\{f > 2^{-n} k\}} \, d\mu \;=\; \int f_n \, d\mu.$$

This shows that $J(f) = \int f \, d\mu$.

Step 5. For any set $S \subset \Omega$,

$$\mu(S) \;=\; \tilde{\mu}(S) := \inf\big\{\mu(U) \colon U \in \mathcal{U}(\mathcal{F}), U \supset S\big\}.$$

Proof: Since $U \supset S$ is equivalent to $1_U \geq 1_S$, and since $1_U \in \mathcal{F}^*$ for any $U \in \mathcal{U}(\mathcal{F})$, we obtain the inequality $\mu(S) \leq \tilde{\mu}(S)$.

On the other hand, for any fixed number $c \in (0, 1)$ and arbitrary functions $g \in \mathcal{F}^*$ with $g \geq S$,

$$c^{-1} g \;\geq\; 1_{\{g > c\}} \;\geq\; 1_S.$$

Since $\{g > c\} \in \mathcal{U}(\mathcal{F})$, this implies that

$$J(g) \;\geq\; c\mu(\{g > c\}) \;\geq\; c\tilde{\mu}(S).$$

Letting $c \uparrow 1$ reveals that $J(g) \geq \tilde{\mu}(S)$, whence $\mu(S) \geq \tilde{\mu}(S)$.

Step 6. A measure μ on $\sigma(\mathcal{F})$ with the stated properties is unique.

Proof: The measure μ is uniquely determined by its values $\mu(U)$, $U \in \mathcal{U}(\mathcal{F})$. But for each $U \in \mathcal{U}(\mathcal{F})$ there exists a pointwise increasing sequence $(f_n)_{n\geq 1}$ in $\mathcal{F}^+$ with limit 1_U, so

$$\mu(U) \;=\; \int 1_U \, d\mu \;=\; \lim_{n\to\infty} \int f_n \, d\mu \;=\; \lim_{n\to\infty} J(f_n).$$

Thus, the measure μ is uniquely determined by the given functional $J : \mathcal{F} \to \mathbb{R}$. □

1.3 Representations of Dual Spaces

Suppose for the moment that $\mathcal{F}$ is an arbitrary real vector space, equipped with a seminorm $\|\cdot\|$. The dual space of $(\mathcal{F}, \|\cdot\|)$ is the space of linear functionals $L : \mathcal{F} \to \mathbb{R}$ which are continuous with respect to the seminorm $\|\cdot\|$. Continuity of L is equivalent to

$$\sup_{f\in\mathcal{F}:\, \|f\|\leq 1} |L(f)| \;<\; \infty.$$

In Functional Analysis, an important question is how the dual space can be represented explicitly. Many of the known results are called a Riesz representation, honoring the hungarian mathematician Frigyes Riesz (1880-1956) who (co-)derived many of these results.

Specifically, let $\mathcal{F}$ be a Stone lattice of functions on a set Ω. Suppose in addition that there is a seminorm $\|\cdot\|$ on $\mathcal{F}$ satisfying the following two properties:

$$\|f\| \;\leq\; \|g\| \quad \text{whenever } f, g \in \mathcal{F}^+ \text{ with } f \leq g.$$

For any pointwise decreasing sequence $(f_n)_{n\geq 1}$ in $\mathcal{F}^+$ with limit 0,

$$\lim_{n\to\infty} \|f_n\| \;=\; 0.$$

The next theorem shows that any continuous linear functional on $(\mathcal{F}, \|\cdot\|)$ can be represented with certain integrals.

Theorem 1.15. *For any functional L in the dual space of $(\mathcal{F}, \|\cdot\|)$ there exist measures μ^+, μ^- on $\sigma(\mathcal{F})$ such that $\mathcal{F} \subset \mathcal{L}^1(\mu^+ + \mu^-)$ and*

$$L(f) \;=\; \int f \, d\mu^+ - \int f \, d\mu^- \quad \textit{for all } f \in \mathcal{F}.$$

Here is a first special case of this theorem:

Theorem 1.16 (Riesz–Markov–Kakutani). *Let (Ω, d) be a compact metric space, and let $\mathcal{C}(\Omega)$ be the family of continuous functions $f : \Omega \to \mathbb{R}$ with respect to d, equipped with the supremum norm $\|\cdot\|_\infty$, that is, $\|f\|_\infty = \max_{\omega\in\Omega} |f(\omega)|$. Let $L : \mathcal{C}(\Omega) \to \mathbb{R}$ be a linear functional which is continuous with respect to $\|\cdot\|_\infty$. Then there exist finite measures μ^+, μ^- on $\mathrm{Borel}(\Omega)$ such that*

$$L(f) \;=\; \int f \, d\mu^+ - \int f \, d\mu^- \quad \textit{for all } f \in \mathcal{C}(\Omega).$$

Once we have learned more about signed measures, the conclusion of Theorem 1.16 can be reformulated as follows: There exists a probability measure P on $\mathrm{Borel}(\Omega)$ and a bounded, measurable function $h\colon \Omega \to \mathbb{R}$ such that

$$L(f) \;=\; \int f h \, dP \quad \text{for all } f \in \mathcal{C}(\Omega).$$

The proof of Theorem 1.15 uses the following two auxiliary results.

Lemma 1.17 (From additive, nonnegative to linear functionals). *Suppose that $J : \mathcal{F}^+ \to [0,\infty)$ is a functional which is additive in the sense that $J(f+g) = J(f) + J(g)$ for all $f, g \in \mathcal{F}^+$. Then,*

$$L(f) \;:=\; J(f^+) - J(f^-)$$

defines a linear functional $L\colon \mathcal{F} \to \mathbb{R}$ such that $L \equiv J$ on $\mathcal{F}^+$.

Lemma 1.18 (From linear to additive, nonnegative functionals). *Suppose that $L : \mathcal{F} \to \mathbb{R}$ is a linear functional. For $f \in \mathcal{F}^+$ let*

$$\begin{aligned} J(f) \;&:=\; \sup\bigl\{L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\}, \\ K(f) \;&:=\; \sup\bigl\{-L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\}. \end{aligned}$$

These two functionals $J, K : \mathcal{F}^+ \to [0,\infty]$ are additive in the sense that $J(f+g) = J(f) + J(g)$ and $K(f+g) = K(f) + K(g)$ for all $f, g \in \mathcal{F}^+$.

They satisfy the equations $J = K + L$ and $K = J - L$. In particular, J is real-valued if and only if K is real-valued.

If J and K are real-valued, then for arbitrary $f \in \mathcal{F}$,

$$L(f) \;=\; L_J(f) - L_K(f),$$

where $L_J(f) := J(f^+) - J(f^-)$ and $L_K(f) := K(f^+) - K(f^-)$.

The proof of Lemma 1.17 is left to the reader as an exercise.

Proof of Lemma 1.18. Since $J(f) \ge L(0) = 0$, the functional J is nonnegative. Let $f_1, f_2 \in \mathcal{F}^+$. For arbitrary $h_1, h_2 \in \mathcal{F}^+$ with $h_1 \le f_1$ and $h_2 \le f_2$,

$$J(f_1 + f_2) \;\ge\; L(h_1 + h_2) \;=\; L(h_1) + L(h_2),$$

and letting $L(h_j) \uparrow J(f_j)$ for $j = 1, 2$ reveals that

$$J(f_1 + f_2) \;\ge\; J(f_1) + J(f_2).$$

On the other hand, if $h \in \mathcal{F}^+$ with $h \le f_1 + f_2$, then $h = h_1 + h_2$ with the functions

$$h_1 \;:=\; h \wedge f_1 \;\le\; f_1 \quad \text{and} \quad h_2 \;:=\; (h - f_1)^+ \;\le\; f_2$$

in $\mathcal{F}^+$. Hence,

$$L(h) \;=\; L(h_1) + L(h_2) \;\le\; J(f_1) + J(f_2),$$

and letting $L(h) \uparrow J(f_1 + f_2)$ reveals that

$$J(f_1 + f_2) \;\le\; J(f_1) + J(f_2).$$

Replacing L with $-L$ reveals that K is also a nonnegative, additive functional on $\mathcal{F}^+$.

For $f \in \mathcal{F}^+$,

$$\begin{aligned} J(f) &= \sup\bigl\{L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \\ &= \sup\bigl\{L(f) - L(f-h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \\ &= L(f) + \sup\bigl\{-L(f-h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \\ &= L(f) + \sup\bigl\{-L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \\ &= L(f) + K(f), \end{aligned}$$

where the second to last step uses the fact that the sets $\{f - h\colon h \in \mathcal{F}^+, h \le f\}$ and $\{h \in \mathcal{F}^+\colon h \le f\}$ coincide. Thus, $J = K + L$ on $\mathcal{F}^+$, and this is equivalent to $K = J - L$.

Now suppose that J and K are real-valued on $\mathcal{F}^+$. Then we know from Lemma 1.17 that $L_J(f) := J(f^+) - J(f^-)$ and $L_K(f) := K(f^+) - K(f^-)$ define linear functionals on $\mathcal{F}$. Since $L = J - K$ on $\mathcal{F}^+$,

$$L(f) \;=\; L(f^+) - L(f^-) \;=\; J(f^+) - K(f^+) - J(f^-) + K(f^-) \;=\; L_J(f) - L_K(f)$$

for any function $f \in \mathcal{F}$. □

Proof of Theorem 1.15. Continuity of L means that $|L(f)| \le C\|f\|$ for all $f \in \mathcal{F}$ and some real constant $C = C(L) \ge 0$. The additional properties of the seminorm $\|\cdot\|$ imply that for $f \in \mathcal{F}^+$,

$$\begin{aligned} J(f) &:= \sup\bigl\{L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \begin{cases} \ge \max\{0, L(f)\}, \\ \le C\|f\|, \end{cases} \\ K(f) &:= \sup\bigl\{-L(h)\colon h \in \mathcal{F}^+, h \le f\bigr\} \begin{cases} \ge \max\{0, -L(f)\}, \\ \le C\|f\|. \end{cases} \end{aligned}$$

Consequently, J and K are additive functionals $\mathcal{F}^+$ with values in $[0, \infty)$, and L_J, L_K are well-defined linear functionals on $\mathcal{F}$ with $L = L_J - L_K$, see Lemmas 1.18 and 1.17. The sequential continuity property of $\|\cdot\|$ implies that L_J, L_K are abstract integrals on $\mathcal{F}$. Indeed, if $(f_n)_{n\ge1}$ is a pointwise increasing sequence in $\mathcal{F}^+$ with limit $f \in \mathcal{F}^+$, then for $M = J, K$,

$$L_M(f_n) \;=\; L_M(f) - L_M(f - f_n) \;=\; L_M(f) - M(f - f_n) \begin{cases} \ge L_M(f) - C\|f - f_n\|, \\ \le L_M(f). \end{cases}$$

As $n \to \infty$, $f - f_n \downarrow 0$, whence $\|f - f_n\| \to 0$ and, consequently, $L_M(f_n) \to L_M(f)$. According to Theorem 1.14, there exist measures μ^+, μ^- on $\mathcal{U}(\mathcal{F})$ such that for arbitrary $f \in \mathcal{F}$, $L_J(f) = \int f\, d\mu^+$ and $L_K(f) = \int f\, d\mu^-$, which implies the asserted representation of L. □

The proof of Theorem 1.16 uses a well-known result from analysis about pointwise and uniform convergence.

Lemma 1.19 (Dini). *Let (Ω, d) be a compact metric space, and let $(f_n)_{n\geq 1}$ be a pointwise decreasing sequence of continuous functions with limit 0. Then $\|f_n - f\|_\infty \to 0$ as $n \to \infty$.*

Proof of Lemma 1.19. For any $\epsilon > 0$, the sets $\{f_n < \epsilon\}$ are open subsets of Ω with $\{f_1 < \epsilon\} \subset \{f_2 < \epsilon\} \subset \{f_3 < \epsilon\} \subset \cdots$, and $\Omega = \bigcup_{n\geq 1}\{f_n < \epsilon\}$. But compactness of Ω implies that $\Omega = \bigcup_{n\leq n(\epsilon)}\{f_n < \epsilon\} = \{f_{n(\epsilon)} < \epsilon\}$ for some integer $n(\epsilon) \geq 1$, and this implies that $\|f_n\|_\infty < \epsilon$ for all $n \geq n(\epsilon)$. □

Proof of Theorem 1.16. In view of Theorem 1.15, note first that the supremum norm $\|\cdot\|_\infty$ has the additional required properties. It is obvious that $\|f\|_\infty \leq \|g\|_\infty$ for functions $0 \leq f \leq g$, and the second property about sequences follows from Lemma 1.19. Thus we can apply Theorem 1.15 and conclude that there exist measures μ^+ and μ^- on $\sigma(\mathcal{C}(\Omega))$ such that $\mathcal{F} \subset \mathcal{L}^1(\mu^+ + \mu^-)$ and $L(f) = \int f\, d\mu^+ - \int f\, d\mu^-$. Since $1 \in \mathcal{F}$, these measures $\mu^\pm$ are finite. Moreover, $\sigma(\mathcal{C}(\Omega))$ coincides with $\mathrm{Borel}(\Omega)$. On the one hand, $\{f > r\}$ is an open subset of Ω for all $f \in \mathcal{C}(\Omega)$ and $r \in \mathbb{R}$, whence $\sigma(\mathcal{C}(\Omega)) \subset \mathrm{Borel}(\Omega)$. On the other hand, if U is a nonvoid proper subset of Ω, then $f(x) := \inf\{d(x, y)\colon y \in \Omega \setminus A\}$ defines a continuous function such that $U = \{f > 0\}$. Hence, $\mathrm{Borel}(\Omega) \subset \sigma(\mathcal{C}(\Omega))$. □

Chapter 2

Signed Measures

2.1 The Hahn–Jordan Decomposition

For simplicity we restrict our attention to finite signed measures. Throughout this section let $(\Omega, \mathcal{A})$ be a measurable space. That means, Ω is a nonvoid set equipped with a σ-field $\mathcal{A}$ over Ω, the family of measurable subsets of Ω.

Definition 2.1 (Finite signed measure). A function $\nu : \mathcal{A} \to \mathbb{R}$ is called a *finite signed measure on* $(\Omega, \mathcal{A})$, if

(SM.1) $\nu(\emptyset) = 0$ and

(SM.2) ν is σ-additive, that is, for arbitrary disjoint sets $A_1, A_2, A_3, \ldots$ in $\mathcal{A}$,

$$\nu\Big(\bigcup_{n=1}^{\infty} A_n\Big) \;=\; \sum_{n=1}^{\infty} \nu(A_n).$$

Note that the only difference to a finite measure is that $\nu(A)$ may be negative for some sets $A \in \mathcal{A}$.

Example 2.2. A standard example of a finite signed measure is $\nu := Q - P$ with finite measures P, Q on $(\Omega, \mathcal{A})$. Properties (SM.1-2) follow immediately from analogous properties of measures.

Example 2.3. Another example is given by

$$\nu(A) \;:=\; \int_A f \, d\mu \;=\; \int 1_A f \, d\mu$$

with a measure μ on $(\Omega, \mathcal{A})$ and a function $f \in \mathcal{L}^1(\mu)$. That means, $f : \Omega \to \bar{\mathbb{R}}$ is $\mathcal{A}$-measurable with $\int |f| \, d\mu < \infty$. Here one can verify properties (SM.1-2) by means of linearity of integrals and dominated convergence.

Later on it will be shown that any finite signed measure may be represented as in Examples 2.2 and 2.3.

Remark 2.4 (Additivity). A finite signed measure ν is additive in the sense that

$$\nu\Big(\bigcup_{n=1}^{N} A_n\Big) \;=\; \sum_{n=1}^{N} \nu(A_n)$$

for arbitrary $N \in \mathbb{N}$ and disjoint sets $A_1, \dots, A_N \in \mathcal{A}$. This follows from (SM.1-2) by setting $A_n = \emptyset$ for $n > N$.

Exercise 2.5 (Continuity properties of signed measures). Let ν be a finite signed measure on $(\Omega, \mathcal{A})$. Show that for arbitrary sets $B_1 \subset B_2 \subset B_3 \subset \cdots$ in $\mathcal{A}$,

$$\nu\Big(\bigcup_{n=1}^{\infty} B_n\Big) \;=\; \lim_{n\to\infty} \nu(B_n).$$

Show that for arbitrary sets $C_1 \supset C_2 \supset C_3 \supset \cdots$ in $\mathcal{A}$,

$$\nu\Big(\bigcap_{n=1}^{\infty} C_n\Big) \;=\; \lim_{n\to\infty} \nu(C_n).$$

The next definition introduces two important concepts for the subsequent results.

Definition 2.6 (Positive and negative sets). Let ν be a finite signed measure on $(\Omega, \mathcal{A})$. A set $A_* \subset \Omega$ is called *ν-positive* if $A_* \in \mathcal{A}$ and

$$\nu(A) \;\geq\; 0 \quad \text{for all } A \in \mathcal{A} \text{ with } A \subset A_*.$$

A set $A_* \subset \Omega$ is called ν-negative, if $A_* \in \mathcal{A}$ and

$$\nu(A) \;\leq\; 0 \quad \text{for all } A \in \mathcal{A} \text{ with } A \subset A_*.$$

Since $-\nu$ is a finite signed measure too, a set $A_* \subset \Omega$ is ν-positive or ν-negative if and only if it is $(-\nu)$-negative or $(-\nu)$-positive, respectively.

Here is a key result for the main theorems in this and the next section.

Proposition 2.7 (Existence of nontrivial positive sets). *Let ν be a finite signed measure on $(\Omega, \mathcal{A})$, and let $A_0 \in \mathcal{A}$ with $\nu(A_0) > 0$. Then there exists a ν-positive set $A_* \subset A_0$ with $\nu(A_*) \geq \nu(A_0)$.*

Proof of Proposition 2.7. We define

$$\delta_0 \;:=\; \sup\big\{-\nu(B) : B \in \mathcal{A}, B \subset A_0\big\} \;\geq\; 0.$$

Then we write $A_0 = A_1 \cup B_1$ with disjoint measurable sets A_1 and B_1 such that

$$-\nu(B_1) \;\geq\; \min\{\delta_0/2, 1\}.$$

This procedure can be iterated. After k steps we have measurable sets $A_0 \supset A_1 \supset \cdots \supset A_k$, and we consider the number

$$\delta_k \;:=\; \sup\big\{-\nu(B) : B \in \mathcal{A}, B \subset A_k\big\} \;\geq\; 0.$$

Then we write $A_k = A_{k+1} \cup B_{k+1}$ with disjoint measurable sets A_{k+1} and B_{k+1} such that

$$-\nu(B_{k+1}) \;\geq\; \min\{\delta_k/2, 1\}.$$

This construction yields a non-increasing sequence $(A_k)_{k\geq 0}$ and the disjoint sets $B_k = A_{k-1}\setminus A_k$, $k \geq 1$. We may write

$$A_0 \;=\; A_* \cup B_*$$

with the disjoint sets

$$A_* \;=\; \bigcap_{k\geq 0} A_k \quad \text{and} \quad B_* \;=\; \bigcup_{k\geq 1} B_k.$$

Since $\nu(B_*) = \sum_{k=1}^{\infty} \nu(B_k)$ with nonpositive summands $\nu(B_k)$, we obtain the inequality

$$\nu(A_*) \;=\; \nu(A_0) - \nu(B_*) \;\geq\; \nu(A_0).$$

Moreover, the sequence $(\nu(B_k))_{k\geq 1}$ converges to 0, so the inequalities $0 \leq \min\{\delta_k/2, 1\} \leq -\nu(B_{k+1})$ imply that $\lim_{k\to\infty} \delta_k = 0$. Consequently, for any measurable set $A \subset A_*$,

$$-\nu(A) \;\leq\; \inf_{k\geq 0} \sup\bigl\{-\nu(A') : A' \in \mathcal{A}, A' \subset A_k\bigr\} \;=\; \inf_{k\geq 0} \delta_k \;=\; 0,$$

whence $\nu(A) \geq 0$. This shows that A_* is a ν-positive set. □

Exercise 2.8 (Unions of positive sets). Let ν be a finite signed measure on $(\Omega, \mathcal{A})$. Let $(A_n)_{n\geq 1}$ be a sequence of ν-positive sets. Show that $A_* := \bigcup_{n\geq 1} A_n$ is also ν-positive and satisfies

$$\nu(A_*) \;\geq\; \sup_{n\geq 1} \nu(A_n).$$

With Proposition 2.7 and Exercise 2.8 one can prove the following representation of signed measures:

Theorem 2.9 (Hahn–Jordan decomposition). *Let ν be a finite signed measure on $(\Omega, \mathcal{A})$. Then $\Omega = \Omega_+ \cup \Omega_-$ with disjoint sets Ω_+, Ω_- such that Ω_+ is ν-positive and Ω_- is ν-negative. In other words,*

$$\nu^+(A) \;:=\; \nu(A \cap \Omega_+) \quad \textit{and} \quad \nu^-(A) \;:=\; -\nu(A \cap \Omega_-)$$

defines two finite measures ν^+, ν^- on $(\Omega, \mathcal{A})$ such that $\nu = \nu^+ - \nu^-$, and these two measures have disjoint support in the sense that $\nu^+(\Omega_-) = \nu^-(\Omega_+) = 0$.

Example 2.10. Before proving Theorem 2.9, we revisit Example 2.3, i.e. $\nu(A) = \int_A f\, d\mu$ for some measure μ on $(\Omega, \mathcal{A})$ and a function $f \in \mathcal{L}^1(\mu)$. Here a Hahn–Jordan decomposition is given by

$$\Omega_+ \;:=\; \{f \geq 0\} \quad \text{and} \quad \Omega_- := \{f < 0\}.$$

The corresponding measures $\nu^{\pm}$ are given by

$$\nu^{\pm}(A) \;=\; \int_A f^{\pm}\, d\mu$$

with $f^{\pm}(\omega) = \max\{\pm f(\omega), 0\}$.

Remark 2.11. Theorem 2.9 implies that *any* finite signed measure ν may be represented as in Example 2.3. Indeed, let ν, Ω_+, Ω_-, ν^+ and ν^- be as in Theorem 2.9. Then

$$\nu(A) \;=\; \int_A f\, d\mu \quad \text{for all } A \in \mathcal{A},$$

with the finite measure $\mu := \nu^+ + \nu^-$ and the measurable function $f := 1_{\Omega_+} - 1_{\Omega_-}$.

Proof of Theorem 2.9. Let $(A_n)_{n\geq 1}$ be a sequence of measurable sets such that

$$\lim_{n\to\infty} \nu(A_n) \;=\; C_+ := \sup\bigl\{\nu(A) : A \in \mathcal{A}\bigr\}.$$

According to Proposition 2.7, we may assume without loss of generality that all sets A_n are ν-positive. But then,

$$\Omega_+ \;:=\; \bigcup_{n\geq 1} A_n$$

is a ν-positive set such that $\nu(\Omega_+) = C_+$, see Exercise 2.8. In particular, $C_+ < \infty$. Moreover, $\Omega_- := \Omega \setminus \Omega_+$ is a ν-negative set. Indeed, if $\nu(A_0) > 0$ for some measurable set $A_0 \subset \Omega_-$, then $\nu(\Omega_+ \cup A_0) = \nu(\Omega_+) + \nu(A_0) > C_+$, a contradition to the definition of C_+. □

Exercise 2.12. Let $\Omega = \Omega_+ \cup \Omega_-$ as in Theorem 2.9. Show that this partition of Ω is essentially unique in the following sense: If $\Omega = B_+ \cup B_-$ with a ν-positive set B_+ and a ν-negative set B_-, then

$$\nu(A) \;=\; 0 \quad \text{for any measurable } A \subset (\Omega_+ \triangle B_+) \cup (\Omega_- \triangle B_-).$$

Remark 2.13. Theorem 2.9 implies, that *any* finite signed measure ν may be represented as the difference of two finite measures. The next exercise shows that the particular measures $\nu^{\pm}$ are minimal in a certain sense.

Exercise 2.14. Let ν be a finite signed measure on $(\Omega, \mathcal{A})$ with Hahn–Jordan decomposition $\nu = \nu^+ - \nu^-$.

(a) Show that $-\nu^-(A) \leq \nu(A) \leq \nu^+(A)$ for arbitrary $A \in \mathcal{A}$.

(b) Let $\nu = \mu^+ - \mu^-$ with finite measures μ^+, μ^- on $(\Omega, \mathcal{A})$. Show that there exists a finite measure μ_o on $(\Omega, \mathcal{A})$ such that $\mu^+ = \nu^+ + \mu_o$ and $\mu^- = \nu^- + \mu_o$. Deduce from the latter fact that ν^+ and ν^- are "minimal" in the sense that

$$\mu^+ + \mu^- \;\geq\; \nu^+ + \nu^-$$

with equality if and only if $\mu^+ = \nu^+$ and $\mu^- = \nu^-$.

Exercise 2.15. Let P and Q be finite measures on $\mathbb{R}$ with densities f and g, respectively, with respect to Lebesgue measure. Determine a Hahn–Jordan decomposition of $\nu := Q - P$ in terms of f and g.

Illustrate your solution graphically in case of $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(0, 2^2)$.

2.2 Radon–Nikodym Derivatives

In this section we consider measures P and Q on $(\Omega, \mathcal{A})$ and investigate under which conditions Q has a density f with respect to P. That means, $f : \Omega \to [0, \infty)$ is an $\mathcal{A}$-measurable function such that

$$Q(A) \;=\; \int_A f \, dP \quad \text{for all } A \in \mathcal{A}. \tag{2.1}$$

Such a function f is also called a *Radon–Nikodym derivative of Q with respect to P*, and (2.1) is sometimes abbreviated as

$$f \;=\; \frac{dQ}{dP}.$$

If a density of Q with respect to P exists, arbitrary integrals with respect to Q may be rewritten as integrals with respect to P.

Lemma 2.16. *Suppose that Q has a density f with respect to P in the sense of* (2.1). *Then for arbitrary $\mathcal{A}$-measurable functions $h : \Omega \to \bar{\mathbb{R}}$,*

$$\int h\, dQ \;=\; \int hf\, dP \tag{2.2}$$

whenever one of the two integrals is well-defined in $\bar{\mathbb{R}}$.

Proof of Lemma 2.16. We follow a standard route in measure theory, also known as "measure-theoretic induction". Assumption (2.1) is equivalent to (2.2) for arbitrary indicator functions $h = 1_A$, $A \in \mathcal{A}$. By linearity of integrals, (2.2) holds true for "simple functions" h, that means, functions $h = \sum_{j=1}^m \lambda_j 1_{A_j}$ with $m \in \mathbb{N}$, constants $\lambda_j \geq 0$ and sets $A_j \in \mathcal{A}$. For an arbitrary measurable function $h : \Omega \to [0, \infty]$, there exists a sequence $(h_n)_n$ of simple functions such that $(h_n)_n \uparrow h$ pointwise. A standard construction is given by

$$h_n(\omega) \;:=\; 2^{-n} \sum_{j=1}^{n2^n} 1_{[h(\omega) \geq 2^{-n} j]}.$$

But then $(h_n f)_n \uparrow hf$, so by monotone convergence,

$$\int h\, dQ \;=\; \lim_{n\to\infty} \int h_n\, dQ \;=\; \lim_{n\to\infty} \int h_n f\, dP \;=\; \int hf\, dP.$$

Finally, any measurable function $h : \Omega \to \bar{\mathbb{R}}$ may be written as $h = h^+ - h^-$ with $h^\pm = (\pm h) \vee 0$. Then $(hf)^\pm = h^\pm f$, because $f \geq 0$, and $\int h^\pm\, dQ = \int (hf)^\pm\, dP$. Hence

$$\int h\, dQ \;=\; \int h^+\, dQ - \int h^-\, dQ \;=\; \int (fh)^+\, dP - \int (hf)^-\, dP \;=\; \int hf\, dP,$$

whenever these differences are well-defined in $\bar{\mathbb{R}}$. □

Corollary 2.17. *Let P, Q and R be measures on $(\Omega, \mathcal{A})$ such that densities $f = dQ/dP$ and $g = dR/dQ$ exist. Then $fg = dR/dP$. Furthermore, if $f > 0$, then $f^{-1} = dP/dQ$.*

Proof of Corollary 2.17. For any set $A \in \mathcal{A}$,

$$\int_A fg\, dP \;=\; \int 1_A g\, f\, dP \;=\; \int 1_A g\, dQ \;=\; \int_A g\, dQ \;=\; R(A),$$

where the second step follows from Lemma 2.16 applied to $h = 1_A g$.

In case of $f > 0$, for any $A \in \mathcal{A}$,

$$P(A) \;=\; \int 1_A\, dP \;=\; \int 1_A f^{-1} f\, dP \;=\; \int 1_A f^{-1}\, dQ \;=\; \int_A f^{-1}\, dQ$$

by Lemma 2.16 applied to $h = 1_A f^{-1}$. □

Exercise 2.18. Let P be a measure on $(\Omega, \mathcal{A})$, and for a fixed $\Omega_o \in \mathcal{A}$ let $P_o(A) := P(A \cap \Omega_o)$. Show that P_o defines a measure on $(\Omega, \mathcal{A})$ with $P_o(\Omega) = P(\Omega_o)$ and $P_o(\Omega \setminus \Omega_o) = 0$. Determine a version of dP_o/dP. Then deduce that for arbitrary measurable functions $h : \Omega \to \bar{\mathbb{R}}$,

$$\int h \, dP_o \;=\; \int_{\Omega_o} h \, dP$$

whenever one (and thus both) of these integrals is well-defined in $\bar{\mathbb{R}}$.

Exercise 2.19 (Transformations). Let $(\Omega, \mathcal{A})$ and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ be measurable spaces, and let $\tau : \Omega \to \tilde{\Omega}$ be a bijective mapping such that τ and τ^{-1} are measurable.

(a) Let M be a measure on $(\Omega, \mathcal{A})$, and let $\tilde{M} := M \circ \tau^{-1}$, a measure on $(\tilde{\Omega}, \tilde{\mathcal{A}})$. Show that

$$\int \tilde{h} \, d\tilde{M} \;=\; \int \tilde{h} \circ \tau \, dM$$

for arbitrary measurable functions $\tilde{h} : \tilde{\Omega} \to \bar{\mathbb{R}}$ such that one of these integrals is well-defined in $\bar{\mathbb{R}}$.

(b) Let P and Q be measures on $(\Omega, \mathcal{A})$ such that a density $f = dQ/dP$ exists. Show that $\tilde{Q} := Q \circ \tau^{-1}$ has density $\tilde{f} := f \circ \tau^{-1}$ with respect to $\tilde{P} := P \circ \tau^{-1}$.

2.2.1 Finite measures

Now we consider the case of finite measures P and Q in more detail. The next lemma implies that a density of Q with respect to P, if it exists, is P-almost everywhere unique.

Lemma 2.20 (Uniqueness of densities). *Let P be a finite measure on $(\Omega, \mathcal{A})$, and for $j = 1, 2$ let $Q_j : \mathcal{A} \to \mathbb{R}$ be given by $Q_j(A) := \int_A f_j \, dP$ with a function $f_j \in \mathcal{L}^1(P)$. If $Q_1 \le Q_2$ on $\mathcal{A}$, then*

$$P(\{f_1 > f_2\}) \;=\; 0.$$

Exercise 2.21. Prove Lemma 2.20.

In what follows, we shall *construct* a density of Q with respect to P. The main idea for the construction results from an elementary consideration in the next exercise.

Exercise 2.22 (Superlevel sets). Let P and Q be finite measures on $(\Omega, \mathcal{A})$ such that Q has a density $f \in \mathcal{L}^1(P)$ with respect to P. Show that for any $\lambda > 0$, a set $A \in \mathcal{A}$ with $\{f > \lambda\} \subset A \subset \{f \ge \lambda\}$ maximizes

$$\mathcal{A} \ni A \;\mapsto\; Q(A) - \lambda P(A).$$

More generally, show that a set $A \in \mathcal{A}$ maximizes $Q - \lambda P$ if and only if

$$P(A \cap \{f < \lambda\}) \;=\; 0 \;=\; P(\{f > \lambda\} \setminus A).$$

Now we state and prove the first main result of this section.

Theorem 2.23. *Let P and Q be finite measures on $(\Omega, \mathcal{A})$. There exist a set $B_* \in \mathcal{A}$ with $P(B_*) = 0$ and a nonnegative function $f \in \mathcal{L}^1(P)$ such that*

$$Q(A) \;=\; Q(A \cap B_*) + \int_A f \, dP \quad \text{for arbitrary } A \in \mathcal{A}.$$

Proof of Theorem 2.23. Motivated by Exercise 2.22, we consider for any $\lambda > 0$ the finite signed measure $Q - \lambda P$. According to Theorem 2.9, there exists a set $A_\lambda \in \mathcal{A}$ such that A_λ is $(Q-\lambda P)$-positive and $\Omega \setminus A_\lambda$ is $(Q - \lambda P)$-negative. Now we could try to combine all these sets and define a function f via $\{f > \lambda\} = A_\lambda$. But this is easier said than done. The problem is that we are dealing with uncountably many sets A_λ.

The precise construction starts from a countable dense subset Λ of $(0, \infty)$, for instance, $\Lambda = \mathbb{Q} \cap (0, \infty)$. For any $\lambda \in \Lambda$ let $A_\lambda \in \mathcal{A}$ be chosen as above. Now we "clean up" these sets as follows: For $t \geq 0$ let

$$B_t \ := \bigcup_{\lambda \in \Lambda : \lambda > t} A_\lambda.$$

Note that $\{A_\lambda : \lambda \in \Lambda, \lambda > t\}$ is a countable family of $(Q - tP)$-positive sets, because $Q - \lambda P \leq Q - tP$ for $\lambda > t$. Hence, B_t is measurable and $(Q-tP)$-positive, too, see Exercise 2.8. Moreover, for any measurable subset A of $\Omega \setminus B_t = \bigcap_{\lambda \in \Lambda : \lambda > t} \Omega \setminus A_\lambda$,

$$(Q - tP)(A) \ = \ \sup_{\lambda \in \Lambda : \lambda > t} (Q - \lambda P)(A) \ \leq \ 0,$$

because $\Omega \setminus A_\lambda$ is $(Q - \lambda P)$-negative. Hence $\Omega \setminus B_t$ is a $(Q - tP)$-negative set. In particular,

$$Q(\Omega \setminus B_0) \ = \ 0.$$

Note also that by construction,

$$B_s \ \supset \ B_t \quad \text{for } 0 \leq s < t$$

and

$$B_s \ = \ \bigcup_{t > s} B_t \quad \text{for } s \geq 0,$$

because

$$B_s \ = \ \bigcup_{\lambda \in \Lambda : \lambda > s} A_\lambda \ = \ \bigcup_{t > s} \bigcup_{\lambda \in \Lambda : \lambda > t} A_\lambda \ = \ \bigcup_{t > s} B_t.$$

The intersection

$$B_* \ := \ \bigcap_{t > 0} B_t$$

satisfies $P(B_*) = 0$, because

$$0 \ \leq \ P(B_*) \ \leq \ \inf_{t > 0} P(B_t) \ \leq \ \inf_{t > 0} Q(B_t)/t \ \leq \ \inf_{t > 0} Q(\Omega)/t \ = \ 0.$$

Next we define a function $\bar{f} : \Omega \to [0, \infty]$ via

$$\bar{f}(\omega) \ := \ \begin{cases} 0 & \text{if } \omega \in \Omega \setminus B_0, \\ \sup\{t \geq 0 : \omega \in B_t\} & \text{if } \omega \in B_0. \end{cases}$$

Note that $\{\bar{f} = \infty\} = B_*$. Note also that

$$\{\bar{f} > t\} \ = \ \bigcup_{s > t} B_s \ = \ B_t \quad \text{for all } t \geq 0,$$

so $\bar{f}$ is measurable. Our goal is to show that $Q(A) = Q(A \cap B_*) + \int_A \bar{f}\, dP$ for arbitrary $A \in \mathcal{A}$. For any fixed parameter $\gamma > 1$, the set Ω may be partitioned into the disjoint sets $\Omega \setminus B_0$, B_* and

$$B_0 \setminus B_* \;=\; \bigcup_{z\in\mathbb{Z}} C_z \quad \text{with} \quad C_z \;:=\; B_{\gamma^z} \setminus B_{\gamma^{z+1}}.$$

Note that $\bar{f} = 0$ on $\Omega \setminus B_0$ and $\gamma^z < \bar{f} \le \gamma^{z+1}$ on C_z, $z \in \mathbb{Z}$. Moreover, C_z is $(Q - \gamma^z P)$-positive and $(Q - \gamma^{z+1}P)$-negative. Hence, for any set $A \in \mathcal{A}$,

$$Q(A \setminus B_0) \;=\; 0 \;=\; \int_{A\setminus B_0} \bar{f}\, dP,$$

and

$$Q(A \cap C_z) \begin{cases} \le\; \gamma^{z+1} P(A \cap C_z) \;\le\; \gamma \displaystyle\int_{A\cap C_z} \bar{f}\, dP, \\ \ge\; \gamma^{z} P(A \cap C_z) \;\ge\; \gamma^{-1} \displaystyle\int_{A\cap C_z} \bar{f}\, dP. \end{cases}$$

Consequently,

$$Q(A \setminus B_*) \;=\; \sum_{z\in\mathbb{Z}} Q(A \cap C_z) \begin{cases} \le\; \gamma \displaystyle\sum_{z\in\mathbb{Z}} \int_{A\cap C_z} \bar{f}\, dP \;=\; \gamma \int_{A\setminus B_*} \bar{f}\, dP, \\ \ge\; \gamma^{-1} \displaystyle\sum_{z\in\mathbb{Z}} \int_{A\cap C_z} \bar{f}\, dP \;=\; \gamma^{-1} \int_{A\setminus B_*} \bar{f}\, dP, \end{cases}$$

and for $\gamma \to 1$ we obtain the desired equation(s)

$$Q(A) \;=\; Q(A \cap B_*) + \int_{A\setminus B_*} \bar{f}\, dP \;=\; Q(A \cap B_*) + \int_A \bar{f}\, dP.$$

The latter equation follows from the fact that $P(B_*) = 0$. Indeed, this representation of Q remains valid if we replace $\bar{f}$ with the real-valued function $f := 1_{\Omega\setminus B_*}\bar{f}$. Setting $A = \Omega$ yields the equation $\int f\, dP = Q(\Omega) - Q(B_*) < \infty$, whence $f \in \mathcal{L}^1(P)$. □

Theorem 2.23 shows that Q is the sum of two measures, a "singular part with respect to P",

$$A \;\mapsto\; Q(A \cap B_*),$$

and an "absolutely continuous part with respect to P",

$$A \;\mapsto\; \int_A f\, dP.$$

The singular part is nontrivial if and only if $Q(B_*) > 0$. Otherwise Q is "absolutely continuous with respect to P" as defined in the next paragraph.

2.2.2 Absolute continuity and σ-finite measures

Definition 2.24 (Absolute continuity). Let P and Q be measures on $(\Omega, \mathcal{A})$. The measure Q is called *absolutely continuous with respect to* P if $Q(A) = 0$ for all sets $A \in \mathcal{A}$ such that $P(A) = 0$.

Definition 2.25 (σ-finiteness). A measure P on $(\Omega, \mathcal{A})$ is called σ-finite if there exists a sequence $(A_n)_{n\geq 1}$ in $\mathcal{A}$ such that $\Omega = \bigcup_{n\geq 1} A_n$ and $P(A_n) < \infty$ for all $n \geq 1$.

A standard example of a σ-finite measure is Lebesgue measure on $\mathbb{R}$. Here $\Omega = \mathbb{R}$ is the union of, say, all intervals $A_n := [-n, n]$, $n \in \mathbb{N}$.

Many results for finite measures may be extended to σ-finite measures by means of the following observation.

Lemma 2.26. *For a nonzero measure P on $(\Omega, \mathcal{A})$, the following two statements are equivalent:*

(i) *P is σ-finite.*

(ii) *There exist a probability measure P_o on $(\Omega, \mathcal{A})$ and an $\mathcal{A}$-measurable function $p : \Omega \to (0, \infty)$ such that*

$$P(A) \;=\; \int_A p\, dP_o \quad \text{for all } A \in \mathcal{A}.$$

Proof of Lemma 2.26. Suppose first that P may represented as in part (ii). Then $\Omega = \bigcup_{n\geq 1} A_n$ with $A_n := \{p \leq n\}$, and $P(A_n) \leq nP_o(A_n) \leq n$ for all n. Thus, P is σ-finite. Moreover, $P(\Omega) > 0$ because $\Omega = \bigcup_{m\geq 1}\{p \geq m^{-1}\}$, so $P(\{p \geq m^{-1}\}) \geq m^{-1}P_o(\{p \geq m^{-1}\}) > 0$ for sufficiently large m.

Now suppose that P is nonzero and σ-finite. In case of $P(\Omega) < \infty$, condition (ii) is satisfied with $P_o := P(\Omega)^{-1}P$ and $p \equiv P(\Omega)$. In case of $P(\Omega) = \infty$, let $\Omega = \bigcup_{n=1}^{\infty} A_n$ with sets $A_n \in \mathcal{A}$ such that $P(A_n) < \infty$, we may assume without loss of generality that these sets A_n are pairwise disjoint with $P(A_n) > 0$. But then

$$P_o(A) \;:=\; \sum_{n=1}^{\infty} 2^{-n} P(A_n)^{-1} P(A \cap A_n)$$

defines a probability measure on $(\Omega, \mathcal{A})$, and for any $A \in \mathcal{A}$,

$$P(A) \;=\; \sum_{n=1}^{\infty} 2^n P(A_n) P_o(A \cap A_n) \;=\; \int_A p\, dP_o$$

with $p(\omega) := \sum_{n=1}^{\infty} 2^n P(A_n) 1_{A_n}(\omega) > 0$. □

Theorem 2.23 implies the following result about σ-finite measures.

Theorem 2.27 (Radon–Nikodym). *Let P and Q be σ-finite measures on $(\Omega, \mathcal{A})$. Then the following two conditions are equivalent:*

(i) *Q is absolutely continuous with respect to P.*

(ii) *There exists a density f of Q with respect to P.*

In case of (i–ii), the density f is P-almost everywhere unique. That means, if $\tilde{f}$ is another density of Q with respect to P, then $P(\{\tilde{f} \neq f\}) = 0$.

Proof of Theorem 2.27. If condition (ii) is satisfied, then $P(A) = 0$ implies that $Q(A)$ equals $\int_A f\, dP = 0$, so condition (i) holds true as well.

Suppose that condition (i) is satisfied. If P and Q are finite, the existence of a density $f = dQ/dP$ is a consequence of Theorem 2.23. In the general case, let P_o and Q_o be probability measures on $(\Omega, \mathcal{A})$ such that there exist strictly positive densities $p = dP/dP_o$ and $q = dQ/dQ_o$. Then $p^{-1} = dP_o/dP$ and $q^{-1} = dQ_o/dQ$, see Corollary 2.17. Hence, for arbitrary sets $A \in \mathcal{A}$,

$$P(A) = 0 \text{ if and only if } P_o(A) = 0,$$
$$Q(A) = 0 \text{ if and only if } Q_o(A) = 0.$$

Thus condition (i) implies that Q_o is absolutely continuous with respect to P_o. Consequently, there exists a density $f_o = dQ_o/dP_o$. But then Lemma 2.16 implies that for any $\mathcal{A}$-measurable function $h : \Omega \to [0, \infty)$,

$$\int h\, dQ \;=\; \int h\, q\, dQ_o \;=\; \int h\, q f_o\, dP_o \;=\; \int h\, q f_o p^{-1}\, dP,$$

so $f := q f_o p^{-1} = dQ/dP$.

Suppose that $\tilde{f}$ is another version of dQ/dP. Then, $\tilde{f}_o := q^{-1}\tilde{f}p$ is another version of dQ_o/dP_o, and $\{\tilde{f}_o \neq f_o\} = \{\tilde{f} \neq f\}$. According to Exercise 2.20, $0 = P_o\{\tilde{f}_o \neq f_o\} = P_o\{\tilde{f} \neq f\}$, whence $P\{\tilde{f} \neq f\} = \int_{\{\tilde{f} \neq f\}} p\, dP_o = 0$. □

Exercise 2.28. Let P be a σ-finite measure on $(\Omega, \mathcal{A})$, and for some measurable function $f : \Omega \to [0, \infty)$, let $Q(A) := \int_A f\, dP$ for $A \in \mathcal{A}$. Show that Q is σ-finite too.

2.3 Another Riesz Representation

Combining the Radon–Nikodym theorem with Theorem 1.15 leads to a well-known representation theorem for L^p-spaces. Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. For $p \in [1, \infty)$ let $\mathcal{L}^p(\mu)$ be the set of measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\|f\|_{p,\mu} := \Big(\int |f|^p\, d\mu\Big)^{1/p} \;<\; \infty.$$

It is well-known that $\mathcal{L}^p(\mu)$ is a Stone lattice and $\|\cdot\|_{p,\mu}$ defines a seminorm on $\mathcal{L}^p(\mu)$. The same is true for the set $\mathcal{L}^\infty(\mu)$ of measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\|f\|_{\infty,\mu} := \inf\big\{r \geq 0 : \mu(\{|f| > r\}) = 0\big\} \;<\; \infty.$$

Theorem 2.29. *For an arbitrary $p \in [1, \infty)$, let $L : \mathcal{L}^p(\mu) \to \mathbb{R}$ be linear and continuous with respect to $\|\cdot\|_{p,\mu}$. Then there exists an function $h \in \mathcal{L}^q(\mu)$, where $q = p/(p-1) \in (1, \infty]$, such that*

$$L(f) \;=\; \int f h\, d\mu \quad \text{for all } f \in \mathcal{L}^p(\mu),$$

and

$$\sup_{f \in \mathcal{L}^p(\mu):\ \|f\|_{p,\mu} \leq 1} |L(f)| \;=\; \|h\|_{q,\mu}.$$

The function h is unique μ-almost everywhere.

In the proof of this theorem, the following result is useful.

Lemma 2.30 (Hölder). *Let $p \in [1,\infty)$ and $q = p/(p-1) \in (1,\infty]$. For any function $h \in \mathcal{L}^q(\mu)$,*

$$\sup_{f\in\mathcal{L}^p(\mu):\ \|f\|_{p,\mu}\le 1} \Big|\int f h\,d\mu\Big| \;=\; \|h\|_{q,\mu}.$$

Proof of Lemma 2.30. If $\|h\|_{q,\mu} = 0$, the assertion is obvious, because $\mu(\{h \neq 0\}) = 0$. Thus let $0 < \|h\|_{q,\mu} < \infty$.

In case of $p = 1$, it follows from $\mu(\{|h| > \|h\|_{\infty,\mu}\}) = 0$ that

$$\Big|\int f h\,d\mu\Big| \;\le\; \|f\|_{1,\infty}\|h\|_{\infty,\mu}.$$

On the other hand, $\mu(\{|h| \ge C\}) > 0$ for any fixed $0 < C < \|h\|_{\infty,\mu}$, and σ-finiteness of μ implies that $0 < \lambda := \mu(\{|h| \ge C\} \cap B) < \infty$ for some $B \in \mathcal{A}$. If we define $f := \lambda^{-1}\operatorname{sign}(h)1_{\{|h|\ge C\}\cap B}$, then $\|f\|_{1,\mu} = 1$ and

$$\int f h\,d\mu \;\ge\; C.$$

Hence, the supremum of $\big|\int f h\,d\mu\big|$ over all $f \in \mathcal{L}^1(\mu)$ with $\|f\|_{1,\mu} \le 1$ equals $\|h\|_{\infty,\mu}$.

In case of $p > 1$, the well-known Hölder inequality states that

$$\Big|\int f h\,d\mu\Big| \;\le\; \|f\|_{p,\mu}\|h\|_{q,\mu}$$

for arbitrary $f \in \mathcal{L}^p(\mu)$. Consequently, $\big|\int f h\,d\mu\big|$ is no larger than $\|h\|_{q,\mu}$ whenever $\|f\|_{p,\mu} \le 1$. It suffices to show that we have equality for a suitable f. To this end, let $f := \lambda\operatorname{sign}(h)|h|^{q-1} = \lambda\operatorname{sign}(h)|h|^{1/(p-1)}$ for some $\lambda > 0$. Then $\|f\|_{p,\mu} = \lambda\|h\|_{q,\mu}^{q/p}$, so $\lambda = \|h\|_{q,\mu}^{-q/p} = \|h\|_{q,\mu}^{-1/(p-1)}$ yields a function with $\|f\|_{p,\mu} = 1$, and

$$\int f h\,d\mu \;=\; \|h\|_{q,\mu}^{-1/(p-1)} \int |h|^q\,d\mu \;=\; \|h\|_{q,\mu}.$$

□

Proof of Theorem 2.29. As in the proof of Theorem 1.15 one can show that there exist two measures M_1 and M_2 on $\sigma(\mathcal{L}^p(\mu))$ such that $\mathcal{L}^p(\mu) \subset \mathcal{L}^1(M_1+M_2)$ and $L(f) = \int f\,dM_1 - \int f\,dM_2$ for any $f \in \mathcal{L}^p(\mu)$. Furthermore, for some constant C, $\int f\,dM_j \le C\|f\|_{p,\mu}$ for all nonnegative functions $f \in \mathcal{L}^p(\mu)$ and $j = 1, 2$.

First of all, $\sigma(\mathcal{L}^p(\mu)) = \mathcal{A}$. On the one hand, $\sigma(\mathcal{L}^p(\mu)) \subset \mathcal{A}$, because all functions $f \in \mathcal{L}^p(\mu)$ are $\mathcal{A}$-measurable. On the other hand, if $\Omega = \bigcup_{m\ge1} B_m$ with sets $B_m \in \mathcal{A}$ such that $\mu(B_m) < \infty$, then any set $A \in \mathcal{A}$ can be written as $A = \bigcup_{m\ge1} A \cap B_m$, and $1_{A\cap B_m} \in \mathcal{L}^p(\mu)$ for all $m \ge 1$, whence $A \in \sigma(\mathcal{L}^p(\mu))$, that is, $\mathcal{A} \subset \sigma(\mathcal{L}^p(\mu))$.

Thus for $j = 1, 2$, M_j is a measure on $\mathcal{A}$ such that $\int f\,dM_j \le C\|f\|_{p,\mu}$ for all nonnegative $f \in \mathcal{L}^p(\mu)$. In particular, if $f = 1_A$ for some $A \in \mathcal{A}$ with $\mu(A) = 0$, then $M_j(A) = 0$. Consequently, by the Radon–Nikodym theorem, there exists a density h_j of M_j with respect to μ, and $\int f\,dM_j = \int f h_j\,d\mu \le C\|f\|_{p,\mu}$ for all nonnegative $f \in \mathcal{L}^p(\mu)$. But this implies

that $\|h_j\|_{q,\mu} < \infty$. Indeed, let $\Omega = \bigcup_{m\geq 1} B_m$ with sets $B_1 \subset B_2 \subset B_3 \subset \cdots$ such that $\mu(B_m) < \infty$. Then $h_{jm} := 1_{B_m}(h_j \wedge m) \in \mathcal{L}^q(\mu)$ satisfies $h_{jm} \uparrow h_j$ pointwise as $m \to \infty$, whence $\|h_{jm}\|_{q,\mu} \to \|h\|_{q,\mu}$ as $m \to \infty$. For any nonnegative $f \in \mathcal{L}^p(\mu)$ with $\|f\|_{p,\mu} \leq 1$,

$$C \;\geq\; \int f h_j \, d\mu \;\geq\; \int f h_{jm} \, d\mu,$$

and for a suitable such function f, the integral on the right hand side equals $\|h_{jm}\|_{q,\mu}$, see the proof of Lemma 2.30. This shows that $\|h_j\|_{q,\mu} \leq C$.

All in all, $h := h_1 - h_2$ is a function in $\mathcal{L}^q(\mu)$ such that

$$L(f) \;=\; \int f h \, d\mu \quad \text{for all } f \in \mathcal{L}^p(\mu),$$

and it follows from Lemma 2.30 that the supremum of $|L(f)|$ over all $f \in \mathcal{L}^p(\mu)$ with $\|f\|_{p,\mu} \leq 1$ is equal to $\|h\|_{q,\mu}$.

Essential uniqueness of h can be verified as follows: If $h, \tilde{h} \in \mathcal{L}^q(\mu)$ such that $L(f) = \int f h \, d\mu = \int f \tilde{h} \, d\mu$ for all $f \in \mathcal{L}^p(\mu)$, then $\int f(h - \tilde{h}) \, d\mu = 0$ for all $f \in \mathcal{L}^p(\mu)$. Then it follows from Lemma 2.30 that $\|h - \tilde{h}\|_{q,\mu} = 0$, that is, $\mu(\{h \neq \tilde{h}\}) = 0$. □

Chapter 3

Conditional Expectations

Throughout this chapter let $(\Omega, \mathcal{A}, P)$ be a probability space, and let X be a random variable in $\mathcal{L}^1(P)$. That means, $X : \Omega \to \mathbb{R}$ is $\mathcal{A}$-measurable, and $\int |X| \, dP < \infty$.

3.1 Conditional expectations with respect to a sub-σ-field

Let $\mathcal{A}_o$ be a sub-σ-field of $\mathcal{A}$. One could think about a rather complex random experiment described by $(\Omega, \mathcal{A}, P)$, and the subfield $\mathcal{A}_o$ represents some partial aspects of it.

Theorem 3.1. *There exists a random variable $X_o \in \mathcal{L}^1(P|_{\mathcal{A}_o})$ such that*

$$\int_{A_o} X_o \, dP \;=\; \int_{A_o} X \, dP \quad \text{for all } A_o \in \mathcal{A}_o. \tag{3.1}$$

This random variable X_o is almost everywhere unique: If $\tilde{X}_o$ is another random variable satisfying (3.1)*, then $P(\tilde{X}_o \neq X_o) = 0$.*

Definition 3.2 (Conditional expectation, I). A random variable X_o as in Theorem 3.1 is called *(a version of the) conditional expectation of X, given the sub-σ-field $\mathcal{A}_o$*. A function X_o satisfying (3.1) is denoted by $\mathbb{E}(X \,|\, \mathcal{A}_o)$.

Proof of Theorem 3.1. Uniqueness of X_o almost everywhere follows from Lemma 2.20 (with $\mathcal{A}_o$ in place of $\mathcal{A}$), so it suffices to prove existence of X_o.

Let $X = X^+ - X^-$ with $X^\pm = \max(\pm X, 0)$. Then

$$Q^\pm(A) \;:=\; \int_A X^\pm \, dP$$

defines finite measures on $(\Omega, \mathcal{A})$ with density $X^\pm$ with respect to P. This implies that $Q^\pm|_{\mathcal{A}_o}$ is absolutely continuous with respect to $P|_{\mathcal{A}_o}$. Hence, by the Radon–Nikodym theorem, there exist densities $X_o^{(\pm)} \in \mathcal{L}^1(P|_{\mathcal{A}_o})$ of $Q^\pm|_{\mathcal{A}_o}$ with respect to $P|_{\mathcal{A}_o}$. This implies that $X_o := X_o^{(+)} - X_o^{(-)}$ has the desired property (3.1). □

Example 3.3 (Countable partitions). Let $\Omega = \bigcup_{n\geq 1} B_n$ with a sequence $(B_n)_{n\geq 1}$ of disjoint measurable sets, and let $\mathcal{A}_o$ be the smallest σ-field containing all these sets B_n, $n \geq 1$. Then

$$\mathbb{E}(X \,|\, \mathcal{A}_o) \;=\; \sum_{n\geq 1} \mathbb{E}(X \,|\, B_n) \cdot 1_{B_n} \quad \text{almost surely}$$

with the numbers

$$\mathbb{E}(X \,|\, B) \;:=\; \begin{cases} P(B)^{-1} \displaystyle\int_B X \, dP & \text{if } P(B) > 0, \\ \mathbb{E}(X) & \text{else.} \end{cases}$$

(The definition $\mathbb{E}(X \,|\, B) := \mathbb{E}(X)$ in case of $P(B) = 0$ is somewhat arbitrary.) Indeed, the function $X_o := \sum_{n\geq 1} \mathbb{E}(X \,|\, B_n) \cdot 1_{B_n}$ is constant on each set B_n, $n \geq 1$, so it is $\mathcal{A}_o$-measurable. Moreover, any set $A_o \in \mathcal{A}_o$ may be written as $A_o = \bigcup_{n\in M} B_n$ with $M \subset \mathbb{N}$, so

$$\int_{A_o} X \, dP \;=\; \sum_{n\in M} \int_{B_n} X \, dP \;=\; \sum_{n\in M} P(B_n)\, \mathbb{E}(X \,|\, B_n) \;=\; \sum_{n\in M} \int_{B_n} X_o \, dP \;=\; \int_{A_o} X_o \, dP.$$

Remark 3.4 (Properties of conditional expectations). In what follows, let $X, Y \in \mathcal{L}^1(P)$.

(a) For real numbers a, b,

$$\mathbb{E}(aX + bY \,|\, \mathcal{A}_o) \;=\; a\, \mathbb{E}(X \,|\, \mathcal{A}_o) + b\, \mathbb{E}(Y \,|\, \mathcal{A}_o) \quad \text{almost surely.}$$

This is a simple consequence of linearity of integrals.

(b) If $X \leq Y$ almost surely, then

$$\mathbb{E}(X \,|\, \mathcal{A}_o) \;\leq\; \mathbb{E}(Y \,|\, \mathcal{A}_o) \quad \text{almost surely.}$$

This is essentially a consequence of Lemma 2.20: It follows from $X \leq Y$ almost surely that

$$\int_{A_o} \mathbb{E}(X \,|\, \mathcal{A}_o)\, dP \;=\; \int_{A_o} X \, dP \;\leq\; \int_{A_o} Y \, dP \;=\; \int_{A_o} \mathbb{E}(Y \,|\, \mathcal{A}_o)\, dP \quad \text{for all } A_o \in \mathcal{A}_o.$$

Hence, $P\big(\mathbb{E}(X \,|\, \mathcal{A}_o) > \mathbb{E}(Y \,|\, \mathcal{A}_o)\big) = 0$.

(c) The mapping $X \mapsto \mathbb{E}(X \,|\, \mathcal{A}_o)$ is a weak contraction in the sense that

$$\big|\mathbb{E}(X \,|\, \mathcal{A}_o)\big| \;\leq\; \mathbb{E}(|X| \,|\, \mathcal{A}_o) \quad \text{almost surely.}$$

This follows from properties (a–b): Writing $X = X^+ - X^-$ with $X^\pm = \max(\pm X, 0)$, one may write $|X| = X^+ + X^-$, and we know that almost everywhere,

$$\begin{aligned}
\mathbb{E}(X^\pm \,|\, \mathcal{A}_o) \;&\geq\; 0 && \text{(property (b)),} \\
\mathbb{E}(X \,|\, \mathcal{A}_o) \;&=\; \mathbb{E}(X^+ - X^- \,|\, \mathcal{A}_o) \;=\; \mathbb{E}(X^+ \,|\, \mathcal{A}_o) - \mathbb{E}(X^- \,|\, \mathcal{A}_o) && \text{(property (a)),} \\
\mathbb{E}(|X| \,|\, \mathcal{A}_o) \;&=\; \mathbb{E}(X^+ + X^- \,|\, \mathcal{A}_o) \;=\; \mathbb{E}(X^+ \,|\, \mathcal{A}_o) + \mathbb{E}(X^- \,|\, \mathcal{A}_o) && \text{(property (a)),}
\end{aligned}$$

whence

$$\big|\mathbb{E}(X \,|\, \mathcal{A}_o)\big| \;\leq\; \mathbb{E}(|X| \,|\, \mathcal{A}_o).$$

(d) It follows from properties (a) and (c) that

$$\int \big|\mathbb{E}(X \,|\, \mathcal{A}_o) - \mathbb{E}(Y \,|\, \mathcal{A}_o)\big|\, dP \;\le\; \int |X - Y|\, dP.$$

(e) For any bounded, $\mathcal{A}_o$-measurable random variable $Z_o : \Omega \to \mathbb{R}$,

$$\int X Z_o \, dP \;=\; \int \mathbb{E}(X \,|\, \mathcal{A}_o) Z_o \, dP. \tag{3.2}$$

Indeed, by definition of $\mathbb{E}(X \,|\, \mathcal{A}_o)$, equation (3.2) is true for indicator functions $Z_o = 1_{A_o}$, $A_o \in \mathcal{A}_o$. By linearity of integrals, (3.2) is even true for $\mathcal{A}_o$-measurable functions $Z_o : \Omega \to \mathbb{R}$ taking only finitely many values. But for any bounded, $\mathcal{A}_o$-measurable function $Z_o : \Omega \to \mathbb{R}$, there exists a sequence $(Z_n)_{n\ge1}$ of $\mathcal{A}_o$-measurable functions taking only finitely many values such that

$$\delta_n \;:=\; \sup_{\omega\in\Omega} \big|Z_n(\omega) - Z_o(\omega)\big| \;\to\; 0 \quad \text{as } n \to \infty.$$

Now (3.2) follows from the fact that for arbitrary $n \ge 1$,

$$\begin{aligned}\Big|\int X Z_o \, dP - \int \mathbb{E}(X \,|\, \mathcal{A}_o) Z_o \, dP\Big| \;&=\; \Big|\int \big(X - \mathbb{E}(X \,|\, \mathcal{A}_o)\big)(Z_o - Z_n)\, dP\Big| \\ &\le\; \delta_n \int \big|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\big|\, dP.\end{aligned}$$

Exercise 3.5. Let P be the exponential distribution on $\Omega = [0,\infty)$ with rate parameter $\lambda > 0$, and let $\mathcal{A}_o$ be the smallest σ-field containing all intervals $[k, k+1)$, $k \in \mathbb{N}_0$. Determine $\mathbb{E}(X \,|\, \mathcal{A}_o)$ in case of $X(\omega) := \omega$.

Exercise 3.6 ("Tower property" of conditional expectations). Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X \in \mathcal{L}^1(P)$. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be σ-fields over Ω such that $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathcal{A}$.

(a) Show that $\mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{A}_1) \,|\, \mathcal{A}_2\big) = \mathbb{E}(X \,|\, \mathcal{A}_1)$ almost surely.

(b) Show the so-called tower property: $\mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{A}_2) \,|\, \mathcal{A}_1\big) = \mathbb{E}(X \,|\, \mathcal{A}_1)$ almost surely.

We end this section with an extension of Jensen's inequality for expectations to conditional expectations.

Lemma 3.7 (Jensen's inequality). *Let $X \in \mathcal{L}^1(P)$, and let $\psi : \mathbb{R} \to \mathbb{R}$ be convex such that $\int \psi(X)\, dP < \infty$. Then*

$$\psi\big(\mathbb{E}(X \,|\, \mathcal{A}_o)\big) \;\le\; \mathbb{E}(\psi(X) \,|\, \mathcal{A}_o) \quad \textit{almost surely},$$

and

$$\int \psi(\mathbb{E}(X \,|\, \mathcal{A}_o))\, dP \;\le\; \int \psi(X)\, dP.$$

Proof of Lemma 3.7. By the definition of conditional expectations and Lemma 2.20, it suffices to show that for any fixed $A_o \in \mathcal{A}_o$,

$$\int_{A_o} \psi\big(\mathbb{E}(X \,|\, \mathcal{A}_o)\big)\, dP \;\le\; \int_{A_o} \psi(X)\, dP.$$

Note that by convexity of ψ, for any fixed $x_o \in \mathbb{R}$ there exists a slope $b(x_o) \in \mathbb{R}$ such that $\psi(x) \geq \psi(x_o) + b(x_o)(x - x_o)$ for all $x \in \mathbb{R}$; see, for instance, Chapter 3 of Dümbgen (2021). Moreover, for arbitrary $x \in \mathbb{R}$,

$$\psi(x) \;=\; \sup_{x_o \in \mathbb{Q}} \big(\psi(x_o) + b(x_o)(x - x_o)\big).$$

Consequently,

$$\begin{aligned}
\int_{A_o} \psi\big(\mathbb{E}(X \,|\, \mathcal{A}_o)\big)\, dP \;&=\; \int_{A_o} \sup_{x_o \in \mathbb{Q}} \big(\psi(x_o) + b(x_o)(\mathbb{E}(X \,|\, \mathcal{A}_o) - x)\big)\, dP \\
&=\; \int_{A_o} \sup_{x_o \in \mathbb{Q}} \mathbb{E}\big(\psi(x_o) + b(x_o)(X - x_o) \,\big|\, \mathcal{A}_o\big)\, dP \\
&\leq\; \int_{A_o} \psi(X)\, dP,
\end{aligned}$$

where the second last step follows from Remark 3.4 (a) and countability of $\mathbb{Q}$, while the last step follows from Remark 3.4 (b) and countability of $\mathbb{Q}$. □

3.2 Conditional expectations as orthogonal projections

In this section we restrict our attention to the space $\mathcal{L}^2(P) \subset \mathcal{L}^1(P)$ of square-integrable random variables. If we identify random variables which are equal almost surely, we obtain the Hilbert space $L^2(P)$ with inner product

$$\langle X, Y \rangle \;:=\; \int XY\, dP$$

and norm

$$\|X\| \;:=\; \langle X, X \rangle^{1/2} \;=\; \Big(\int X^2\, dP\Big)^{1/2}.$$

One may view $\mathbb{E}(\cdot \,|\, \mathcal{A}_o)$ as a continuous linear mapping from $L^2(P)$ to its closed linear subspace $L^2(P|_{\mathcal{A}_o})$. Indeed, applying Lemma 3.7 with $\psi(x) := x^2$ leads to the inequality

$$\int \mathbb{E}(X \,|\, \mathcal{A}_o)^2\, dP \;\leq\; \int X^2\, dP.$$

Theorem 3.8. *The mapping $X \mapsto \mathbb{E}(X \,|\, \mathcal{A}_o)$ is the orthogonal linear projection of $L^2(P)$ onto $L^2(P|_{\mathcal{A}_o})$. In particular, for arbitrary random variables $X \in L^2(P)$ and $Y_o \in L^2(P|_{\mathcal{A}_o})$,*

$$\|X - Y_o\|^2 \;=\; \big\|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\big\|^2 + \big\|\,\mathbb{E}(X \,|\, \mathcal{A}_o) - Y_o\big\|^2 \;\geq\; \big\|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\big\|^2$$

with equality if and only if $Y_o = \mathbb{E}(X \,|\, \mathcal{A}_o)$. Moreover,

$$\|X\|^2 \;=\; \big\|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\big\|^2 + \big\|\,\mathbb{E}(X \,|\, \mathcal{A}_o)\big\|^2 \;\geq\; \big\|\mathbb{E}(X \,|\, \mathcal{A}_o)\big\|^2$$

with equality if and only if $X = \mathbb{E}(X \,|\, \mathcal{A}_o)$.

Proof of Theorem 3.8. It is well-known from general results about Hilbert spaces that $X_o := \mathbb{E}(X \,|\, \mathcal{A}_o)$ is the orthogonal projection of X onto $L^2(P|_{\mathcal{A}_o})$ if and only if

$$X - X_o \;\perp\; L^2(P|_{\mathcal{A}_o});$$

see, for instance, Chapter 2 of Dümbgen (2021). That means,

$$\langle X - X_o, Z_o \rangle \;=\; 0 \quad \text{for all } Z_o \in L^2(P|_{\mathcal{A}_o}). \tag{3.3}$$

Note first, that the left hand side of (3.3) equals

$$\int X Z_o \, dP - \int \mathbb{E}(X \,|\, \mathcal{A}_o) Z_o \, dP.$$

This is equal to 0 whenever Z_o is bounded, see Remark 3.4 (e). But for arbitrary $Z_o \in L^2(P|_{\mathcal{A}_o})$,

$$Z_n \;:=\; \mathrm{sign}(Z_o) \min(|Z_o|, n)$$

defines a sequence $(Z_n)_{n \geq 1}$ of bounded random variables $Z_n \in L^2(P|_{\mathcal{A}_o})$ such that

$$\|Z_o - Z_n\| \;\to\; 0 \quad \text{as } n \to \infty,$$

which may be verified by dominated convergence. Now (3.3) follows from the fact that for arbitrary $n \geq 1$,

$$\big|\langle X - X_o, Z_o \rangle\big| \;=\; \big|\langle X - X_o, Z_o - Z_n \rangle\big| \;\leq\; \|X - X_o\| \|Z_o - Z_n\|$$

by the Cauchy–Schwarz inequality.

Orthogonality of $X - X_o$ onto $L^2(P|_{\mathcal{A}_o})$ implies that for any $Y_o \in L^2(P|_{\mathcal{A}_o})$,

$$\begin{aligned} \|X - Y_o\|^2 \;&=\; \|X - X_o + X_o - Y_o\|^2 \\ &=\; \|X - X_o\|^2 + \|X_o - Y_o\|^2 + 2\langle X - X_o, X_o - Y_o \rangle \\ &=\; \|X - X_o\|^2 + \|X_o - Y_o\|^2, \end{aligned}$$

and in the special case $Y_o = 0$ we obtain

$$\|X\|^2 \;=\; \|X - X_o\|^2 + \|X_o\|^2.$$

□

Exercise 3.9. As in Exercise 3.5 let P be the exponential distribution on $\Omega = [0, \infty)$ with rate parameter $\lambda > 0$, let $\mathcal{A}_o$ be the smallest σ-field containing all intervals $[k, k+1)$, $k \in \mathbb{N}_0$, and let $X(\omega) := \omega$. Determine the value of

$$\|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\|.$$

Hint for checking your solution: $\|X - \mathbb{E}(X \,|\, \mathcal{A}_o)\|^2$ equals $\lambda^{-2} - e^{\lambda}(e^{\lambda} - 1)^{-2}$.

3.3 Conditional expectations given another random variable

Previously we referred to $\mathcal{A}_o$ as describing partial aspects of the random experiment $(\Omega, \mathcal{A}, P)$. Specifically, suppose that instead of the random outcome $\omega \in \Omega$ we only know $T(\omega)$ for some given measurable function $T : (\Omega, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$. This corresponds to the sub-σ-field

$$\sigma(T) \;:=\; \big\{T^{-1}(B) : B \in \mathcal{B}\big\}$$

of $\mathcal{A}$. It is the smallest σ-field $\mathcal{A}_o$ over Ω such that T is $\mathcal{A}_o$-$\mathcal{B}$-measurable. There is a simple result about $\sigma(T)$-measurable functions $X : \Omega \to \mathbb{R}$.

Lemma 3.10 (Lifting). *A mapping $X : \Omega \to \mathbb{R}$ is $\sigma(T)$-measurable if and only if*

$$X \;=\; V \circ T$$

for some $\mathcal{B}$-measurable function $V : \mathcal{T} \to \mathbb{R}$.

Here $V \circ T$ stands for the mapping $\Omega \ni \omega \mapsto V(T(\omega)) \in \mathbb{R}$. In connection with integrals we often write $V(T)$ instead of $V \circ T$.

Proof of Lemma 3.10. On the one hand, if $X = V \circ T$ with a $\mathcal{B}$-measurable function $V : \mathcal{T} \to \mathbb{R}$, then $V^{-1}(C) \in \mathcal{B}$ and $X^{-1}(C) = T^{-1}(V^{-1}(C)) \in \sigma(T)$ for any Borel set $C \subset \mathbb{R}$.

On the other hand, let X be $\sigma(T)$-measurable. For $n \in \mathbb{N}$ let $X_n := 2^{-n}\lfloor 2^n X \rfloor$. That means, for $z \in \mathbb{Z}$,

$$A_{n,z} \;:=\; \{X_n = 2^{-n}z\} \;=\; \{2^{-n}z \le X < 2^{-n}(z+1)\}.$$

By assumption on X, the latter event may be written as $\{T \in B_{n,z}\}$ for some set $B_{n,z} \in \mathcal{B}$. Since the sets $A_{n,z}$, $z \in \mathbb{Z}$, are pairwise disjoint, we may assume without loss of generality that the sets $B_{n,z}$, $z \in \mathbb{Z}$, are pairwise disjoint too. Just use some enumeration $z(1), z(2), z(3), \ldots$ of $\mathbb{Z}$, and then replace $B_{n,z(k)}$ with its subset $B_{n,z(k)} \setminus \bigcup_{\ell<k} B_{n,z(\ell)}$. Then,

$$X_n \;=\; V_n \circ T$$

with the $\mathcal{B}$-measurable function $V_n : \mathcal{T} \to \mathbb{R}$,

$$V_n(t) \;:=\; \begin{cases} 2^{-n}z & \text{if } t \in B_{n,z},\, z \in \mathbb{Z}, \\ 0 & \text{if } t \notin \bigcup_{z\in\mathbb{Z}} B_{n,z}. \end{cases}$$

Since $X_n \le X < X_n + 2^{-n}$,

$$X \;=\; \lim_{n\to\infty} V_n \circ T \;=\; \bar{V} \circ T$$

with the $\mathcal{B}$-measurable function $\bar{V} : \mathcal{T} \to \bar{\mathbb{R}}$,

$$\bar{V}(t) \;:=\; \limsup_{n\to\infty}\, V_n(t).$$

But $X = \bar{V} \circ T$ being real-valued implies that $T^{-1}\big(\bar{V}^{-1}(\{-\infty,\infty\})\big) = \emptyset$. Hence, $X = V \circ T$ with the $\mathcal{B}$-measurable function $V : \mathcal{T} \to \mathbb{R}$ given by, say,

$$V(t) \;:=\; \begin{cases} \bar{V}(t) & \text{if } \bar{V}(t) \in \mathbb{R}, \\ 0 & \text{else.} \end{cases}$$

□

In addition to Lemma 3.10 we need an elementary result about the distribution P^T of T, i.e. the probability measure

$$\mathcal{B} \ni B \;\mapsto\; P^T(B) := P(T \in B).$$

Exercise 3.11 (Change of variables). Show that for any real-valued random variable V on the probability space $(\mathcal{T}, \mathcal{B}, P^T)$ and any set $B \in \mathcal{B}$,

$$\int_{T^{-1}(B)} V(T)\, dP \;=\; \int_B V\, dP^T,$$

provided that one of these two integrals is well-defined in $\bar{\mathbb{R}}$. Hint: Use approximations of V as in Lemma 2.16.

Now we can prove a generalization of Theorem 3.1:

Theorem 3.12. *For any random variable $X \in \mathcal{L}^1(P)$ there exists a random variable $V \in \mathcal{L}^1(P^T)$ such that*

$$\int_{T^{-1}(B)} X\, dP \;=\; \int_B V\, dP^T \quad \text{for arbitrary } B \in \mathcal{B}. \tag{3.4}$$

This random variable V is almost everywhere unique in the following sense: If $\tilde{V}$ is another random variable satisfying (3.4)*, then $P^T(\tilde{V} \neq V) = 0$.*

Definition 3.13 (Conditional expectation, II). A random variable V as in Theorem 3.12 is called *(a version of the) conditional expectation of X, given the random variable T*. For any such random variable V we write $\mathbb{E}(X \,|\, T)$ instead of V and $\mathbb{E}(X \,|\, T = t)$ instead of $\mathbb{E}(V \,|\, T)(t)$ or $V(t)$.

Note that $\mathbb{E}(X \,|\, T)$ is a function on $\mathcal{T}$. But sometimes we and other people abuse notation slightly to write $\mathbb{E}(X \,|\, T)$ instead of $\mathbb{E}(X \,|\, \sigma(T))$, which is a function on Ω.

Example 3.14. Let $\Omega = [0, \infty)$, and let P be the exponential distribution with rate parameter $\lambda > 0$, i.e. with density $f_\lambda(\omega) := \lambda e^{-\lambda\omega}$ with respect to Lebesgue measure on Ω. Let $X(\omega) := \omega$, and let $\mathcal{A}_o$ be the smallest σ-field over Ω containing all intervals $[k, k+1)$, $k \in \mathbb{N}_0$. One can easily verify that $\mathcal{A}_o = \sigma(T)$ with

$$T : [0, \infty) \to \mathbb{N}_0, \quad T(\omega) := \lfloor \omega \rfloor.$$

And the result of Exercise 3.5 may be reformulated as

$$\begin{aligned} \mathbb{E}(X \,|\, \mathcal{A}_o) \;&=\; \lfloor X \rfloor + a(\lambda), \\ \mathbb{E}(X \,|\, T = t) \;&=\; t + a(\lambda) \quad \text{for } t \in \mathbb{N}_0, \end{aligned}$$

with $a(\lambda) = \lambda^{-1}(e^\lambda - 1 - \lambda)/(e^\lambda - 1)$.

Remark 3.15. Theorem 3.1 may be viewed as a special case of Theorem 3.12 if we define $(\mathcal{T}, \mathcal{B}) := (\Omega, \mathcal{A}_o)$ and $T(\omega) := \omega$.

Proof of Theorem 3.12. We may apply Theorem 3.1 to the sub-σ-field $\mathcal{A}_o = \sigma(T)$. This yields a random variable $X_o \in \mathcal{L}^1(P|_{\sigma(T)})$ such that

$$\int_{T^{-1}(B)} X\, dP \;=\; \int_{T^{-1}(B)} X_o\, dP \quad \text{for all } B \in \mathcal{B}.$$

Lemma 3.10 implies that $X_o = V \circ T$ for some $\mathcal{B}$-measurable function $V : \mathcal{T} \to \mathbb{R}$. Now it follows from Exercise 3.11 that

$$\int_{T^{-1}(B)} V(T)\, dP \;=\; \int_B V\, dP^T \quad \text{for all } B \in \mathcal{B}.$$

This proves existence of a random variable V with the desired property (3.4). Essential uniqueness of V is a consequence of Lemma 2.20. □

Remark 3.16 (Properties of conditional expectations). The properties listed in Remark 3.4 carry over to the present setting with only few modifications. In what follows, let $X, Y \in \mathcal{L}^1(P)$.

(a) For real numbers a, b,

$$\mathbb{E}(aX + bY \,|\, T) \;=\; a\,\mathbb{E}(X \,|\, T) + b\,\mathbb{E}(Y \,|\, T) \quad P^T\text{-almost surely.}$$

(b) If $X \le Y$ almost surely, then

$$\mathbb{E}(X \,|\, T) \;\le\; \mathbb{E}(Y \,|\, T) \quad P^T\text{-almost surely.}$$

(c) The mapping $X \mapsto \mathbb{E}(X \,|\, T)$ is a weak contraction in the sense that

$$\big|\mathbb{E}(X \,|\, T)\big| \;\le\; \mathbb{E}(|X| \,|\, T) \quad P^T\text{-almost surely.}$$

(d) It follows from properties (a) and (c) that

$$\int \big|\mathbb{E}(X \,|\, T) - \mathbb{E}(Y \,|\, T)\big|\, dP^T \;\le\; \int |X - Y|\, dP.$$

(e) For any bounded, $\mathcal{B}$-measurable function $h : \mathcal{T} \to \mathbb{R}$,

$$\int X h(T)\, dP \;=\; \int \mathbb{E}(X \,|\, T) h\, dP^T. \tag{3.5}$$

This follows again from an approximation argument. The definition of $\mathbb{E}(X \,|\, T)$ implies that equation (3.5) is true for indicator functions $h = 1_B$, $B \in \mathcal{B}$, because $1_{T^{-1}(B)} = 1_B(T)$. The remaining arguments are analogous to the arguments for Remark 3.4 (e).

An extension of the lifting lemma. Although it is not needed in this lecture, it is worthwhile to mention that Lemma 3.10 can be generalised as follows:

Lemma 3.17 (Lifting). *Let $(\mathcal{X}, d)$ be a complete, separable metric space, equipped with its Borel-σ-field. A mapping $X : \Omega \to \mathcal{X}$ is $\sigma(T)$-measurable if and only if*

$$X \;=\; V \circ T$$

for some $\mathcal{B}$-measurable function $V : \mathcal{T} \to \mathcal{X}$.

The proof is split into the following two exercises.

Exercise 3.18. Let $(\Omega, \mathcal{A})$ be a measurable and $(\mathcal{X}, d)$ a metric space. Further let $(V_n)_{n\geq 1}$ be a sequence of measurable functions $V_n : \Omega \to \mathcal{X}$, where $\mathcal{X}$ is equipped with its Borel-σ-field with respect to d.

(a) Suppose that $(V_n)_{n\geq 1}$ converges pointwise to a function $V : \Omega \to \mathcal{X}$. Show that V is measurable.

(b) Suppose that $(\mathcal{X}, d)$ is separable and complete. Without any further assumptions on $(V_n)_{n\geq 1}$, show that the set C of all $\omega \in \Omega$ such that $(V_n(\omega))_{n\geq 1}$ converges is a measurable subset of Ω.

Exercise 3.19. Imitate the proof of Lemma 3.10 to construct a sequence $(V_n)_{n\geq 1}$ of measurable functions $V_n : \mathcal{T} \to \mathcal{X}$ converging pointwise to a function V such that $X = V \circ T$. Conclude from Exercise 3.18 that V is measurable.

Chapter 4

Stochastic Kernels

In this section we collect some useful results about finite measures on product spaces. Throughout let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, and we consider the Cartesian product $\Omega := \mathcal{X} \times \mathcal{Y}$ equipped with the product σ-field

$$\mathcal{C} \ := \ \mathcal{A} \otimes \mathcal{B},$$

i.e. the smallest σ-field containing all sets $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. An important property of $\mathcal{C}$ is that for any $C \in \mathcal{C}$,

$$\{x \in \mathcal{X} : (x, y_o) \in C\} \ \in \ \mathcal{A} \quad \text{for all } y_o \in \mathcal{Y}, \tag{4.1}$$

$$\{y \in \mathcal{Y} : (x_o, y) \in C\} \ \in \ \mathcal{B} \quad \text{for all } x_o \in \mathcal{X}. \tag{4.2}$$

The reason is that the family of all sets $C \subset \Omega$ satisfying (4.1) and (4.2) is easily verified to be a σ-field over Ω containing the generator

$$\mathcal{A} \boxtimes \mathcal{B} \ := \ \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

of $\mathcal{C}$.

4.1 Stochastic Kernels and Fubini's Theorem

Definition 4.1 (Stochastic kernel). A stochastic kernel from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ is a mapping

$$K : \mathcal{X} \times \mathcal{B} \to [0, 1]$$

such that

- for any fixed $x \in \mathcal{X}$, the mapping $K(x, \cdot)$ defines a probability measure on $(\mathcal{Y}, \mathcal{B})$,
- for any fixed $B \in \mathcal{B}$, the mapping $K(\cdot, B)$ is $\mathcal{A}$-measurable on $\mathcal{X}$.

Remark 4.2 (Randomized mappings). One may interpret a stochastic kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ as a "randomized measurable mapping" from $\mathcal{X}$ to $\mathcal{Y}$. Instead of mapping a point $x \in \mathcal{X}$ to a unique point $k(x) \in \mathcal{Y}$, we choose a random point in $\mathcal{Y}$ with distribution $K(x, \cdot)$. Indeed, any $\mathcal{A}$-$\mathcal{B}$-measurable mapping $k : \mathcal{X} \to \mathcal{Y}$ corresponds to the (degenerate) stochastic kernel K given by

$$K(x, \cdot) \ := \ \delta_{k(x)}, \quad K(x, B) \ = \ 1_{k^{-1}(B)}(x).$$

Theorem 4.3 (Product of a finite measure and a stochastic kernel). *Let P be a finite measure on $(\mathcal{X}, \mathcal{A})$, and let K be a stochastic kernel from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$. Then for arbitrary $C \in \mathcal{C}$, the number*

$$P \otimes K(C) \;:=\; \int_{\mathcal{X}} K(x, C_x)\, P(dx) \quad \text{with } C_x := \{y \in \mathcal{Y} : (x, y) \in C\}$$

is well-defined, and $P \otimes K$ is a finite measure on $(\Omega, \mathcal{C})$ with $P \otimes K(\Omega) = P(\mathcal{X})$.

Remark 4.4 (Product measures). Let Q be a probability measure on $(\mathcal{Y}, \mathcal{B})$. If we set $K(x, B) := Q(B)$ for arbitrary $x \in \mathcal{X}$ and $B \in \mathcal{B}$, then $P \otimes K$ is the usual product measure $P \otimes Q$.

In the subsequent proofs, we use repeatedly the notion of a Dynkin system, see Section A.1 for the basic concepts.

Proof of Theorem 4.3. One can easily verify that the family $\mathcal{D}$ of sets $C \in \mathcal{C}$ such that $x \mapsto K(x, C_x)$ is $\mathcal{A}$-measurable is a Dynkin system containing the generator $\mathcal{A} \boxtimes \mathcal{B}$ of $\mathcal{C}$. Since $\mathcal{A} \boxtimes \mathcal{B}$ is closed under intersections, the smallest Dynkin system containing $\mathcal{A} \boxtimes \mathcal{B}$ coincides with the product σ-field $\mathcal{A} \otimes \mathcal{B} = \mathcal{C}$. Hence $P \otimes K(C)$ is well-defined for any $C \in \mathcal{C}$.

It follows from the properties of K that $P \otimes K$ defines a content, i.e. a finitely additive function on $\mathcal{A} \otimes \mathcal{B}$ with $P \otimes K(\emptyset) = 0$ and $P \otimes K(\Omega) = P(\mathcal{X})$. It is a measure, because for arbitrary sets $C^{(1)} \subset C^{(2)} \subset C^{(3)} \subset \cdots$ in $\mathcal{C}$ and $C := \bigcup_{n \geq 1} C^{(n)}$,

$$\begin{aligned} \lim_{n\to\infty} P \otimes K(C^{(n)}) \;&=\; \lim_{n\to\infty} \int K(x, C_x^{(n)})\, P(dx) \\ &=\; \int \lim_{n\to\infty} K(x, C_x^{(n)})\, P(dx) \\ &=\; \int K(x, C_x)\, P(dx) \;=\; P \otimes K(C) \end{aligned}$$

by monotone convergence. □

If a distribution on a product space is a product of a marginal distribution and a stochastic kernel, there is a generalization of Fubini's theorem for integrals with respect to product measures.

Theorem 4.5 (Fubini's theorem). *Let $\mathbb{P} = P \otimes K$ with a finite measure P on $(\mathcal{X}, \mathcal{A})$ and a stochastic kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$.*

(i) *For any $\mathcal{C}$-measurable function $f : \Omega \to [0, \infty]$,*

$$h(x) \;:=\; \int_{\mathcal{Y}} f(x, y)\, K(x, dy)$$

defines an $\mathcal{A}$-measurable mapping $h : \mathcal{X} \to [0, \infty]$, and

$$\int_{\Omega} f\, d\,\mathbb{P} \;=\; \int_{\mathcal{X}} h\, dP.$$

(ii) *If $f \in \mathcal{L}^1(\mathbb{P})$, then there exists a set $A(f) \in \mathcal{A}$ with $P(A(f)) = 0$ such that for all $x \in \mathcal{X} \setminus A(f)$,*

$$h(x) \;:=\; \int f(x, y)\, K(x, dy)$$

is well-defined in $\mathbb{R}$*, and*

$$\int_\Omega f \, d\,\mathbb{P} \;=\; \int_{\mathcal{X}} h \, dP.$$

Proof of Theorem 4.5. Part (ii) is an immediate consequence of part (i). Hence we prove only part (i). For a measurable function $f : \Omega \to [0, \infty]$ and $n \in \mathbb{N}$ let

$$f_n \;:=\; 2^{-n} \sum_{k=1}^{n2^n} 1_{[f \geq k2^{-n}]}.$$

This defines a sequence of measurable functions $f_n \geq 0$ such that $f_n \uparrow f$ as $n \to \infty$ pointwise. Indeed, for any integer $\ell \geq 0$,

$$f_n \;=\; \min(\ell, n2^n)2^{-n} \quad \text{on} \quad \{\ell 2^{-n} \leq f < (\ell+1)2^{-n}\}.$$

For any fixed n,

$$\begin{aligned}
\int_\Omega f_n \, d\,\mathbb{P} \;&=\; 2^{-n} \sum_{k=1}^{n2^n} \mathbb{P}(f \geq k2^{-n}) \\
&=\; 2^{-n} \sum_{k=1}^{n2^n} \int_{\mathcal{X}} K(x, \{f \geq k2^{-n}\}_x) \, P(dx) \\
&=\; \int_{\mathcal{X}} h_n \, dP
\end{aligned}$$

with

$$h_n(x) \;:=\; 2^{-n} \sum_{k=1}^{n2^n} K(x, \{f \geq k2^{-n}\}_x) \;=\; \int_{\mathcal{Y}} f_n(x, y) \, K(x, dy).$$

This is obviously a measurable function on $(\mathcal{X}, \mathcal{A})$. By monotone convergence, for any $x \in \mathcal{X}$,

$$h_n(x) \;\uparrow\; h(x) = \int_{\mathcal{Y}} f(x, y) \, K(x, dy) \quad \text{as } n \to \infty,$$

whence h is measurable as well. Another application of monotone convergence yields that

$$\int_{\mathcal{X}} h_n \, dP \;\uparrow\; \int_{\mathcal{X}} h \, dP \quad \text{as } n \to \infty.$$

On the other hand, by monotone convergence,

$$\int_\Omega f_n \, d\,\mathbb{P} \;\uparrow\; \int_\Omega f \, d\,\mathbb{P} \quad \text{as } n \to \infty,$$

so the asserted equation is true. □

Remark 4.6 (Fubini's theorem for products of σ-finite measures). The present results also imply a more traditional version of Fubini's theorem. Let P and Q be σ-finite measures on $(\mathcal{X}, \mathcal{A})$ and on $(\mathcal{Y}, \mathcal{B})$, respectively. For $C \in \mathcal{C}$, the number

$$\mathbb{P}(C) \;:=\; \int_{\mathcal{X}} Q(C_x) \, P(dx) \quad \text{with } C_x := \{y \in \mathcal{Y} : (x, y) \in C\} \tag{4.3}$$

is well-defined in $[0,\infty]$ and defines a σ-finite measure on $(\Omega,\mathcal{C})$. Moreover, for arbitrary $\mathcal{C}$-measurable functions $f:\Omega\to[0,\infty]$,

$$h(x) \;:=\; \int_{\mathcal{Y}} f(x,y)\,Q(dy)$$

defines a $\mathcal{A}$-measurable function $h:\mathcal{X}\to[0,\infty]$, and

$$\int_\Omega f\,d\,\mathbb{P} \;=\; \int_{\mathcal{X}} h\,dP.$$

To verify this, we apply Lemma 2.26 to find probability measures P_o on $(\mathcal{X},\mathcal{A})$ and Q_o on $(\mathcal{Y},\mathcal{B})$ such that densities $p=dP/dP_o$ and $q=dQ/dQ_o$ exist. Then the product measure $\mathbb{P}_o := P_o\otimes Q_o$ is well-defined, and

$$\mathbb{P}(C) \;:=\; \int_C p(x)q(y)\;\mathbb{P}_o(d(x,y))$$

defines a σ-finite measure on $(\Omega,\mathcal{C})$. This definition coincides with (4.3), because Theorem 4.5 implies that

$$\begin{aligned}\mathbb{P}(C) \;&=\; \int_{\mathcal{X}}\int_{\mathcal{Y}} 1_C(x,y)p(x)q(y)\,Q_o(dy)\,P_o(dx)\\ &=\; \int_{\mathcal{X}}\int_{C_x} q(y)\,Q_o(dy)\,p(x)\,P_o(dx)\\ &=\; \int_{\mathcal{X}} Q(C_x)\,p(x)\,P_o(dx)\\ &=\; \int_{\mathcal{X}} Q(C_x)\,P(dx).\end{aligned}$$

More generally, for $\mathcal{C}$-measurable functions $f:\Omega\to[0,\infty]$,

$$\begin{aligned}\int f\,d\,\mathbb{P} \;&=\; \int_\Omega f(x,y)p(x)q(y)\;\mathbb{P}_o(d(x,y))\\ &=\; \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)p(x)q(y)\,Q_o(dy)\,P_o(dx)\\ &=\; \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)\,Q(dy)\,p(x)\,P_o(dx)\\ &=\; \int_{\mathcal{X}} h(x)\,P(dx)\end{aligned}$$

with $h(x):=\int_{\mathcal{Y}} f(x,y)\,Q(dy)$.

The next result concerns uniqueness of stochastic kernels.

Lemma 4.7 (Uniqueness in product measures). *Let $P,\tilde P$ be finite measures on $(\mathcal{X},\mathcal{A})$, and let $K,\tilde K$ be stochastic kernels from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ such that*

$$P\otimes K \;=\; \tilde P\otimes\tilde K.$$

Then $P\equiv\tilde P$, and for any $B\in\mathcal{B}$, $K(\cdot,B)=\tilde K(\cdot,B)$ P-almost everywhere. If $\mathcal{B}$ has a countable generating family $\mathcal{E}$, then even

$$K(x,\cdot)\;\equiv\;\tilde K(x,\cdot)\quad \text{for } P\text{-almost all } x\in\mathcal{X}.$$

Proof of Lemma 4.7. The construction of the product measures implies that for any $A \in \mathcal{A}$,

$$P(A) \;=\; P \otimes K(A \times \mathcal{Y}) \;=\; \tilde{P} \otimes \tilde{K}(A \times \mathcal{Y}) \;=\; \tilde{P}(A),$$

i.e. $P \equiv \tilde{P}$. For any fixed $B \in \mathcal{B}$, the functions $K(\cdot, B), \tilde{K}(\cdot, B) \in \mathcal{L}^1(P)$ satisfy

$$\int_A K(x, B)\, P(dx) \;=\; P \otimes K(A \times B) = P \otimes \tilde{K}(A \times B) \;=\; \int_A \tilde{K}(x, B)\, P(dx)$$

for arbitrary $A \in \mathcal{A}$. Hence $K(\cdot, B) = \tilde{K}(\cdot, B)$ P-almost everywhere by Lemma 2.20.

Finally, suppose that $\mathcal{E}$ is a countable family of subsets of $\mathcal{Y}$ generating $\mathcal{B}$. The family $\mathcal{E}'$ of intersections of finitely many sets in $\mathcal{E}$ is countable as well, see Exercise 4.8. Then the set

$$A_* \;:=\; \big\{x \in \mathcal{X} : K(x, B) \neq \tilde{K}(x, B) \text{ for some } B \in \mathcal{E}'\big\}$$

satisfies $P(A_*) = 0$. But for any fixed $x \in \mathcal{X} \setminus A_*$,

$$\mathcal{D}(x) \;:=\; \big\{B \in \mathcal{B} : K(x, B) = \tilde{K}(x, B)\big\}$$

defines a Dynkin system with $\mathcal{E}' \subset \mathcal{D}(x) \subset \mathcal{B}$. Since $\mathcal{E}'$ is closed under (finite) intersections, the smallest Dynkin system containing $\mathcal{E}'$ is a σ-field, so $\mathcal{D}(x) = \mathcal{B}$ and $K(x, \cdot) \equiv \tilde{K}(x, \cdot)$. □

Exercise 4.8 (Countable set families). Let $\mathcal{E}$ be a countable family of subsets of $\mathcal{Y}$.

(a) Show that the family $\mathcal{E}'$ of all intersections of finitely many sets in $\mathcal{E}$ is countable as well, and that $\mathcal{E}'$ is closed under (finite) intersections.

(b) Let $\mathcal{B}_o$ be the family of all sets of the form

$$\bigcup_{i=1}^{m} \bigcap_{j=1}^{n} E_{ij}$$

with $m, n \in \mathbb{N}$ and $E_{ij} \in \mathcal{E} \cup \{\mathcal{Y} \setminus E : E \in \mathcal{E}\} \cup \{\emptyset, \mathcal{Y}\}$. Show that $\mathcal{B}_o$ is countable, and that it is the smallest field over $\mathcal{Y}$ containing $\mathcal{E}$.

4.2 Decomposing Measures on Product Spaces

An interesting question is whether any finite measure $\mathbb{P}$ on $(\Omega, \mathcal{C})$ may be decomposed into the product $P \otimes K$ of a finite measure and a stochastic kernel. Let us start with two examples for such a decomposition.

Example 4.9. Suppose that $\mathcal{X}$ is a countable set and $\mathcal{A} = \mathcal{P}(\mathcal{X})$. Then any finite measure $\mathbb{P}$ on Ω may be represented as $P \otimes K$, where

$$P(\{x\}) \;:=\; \mathbb{P}(\{x\} \times \mathcal{Y})$$

and

$$K(x, B) \;:=\; \begin{cases} P(\{x\})^{-1}\, \mathbb{P}(\{x\} \times B) & \text{if } P(\{x\}) > 0, \\ \mathbb{P}(\mathcal{X})^{-1}\, \mathbb{P}(\mathcal{X} \times B) & \text{if } P(\{x\}) = 0. \end{cases}$$

Example 4.10. Let L and M be σ-finite measures on $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, respectively. Suppose that $\mathbb{P}$ is a finite measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ with density $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ with respect to the product measure $L \otimes M$. By Fubini's theorem (Remark 4.6),

$$\mathbb{P}(C) \;=\; \int_{\mathcal{X}} \int_{\mathcal{Y}} 1_C(x, y) f(x, y)\, M(dy)\, L(dx) \;=\; \int_{\mathcal{X}} \int_{C_x} f(x, y)\, M(dy)\, L(dx)$$

for arbitrary $C \in \mathcal{C} = \mathcal{A} \otimes \mathcal{B}$. If we define

$$P(A) \;:=\; \mathbb{P}(A \times \mathcal{Y})$$

for $A \in \mathcal{A}$, then P is a finite measure on $(\mathcal{X}, \mathcal{A})$ with $P(\mathcal{X}) = \mathbb{P}(\Omega)$ and

$$P(A) \;=\; \int_A f_1(x)\, P(dx) \quad \text{with} \quad f_1(x) \;:=\; \int_{\mathcal{Y}} f(x, y)\, M(dy).$$

Thus $f_1 = dP/dL$. Note also that $P(\{f_1 = 0\}) = 0 = P(\{f_1 = \infty\})$.

Analogously, $Q(B) := \mathbb{P}(\Omega)^{-1}\, \mathbb{P}(\mathcal{X} \times B)$ defines a probability measure on $(\mathcal{Y}, \mathcal{B})$ with density $f_2 = dQ/dM$ given by

$$f_2(y) \;:=\; \mathbb{P}(\Omega)^{-1} \int_{\mathcal{X}} f(x, y)\, L(dx).$$

For $C \in \mathcal{C}$, we may rewrite $\mathbb{P}(C)$ as

$$\mathbb{P}(C) \;=\; \int_{\mathcal{X}} f_1(x) K(x, C_x)\, L(dx) \;=\; \int_{\mathcal{X}} K(x, C_x)\, P(dx)$$

with

$$K(x, B) \;:=\; \int_B f_{2|1}(y \,|\, x)\, M(dy)$$

and

$$f_{2|1}(y \,|\, x) \;:=\; \begin{cases} f_1(x)^{-1} f(x, y) & \text{if } 0 < f_1(x) < \infty, \\ f_2(y) & \text{else.} \end{cases}$$

This defines a stochastic kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ such that $\mathbb{P} = P \otimes K$.

Exercise 4.11 (Decomposition of a probability distribution). Let $\mathbb{P}$ be the uniform distribution on the unit sphere in $\mathbb{R} \times \mathbb{R}$. That is, it describes the distribution of $\big(\cos(U), \sin(U)\big)$, where $U \sim \mathrm{Unif}[0, 2\pi]$. Determine a decomposition $\mathbb{P} = P \otimes K$ with a probability measure P on $\mathbb{R}$ given by some density, and a stochastic kernel K from $\mathbb{R}$ to $\mathbb{R}$.

Our final result in this section will show that for "nice" measurable spaces $(\mathcal{Y}, \mathcal{B})$, any finite measure on $\mathcal{C}$ may be decomposed into the product of a finite measure on $(\mathcal{X}, \mathcal{A})$ and a stochastic kernel from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$.

Theorem 4.12 (Existence of stochastic kernels). *Let $(\mathcal{Y}, d)$ be a separable and complete metric space*[1]*, and let $\mathcal{B} = \mathrm{Borel}(\mathcal{Y}, d)$. Then for any finite measure $\mathbb{P}$ on $(\Omega, \mathcal{C})$ there exists a finite measure P on $(\mathcal{X}, \mathcal{A})$ and a stochastic kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ such that*

$$\mathbb{P} \;=\; P \otimes K.$$

[1]The corresponding topological space is called a *Polish space*.

For the proof of this theorem, we need some basic facts about finite measures on metric spaces.

Lemma 4.13. *Let $(\mathcal{Y}, d)$ be a metric space, and let Q be a finite measure on $\mathcal{B} = \mathrm{Borel}(\mathcal{Y}, d)$.*

(i) *Any set $B \in \mathcal{B}$ may be approximated in the following sense: For each $\epsilon > 0$ there exist a closed set $A \subset \mathcal{Y}$ and an open set $U \subset \mathcal{Y}$ such that $A \subset B \subset U$ and $Q(U \setminus A) \leq \epsilon$.*

(ii) *If $(\mathcal{Y}, d)$ is separable and complete, the closed sets A in part (i) may even be chosen to be compact.*

Proof of Lemma 4.13. The proof of part (i) is left to the reader as an exercise. One can proceed as follows: Let $\mathcal{B}'$ be the set of all $B \in \mathcal{B}$ which may by approximated by closed and open sets as stated. One can show that $\mathcal{B}'$ is a field over $\mathcal{Y}$. Then one can show that it is even a σ-field over $\mathcal{Y}$ by showing that $\bigcup_{n\geq 1} B_n \in \mathcal{B}'$ for arbitrary sets $B_1 \subset B_2 \subset B_3 \subset \cdots$ in $\mathcal{B}'$. Any closed set $A \subset \mathcal{Y}$ belongs to $\mathcal{B}$. To this end consider the neighborhoods $U_\epsilon(A) := \{y \in \mathcal{Y} : d(y, A) < \epsilon\}$ with $d(y, A) := \inf\{d(y, z) : z \in A\}$. Consequently, $\mathcal{B}'$ is a σ-field containing all closed sets and contained in $\mathcal{B}$. But this implies that $\mathcal{B}' = \mathcal{B}$.

As to part (ii), let $\{y_1, y_2, y_3, \ldots\}$ be a dense subset of $\mathcal{Y}$. For any fixed $\epsilon > 0$ and arbitrary $k \in \mathbb{N}$ let $N(\epsilon, k) \in \mathbb{N}$ such that

$$Q\Big(\mathcal{Y} \setminus \bigcup_{n=1}^{N(\epsilon,k)} B(y_n, 1/k)\Big) \;\leq\; 2^{-k}\epsilon,$$

where $B(y, \delta)$ is the closed ball with center y and radius $\delta \geq 0$. Then

$$K_\epsilon \;:=\; \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{N(k,\epsilon)} B(y_n, 1/k)$$

is a closed subset of $\mathcal{Y}$ such that $Q(\mathcal{Y} \setminus K_\epsilon) \leq \epsilon$. Indeed, the set K_ϵ is even compact. To show this, let $(x_m)_{m\geq 1}$ be an arbitrary sequence in K_ϵ. Since K_ϵ is closed and $(\mathcal{Y}, d)$ is complete, it suffices to show that $(x_m)_{m\geq 1}$ has a Cauchy subsequence. There exists an index $n(1) \in \{1, \ldots, N(\epsilon, k)\}$ such that $M_1 := \{m \geq 1 : x_m \in B(y_{n(1)}, 1)\}$ is infinite. Suppose that for some $k \geq 1$ we have chosen an infinit set $M_k \subset \mathbb{N}$ and an index $n(k) \in \{1, \ldots, N(\epsilon, k)$ such that $x_m \in B(y_{n(k)}, 1/k)$ for $m \in M_k$. Then for a suitable index $n(k+1) \in \{1, \ldots, N(\epsilon, k+1)\}$, the set

$$M_{k+1} \;:=\; \big\{m \in M_k : x_m \in B(y_{n(k+1)}, 1/(k+1))\big\}$$

is infinite too. This leads to infinite subsets $M_1 \supset M_2 \supset M_3 \supset \cdots$ of $\mathbb{N}$ such that for each $k \geq 1$,

$$x_m \;\in\; B(y_{n(k)}, 1/k) \quad \text{for all } m \in M_k.$$

Now let $m(k)$ be the k-th smallest element of M_k. Then $m(1) < m(2) < m(3) < \cdots$, and for integers $1 \leq k \leq \ell$, $m(k), m(\ell) \in M_k$, whence

$$d(x_k, x_\ell) \;\leq\; d(x_k, y_{n(k)}) + d(x_\ell, y_{n(k)}) \;\leq\; 2/k.$$

This implies that $(x_{m(k)})_{k\geq 1}$ is a Cauchy subsequence of $(x_m)_{m\geq 1}$, whence K_ϵ is compact.

For any set $B \in \mathcal{B}$ and $\epsilon > 0$ there exist a closed set A and an open set U such that $A \subset B \subset U$ and $Q(U \setminus A) \leq \epsilon/2$. But then, $\tilde{A} := A \cap K_{\epsilon/2}$ is a compact subset of B such that $Q(U \setminus \tilde{A}) \leq \epsilon$. $\square$

Proof of Theorem 4.12. Without loss of generality let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{C})$. For $A \in \mathcal{A}$ and $B \in \mathcal{B}$ let $P(A) := \mathbb{P}(A \times \mathcal{Y})$ and $Q(B) := \mathbb{P}(\mathcal{X} \times B)$. This defines probability measures P and Q on $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, respectively.

Since $(\mathcal{Y}, d)$ is separable, there exists a countable dense subset $\mathcal{Y}_o$ of $\mathcal{Y}$. Let $\mathcal{E}$ be the family of all open balls with center in $\mathcal{Y}_o$ and rational radius. This family $\mathcal{E}$ is countable, and any open subset of $\mathcal{Y}$ is the union of balls in $\mathcal{E}$. As shown in Exercise 4.8, the smallest field $\mathcal{B}_o$ over $\mathcal{Y}$ containing $\mathcal{E}$ is countable, too. Hence, $\mathcal{B} = \sigma(\mathcal{E}) = \sigma(\mathcal{B}_o)$. Since $(\mathcal{Y}, d)$ is complete, for each $B_o \in \mathcal{B}_o$ there exist compact sets

$$D_1(B_o) \subset D_2(B_o) \subset D_3(B_o) \subset \cdots \subset B_o$$

such that $\lim_{k\to\infty} Q(D_k(B_o)) = Q(B_o)$, see Lemma 4.13. Finally, we consider the smallest field $\bar{\mathcal{B}}_o$ over $\mathcal{Y}$ containing $\mathcal{B}_o$ and $\{D_k(B_o) : B_o \in \mathcal{B}_o, k \in \mathbb{N}\}$, which is also countable.

For each $B \in \bar{\mathcal{B}}_o$ let $K(\cdot, B)$ be an explicit version of $\mathbb{E}(1_B(Y) \,|\, X)$, where $X(x, y) := x$ and $Y(x, y) := y$ for $(x, y) \in \Omega$. That means, for arbitrary $A \in \mathcal{A}$,

$$\mathbb{P}(A \times B) = \mathbb{E}\big(1_A(X)1_B(Y)\big) \;=\; \mathbb{E}\big(1_A(X)K(X, B)\big) = \int_A K(x, B)\, P(dx).$$

The properties of conditional expectations imply that P-almost everywhere,

$$\begin{aligned} K(\cdot, \emptyset) \;&=\; 0, \\ K(\cdot, \mathcal{Y}) \;&=\; 1, \\ K(\cdot, B \cup B') \;&=\; K(\cdot, B) + K(\cdot, B') \quad \text{for disjoint sets } B, B' \in \bar{\mathcal{B}}_o. \end{aligned}$$

The collection of all these equations is countable, so there exists a set $A_1 \in \mathcal{A}$ with $P(A_1) = 0$ such that for all $x \in \mathcal{X} \setminus A_1$, the mapping $K(x, \cdot)$ defines a probability content on $\bar{\mathcal{B}}_o$, i.e. $0 \le K(x, \cdot) \le 1$ with $K(x, \emptyset) = 0$, $K(x, \mathcal{Y}) = 1$ and $K(x, B \cup B') = K(x, B) + K(x, B')$ for disjoint sets $B, B' \in \bar{\mathcal{B}}_o$.

For any set B_o in the smaller set algebra $\mathcal{B}_o$, we know that

$$\lim_{k\to\infty} Q(D_k(B_o)) \;=\; Q(B_o).$$

But $Q(D_k(B_o)) \;=\; \int K(x, D_k(B_o))\, P(dx)$, $Q(B_o) \;=\; \int K(x, B_o)\, P(dx)$, and for any $x \in \mathcal{X} \setminus A_1$, the sequence $\big(K(x, D_k(B_o))\big)_{k\ge1}$ is increasing with a limit $h(x, B_o) \le K(x, B_o)$. By monotone convergence, $\int h(x, B_o)\, P(dx) = Q(B_o) = \int K(x, B_o)\, P(dx)$, whence $K(\cdot, B_o) = h(\cdot, B_o)$ for P-almost all $x \in \mathcal{X}$. Since $\mathcal{B}_o$ is countable, there exists a set $A_2 \in \mathcal{A}$ with $A_2 \subset \mathcal{X} \setminus A_1$ and $P(A_2) = 0$ such that for any $x \in \mathcal{X} \setminus (A_1 \cup A_2)$,

$$\lim_{k\to\infty} K(x, D_k(B_o)) \;=\; K(x, B_o) \quad \text{for each } B_o \in \mathcal{B}_o.$$

These properties imply that for each $x \in \mathcal{X} \setminus (A_1 \cup A_2)$, the mapping $K(x, \cdot)$ is a probability *measure* on the smaller set algebra $\mathcal{B}_o$. To prove this it suffices to show that for any sequence $(B_n)_n$ of sets $B_1 \supset B_2 \supset B_3 \supset \cdots$ in $\mathcal{B}_o$ with $\bigcap_{n\ge1} B_n = \emptyset$,

$$\lim_{n\to\infty} K(x, B_n) \;=\; 0.$$

Indeed, for any fixed $\epsilon > 0$ there exist indices $k(n) \geq 1$ such that

$$K(x, B_n \setminus D_n) \;=\; K(x, B_n) - K(x, D_n) \;\leq\; 2^{-n}\epsilon \quad \text{with } D_n := D_{k(n)}(B_n)).$$

Note that D_n is a subset of B_n, so $\bigcap_{n\geq 1} D_n = \emptyset$. Since all sets D_n are compact, there exists an integer $n_o \geq 1$ such that $\bigcap_{m=1}^{n_o} D_m = \emptyset$. Consequently, for $n \geq n_o$,

$$\begin{aligned} K(x, B_n) \;\leq\; K(x, B_{n_o}) \;&=\; K(x, B_{n_o} \setminus \bigcap_{m=1}^{n_o} D_m) \\ &=\; K\Big(x, \bigcup_{m=1}^{n_o} B_{n_o} \setminus D_m\Big) \\ &\leq\; K\Big(x, \bigcup_{m=1}^{n_o} B_m \setminus D_m\Big) \\ &\leq\; \sum_{m=1}^{n_o} K(x, B_m \setminus D_m) \;\leq\; \sum_{m=1}^{n_o} 2^{-m}\epsilon \;<\; \epsilon. \end{aligned}$$

This shows that $K(x, B_n) \to 0$ as $n \to \infty$.

For each $x \in \mathcal{X} \setminus (A_1 \cup A_2)$, there exists a unique extension of $K(x, \cdot)$ to a probability measure $K^*(x, \cdot)$ on $(\mathcal{Y}, \mathcal{B})$. If we set

$$K^*(x, B) \;:=\; Q(B) \quad \text{for } x \in A_1 \cup A_2 \text{ and } B \in \mathcal{B},$$

then $K^* : \mathcal{X} \times \mathcal{B}$ has the following properties:

- For each $x \in \mathcal{X}$, $K^*(x, \cdot)$ is a probability measure on $(\mathcal{Y}, \mathcal{B})$;
- for each $B_o \in \mathcal{B}_o$, $K^*(\cdot, B_o)$ is $\mathcal{A}$-measurable and a version of $\mathbb{E}(1_{B_o}(Y) \,|\, X)$.

But the set of all $B \in \mathcal{B}$ such that $K^*(\cdot, B)$ is $\mathcal{A}$-measurable is easily seen to be a Dynkin system. Since it contains the set algebra $\mathcal{B}_o$, it coincides with $\mathcal{B}$. Hence K^* is indeed a stochastic kernel, and the distribution $P \otimes K^*$ coincides with $\mathbb{P}$ on the family $\{A \times B_o : A \in \mathcal{A}, B_o \in \mathcal{B}_o\}$. The latter family is closed under intersections and generates $\mathcal{C}$, so $P \otimes K^* \equiv \mathbb{P}$. □

4.3 Conditional Expectations and Distributions

So far, we know conditional expectations as functions with certain properties. But the term "expectation" refers to integrals, so an obvious question is whether a conditional expectation may be viewed as an integral with respect to some "conditional distribution". As shown below, the answer is yes.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and consider random variables $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{A})$ and $Y : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \mathcal{B})$. This gives rise to three different distributions:

$$\begin{aligned} \mathcal{A} \ni A \;\mapsto\; \mathbb{P}^X(A) \;&:=\; \mathbb{P}(X \in A) && \text{(distribution of } X\text{)}, \\ \mathcal{B} \ni B \;\mapsto\; \mathbb{P}^Y(B) \;&:=\; \mathbb{P}(Y \in B) && \text{(distribution of } Y\text{)}, \\ \mathcal{A} \otimes \mathcal{B} \ni C \;\mapsto\; \mathbb{P}^{(X,Y)}(C) \;&:=\; \mathbb{P}((X, Y) \in C) && \text{(joint distribution of } X \text{ and } Y\text{)}. \end{aligned}$$

Now suppose that there exists a stochastic kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ such that

$$\mathbb{P}^{(X,Y)} \;=\; \mathbb{P}^X \otimes K.$$

Then $K(x, \cdot)$ may be interpreted as the *conditional distribution of Y given $X = x$*. Indeed, let $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ be $\mathcal{A} \otimes \mathcal{B}$-measurable. Then it follows from Theorem 4.5 that

$$\mathbb{E}\, f(X, Y) \;=\; \int_{\mathcal{X}\times\mathcal{Y}} f \, d\, \mathbb{P}^{(X,Y)} \;=\; \int_{\mathcal{X}} h \, d\, \mathbb{P}^X$$

with the $\mathcal{A}$-measurable function $h : \mathcal{X} \to [0, \infty]$,

$$h(x) \;:=\; \int_{\mathcal{Y}} f(x, y)\, K(x, dy).$$

The same conclusion is true if $f \in \mathcal{L}^1(\mathbb{P}^{(X,Y)})$, the only caveat being that the integral $h(x)$ may not exist for points x in a set $A(f) \in \mathcal{A}$ such that $\mathbb{P}^X(A(f)) = 0$.

In particular, for $g \in \mathcal{L}^1(\mathbb{P}^Y)$ and arbitrary sets $A \in \mathcal{A}$,

$$\begin{aligned}
\int_{X^{-1}(A)} g(Y)\, d\, \mathbb{P} \;&=\; \int_{\Omega} 1_A(X) g(Y)\, d\, \mathbb{P} \\
&=\; \int_{\mathcal{X}\times\mathcal{Y}} 1_A(x) g(y)\; \mathbb{P}^{(X,Y)}(d(x, y)) \\
&=\; \int_{\mathcal{X}} \int_{\mathcal{Y}} 1_A(x) g(y)\, K(x, dy)\; \mathbb{P}^X(dx) \\
&=\; \int_{A} \int_{\mathcal{Y}} g(y)\, K(x, dy)\; \mathbb{P}^X(dx),
\end{aligned}$$

so

$$x \;\mapsto\; \int_{\mathcal{Y}} g(y)\, K(x, dy)$$

is a version of $\mathbb{E}(g(Y) \,|\, X)$. One often writes

$$\left.\begin{array}{c} \mathbb{P}(Y \in B \,|\, X = x) \\ \mathbb{P}^Y(B \,|\, X = x) \end{array}\right\} \text{ instead of } K(x, B),$$

$$\mathbb{E}(g(Y) \,|\, X = x) \text{ instead of } \int_{\mathcal{Y}} g(y)\, K(x, dy).$$

Optimal predictions. In the special case that $(\mathcal{Y}, d)$ is a separable, complete metric space and $\mathcal{B} = \mathrm{Borel}(\mathcal{Y}, d)$, it follows from Theorem 4.12 that a kernel K with the stated properties does exist. In particular, let $\mathcal{Y} = \mathbb{R}$. If $Y \in \mathcal{L}^1(\mathbb{P})$, then a version of $\mathbb{E}(Y \,|\, X)$ is given by

$$\mathbb{E}(Y \,|\, X = x) \;:=\; \int_{\mathbb{R}} y\, K(x, dy).$$

Note that

$$\infty \;>\; \int |Y|\, d\, \mathbb{P} \;=\; \int_{\mathcal{X}} |y|\, K(x, dy)\; \mathbb{P}^X(dx),$$

so we may redefine $K(x, \cdot) := \mathbb{P}^Y$ for the exceptional points $x \in \mathcal{X}$ such that $\int |y|\, K(x, dy) = \infty$. With the kernel K at hand, we can solve various prediction problems. The goal is to find a *predictor $g(X)$ of Y*, determined by a measurable function $g : (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$, such that the approximation error $Y - g(X)$ is "small". The solution of this prediction problem depends on the way how we measure the approximation error precisely.

Case 1: Mean squared prediction error. Suppose that $Y \in \mathcal{L}^2(\mathbb{P})$. If we want to minimize $\mathbb{E}\big((Y - g(X))^2\big)$, one can deduce from the considerations in Section 3.2 that $\mathbb{E}(Y \,|\, X)$ solves this problem. But let us follow a more direct route using the kernel K. Recall first that for any constant $q \in \mathbb{R}$, the expectation of $(Y - q)^2$ equals $\mathrm{Var}(Y) + (\mathbb{E}(Y) - q)^2$, where $\mathrm{Var}(Y) = \mathbb{E}\big((Y - \mathbb{E}(Y))^2\big)$. Note also that

$$\infty \;>\; \int Y^2 \, d\,\mathbb{P} \;=\; \int_{\mathcal{X}} \int_{\mathbb{R}} y^2 \, K(x, dy) \; \mathbb{P}^X(dx),$$

so we may redefine $K(x, \cdot) := \mathbb{P}^Y$ whenever $\int y^2 \, K(x, dy) = \infty$. Now we can write

$$\begin{aligned} \mathbb{E}\big((Y - g(X))^2\big) \;&=\; \int_{\mathcal{X}} \int_{\mathbb{R}} (y - g(x))^2 \, K(x, dy) \; \mathbb{P}^X(dx) \\ &=\; \int_{\mathcal{X}} \Big(\mathrm{Var}(Y \,|\, X = x) + \big(\mathbb{E}(Y \,|\, X = x) - g(x)\big)^2 \Big) \; \mathbb{P}^X(dx), \end{aligned}$$

where

$$\mathrm{Var}(Y \,|\, X = x) \;:=\; \int_{\mathbb{R}} (y - \mathbb{E}(Y \,|\, X = x))^2 \, K(x, dy).$$

Consequently, a predictor $g(X)$ of Y minimizes the mean squared prediction error if and only if $g(X) = \mathbb{E}(Y \,|\, X)$ almost surely.

Case 2: Mean absolute prediction error. Suppose we want to minimize $\mathbb{E}\,|Y - g(X)|$. It is well-known that for a real constant q, the expectation of $|Y - q|$ is minimal if and only if q is a median of (the distribution of) Y, that means $\mathbb{P}(Y < q) \le 1/2 \le \mathbb{P}(Y \ge q)$. This leads to an optimal predictor $g(X)$:

$$\mathbb{E}\,|Y - g(X)| \;=\; \int_{\mathcal{X}} \int_{\mathbb{R}} |y - g(x)| \, K(x, dy) \; \mathbb{P}^X(dx)$$

is minimal if and only if $g(x)$ is a median for $\mathbb{P}(Y \in \cdot \,|\, X = x)$, that means,

$$\mathbb{P}(Y < g(x) \,|\, X = x) \;\le\; 1/2 \;\le\; \mathbb{P}(Y \le g(x) \,|\, X = x),$$

for $\mathbb{P}^X$-almost all $x \in \mathcal{X}$.

Chapter 5

Haar Measure

5.1 Locally Compact Topological Groups

Groups. Recall the definition of a group $(\mathcal{X}, \cdot)$. This is a set $\mathcal{X}$ with a binary operation

$$\mathcal{X} \times \mathcal{X} \ni (x, y) \ \mapsto \ x \cdot y \in \mathcal{X}$$

satisfying the following three conditions:

- (Associativity) For arbitrary $x, y, z \in \mathcal{X}$,

$$x \cdot (y \cdot z) \ = \ (x \cdot y) \cdot z.$$

- (Identity element) There exists an element $e \in \mathcal{X}$ such that for arbitrary $x \in \mathcal{X}$,

$$x \cdot e \ = \ e \cdot x \ = \ x.$$

- (Inverse elements) For each $x \in \mathcal{X}$ there exists an element $x^{-1} \in \mathcal{X}$ such that

$$x \cdot x^{-1} \ = \ x^{-1} \cdot x \ = \ e.$$

The identity element e is unique, and for each $x \in \mathcal{X}$, its inverse element x^{-1} is unique as well. Instead of $x \cdot y$ one often writes xy.

Exercise 5.1. Let $(\mathcal{X}, \cdot)$ be a group.

(a) For $x, y \in \mathcal{X}$ let $x \star y := yx$. Show that $(\mathcal{X}, \star)$ is also a group with the same identity element and the same inverse elements.

(b) Show that for any fixed $x_o \in \mathcal{X}$, the mapping $x \mapsto x_o x$ as well as the mapping $x \mapsto x x_o$ is bijective from $\mathcal{X}$ to $\mathcal{X}$.

Locally compact topological groups. Suppose that there is also a metric d on $\mathcal{X}$ such that the following two conditions are satisfied:[1]

[1]For simplicity of exposition, we restrict our attention to metric spaces, although the subsequent definitions and main results may be extended to non-metric topological spaces.

- The mapping $x \mapsto x^{-1}$, from $\mathcal{X}$ to $\mathcal{X}$ is continuous.
- The mapping $(x, y) \mapsto xy$, from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{X}$ is continuous.[2]

Then one calls $(\mathcal{X}, \cdot, d)$ a *topological group*.

Suppose that in addition,

- for each $x \in \mathcal{X}$ there exists an $\epsilon > 0$ such that the closed ball $B_\epsilon(x)$ with center x and radius ϵ is compact.

Then one calls $(\mathcal{X}, \cdot, d)$ a *locally compact topological group*. If the group operation and metric are clear from the context, we just talk about the (locally compact) topological group $\mathcal{X}$.

Now we provide three examples of locally compact topological groups.

Example 5.2 (Euclidean spaces). Let $\mathcal{X} = \mathbb{R}^d$. With the usual addition of vectors and the usual Euclidean distance, $\mathbb{R}^d$ becomes a locally compact topological group which is also commutative.

The following two examples involve sets of invertible matrices in $\mathbb{R}^{d\times d}$ for some $d \in \mathbb{N}$, and the binary operation is matrix multiplication. To stay coherent with the general theory in this section, we denote matrices temporarily with lower-case, boldface letters, and the identity matrix is denoted with $\boldsymbol{e}$. Note that $\mathbb{R}^{d\times d}$ may be identified with $\mathbb{R}^{d^2}$ by identifying a matrix $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d]$ having columns $\boldsymbol{x}_j = (x_{ij})_{i=1}^d \in \mathbb{R}^d$ with the vector $(\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_d^\top)^\top$. Thus the usual Euclidean norm on $\mathbb{R}^{d^2}$ corresponds to Frobenius norm on $\mathbb{R}^{d\times d}$, i.e.

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \;:=\; \sum_{i,j=1}^{d} x_{ij} y_{ij} \quad \text{and} \quad \|\boldsymbol{x}\| \;:=\; \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}.$$

Note also that any other norm on $\mathbb{R}^{d\times d}$ or $\mathbb{R}^{d^2}$ induces the same topology. Finally, a basic fact is that matrix multiplication is a continuous mapping from $\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d}$ to $\mathbb{R}^{d\times d}$.

Example 5.3 (Linear groups). Let $\mathcal{X}$ be the set of all invertible matrices $\boldsymbol{x} \in \mathbb{R}^{d\times d}$. With the usual matrix multiplication, $\mathcal{X}$ is a non-commutative[3] group. Moreover, the set $\mathcal{X}$ is an open subset of $\mathbb{R}^{d\times d}$, and the mapping $x \mapsto x^{-1}$ is continuous from $\mathcal{X}$ to $\mathcal{X}$. Indeed, for $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\delta} \in \mathbb{R}^{d\times d}$, it is well-known that $\boldsymbol{x} + \boldsymbol{\delta}$ is invertible, provided that $\|\boldsymbol{\delta}\|$ is sufficiently small, and

$$(\boldsymbol{x} + \boldsymbol{\delta})^{-1} \;=\; \boldsymbol{x}^{-1} - \boldsymbol{x}^{-1}\boldsymbol{\delta}\boldsymbol{x}^{-1} + O(\|\boldsymbol{\delta}\|^2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0}.$$

This follows from $\boldsymbol{x} + \boldsymbol{\delta} = \boldsymbol{x}(\boldsymbol{e} + \boldsymbol{x}^{-1}\boldsymbol{\delta})$ and from von Neumann's series expansion

$$(\boldsymbol{e} + \boldsymbol{a})^{-1} \;=\; \sum_{k=0}^{\infty} (-1)^k \boldsymbol{a}^k$$

for $\boldsymbol{a} \in \mathbb{R}^{d\times d}$ with $\|\boldsymbol{a}\|$ sufficiently small. Hence $\mathcal{X}$ is a topological group.

Example 5.4 (Orthogonal groups). Let $\mathcal{X}_{\mathrm{orth}}$ be the set of orthogonal matrices $\boldsymbol{x} \in \mathbb{R}^{d\times d}$, that means, $\boldsymbol{x}^\top \boldsymbol{x} = \boldsymbol{e}$. This is a subgroup of the set $\mathcal{X}$ of all invertible matrices and thus a locally compact topological group. Indeed, $\mathcal{X}_{\mathrm{orth}}$ is even a compact set and a smooth $d(d-1)/2$-dimensional

[2] Here $\mathcal{X} \times \mathcal{X}$ is equipped with the metric $d\big((x, x'), (y, y')\big) := \max\big\{d(x, y), d(x', y')\big\}$.

[3] unless $d = 1$

manifold. To see this, note first that

$$\mathcal{X}_{\rm orth} \;=\; F^{-1}(\boldsymbol{e})$$

with the continuous mapping

$$F : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}, \quad F(\boldsymbol{x}) \;:=\; \boldsymbol{x}^\top \boldsymbol{x}.$$

Hence, $\mathcal{X}_{\rm orth}$ is a closed subset of $\mathbb{R}^{d\times d}$. Moreover, with the Frobenius norm $\|\cdot\|$, all matrices $\boldsymbol{x} \in \mathcal{X}_{\rm orth}$ satisfy $\|\boldsymbol{x}\| = \sqrt{d}$, so $\mathcal{X}_{\rm orth}$ is a closed and bounded subset of $\mathbb{R}^{d\times d}$. Hence it is compact.

That $\mathcal{X}_{\rm orth}$ is a smooth $d(d-1)$-dimensional manifold can be verified as follows: Note that

$$\mathbb{R}^{d\times d} \;=\; \mathbb{R}^{d\times d}_{\rm sym} + \mathbb{R}^{d\times d}_{\rm skew},$$

where $\mathbb{R}^{d\times d}_{\rm sym}$ and $\mathbb{R}^{d\times d}_{\rm skew}$ are the linear spaces of symmetric and skew-symmetric matrices in $\mathbb{R}^{d\times d}$, respectively:

$$\begin{aligned}
\mathbb{R}^{d\times d}_{\rm sym} &:=\; \big\{\boldsymbol{a} \in \mathbb{R}^{d\times d} : \boldsymbol{a}^\top = \boldsymbol{a}\big\} && \text{with dimension } d(d+1)/2,\\
\mathbb{R}^{d\times d}_{\rm skew} &:=\; \big\{\boldsymbol{a} \in \mathbb{R}^{d\times d} : \boldsymbol{a}^\top = -\boldsymbol{a}\big\} && \text{with dimension } d(d-1)/2.
\end{aligned}$$

Moreover, $\mathbb{R}^{d\times d}_{\rm sym} \perp \mathbb{R}^{d\times d}_{\rm skew}$, and any matrix $\boldsymbol{a}$ may be written as $\boldsymbol{a} = \boldsymbol{a}_{\rm sym} + \boldsymbol{a}_{\rm skew}$ with

$$\boldsymbol{a}_{\rm sym} := 2^{-1}(\boldsymbol{a} + \boldsymbol{a}^\top) \in \mathbb{R}^{d\times d}_{\rm sym}, \quad \boldsymbol{a}_{\rm skew} := 2^{-1}(\boldsymbol{a} - \boldsymbol{a}^\top) \in \mathbb{R}^{d\times d}_{\rm skew}.$$

Note that $F(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{x}$ defines a continuously differentiable mapping

$$F : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}_{\rm sym}.$$

For any fixed $\boldsymbol{x} \in \mathcal{X}_{\rm orth}$ and arbitrary $\boldsymbol{\delta} \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}
F(\boldsymbol{x}+\boldsymbol{\delta}) \;&=\; \boldsymbol{e} + \boldsymbol{x}^\top\boldsymbol{\delta} + \boldsymbol{\delta}^\top\boldsymbol{x} + \boldsymbol{\delta}^\top\boldsymbol{\delta}\\
&=\; \boldsymbol{e} + 2(\boldsymbol{x}^\top\boldsymbol{\delta})_{\rm sym} + O(\|\boldsymbol{\delta}\|^2) \quad \text{as } \boldsymbol{\delta} \to \boldsymbol{0}.
\end{aligned}$$

Hence the derivative of F at $\boldsymbol{x} \in \mathcal{X}_{\rm orth}$ is given by the linear mapping

$$DF(\boldsymbol{x}) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}_{\rm sym}, \quad DF(\boldsymbol{x})\boldsymbol{a} \;=\; 2(\boldsymbol{x}^\top\boldsymbol{a})_{\rm sym},$$

and the null space of this mapping equals

$$\big\{\boldsymbol{a} \in \mathbb{R}^{d\times d} : DF(\boldsymbol{x})\boldsymbol{a} = \boldsymbol{0}\big\} \;=\; \big\{\boldsymbol{x}\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^{d\times d}, \boldsymbol{b}_{\rm sym} = \boldsymbol{0}\big\} \;=\; \big\{\boldsymbol{x}\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^{d\times d}_{\rm skew}\big\}$$

whereas

$$DF(\boldsymbol{x})\boldsymbol{x}\boldsymbol{b} \;=\; 2\boldsymbol{b} \quad \text{for all } \boldsymbol{b} \in \mathbb{R}^{d\times d}_{\rm sym}.$$

Consequently, the linear mapping $DF(\boldsymbol{x})$ has full rank $d(d+1)/2$, and the implicit function theorem shows that $\mathcal{X}_{\rm orth}$ is a differentiable manifold of dimension $d(d-1)/2$. Its tangent space at any point $\boldsymbol{x} \in \mathcal{X}_{\rm orth}$ is the null space of $DF(\boldsymbol{x})$. That means, for small $\epsilon > 0$,

$$\big\{\boldsymbol{y} \in \mathcal{X}_{\rm orth} : \|\boldsymbol{y} - \boldsymbol{x}\| < \epsilon\big\} \;\approx\; \big\{\boldsymbol{x} + \boldsymbol{x}\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^{d\times d}_{\rm skew}, \|\boldsymbol{b}\| < \epsilon\big\}.$$

Exercise 5.5 (An embedding for $\mathcal{X}_{\mathrm{orth}}$). Consider the mapping $\Psi : \mathcal{X} \to \mathbb{R}^{d\times d}$,

$$\Psi(\boldsymbol{x}) \;:=\; \boldsymbol{x}\big(\boldsymbol{x}^\top\boldsymbol{x}\big)^{-1/2}.$$

Show that Ψ has the following five properties:

(a) $\Psi(\boldsymbol{x}) \in \mathcal{X}_{\mathrm{orth}}$ for all $\boldsymbol{x} \in \mathcal{X}$;

(b) $\Psi(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x} \in \mathcal{X}_{\mathrm{orth}}$;

(c) $\Psi(\boldsymbol{x}\boldsymbol{y}) = \boldsymbol{x}\Psi(\boldsymbol{y})$ for all $\boldsymbol{x} \in \mathcal{X}_{\mathrm{orth}}$ and $\boldsymbol{y} \in \mathcal{X}$;

(d) $\Psi(\boldsymbol{e} + \boldsymbol{\delta}) = \boldsymbol{e}$ for all $\boldsymbol{\delta} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ such that $\lambda_{\min}(\boldsymbol{\delta}) > -1$;

(e) for $\boldsymbol{\delta} \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{e} + \boldsymbol{\delta} \in \mathcal{X}$,

$$\Psi(\boldsymbol{e} + \boldsymbol{\delta}) \;=\; \boldsymbol{e} + \boldsymbol{\delta}_{\mathrm{skew}} + O(\|\boldsymbol{\delta}\|^2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0}.$$

In part (e) one should use (and possibly verify) the fact that for $\boldsymbol{a} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ with minimal eigenvalue $\lambda_{\min}(\boldsymbol{a}) > -1$ and any fixed $b \in \mathbb{R}$,

$$(\boldsymbol{e} + \boldsymbol{a})^b \;=\; \boldsymbol{e} + b\boldsymbol{a} + O(\|\boldsymbol{a}\|^2) \quad \text{as } \boldsymbol{a} \to \mathbf{0}.$$

5.2 Left- and Right-Invariant Measures

Throughout this section let $\mathcal{X}$ be a locally compact topological group. In what follows, for $A, B \subset \mathcal{X}$ we write

$$AB \;:=\; \{ab : a \in A, b \in B\} \quad \text{and} \quad B^{-1} \;:=\; \{b^{-1} : b \in B\}.$$

Moreover, for $x \in \mathcal{X}$ we set $xB := \{xb : b \in B\}$ and $Ax := \{ax : a \in A\}$. We also write

$$\begin{aligned}
\mathcal{C} &:= \{\text{compact subsets of } \mathcal{X}\},\\
\mathcal{U} &:= \{\text{open subsets of } \mathcal{X}\},\\
\mathcal{B} &:= \{\text{Borel subsets of } \mathcal{X}\}.
\end{aligned}$$

Our first main result is about the existence and (essential) uniqueness of a left-invariant measure on (the Borel subsets of) $\mathcal{X}$.

Theorem 5.6 (Haar–Weil). *There exists a measure μ on $(\mathcal{X}, \mathcal{B})$ with the following five properties:*

$$\mu(xB) \;=\; \mu(B) \qquad \text{for all } x \in \mathcal{X}, B \in \mathcal{B}, \tag{5.1}$$

$$\mu(B) \;=\; \inf_{U\in\mathcal{U}:U\supset B} \mu(U) \quad \text{for all } B \in \mathcal{B}, \tag{5.2}$$

$$\mu(U) \;=\; \sup_{C\in\mathcal{C}:C\subset U} \mu(C) \quad \text{for all } U \in \mathcal{U}, \tag{5.3}$$

$$\mu(C) \;<\; \infty \qquad \text{for all } C \in \mathcal{C}, \tag{5.4}$$

$$\mu(U) \;>\; 0 \qquad \text{for all } U \in \mathcal{U} \setminus \{\emptyset\}. \tag{5.5}$$

If $\tilde{\mu}$ is another measure with these properties, then there exists a constant $\gamma > 0$ such that $\tilde{\mu} = \gamma\mu$.

Definition 5.7 (Left Haar measure). A measure μ with the properties listed in Theorem 5.6 is called a *left Haar measure on* $\mathcal{X}$.

The proof presented here is essentially the one of Haar (1933). Since it uses Tikhonov's theorem and thus the axiom of choice, some mathematicians came up with constructive proofs, see Alfsen (1963) and the references therein.

Proof of Theorem 5.6. In this proof, some facts are stated without proof, and the reader should fill in these details. For arbitrary sets $A, B \subset \mathcal{X}$ with $B \neq \emptyset$ we define a "covering number"

$$N(A \,|\, B) \;:=\; \min\bigl\{\#\mathcal{X}_o : \mathcal{X}_o \subset \mathcal{X}, A \subset \mathcal{X}_o B\bigr\} \;\in\; \mathbb{N}_0 \cup \{\infty\}.$$

These covering numbers have the following properties:

$$\begin{aligned}
&N(B \,|\, B) = 1.\\
&N(yA \,|\, zB) = N(A \,|\, B) \text{ for arbitrary } y, z \in \mathcal{X}.\\
&N(A \,|\, B) > 0 \text{ if and only if } A \neq \emptyset.\\
&N(A \,|\, B) < \infty \text{ if } A \in \mathcal{C} \text{ and } B \text{ has non-empty interior.}
\end{aligned}$$

Furthermore, for $A_1, A_2 \subset \mathcal{X}$,

$$\begin{aligned}
&N(A_1 \,|\, B) \;\leq\; N(A_2 \,|\, B) \quad \text{if } A_1 \subset A_2,\\
&N(A_1 \cup A_2 \,|\, B) \begin{cases} \leq\; N(A_1 \,|\, B) + N(A_2 \,|\, B), \\ =\; N(A_1 \,|\, B) + N(A_2 \,|\, B) \quad \text{if } A_1 B^{-1} \cap A_2 B^{-1} = \emptyset. \end{cases}
\end{aligned}$$

Finally, if B' is another non-empty subset of $\mathcal{X}$, then

$$N(A \,|\, B') \;\leq\; N(A \,|\, B) N(B \,|\, B'). \tag{5.6}$$

The rough idea behind Haar's and Weil's construction is that $N(\cdot \,|\, B)$ behaves almost like a left-invariant measure if B is a "small" neighborhood of the identity element e. To make this precise, let B_ϵ be the closed ball around e with radius $\epsilon > 0$,

$$B_\epsilon \;:=\; \{x \in \mathcal{X} : d(x, e) \leq \epsilon\}.$$

Without loss of generality let B_1 be compact[4]. For $0 < \epsilon \leq 1$ we define $\lambda_\epsilon : \mathcal{C} \to [0, \infty)$ via

$$\lambda_\epsilon(C) \;:=\; \frac{N(C \,|\, B_\epsilon)}{N(B_1 \,|\, B_\epsilon)}.$$

It follows from the properties of $N(\cdot \,|\, \cdot)$ that $\lambda_\epsilon(C)$ is well-defined in $[0, \infty)$ for any $C \in \mathcal{C}$, and

$$\lambda_\epsilon(xC) \;=\; \lambda_\epsilon(C) \quad \text{for arbitrary } x \in \mathcal{X}, C \in \mathcal{C}.$$

Moreover, (5.6) implies that

$$\lambda_\epsilon(C) \begin{cases} \leq\; \dfrac{N(C \,|\, B_1) N(B_1 \,|\, B_\epsilon)}{N(B_1 \,|\, B_\epsilon)} \;=\; N(C \,|\, B_1) \\ \geq\; \dfrac{N(C \,|\, B_\epsilon)}{N(B_1 \,|\, A) N(C \,|\, B_\epsilon)} \;=\; N(B_1 \,|\, A)^{-1} \end{cases}$$

[4]Otherwise, replace $d(\cdot, \cdot)$ with $\epsilon_o^{-1} d(\cdot, \cdot)$ for sufficiently small $\epsilon_o > 0$.

with the conventions $N(B_1 \,|\, \emptyset) := \infty$ and $1/\infty := 0$. In particular, $\lambda_\epsilon(B_1) = 1$. Finally, for $C_1, C_2 \in \mathcal{C}$,

$$\begin{aligned} &\lambda_\epsilon(C_1) \;\le\; \lambda_\epsilon(C_2) \quad \text{if } C_1 \subset C_2, \\ &\lambda_\epsilon(C_1 \cup C_2) \begin{cases} \le\; \lambda_\epsilon(C_1) + \lambda_\epsilon(C_2), \\ =\; \lambda_\epsilon(C_1) + \lambda_\epsilon(C_2) \quad \text{if } C_1 B_\epsilon^{-1} \cap C_2 B_\epsilon^{-1} = \emptyset, \end{cases} \end{aligned} \tag{5.7}$$

Note that each λ_ϵ, viewed as a tuple $(\lambda_\epsilon(C))_{C \in \mathcal{C}}$, is a point in the infinite Cartesian product

$$\prod_{C \in \mathcal{X}} K_C$$

with the compact intervals $K_C := \big[N(B_1 \,|\, C)^{-1}, N(C \,|\, B_1)\big]$. It follows from Tikhonov's theorem that this product, equipped with the corresponding product topology, is a compact topological space. Hence, the functions λ_ϵ, $\epsilon \in (0,1]$, have a cluster point $\lambda \in \prod_{C \in \mathcal{C}} K_C$ as $\epsilon \downarrow 0$. Interpreting λ as a function $\lambda : \mathcal{C} \to [0, \infty)$, this cluster point property means that for any finite set $\mathcal{C}_o \subset \mathcal{C}$,

$$\liminf_{\epsilon \downarrow 0} \max_{C \in \mathcal{C}_o} \big|\lambda_\epsilon(C) - \lambda(C)\big| \;=\; 0.$$

In particular, λ inherits various nice properties of the functions λ_ϵ:

$$\begin{aligned} &\lambda(C) \in K_C \;\text{ for any } C \in \mathcal{C}, \\ &\lambda(B_1) = 1, \\ &\lambda(xC) = \lambda(C) \;\text{ for arbitrary } x \in \mathcal{X}, C \in \mathcal{C}, \end{aligned}$$

and for $C_1, C_2 \in \mathcal{C}$,

$$\begin{aligned} &\lambda(C_1) \;\le\; \lambda(C_2) \quad \text{if } C_1 \subset C_2, \\ &\lambda(C_1 \cup C_2) \begin{cases} \le \lambda(C_1) + \lambda(C_2), \\ = \lambda(C_1) + \lambda(C_2) \quad \text{if } C_1 \cap C_2 = \emptyset. \end{cases} \end{aligned}.$$

The latter equality follows from (5.7) and the fact that for disjoint compact sets C_1 and C_2, the neighborhoods $C_1 B_\epsilon^{-1}$ and $C_2 B_\epsilon^{-1}$ are disjoint for sufficiently small $\epsilon > 0$.

Now we are ready to define the measure μ. At first we set

$$\mu(U) \;:=\; \sup_{C \in \mathcal{C} : C \subset U} \lambda(C) \quad \text{for } U \in \mathcal{U}.$$

This definition implies that

$$\mu(\emptyset) = 0 \quad \text{and} \quad \mu(U) \le \mu(U') \;\text{ for open sets } U \subset U' \subset \mathcal{X}.$$

This allows us to extend μ to a function on $\mathcal{P}(\mathcal{X})$ via

$$\mu(M) \;:=\; \inf_{U \in \mathcal{U} : U \supset M} \mu(U) \quad \text{for } M \subset \mathcal{X}.$$

Note that

$$\mu(M) \;\le\; \mu(M') \quad \text{for } M \subset M' \subset \mathcal{X}.$$

Since $\lambda(\cdot)$ is left-invariant, the same is true for μ, that means,

$$\mu(xM) \;=\; \mu(M) \quad \text{for all } x \in \mathcal{X} \text{ and } M \subset \mathcal{X}.$$

Note also that

$$\lambda(C) \;\leq\; \inf_{U \in \mathcal{U}: C \subset U} \mu(U) \;=\; \mu(C) \quad \text{for all } C \in \mathcal{C}.$$

The first step follows from the definition of $\mu(U)$ for open sets U, the second step is just the definition of $\mu(C)$. Together with the monotonicity of $\mu(\cdot)$ this implies that

$$\mu(U) \;=\; \sup_{C \in \mathcal{C}: C \subset U} \mu(C) \quad \text{for all } U \in \mathcal{U}.$$

To show that $\mu > 0$ on $\mathcal{U} \setminus \{\emptyset\}$ and $\mu < \infty$ on $\mathcal{C}$, consider arbitrary sets $C \in \mathcal{C} \setminus \{\emptyset\}$ and $U \in \mathcal{U}$ such that $C \subset U$. For each $x \in C$ there exist $U_x \in \mathcal{U}$ and $B_x \in \mathcal{C}$ such that $x \in U_x \subset B_x \subset U$. Then $(U_x)_{x \in C}$ defines a covering of C with open sets, so by compactness of C, there exists a finite set $C_o \subset C$ such that

$$C \;\subset\; \tilde{U} := \bigcup_{x \in C_o} U_x \;\subset\; \tilde{C} := \bigcup_{x \in C_o} B_x \;\subset\; U.$$

Note that $\tilde{U} \in \mathcal{U}$ and $\tilde{C} \in \mathcal{C}$. Hence,

$$\mu(C) \;\leq\; \mu(\tilde{U}) \;\leq\; \lambda(\tilde{C}) \;<\; \infty$$

and

$$\mu(U) \;\geq\; \lambda(\tilde{C}) \;\geq\; N(B_1 \,|\, \tilde{C})^{-1} \;>\; 0,$$

because $\tilde{C}$ has non-empty interior.

It remains to show that μ defines a measure on $\mathcal{B}$. To this end, we apply Carathéodory's theory of outer measures. It suffices to show that μ defines an outer measure on $\mathcal{X}$ and that any open set $U \in \mathcal{U}$ is μ-measurable in the sense that

$$\mu(M) \geq \mu(M \cap U) + \mu(M \setminus U) \quad \text{for any } M \subset \mathcal{X}. \tag{5.8}$$

Proof that μ is an outer measure: Let $(M_n)_{n \geq 1}$ be an arbitrary sequence of sets $M_n \subset \mathcal{X}$, and let $M \subset \bigcup_{n \geq 1} M_n$. We have to show that $\mu(M) \leq \sum_{n \geq 1} \mu(M_n)$. We may assume that the right hand side is finite, because otherwise the claim is trivial. For arbitrary fixed $\epsilon > 0$ let $U_n \in \mathcal{U}$ with $M_n \subset U_n$ and $\mu(U_n) \leq \mu(M_n) + 2^{-n}\epsilon$. Then $U := \bigcup_{n \geq 1} U_n$ is an open set containung M, whence

$$\mu(M) \;\leq\; \mu(U) \quad \text{while} \quad \sum_{n \geq 1} \mu(U_n) \;\leq\; \sum_{n \geq 1} \mu(M_n) + \epsilon.$$

Consequently, it suffices to show that $\mu(U) \leq \sum_{n \geq 1} \mu(U_n)$. For any compact set $C \subset U$ and $x \in C$ there exist $n(x) \in \mathbb{N}$ and a compact neighborhood B_x of x such that $B_x \subset U_{n(x)}$. But compactness of C implies that for some finite set $C_o \subset C$,

$$C \;\subset\; \bigcup_{x \in C_o} B_x \;=\; \bigcup_{n=1}^{m} \bigcup_{x \in C_o : n(x) = n} B_x \;=\; \bigcup_{n=1}^{m} C_n$$

with $m := \max\{n(x) : x \in C_o\}$ and the compact sets

$$C_n \;:=\; \bigcup_{x \in C_o : n(x) = n} B_x \;\subset\; U_n.$$

Consequently, the properties of λ imply that

$$\lambda(C) \;\le\; \lambda\Big(\bigcup_{n=1}^{m} C_n\Big) \;\le\; \sum_{n=1}^{m} \lambda(C_n) \;\le\; \sum_{n=1}^{m} \mu(U_n),$$

and for $\lambda(C) \uparrow \mu(U)$ the asserted inequality follows.

Proof of (5.8): It suffices to consider the case $\mu(M) < \infty$. For any fixed $\epsilon > 0$, there exists an open set $V \supset M$ such that $\mu(V) \le \mu(M) + \epsilon$. On the other hand,

$$\mu(M \cap U) \;\le\; \mu(V \cap U) \;\le\; \lambda(C) + \epsilon$$

for some compact set $C \subset V \cap U$. But then

$$\mu(M \setminus U) \;\le\; \mu(V \setminus U) \;=\; \mu(V \setminus (V \cap U)) \;\le\; \mu(V \setminus C) \;\le\; \lambda(\tilde{C}) + \epsilon$$

for some compact set $\tilde{C} \subset V \setminus C$. Hence,

$$\mu(M) + \epsilon \;\ge\; \mu(V) \;\ge\; \lambda(C \cup \tilde{C}) \;=\; \lambda(C) + \lambda(\tilde{C}) \;\ge\; \mu(M \cap U) + \mu(M \setminus U) - 2\epsilon.$$

As $\epsilon \downarrow 0$, this yields (5.8).

It remains to prove uniqueness of μ up to positive multiples. To this end, let $\tilde{\mu}$ be a second measure on $\mathcal{B}$ with the stated properties. It suffices to show that $\tilde{\mu} = \gamma\mu$ on $\mathcal{C}_*$ for some constant $\gamma > 0$, where $\mathcal{C}_*$ is the set of compact subsets of $\mathcal{X}$ with non-empty interior. To this end, consider arbitrary compact sets $A, B \subset \mathcal{X}$. Then

$$\begin{aligned}
\mu(A)\tilde{\mu}(B) \;&=\; \int \mu(A) 1_B(y)\, \tilde{\mu}(dy) \\
&=\; \int \mu(yA) 1_B(y)\, \tilde{\mu}(dy) && \text{(left-invariance of } \mu) \\
&=\; \int\int 1_{yA}(x) 1_B(y)\, \mu(dx)\, \tilde{\mu}(dy) \\
&=\; \int\int 1_{yA}(x) 1_B(y)\, \tilde{\mu}(dy)\, \mu(dx) && \text{(Fubini's theorem)} \\
&=\; \int\int 1_{xA^{-1}}(y) 1_B(y)\, \tilde{\mu}(dy)\, \mu(dx) && (x \in yA \;\text{ iff }\; y \in xA^{-1}) \\
&=\; \int \tilde{\mu}\big((xA^{-1}) \cap B\big)\, \mu(dx) \\
&=\; \int \tilde{\mu}\big(A^{-1} \cap (x^{-1}B)\big)\, \mu(dx) && \text{(left-invariance of } \tilde{\mu}) \\
&\le\; \int 1_{BA}(x) \tilde{\mu}(A^{-1})\, \mu(dx) && (A^{-1} \cap (x^{-1}B) = \emptyset \;\text{ if } x \notin BA) \\
&=\; \mu(BA)\tilde{\mu}(A^{-1}).
\end{aligned}$$

Note that the application of Fubini's theorem was justified because the integrand, $1_{yA}(x)1_B(y)$, equals zero if $x \notin BA$ or $y \notin B$, so μ and $\tilde{\mu}$ may be viewed temporarily as finite measures on the compact sets BA and B, respectively. Interchanging the roles of μ and $\tilde{\mu}$ yields the inequalities

$$\mu(A)\tilde{\mu}(B) \;\le\; \mu(BA)\tilde{\mu}(A^{-1}) \quad \text{and} \quad \tilde{\mu}(A)\mu(C) \;\le\; \tilde{\mu}(CA)\mu(A^{-1})$$

for $A, B, C \in \mathcal{C}$. Specifically let B, C be arbitrary fixed sets in $\mathcal{C}_*$, and let $A := B_\epsilon \cap B_\epsilon^{-1}$ for some $\epsilon \in (0, 1]$, so $A^{-1} = A$. Then

$$\frac{\tilde{\mu}(B)}{\mu(BA)} \;\le\; \frac{\tilde{\mu}(A)}{\mu(A)} \;\le\; \frac{\tilde{\mu}(CA)}{\mu(C)}.$$

If we let $\epsilon \downarrow 0$, then the left-hand and right-hand side converge to $\tilde{\mu}(B)/\mu(B)$ and $\tilde{\mu}(C)/\mu(C)$, respectively, because for arbitrary $U, V \in \mathcal{U}$ with $B \subset U$ and $C \subset V$, $BA \subset U$ and $CA \subset V$ for sufficiently small $\epsilon > 0$. This shows that $\tilde{\mu}/\mu$ is constant on $\mathcal{C}_*$. □

By means of Exercise 5.1 (a), one can easily deduce from Theorem 5.6 that there exists a *right Haar measure* μ on $\mathcal{X}$, that means, Condition 5.1 in Theorem 5.6 can be replaced with

$$\mu(Bx) \;=\; \mu(B) \quad \text{for all } x \in \mathcal{X} \text{ and } B \in \mathcal{B}. \tag{5.9}$$

Note that in case of a commutative group, there is no difference between left and right Haar measures, so we talk about *Haar measures*.

An obvious question is how left and right Haar measures are related. The next result is a first step to clarify this.

Theorem 5.8. *There exists a unique function $J : \mathcal{X} \to (0, \infty)$ such that for any left Haar measure μ on $\mathcal{X}$,*

$$\mu(Bx) \;=\; J(x)\mu(B) \quad \text{for all } x \in \mathcal{X} \text{ and } B \in \mathcal{B}.$$

Moreover, J is continuous, and

$$J(xy) \;=\; J(x)J(y) \quad \text{for arbitrary } x, y \in \mathcal{X}.$$

Here is an immediate consequence of this result:

Corollary 5.9. *Suppose that $\mathcal{X}$ is a compact topological group. Then any left Haar measure on $\mathcal{X}$ is also a right Haar measure.*

The reason is that $J(x^z) = J(x)^z$ for arbitrary $x \in \mathcal{X}$ and integers z. Consequently, if $J(x) \neq 1$ for some $x \in \mathcal{X}$, then J is unbounded on $\mathcal{X}$. But a continuous function on a compact set is bounded, whence $J \equiv 1$ in case of a compact topological group $\mathcal{X}$.

Proof of Theorem 5.8. If μ is a left Haar measure on $\mathcal{X}$, then for fixed $x \in \mathcal{X}$,

$$B \;\mapsto\; \mu(Bx)$$

is easily seen to define a left Haar measure on $\mathcal{X}$, too. But this means that there exists a unique constant $J(x) > 0$ such that

$$\mu(Bx) \;=\; J(x)\mu(B) \quad \text{for all } B \in \mathcal{B}.$$

For $x, y \in \mathcal{X}$ and $B \in \mathcal{B}$,

$$J(xy)\mu(B) \;=\; \mu(Bxy) \;=\; J(y)\mu(Bx) \;=\; J(y)J(x)\mu(B).$$

Since $0 < \mu(B) < \infty$ in case of B being compact with non-empty interior, this proves the equation $J(xy) = J(x)J(y)$ for $x, y \in \mathcal{X}$.

To show continuity of J, let $A_\epsilon := B_\epsilon \cap B_\epsilon^{-1}$ with $B_\epsilon = \{x \in \mathcal{X} : d(x, e) \le \epsilon\}$. Further, let $C \in \mathcal{C}$ with non-empty interior. Then for any $\kappa > 1$ there exists an $\epsilon > 0$ such that $\mu(CA_\epsilon) \le \mu(C)(1 + \kappa)$. Thus for $x \in A_\epsilon$,

$$J(x) \;=\; \frac{\mu(Cx)}{\mu(C)} \;\le\; \frac{\mu(CA_\epsilon)}{\mu(C)} \;\le\; 1 + \kappa$$

and

$$J(x) \;=\; \frac{1}{J(x^{-1})} \;=\; \frac{\mu(C)}{\mu(Cx^{-1})} \;\ge\; \frac{\mu(C)}{\mu(CA_\epsilon)} \;\ge\; \frac{1}{1+\kappa}.$$

This proves continuity of J at e, and continuity at arbitrary points in $\mathcal{X}$ is easily deduced from the equation $J(xy) = J(x)J(y)$ for arbitrary $x, y \in \mathcal{X}$. □

The function J in Theorem 5.8 is called the *modular function* of the locally compact topological group $\mathcal{X}$. It yields a Radon–Nikodym derivative of a right-invariant with respect to a left-invariant measure.

Theorem 5.10. *Let μ be a left Haar measure on $\mathcal{X}$. Then*

$$\tilde{\mu}(B) \;:=\; \int_B J(x)^{-1}\, \mu(dx)$$

defines a right-invariant measure on $\mathcal{X}$. That means, $\tilde{\mu}$ satisifies the same properties as μ, except that (5.1) *is replaced with* (5.9).

As a preparation for the proof of Theorem 5.10 the reader should verify the following two facts about integrals with respect to left-invariant measures.

Exercise 5.11. Let μ be a left Haar measure on $\mathcal{X}$. Show that for arbitrary $y \in \mathcal{X}$ and measurable functions $h : \mathcal{X} \to \bar{\mathbb{R}}$,

$$\int h(yx)\, \mu(dx) \;=\; \int h\, d\mu \quad \text{and} \quad \int h(xy)\, \mu(dx) \;=\; J(y)^{-1} \int h\, d\mu,$$

provided that $\int h\, d\mu$ is well-defined in $\bar{\mathbb{R}}$.

Proof of Theorem 5.10. We start with an arbitrary measure μ satisfying the regularity conditions (5.2) and (5.3) and consider any measure $\tilde{\mu}$ with a continuous density $f = d\tilde{\mu}/d\mu > 0$.

Since $f > 0$, $\tilde{\mu}(B) > 0$ whenever $\mu(B) > 0$. Hence, $\tilde{\mu}$ satisfies property (5.5) whenever μ does. Note also that for $C \in \mathcal{C}$,

$$\tilde{\mu}(C) \;\leq\; \max_{x \in C} f(x)\, \mu(C),$$

so $\tilde{\mu}$ satisfies property (5.4) whenever μ does.

To show that $\tilde{\mu}$ satisfies (5.3), note first that for any $U \in \mathcal{U}$,

$$\tilde{\mu}(U) \;=\; \lim_{R \to \infty} \tilde{\mu}(U \cap \{R^{-1} < f < R\}).$$

Since each set $\{R^{-1} < f < R\}$ is open itself, it suffices to consider open sets U such that $a \leq f \leq b$ on U with $0 < a < b$. In particular, $a\mu(B) \leq \tilde{\mu}(B) \leq b\mu(B)$ for all measurable sets $B \subset U$. By assumption, there exist compact subsets $C_1 \subset C_2 \subset C_3 \subset \cdots$ of U such that $\mu(C_n) \to \mu(U)$ as $n \to \infty$. Now one can easily show that $\tilde{\mu}(C_n) \to \tilde{\mu}(U)$ as $n \to \infty$, distinguishing the cases $\mu(U) = \infty$ and $\mu(U) < \infty$. Hence $\tilde{\mu}$ satisfies (5.3), too.

To show that $\tilde{\mu}$ satisfies (5.2), we fix an arbitrary constant $\gamma > 1$. Then any set $B \in \mathcal{B}$ may be decomposed as

$$B \;=\; \bigcup_{z \in \mathbb{Z}} B_z \quad \text{with} \quad B_z \;:=\; B \cap \{\gamma^z \leq f < \gamma^{z+1}\}.$$

By assumption, for any integer z there exists an open set U_z such that $B_z \subset U_z \subset \{f < \gamma^{z+1}\}$ and $\mu(U_z) \leq \gamma\mu(B_z)$. Consequently, $U := \bigcup_{z \in \mathbb{Z}} U_z$ is an open set containing B such that

$$\tilde{\mu}(U) \;\leq\; \sum_{z \in \mathbb{Z}} \tilde{\mu}(U_z) \;\leq\; \sum_{z \in \mathbb{Z}} \gamma^{z+1} \mu(U_z) \;\leq\; \sum_{z \in \mathbb{Z}} \gamma^{z+2} \mu(B_z) \;\leq\; \sum_{z \in \mathbb{Z}} \gamma^2 \tilde{\mu}(B_z) \;=\; \gamma^2 \tilde{\mu}(B).$$

Hence, $\tilde{\mu}$ satisfies (5.2), too.

It remains to verify (5.9) for $\tilde{\mu}$ in case of $f(x) = J(x)^{-1}$ and μ being a left Haar measure on $\mathcal{X}$. Indeed, for measurable functions $h : \mathcal{X} \to [0, \infty)$ and $y \in \mathcal{X}$,

$$\begin{aligned}
\int h(xy)\, \tilde{\mu}(dx) \;&=\; \int h(xy) J(x)^{-1}\, \mu(dx) \\
&=\; J(y) \int h(xy) J(xy)^{-1}\, \mu(dx) \\
&=\; \int h(x) J(x)^{-1}\, \mu(dx) \\
&=\; \int h(x)\, \tilde{\mu}(dx),
\end{aligned}$$

where the second last step follows from Exercise 5.11. □

5.3 Some Explicit Constructions

In this section we present a few examples of Haar measure.

Discrete groups. Let $(\mathcal{X}, \cdot)$ be an arbitrary group. With the metric $d(x, y) := 1_{[x \neq y]}$, the topology $\mathcal{U}$ and the Borel σ-field $\mathcal{B}$ coincide with $\mathcal{P}(\mathcal{X})$, and $\mathcal{C}$ is the family of finite subsets of $\mathcal{X}$. Thus $(\mathcal{X}, \cdot, d)$ is a locally compact topological group. Here, the counting measure μ,

$$\mu(B) \ := \ \#B,$$

is a left and right Haar measure on $\mathcal{X}$. This follows essentially from the fact that for any fixed $x_o \in \mathcal{X}$, the mappings $x \mapsto x_o x$ and $x \mapsto x x_o$ are bijective from $\mathcal{X}$ to $\mathcal{X}$, see Exercise5.1 (b).

Lebesgue measure on $\mathbb{R}^d$. Lebesgue measure Leb on (the Borel subsets of) $\mathbb{R}^d$ satisfies

$$\mathrm{Leb}(x + B) \ = \ \mathrm{Leb}(B)$$

for arbitrary $x \in \mathbb{R}^d$ and $B \in \mathrm{Borel}(\mathbb{R}^d)$. Consequently, it is the unique Haar measure μ on the additive group $\mathbb{R}^d$ such that $\mu([0, 1]^d) = 1$.

Haar measure on linear groups. Let $(\mathcal{X}, \cdot)$ be the set of nonsingular matrices $\boldsymbol{x} \in \mathbb{R}^{d \times d}$ with matrix multiplication. As mentioned before, $\mathcal{X}$ is an open subset of $\mathbb{R}^{d \times d}$. For any fixed $\boldsymbol{x}_o \in \mathcal{X}$,

$$L_{\boldsymbol{x}_o}(\boldsymbol{x}) \ := \ \boldsymbol{x}_o \boldsymbol{x} \quad \text{and} \quad R_{\boldsymbol{x}_o}(\boldsymbol{x}) \ := \ \boldsymbol{x} \boldsymbol{x}_o$$

define bijective mappings from $\mathcal{X}$ to $\mathcal{X}$. They may also be viewed as bijective linear mappings from $\mathbb{R}^{d \times d}$ to $\mathbb{R}^{d \times d}$. If we identify $R^{d \times d}$ with $\mathbb{R}^{d^2}$ in the usual fashion, then the determinant of these linear mappings equals

$$\det(L_{\boldsymbol{x}_o}) \ = \ \det(R_{\boldsymbol{x}_o}) \ = \ \det(\boldsymbol{x}_o)^d,$$

see Exercise 5.12. With Lebesgue measure Leb on $\mathbb{R}^{d \times d}$, this means that for any measurable function $h : \mathcal{X} \to [0, \infty)$,

$$\int_{\mathcal{X}} h(\boldsymbol{x}_o \boldsymbol{x}) \, d\boldsymbol{x} \ = \ \int_{\mathcal{X}} h(\boldsymbol{x} \boldsymbol{x}_o) \, d\boldsymbol{x} \ = \ |\det(\boldsymbol{x}_o)|^{-d} \int_{\mathcal{X}} h(\boldsymbol{x}) \, d\boldsymbol{x},$$

where $d\boldsymbol{x}$ stands for $\mathrm{Leb}(d\boldsymbol{x})$. This implies that

$$\mu(B) \ := \ \int_B |\det(\boldsymbol{x})|^{-d} \, d\boldsymbol{x}$$

defines a left and right Haar measure on $\mathcal{X}$. For if $\boldsymbol{x}_o \in \mathcal{X}$, then

$$\begin{aligned} \mu(\boldsymbol{x}_o^{-1} B) \ &= \ \int_{\mathcal{X}} 1_B(\boldsymbol{x_o x}) |\det(\boldsymbol{x})|^{-d} \, d\boldsymbol{x} \\ &= \ |\det(\boldsymbol{x}_o)|^d \int_{\mathcal{X}} 1_B(\boldsymbol{x}_o \boldsymbol{x}) |\det(\boldsymbol{x}_o \boldsymbol{x})|^{-d} \, d\boldsymbol{x} \\ &= \ \int_{\mathcal{X}} 1_B(\boldsymbol{x}) |\det(\boldsymbol{x})|^{-d} \, d\boldsymbol{x} \\ &= \ \mu(B), \end{aligned}$$

and analogously one can show that $\mu(B \boldsymbol{x}_o^{-1}) = \mu(B)$.

Exercise 5.12. Let $\mathrm{vec} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d^2}$ be given by

$$\mathrm{vec}(\boldsymbol{x}) \;:=\; (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_d^\top)^\top$$

for $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d]$ with columns $\boldsymbol{x}_j \in \mathbb{R}^d$. Show that for any fixed $\boldsymbol{x}_o \in \mathcal{X}$ and arbitrary $\boldsymbol{x} \in \mathbb{R}^{d\times d}$,

$$\mathrm{vec}(\boldsymbol{x}_o\boldsymbol{x}) \;=\; \boldsymbol{L}_{\boldsymbol{x}_o}\mathrm{vec}(\boldsymbol{x}) \quad \text{and} \quad \mathrm{vec}(\boldsymbol{x}\boldsymbol{x}_o) \;=\; \boldsymbol{R}_{\boldsymbol{x}_o}\mathrm{vec}(\boldsymbol{x})$$

with matrices $\boldsymbol{L}_{\boldsymbol{x}_o}, \boldsymbol{R}_{\boldsymbol{x}_o} \in \mathbb{R}^{d^2\times d^2}$ such that

$$\det(\boldsymbol{L}_{\boldsymbol{x}_o}) \;=\; \det(\boldsymbol{x}_o)^d \;=\; \det(\boldsymbol{R}_{\boldsymbol{x}_o}).$$

A general consideration about compact topological groups. Suppose that $\mathcal{X}$ is a compact topological group, and suppose that P is a left-invariant probability measure on $\mathcal{X}$, that means,

$$P(xB) \;=\; P(B) \quad \text{for all } x \in \mathcal{X} \text{ and } B \in \mathcal{B}.$$

Then P is also right-invariant and inversion-invariant, that means,

$$P(xB) \;=\; P(Bx) \;=\; P(B^{-1}) \;=\; P(B) \quad \text{for all } x \in \mathcal{X} \text{ and } B \in \mathcal{B}.$$

If $\tilde{P}$ is another left-invariant probability measure on $\mathcal{X}$, then $\tilde{P} \equiv P$.

Proof. Let X and Y be stochastically independent random variables on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(\mathcal{X}, \mathcal{B})$, where $\mathbb{P}^X = P$ while $\mathbb{P}^Y$ is an arbitrary distribution Q. Then $Y^{-1}X$ has distribution P, too, because for any $B \in \mathcal{B}$,

$$\mathbb{P}(Y^{-1}X \in B) \;=\; \mathbb{P}(X \in YB) \;=\; \mathbb{E}\underbrace{\mathbb{P}(X \in YB \,|\, \sigma(Y))}_{=P(YB)=P(B) \text{ a.s.}} \;=\; P(B).$$

In particular, if $Q = P$, then for any $B \in \mathcal{B}$,

$$P(B) \;=\; \mathbb{P}(Y^{-1}X \in B) \;=\; \mathbb{P}(X^{-1}Y \in B^{-1}) \;=\; \mathbb{P}(Y \in XB^{-1}) \;=\; P(B^{-1}).$$

But this implies right-invariance of P, because for all $x \in \mathcal{X}$ and $B \in \mathcal{B}$,

$$P(Bx) \;=\; P((Bx)^{-1}) \;=\; P(x^{-1}B^{-1}) \;=\; P(B^{-1}) \;=\; P(B).$$

Finally, if Q is left-invariant itself, then for all $B \in \mathcal{B}$,

$$Q(B) \;=\; \mathbb{P}(X^{-1}Y \in B) \;=\; \mathbb{P}(Y^{-1}X \in B^{-1}) \;=\; P(B^{-1}) \;=\; P(B).$$

□

Haar measure on orthogonal groups. Let $\mathcal{X}$ and $\mathcal{X}_{\mathrm{orth}}$ be the groups of invertible and orthogonal matrices in $\mathbb{R}^{d\times d}$, respectively, the binary operation being matrix multiplication. In view of the previous consideration, it suffices to construct a random variable $\boldsymbol{X}$ with values in $\mathcal{X}_{\mathrm{orth}}$ such that

$$\mathbb{P}(\boldsymbol{X} \in \boldsymbol{x}B) \;=\; \mathbb{P}(\boldsymbol{X} \in B) \quad \text{for all } \boldsymbol{x} \in \mathcal{X}_{\mathrm{orth}} \text{ and } B \in \mathcal{B}.$$

Then the distribution $P := \mathbb{P}^{\boldsymbol{X}}$ is the unique left Haar probability measure on $\mathcal{X}_{\mathrm{orth}}$, and it happens to be left-, right- and inversion-invariant.

A particular construction starts with a random matrix $\boldsymbol{Z} = (Z_{ij})_{i,j=1}^{d}$ with d^2 stochastically independent random variables $Z_{ij} \sim \mathcal{N}(0,1)$. Writing $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_d]$, the columns $\boldsymbol{Z}_j$ are stochastically independent with standard Gaussian distribution $\mathcal{N}_d(\boldsymbol{0}, \boldsymbol{e})$. For $2 \le k \le d$,

$$\mathbb{P}\big(\boldsymbol{Z}_k \in \operatorname{span}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{k-1})\big) \;=\; \mathbb{E}\,\mathbb{P}\big(\boldsymbol{Z}_k \in \operatorname{span}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{k-1}) \,\big|\, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{k-1}\big) \;=\; 0,$$

because $\mathbb{P}(\boldsymbol{Z}_k \in \mathbb{V}) = 0$ for any fixed linear space $\mathbb{V} \subset \mathbb{R}^d$ with $\dim(\mathbb{V}) < d$. Consequently, the matrix $\boldsymbol{Z}$ is invertible almost surely. This allows us to apply the mapping $\Psi : \mathcal{X} \to \mathcal{X}_{\mathrm{orth}}$, $\Psi(\boldsymbol{z}) := \boldsymbol{z}(\boldsymbol{z}^\top \boldsymbol{z})^{-1/2}$. Defining $\Psi(\boldsymbol{z}) := \boldsymbol{e}$ for singular matrices $\boldsymbol{z} \in \mathbb{R}^{d\times d}$, the distribution P of the random matrix

$$\boldsymbol{X} \;:=\; \Psi(\boldsymbol{Z})$$

is left-invariant. Indeed, it is well-known from probability theory that the distribution of $\boldsymbol{Z}$ does not change if we multiply it from the left with an arbitrary matrix $\boldsymbol{u} \in \mathcal{X}_{\mathrm{orth}}$. Hence, P is also the distribution of

$$\Psi(\boldsymbol{u}\boldsymbol{Z}) \;=\; \boldsymbol{u}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{u}^\top \boldsymbol{u}\boldsymbol{Z})^{-1/2} \;=\; \boldsymbol{u}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2} \;=\; \boldsymbol{u}\boldsymbol{X}.$$

Another explicit construction would be to apply the Gram–Schmidt orthogonalization procedure to the columns of $\boldsymbol{Z}$. That means, we define $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d]$ via

$$\boldsymbol{X}_1 \;:=\; \|\boldsymbol{Z}_1\|^{-1} \boldsymbol{Z}_1$$

and inductively

$$\boldsymbol{X}_k \;:=\; \Big\|\boldsymbol{Z}_k - \sum_{i=1}^{k-1} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{Z}_k\Big\|^{-1} \Big(\boldsymbol{Z}_k - \sum_{i=1}^{k-1} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{Z}_k\Big)$$

for $k = 2, \ldots, d$. Again, one can show that replacing $\boldsymbol{Z}$ with $\boldsymbol{u}\boldsymbol{Z}$ for some $\boldsymbol{u} \in \mathcal{X}_{\mathrm{orth}}$ results in replacing $\boldsymbol{X}$ with $\boldsymbol{u}\boldsymbol{X}$.

Note that the construction via Gram–Schmidt implies that for a random matrix $\boldsymbol{X} \in \mathcal{X}_{\mathrm{orth}}$ with left-invariant distribution, any column is uniformly distributed on the unit sphere of $\mathbb{R}^d$ (with respect to the standard Euclidean norm).

Part II

Mathematical Statistics

Chapter 6

Measurement Series and Estimators of Location

6.1 Statistical Experiments and Point Estimators

Before we discuss estimation of a location parameter, let us intruduce some general terminology.

Definition 6.1 (Statistical experiment). A *statistical experiment* is a triplet $\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ consisting of a measurable space $(\Omega, \mathcal{A})$, the *sample space*, and a family of probability distributions $\mathbb{P}_\theta$ on $(\Omega, \mathcal{A})$, depending on a *parameter* θ in a *parameter space* Θ.

The sample space $(\Omega, \mathcal{A})$ represents all possible data sets one could observe. The elements of the parameter space Θ represent potential values of an *unknown true parameter* $\theta \in \Theta$. We assume that the observed data are a (realization of a) random variable with distribution $\mathbb{P}_\theta$. 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 be denoted by a corresponding subscript, leading to $\mathbb{P}_\theta$, $\mathbb{E}_\theta$, Var_θ etc.

Of course there is some redundancy in the definition of a statistical experiment $\mathcal{E}$, because specifying the family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ implies the specification of the measurable space $(\Omega, \mathcal{A})$. But sometimes we shall replace $\mathcal{A}$ with certain sub-σ-fields, so the current definition is useful.

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 observed data $\omega \in \Omega$ a simple guess $\widehat{g}(\omega) \in \mathbb{G}$ of $g(\theta)$.

Definition 6.2 (Point estimator). A *point estimator of* $g(\theta)$ (short: an *estimator*) is a measurable[1] function

$$\widehat{g} : \Omega \to \mathbb{G}.$$

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

[1] $\mathcal{A}$-Borel$(\mathbb{R}^q)$-measurable

squared error,

$$\mathbb{E}_\theta\big(d(\widehat{g}, g(\theta))^2\big).$$

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 in another region.

6.2 Estimators of Location

Simple location families. Let $\theta \in \mathbb{R}$ be an unknown parameter which has to be estimated in an experiment which yields n single measurements $X_1, \ldots, X_n$. Suppose that

$$X_i \;=\; \theta + \epsilon_i, \quad 1 \le i \le n,$$

with independent random measurement errors $\epsilon_1, \ldots, \epsilon_n$ with a known distribution P_0. This leads to the statistical experiment

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

with

$$\mathbb{P}_\theta \;:=\; P_\theta^{\otimes n} \quad \text{and} \quad P_\theta \;=\; P_0 \star \delta_\theta.$$

Here '$\star$' denotes convolution[2], and δ_θ is the Dirac measure at the point θ. In other words, $\mathbb{P}_\theta$ describes the distribution of a random vector with independent components having distribution function F_θ, where

$$F_\theta(x) \;:=\; F_0(x-\theta),$$

and F_0 is the distribution function of the error distribution P_0. Such a statistical experiment is called a *simple location family*.

In what follows, we write $\boldsymbol{x}$ instead of ω for a sample in $\mathbb{R}^n$. The random variables X_i are just the coordinate functions $\boldsymbol{x} = (x_i)_{i=1}^n \mapsto X_i(\boldsymbol{x}) := x_i$, and $\boldsymbol{X} := (X_i)_{i=1}^n$ is the identity function.

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

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

[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 $X \sim P$.

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 $\int x\, P_0(dx) = 0$,

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

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

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

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.

6.3 Constructing an Optimal Equivariant Estimator

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

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

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 6.4 (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 6.5 (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 6.4 carefully reveals that Theorem 6.4 remains valid with this definition $\boldsymbol{T}$. Our particular version corresponds to $\tilde{\theta} = X_1$ and is convenient for explicit calculations.

Exercise 6.6. 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 6.7 (Characterization of optimality). The particular construction in Theorem 6.4 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 6.8. Let $\widehat{\theta}_* : \mathbb{R}^n \to \mathbb{R}$ be an optimal equivariant estimator for θ. Show that $\widehat{\theta}_*$ is unbiased, that means,

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

Exercise 6.9. 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 6.4. The general theory in Sections 3.2 and 3.3 shows that $\mathbb{E}_0(\widehat{\theta} \,|\, \sigma(\boldsymbol{T})) = g_*(\boldsymbol{T})$ 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

$$\mathbb{E}_0(\widehat{\theta}^2) \;=\; \mathbb{E}_0\big((\widehat{\theta} - g_*(\boldsymbol{T}))^2\big) + \mathbb{E}_0(g_*(\boldsymbol{T})^2). \tag{6.1}$$

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 (6.1) 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 $\theta \in \mathbb{R}$. □

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 6.10. *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*

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

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

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

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

Example 6.11 (Gaussian distributions). Suppose that $P_0 = \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Then

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

This follows from Corollary 6.10 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 6.10 is equal to $\mathcal{N}(\bar{x}, \sigma^2/n)$. Hence

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

Example 6.12 (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 $f_0(x) = (2\sigma)^{-1}1_{[-\sigma\le x\le\sigma]}$,

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

Hence the distribution $Q_{\boldsymbol{x}}$ in Corollary 6.10 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,

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

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

Exercise 6.13. Suppose that $P_0 = \mathrm{Unif}[-\sigma, \sigma]$. As shown before, the optimal equivariant estimator of θ is given by $\widehat{\theta}_*(\boldsymbol{x}) = \bigl(\min(\boldsymbol{x}) + \max(\boldsymbol{x})\bigr)/2$.

(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 6.14 (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 $Q_{\boldsymbol{X}} \approx \mathcal{N}(\widehat{\theta}_{\mathrm{ML}}, \widehat{\gamma}^2)$ for large sample sizes n, and $\widehat{\theta}_{\mathrm{ML}}$ seems to be a good surrogate for $\widehat{\theta}_*$.

Exercise 6.15. 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 6.10. 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,

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

for any Borel set $B \subset \mathbb{R}^n$. This shows that the distribution of $\boldsymbol{AX}$ 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 $\boldsymbol{AX} = (X_1, T_2, \ldots, T_n)^\top$ while $T_1 \equiv 0$. Thus the considerations in Example 4.10 imply the following: 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

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

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

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

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

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 6.16 (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 $P_0 = \mathcal{N}(0,\sigma^2)$ for some $\sigma \geq 0$.*

Proof of Theorem 6.16. 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 6.4 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,

$$\mathbb{E}\Big(\sum_{i=1}^n X_i g(\boldsymbol{T})\Big) \;=\; 0 \quad \text{whenever } g(\boldsymbol{T}) \in L^2(\mathbb{P}). \tag{6.2}$$

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

$$\mathbb{E}\big((X_1 + X_2 + X_3)h(T_2,T_3)\big) \;=\; 0 \quad \text{whenever } h(T_2,T_3) \in L^2(\mathbb{P}). \tag{6.3}$$

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 (6.3) 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 6.17. 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 $\mathcal{N}(\mu, \sigma^2)$ is given by $t \mapsto \exp(\boldsymbol{i}t\mu - \sigma^2 t^2/2)$. Hence the characteristic function of P_0 coincides with the characteristic function of $\mathcal{N}(0, \sigma^2)$, where $\sigma := \sqrt{-\alpha}$. Since any probability distribution is uniquely determined by its characteristic distribution, this shows that $P_0 = \mathcal{N}(0, \sigma^2)$. $\square$

Exercise 6.17. 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

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

Exercise 6.18 (Distribution of order statistics). The contents of this exercise are probably known from other courses in Statistics. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be the order statistics of independent random variables $X_1, \ldots, X_n$ with distribution function F on $\mathbb{R}$.

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

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

where $B_{n,p}(\cdot)$ denotes the distribution function of the binomial distribution $\mathrm{Bin}(n, p)$.

(b) Show that for $c \in \{0, 1, \ldots, n-1\}$, $B_{n,0}(c) = 1$, $B_{n,1}(c) = 0$ and

$$B_{n,p}(c) \;=\; n\binom{n-1}{c} \int_p^1 u^c (1-u)^{n-1-c}\, du \;\text{ for } p \in [0,1].$$

(c) Deduce from parts (a-b) that

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

Exercise 6.19 (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 $m \ge 1$.

(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 $n\,\mathbb{E}(M_n^2) \to \pi^2/4$.

6.4 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 $\mathcal{N}(\nu, \tau^2)$ 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 6.20 (Gaussian model and prior). Suppose that π is the density of $\mathcal{N}(0, \tau^2)$ for some $\tau > 0$, and let $P_0 = \mathcal{N}(0, \sigma^2)$ 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, $Q^{\mathrm{B}}_{\boldsymbol{x}} = \mathcal{N}(\beta\bar{x}, \gamma^2)$, 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 6.21 (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 6.22. 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 $R(\tilde{\theta},\cdot) \le R(\widehat{\theta},\cdot)$, 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 6.23. 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 $\max(\boldsymbol{X})-1\le\theta\le\min(\boldsymbol{X})+1$ 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 6.24. *If $P_0 = \mathcal{N}(0,\sigma^2)$ for some $\sigma > 0$, then $\bar{X}$ is an admissible estimator of θ.*

Proof of Theorem 6.24. The risk function of $\bar{X}$ is constant σ^2/n. Suppose that $\widehat{\theta}$ is an arbitrary estimator such that $R(\widehat{\theta},\cdot)\le\sigma^2/n$ on the whole real line. As shown in Exercises 6.25 and 6.26, 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 $R(\widehat{\theta},\theta)\le\sigma^2/n-\delta$ for $\theta\in[a,b]$. Now we evaluate the performance of $\widehat{\theta}$ in a Bayesian framework

with $\theta \sim \mathcal{N}(0, \tau^2)$ 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 6.25 (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 6.26 (Continuity of risk functions in simple Gaussian location families). Consider the simple location family with $P_0 = \mathcal{N}(0, \sigma^2)$ 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 6.27. Let $Z \sim \mathcal{N}(0,1)$ 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 6.28. Suppose that the error distribution equals $P_0 = \mathcal{N}(0, \sigma^2)$ 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 6.27.)

(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 $\Big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), \big(\mathcal{N}(\theta, \sigma^2)^{\otimes n}\big)_{\theta \geq 0}\Big)$, because $\bar{X}^+$ has strictly smaller risk than $\bar{X}$. Whether or not $\bar{X}^+$ is admissible itself is a different question.

6.5 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

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

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

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

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 6.29 (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, $\mathrm{mean}(\cdot)$ is an equivariant location functional on $\mathcal{P}^1$, and $\mathrm{median}(\cdot)$ 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

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

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 6.30 (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 6.30. 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

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

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

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

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 $S(P^{(1)}) \leq -x_\epsilon$ or $S(P^{(2)}) \geq x_\epsilon$, whence

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

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

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

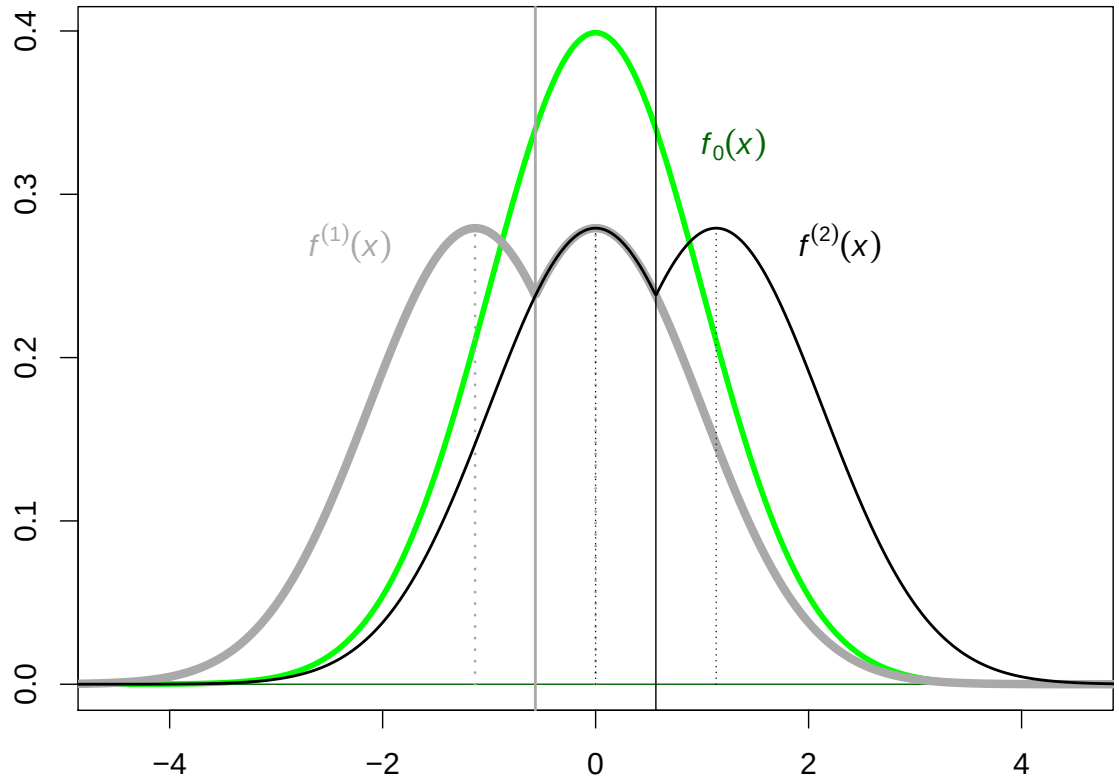

Figure 6.1: Construction of two particular distributions $P^{(1)}, P^{(2)} \in \mathcal{U}_{0.3}(0)$ with densities $f^{(1)}, f^{(2)}$ in case of $P_0 = \mathcal{N}(0,1)$.

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

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

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

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

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

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

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 6.1 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 $n \to \infty$.

Chapter 7

Statistical Tests

In this chapter we consider a general statistical experiment (also called statistical model)

$$\big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$$

consisting of a sample space $(\Omega, \mathcal{A})$, a parameter space Θ and given probability distributions $\mathbb{P}_\theta$ on $(\Omega, \mathcal{A})$ for arbitrary parameters $\theta \in \Theta$.

Recall that $(\mathbb{P}_\theta)_{\theta\in\Theta}$ describes potential distributions of the observed data $\omega \in \Omega$. Suppose for the moment that the observed data are indeed a realization of a random variable with distribution $\mathbb{P}_\theta$ 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

$$\varphi : \Omega \to \{0, 1\}.$$

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

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

$$\varphi : \Omega \to [0, 1].$$

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

$$\mathbb{E}_\theta(\varphi) \;=\; \int \varphi \, d\,\mathbb{P}_\theta \,.$$

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

Definition 7.1 (Statistical test, power function). A *(statistical) test* is a measurable mapping $\varphi : \Omega \to [0, 1]$. 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

$$\Theta \ni \theta \;\mapsto\; \mathbb{E}_\theta(\varphi) = \int \varphi \, d\,\mathbb{P}_\theta,$$

and $\mathbb{E}_\theta(\varphi)$ 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 ω coming from $\mathbb{P}_\theta$ 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

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

whereas an error of the second kind occurs with probability

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

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

Definition 7.2 (Test level). Let $\emptyset \subsetneq \Theta_o \subsetneq \Theta$, and let $\alpha \in (0,1)$. Suppose that φ is a test such that

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

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*.

Example 7.3 (Quality control). The producer of a certain gadget wants to learn something about the unknown probability p 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 $\omega = (\omega_i)_{i=1}^n$ in $\{0,1\}^n$, where ω_i specifies whether the i-th gadget fails ($\omega_i = 1$) or not ($\omega_i = 0$). Assuming that the n gadgets perform independently, this leads to the statistical model

$$\big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathbb{P}_p)_{p\in[0,1]}\big)$$

with $\mathbb{P}_p$ given by

$$\mathbb{P}_p(\{\omega\}) \;=\; \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i} \;=\; p^{T(\omega)}(1-p)^{n-T(\omega)}.$$

Here $T(\omega) := \sum_{i=1}^n \omega_i$ is the total number of failures in the experiment.

Alternatively, the producer could focus immediately on the total number of failures in his experiment. Indeed, in a later chapter it will be shown that this reduction is well justified. This leads to the statistical model

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

Suppose the producer wants to verify that the unknown probability p is smaller than a given (small) number p_o. Then he should test the null hypothesis $\Theta_o = [p_o, 1]$. 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 p is smaller than p_o.

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 $\mathbb{E}_\theta(\varphi)$ for $\theta \in \Theta_A$.

Exercise 7.4 (De-randomisation). Let $\phi : \Omega \to [0, 1]$ be a statistical test. Show that for any fixed $\beta \in (0, 1)$,

$$\tilde{\phi} \;:=\; 1_{[\phi \geq \beta]}$$

is a non-randomized test satisfying

$$\frac{\mathbb{E}_\theta(\phi)}{\beta} \;\geq\; \mathbb{E}_\theta(\tilde{\phi}) \;\geq\; \frac{\mathbb{E}_\theta(\phi) - \beta}{1 - \beta} \quad \text{for all } \theta \in \Theta.$$

7.1 The Neyman–Pearson Lemma

We start with the very simple setting of $\Theta = \{0, 1\}$ and $\Theta_o = \{0\}$.

Theorem 7.5 (Neyman–Pearson). *Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ have densities f_0 and f_1, respectively, with respect to some measure M on $(\Omega, \mathcal{A})$. For any $\alpha \in (0, 1)$ there exist constants $k_\alpha \geq 0$ and $\gamma_\alpha \in [0, 1]$ such that*

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

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

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

$$\mathbb{E}_0(\varphi_*) \;=\; \alpha.$$

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

$$\mathbb{E}_1(\varphi) \;\leq\; \mathbb{E}_1(\varphi_*).$$

(iii) If φ is a level-α test of $\{0\}$ with $\mathbb{E}_1(\varphi) = \mathbb{E}_1(\varphi_*)$, 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 $\mathbb{E}_0(\varphi) = \alpha$.

(iv) If $\mathbb{P}_0 \neq \mathbb{P}_1$, then

$$\mathbb{E}_1(\varphi_*) \;>\; \alpha.$$

Note that the optimal level-α test φ_* 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 $0/0 := 0$:

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

Remark 7.6 (Existence and choice of M). The assumption that $\mathbb{P}_0$ and $\mathbb{P}_1$ have densities with respect to some measure M on $(\Omega, \mathcal{A})$ is not a real restriction. If we take $M^o := \mathbb{P}_0 + \mathbb{P}_1$, then it follows from the theorem of Radon–Nikodym that there exist densities $f_\theta^o = d\,\mathbb{P}_\theta\,/dM^o$ for $\theta = 0, 1$. If M is an arbitrary measure such that a density $f_\theta = d\,\mathbb{P}_\theta\,/dM$ 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_* would be essentially the same, no matter which measure M we start from.

Proof of Theorem 7.5. For the construction of our special test φ_* we consider the auxiliary function $H : [0, \infty) \to [0, 1]$ given by

$$H(r) \;=\; \mathbb{P}_0(f_1 \le r f_0).$$

One can easily verify that $H(r) = \mathbb{P}_0(f_1/f_0 \le r)$, where $a/0 := \infty$ for $a > 0$ and $0/0 := 0$. Since $\mathbb{P}_0(f_1/f_0 = \infty) \le \mathbb{P}_0(f_0 = 0) = 0$, H is a distribution function on $[0, \infty)$. That means, 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

$$\mathbb{P}_0(f_1 > k_\alpha f_0) \;=\; 1 - H(k_\alpha) \;\le\; \alpha \;\le\; 1 - H(k_\alpha -) \;=\; \mathbb{P}_0(f_1 \ge k_\alpha f_0).$$

If $\mathbb{P}_0(f_1 = k_\alpha f_0) = 0$, we set $\gamma_\alpha := 1$. Otherwise we define

$$\gamma_\alpha \;:=\; \frac{\alpha - \mathbb{P}_0(f_1 > k_\alpha f_0)}{\mathbb{P}_0(f_1 \ge k_\alpha f_0) - \mathbb{P}_0(f_1 > k_\alpha f_0)} \;=\; \frac{\alpha - \mathbb{P}_0(f_1 > k_\alpha f_0)}{\mathbb{P}_0(f_1 = k_\alpha f_0)} \;\in\; [0, 1].$$

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

$$\mathbb{E}_0(\varphi_*) \;=\; \mathbb{P}_0(f_1 = k_\alpha f_0)\gamma_\alpha + \mathbb{P}_0(f_1 > k_\alpha f_0) \;=\; \alpha.$$

This proves property (i).

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

$$(\varphi - \varphi_*)\,(f_1 - k_\alpha f_0) \;\le\; 0,$$

because

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

Consequently,

$$\begin{aligned} 0 \;&\geq\; \int (\varphi - \varphi_*)(f_1 - k_\alpha f_0)\, dM \\ &=\; \mathbb{E}_1(\varphi) - \mathbb{E}_1(\varphi_*) - k_\alpha\big(\mathbb{E}_0(\varphi) - \mathbb{E}_0(\varphi_*)\big) \\ &=\; \mathbb{E}_1(\varphi) - \mathbb{E}_1(\varphi_*) - k_\alpha\big(\mathbb{E}_0(\varphi) - \alpha\big). \end{aligned}$$

In other words, for any test φ,

$$\mathbb{E}_1(\varphi) - \mathbb{E}_1(\varphi_*) \;\leq\; k_\alpha\big(\mathbb{E}_0(\varphi) - \alpha\big) \tag{7.1}$$

with equality if and only if

$$M(f_1 > k_\alpha f_0 \text{ and } \varphi < 1) \;=\; 0 \;=\; M(f_1 < k_\alpha f_0 \text{ and } \varphi > 0). \tag{7.2}$$

If φ is a level-α test of $\{0\}$, then the right hand side of (7.1) is non-positive, so $\mathbb{E}_1(\varphi) \leq \mathbb{E}_1(\varphi_*)$. This proves property (ii).

If φ is a level-α test of $\{0\}$ with $\mathbb{E}_1(\varphi) = \mathbb{E}_1(\varphi_*)$, then the right hand side of (7.1) has to be zero, and the first half of property (iii) is just (7.2). Moreover, if $k_\alpha > 0$, then the right hand side of (7.1) being zero means that $\mathbb{E}_0(\varphi) = \alpha$, which proves the second half of property (iii).

Finally, we may compare φ_* with the trivial test $\varphi \equiv \alpha$, so $\mathbb{E}_0(\varphi) = \mathbb{E}_1(\varphi) = \alpha$. Then (7.1) and (7.2) show that $\mathbb{E}_1(\varphi_*) \geq \alpha$ with equality if and only if

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

That means, $\mathbb{P}_1$ has density $k_\alpha f_0$ with respect to M. But then $1 = \mathbb{P}_1(\Omega) = k_\alpha\, \mathbb{P}_0(\Omega) = k_\alpha$, so $\mathbb{P}_1 = \mathbb{P}_0$. This proves property (iv). □

Example 7.7. Let $\Omega = (0, \infty)$ and $\mathbb{P}_\theta := \mathrm{Gamma}(a_\theta, b)$ with shape parameters $a_1 > a_0 > 0$ and a common scale parameter $b > 0$. Then the density f_θ of $\mathbb{P}_\theta$ with respect to Lebesgue measure on Ω equals $f_\theta(x) = \Gamma(a_\theta)^{-1} b^{-a_\theta} x^{a_\theta - 1} e^{-x/b}$, so

$$\frac{f_1}{f_0}(x) \;=\; \frac{\Gamma(a_0) b^{a_0 - a_1}}{\Gamma(a_1)}\, x^{a_1 - a_0}$$

is strictly increasing in $x > 0$. Hence the optimal level-α test of $\{0\}$ versus $\{1\}$ – in other words: of $\mathrm{Gamma}(a_0, b)$ versus $\mathrm{Gamma}(a_1, b)$ – is given by

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

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

Exercise 7.8. Let $\mathbb{P}_0 = \mathcal{N}(0, \sigma^2)$ with $\sigma \leq \sqrt{2}$. The corresponding distribution function is $F_0(x) = \Phi(x/\sigma)$. Further let $\mathbb{P}_1$ be the standard logistic distribution with distribution function

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

Show that the Neyman–Pearson test of $\mathbb{P}_0$ versus $\mathbb{P}_1$ (i.e. of $\{0\}$ versus $\{1\}$) at level $\alpha \in (0,1)$ is given by

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

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 $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively.

Exercise 7.9. Let be $\mathbb{P}_0 = \mathcal{N}(0,1)$ and $\mathbb{P}_1 = \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 1$.

(a) Show that the Neyman–Pearson test of $\mathbb{P}_0$ versus $\mathbb{P}_1$ at level α has the form

$$\varphi_*(x) \;=\; 1_{\left[x \notin (\mu/(1-\sigma^2) \pm \delta_*)\right]}$$

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

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

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

$$\int \varphi_* \, d\mathcal{N}(\mu, \sigma^2) \;\geq\; \int \varphi_* \, d\mathcal{N}(0, \sigma^2) \quad \text{for arbitrary } \mu \in \mathbb{R}.$$

Determine the latter power.

7.2 Monotone Density Ratios

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

$$T : \Omega \to \mathbb{R}$$

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

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

Example 7.10 (Bernoulli experiments). Motivated by Example 7.3, let $\Omega = \{0,1\}^n$ and $\Theta = (0,1)$, and let $\mathbb{P}_\theta$ 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 Ω, $\mathbb{P}_\theta$ has density

$$f_\theta(\omega) \;=\; \prod_{i=1}^n \theta^{\omega_i}(1-\theta)^{1-\omega_i} \;=\; \theta^{T(\omega)}(1-\theta)^{n-T(\omega)}$$

with $T(\omega) := \sum_{i=1}^n \omega_i$. Then for $0 < \theta_1 < \theta_2 < 1$ and $\omega \in \Omega$,

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

with and

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

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 7.11 (Gaussian location family). Let $\Omega = \mathbb{R}^n$, $\Theta = \mathbb{R}$ and $\mathbb{P}_\theta = \mathcal{N}(\theta,\sigma^2)^{\otimes n}$ for a fixed standard deviation $\sigma > 0$. Recall that the density f_θ of $\mathbb{P}_\theta$ 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 7.12 (Gamma families). As in Example 7.7, let $\Omega = (0,\infty)$, and let $\mathrm{Gamma}(a,b)$ be the gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Its density with respect to Lebesgue measure on Ω equals $f_{a,b}(\omega) := f_a(\omega/b)/b$ with $f_a(y) := \Gamma(a)^{-1}\omega^{a-1}e^{-\omega}$. Hence, for parameters $(a_1,b_1),(a_2,b_2) \in (0,\infty)\times(0,\infty)$,

$$\frac{f_{a_2,b_2}}{f_{a_1,b_1}}(\omega) \;=\; \frac{\Gamma(a_1)b_1^{a_1}}{\Gamma(a_2)b_2^{a_2}}\,\omega^{a_2-a_1}\exp((1/b_1-1/b_2)\omega),$$

which is strictly increasing in $T(\omega) := \omega$ whenever $a_1 \le a_2$, $b_1 \le b_2$ and $(a_1,b_1) \ne (a_2,b_2)$. Consequently, our assumption is satisfied if, for instance, $\Theta = (0,\infty)$ and $\mathbb{P}_\theta = \mathrm{Gamma}(\theta,b)$ for a given $b > 0$ or $\mathbb{P}_\theta = \mathrm{Gamma}(a,\theta)$ for a given $a > 0$.

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 $\theta_o \in \Theta$.

Theorem 7.13 (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*

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

satisfies

$$\mathbb{E}_{\theta_o}(\varphi_*) \;=\; \alpha.$$

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

(ii.1) The power function $\theta \mapsto \mathbb{E}_\theta(\varphi_*)$ is non-decreasing on Θ with values in $(0,1)$. In particular,

$$\mathbb{E}_\theta(\varphi_*) \;\le\; \alpha \quad \text{for all } \theta \in \Theta \cap (-\infty, \theta_o].$$

(ii.2) For any test φ with $\mathbb{E}_{\theta_o}(\varphi) \le \alpha$,

$$\mathbb{E}_\theta(\varphi) \;\le\; \mathbb{E}_\theta(\varphi_*) \quad \text{for all } \theta \in \Theta \cap (\theta_o, \infty).$$

(ii.3) For arbitrary parameters $\theta_1 < \theta_2$ with $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$,

$$\mathbb{E}_{\theta_1}(\varphi_*) \;<\; \mathbb{E}_{\theta_2}(\varphi_*).$$

Remark 7.14 (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

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

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

Proof of Theorem 7.13. The existence of $\gamma_\alpha \in [0,1]$ and $k_\alpha \in \mathbb{R}$ such that $\varphi_* := 1_{[T=k_\alpha]}\gamma_\alpha + 1_{[T>k_\alpha]}$ satisfies $\mathbb{E}_{\theta_o}(\varphi_*) = \alpha$ 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

$$H(r) \;:=\; \mathbb{P}_{\theta_o}(T \le r).$$

Then we define

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

so

$$\mathbb{P}_{\theta_o}(T > k_\alpha) \;\le\; \alpha \;\le\; \mathbb{P}_{\theta_o}(T \ge k_\alpha).$$

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

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

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

As to part (ii), we start with a rather general consideration. Let $h : \mathbb{R} \to \mathbb{R}$ be a non-decreasing and bounded function. Then

$$\int h(T)\, d\,\mathbb{P}_\theta$$

is a non-decreasing function of $\theta \in \Theta$. For if $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$, then with $g := g_{\theta_1,\theta_2}$,

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

For $h \equiv 1$ we obtain

$$\int (g(T) - 1)\, d\,\mathbb{P}_{\theta_1} \;=\; 0.$$

Since g is non-decreasing, the latter equation implies that for some $t_o \in \mathbb{R}$,

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

But then we may conclude that

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

because the latter integrand is everywhere non-negative.

Note that $\varphi_* = h(T)$ with the nondecreasing function $h : \mathbb{R} \to [0, 1]$, $h(t) := 1_{[t=k_\alpha]}\gamma_\alpha + 1_{[t>k_\alpha]}$. Consequently, the power function of φ_* is non-decreasing on Θ. Moreover, for arbitrary $\theta \in \Theta$ it follows from $g_{\theta_o,\theta} > 0$ that

$$\int \varphi_*\, d\,\mathbb{P}_\theta \;=\; \begin{cases} \displaystyle\int \varphi_*\, g_{\theta_o,\theta}\, d\,\mathbb{P}_{\theta_o} \;>\; 0, \\ \displaystyle 1 - \int (1 - \varphi_*)\, g_{\theta_o,\theta}\, d\,\mathbb{P}_{\theta_o} \;<\; 1, \end{cases}$$

because otherwise $\mathbb{P}_{\theta_o}(\varphi_* > 0) = 0$ or $\mathbb{P}_{\theta_o}(\varphi_* < 1) = 0$, a contradiction to $\mathbb{E}_{\theta_o}(\varphi_*) = \alpha \in (0, 1)$. These considerations prove property (ii.1).

For arbitrary $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$, the function $g := g_{\theta_1,\theta_2}$ is non-decreasing. Hence any test φ satisfies the inequality

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

Consequently

$$\begin{aligned} 0 \;&\geq\; \int (\varphi - \varphi_*)\big(f_{\theta_2} - g(k_\alpha) f_{\theta_1}\big)\, dM \\ &=\; \mathbb{E}_{\theta_2}(\varphi) - \mathbb{E}_{\theta_2}(\varphi_*) - g_{\theta_1,\theta_2}(k_\alpha)\big(\mathbb{E}_{\theta_1}(\varphi) - \mathbb{E}_{\theta_1}(\varphi_*)\big), \end{aligned}$$

so

$$\mathbb{E}_{\theta_2}(\varphi) - \mathbb{E}_{\theta_2}(\varphi_*) \;\leq\; g_{\theta_1,\theta_2}(k_\alpha)\big(\mathbb{E}_{\theta_1}(\varphi) - \mathbb{E}_{\theta_1}(\varphi_*)\big) \tag{7.3}$$

with $g_{\theta_1,\theta_2}(k_\alpha) > 0$ by assumption.

In the special case of $\theta_1 = \theta_o$, it follows from (7.3) that $\mathbb{E}_{\theta_2}(\varphi) \leq \mathbb{E}_{\theta_2}(\varphi_*)$ for arbitrary $\theta_2 > \theta_o$ and any test φ satisfying $\mathbb{E}_{\theta_o}(\varphi) \leq \alpha = \mathbb{E}_{\theta_o}(\varphi_*)$. This proves property (ii.2).

As to property (ii.3), (7.3) shows that for arbitrary parameters $\theta_1 < \theta_2$, the test φ_* is an optimal test of the simple null hypothesis $\{\theta_1\}$ versus the simple alternative hypothesis $\{\theta_2\}$ at level $\mathbb{E}_{\theta_1}(\varphi_*) \in (0, 1)$. Thus it follows from the Neyman–Pearson lemma that $\mathbb{E}_{\theta_2}(\varphi_*) > \mathbb{E}_{\theta_1}(\varphi_*)$ unless $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$. □

Example 7.10 (Bernoulli experiments, cont.) Note that the distribution $\mathbb{P}_\theta^T$ of T equals the binomial distribution $\mathrm{Bin}(n,\theta)$. Let $b_{n,\theta}$ and $B_{n,\theta}$ denote the weight and distribution function of $\mathrm{Bin}(n,\theta)$, respectively, i.e.

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

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

$$B_{n,p}^{-1}(u) := \min\{x : B_{n,p}(x) \ge u\}, \quad u \in (0,1],$$

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

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

where

$$k_\alpha = B_{n,\theta_o}^{-1}(1-\alpha) \quad \text{and} \quad \gamma_\alpha = \frac{B_{n,\theta_o}(k_\alpha) - 1 + \alpha}{b_{n,\theta_o}(k_\alpha)}.$$

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

$$\mathbb{E}_\theta(\varphi_*) = 1 - B_{n,\theta}(k_\alpha) + b_{n,\theta}(k_\alpha)\gamma_\alpha.$$

Example 7.11 (Gaussian location family, cont.) Note that in case of $\boldsymbol{X} \sim \mathcal{N}(\theta,\sigma^2)^{\otimes n}$, the sample mean $\bar{X} = T(\boldsymbol{X})$ has distribution $\mathcal{N}(\theta,\tau^2)$ with $\tau := \sigma/\sqrt{n}$. Hence,

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

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

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

The power of this test φ_* at an arbitrary parameter θ equals

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

Exercise 7.15. Motivated by the Hardy–Weinberg law in genetics, consider the statistical model $(\mathbb{P}_\theta)_{\theta\in(0,1)}$ with

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

(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 $\mathbb{P}_\theta^T$ of this test statistic T.

7.3 Stochastic Order, P-Values, Confidence Bounds

In practice, only non-randomized tests, i.e. tests with values in $\{0,1\}$, are used. Nevertheless the results in Section 7.2 show that certain non-randomized tests based on so-called p-values are essentially optimal. In the present section we describe non-randomized tests, p-values and confidence regions which are valid under a weaker condition on our statistical model $(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$.

Stochastically ordered distributions. We still assume that Θ is a subset of $\mathbb{R}$. Further we assume that there exists a measurable function $T : \Omega \to \mathbb{R}$ such that the corresponding distribution functions $F_\theta : \mathbb{R} \to [0,1]$ with

$$F_\theta(t) \;:=\; \mathbb{P}_\theta(T \le t)$$

satisfy the following conditions, which are equivalent:

(SO.1) For any fixed $t \in \mathbb{R}$, $F_\theta(t)$ is non-increasing in $\theta \in \Theta$.

(SO.2) For any fixed $t \in \mathbb{R}$, $F_\theta(t-) = \mathbb{P}_\theta(T < t)$ is non-increasing in $\theta \in \mathbb{R}$.

(SO.3) For any fixed $u \in (0,1)$, $F_\theta^{-1}(u) = \min\{t \in \mathbb{R} : F_\theta(t) \ge u\}$ is non-decreasing in $\theta \in \mathbb{R}$.

(SO.4) For any non-decreasing function $h : \mathbb{R} \to [0,\infty)$, $\int h(T)\, d\,\mathbb{P}_\theta$ is non-decreasing in $\theta \in \mathbb{R}$.

If (SO.1-4) are satisfied, we say that the distribution functions F_θ are stochastically ordered in the sense that $F_{\theta_1} \le_{\text{st.}} F_{\theta_2}$ whenever $\theta_1 < \theta_2$.

Exercise 7.16. Show that Conditions (SO.1-4) are equivalent.

The proof of Theorem 7.13 (ii) shows that any statistical model with monotone density ratios satisfies condition (SO.4) and hence (SO.1-4). Note also that in Example 7.10 one may extend the parameter space to $\Theta = [0,1]$, and the stochastic order constraint remains valid.

Tests in terms of p-values. To test whether a hypothetical parameter θ_o is plausible for a particular data set $\omega \in \Omega$, we could compute

- the left-sided p-value

$$F_{\theta_o}(T(\omega)) \;=\; \mathbb{P}_{\theta_o}(T \le T(\omega)),$$

- the right-sided p-value

$$1 - F_{\theta_o}(T(\omega)-) \;=\; \mathbb{P}_{\theta_o}(T \ge T(\omega)).$$

The left-sided p-value is non-decreasing in $T(\omega)$, and small values indicate that $T(\omega)$ is "suspiciously small" for the parameter θ_o. The right-sided p-value is non-increasing in $T(\omega)$ with small values indicating that $T(\omega)$ is "suspiciously large" for θ_o.

These p-values lead to tests which are similar to the UMP tests in Section 7.2: Let us fix a test level $\alpha \in (0,1)$. On the one hand, with the right-sided critical value

$$k^{(r)}_{\alpha,\theta_o} \;:=\; \min\bigl\{t \in \mathbb{R} : F_{\theta_o}(t) \ge 1-\alpha\bigr\}$$

we can conclude that

$$1 - F_{\theta_o}(T-) \;\le\; \alpha \quad \text{if and only if} \quad T \begin{cases} > \; k^{(r)}_{\alpha,\theta_o} & \text{if } F_{\theta_o}(k^{(r)}_{\alpha,\theta_o}-) < 1-\alpha, \\ \ge \; k^{(r)}_{\alpha,\theta_o} & \text{if } F_{\theta_o}(k^{(r)}_{\alpha,\theta_o}-) = 1-\alpha. \end{cases}$$

Thus by stochastic ordering,

$$\mathbb{P}_\theta(1 - F_{\theta_o}(T-) \le \alpha) \;\le\; \mathbb{P}_{\theta_o}(1 - F_{\theta_o}(T-) \le \alpha) \;\le\; \alpha \quad \text{for arbitrary } \theta \le \theta_o.$$

Consequently,

$$\varphi^{(r)}_{\alpha,\theta_o} \;:=\; 1_{[1-F_{\theta_o}(T-)\le\alpha]} \;=\; 1_{[F_{\theta_o}(T-)\ge 1-\alpha]}$$

defines a level-α test of $\Theta \cap (-\infty, \theta_o]$.

Similarly, with the left-sided critical value

$$k^{(\ell)}_{\alpha,\theta_o} \;:=\; \max\bigl\{t \in \mathbb{R} : F_{\theta_o}(t-) \le \alpha\bigr\}$$

we can conclude that

$$F_{\theta_o}(T) \;\le\; \alpha \quad \text{if and only if} \quad T \begin{cases} < \; k^{(\ell)}_{\alpha,\theta_o} & \text{if } F_{\theta_o}(k^{(\ell)}_{\alpha,\theta_o}) > \alpha, \\ \le \; k^{(\ell)}_{\alpha,\theta_o} & \text{if } F_{\theta_o}(k^{(\ell)}_{\alpha,\theta_o}) = \alpha, \end{cases}$$

whence by stochastic ordering,

$$\mathbb{P}_\theta(F_{\theta_o}(T) \le \alpha) \;\le\; \mathbb{P}_{\theta_o}(F_{\theta_o}(T) \le \alpha) \;\le\; \alpha \quad \text{for arbitrary } \theta \ge \theta_o.$$

Consequently,

$$\varphi^{(\ell)}_{\alpha,\theta_o} \;:=\; 1_{[F_{\theta_o}(T)\le\alpha]}$$

defines a level-α test of $\Theta \cap [\theta_o, \infty)$.

Confidence bounds. By means of the p-values just constructed, one can also construct confidence regions for the parameter $\theta \in \Theta$:

For given test level $\alpha \in (0,1)$ and data set $\omega \in \Omega$ let

$$\mathcal{C}^{(\ell)}_\alpha(\omega) \;:=\; \bigl\{\theta \in \Theta : F_\theta(T(\omega)) > \alpha\bigr\}.$$

That means, $\mathcal{C}^{(\ell)}_\alpha(\omega)$ is the set of all parameters such that the corresponding left-sided p-value $F_\theta(T(\omega))$ is larger than α. Since $F_\theta(T)$ is non-increasing in $\theta \in \Theta$, the set $\mathcal{C}^{(\ell)}_\alpha(\omega)$ is always an interval $\Theta \cap (-\infty, b_\alpha(\omega))$ or $\Theta \cap (-\infty, b_\alpha(\omega)]$ for some $b_\alpha(\omega) \in [-\infty, \infty]$.

Consequently, $\mathcal{C}^{(\ell)}_\alpha$ and b_α comprise a $(1-\alpha)$-confidence interval and an upper $(1-\alpha)$-confidence bound in the folling sense: For arbitrary $\theta \in \Theta$,

$$\mathbb{P}_\theta(b_\alpha \ge \theta) \;\ge\; \mathbb{P}_\theta(\mathcal{C}^{(\ell)}_\alpha \ni \theta) \;\ge\; 1-\alpha.$$

In other words, assuming that an observed data set ω is a realization of a random variable with distribution $\mathbb{P}_\theta$ for some unknown true parameter $\theta \in \Theta$, we may claim with confidence $1-\alpha$ that $\theta \in \mathcal{C}^{(\ell)}_\alpha(\omega)$ and $\theta \le b_\alpha(\omega)$.

Similarly let

$$\mathcal{C}_\alpha^{(r)}(\omega) \;:=\; \big\{\theta \in \Theta : F_\theta(T(\omega)-) < 1-\alpha\big\},$$

the set of all parameters such that the corresponding right-sided p-value is larger than α. Since $F_\theta(T(\omega)-)$ is non-increasing in $\theta \in \Theta$, the set $\mathcal{C}_\alpha^{(r)}(\omega)$ is always an interval $\Theta \cap (a_\alpha(\omega), \infty)$ or $\Theta \cap [a_\alpha(\omega), \infty)$ for some $a_\alpha(\omega) \in [-\infty, \infty]$.

Consequently, $\mathcal{C}_\alpha^{(r)}$ and a_α comprise a $(1-\alpha)$-confidence interval and a lower $(1-\alpha)$-confidence bound in the folling sense: For arbitrary $\theta \in \Theta$,

$$\mathbb{P}_\theta(a_\alpha \le \theta) \;\ge\; \mathbb{P}_\theta(\mathcal{C}_{\alpha,r} \ni \theta) \;\ge\; 1-\alpha.$$

That means, assuming that an observed data set ω is a realization of a random variable with distribution $\mathbb{P}_\theta$ for some unknown true parameter $\theta \in \Theta$, we may claim with confidence $1-\alpha$ that $\theta \in \mathcal{C}_\alpha^{(r)}(\omega)$ and $\theta \ge a_\alpha(\omega)$.

Example 7.10 (Bernoulli sequences, cont.) For any $\theta_o \in [0,1]$ the left- and right-sided p-values are given by

$$B_{n,\theta_o}(T(\omega)) \quad \text{and} \quad 1 - B_{n,\theta_o}(T(\omega)-1),$$

respectively. The resultung confidence intervals are

$$\mathcal{C}_\alpha^{(\ell)}(\omega) \;=\; \begin{cases} [0, b_\alpha(\omega)) & \text{if } T(\omega) < n, \\ [0,1] & \text{if } T(\omega) = n, \end{cases}$$

$$\mathcal{C}_\alpha^{(r)}(\omega) \;=\; \begin{cases} [0,1] & \text{if } T(\omega) = 0, \\ (a_\alpha(\omega), 1] & \text{if } T(\omega) > 0, \end{cases}$$

where $b_\alpha(\omega)$ is the unique $p \in (0,1)$ such that $B_{n,p}(T(\omega)) = \alpha$ (if $T(\omega) < n$) while $a_\alpha(\omega)$ is the unique $p \in (0,1)$ such that $B_{n,p}(T(\omega)-1) = 1-\alpha$ (if $T(\omega) > 0$). Here we utilize the fact that for any integer $x \in \{0,1,\ldots,n-1\}$, the function $p \mapsto B_{n,p}(x)$ is continuous and strictly decreasing with boundary values $B_{n,0}(x) = 1$ and $B_{n,1}(x) = 0$.

Example 7.17 (Gaussian location family, cont.). Since

$$F_\theta(t) \;=\; \Phi\big((t-\theta)/\tau\big),$$

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 7.18. 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 $F_0 \le_{\text{st.}} F_1$ (i.e. $F_0 \ge F_1$) but $g(x) := P_1(\{x\})/P_0(\{x\})$ is not monotone in $x \in \mathcal{X}$.

Example 7.19 (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 7.20 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

$$a_\alpha(x) \;:=\; \min\big\{N \geq \max(n_1, n_2) : F_N(x) > \alpha\big\}$$

while an upper bound is given by

$$b_\alpha(x) \;=\; \begin{cases} \infty & \text{if } x = 0, \\ \max\big\{N \geq \max(n_1, n_2) : F_N(x-1) < 1-\alpha\big\} & \text{if } x > 0. \end{cases}$$

Exercise 7.20. 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 7.21. 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 $\mathbb{E}_N(\widehat{N})$ for the point estimator $\widehat{N} := (n_1+1)(n_2+1)/(X+1)$.

7.4 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 7.22 (Generalized Neyman–Pearson lemma). *Let $\mathcal{T}$ be the set of all statistical tests $\varphi : \Omega \to [0,1]$. Let M be a σ-finite measure on $(\Omega, \mathcal{A})$, 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*

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

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

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

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

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

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

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

$$\int \varphi_* f_{m+1}\, dM \;\geq\; \int \varphi f_{m+1}\, dM$$

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

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

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

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

Then there exists a test φ_* as described in part (ii).

Proof of Theorem 7.22. In what follows let

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

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

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

is a compact and convex subset of $\mathbb{R}^{\mathcal{F}}$, equipped with the product topology, see Theorem A.12 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.14 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 $\varphi_* \in \mathcal{T}(\boldsymbol{\alpha})$ such that

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

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

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

whence

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

The right hand side equals 0, if $\varphi \in \mathcal{T}(\boldsymbol{\alpha})$, so φ_* maximizes $\int \varphi f_{m+1} \, dM$ 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 $1 \le j \le m$,

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

Proof of part (iii): Let

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

with

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

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 $\varphi \in \mathcal{T}$,

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

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

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

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

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

shows that

$$M(\varphi_* \neq \varphi_{**}) \;=\; 0.$$

Hence we may replace φ_* with φ_{**}. □

7.5 Tests of Two-Sided Hypotheses

7.5.1 One-parameter exponential families (with natural parametrization)

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

$$f_\theta(\omega) = \frac{d\,\mathbb{P}_\theta}{dM}(\omega) \;=\; C(\theta) h(\omega) \exp(\theta T(\omega))$$

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

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

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

Example 7.10 (Bernoulli sequences, cont.) This statistical model is an exponential family with $\Omega = \{0,1\}^n$, $M =$ counting measure on Ω, and

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

Example 7.23 (Poisson distributions). The family of Poisson distributions $\mathrm{Poiss}(\lambda)$, $\lambda > 0$, is a exponential family with $\Omega = \mathbb{N}_0$, $M =$ counting measure on Ω, and

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

Example 7.11 (Gaussian location family, cont.) The family of distributions $\mathcal{N}(\mu,\sigma^2)^{\otimes n}$, $\mu \in \mathbb{R}$, is an exponential family with $\Omega = \mathbb{R}^n$, $M =$ 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 $\mathbb{P}_\theta^T = \mathcal{N}(\theta, 1)$.

7.5.2 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 7.24. ***(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*

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

is a test satisfying

$$\mathbb{E}_{\theta_1}(\varphi_*) \;=\; \mathbb{E}_{\theta_2}(\varphi_*) \;=\; \alpha.$$

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

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

$$\mathbb{E}_\theta(\varphi_*) \;\ge\; \mathbb{E}_\theta(\varphi) \quad \text{for all } \theta \in (\theta_1, \theta_2).$$

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

$$\mathbb{E}_\theta(\varphi_*) \;\le\; \mathbb{E}_\theta(\varphi) \quad \text{for all } \theta \in \Theta_o.$$

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

$$\mathbb{E}_\theta(\varphi_*) \;\le\; \alpha \quad \text{for all } \theta \in \Theta_o.$$

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

Lemma 7.25. *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

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

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 7.25. Suppose first that $c_1 < c_2$. 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.

In 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 7.24. Since $\mathbb{P}_\theta(h = 0) = 0$ for all $\theta \in \Theta$, we may replace Ω with $\{h > 0\}$ and $M(d\omega)$ with $h(\omega)M(d\omega)$, so $h \equiv 1$.

Proof of part (i). We fix an arbitrary parameter $\theta_3 \in (\theta_1, \theta_2)$ and construct an optimal level-α test φ_* of $\{\theta_1, \theta_2\}$ versus $\{\theta_3\}$ by means of the generalized Neyman–Pearson lemma. To this end we consider the point $\boldsymbol{\alpha} := (\alpha, \alpha)^\top$ and the set

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

with $\boldsymbol{f} := (f_{\theta_1}, f_{\theta_2})^\top : \Omega \to \mathbb{R}^2$. 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

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

This implies that $\boldsymbol{\alpha}$ is an interior point of the set $\mathcal{K}_2$, see also Exercise 7.26 below. Consequently, there exists a test φ_* such that $\mathbb{E}_{\theta_1}(\varphi_*) = \mathbb{E}_{\theta_2}(\varphi_*) = \alpha$ with

$$\varphi_* \;=\; \begin{cases} 1 & \text{if } f_{\theta_3} > k_1 f_{\theta_1} + k_2 f_{\theta_2} \\ 0 & \text{if } f_{\theta_3} < k_1 f_{\theta_1} + k_2 f_{\theta_2} \end{cases}$$

for certain constants $k_1, k_2 \in \mathbb{R}$. With $b_j := k_j C(\theta_j)/C(\theta_3)$ we may also write

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

with

$$A(t) \;:=\; b_1 e^{(\theta_1 - \theta_3)t} + b_2 e^{(\theta_2 - \theta_3)t}.$$

Since $\mathbb{E}_{\theta_1}(\varphi_*), \mathbb{E}_{\theta_2}(\varphi_*) < 1$, 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_* with

$$\bar{\varphi}_* \;:=\; \begin{cases} 1 & \text{if } A(T) < 1 \\ \displaystyle\int_{\{T=c\}} \varphi_* h\, dM \Big/ \int_{\{T=c\}} h\, dM & \text{if } T = c \in \{A = 1\} \\ 0 & \text{else} \end{cases}$$

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

$$\int_{\{T=c\}} \varphi_*\, d\,\mathbb{P}_\theta \;=\; C(\theta) e^{\theta c} \int_{\{T=c\}} \varphi_* h\, dM \;=\; C(\theta) e^{\theta c} \int_{\{T=c\}} \bar{\varphi}_*\, dM \;=\; \int_{\{T=c\}} \bar{\varphi}_*\, d\,\mathbb{P}_\theta$$

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

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

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 $\mathbb{E}_\theta(\varphi_*)$ would be strictly increasing or strictly decreasing in $\theta \in \Theta$, see Theorem 7.13 (ii). But this would contradict the equation $\mathbb{E}_{\theta_1}(\varphi_*) = \mathbb{E}_{\theta_2}(\varphi_*)$. 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_* has the asserted form

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

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_* 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 7.25 shows that there exist constants $b_{\theta 1}, b_{\theta 2} > 0$ 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 7.25 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

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

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 7.25 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

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

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

Exercise 7.26. 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.

7.5.3 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 7.24.

Theorem 7.27. ***(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*

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

is a test satisfying

$$\mathbb{E}_{\theta_1}(\varphi_*) \;=\; \mathbb{E}_{\theta_2}(\varphi_*) \;=\; \alpha.$$

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

(ii.1)

$$\mathbb{E}_{\theta}(\varphi_*) \;\le\; \alpha \quad \text{for all } \theta \in \Theta_o.$$

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

$$\mathbb{E}_{\theta}(\varphi_*) \;\ge\; \mathbb{E}_{\theta}(\varphi) \quad \text{for all } \theta \in \Theta_A.$$

Proof of Theorem 7.27. We may apply Theorem 7.24 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 7.24 imply that $\varphi_* := 1 - \varphi_{**}$ has the properties stated in Theorem 7.27. □

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 $\mathbb{E}_{\theta_o}(\varphi_{\alpha,l}) = \mathbb{E}_{\theta_o}(\varphi_{\alpha,r}) = \alpha$. Then by Theorem 7.13 (ii), the power functions of $\varphi_{\alpha,l}$ and $\varphi_{\alpha,r}$ are strictly decreasing and strictly increasing, respectively, and any test φ with $\mathbb{E}_{\theta_o}(\varphi) \le \alpha$ satisfies

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

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

$$\mathbb{E}_{\theta}(\varphi) \;\ge\; \alpha \;=\; \mathbb{E}_{\theta_o}(\varphi),$$

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 6.25 that for any test φ, the power function

$$\Theta \ni \theta \;\mapsto\; \mathbb{E}_\theta(\varphi)$$

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

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

Hence, if φ is a test such that

$$\mathbb{E}_\theta(\varphi) \;\geq\; \alpha \;\geq\; \mathbb{E}_{\theta_o}(\varphi) \quad \text{for all } \theta \in \Theta \setminus \{\theta_o\}, \tag{7.4}$$

then

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

Theorem 7.28. ***(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*

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

is a test satisfying

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

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

$$\mathbb{E}_\theta(\varphi_*) \;\geq\; \mathbb{E}_\theta(\varphi) \quad \textit{for all } \theta \in \Theta.$$

In particular, φ_* has property (7.4)*, and its power function is pointwise maximal among all tests with that property.*

Proof of Theorem 7.28. Without loss of generality we may assume that $\theta_o = 0$ and $M = \mathbb{P}_0$, $h \equiv 1$, 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 $\mathbb{P}_\theta$ with respect to $\mathbb{P}_{\theta_o}$. We may further assume without loss of generality that $\mathbb{E}_0(T) = 0$, because

$$C(\theta)\cdot\exp(\theta T) \;=\; C(\theta)\exp(\theta\,\mathbb{E}_0(T))\cdot\exp\bigl(\theta(T - \mathbb{E}_0(T))\bigr).$$

Now we construct for a fixed $\theta_1 \in \Theta \setminus \{0\}$ a test φ_* maximizing $\mathbb{E}_{\theta_1}(\varphi_*)$ under the constraints that

$$\mathbb{E}_0(\varphi_*) \;=\; \int \varphi_* \, d\,\mathbb{P}_0 \;=\; \alpha \quad \text{and} \quad \mathbb{E}_0(\varphi_* T) \;=\; \int \varphi_* T \, d\,\mathbb{P}_0 \;=\; 0.$$

This may be achieved with the generalized Neyman–Pearson lemma applied to $f_1 := 1$, $f_2 := T$, $f_3 := f_{\theta_1}$ and $M := \mathbb{P}_0$. With the four tests $\varphi := 0$, $\varphi := 1$, $\varphi := 1_{[T>0]}$ and $\varphi := 1_{[T\le 0]}$ it follows that

$$\mathcal{K}_2 \;:=\; \Big\{ \int \varphi \boldsymbol{f} \, d\,\mathbb{P}_0 : \varphi \in \mathcal{T} \Big\}$$

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 $\beta = \mathbb{P}_0(T>0) \in (0,1)$ and $\gamma = \int T^+ \, d\,\mathbb{P}_0 = \int T^- \, d\,\mathbb{P}_0 > 0$. Hence $(\alpha,0)^\top$ is an interior point of $\mathcal{K}_2$. Consequently there exist a test φ_* and real constants k_1, k_2 such that

$$\int \varphi_* \, d\,\mathbb{P}_0 \;=\; \alpha, \quad \int \varphi_* T \, d\,\mathbb{P}_0 \;=\; 0$$

and

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

Note that $f_1 = C(\theta_1) e^{\theta_1 T}$. Since $t \mapsto C(\theta_1) e^{\theta_1 t}$ is strictly convex, and since $\varphi_* \not\equiv 1$, there have to exist points $c_1 \le c_2$ such that $C(\theta_1) e^{\theta_1 c_j} = k_1 + k_2 c_j$ for $j = 1, 2$, and

$$\varphi_* \;=\; \begin{cases} 1 & \text{if } T \in \mathbb{R} \setminus [c_1, c_2], \\ 0 & \text{if } T \in (c_1, c_2). \end{cases}$$

Finally, we may replace φ_* with

$$\bar{\varphi}_* \;:=\; \begin{cases} 1 & \text{if } T \in \mathbb{R} \setminus [c_1, c_2] \\ \displaystyle\int_{\{T=c\}} \varphi_* \, d\,\mathbb{P}_0 \,\big/\, \mathbb{P}_0(T=c) & \text{if } T = c \in \{c_1, c_2\} \\ 0 & \text{else} \end{cases}$$

with the convention that $0/0 := 0$. This does not change the power function of φ_*, and then φ_* has the properties mentioned 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

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

Consequently, it follows from part (ii) of the generalized Neyman–Pearson lemma that $\mathbb{E}_\theta(\varphi_*) \ge \mathbb{E}_\theta(\varphi)$ for any test φ such that $\int \varphi \, d\,\mathbb{P}_0 = \alpha$ and $\int \varphi T \, d\,\mathbb{P}_0 = 0$. Taking $\varphi \equiv \alpha$ reveals that φ_* has property (7.4) and is optimal among all tests with that property. $\square$

Exercise 7.29. This exercise provides further details about the power function of statistical tests in a one-parameter exponential family with natural parametrization and test statistic $T : \Omega \to \mathbb{R}$.

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

$$h_f(\theta) \;:=\; \mathbb{E}_\theta(f)$$

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

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

Hint: Consider Exercise 6.25.

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

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

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

$$h_f''(\theta) \;=\; \mathbb{E}_\theta\big((T-c_1)(T-c_2)(f-\mathbb{E}_\theta(f))\big)$$

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

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

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

with $c_1 \le c_2$ and $\gamma_1, \gamma_2 \in [0,1]$. Show that

$$(\varphi_* - \mathbb{E}_\theta(\varphi_*))(T-c_1)(T-c_2) \begin{cases} \le 0 & \text{(Type 1)} \\ \ge 0 & \text{(Type 2)} \end{cases}$$

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

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

unless $\mathbb{P}_\theta^T$ is concentrated on $\{c_1, c_2\}$.

7.5.4 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 7.5.1, 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 7.30 (One-parameter exponential family). A one-parameter exponential family is a statistical experiment

$$\big(\Omega, \mathcal{A}, (\mathbb{P}_\lambda)_{\lambda\in\Lambda}\big)$$

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

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

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 $\mathbb{P}_{\lambda_1} \neq \mathbb{P}_{\lambda_2}$ 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"

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

Now we consider tests φ_* of one of the following types:

$$\varphi_* \;=\; 1_{[T=c_1]}\gamma_1 + 1_{[c_1<T<c_2]} + 1_{[T=c_2]}\gamma_2, \tag{7.5}$$

$$\varphi_* \;=\; 1_{[T=c_1]}\gamma_1 + 1_{[T=c_2]}\gamma_2 + 1_{[T<c_1 \text{ or } T>c_2]}, \tag{7.6}$$

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_* of the following type:

$$\varphi_* \;=\; 1_{[c_1\leq T\leq c_2]}, \tag{7.7}$$

$$\varphi_* \;=\; 1_{[T\leq c_1 \text{ or } T\geq c_2]}, \tag{7.8}$$

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 (7.5) or (7.7), provided that

$$\mathbb{E}_{\lambda_1}(\varphi_*) \;=\; \alpha \;=\; \mathbb{E}_{\lambda_2}(\varphi_*).$$

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 (7.6) or (7.8), provided that

$$\mathbb{E}_{\lambda_1}(\varphi_*) \;=\; \alpha \;=\; \mathbb{E}_{\lambda_2}(\varphi_*).$$

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 (7.6) or (7.8), provided that

$$\mathbb{E}_{\lambda_o}(\varphi_*) \;=\; \alpha \quad \text{and} \quad \frac{d}{d\lambda}\Big|_{\lambda=\lambda_o} \mathbb{E}_\lambda(\varphi_*) \;=\; 0.$$

Example 7.11 (Gaussian location family, cont.) Suppose we observe a random vector $\boldsymbol{X} \in \mathbb{R}^n$ with distribution $\mathbb{P}_\mu := \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ for a given $\sigma > 0$ and some unknown $\mu \in \mathbb{R}$. As shown before, the model $\big(\mathcal{N}(\mu, \sigma^2)^{\otimes n}\big)_{\mu\in\mathbb{R}}$ 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

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

where $\tau := \sigma/\sqrt{n}$ is the standard deviation of $\bar{X}$. This follows from the fact that φ_* is of type (7.8), and with $Z := (\bar{X} - \mu)/\tau \sim \mathcal{N}(0,1)$, the power function of φ_* is given by

$$\begin{aligned} \mathbb{E}_\mu(\varphi_*) \;&=\; \mathbb{P}\Big(Z \geq \frac{\mu_o - \mu}{\tau} + \Phi^{-1}(1-\alpha/2)\Big) + \mathbb{P}\Big(Z \leq \frac{\mu_o - \mu}{\tau} - \Phi^{-1}(1-\alpha/2)\Big) \\ &=\; 1 - \Phi\Big(\frac{\mu_o - \mu}{\tau} + \Phi^{-1}(1-\alpha/2)\Big) + \Phi\Big(\frac{\mu_o - \mu}{\sigma/\sqrt{n}} - \Phi^{-1}(1-\alpha/2)\Big) \\ &=\; \Phi\Big(\Phi^{-1}(\alpha/2) + \frac{\mu - \mu_o}{\tau}\Big) + \Phi\Big(\Phi^{-1}(\alpha/2) - \frac{\mu - \mu_o}{\tau}\Big). \end{aligned}$$

This equals α for $\mu = \mu_o$, and it is an even function of $\mu - \mu_o$, so the derivative of the power function at μ_o equals 0.

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_* would be

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

for a suitable constant $c_\alpha > 0$. Indeed, this defines a test of type (7.7), and with Z as above we may write

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

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

$$\mathbb{E}_{\mu_o \pm \delta}(\varphi_*) \;=\; \alpha.$$

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

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

In this case, we may claim with confidence $1-\alpha$ that $\mu = \mu_o$. This sounds almost miraculous, but note that the inequality $|\bar{X} - \mu_o| \le \tau\Phi^{-1}((1+\alpha)/2)$ 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 7.31. 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 (b) to become an optimal test of

$$\{b_o\} \quad \text{versus} \quad (0,\infty) \setminus \{b_o\}.$$

for arbitrary fixed $b_o > 0$.

Exercise 7.32. 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 $\delta = 0.1$.

7.6 Tests and Confidence regions

For any statistical model $(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ there is a close relationship between tests and confidence regions. Let uns first clarify what we mean by confidence regions.

Definition 7.33 (Randomized confidence region). A *(randomized) confidence region* is a mapping $C : \Omega \times \Theta \to [0, 1]$ such that for any fixed $\theta \in \Theta$, the mapping $C(\cdot, \theta)$ is $\mathcal{A}$-measurable.

Suppose that for a given test level $\alpha \in (0, 1)$,

$$\int C(\omega, \theta)\ \mathbb{P}_\theta(d\omega)\ \geq\ 1 - \alpha \quad \text{for arbitrary } \theta \in \Theta.$$

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 $X \in \Omega$ with distribution $\mathbb{P}_\theta$, 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

$$\mathcal{C}(X, U)\ :=\ \big\{\theta \in \Theta : C(X, \theta) \geq U\big\}\ \subset\ \Theta.$$

In case of C being non-randomized, i.e. C taking only values in $\{0, 1\}$, we don't need the extra random variable U, because the set $\mathcal{C}(X, U)$ equals

$$\mathcal{C}(X)\ :=\ \big\{\theta \in \Theta : C(X, \theta) = 1\big\}\ \subset\ \Theta$$

almost surely.

If C is a $(1 - \alpha)$-confidence region, then

$$\left.\begin{array}{c}\mathbb{P}_\theta\big(\theta \in \mathcal{C}(X)\big)\\ \mathbb{P}_\theta\big(\theta \in \mathcal{C}(X, U)\big)\end{array}\right\}\ \geq\ 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

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 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 $C : \Omega \times \Theta \to [0, 1]$ 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 $C(\omega, \theta) := 1 - \varphi(\omega, \theta)$ 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 7.34 (Simple Gaussian shift model). Suppose we observe $X \sim \mathbb{P}_\mu = \mathcal{N}(\mu, 1)$ for some unknown parameter $\mu \in \mathbb{R}$. One can easily verify that for any hypothetical parameter $\mu_o \in \mathbb{R}$, an optimal level-α test of $\{\mu_o\}$ with power function at least α everywhere is given by

$$\varphi(X, \mu_o) \ := \ 1_{\left[|X-\mu_o| > \Phi^{-1}(1-\alpha/2)\right]}.$$

This leads to the $(1-\alpha)$-confidence region C given by

$$C(X, \mu) \ = \ 1_{\left[|X-\mu| \le \Phi^{-1}(1-\alpha/2)\right]} \ = \ 1_{\left[\mu \in [X \pm \Phi^{-1}(1-\alpha/2)]\right]},$$

i.e. $\mathcal{C}(X) = [X \pm \Phi^{-1}(1-\alpha/2)]$.

Exercise 7.35. As in Exercise 7.31, 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 $\mathbb{P}_p(p \in \mathcal{C}(X))$ 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 7.31 (b)?

Duality between confidence regions and tests of composite hypotheses. Sometimes we are not interested in a confidence region for the unknown true parameter θ but only in an upper bound for $g(\theta)$ with a given function

$$g : \Theta \to \mathbb{R}.$$

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 $b_\alpha(X)$ for $g(\theta)$ corresponds to a measurable function $b_\alpha : \Omega \to \mathbb{R}$ such that

$$\mathbb{P}_\theta(b_\alpha \ge g(\theta)) \ \ge \ 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

Such a confidence bound gives rise to the level-α test

$$\varphi_\alpha(X, \delta) \ := \ 1_{[b_\alpha(X) < \delta]}$$

of the null hypothesis

$$\Theta(\delta) \;:=\; \big\{\theta \in \Theta : g(\theta) \geq \delta\big\}$$

for arbitrary numbers $\delta \geq 0$. Indeed, if $g(\theta) \geq \delta$, then

$$\mathbb{P}_\theta(\varphi_\alpha(\cdot,\delta) = 1) \;=\; \mathbb{P}_\theta(b_\alpha < \delta) \;\leq\; \mathbb{P}_\theta(b_\alpha < g(\theta)) \;\leq\; \alpha.$$

On the other hand, suppose that for any number $\delta \geq 0$, we have constructed a level-α test $\varphi_\alpha(\cdot,\delta) : \Omega \to \{0,1\}$ 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 $g(\theta)$, because for any $\theta \in \Theta$ and $\delta := g(\theta)$,

$$\mathbb{P}_\theta(g(\theta) \leq b_\alpha) \;\geq\; \mathbb{P}_\theta(\varphi_\alpha(\cdot,\delta) = 0) \;=\; 1 - \mathbb{P}_\theta(\varphi_\alpha(\cdot,\delta) = 1) \;\geq\; 1-\alpha.$$

Finding an "optimal" upper confidence bound b_α may be interpreted as finding optimal level-α tests $\varphi_\alpha(\cdot,\delta)$, $\delta \geq 0$, of the null hypotheses $\Theta(\delta)$. Ideally, the function $\varphi_\alpha(\cdot,\cdot)$ is even non-decreasing in its second argument.

Example 7.34 (Simple Gaussian shift model, 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 $|X - \mu_o| + \Phi^{-1}(1-\alpha)$. But if we think about the duality of tests and confidence regions, a good upper $(1-\alpha)$-confidence bound $b_\alpha(X)$ for $|\mu - \mu_o|$ should satisfy the following condition: For any $\delta \geq 0$,

$$\varphi_\alpha(X,\delta) \;:=\; 1_{[b_\alpha(X) < \delta]}$$

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

$$\varphi_\alpha(X,\delta) \;:=\; 1_{\left[|X-\mu_o| < c_\alpha(\delta)\right]}$$

where $c_\alpha(\delta) > 0$ is the unique number such that

$$\Phi(c_\alpha(\delta) + \delta) + \Phi(c_\alpha(\delta) - \delta) \;=\; 1 + \alpha.$$

Recall that $c_\alpha(0) = \Phi^{-1}((1+\alpha)/2)$, but for $\delta > 0$, there is no simple formula for $c(\delta)$. An optimal upper confidence bound for $|\mu - \mu_o|$ is given by

$$\begin{aligned} b_\alpha(X) \;&:=\; \max\big(\{0\} \cup \big\{\delta \geq 0 : \varphi_\alpha(X,\delta) = 0\big\}\big) \\ &=\; \max\big(\{0\} \cup \big\{\delta \geq 0 : c(\delta) \leq |X - \mu_o|\big\}\big) \\ &=\; \max\big(\{0\} \cup \big\{\delta \geq 0 : \Phi(|X-\mu_o| + \delta) + \Phi(|X-\mu_o| - \delta) \geq 1 + \alpha\big\}\big). \end{aligned}$$

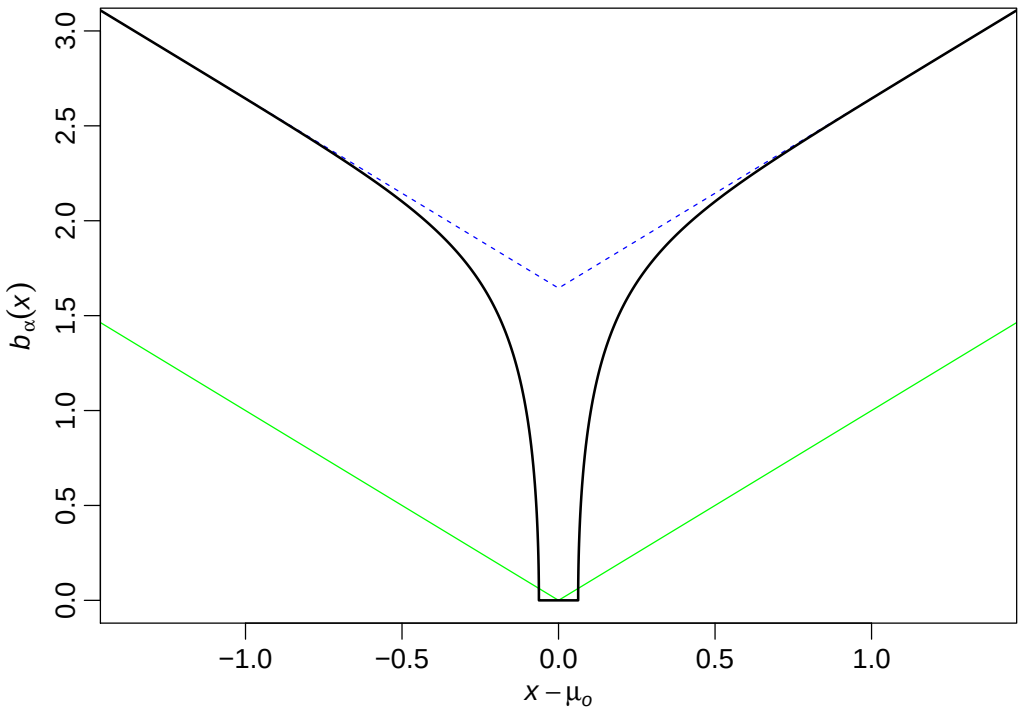

Figure 7.1: Optimal (black) and ad hoc (blue) upper 95%-confidence bound for $|\mu|$ in the simple Gaussian shift model.

If $|X - \mu_o| \leq \Phi^{-1}((1+\alpha)/2)$, then $b_\alpha(X) = 0$. Otherwise, $b_\alpha(X)$ is the unique solution $\delta > 0$ of the equation

$$\Phi(|X - \mu_o| + \delta) + \Phi(|X - \mu_o| - \delta) \;=\; 1 + \alpha,$$

which can be computed numerically.

Figure 7.1 depicts the optimal upper 95%-confidence bound (black curve) for $|\mu - \mu_o|$ and the simple upper bound $|X - \mu_o| + \Phi^{-1}(1-\alpha)$ (blue broken line).

Exercise 7.36 (Simple Gaussian shift model). Let $X \sim \mathbb{P}_\mu := \mathcal{N}(\mu, 1)$ for some unknown parameter $\mu \in \mathbb{R}$. A standard $(1-\alpha)$-confidence interval for μ is given by $[X \pm \Phi^{-1}(1-\alpha/2)]$.

(a) Show that for any given $\mu_o \in \mathbb{R}$,

$$\mathcal{C}_\alpha(X) \;:=\; \big[\min(X - \Phi^{-1}(1-\alpha), \mu_o), \max(X + \Phi^{-1}(1-\alpha), \mu_o)\big]$$

is also a $(1-\alpha)$-confidence interval for μ, i.e.

$$\mathbb{P}_\mu(\mu \in \mathcal{C}_\alpha) \;\geq\; 1 - \alpha \quad \text{for all } \mu \in \mathbb{R}.$$

(b) Show that

$$b_\alpha(X) \;:=\; |X - \mu_o| + \Phi^{-1}(1-\alpha)$$

is an upper $(1-\alpha)$-confidence bound for $|\mu - \mu_o|$, i.e.

$$\mathbb{P}_\mu(|\mu - \mu_o| \leq b_\alpha) \;\geq\; 1 - \alpha \quad \text{for all } \mu \in \mathbb{R}.$$

Exercise 7.37. Suppose we observe $X \sim \mathrm{Bin}(n, p)$ with given $n \in \mathbb{N}$ and unknown $p \in [0, 1]$.

(a) Show that for arbitrary integers $0 \le c_1 \le c_2 \le n$ with $c_2 - c_1 < n$,

$$[0, 1] \ni p \;\mapsto\; \log \mathbb{P}_p(c_1 \le X \le c_2) \in [-\infty, 0]$$

is strictly concave.

(b) Describe an optimal upper $(1 - \alpha)$-confidence bound $b_\alpha(X)$ for $|p - 0.5|$.
Hint: Consider $\mathbb{P}_p(k \le X \le n - k)$ for $k = 0, 1, \ldots, \lfloor n/2 \rfloor$.

(c) Write a computer program to compute the confidence bound $b_\alpha(X)$ in part (b).

Chapter 8

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*.

8.1 Decision Problems and Procedures

As in the previous chapters, we consider a statistical experiment

$$\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big).$$

If there is no doubt about the sample space $(\Omega, \mathcal{A})$, we just write $\mathcal{E} = (\mathbb{P}_\theta)_{\theta\in\Theta}$.

Decision spaces. A *decision space* is a measurable space $(\mathbb{V}, \mathcal{B})$ representing the possible conclusions we could draw about the unknown true parameter θ.

Loss functions. A *loss function* is a mapping

$$L : \mathbb{V} \times \Theta \to (-\infty, \infty]$$

such that $L(\cdot, \theta)$ is $\mathcal{B}$-measurable for any fixed $\theta \in \Theta$. Here $L(v, \theta)$ quantifies the loss (e.g. the costs) when drawing the conclusion $v \in \mathbb{V}$ while the true parameter equals $\theta \in \Theta$.

Example 8.1 (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 8.2 (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 to 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

$$L(v, \theta) \ := \ 1_{[v=1, \theta \in \Theta_0]} \lambda_1 + 1_{[v=0, \theta \in \Theta_1]} \lambda_2. \tag{8.1}$$

Decision problems. A triplet $(\mathcal{E}, (\mathbb{V}, \mathcal{B}), L)$, consisting of a statistical experiment, a decision space and a loss function is called a *decision problem*. If there is a standard σ-field $\mathcal{B}$ on $\mathbb{V}$, we just write $(\mathcal{E}, \mathbb{V}, L)$.

Decision procedures. A *non-randomized decision procedure* is a measurable mapping $\rho : \Omega \to \mathbb{V}$. That means, if we observe $X \sim \mathbb{P}_\theta$ with unknown $\theta \in \Theta$, then we draw the conclusion $\rho(X) \in \mathbb{V}$ about θ.

More generally, a *decision procedure* is a stochastic kernel ρ from $(\Omega, \mathcal{A})$ to $(\mathbb{V}, \mathcal{B})$. That means,

$$\rho : \Omega \times \mathcal{B} \to [0, 1]$$

is a mapping such that

$$\begin{aligned} &\text{for any } \omega \in \Omega, \quad \rho(\omega, \cdot) \text{ is a probability measure on } (\mathbb{V}, \mathcal{B}), \\ &\text{for any } B \in \mathcal{B}, \quad \rho(\cdot, B) \text{ is } \mathcal{A}\text{-measurable on } \Omega. \end{aligned}$$

Now the interpretation is that having observed $X \sim \mathbb{P}_\theta$, we draw a random conclusion about θ from the probability measure $\rho(X, \cdot)$.

With slightly ambiguous notation, a non-randomized decision procedure $\rho : \Omega \to \mathbb{V}$ 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 $R(\rho, \cdot) : \Theta \to [-\infty, \infty]$,

$$R(\rho, \theta) \ := \ \int_\Omega \int_{\mathbb{V}} L(v, \theta) \, \rho(\omega, dv) \ \mathbb{P}_\theta(d\omega).$$

In our explicit examples, the loss functions L are non-negative, so $R(\rho, \theta)$ is well-defined in $[0, \infty]$.

Example 8.1 (Point estimation, continued). A point estimator for $g(\theta)$ is a measurable mapping $\widehat{g} : \Omega \to \mathbb{R}^q$. This corresponds to the (non-randomized) decision procedure

$$\rho(\omega, \cdot) \;:=\; \delta_{\widehat{g}(\omega)}.$$

Its risk function is given by

$$R(\widehat{g}, \theta) \;=\; \int_\Omega \|\widehat{g} - g(\theta)\|^r \, d\,\mathbb{P}_\theta \;=\; \mathbb{E}_\theta\big(\|\widehat{g} - g(\theta)\|^r\big)$$

or

$$R(\widehat{g}, \theta) \;=\; \int_\Omega \psi(\widehat{g} - g(\theta)) \, d\,\mathbb{P}_\theta \;=\; \mathbb{E}_\theta\, \psi(\|\widehat{g} - g(\theta)).$$

Exercise 8.3 (De-randomisation for point estimation). This exercise shows that in the context of point estimation, it is often sufficient to consider non-randomized decision procedures, i.e. simple point estimators. Let $L(v, \theta) = \psi(v - g(\theta))$ with a convex function $\psi : \mathbb{R}^q \to \mathbb{R}$ such that $\psi(v) \to \infty$ as $\|v\| \to \infty$. Show that for any decision procedure ρ, i.e. a stochastic kernel from $(\Omega, \mathcal{A})$ to $(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d))$, there exists a point estimator $\widehat{g} : \Omega \to \mathbb{R}^d$ such that

$$R(\widehat{g}, \theta) \;=\; \int_\Omega L(\widehat{g}(\omega), \theta) \; \mathbb{P}_\theta(d\omega)$$

is no larger than

$$R(\rho, \theta) \;=\; \int_\Omega \int_{\mathbb{R}^d} L(v, \theta) \, \rho(\omega, dv) \; \mathbb{P}_\theta(d\omega).$$

Example 8.2 (Statistical tests, continued). With the decision space $\mathbb{V} = \{0, 1\}$, any decision procedure ρ may be written as

$$\rho(\omega, \cdot) \;=\; (1 - \varphi(\omega))\delta_0 + \varphi(\omega)\delta_1$$

for some measurable function $\varphi : \Omega \to [0, 1]$, i.e. a test on Ω. With the power function $\theta \mapsto \mathbb{E}_\theta(\varphi)$ of φ and the loss function L in (8.1),

$$R(\varphi, \theta) \;=\; 1_{[\theta \in \Theta_0]} \lambda_1 \, \mathbb{E}_\theta(\varphi) + 1_{[\theta \in \Theta_1]} \lambda_2 (1 - \mathbb{E}_\theta(\varphi)).$$

Bayes risks. Suppose that Θ itself is equipped with a σ-field $\mathcal{C}$, and suppose that the loss function $L : \Theta \times \mathbb{V} \to (-\infty, \infty]$ is $\mathcal{C} \otimes \mathcal{B}$-measurable. Further suppose that $\theta \mapsto \mathbb{P}_\theta(A)$ is $\mathcal{C}$-measurable for any fixed $A \in \mathcal{A}$. Then we may view $(\mathbb{P}_\theta)_{\theta \in \Theta}$ as a stochastic kernel, too, and consider the following Bayesian model: Let Π be a probability distribution on $(\Theta, \mathcal{C})$, a so-called *prior (distribution)*. One could imagine "mother nature" choosing a parameter $\theta \in \Theta$ from that distribution Π, and then, conditional on θ, we observe a random variable $X \sim \mathbb{P}_\theta$. The Bayes risk of a decision procedure ρ for the prior Π is defined as

$$R(\rho, \Pi) \;:=\; \int_\Theta R(\rho, \theta) \, \Pi(d\theta) \;=\; \int_\Theta \int_\Omega \int_{\mathbb{V}} L(\theta, v) \, \rho(\omega, dv) \; \mathbb{P}_\theta(d\omega) \, \Pi(d\theta).$$

8.2 Some Optimality Concepts and Results

Let $(\mathcal{E}, (\mathbb{V}, \mathcal{B}), L)$ 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, "$\min_\rho$" and "$\inf_\rho$" stand for the minimum and infimum, respectively, over all decision procedures ρ.

Admissibility. A decision procedure ρ_* is called *admissible*, if there exists no decision procedure ρ satisfying

$$R(\rho, \theta) \;\leq\; R(\rho_*, \theta) \quad \text{for all } \theta \in \Theta$$

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 8.4. *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 8.4. 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,

$$\inf_\rho R(\rho,\Pi) \;\le\; R(\rho_*,\Pi) \;\le\; \sup_{\theta\in\Theta} R(\rho_*,\theta) \;=\; R(\rho_*,\Pi_*) \;=\; \inf_\rho R(\rho,\Pi_*)$$

by assumption. Hence Π_* is least favourable. □

Lemma 8.5. *Let ρ_* be a decision procedure such that for a sequence $(\Pi_k)_{k\ge1}$ 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 8.5. For any decision procedure ρ_o,

$$\sup_{\theta\in\Theta} R(\rho_o,\theta) \;\ge\; \limsup_{k\to\infty} R(\rho_o,\Pi_k) \;\ge\; \lim_{k\to\infty} \inf_\rho R(\rho,\Pi_k) \;=\; \sup_{\theta\in\Theta} R(\rho_*,\theta).$$

□

Lemma 8.6. *Let Θ be a topological space equipped with its Borel-σ-field. Suppose that any decision procedure ρ has continuous, real-valued risk function. If Π_o is a prior on Θ such that $\Pi_o(U) > 0$ for any non-void open set $U \subset \Theta$, and if ρ_o is a Bayes-optimal decision procedure for Π_o with $R(\rho_o,\Pi_o) < \infty$, then ρ_o is admissible.*

Exercise 8.7. Suppose that each distribution $\mathbb{P}_\theta$ has a density f_θ with respect to some σ-finite measure M on $(\Omega,\mathcal{A})$ such that for every $\omega \in \Omega$, $\theta \to f_\theta(\omega)$ is continuous on Θ. Further, suppose that the loss function L is bounded, and for any $v \in \mathbf{V}$, let $\theta \mapsto L(v,\theta)$ be continuous on Θ. Show that each decision procedure has bounded and continuous risk function. (Hint: Use Scheffé's theorem and dominated convergence.)

Proof of Lemma 8.6. Suppose that ρ_o is not admissible. That means, there exists a decision procedure ρ such that $R(\rho,\cdot) \le R(\rho_o,\cdot)$ and $R(\rho,\theta_o) < R(\rho_o,\theta_o)$ for some $\theta_o \in \Theta$. Since both risk functions $R(\rho,\cdot)$ and $R(\rho_o,\cdot)$ are continuous, there exist an open set $U \subset \Theta$ and an $\epsilon > 0$ such that $R(\rho,\cdot) \le R(\rho_o,\cdot) - \epsilon$ on U. But then

$$\begin{aligned} R(\rho,\Pi_o) \;&=\; \int_{\Theta\setminus U} R(\rho,\theta)\,\Pi(d\theta) + \int_U R(\rho,\theta)\,\Pi(d\theta) \\ &\le\; \int_{\Theta\setminus U} R(\rho_o,\theta)\,\Pi(d\theta) + \int_U (R(\rho_o,\theta)-\epsilon)\,\Pi(d\theta) \\ &=\; R(\rho_o,\Pi_o) - \epsilon\Pi_o(U) \\ &<\; R(\rho_o,\Pi_o), \end{aligned}$$

a contradiction to Bayes-optimality of ρ_o. □

Example 8.8 (Gaussian location model). For a given sample size $n \in \mathbb{N}$ and a given standard deviation $\sigma > 0$, let $\mathcal{E} = \left(\mathcal{N}(\mu,\sigma^2)^{\otimes n}\right)_{\mu\in\mathbb{R}}$. The sample mean $\widehat{\mu}_*(\boldsymbol{x}) := \bar{x}$ is a minimax-optimal point estimator of $g(\mu) = \mu$, if

$$L(v,\mu) \;:=\; (v-\mu)^2.$$

To see this, note first that $R(\widehat{\mu}_*, \cdot) \equiv \sigma^2/n$. By Exercise 8.3, it suffices to consider non-randomized point estimators. Moreover, if $\Pi_k = \mathcal{N}(0, k)$, we know from Example 6.20 that the Bayes-optimal estimator is given by $\widehat{\mu}_k(\boldsymbol{x}) = n\bar{x}/(n + \sigma^2/k)$ with Bayes-risk $R(\widehat{\mu}_k, \Pi_k) = \sigma^2/(n + \sigma^2/k)$. Since this converges to σ^2/n as $k \to \infty$, Lemma 8.5 shows that $\widehat{\mu}_*$ is minimax-optimal.

Exercise 8.9 (Point estimation of a binomial parameter). For $p \in [0,1]$ let $\mathbb{P}_p := \mathrm{Bin}(n, p)$. We consider point estimators of p with loss function $L : [0,1] \times [0,1] \to [0, \infty)$ given by

$$L(v, p) \;:=\; (v - p)^2.$$

(a) Consider the Bayesian model with a random parameter $p \sim \mathrm{Beta}(a, b)$ with given "hyper-parameters" $a, b > 0$, and a random observation X with $\mathcal{L}(X \,|\, p) = \mathbb{P}_p$. Here $\mathrm{Beta}(a, b)$ is the distribution on $(0, 1)$ with Lebesgue density

$$\beta_{a,b}(p) \;=\; \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \quad B(a,b) \;:=\; \int_0^1 u^{a-1}(1-u)^{b-1}\, du \;=\; \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Show that

$$\mathcal{L}(p \,|\, X) \;=\; \mathrm{Beta}(a + X, b + n - X) \quad \text{and} \quad \mathbb{E}(p \,|\, X) \;=\; \frac{a + X}{a + b + n}.$$

(b) Adapt arguments from previous chapters to show that the estimator

$$\widehat{p}_{a,b}(x) \;:=\; \frac{a + x}{a + b + n}$$

minimizes the Bayes risk

$$R(\widehat{p}, \mathrm{Beta}(a,b)) \;:=\; \int_0^1 R(\widehat{p}, p)\, \mathrm{Beta}(a,b)(dp).$$

(c) Determine the risk function $R(\widehat{p}_{a,b}, \cdot)$ of the estimator $\widehat{p}_{a,b}$ in part (b) explicitly.

(d) Now find parameters a, b such that the risk function in part (c) is constant. What are the consequences for the corresponding prior $\mathrm{Beta}(a, b)$ and the corresponding estimator $\widehat{p}_{a,b}$?

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 $\theta, \eta \in \Theta$,

$$R(\rho, \theta) \;=\; \int_\Omega \int_{\mathbb{V}} L(v, \theta)\, \rho(\omega, dv)\; \mathbb{P}_\theta(d\omega) \;\le\; \int_\Omega \int_{\mathbb{V}} L(v, \eta)\, \rho(\omega, dv)\; \mathbb{P}_\theta(d\omega). \tag{8.2}$$

Example 8.1 (Point estimation, continued). Let $L(v, \theta) := \|v - g(\theta)\|^2$ with some Euclidean norm $\|\cdot\|$ on $\mathbb{R}^q$, that is, $\|x\| = \sqrt{x^\top A x}$ for some symmetric, positive definite matrix $A \in \mathbb{R}^{q \times q}$.

For any estimator $\widehat{g} : \Omega \to \mathbb{R}^q$, the risk $R(\widehat{g}, \theta) = \mathbb{E}_\theta\big(\|\widehat{g} - g(\theta)\|^2\big)$ is finite if, and only if, $\mathbb{E}_\theta\big(\|\widehat{g}\|^2\big) < \infty$. In the latter case,

$$\begin{aligned}\mathbb{E}_\theta\big(\|\widehat{g} - g(\eta)\|^2\big) &= \mathbb{E}_\theta\big(\|\widehat{g} - \mathbb{E}_\theta(\widehat{g})\|^2\big) + \big\|\mathbb{E}_\theta(\widehat{g}) - g(\eta)\big\|^2,\\ \mathbb{E}_\theta\big(\|\widehat{g} - g(\theta)\|^2\big) &= \mathbb{E}_\theta\big(\|\widehat{g} - \mathbb{E}_\theta(\widehat{g})\|^2\big) + \big\|\mathbb{E}_\theta(\widehat{g}) - g(\theta)\big\|^2,\end{aligned}$$

so $\widehat{g}$ is unbiased if an only if for all $\theta \in \Theta$,

$$\big\|\mathbb{E}_\theta(\widehat{g}) - g(\theta)\big\| = \min_{\eta\in\Theta} \big\|\mathbb{E}_\theta(\widehat{g}) - g(\eta)\big\|.$$

Consequently, if $\widehat{g}$ is an estimator such that

$$\mathbb{E}_\theta\big(\|\widehat{g}\|^2\big) < \infty \quad \text{and} \quad \mathbb{E}_\theta(\widehat{g}) \in \text{closure}\{g(\eta) : \eta \in \Theta\}$$

for all $\theta \in \Theta$, then $\widehat{g}$ is unbiased if, and only if,

$$\mathbb{E}_\theta(\widehat{g}) = g(\theta) \quad \text{for all } \theta \in \Theta.$$

Example 8.2 (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

$$\int_\Omega \int_{\mathbb{V}} L(v, \eta)\, \rho(\omega, dv)\ \mathbb{P}_\theta(d\omega) \geq R(\varphi, \theta) \quad \text{for arbitrary } \theta, \eta \in \Theta$$

is easily shown to be equivalent to

$$\begin{cases}\lambda_2(1 - \mathbb{E}_{\theta_0}(\varphi)) \geq \lambda_1\, \mathbb{E}_{\theta_0}(\varphi) & \text{for any } \theta_0 \in \Theta_0,\\ \lambda_1\, \mathbb{E}_{\theta_1}(\varphi) \geq \lambda_2(1 - \mathbb{E}_{\theta_1}(\varphi)) & \text{for any } \theta_1 \in \Theta_1.\end{cases}$$

In other words,

$$\mathbb{E}_\theta(\varphi) \begin{cases}\leq \alpha & \text{for } \theta \in \Theta_0\\ \geq \alpha & \text{for } \theta \in \Theta_1\end{cases}$$

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_o with power at least α for each $\theta \in \Theta_A$ is unbiased in the sense of (8.2) if, say, $\lambda_1 = 1 - \alpha$ and $\lambda_2 = \alpha$.

Exercise 8.10 (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 $\big(\text{Bin}(n, p)\big)_{p\in[0,1]}$ 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 $g(p)$, i.e. $\mathbb{E}_p(\widehat{g}) = g(p)$ for all $p \in [0, 1]$. Show that $g(p)$ is a polynomial in p 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 $g(p) = p^k$ for some $k \in \{1, \ldots, n\}$. Hint: Consider factorials $[x]_k := \prod_{0\leq i<k}(x - i)$.

8.3 Informativity and Sufficiency

This section is about the comparison of two statistical experiments with one and the same parameter space Θ.

8.3.1 Informativity

Let $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ and $\tilde{\mathcal{E}} = (\tilde\Omega, \tilde{\mathcal{A}}, (\tilde{\mathbb{P}}_\theta)_{\theta\in\Theta})$ be statistical experiments. Experiment $\tilde{\mathcal{E}}$ is called *at least as informative as* experiment $\mathcal{E}$, if the following condition is satisfied: Let $(\mathbb{V}, \mathcal{B})$ be an arbitrary decision space and $L : \mathbb{V} \times \Theta \to (-\infty, \infty]$ be any loss function. Then for any decision procedure $\rho : \Omega \times \mathcal{B} \to [0, 1]$, there exists a decision procedure $\tilde\rho : \tilde\Omega \times \mathcal{B} \to [0, 1]$ 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 8.11 (Blackwell's criterion). *Let $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ and $\tilde{\mathcal{E}} = (\tilde\Omega, \tilde{\mathcal{A}}, (\tilde{\mathbb{P}}_\theta)_{\theta\in\Theta})$ be statistical experiments. Suppose there exists a stochastic kernel K from $(\tilde\Omega, \tilde{\mathcal{A}})$ to $(\Omega, \mathcal{A})$ such that for any $\theta \in \Theta$ and $A \in \mathcal{A}$,*

$$\mathbb{P}_\theta(A) \;=\; \tilde{\mathbb{P}}_\theta \otimes K(\tilde\Omega \times A) = \int_{\tilde\Omega} K(\tilde\omega, A)\, \tilde{\mathbb{P}}_\theta(d\tilde\omega).$$

Then $\tilde{\mathcal{E}}$ is at least as informative as $\mathcal{E}$.

The intuition behind Blackwell's criterion is as follows: 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: "$(\mathbb{V}, \mathcal{B})$ 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 experiments yields data $\tilde X \sim \tilde{\mathbb{P}}_\theta$, 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 $V \sim \rho(X, \cdot)$!" Bob: "That's cheap!"

Proof of Lemma 8.11. The assumption implies that for any measurable function $h : (\Omega, \mathcal{A}) \to [0, \infty]$,

$$\int_\Omega h\, d\,\mathbb{P}_\theta \;=\; \int_{\tilde\Omega} \int_\Omega h(\omega)\, K(\tilde\omega, d\omega)\, \tilde{\mathbb{P}}_\theta(d\tilde\omega).$$

By assumption, this is true for indicator functions $h = 1_A$ with $A \in \mathcal{A}$. 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 $h : (\Omega, \mathcal{A}) \to [0, \infty]$.

For a given decision space $(\mathbb{V}, \mathcal{B})$ and decision procedure $\rho : \Omega \times \mathcal{B} \to [0,1]$ we define a decision procedure $\tilde{\rho} : \tilde{\Omega} \times \mathcal{B} \to [0,1]$ as follows: For $B \in \mathcal{B}$ we set

$$\tilde{\rho}(\tilde{\omega}, B) \;:=\; \int_\Omega \rho(\omega, B)\, K(\tilde{\omega}, d\omega).$$

This construction of $\tilde{\rho}$ implies that for any measurable function $h : (\mathbb{V}, \mathcal{B}) \to [0,\infty]$ and $\tilde{\omega} \in \tilde{\Omega}$,

$$\int_{\mathbb{V}} h(v)\, \tilde{\rho}(\tilde{\omega}, dv) \;=\; \int_\Omega \int_{\mathbb{V}} h(v)\, \rho(\omega, dv)\, K(\tilde{\omega}, d\omega).$$

Hence for loss functions $L : \mathbb{V} \times \Theta \to [0,\infty]$ and any $\theta \in \Theta$,

$$\begin{aligned} R(\tilde{\rho}, \theta) \;&=\; \int_{\tilde{\Omega}} \int_{\mathbb{V}} L(v,\theta)\, \tilde{\rho}(\tilde{\omega}, dv)\, \tilde{\mathbb{P}}_\theta(d\tilde{\omega}) \\ &=\; \int_{\tilde{\Omega}} \int_\Omega \int_{\mathbb{V}} L(v,\theta)\, \rho(\omega, dv)\, K(\tilde{\omega}, d\omega)\, \tilde{\mathbb{P}}_\theta(d\tilde{\omega}) \\ &=\; \int_\Omega \int_{\mathbb{V}} L(v,\theta)\, \rho(\omega, dv)\; \mathbb{P}_\theta(d\omega) \\ &=\; R(\rho, \theta). \end{aligned}$$

The same equation is true for loss functions L with values in $(-\infty,\infty]$, provided that $R(\rho',\theta)$ or $R(\rho,\theta)$ is well-defined in $(-\infty,\infty]$. Just write $L = L^+ - L^-$ with $L^\pm := \max(\pm L, 0)$. □

Example 8.12 (Sampling with and without replacement). Let $\mathcal{M}$ be a population of *known* size $N = \#\mathcal{M}$, but with unknown characteristics $\theta \in \Theta$, the latter describing certain properties of the individuals in $\mathcal{M}$. Suppose we draw a random sample of size $n \geq 2$ from $\mathcal{M}$ *with* replacement. That means, we obtain a sample $\omega = (\omega_1, \ldots, \omega_n)$ in the set $\Omega := \mathcal{M}^n$ with N^n different elements. Here the corresponding distributions $\mathbb{P}_\theta$, $\theta \in \Theta$, are all identical: the uniform distribution on Ω. The subscript θ just indicates that the potential samples may differ in terms of certain characteristics of the individuals.

In case of $N \geq n$, sampling *without* replacement would be an alternative strategy, leading to the experiment $\tilde{\mathcal{E}} = (\tilde{\Omega}, \mathcal{P}(\tilde{\Omega}), (\tilde{\mathbb{P}}_\theta)_{\theta\in\Theta})$ with $\tilde{\mathbb{P}}_\theta$ denoting the uniform distribution on

$$\tilde{\Omega} \;:=\; \big\{\tilde{\omega} \in \Omega : \tilde{\omega}_i \neq \tilde{\omega}_j \text{ for } 1 \leq i < j \leq n\big\},$$

a set of $[N]_n = N(N-1)\cdots(N-n+1)$ different samples. Indeed the experiment $\tilde{\mathcal{E}}$ is at least as informative as $\mathcal{E} = (\Omega, \mathcal{P}(\Omega), (\mathbb{P}_\theta)_{\theta\in\Theta})$. For if we draw a sample $\tilde{\omega} \in \tilde{\Omega}$ from the uniform distribution on $\tilde{\Omega}$, we may generate a new sample $\omega \in \Omega$ as follows: We choose $\omega_1 := \tilde{\omega}_1$ and independent random points $\omega_2, \ldots, \omega_n$ with distribution

$$\omega_j \;\sim\; \sum_{i=1}^{j-1} \frac{1}{N} \delta_{\tilde{\omega}_i} + \frac{N-j+1}{N} \delta_{\tilde{\omega}_j}.$$

In other words, we go through the "list" $\tilde{\omega}$, and each component $\tilde{\omega}_j$ is kept with probability $1 - (j-1)/N$ or replaced with a randomly chosen predecessor $\tilde{\omega}_i$, $i < j$. This defines a probability distribution $K(\tilde{\omega}, \cdot)$ on Ω such that the resulting ω is uniformly distributed on Ω.

Exercise 8.13 (From sampling without to sampling with replacement). Show that the construction of ω from $\tilde{\omega}$ in Example 8.12 corresponds to a stochastic kernel K from $\tilde{\Omega}$ to Ω such that

$$K\big(\tilde{\omega}, \{\tilde{\omega}_1, \ldots, \tilde{\omega}_n\}^n\big) \;=\; 1$$

and for any $y \in \mathcal{M}^n$,

$$\frac{1}{[N]_n} \sum_{\tilde{\omega} \in \tilde{\Omega}} K(\tilde{\omega}, \{y\}) \;=\; \frac{1}{N^n}.$$

Exercise 8.14 (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 $n \leq N$ without replacement from that population reveals nothing about the population size 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 $g(\mathcal{M}) := 1/N$.

(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).

8.3.2 Sufficiency

The concept of sufficiency is a special instance of Blackwell's criterion. We consider a statistical experiment $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$. 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 $X \sim \mathbb{P}_\theta$ with a given function $T(X)$, or we want to restrict our attention to events A_o in a sub-σ-field $\mathcal{A}_o$ of $\mathcal{A}$.

Definition 8.15 (Sufficient statistic). A measurable mapping $T : (\Omega, \mathcal{A}) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$ is called a *sufficient statistic for* $\mathcal{E}$, if there exists a stochastic kernel K from $(\tilde{\Omega}, \tilde{\mathcal{A}})$ to $(\Omega, \mathcal{A})$ describing the conditional distribution of $X \sim \mathbb{P}_\theta$, given $T(X)$, for any $\theta \in \Theta$. In other words, for arbitrary $\theta \in \Theta$, $A \in \mathcal{A}$ and $\tilde{A} \in \tilde{\mathcal{A}}$,

$$\mathbb{P}_\theta(\{T \in \tilde{A}\} \cap A) \;=\; \int_{\tilde{A}} K(t, A)\; \mathbb{P}_\theta^T(dt).$$

Sufficiency of T implies that the experiment $\mathcal{E}^T := (\tilde{\Omega}, \tilde{\mathcal{A}}, (\mathbb{P}_\theta^T)_{\theta \in \Theta})$ 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 $T(\boldsymbol{X})$, one could generate an artificial observation $\tilde{X} \sim K(T(X), \cdot)$, and $\tilde{X} \sim \mathbb{P}_\theta$, too, no matter what value the unknown parameter $\theta \in \Theta$ has. Sufficiency can often be verified with the following criterion:

Theorem 8.16 (Neyman's factorization criterion). *Suppose that (Ω, d) is a separable and complete metric space, and let $\mathcal{A} = \mathrm{Borel}(\Omega, d)$. Further let M be a σ-finite measure on $(\Omega, \mathcal{A})$ such that each distribution $\mathbb{P}_\theta$ has a density f_θ with respect to M. Suppose that there exists a measurable function $h : (\Omega, \mathcal{A}) \to [0, \infty)$ such that for any $\theta \in \Theta$ and $\omega \in \Omega$,*

$$f_\theta(\omega) \;=\; g_\theta(T(\omega))h(\omega)$$

with $g_\theta : (\tilde{\Omega}, \tilde{\mathcal{A}}) \to [0, \infty)$ measurable. Then T is a sufficient statistic for $\mathcal{E}$.

Corollary 8.17. *Let Ω and $\tilde{\Omega}$ be countable sets equipped with $\mathcal{A} = \mathcal{P}(\Omega)$ and $\tilde{\mathcal{A}} = \mathcal{P}(\tilde{\Omega})$. Suppose that there exists a function $h : \Omega \to [0, \infty)$ such that for any $\theta \in \Theta$ and $\omega \in \Omega$,*

$$\mathbb{P}_\theta(\{\omega\}) \;=\; g_\theta(T(\omega))h(\omega)$$

with $g_\theta : \tilde{\Omega} \to [0, \infty)$. Then T is a sufficient statistic for $\mathcal{E}$.

For readers feeling uneasy about measure theory it may be instructive to prove the latter corollary directly. It is a consequence of Theorem 8.16 if we use the metric $d(\tilde{\omega}, \tilde{\omega}') := 1_{[\tilde{\omega} \neq \tilde{\omega}']}$ on Ω and the counting measure M on Ω, that is, $M(\{\omega\}) = 1$ for all $\omega \in \Omega$.

Proof of Theorem 8.16. A measure M on $(\Omega, \mathcal{A})$ is σ-finite if, and only if, there exists a measurable function $J : (\Omega, \mathcal{A}) \to (0, \infty)$ such that $\int J \, dM = 1$. But then we could replace M with the measure $J \cdot M$, i.e. $A \mapsto \int_A J \, dM$, and h with h/J. Hence we may and do assume without loss of generality that M is a probability measure. By our assumption on $(\Omega, \mathcal{A})$ there exists a stochastic kernel K_o from $(\tilde{\Omega}, \tilde{\mathcal{A}})$ to $(\Omega, \mathcal{A})$ such that for arbitrary $\tilde{A} \in \tilde{\mathcal{A}}$ and $A \in \mathcal{A}$,

$$M(\{T \in \tilde{A}\} \cap A) \;=\; \int_{\tilde{A}} K_o(t, A) \, M^T(dt),$$

see Chapter 4. More generally, for arbitrary measurable and non-negative functions f on $(\tilde{\Omega} \times \Omega, \tilde{\mathcal{A}} \otimes \mathcal{A})$,

$$\int_\Omega f(T(\omega), \omega) \, M(d\omega) \;=\; \int_{\tilde{\Omega}} \int_\Omega f(t, \omega) \, K_o(t, d\omega) \, M^T(dt).$$

This implies that for arbitrary $\theta \in \Theta$, $\tilde{A} \in \tilde{\mathcal{A}}$ and $A \in \mathcal{A}$,

$$\begin{aligned} \mathbb{P}_\theta(\{T \in \tilde{A}\} \cap A) &= \int_\Omega 1_{\tilde{A}}(T(\omega)) 1_A(\omega) g_\theta(T(\omega)) h(\omega) \, M(d\omega) \\ &= \int_{\tilde{\Omega}} \int_\Omega 1_{\tilde{A}}(t) g_\theta(t) 1_A(\omega) h(\omega) \, K_o(t, d\omega) \, M^T(dt) \\ &= \int_{\tilde{A}} g_\theta(t) \int_A h(\omega) \, K_o(t, d\omega) \, M^T(dt). \end{aligned}$$

Taking $A = \Omega$ shows that the density of $\mathbb{P}_\theta^T$ with respect to M^T is given by

$$f_\theta^T(t) \;:=\; g_\theta(t) H(t) \quad \text{with} \quad H(t) \;:=\; \int_\Omega h(\omega) \, K_o(t, d\omega).$$

In particular, the set $N := \{t \in \tilde{\Omega} : H(t) = 0 \text{ or } H(t) = \infty\}$ satisfies $\mathbb{P}_\theta^T(N) = 0$ for any $\theta \in \Theta$, because $\{H = 0\} \subset \{f_\theta^T = 0\}$ and $\{H = \infty\} \subset \{f_\theta^T = 0\} \cup \{f_\theta^T = \infty\}$, and both $\mathbb{P}_\theta^T(f_\theta^T = 0)$ and $\mathbb{P}_\theta^T(f_\theta^T = \infty)$ are equal to zero. Hence

$$K(t, A) \;:=\; \begin{cases} H(t)^{-1} \displaystyle\int_A h(\omega)\, K_o(t, d\omega) & \text{if } 0 < H(t) < \infty \\ M(A) & \text{else} \end{cases}$$

defines a stochastic kernel K from $(\tilde{\Omega}, \tilde{\mathcal{A}})$ to $(\Omega, \mathcal{A})$ such that

$$\mathbb{P}_\theta(\{T \in \tilde{A}\} \cap A) \;=\; \int_{\tilde{A}} f_\theta^T(t) K(t, A)\, M^T(dt) \;=\; \int_{\tilde{A}} K(t, A)\; \mathbb{P}_\theta^T(dt)$$

for arbitrary $\theta \in \Theta$, $\tilde{A} \in \tilde{\mathcal{A}}$ and $A \in \mathcal{A}$. □

Example 8.18 (Bernoulli sequences). Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with values in $\{0, 1\}$ and unknown parameter $p = \mathbb{P}(X_i = 1) = \mathbb{E}(X_i) \in [0, 1]$. This leads to the statistical experiment $\mathcal{E} = \big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathbb{P}_p)_{p\in[0,1]}\big)$ with $\mathbb{P}_p := ((1-p)\delta_0 + p\delta_1)^{\otimes n}$. The density f_p of $\mathbb{P}_p$ with respect to counting measure on $\{0, 1\}^n$ is given by

$$f_p(\omega) \;=\; \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i} \;=\; p^{T(\omega)}(1-p)^{n-T(\omega)},$$

i.e. a function of $T(\omega) := \sum_{i=1} \omega_i$ only. Hence T is a sufficient statistic for $\mathcal{E}$. Indeed, T has distribution $\mathrm{Bin}(n, p)$, and for any $t \in \{0, 1, \ldots, n\}$, the conditional distribution $\mathbb{P}_p(\cdot \,|\, T = t)$ is the uniform distribution on the set of all $\omega \in \{0, 1\}^n$ with $T(\omega) = t$.

Example 8.19 (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 $\mathcal{E} = \big(\mathbb{R}^n, \mathrm{Borel}(\mathbb{R}^n), (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ with $\Theta = \mathbb{R} \times (0, \infty)$ and $\mathbb{P}_{\mu,\sigma} := \mathcal{N}(\mu, \sigma^2)^{\otimes n}$. Setting $X_i(\omega) := \omega_i$, the density $f_{\mu,\sigma}$ of $\mathbb{P}_{\mu,\sigma}$ 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 $\mathbb{P}_{\mu,\sigma}$ coincides with the distribution of

$$\tilde{\boldsymbol{X}} \;:=\; \mu \mathbf{1}_n + \sigma \sum_{i=1}^n Z_i \boldsymbol{b}_i$$

with stochastically independent random variables $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$. The statistic $T(\tilde{\boldsymbol{X}})$ 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

$$\tilde{\boldsymbol{X}} \;=\; T_1(\tilde{\boldsymbol{X}})\mathbf{1}_n + \sqrt{T_2(\tilde{\boldsymbol{X}})} \sum_{i=2}^n W_i \boldsymbol{b}_i$$

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 8.20 (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 8.21 (Markov chains with finite state space). Let $\boldsymbol{X} = (X_t)_{t=0}^n$ be a Markov chain with values in a finite set $\mathcal{X}$ and fixed starting point $X_0 = x_0 \in \mathcal{X}$. 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" $\theta \in [0,1]^{\mathcal{X}\times\mathcal{X}}$ such that

$$\sum_{z\in\mathcal{X}} \theta_{y,z} \;=\; 1 \quad \text{for all } y \in \mathcal{X}.$$

Let Θ be the set of all such "matrices" θ. Determine a sufficient statistic for the statistical experiment $\big(\Omega, \mathcal{P}(\Omega), (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$, where $\Omega = \{\boldsymbol{y} \in \mathcal{X}^{\{0,1,\ldots,n\}} : y_0 = x_0\}$.

Exercise 8.22. 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 $b - a > n$ 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 $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ of identification numbers.

Show that $T(\omega) = \big(\min(\omega), \max(\omega)\big)$ is a sufficient statistic for this experiment. Describe the conditional distribution of ω given $T(\omega) = (s_1, s_2)$.

Hint: The formal definition of the experiment is

$$\begin{aligned} \Omega &= \{\omega \in \mathbb{Z}^n : \omega_i \neq \omega_j \text{ whenever } i \neq j\}, \\ \mathbb{P}_{a,b} &= \mathrm{Unif}\big(\Omega \cap \{a, \dots, b\}^n\big), \\ \Theta &= \big\{(a,b) \in \mathbb{Z}^n : b - a > n\big\}. \end{aligned}$$

Definition 8.23 (Sufficient sub-σ-fields). Let $\mathcal{A}_o$ be a σ-field over Ω such that $\mathcal{A}_o \subset \mathcal{A}$. It is called a *sufficient sub-σ-field for $\mathcal{E}$*, if there exists a stochastic kernel K from $(\Omega, \mathcal{A}_o)$ to $(\Omega, \mathcal{A})$ describing the conditional distribution of $X \sim \mathbb{P}_\theta$, given $\mathcal{A}_o$, for any $\theta \in \Theta$. In other words, for arbitrary $\theta \in \Theta$, $A \in \mathcal{A}$ and $A_o \in \mathcal{A}_o$,

$$\mathbb{P}_\theta(A_o \cap A) \;=\; \int_{A_o} K(\omega, A)\; \mathbb{P}_\theta(d\omega).$$

Note that sufficiency of $\mathcal{A}_o$ is equivalent to sufficiency of the statistic

$$T : (\Omega, \mathcal{A}) \;\to\; (\Omega, \mathcal{A}_o), \quad T(\omega) := \omega.$$

Sufficiency of $\mathcal{A}_o$ implies that the experiment $\mathcal{E}_o := (\Omega, \mathcal{A}_o, (\mathbb{P}_\theta)_{\theta \in \Theta})$ is at least as informative as $\mathcal{E}$. In other words, when analyzing raw data $X \sim \mathbb{P}_\theta$, we may restrict our attention to decision procedures which are $\mathcal{A}_o$-measureable rather than $\mathcal{A}$-measurable.

Example 8.24 (Invariant distributions). Let $\mathcal{G}$ be a finite group of measurable bijective mappings $g : (\Omega, \mathcal{A}) \to (\Omega, \mathcal{A})$. That means, for arbitrary $g, h \in \mathcal{G}$, both $h \circ g$ and g^{-1} belong to $\mathcal{G}$, too. Now let $\mathcal{A}_\mathcal{G}$ be the set of $\mathcal{G}$-invariant sets $A \in \mathcal{A}$, i.e.

$$g(A) = \{g(\omega) : \omega \in A\} \;=\; A \quad \text{for all } g \in \mathcal{G}.$$

This is obviously a sub-σ-field of $\mathcal{A}$.

Now suppose that all distributions $\mathbb{P}_\theta$, $\theta \in \Theta$, are $\mathcal{G}$-invariant in the sense that

$$\mathbb{P}_\theta^g \;=\; \mathbb{P}_\theta \quad \text{for all } g \in \mathcal{G},$$

where $\mathbb{P}_\theta^g$ is the image measure $\mathbb{P}_\theta \circ g^{-1}$. Then $\mathcal{A}_\mathcal{G}$ is sufficient for $\mathcal{E}$, and the conditional distribution $\mathbb{P}_\theta(\cdot \,|\, \mathcal{A}_o)$ is given by the stochastic kernel

$$K(\omega, A) \;:=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} 1_A(g(\omega))$$

i.e.

$$K(\omega, \cdot) \;=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} \delta_{g(\omega)}.$$

Hence the experiment $\mathcal{E}_{\mathcal{G}} := \big(\Omega, \mathcal{A}_{\mathcal{G}}, (\mathbb{P}_\theta)_{\theta \in \Theta}\big)$ is at least as informative as $\mathcal{E}$.

Proof: Obviously, $K(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{A})$ for any $\omega \in \Omega$. For fixed $A \in \mathcal{A}$, the function $\omega \to K(\omega, A)$ is certainly $\mathcal{A}$-measurable. To verify $\mathcal{A}_{\mathcal{G}}$-measurability it suffices to show that $K(h(\omega), A) = K(\omega, A)$ for arbitrary $\omega \in \Omega$ and $h \in \mathcal{G}$, see Exercise 8.25 below. But

$$K(h(\omega), A) \;=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} 1_A(g \circ h(\omega)) \;=\; \frac{1}{\#\mathcal{G}} \sum_{\tilde{g} \in \mathcal{G}} 1_A(\tilde{g}(\omega)),$$

because for any fixed $h \in \mathcal{G}$, the mapping $\mathcal{G} \ni\!\mapsto g \circ h \in \mathcal{G}$ is bijective, see Exercise 5.1. It remains to be shown that for arbitrary $\theta \in \Theta$, $A_o \in \mathcal{A}_{\mathcal{G}}$ and $A \in \mathcal{A}$,

$$\mathbb{P}_\theta(A_o \cap A) \;=\; \int_{A_o} K(\omega, A)\; \mathbb{P}_\theta(d\omega).$$

But the right hand side equals

$$\frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} \int_{A_o} 1_A(g(\omega))\; \mathbb{P}_\theta(d\omega),$$

and each summand equals

$$\begin{aligned}
\int_{A_o} 1_A(g(\omega))\; \mathbb{P}_\theta(d\omega) \;&=\; \mathbb{P}_\theta(A_o \cap g^{-1}(A)) \\
&=\; \mathbb{P}_\theta(g^{-1}(g(A_o) \cap A)) \\
&=\; \mathbb{P}_\theta(g^{-1}(A_o \cap A)) \qquad\qquad \text{(by } \mathcal{G}\text{-invariance of } A_o) \\
&=\; \mathbb{P}_\theta^g(A_o \cap A) \\
&=\; \mathbb{P}_\theta(A_o \cap A) \qquad\qquad \text{(by } \mathcal{G}\text{-invariance of } \mathbb{P}_\theta).
\end{aligned}$$

Exercise 8.25. In the setting of Example 8.24, let $(\mathbb{V}, \mathcal{B})$ be another measurable space, and let $\rho : (\Omega, \mathcal{A}) \to (\mathbb{V}, \mathcal{B})$ 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 $\mathcal{A}_{\mathcal{G}}$-$\mathcal{B}$-measurable.

(a) Suppose that ρ is $\mathcal{A}_{\mathcal{G}}$ measurable, and suppose that $\mathcal{B}$ "seqarates points in $\mathbb{V}$". That means, for arbitrary different points $v_1, v_2 \in \mathbb{V}$ there exists a set $B \in \mathcal{B}$ such that $v_1 \in B$ but $v_2 \notin B$. Show that ρ is $\mathcal{G}$-invariant.

Example 8.26 (Permutation-invariance). For some integer $n \geq 2$ let $(\Omega, \mathcal{A}) = (\mathcal{X}^n, \mathcal{B}^{\otimes n})$. 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 $g_\pi : \Omega \to \Omega$,

$$\omega = (\omega_i)_{i=1}^n \;\mapsto\; g_\pi(\omega) \;:=\; (\omega_{\pi(i)})_{i=1}^n.$$

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 $(\Omega, \mathcal{A})$.

A distribution $\mathbb{P}$ on $(\Omega, \mathcal{A})$ 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 $\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ consists of permutation-invariant distributions, then the sub-σ-field $\mathcal{A}_\mathcal{G}$ of all permutation-invariant sets is sufficient for $\mathcal{E}$. The corresponding stochastic kernel K from $(\Omega, \mathcal{A}_\mathcal{G})$ to $(\Omega, \mathcal{A})$ is given by

$$K(\omega, \cdot) \;=\; \frac{1}{n!} \sum_{\pi\in\mathcal{S}_n} \delta_{g_\pi(\omega)}.$$

Example 8.27 (Sign invariance). For some integer $n \geq 1$ let $\Omega = \mathbb{R}^n$. Any sign vector $\xi \in \{-1, 1\}^n$ induces a measurable bijection $g_\xi : \mathbb{R}^n \to \mathbb{R}^n$,

$$\omega = (\omega_i)_{i=1}^n \;\mapsto\; g_\xi(\omega) \;:=\; (\xi_i\omega_i)_{i=1}^n.$$

Note that $\{-1, 1\}^n$ with component-wise multiplication is an Abelian group. The corresponding family $\mathcal{G} := \big\{g_\xi, \xi \in \{-1, 1\}\big\}$ 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 $X = (X_i)_{i=1}^n$ 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 sets is sufficient for $\mathcal{E}$. The corresponding stochastic kernel K from $(\Omega, \mathcal{A}_\mathcal{G})$ to $(\Omega, \mathcal{A})$ is given by

$$K(\omega, \cdot) \;=\; \frac{1}{2^n} \sum_{\xi\in\{-1,1\}^n} \delta_{g_\xi(\omega)}.$$

8.4 Complete Statistical Experiments

Definition 8.28 (Complete statistical experiment). A statistical experiment $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ is called *boundedly complete*, if for any bounded and measurable function $f : (\Omega, \mathcal{A}) \to \mathbb{R}$,

$$\int f \, d\,\mathbb{P}_\theta \;=\; 0 \quad \text{for all } \theta \in \Theta$$

implies that

$$\mathbb{P}_\theta(f \neq 0) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

The experiment $\mathcal{E}$ is called *complete* if for any function $f \in \bigcap_{\theta\in\Theta} \mathcal{L}^1(\mathbb{P}_\theta)$,

$$\int f \, d\,\mathbb{P}_\theta \;=\; 0 \quad \text{for all } \theta \in \Theta$$

implies that

$$\mathbb{P}_\theta(f \neq 0) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

Obviously, completeness of $\mathcal{E}$ implies bounded completeness.

Example 8.29 (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 $\Theta \subset \mathbb{R}^d$,

$$0 \;<\; \int \exp(\theta^\top x) h(x)\, M(dx) \;<\; \infty.$$

Then we consider the statistical experiment $\mathcal{E} = \big(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (\mathbb{P}_\theta)_{\theta\in\mathbb{R}^d}\big)$ where

$$\frac{d\,\mathbb{P}_\theta}{dM}(x) \;=\; C(\theta) \exp(\theta^\top x) h(x)$$

with $C(\theta) := \big(\int \exp(\theta^\top x) h(x)\, M(dx)\big)^{-1}$. If Θ has nonempty interior, then $\mathcal{E}$ is complete.

This follows immediately from Theorem A.15 in Appendix A.3.

Example 8.30 (Product measures). Let $(\mathcal{X}, \mathcal{B}, M)$ be a σ-finite measure space, and let Θ be a family of probability densities with respect to M. For some integer $n \geq 2$ consider $\mathcal{E} = (\mathcal{X}^n, \mathcal{B}^{\otimes n}, (Q_\theta^{\otimes n})_{\theta\in\Theta})$ with Q_θ given by $dQ_\theta/dM = \theta$. This experiment is *not* boundedly complete. To see this, consider some bounded measurable function $h : \mathcal{X} \to \mathbb{R}$ with $\mathrm{Var}_{\theta_o}(h) > 0$ for at least one $\theta_o \in \Theta$. Then $f(\omega) := h(\omega_1) - h(\omega_2)$ defines a bounded and measurable function $f : \mathcal{X}^n \to \mathbb{R}$ such that $\int f\, dQ_\theta^{\otimes n} = 0$ for any $\theta \in \Theta$, but $Q_{\theta_o}^{\otimes n}(f \neq 0) > 0$, because $\int f^2\, dQ_{\theta_o}^{\otimes n} = 2\,\mathrm{Var}_{\theta_o}(h) > 0$.

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 $\mathcal{E}_o = (\mathcal{X}^n, \mathcal{A}_o, (Q_\theta^{\otimes n})_{\theta\in\Theta})$ is often boundedly complete or even complete:

Special case 1. Suppose that Θ contains the convex hull of all probability densities $M(B)^{-1}1_B$ with $B \in \mathcal{B}$ satisfying $0 < M(B) < \infty$. Then $\mathcal{E}_o$ is boundedly complete.

Proof: For arbitrary fixed sets $B_1, \ldots, B_n \in \mathcal{B}$ with $0 < M(B_i) < \infty$ and arbitrary tupels $\gamma \in (0, \infty)^n$, consider the probability density

$$\theta_\gamma \;:=\; C_\gamma^{-1} \sum_{i=1}^n \gamma_i 1_{B_i}$$

with $C_\gamma := \sum_{i=1}^n \gamma_i M(B_i)$. Then by assumption, $\theta_\gamma \in \Theta$. If $f : (\Omega, \mathcal{A}) \to \mathbb{R}$ is bounded and measurable, it follows from

$$\int f\, dQ_\theta^{\otimes n} \;=\; 0 \quad \text{for all } \theta \in \Theta$$

that

$$\int f(x) \prod_{i=1}^n \Big(\sum_{j=1}^n \gamma_j 1_{B_j}(x_i) \Big)\, M^{\otimes n}(dx) \;=\; 0 \quad \text{for all } \gamma \in (0,\infty)^n. \tag{8.3}$$

The integral on the left hand side equals

$$\sum_{j(1),\ldots,j(n)=1}^n \prod_{i=1}^n \gamma_{j(i)} \int_{B_{j(1)}\times\cdots\times B_{j(n)}} f\, dM^{\otimes n},$$

and this is an n-variate polynomial in γ of degree n. Since

$$\frac{\partial^n}{\partial\gamma_1\partial\gamma_2\cdots\partial\gamma_n}\prod_{i=1}^n\gamma_{j(i)} \;=\; \begin{cases}1 & \text{if } \{j(1),\ldots,j(n)\}=\{1,\ldots,n\}\\ 0 & \text{else,}\end{cases}$$

it follows from (8.3) that

$$\sum_{\pi\in\mathcal{S}_n}\int_{B_{\pi(1)}\times\cdots\times B_{\pi(n)}} f\,dM^{\otimes n} \;=\; 0.$$

If f is permutation-invariant, that means, $\mathcal{A}_o$-measurable rather than just $\mathcal{A}$-measurable, then the latter sum equals

$$n!\int_{B_1\times\cdots\times B_n} f\,dM^{\otimes n}.$$

All in all, we know that

$$\int_{B_1\times\cdots\times B_n} f\,dM^{\otimes n} \;=\; 0$$

for arbitrary sets $B_1,\ldots,B_n\in\mathcal{B}$ with finite measure $M(B_i)$. As shown in Exercise 8.31, this implies that $M^{\otimes n}(f\neq 0)=0$. □

Exercise 8.31. Let M be a σ-finite measure on $(\mathcal{X},\mathcal{B})$, and let $f:\mathcal{X}^n\to\mathbb{R}$ be bounded and measurable such that

$$\int_{B_1\times\cdots\times B_n} f\,dM^{\otimes n} \;=\; 0$$

for arbitrary sets $B_1,\ldots,B_n\in\mathcal{B}$ with finite measure $M(B_i)$. Show that

$$M^{\otimes n}(f\neq 0) \;=\; 0.$$

Hint: Consider first the case of $M(\mathcal{X})<\infty$ and the two finite measures $Q^{\pm}$ given by $Q^{\pm}(A):=\int_A f^{\pm}\,dM^{\otimes n}$.

Special case 2. Let M be Lebesgue measure on $\mathcal{X}=\mathbb{R}$. Suppose that for an arbitrary fixed $\sigma>0$, the set Θ contains all finite convex combinations of Gaussian densities $\phi_{\mu,\sigma}$, $\mu\in\mathbb{R}$, where

$$\phi_{\mu,\sigma}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

Then $\mathcal{E}_o$ is complete.

Proof: With the same arguments as in the previous special case, one can show that for any $\mathcal{A}_o$-measurable function $f:\mathbb{R}^n\to\mathbb{R}$, it follows from

$$\int f\,dQ_\theta^{\otimes n} \;=\; 0 \quad \text{for arbitrary } \theta\in\Theta$$

that

$$\int f(\boldsymbol{x})\prod_{i=1}^n\phi_{\mu_i,\sigma}(x_i)\,d\boldsymbol{x} \;=\; 0 \quad \text{for arbitrary } \boldsymbol{\mu}\in\mathbb{R}^n,$$

and this is equivalent to

$$\int f(\boldsymbol{x})\exp\big(-\|\boldsymbol{x}\|^2/(2\sigma^2)\big)\exp(\boldsymbol{\mu}^\top\boldsymbol{x})\,d\boldsymbol{x} \;=\; 0 \quad \text{for arbitrary } \boldsymbol{\mu}\in\mathbb{R}^n.$$

But then it follows from Theorem A.15 in the Appendix that the two finite measures $A\mapsto\int_A f^{\pm}\,dM^{\otimes n}$ are identical, whence $M^{\otimes n}(f\neq 0)=0$. □

Unbiased estimation. With the concepts of sufficiency and completeness one can say something about unbiased point estimators. This was first noted by P.R. Halmos [7].

Theorem 8.32 (Halmos). *Let $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical experiment with a sufficient statistic $T : (\Omega, \mathcal{A}) \to (\tilde\Omega, \tilde{\mathcal{A}})$. 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 for any fixed $\theta \in \Theta$.*

(a) *Suppose $\widehat g : (\Omega, \mathcal{A}) \to \mathbb{R}^q$ is an unbiased estimator of $g(\theta)$. That means,*

$$\mathbb{E}_\theta(\widehat g) \;=\; g(\theta) \quad \text{for any } \theta \in \Theta.$$

Then

$$\check g(\omega) \;:=\; \mathbb{E}(\widehat g \,|\, T = T(\omega))$$

defines another unbiased estimator with

$$\mathbb{E}_\theta\, L(\check g, \theta) \;\le\; \mathbb{E}_\theta\, L(\widehat g, \theta) \quad \text{for any } \theta \in \Theta.$$

If $L(\cdot, \theta)$ is strictly convex, then the latter inequality is strict unless $\mathbb{P}_\theta(\widehat g \neq \check g) = 0$.

(b) *If $\mathcal{E}^T = (\tilde\Omega, \tilde{\mathcal{A}}, (\mathbb{P}^T_\theta)_{\theta\in\Theta})$ is complete, then the unbiased estimator $\check g$ in part (a) is essentially unique. That means, if $\tilde g = \tilde h(T)$ is another unbiased estimator of $g(\theta)$, then $\mathbb{P}_\theta(\tilde g \neq \check g) = 0$ for arbitrary $\theta \in \Theta$.*

Proof of Theorem 8.32. By sufficiency of T, there exists a stochastic kernel K from $(\tilde\Omega, \tilde{\mathcal{A}})$ to $(\Omega, \mathcal{A})$ describing the conditional distribution of $X \sim \mathbb{P}_\theta$, given $T(X)$, for any parameter $\theta \in \Theta$. This gives rise to conditional expectations $\mathbb{E}(h \,|\, T = t) := \int h(\omega)\, K(t, d\omega)$. Note that by assumption, for any $\theta \in \Theta$,

$$\infty \;>\; \mathbb{E}_\theta \,\|\widehat g\| \;=\; \int_{\tilde\Omega} \mathbb{E}\big(\|\widehat g\| \,\big|\, T = t\big)\; \mathbb{P}^T_\theta(dt),$$

so the set $\tilde N := \big\{t \in \tilde\Omega : \mathbb{E}\big(\|\widehat g\| \,\big|\, T = t\big) = \infty\big\}$ satisfies $\mathbb{P}^T_\theta(\tilde N) = 0$. In particular,

$$\check h(t) \;:=\; \mathbb{E}(\widehat g \,|\, T = t)$$

is well-defined in $\mathbb{R}^d$ for any $t \in \tilde\Omega \setminus \tilde N$.

As to part (a), $\check g = \check h(T)$ with $\check h(t)$ defined before. Since $\widehat g$ is unbiased,

$$g(\theta) \;=\; \mathbb{E}_\theta(\widehat g) \;=\; \int_{\tilde\Omega} \mathbb{E}(\widehat g \,|\, T = t)\; \mathbb{P}^T_\theta(dt) \;=\; \int_{\tilde\Omega} \check h(t)\; \mathbb{P}^T_\theta(dt) \;=\; \mathbb{E}_\theta(\check h(T)) \;=\; \mathbb{E}_\theta(\check g).$$

Hence, $\check g$ is unbiased, too. Moreover, applying Jensen's inequality to the conditional expectations $\mathbb{E}(\cdot \,|\, T = t)$ leads to

$$\begin{aligned}
\mathbb{E}_\theta\, L(\widehat g, \theta) \;&=\; \int_{\tilde\Omega} \mathbb{E}\big(L(\widehat g, \theta) \,\big|\, T = t\big)\; \mathbb{P}^T_\theta(dt) \\
&\ge\; \int_{\tilde\Omega} L\big(\mathbb{E}(\widehat g \,|\, T = t), \theta\big)\; \mathbb{P}^T_\theta(dt) \\
&=\; \int_{\tilde\Omega} L\big(\check h(t), \theta\big)\; \mathbb{P}^T_\theta(dt) \\
&=\; \mathbb{E}_\theta\, L(\check g, \theta).
\end{aligned}$$

If $L(\cdot,\theta)$ is strictly convex, the inequality

$$\mathbb{E}\big(L(\widehat{g},\theta)\,\big|\,T=t\big)\;\geq\;L\big(\mathbb{E}(\widehat{g}\,|\,T=t),\theta\big)$$

is strict, unless $P(\widehat{g}\neq\check{g}\,|\,T=t)=0$. Hence $\mathbb{E}_\theta\,L(\widehat{g},\theta)>\mathbb{E}_\theta\,L(\check{g},\theta)$, unless $\mathbb{P}_\theta(\widehat{g}\neq\check{g})=0$.

As to part (b), suppose that $\tilde{g}=\tilde{h}(T)$ is another unbiased estimator of $g(\theta)$. That means, the difference $\Delta:=\tilde{h}-\check{h}$ satisfies

$$\int_{\tilde{\Omega}}\Delta\,d\,\mathbb{P}_\theta^T\;=\;0\quad\text{for all }\theta\in\Theta.$$

But then completeness of $\mathcal{E}^T$ implies that $0=\mathbb{P}_\theta^T(\Delta\neq 0)=\mathbb{P}_\theta(\tilde{g}\neq\check{g})$ for arbitrary $\theta\in\Theta$. □

8.5 U-Statistics

The material in this section is based on the famous paper [8] by W. Hoeffding. Let $X_1,X_2,\ldots,X_n$ be independent random variables with unknown distribution P on a measurable space $(\mathcal{X},\mathcal{B})$. To establish a link to the previous sections, let $\mathcal{P}$ be a given family of probability distributions on $(\mathcal{X},\mathcal{B})$. Assuming that the unknown distribution P belongs to $\mathcal{P}$, the corresponding statistical experiment is

$$\big(\mathcal{X}^n,\mathcal{B}^{\otimes n},(P^{\otimes n})_{P\in\mathcal{P}}\big).$$

Note that all distributions $P^{\otimes n}$, $P\in\mathcal{P}$, are exchangeable (permutation-invariant). Hence for any given function $g:\mathcal{P}\to\mathbb{R}^q$, an unbiased point estimator $\widehat{g}$ of $g(P)$ can be improved by replacing $\widehat{g}(X_1,X_2,\ldots,X_n)$ with

$$\frac{1}{n!}\sum_{\pi\in\mathcal{S}_n}\widehat{g}(X_{\pi(1)},X_{\pi(2)},\ldots,X_{\pi(n)}).$$

And if $\mathcal{P}$ is sufficiently rich, the latter estimator is essentially unique, see Example 8.30 and Theorem 8.32.

U-Statistics. Now we consider a particular type of parameter $g(P)$. Let $h:(\mathcal{X}^m,\mathcal{B}^{\otimes m})\to\mathbb{R}$ be a given measurable function such that

$$g(P)\;:=\;\int h\,dP^{\otimes m}\;\in\;\mathbb{R}.$$

In case of $m\geq 2$, we assume without loss of generality that h is permutation-invariant, i.e. symmetric in its m arguments. Otherwise we could replace $h(x_1,\ldots,x_m)$ with

$$\tilde{h}(x_1,\ldots,x_m)\;:=\;\frac{1}{m!}\sum_{\pi\in\mathcal{S}_m}h(x_{\pi(1)},\ldots,x_{\pi(m)})$$

which defines a permutation-invariant function $\tilde{h}$ satisfying $\int\tilde{h}\,dP^{\otimes m}=\int h\,dP^{\otimes m}$.

Obviously, for $n\geq m$, a naive unbiased estimator for $g(P)$ is given by $\widehat{g}:=h(X_1,\ldots,X_m)$. Averaging this naive estimator over all permutations of $\boldsymbol{X}_n=(X_i)_{i=1}^n$ yields the following unbiased estimator of $g(P)$:

$$\check{g}_n\;:=\;\binom{n}{m}^{-1}\sum_{1\leq i_1<i_2<\cdots<i_m\leq n}h(X_{i_1},X_{i_2},\ldots,X_{i_m}).$$

Such an estimator is called a *U-statistic with kernel h* or a *U-statistic of order m*.

Example 8.33 (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:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \;=\; \binom{n}{1}^{-1} \sum_{1\le i\le n} h(X_i).$$

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:

$$\begin{aligned}
\binom{n}{2}^{-1} \sum_{1\le i_1<i_2\le n} (X_{i_1}-X_{i_2})^2/2 \;&=\; \frac{1}{n(n-1)} \sum_{i,j=1}^n (X_i-X_j)^2/2 \\
&=\; \frac{1}{n(n-1)}\Big(n\sum_{i=1}^n X_i^2 - \Big(\sum_{i=1}^n X_i\Big)^2\Big) \\
&=\; \frac{1}{n-1}\Big(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2\Big) \\
&=\; \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X}_n)^2.
\end{aligned}$$

Example 8.34. 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

$$g(P) \;:=\; \mathbb{P}\Big(X_3 > \sqrt{X_1^2+X_2^2}\Big).$$

(Maybe this is not such a brilliant proposal; the main point is to illustrate the construction of U-statistics.) This corresponds to $\mathbb{E}\, h(X_1,X_2,X_3)$ with $h(x_1,x_2,x_3) := 1_{\left[x_3>\sqrt{x_1^2+x_2^2}\right]}$. Symmetrizing this kernel leads to

$$h(x_1,x_2,x_3) \;=\; \frac{1}{3}\sum_{i=1}^3 1_{\left[x_i>\sqrt{\|x\|^2-x_i^2}\right]},$$

where $\|x\|^2 = x_1^2+x_2^2+x_3^2$, and the resulting unbiased estimator for $g(P)$ would be the U-statistic with this symmetric kernel h.

Here is a first result about the variance of a U-statistic which will suffice for our purposes:

Lemma 8.35 (Hoeffding). *Suppose that $\int h^2\, dP^{\otimes m} < \infty$. Let*

$$\begin{aligned}
h_m(x_1,\dots,x_m) \;&:=\; h(x_1,\dots,x_m) - g(P), \\
h_k(x_1,\dots,x_k) \;&:=\; \mathbb{E}\, h_m(x_1,\dots,x_k,X_{k+1},\dots,X_m) \quad \textit{for } 1\le k<m, \\
\sigma_k^2 \;&:=\; \mathbb{E}\big(h_k(X_1,\dots,X_k)^2\big) \quad \textit{for } 1\le k\le m.
\end{aligned}$$

Then $\sigma_1^2 \le \cdots \le \sigma_m^2$, *and*

$$\mathrm{Var}(\check{g}_n) \;=\; \sum_{k=1}^{m} \mathbb{P}(Y=k)\,\sigma_k^2$$

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 [8] findings. Stronger statements about the variances σ_k^2 and the distribution of $\check{g}_n$ will be derived in Exercise 8.43 from the so-called Hoeffding decomposition, introduced in Section A.4.

Proof of Lemma 8.35. For any index set $J \subset \{1,\dots,n\}$ with $1 \le \#J \le m$ let $X_J := (X_i)_{i\in J}$. Then we may write

$$\check{g}_n - g(P) \;=\; U_n \;:=\; \binom{n}{m}^{-1} \sum_{J\subset\{1,\dots,n\}:\#J=m} h_m(X_J).$$

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

$$\mathbb{E}\big(h_m(X_I)h_m(X_J)\big) \;=\; \sigma^2_{\#(I\cap J)} \quad \text{for } I,J \subset \{1,\dots,n\}, \#I = \#J = m, \tag{8.4}$$

where $\sigma_0^2 := 0$. Indeed, $\mathbb{E}\, h_m(X_1,\dots,X_m) = 0$ 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

$$\begin{aligned}
\mathbb{E}\big(h_m(X_I)h_m(X_J)\big) \;&=\; \mathbb{E}\big(h_m(X_{I\cap J}, X_{I\setminus J})h_m(X_{I\cap J}, X_{J\setminus I})\big) \\
&=\; \mathbb{E}\,\mathbb{E}\big(h_m(X_{I\cap J}, X_{I\setminus J})h_m(X_{I\cap J}, X_{J\setminus I}) \,\big|\, X_{I\cap J}\big) \\
&=\; \mathbb{E}\big(h_k(X_{I\cap J})^2\big) \\
&=\; \sigma_k^2.
\end{aligned}$$

Equation (8.4) yields the specific formula for $\mathrm{Var}(\check{g}_n)$, because

$$\mathrm{Var}(\check{g}_n) = \mathbb{E}(U_nU_n) \;=\; \binom{n}{m}^{-1} \sum_{I\subset\{1,\dots,n\}:\#I=m} \mathbb{E}\big(h_m(X_I)U_n\big),$$

and for any fixed index set $I \subset \{1, \dots, n\}$ with m elements,

$$\begin{aligned}
\mathbb{E}\big(h_m(X_I)U_n\big) &= \binom{n}{m}^{-1} \sum_{J\subset\{1,\dots,n\}:\#J=m} \mathbb{E}\big(h_m(X_I)h_m(X_J)\big) \\
&= \binom{n}{m}^{-1} \sum_{k=1}^{m} \#\big\{J \subset \{1,\dots,n\} : \#J = m, \#(J\cap I) = k\big\}\, \sigma_k^2 \\
&= \sum_{k=1}^{m} \binom{n}{m}^{-1} \binom{m}{k}\binom{n-m}{m-k}\sigma_k^2 \\
&= \sum_{k=1}^{m} \mathbb{P}(Y=k)\sigma_k^2.
\end{aligned}$$

Since this value does not depend on the particular choice of I, this proves the asserted equation for $\mathrm{Var}(\check{g}_n)$.

It remains to prove the inequality $\sigma_k^2 \le \sigma_{k+1}^2$ for $1 \le k < m$. Note first that

$$h_k(x_1,\dots,x_k) \;=\; \mathbb{E}\, h_{k+1}(x_1,\dots,x_k,X_{k+1}).$$

In case of $k = m-1$ this is just the definition of $h_k(x_1,\dots,x_k)$, and otherwise it is a consequence of Fubini's theorem, because

$$\begin{aligned}
h_k(x_1,\dots,x_k) &= \mathbb{E}\, h_m(x_1,\dots,x_k,X_{k+1},X_{k+2},\dots,X_m) \\
&= \mathbb{E}\,\mathbb{E}\big(h_m(x_1,\dots,x_k,X_{k+1},X_{k+2},\dots,X_m)\,\big|\,X_{k+1}\big) \\
&= \mathbb{E}\, h_{k+1}(x_1,\dots,x_k,X_{k+1}).
\end{aligned}$$

But then the Cauchy–Schwarz inequality implies that

$$\begin{aligned}
\sigma_k^2 &= \mathbb{E}\big(\mathbb{E}\big(h_{k+1}(X_1,\dots,X_k,X_{k+1})\,\big|\,X_1,\dots,X_k\big)^2\big) \\
&\le \mathbb{E}\big(\mathbb{E}\big(h_{k+1}(X_1,\dots,X_k,X_{k+1})^2\,\big|\,X_1,\dots,X_k\big)\big) \\
&= \mathbb{E}\big(h_{k+1}(X_1,\dots,X_k,X_{k+1})^2\big) \\
&= \sigma_{k+1}^2.
\end{aligned}$$

□

Corollary 8.36. *In case of* $\int h^2\, dP^{\otimes m} < \infty$,

$$\mathrm{Var}(\check{g}_n) \;=\; \frac{m^2\sigma_1^2}{n} + O(n^{-2}).$$

Proof of Corollary 8.36. This expansion follows immediately from Lemma 8.35 and the fact that

for $Y \sim \mathrm{Hyp}(n, m, m)$,

$$\begin{aligned}
\mathbb{P}(Y = k) &= \binom{n}{m}^{-1} \binom{m}{k} \binom{n-m}{m-k} \\
&= \frac{m! [m]_k \, [n-m]_{m-k}}{[n]_m \, k!(m-k)!} \\
&= \frac{[m]_k^2 \, n^{m-k}(1 + O(n^{-1}))}{k! \, n^m (1 + O(n^{-1}))} \\
&= \frac{[m]_k^2 (1 + O(n^{-1}))}{k! \, n^k} \\
&= \begin{cases} m^2/n + O(n^{-2}) & \text{if } k = 1, \\ O(n^{-2}) & \text{if } k \geq 2. \end{cases}
\end{aligned}$$

□

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 8.37 (Hoeffding). *Under the conditions of Lemma 8.35,*

$$\check{g}_n = g(P) + \frac{m}{n} \sum_{i=1}^{n} h_1(X_i) + R_n$$

where

$$\mathbb{E}(R_n^2) = O(n^{-2}).$$

Moreover,

$$\sqrt{n}(\check{g}_n - g(P)) \to_{\mathcal{L}} \mathcal{N}(0, m^2 \sigma_1^2)$$

as $n \to \infty$.

Our proof of Theorem 8.37 utilizes a general approximation result of Jaroslav Hájek:

Lemma 8.38 (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 8.38. 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 8.37. It follows from Lemma 8.38 that

$$\check{g}_n - g(P) \;=\; U_n \;=\; \sum_{i=1}^n \mathbb{E}(U_n \mid X_i) + R_n$$

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

$$\begin{aligned}
\mathbb{E}(U_n \mid X_i) &= \binom{n}{m}^{-1} \sum_{J \subset \{1,\ldots,m\}:\#J=m} \mathbb{E}\big(h_m(X_J) \mid X_i\big) \\
&= \binom{n}{m}^{-1} \sum_{J \subset \{1,\ldots,m\}:\#J=m} 1_{[i \in J]} h_1(X_i) \\
&= \binom{n}{m}^{-1} \#\big\{J \subset \{1,\ldots,m\} : \#J = m, i \in J\big\}\, h_1(X_i) \\
&= \binom{n}{m}^{-1} \binom{n-1}{m-1} h_1(X_i) \\
&= \frac{m}{n}\, h_1(X_i).
\end{aligned}$$

Hence

$$U_n \;=\; \frac{m}{n} \sum_{i=1}^n h_1(X_i) + R_n$$

with

$$\mathbb{E}(R_n^2) \;=\; \mathbb{E}(U_n^2) - \frac{m^2\sigma_1^2}{n} \;=\; \mathrm{Var}(\check{g}_n) - \frac{m^2\sigma_1^2}{n} \;=\; O(n^{-2})$$

according to Corollary 8.36. Consequently, $R_n = O_p(n^{-1})$, whence

$$\sqrt{n}(\check{g}_n - g(P)) \;=\; \frac{m}{\sqrt{n}} \sum_{i=1}^n h_1(X_i) + O_p(n^{-1/2}),$$

and it follows from the usual Central Limit Theorem that the right hand side converges in distribution to a Gaussian random variable with mean 0 and variance $m^2\,\mathbb{E}(h_1(X_1)^2) = m^2\sigma_1^2$. $\square$

Remark 8.39 (Hoeffding's decomposition in case of $m = 2$). In case of $m = 2$ one may write

$$\check{g}_n \;=\; g(P) + \frac{2}{n}\sum_{i=1}^{n} h_1(X_i) + \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_2^o(X_i, X_j)$$

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 8.33 (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 $n \to \infty$,

$$\sqrt{n}(S_X^2 - \sigma(P)^2) \;\to_{\mathcal{L}}\; \mathcal{N}\big(0, \mathbb{E}[(X_1 - \mu(P))^4] - \sigma(P)^4\big).$$

Example 8.40 (Kendall's τ). Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ 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 [9], 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 $X_1 X_2, Y_1, Y_2$ 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

$$\check{\tau}_n \;:=\; \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \operatorname{sign}(X_j - X_i)\operatorname{sign}(Y_j - Y_i),$$

which is a U-statistic of order 2 with kernel h.

Since $|h| \le 1$, we may apply Theorem 8.37 and conclude that

$$\check{\tau}_n \;=\; \tau(P) + \frac{2}{n}\sum_{i=1}^{n} h_1(X_i, Y_i) + R_n$$

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 $n \to \infty$,

$$\sqrt{n}(\check{\tau}_n - \tau(P)) \;\to_{\mathcal{L}}\; \mathcal{N}\big(0, 4\,\mathbb{E}(h_1(X_1, Y_1)^2)\big).$$

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

$$\sqrt{n}\,\check{\tau}_n \;\to_{\mathcal{L}}\; \mathcal{N}(0, 4/9).$$

Exercise 8.41. 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

$$g_m(P) \;:=\; \mathbb{E}_P \operatorname{Med}(X_1, \ldots, X_m)$$

with sample median function $\operatorname{Med}(\ldots)$.

Show that for $n \geq m$, the corresponding U-statistic

$$\breve{g} \;=\; \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} \operatorname{Med}(X_{i_1}, \ldots, X_{i_m})$$

is a L-statistic, that means,

$$\breve{g} \;=\; \sum_{i=1}^{n} w_i X_{(i)}$$

with suitable weights $w_1, w_2, \ldots, w_n \geq 0$.

Hint: Distinguish the cases of odd and even m.

Exercise 8.42. Suppose that $X_1, X_2, X_3, \ldots$ 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 $h : \mathbb{R}^3 \to \mathbb{R}$ such that $g(P) = \mathbb{E}\, h(X_1, X_2, X_3)$, and construct a corresponding U-statistic $\breve{g}$.

(b) A naive estimator for $g(P)$ is given by

$$\widehat{g}_{\text{naive}} \;:=\; \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^3.$$

Show that this estimator can be written as a function of the sums $S_\ell := \sum_{i=1}^n X_i^\ell$, $1 \leq \ell \leq 3$, and that the computation of $\widehat{g}_{\text{naive}}$ requires $O(n)$ steps.

(c) Show that the unbiased estimator $\breve{g}$ is also a simple function of S_1, S_2, S_3 and can be computed in $O(n)$ steps.

(d) Show that for some constant $C(P)$,

$$\mathbb{E}\,|\breve{g} - \widehat{g}_{\text{naive}}| \;\leq\; C(P) n^{-1}$$

for all $n \geq 3$. Proposal: One can write $\breve{g} - \widehat{g}_{\text{naive}}$ as a linear combination of the three sums S_3, $S_{21} := \sum_{i,j=1}^n 1_{[i \neq j]} X_i^2 X_j$ and $S_{111} := \sum_{i,j,k=1}^n 1_{[i,j,k \text{ different}]} X_i X_j X_k$ with coefficients depending on n.

Exercise 8.43 (Refinements via Hoeffding's decomposition). With the notation of Lemma 8.35 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 $1 \leq k \leq m$. The general Hoeffding decomposition, presented in Section A.4, implies that the random variables

$$h_k^o(X_J), \quad k \in \{1, \ldots, m\} \text{ and } J \subset \{1, \ldots, n\}, \#J = k,$$

are centered and uncorrelated, where

$$h_k(x_1, \ldots, x_k) \;=\; \sum_{\ell=1}^{k} \sum_{J \subset \{1,\ldots,k\}:\#J=\ell} h_\ell^o(x_J).$$

Let

$$\tau_k^2 \;:=\; \mathbb{E}\big(h_k^o(X_1, \ldots, X_k)^2\big).$$

(a) Show that $\sigma_k^2 = \mathbb{E}\big(h_k(X_1, \ldots, X_k)^2\big)$ equals

$$\sigma_k^2 \;=\; \sum_{\ell=1}^{k} \binom{k}{\ell} \tau_\ell^2.$$

Deduce from this representation that σ_k^2/k is non-decreasing in $k \in \{1, \ldots, m\}$.

(b) Show that $U_n := \check{g}_n - g(P)$ equals

$$U_n \;=\; \sum_{k=1}^{m} \binom{n}{k}^{-1} \sum_{J \subset \{1,\ldots,n\}:\#J=k} \binom{m}{k} h_k^o(X_J),$$

and that

$$\mathbb{E}(U_n^2) \;=\; \sum_{k=1}^{m} \frac{[m]_k}{[n]_k} \binom{m}{k} \tau_k^2.$$

(c) Deduce from (a) and (b) that

$$\frac{m^2\sigma_1^2}{n} \;\leq\; \mathbb{E}(U_n^2) \;\leq\; \frac{m^2\sigma_1^2}{n} + \frac{m^2\sigma_m^2}{n^2}.$$

Show that $n\,\mathbb{E}(U_n^2)$ is non-increasing in $n \geq m$ with limit $m^2\sigma_1^2$ as $n \to \infty$.

Chapter 9

Exponential Families

9.1 Definitions and Basic Properties

Definition 9.1 (Exponential families). A statistical experiment $\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ is called an *exponential family* if there exist a σ-finite measure M on $(\Omega, \mathcal{A})$, a measurable mapping $h : \Omega \to [0, \infty)$, a mapping $\alpha : \Theta \to \mathbb{R}^d$ and a measurable mapping $T : \Omega \to \mathbb{R}^d$ such that for any $\theta \in \Theta$,

$$\frac{d\,\mathbb{P}_\theta}{dM}(\omega) \;=\; h(\omega)\exp\big(\alpha(\theta)^\top T(\omega) - \kappa(\theta)\big)$$

with

$$\kappa(\theta) \;:=\; \log\int h\exp(\alpha(\theta)^\top T)\,dM.$$

(In particular, we assume that $\int h\exp(\alpha(\theta)^\top T)\,dM < \infty$ for all $\theta \in \Theta$.)

Definition 9.2 (Natural exponential families). For a given measure space $(\Omega, \mathcal{A}, M)$ and measurable mappings $h : \Omega \to [0, \infty)$, $T : \Omega \to \mathbb{R}^d$, the corresponding *natural exponential family* is given by $\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta_{\rm nat}}\big)$ with the *natural parameter space*

$$\Theta_{\rm nat} \;:=\; \Big\{\theta \in \mathbb{R}^d : \int h\exp(\theta^\top T)\,dM < \infty\Big\},$$

and the probability distributions $\mathbb{P}_\theta$ are given by

$$\begin{aligned}\frac{d\,\mathbb{P}_\theta}{dM}(\omega) \;&:=\; h(\omega)\exp\big(\theta^\top T(\omega) - \kappa(\theta)\big),\\ \kappa(\theta) \;&:=\; \log\int h\exp(\theta^\top T)\,dM.\end{aligned}$$

Example 9.3 (Gaussian samples). Let $\Omega = \mathbb{R}^n$, equipped with its Borel σ-field, let $\Theta = \mathbb{R}\times(0,\infty)$, and for $\theta = (\mu, \sigma) \in \Theta$, let

$$\mathbb{P}_{\mu,\sigma} \;:=\; \mathcal{N}(\mu, \sigma^2)^{\otimes n}.$$

With M denoting Lebesgue measure on $\mathbb{R}^n$, $h(\omega) := (2\pi)^{-n/2}$ and $X_i(\omega) := \omega_i$ for $\omega \in \Omega$, the

log-density of $\mathbb{P}_{\mu,\sigma}$ with respect to M is given by

$$\begin{aligned}\frac{d\,\mathbb{P}_{\mu,\sigma}}{dM} &= h(\omega)\exp\Big(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} - n\log\sigma\Big)\\ &= h(\omega)\exp\Big(\frac{\mu}{\sigma^2}\sum_{i=1}^n X_i + \frac{-1}{2\sigma^2}\sum_{i=1}^n X_i^2 - n\log\sigma\Big)\\ &= h(\omega)\exp\big(\alpha(\mu,\sigma)^\top T - \kappa(\mu,\sigma)\big)\end{aligned}$$

with

$$\begin{aligned}\alpha(\mu,\sigma) &:= \Big(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\Big)^\top,\\ T &:= \Big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\Big)^\top,\\ \kappa(\nu,\gamma) &:= n\log\sigma.\end{aligned}$$

Here one easily verifies that

$$\Theta_{\mathrm{nat}} = \mathbb{R}\times(-\infty,0).$$

Example 9.4 (Gamma distributions). Let $\Omega = (0,\infty)$, 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

$$\begin{aligned}f_{a,b}(\omega) &= \frac{1}{\Gamma(a)b}\Big(\frac{\omega}{b}\Big)^{a-1}\exp\Big(-\frac{\omega}{b}\Big)\\ &= \exp\big(\alpha(a,b)^\top T(\omega) - \kappa(a,b)\big)\end{aligned}$$

with

$$\begin{aligned}\alpha(a,b) &:= (a-1,-1/b)^\top,\\ T(\omega) &:= (\log\omega,\omega)^\top,\\ \kappa(a,b) &:= a\log b + \log\Gamma(a).\end{aligned}$$

Here one easily verifies that

$$\Theta_{\mathrm{nat}} = (-1,\infty)\times(-\infty,0).$$

Remark 9.5 (Convexity and smoothness). Let $\mathcal{E}$ be a natural exponential family as in Definition 9.2. It follows from convexity of the exponential function that the set Θ_{nat} is a convex subset of $\mathbb{R}^d$. Moreover, if $f\in\bigcap_{\theta\in\Theta_{\mathrm{nat}}} L^1(\mathbb{P}_\theta)$, then the function $h:\Theta\to\mathbb{R}$,

$$h(\theta) := \int f\,d\,\mathbb{P}_\theta,$$

is twice continuously differentiable on the interior of Θ_{nat} with gradient

$$\nabla h(\theta) = \mathrm{Cov}_\theta(f,T) = \int fT\,d\,\mathbb{P}_\theta - h(\theta)\int T\,d\,\mathbb{P}_\theta\,.$$

Remark 9.6 (Sufficiency). Suppose that $\mathcal{E}$ is an exponential family as in Definition 9.1, where (Ω,d) is a separable and complete metric space and $\mathcal{A}=\mathrm{Borel}(\Omega,d)$. Then T is a sufficient statistic for $\mathcal{E}$. This follows immediately from Neyman's factorization criterion.

Theorem 9.7 (Completeness in exponential families). *Let $\mathcal{E}$ be an exponential family as in Definition 9.1. Suppose that the set $\{\alpha(\theta) : \theta \in \Theta\} \subset \mathbb{R}^d$ contains an interior point. Then the statistical model $\big(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (\mathbb{P}_\theta^T)_{\theta\in\Theta}\big)$ is complete.*

Proof of Theorem 9.7. We may assume without loss of generality that $h \equiv 1$. Otherwise we could replace M with $\tilde{M}$, where $\tilde{M}(B) := \int_B h\, dM$. Note that the image measures[1] M^T and $\mathbb{P}_\theta^T$ on $\mathbb{R}^d$ satisfy

$$\frac{d\,\mathbb{P}_\theta^T}{dM^T}(x) \;=\; \exp\big(\alpha(\theta)^\top x - \kappa(\theta)\big).$$

Hence for a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, the property

$$\int_{\mathbb{R}^d} f\, d\,\mathbb{P}_\theta^T \;=\; 0 \quad \text{for all } \theta \in \Theta$$

is equivalent to

$$\int_{\mathbb{R}^d} f(x) \exp(\alpha(\theta)^\top x)\, M^T(dx) \;=\; 0 \quad \text{for all } \theta \in \Theta.$$

Since $\alpha(\Theta)$ contains an interior point, it follows from Theorem A.15 in Section A.3 that $f(x) = 0$ for M^T-almost all $x \in \mathbb{R}^d$. In particular, $\mathbb{P}_\theta^T(f \neq 0) = 0$ for all $\theta \in \Theta$. □

9.2 Nuisance Parameters

In this section we consider statistical experiments $\mathcal{E} = \big(\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ 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 $(\Omega, \mathcal{A})$ and measurable functions $S : (\Omega, \mathcal{A}) \to \mathbb{R}^d$, $Y : (\Omega, \mathcal{A}) \to \mathbb{R}$ such that for arbitrary $(\nu,\gamma) \in \Theta$,

$$\frac{d\,\mathbb{P}_{\nu,\gamma}}{dM} \;=\; \exp\big(\nu^\top S + \gamma Y - \kappa(\nu,\gamma)\big)$$

with

$$\kappa(\nu,\gamma) \;:=\; \log \int \exp\big(\nu^\top S + \gamma Y\big)\, dM \;<\; \infty.$$

Here the pair (S, Y) is a sufficient statistic, provided that (Ω, 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 $\mathbb{P}_{\nu,\gamma}$, 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.

[1] $M^T(B) := M(T \in B)$ and $\mathbb{P}_\theta^T(B) := \mathbb{P}_\theta(T \in B)$

Proposition 9.8. *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 $\mathbb{P}_{\nu_o,\gamma_o}$. That means,*

$$\mathbb{P}_{\nu_o,\gamma_o}(Y \in B \,|\, S = \cdot) \;=\; K(\cdot, B) \quad \text{for any Borel set } B \subset \mathbb{R}.$$

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: $M(S \in C) = 0$, and for any $\gamma \in \Gamma$,

$$\tilde{\kappa}(s, \gamma) \;:=\; \log \int_{\mathbb{R}} \exp(\gamma y)\, w(y)\, K(s, dy) \;<\; \infty \quad \text{for all } s \in \mathbb{R}^d \setminus C.$$

Moreover,

$$K_\gamma(s, B) \;:=\; \begin{cases} \displaystyle\int_B \exp\bigl(\gamma y - \tilde{\kappa}(s,\gamma)\bigr)\, w(y)\, K(s, dy) & \text{if } s \in \mathbb{R}^d \setminus C, \\ 1_{[0 \in B]} & \text{if } s \in C, \end{cases}$$

defines a stochastic kernel K_γ from $\mathbb{R}^d$ to $\mathbb{R}$ describing the conditional distribution of Y, given S, under any measure $\mathbb{P}_{\nu,\gamma}$, $\nu \in N$. That means,

$$\mathbb{P}_{\nu,\gamma}(Y \in B \,|\, S = \cdot) \;=\; K_\gamma(\cdot, B) \quad \text{for all } \nu \in N \text{ and any Borel set } B \subset \mathbb{R}.$$

Proof of Proposition 9.8. The assumption on K is equivalent to

$$\mathbb{E}_{\nu_o,\gamma_o}\bigl(g(S)h(Y)\bigr) \;=\; \mathbb{E}_{\nu_o,\gamma_o}\Bigl(g(S) \int h(y)\, K(S, dy)\Bigr)$$

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

$$\begin{aligned} &\mathbb{E}_{\nu,\gamma}\bigl(g(S)h(Y)\bigr) \\ &= \mathbb{E}_{\nu_o,\gamma_o}\Bigl(g(S)h(Y)\frac{\exp(\nu^\top S + \gamma Y - \kappa(\nu,\gamma))}{\exp(\nu_o^\top S + \gamma_o Y - \kappa(\nu_o,\gamma_o))}\Bigr) \qquad (9.1) \\ &= \mathbb{E}_{\nu_o,\gamma_o}\Bigl(g(S)\exp\bigl((\nu-\nu_o)^\top S - \kappa(\nu,\gamma)\bigr) \int h(y)\exp(\gamma y)\, w(y)\, K(S, dy)\Bigr) \qquad (9.2) \end{aligned}$$

with

$$w(y) \;:=\; \exp\bigl(\kappa(\nu_o,\gamma_o) - \gamma_o y\bigr).$$

Taking $h \equiv 1$ shows that

$$\mathbb{E}_{\nu,\gamma}(g(S)) \;=\; \mathbb{E}_{\nu_o,\gamma_o}\bigl(g(S) f_{\nu,\gamma}(S)\bigr)$$

with

$$f_{\nu,\gamma}(s) \;:=\; \exp\bigl((\nu-\nu_o)^\top s - \kappa(\nu,\gamma)\bigr) \int \exp(\gamma y)\, w(y)\, K(s, dy).$$

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 $\mathbb{P}_{\nu,\gamma}(S \in C(\gamma)) = 0$ for all $\nu \in N$. Since $\mathbb{P}_{\nu,\gamma}$ has a strictly positive density with respect to M, we may even conclude that

$$M(S \in C(\gamma)) \;=\; 0.$$

But then the set $C := \bigcup_{\gamma\in\Gamma\cap\mathbb{Q}} C(\gamma)$ satisfies

$$M(S \in C) \;=\; 0,$$

and by convexity of the exponential function,

$$\int \exp(\gamma y)\, w(y)\, K(s, dy) \;<\; \infty \quad \text{for all } s \in \mathbb{R}^d \setminus C \text{ and } \gamma \in \Gamma.$$

Hence the stochastic kernel K_γ described in the proposition is well-defined, and it follows from (9.2) that for arbitrary measurable functions $g : \mathbb{R}^d \to [0,\infty]$ and $h : \mathbb{R} \to [0,\infty]$,

$$\begin{aligned}\mathbb{E}_{\nu,\gamma}\big(g(S)h(Y)\big) \;&=\; \mathbb{E}_{\nu_o,\gamma_o}\Big(g(S)\int h(y)\, K_\gamma(S, dy)\, f_{\nu,\gamma}(S)\Big)\\ &=\; \mathbb{E}_{\nu,\gamma}\Big(g(S)\int h(y)\, K_\gamma(S, dy)\Big).\end{aligned}$$

This shows that the kernel K_γ describes the conditional distribution of Y, given S, under the measure $\mathbb{P}_{\nu,\gamma}$. □

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 $\phi_\alpha : \mathbb{R}^d \times \mathbb{R} \to [0,1]$ with

$$\phi_\alpha(s,y) \;:=\; \begin{cases} 0 & \text{if } y < k_\alpha(s)\\ \gamma_\alpha(s) & \text{if } y = k_\alpha(s)\\ 1 & \text{if } y > k_\alpha(s)\end{cases}$$

satisfies

$$\mathbb{E}_{\nu_o,\gamma_o}\big(\phi_\alpha(S,Y)\,\big|\,S\big) \;=\; \alpha \quad \text{almost surely.}$$

Then ϕ_α is a level-α test of Θ_o, and has a certain optimality property:

Theorem 9.9 (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 $\phi : \mathbb{R}^d \times \mathbb{R} \to [0,1]$ such that*

$$\mathbb{E}_{\nu,\gamma}\,\phi(S,Y) \;\begin{cases} \le \alpha & \text{if } \gamma \le \gamma_o,\\ \ge \alpha & \text{if } \gamma > \gamma_o.\end{cases}$$

For arbitrary $\phi \in \Phi_\alpha$ and $(\nu,\gamma) \in \Theta \setminus \Theta_o$,

$$\mathbb{E}_{\nu,\gamma}\,\phi_\alpha(S,Y) \;\ge\; \mathbb{E}_{\nu,\gamma}\,\phi(S,Y).$$

Proof of Theorem 9.9. With the stochastic kernels K_γ, $\gamma \in \Gamma$, and the Borel set $C \subset \mathbb{R}^d$ as in Proposition 9.8, we may alter the set C if necessary such that

$$\int_{\mathbb{R}} \phi_\alpha(s,y)\, K_{\gamma_o}(s,dy) \;=\; \alpha \quad \text{for all } s \in \mathbb{R}^d \setminus C.$$

For any test $\phi : \mathbb{R}^d \times \mathbb{R} \to [0,1]$, its power

$$\mathbb{E}_{\nu,\gamma}\, \phi(S,Y) \;=\; \int_\Omega \phi(S,Y) \exp\bigl(\nu^\top S + \gamma Y - \kappa(\nu,\gamma)\bigr)\, dM$$

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 $\phi \in \Phi_\alpha$, this implies that

$$\mathbb{E}_{\nu,\gamma_o}\, \phi(S,Y) \;=\; \alpha \quad \text{for arbitrary } \nu \in N.$$

We may also write

$$\mathbb{E}_{\nu,\gamma}\, \phi(S,Y) \;=\; \mathbb{E}_{\nu,\gamma}\Bigl(\int_{\mathbb{R}} \phi(S,y)\, K_\gamma(S,dy)\Bigr).$$

But the restricted statistical experiment $\mathcal{E}_o := \bigl(\Omega, \mathcal{A}, (\mathbb{P}_{\nu,\gamma_o})_{\nu\in N}\bigr)$ is an exponential family with natural parametrization, because

$$\frac{d\,\mathbb{P}_{\nu,\gamma_o}}{dM} \;=\; \exp\bigl(\nu^\top S - \kappa(\nu,\gamma_o)\bigr) \exp(\gamma_o Y),$$

i.e. with the modified measure

$$M_o(d\omega) \;:=\; \exp(\gamma_o Y(\omega))\, M(d\omega)$$

we may write

$$\frac{d\,\mathbb{P}_{\nu,\gamma_o}}{dM_o} \;=\; \exp\bigl(\nu^\top S - \kappa(\nu,\gamma_o)\bigr).$$

Since N is open, the corresponding family $\bigl(\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (\mathbb{P}^S_{\nu,\gamma_o})_{\nu\in N}\bigr)$ is complete, that means, it follows from

$$\alpha \;=\; \mathbb{E}_{\nu,\gamma_o}\Bigl(\int_{\mathbb{R}} \phi(S,y)\, K_{\gamma_o}(S,dy)\Bigr) \quad \text{for all } \nu \in N$$

that

$$\int_{\mathbb{R}} \phi(S,y)\, K_{\gamma_o}(S,dy) \;=\; \alpha \quad \text{almost surely}$$

under any measure $\mathbb{P}_{\nu,\gamma_o}$, $\nu \in N$. But our special construction of ϕ_α and Theorem 7.13 imply that for $\gamma > \gamma_o$ and $s \in \mathbb{R}^d \setminus C$,

$$\int_{\mathbb{R}} \phi(s,y)\, K_\gamma(s,dy) \;\le\; \int_{\mathbb{R}} \phi_\alpha(s,y)\, K_\gamma(s,dy) \quad \text{whenever } \int_{\mathbb{R}} \phi(s,y)\, K_{\gamma_o}(s,dy) \le \alpha.$$

Thus for arbitrary $(\nu,\gamma) \in \Theta$ with $\gamma > \gamma_o$,

$$\begin{aligned}\mathbb{E}_{\nu,\gamma}\, \phi(S,Y) \;&=\; \mathbb{E}_{\nu,\gamma}\Bigl(\int_{\mathbb{R}} \phi(S,y)\, K_\gamma(S,dy)\Bigr) \\ &\le\; \mathbb{E}_{\nu,\gamma}\Bigl(\int_{\mathbb{R}} \phi_\alpha(S,y)\, K_\gamma(S,dy)\Bigr) \;=\; \mathbb{E}_{\nu,\gamma}\, \phi_\alpha(S,Y).\end{aligned}$$

□

Example 9.10 (Fisher's exact test and odds ratios). Suppose that $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent, identically distributed random variables with values in $\{0,1\} \times \{0,1\}$ such that all four probabilities

$$p_{xy} \;:=\; \mathbb{P}(X = x, Y = y), \quad x, y \in \{0,1\},$$

are strictly positive; here (X, Y) denotes any of the n pairs (X_i, Y_i). With the parameter $\boldsymbol{p} := (p_{00}, p_{01}, p_{10}, p_{11})$, the "correlation" between X and Y may be quantified in terms of the odds ratio

$$\rho = \rho(\boldsymbol{p}) \;:=\; \frac{p_{11}p_{00}}{p_{01}p_{10}} \;=\; \frac{\operatorname{odds}(X = 1 \,|\, Y = 1)}{\operatorname{odds}(X = 1 \,|\, Y = 0)} \;=\; \frac{\operatorname{odds}(Y = 1 \,|\, X = 1)}{\operatorname{odds}(Y = 1 \,|\, X = 0)}.$$

Suppose we are only interested in ρ. To design good tests or confidence regions for ρ let us rewrite the model as a suitable natural exponential family:

Viewing the X_i and Y_i as random variables on $\Omega := (\{0,1\} \times \{0,1\})^n$, equipped with counting measure M, we obtain a statistical experiment with distributions $\mathbb{P}_{\boldsymbol{p}}$ on Ω such that

$$\begin{aligned} \log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \;&=\; \log\bigl(p_{00}^{H_{00}} p_{01}^{H_{01}} p_{10}^{H_{10}} p_{11}^{H_{11}}\bigr) \\ &=\; H_{00} \log p_{00} + H_{01} \log p_{01} + H_{10} \log p_{10} + H_{11} \log p_{11} \end{aligned}$$

with the absolute frequencies

$$H_{xy} \;:=\; \#\{i \le n : X_i = x, Y_i = y\}.$$

With the marginal frequencies

$$\begin{aligned} H_{+1} \;&:=\; H_{01} + H_{11} \;=\; \#\{i \le n : Y_i = 1\}, \\ H_{1+} \;&:=\; H_{10} + H_{11} \;=\; \#\{i \le n : X_i = 1\}, \end{aligned}$$

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

$$\log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \;=\; H_{+1} \log \frac{p_{01}}{p_{00}} + H_{1+} \log \frac{p_{10}}{p_{00}} + H_{11} \log \frac{p_{11}p_{00}}{p_{01}p_{10}} + n \log p_{00}.$$

With

$$\begin{aligned} \nu = \nu(\boldsymbol{p}) \;&:=\; \Bigl(\log \frac{p_{01}}{p_{00}}, \log \frac{p_{10}}{p_{00}}\Bigr)^\top \;\in\; \mathbb{R}^2, \\ \gamma = \gamma(\boldsymbol{p}) \;&:=\; \log \frac{p_{11}p_{00}}{p_{01}p_{10}} \;=\; \log \rho \;\in\; \mathbb{R}, \end{aligned}$$

we may write

$$\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \;=\; (1 + e^{\nu_1} + e^{\nu_2} + e^{\nu_1+\nu_2+\gamma})^{-1} \begin{bmatrix} 1 & e^{\nu_1} \\ e^{\nu_2} & e^{\nu_1+\nu_2+\gamma} \end{bmatrix}.$$

In particular, for any choice of $\nu \in \mathbb{R}^2$ and $\gamma \in \mathbb{R}$ there exists a probability vector $\boldsymbol{p}$ such that $\nu = \nu(\boldsymbol{p})$ and $\gamma = \gamma(\boldsymbol{p})$. Moreover,

$$\log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \;=\; \nu^\top S + \gamma Y - \kappa(\nu, \gamma)$$

with

$$\begin{aligned} S &:= (H_{+1}, H_{1+})^\top, \\ Y &:= H_{11}, \\ \kappa(\nu,\gamma) &:= -n \log p_{00} \;=\; n \log(1 + e^{\nu_1} + e^{\nu_2} + e^{\nu_1+\nu_2+\gamma}). \end{aligned}$$

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 $s, z \geq 0$, 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,

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{p}}(H_{11} = k, H_{+1} = s, H_{1+} = z) \\ &\quad= \mathbb{P}_{\boldsymbol{p}}(H_{11} = k, H_{01} = s-k, H_{10} = z-k, H_{00} = n-s-z+k) \\ &\quad= \frac{n!}{k!(s-k)!(z-k)!(n-z-s+k)!}\, p_{11}^k p_{01}^{s-k} p_{10}^{z-k} p_{00}^{n-z-s+k} \\ &\quad= n!\, p_{01}^s p_{10}^z p_{00}^{n-z-s}\, \frac{\rho^k}{k!(s-k)!(z-k)!(n-z-s+k)!}. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}_{\boldsymbol{p}}(H_{11} = k \,|\, H_{+1} = s, H_{1+} = z) \;&=\; \frac{\mathbb{P}_{\boldsymbol{p}}(H_{11} = k, H_{+1} = s, H_{1+} = z)}{\sum_\ell \mathbb{P}_{\boldsymbol{p}}(H_{11} = \ell, H_{+1} = s, H_{1+} = z)} \\ &=\; C(n,s,z,\rho)^{-1}\, \frac{\rho^k}{k!(s-k)!(z-k)!(n-z-s+k)!} \end{aligned}$$

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

$$\mathbb{P}_{\boldsymbol{p}}(H_{11} = k \,|\, H_{+1} = s, H_{1+} = z) \;=\; \tilde{C}(n,s,z,\rho)^{-1}\, \mathrm{Hyp}_{n,s,z}(\{k\})\, \rho^k$$

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 9.11 (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

a finite set $\mathcal{X} := \{x_1, x_2, \ldots, x_K\}$ with $K \geq 3$ different elements. The parameter vector $\boldsymbol{p} = (p_k)_{k=1}^K$ with $p_k := \mathbb{P}(X = x_k) > 0$ is unknown. Suppose we are mainly interested in the ratio

$$\rho = \rho(\boldsymbol{p}) \ := \ \frac{p_1}{p_2}.$$

To this end we consider the X_i as random variables on the finite set $\Omega := \mathcal{X}^n$, equipped with counting measure M. This leads to a statistical experiment with distributions $\mathbb{P}_{\boldsymbol{p}}$ on Ω given by

$$\log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \ = \ \log \prod_{j=1}^K p_j^{H_j} \ = \ \sum_{j=1}^K H_j \log(p_j)$$

with the absolute frequencies

$$H_j \ := \ \#\{i \leq n : X_i = x_j\}.$$

Since $H_K = n - \sum_{j<K} H_j$, we may rewrite this as

$$\begin{aligned} \log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \ &= \ \sum_{j=1}^{K-1} H_j \log(p_j/p_K) + n \log(p_K) \\ &= \ H_1 \log \frac{p_1}{p_2} + (H_1 + H_2) \log \frac{p_2}{p_K} + \sum_{2<j<K} H_j \log \frac{p_j}{p_K} + n \log(p_K). \end{aligned}$$

With

$$\begin{aligned} \nu = \nu(\boldsymbol{p}) \ &:= \ \Big(\log \frac{p_{\ell+1}}{p_K}\Big)_{\ell=1}^{K-2} \ \in \ \mathbb{R}^{K-2}, \\ \gamma = \gamma(\boldsymbol{p}) \ &:= \ \log \frac{p_1}{p_2} \ = \ \log \rho \end{aligned}$$

we may write

$$\boldsymbol{p} \ = \ \frac{(e^{\gamma+\nu_1}, e^{\nu_1}, \ldots, e^{\nu_{K-2}}, 1)^\top}{e^{\gamma+\nu_1} + e^{\nu_1} + \cdots + e^{\nu_{K-2}} + 1}.$$

In particular, for any choice of $\nu \in \mathbb{R}^{K-2}$ and $\gamma \in \mathbb{R}$ there exists a probability vector $\boldsymbol{p}$ such that $\nu = \nu(\boldsymbol{p})$ and $\gamma = \gamma(\boldsymbol{p})$. Moreover,

$$\log \frac{d\,\mathbb{P}_{\boldsymbol{p}}}{dM} \ = \ \nu^\top S + \gamma Y - \kappa(\nu, \gamma)$$

with

$$\begin{aligned} S_1 \ &:= \ H_1 + H_2, \\ S_\ell \ &:= \ H_{\ell+1} \quad \text{for } 2 \leq \ell \leq K-2, \\ Y \ &:= \ H_1, \\ \kappa(\nu, \gamma) \ &:= \ -n \log(p_K) \ = \ n \log(e^{\gamma+\nu_1} + e^{\nu_1} + \cdots + e^{\nu_{K-2}} + 1). \end{aligned}$$

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

$m, s_3, \ldots, s_K \geq 0$ with $m + \sum_{j=3}^K s_j = n$ and $k \in \{0, \ldots, m\}$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{p}}(H_1 = k, H_1 + H_2 = m, H_j = s_j \text{ for } j \geq 3) \\
&\quad = \mathbb{P}_{\boldsymbol{p}}(H_1 = k, H_2 = m - k, H_j = s_j \text{ for } j \geq 3) \\
&\quad = \frac{n!}{k!(m-k)! \prod_{j\geq 3} s_j!} p_1^k p_2^{m-k} \prod_{j\geq 3} p_j^{s_j} \\
&\quad = \binom{m}{k} \pi^k (1-\pi)^{m-k} \frac{n!}{m! \prod_{j\geq 3} s_j!} (p_1 + p_2)^m \prod_{j\geq 3} p_j^{s_j}
\end{aligned}$$

with

$$\pi = \pi(\boldsymbol{p}) \;:=\; \frac{p_1}{p_1 + p_2} \;=\; \frac{\rho}{1+\rho} \;\in\; (0, 1).$$

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

$$\rho \;=\; \frac{\pi}{1-\pi}.$$

Example 9.12 (Comparing two Poisson parameters). Suppose we observe independent random variables $X \sim \mathrm{Poiss}(\lambda)$ and $Y \sim \mathrm{Poiss}(\mu)$ with unknown parameters $\lambda, \mu > 0$. Suppose further that we are mainly interested in the ratio $\rho := \lambda/\mu$. With $\Omega := \mathbb{N}_0 \times \mathbb{N}_0$, $X(\omega) := \omega_1$, $Y(\omega) := \omega_2$ and counting measure M on Ω, this situation corresponds to a statistical model with distributions $\mathbb{P}_{\lambda,\mu}$ given by

$$\frac{d\,\mathbb{P}_{\lambda,\mu}}{dM} \;=\; e^{-\lambda}\frac{\lambda^X}{X!} e^{-\mu}\frac{\mu^Y}{Y!} \;=\; \frac{\exp(X\log\lambda + Y\log\mu - \lambda - \mu)}{X!Y!}.$$

Replacing M with the measure M_o given by $M_o(\{\omega\}) := (\omega_1!\omega_2!)^{-1}$ we may write

$$\begin{aligned}
\log\frac{d\,\mathbb{P}_{\lambda,\mu}}{dM_o} &= X\log\lambda + Y\log\mu - (\lambda+\mu) \\
&= X\log(\lambda/\mu) + (X+Y)\log\mu - (\lambda+\mu) \\
&= \nu(X+Y) + \gamma X - \kappa(\nu,\gamma)
\end{aligned}$$

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 X, given $X + Y$. But for arbitrary integers $m \geq 0$ and $k \in \{0, \ldots, m\}$,

$$\begin{aligned}
\mathbb{P}_{\lambda,\mu}(X = k, X + Y = m) &= \mathbb{P}_{\lambda,\mu}(X = k)\,\mathbb{P}_{\lambda,\mu}(Y = m - k) \\
&= e^{-(\lambda+\mu)}\frac{\lambda^k \mu^{m-k}}{k!(m-k)!} \\
&= \binom{m}{k}\pi^k(1-\pi)^{m-k}\, e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^m}{m!} \\
&= \mathrm{Bin}_{m,\pi}(\{k\})\mathrm{Poiss}_{\lambda+\mu}(\{m\})
\end{aligned}$$

with

$$\pi \;:=\; \frac{\lambda}{\lambda+\mu} \;=\; \frac{\rho}{1+\rho} \;\in\; (0,1).$$

This shows that the conditional distribution of X, given $X+Y$, is equal to

$$\mathrm{Bin}(X+Y,\pi) \;=\; \mathrm{Bin}\Big(X+Y,\frac{\rho}{1+\rho}\Big).$$

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 9.13 (Basú). *Let $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical experiment, and let $S : (\Omega,\mathcal{A}) \to (\Omega',\mathcal{A}')$ be a sufficient statistic for $\mathcal{E}$ such that the experiment $\mathcal{E}^S = (\Omega',\mathcal{A}',(\mathbb{P}_\theta^S)_{\theta\in\Theta})$ is boundedly complete. If $V : (\Omega,\mathcal{A}) \to (\Omega'',\mathcal{A}'')$ is a measurable mapping such that its distributions $\mathbb{P}_\theta^V$, $\theta\in\Theta$, are identical, then S and V are stochastically independent under each measure $\mathbb{P}_\theta$, $\theta\in\Theta$.*

Proof of Lemma 9.13. By sufficency of S there exists a stochastic kernel K from $(\Omega,\mathcal{A})$ to $(\Omega',\mathcal{A}')$ describing the conditional distribution of $X\sim\mathbb{P}_\theta$, given S, simultaneously for all $\theta\in\Theta$. That means, for arbitrary $A'\in\mathcal{A}'$, $A''\in\mathcal{A}''$ and $\theta\in\Theta$,

$$\begin{aligned}\mathbb{P}_\theta(S\in A', V\in A'') \;&=\; \mathbb{E}_\theta\big(\mathbb{E}_\theta(1_{A'}(S)1_{A''}(V)\,|\,S)\big)\\ &=\; \mathbb{E}_\theta\big(1_{A'}(S)\,\mathbb{P}_\theta(V\in A''\,|\,S)\big)\\ &=\; \mathbb{E}_\theta\big(1_{A'}(S)K(S,\{V\in A''\})\big).\end{aligned}$$

Setting $A'=\Omega'$, we obtain the equation

$$\mathbb{P}_\theta(V\in A'') \;=\; \mathbb{E}_\theta\, K(S,\{V\in A''\}).$$

By our assumption on V, the left hand side does not depend on $\theta\in\Theta$, and we denote this number with $P(V\in A'')$. Hence $f(s) := K(s,\{V\in A''\}) - P(V\in A'')$ defines a bounded measurable function on $(\Omega',\mathcal{A}'')$ such that $\mathbb{E}_\theta\, f(S) = 0$ for all $\theta\in\Theta$. By completeness of $\mathcal{E}^S$,

$$\mathbb{P}_\theta(f(S)\neq 0) \;=\; \mathbb{P}_\theta\big(K(S,\{V\in A''\})\neq P(V\in A'')\big) \;=\; 0 \quad \text{for all } \theta\in\Theta.$$

Hence for arbitrary $A'\in\mathcal{A}'$, $A''\in\mathcal{A}''$ and $\theta\in\Theta$,

$$\begin{aligned}\mathbb{P}_\theta(S\in A', V\in A'') \;&=\; \mathbb{E}_\theta\big(1_{A'}(S)K(S,\{V\in A''\})\big)\\ &=\; \mathbb{E}_\theta\big(1_{A'}(S)P(V\in A'')\big)\\ &=\; \mathbb{P}_\theta(S\in A')P(V\in A''),\end{aligned}$$

which proves stochastic independence of S and V under $\mathbb{P}_\theta$. □

Application to exponential families. Let $\mathcal{E} = (\Omega, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ 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 "$\gamma \le \gamma_o$" or "$\gamma \ge \gamma_o$" or "$\gamma = \gamma_o$" for any given value $\gamma_o \in \Gamma$.

Note that for the restricted experiment $\mathcal{E}_o = (\Omega, \mathcal{A}, (\mathbb{P}_{\nu,\gamma_o})_{\nu\in N})$ the statistic S is sufficient, and the family $\mathcal{E}_o^S = (\mathbb{R}^d, \mathrm{Borel}(\mathbb{R}^d), (\mathbb{P}^S_{\nu,\gamma_o})_{\nu\in N})$ is complete. Suppose we can identify a real-valued statistic

$$V \;=\; f(S, Y)$$

with the following two properties:

- $V = f(S, Y)$ is strictly increasing in Y almost surely,
- the distribution of V under $\mathbb{P}_{\nu,\gamma_o}$ does not depend on $\nu \in N$.

Then optimal unbiased level-α tests of the null hypotheses above may be constructed in terms of V and its unconditional distribution under $\mathbb{P}_{\nu,\gamma_o}$, where $\nu \in N$ is arbitrary.

For instance if V has continuous distribution function F_{γ_o} in case of $\gamma = \gamma_o$, then the right-sided p-value

$$1 - F_{\gamma_o}(V)$$

yields an optimal unbiased level-α test of "$\gamma \le \gamma_o$", whereas the left-sided p-value

$$F_{\gamma_o}(V)$$

is optimal for the null hypothesis "$\gamma \ge \gamma_o$".

Example 9.14 (Student's t-test for a Gaussian mean). As in Example 9.3 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 "$\mu \le \mu_o$" or "$\mu \ge \mu_o$" 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 $(2\pi)^{-n/2}$ and $X_i(\omega) := \omega_i$, we may write

$$\begin{aligned}
\log \frac{d\,\mathbb{P}_{\mu,\sigma}}{dM} \;&=\; -\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} - n\log\sigma \\
&=\; -\sum_{i=1}^n \frac{(X_i-\mu_o)^2}{2\sigma^2} + \frac{n(\mu-\mu_o)(\bar X-\mu_o)}{\sigma^2} - n\log\sigma \\
&=\; \nu S + \gamma Y - \kappa(\nu,\gamma),
\end{aligned}$$

where

$$\begin{aligned} \nu &:= \frac{-1}{2\sigma^2} \in (-\infty, 0), \\ \gamma = \gamma(\mu_o) &:= \frac{\sqrt{n}(\mu - \mu_o)}{\sigma^2} \in \mathbb{R}, \\ \kappa(\nu, \gamma) &:= n \log \sigma, \\ S = S(\mu_o) &:= \sum_{i=1}^{n} (X_i - \mu_o)^2, \\ Y = Y(\mu_o) &:= \sqrt{n}(\bar{X} - \mu_o). \end{aligned}$$

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 $\mathbb{P}_{\mu_o,\sigma}$ for some $\sigma > 0$ directly, we apply Basú's lemma and recall student's method from introductory statistics courses: With the sample standard deviation

$$S_X := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

it is well-known that

$$V = V(\mu_o) := \frac{\sqrt{n}(\bar{X} - \mu_o)}{S_X} \sim t_{n-1}$$

whenever $\mu = \mu_o$, irrespective of $\sigma > 0$. But

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i - \mu_o)^2 - n(\bar{X} - \mu_o)^2 = S - Y^2,$$

so

$$V = \frac{\sqrt{n-1}\, Y}{\sqrt{S - Y^2}},$$

which is monotone increasing in Y almost surely. (Note that $S > 0$ and $Y^2 < S$ almost surely.) Hence with F_{n-1} denoting the distribution function of t_{n-1}, the left-sided p-value

$$F_{n-1}(V(\mu_o))$$

yields optimal unbiased tests of "$\mu \geq \mu_o$", the right-sided p-value

$$1 - F_{n-1}(V(\mu_o))$$

yields optimal unbiased tests of "$\mu \leq \mu_o$", and the traditional $(1-\alpha)$-confidence interval

$$\left[\bar{X} \pm \frac{S_X}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\right]$$

for μ is based on optimal unbiased level-α tests of one-point hypotheses "$\mu = \mu_o$", $\mu_o \in \mathbb{R}$.

Example 9.15 (Comparing two Gamma scale parameters). Suppose we observe independent random variables

$$Y_1 \sim \text{Gamma}(a_1, \beta_1) \quad \text{and} \quad Y_2 \sim \text{Gamma}(a_2, \beta_2)$$

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 $\Omega := (0,\infty)^2$, Lebesgue measure M on Ω and $Y_i(\omega) := \omega_i$, this corresponds to the statistical model with distributions $\mathbb{P}_{\boldsymbol{\beta}}$ given by

$$\frac{d\,\mathbb{P}_{\boldsymbol{\beta}}}{dM} = \frac{(Y_1/\beta_1)^{a_1-1}\exp(-Y_1/\beta_1)}{\Gamma(a_1)\beta_1}\,\frac{(Y_2/\beta_2)^{a_2-1}\exp(-Y_2/\beta_2)}{\Gamma(a_2)\beta_2}.$$

With the modified measure M_o given by

$$\frac{dM_o}{dM} := \frac{Y_1^{a_1-1}Y_2^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)}$$

we may write

$$\log\frac{d\,\mathbb{P}_{\boldsymbol{\beta}}}{dM_o} = \frac{-1}{\beta_1}\,Y_1 + \frac{-1}{\beta_2}\,Y_2 - a_1\log\beta_1 - a_2\log\beta_2.$$

For a hypothetical value ρ_o of ρ we may rewrite the log-likelihood as

$$\begin{aligned}\log\frac{d\,\mathbb{P}_{\boldsymbol{\beta}}}{dM_o} &= \frac{\rho-\rho_o}{\beta_2\rho}\,Y_1/\rho_o + \frac{-1}{\beta_2}\,(Y_1/\rho_o + Y_2) - a_1\log\beta_1 - a_2\log\beta_2\\ &= \nu S + \gamma Y - \kappa(\nu,\gamma)\end{aligned}$$

with

$$\begin{aligned}\nu = \nu(\boldsymbol{\beta}) &:= \frac{-1}{\beta_2} \in (-\infty, 0),\\ \gamma = \gamma(\boldsymbol{\beta},\rho_o) &:= \frac{\rho-\rho_o}{\beta_2\rho} \in \mathbb{R},\\ S = S(\rho_o) &:= Y_1/\rho_o + Y_2,\\ Y = Y(\rho_o) &:= Y_1/\rho_o\end{aligned}$$

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 $\mathbb{P}_{\beta_2\rho_o,\beta_2}$.

In case of $\rho = \rho_o$, the distribution of $(Y_1/\rho_o, Y_2)$ coincides with the distribution of $\beta_2(Z_1, Z_2)$ with independent random variables

$$Z_1 \sim \text{Gamma}(a_1, 1) \quad \text{and} \quad Z_2 \sim \text{Gamma}(a_2, 1).$$

Then the test statistic

$$V(\rho_o) \;:=\; \frac{Y(\rho_o)}{S(\rho_o)} \;=\; \frac{Y_1/\rho_o}{Y_1/\rho_o + Y_2}$$

has the same distribution as

$$\frac{Z_1}{Z_1 + Z_2} \;\sim\; \mathrm{Beta}(a_1, a_2),$$

irrespective of $\beta_2 = -1/\nu$. Thus by Basú's lemmaor a direct argument, $V(\rho_o)$ and $S(\rho_o)$ are stochastically independent in case of $\rho = \rho_o$, and optimal tests are obtained by comparing $V(\rho_o)$ with $\mathrm{Beta}(a_1, a_2)$.

Specifically, if we denote the u-quantile of $\mathrm{Beta}(a_1, a_2)$ with $\mathrm{Beta}_{a_1,a_2;u}$, then an optimal unbiased level-α test of "$\rho \geq \rho_o$" rejects this null hypothesis if

$$V(\rho_o) \;\leq\; \mathrm{Beta}_{a_1,a_2;\alpha}.$$

This leads to the $(1-\alpha)$-confidence region

$$\begin{aligned} C_\alpha(Y_1, Y_2) \;&:= \big\{\rho_o > 0 : V(\rho_o) > \mathrm{Beta}_{a_1,a_2;\alpha}\big\} \\ &= \Big\{\rho_o > 0 : \frac{Y_1}{Y_1 + \rho_o Y_2} > \mathrm{Beta}_{a_1,a_2;\alpha}\Big\} \\ &= \Big(0,\; \frac{Y_1}{Y_2}\big(B^{-1}_{a_1,a_2;\alpha} - 1\big)\Big). \end{aligned}$$

for ρ. Analogously, an optimal unbiased level-α test of "$\rho \leq \rho_o$" rejects this null hypothesis if

$$V(\rho_o) \;\geq\; \mathrm{Beta}_{a_1,a_2;1-\alpha},$$

and this leads to the $(1-\alpha)$-confidence region

$$C_\alpha(Y_1, Y_2) \;:=\; \Big(\frac{Y_1}{Y_2}\big(B^{-1}_{a_1,a_2;1-\alpha} - 1\big),\; \infty\Big).$$

for ρ.

Chapter 10

Some Asymptotics

10.1 Testing, Total Variation and Hellinger Distances

Let P, Q be two probability distributions on a measurable space $(\Omega, \mathcal{A})$. 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 $\varphi : \Omega \to [0, 1]$ 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

$$\mathbb{P}(\text{error of 1st kind}) + \mathbb{P}(\text{error of 2nd kind}) \;=\; \int \varphi \, dP + \int (1 - \varphi) \, dQ.$$

Hence a natural measure of similarity between P and Q is given by the *testing affinity*

$$\eta_T(P, Q) \;:=\; \inf_{\text{tests } \varphi} \Big(\int \varphi \, dP + \int (1 - \varphi) \, dQ \Big).$$

As shown in the next lemma, the latter infimum is always a minimum.

Lemma 10.1 (Testing affinity and distance). *Suppose that P and Q have densities f and g, respectively, with respect to some measure M on $(\Omega, \mathcal{A})$. Then*

$$\eta_T(P, Q) \;=\; \int \min(f, g) \, dM.$$

The minimum is attained by any test φ such that $\varphi = 0$ on $\{f > g\}$ and $\varphi = 1$ on $\{f < g\}$.

There exists always a measure M dominating both P and Q, and the testing affinity $\eta_T(P, Q)$ does not depend on the choice of M. The related quantity

$$D_T(P, Q) \;:=\; 1 - \int \min(f, g) \, dM \;=\; \frac{1}{2} \int |f - g| \, dM$$

is the so-called testing distance between P and Q. It defines a metric on the space of probability measures on (Ω, M) with values in $[0, 1]$.

Proof of Lemma 10.1. Note first that for any test φ,

$$\begin{aligned}\int \varphi\, dP + \int (1-\varphi)\, dQ \;&=\; \int \big(\varphi(\omega) f(\omega) + (1-\varphi(\omega)) g(\omega)\big)\, M(d\omega) \\ &\geq\; \int \min\big(f(\omega), g(\omega)\big)\, M(d\omega)\end{aligned}$$

with equality if

$$\varphi(\omega) \;=\; \begin{cases} 0 & \text{if } f(\omega) > g(\omega), \\ 1 & \text{if } f(\omega) < g(\omega). \end{cases}$$

This proves already the first part of the lemma.

As to the second part, consider the particular measure $M_o := P + Q$. Then $M_o(\Omega) = 2$ and $P(A) = Q(A) = 0$ whenever $M_o(A) = 0$. Hence by the theorem of Radon–Nikodym there exist densities f_o and g_o of P and Q, respectively, with respect to M_o.

Now let M be an arbitrary measure such that $f := dP/dM$ and $g := dQ/dM$ exist. Then M_o has density $h := f + g$ with respect to M. Defining $f_o := f/h$ and $g_o := g/h$ with the convention $0/0 := 0$, we have

$$\int \min(f_o, g_o)\, dM_o \;=\; \int \min(f/h, g/h) h\, dM \;=\; \int \min(f, g)\, dM.$$

This shows that the infimum in $\eta_T(P, Q)$ is always a minimum, and its value does not depend on the choice of M.

Since $f, g \geq \min(f, g) \geq 0$, it is clear that $0 \leq \int \min(f, g)\, dM \leq 1$. Moreover,

$$\begin{aligned} 1 - \eta_T(P, Q) \;&=\; \int \Big(\frac{f+g}{2} - \min(f, g\Big)\, dM \\ &=\; \frac{1}{2} \int \big(f + g - 2\min(f, g)\big)\, dM \\ &=\; \frac{1}{2} \int \big(\max(f, g) - \min(f, g)\big)\, dM \\ &=\; \frac{1}{2} \int |f - g|\, dM. \end{aligned}$$

Obviously this equals 0 if, and only if, $f = g$ M-almost everywhere, which is equivalent to $P \equiv Q$.

Finally, if P, Q, R are probability distributions on $(\Omega, \mathcal{A})$, we may assume that there exists a finite measure M on $(\Omega, \mathcal{A})$ such that $f = dQ/dM$, $g = dQ/dM$ and $h = dR/dM$ exist. For instance, $M := P + Q + R$ would do the job. But then by the triangle inequality for the norm $\|\cdot\|_{M,1}$ in $L^1(M)$,

$$D_T(P, R) \;=\; \frac{\|f - h\|_{M,1}}{2} \;\leq\; \frac{\|f - g\|_{M,1} + \|g - h\|_{M,1}}{2} \;=\; D_T(P, Q) + D_T(Q, R).$$

□

Total variation distance. Another measure of distance is given by the total variation distance

$$D_{TV}(P,Q) \;:=\; \sup_{A\in\mathcal{A}}\big|P(A)-Q(A)\big|.$$

One could easily verify directly that this defines a metric on the space of probability measures on $(\Omega,\mathcal{A})$. But the next lemma shows that $D_{TV}=D_T$.

Lemma 10.2. *For arbitrary probability distributions P and Q on $(\Omega,\mathcal{A})$,*

$$D_{TV}(P,Q) \;=\; \max_{A\in\mathcal{A}}\big(P(A)-Q(A)\big) \;=\; \max_{B\in\mathcal{A}}\big(Q(B)-P(B)\big) \;=\; D_T(P,Q).$$

If P and Q have densities f and g as in Lemma 10.1, then the latter two maxima are attained for arbitrary sets A,B such that $\{f>g\}\subset A\subset\{f\geq g\}$ and $\{f<g\}\subset B\subset\{f\leq g\}$.

Proof of Lemma 10.2. Since $P(A)-Q(A)=Q(\Omega\setminus A)-P(\Omega\setminus A)$,

$$D_{TV}(P,Q) \;=\; \sup_{A\in\mathcal{A}}\big(P(A)-Q(A)\big) \;=\; \sup_{B\in\mathcal{A}}\big(Q(B)-P(B)\big).$$

In case of P and Q having densities f and g, respectively, with respect to some measure M on $(\Omega,\mathcal{A})$,

$$P(A)-Q(A) \;=\; \int 1_A(f-g)\,dM \;\leq\; \int (f-g)^+\,dM$$

with equality if $\{f>g\}\subset A\subset\{f\geq g\}$. Analogously,

$$Q(B)-P(B) \;\leq\; \int (g-f)^+\,dM \;=\; \int (f-g)^-\,dM$$

with equality if $\{f<g\}\subset B\subset\{f\leq g\}$. Consequently,

$$D_{TV}(P,Q) \;=\; \frac{1}{2}\int\big((f-g)^+ + (f-g)^-\big)\,dM \;=\; \frac{1}{2}\int|f-g|\,dM \;=\; D_T(P,Q).$$

□

Hellinger affinity and distance. In many situations it turns out that $D_T(P,Q)$ is difficult to compute explicitly. As we shall see later, interesting proxys are given by the

$$\text{Hellinger affinity}\quad \eta_H(P,Q) \;:=\; \int\sqrt{fg}\,dM,$$

and

$$\text{Hellinger distance}\quad D_H(P,Q) \;:=\; \sqrt{\frac{1}{2}\int\big(\sqrt{f}-\sqrt{g}\big)^2\,dM},$$

where M is some measure on $(\Omega,\mathcal{A})$ such that $f:=dP/dM$ and $g:=dQ/dM$ exist. As in case of $\eta_T(P,Q)$ and $D_T(P,Q)$, the choice of M is irrelevant. Note also that

$$D_H(P,Q)^2 \;=\; \frac{1}{2}\int\big(f+g-2\sqrt{fg}\big)\,dM \;=\; 1-\eta_H(P,Q).$$

The following lemma shows that testing and Hellinger distance induce the same topology on the space of probability distributions on $(\Omega,\mathcal{A})$.

Lemma 10.3 (Relationships between testing and Hellinger distance).

$$1 - \sqrt{1 - \eta_H^2} \;\le\; \eta_T \;\le\; \eta_H$$

and

$$D_H^2 \;\le\; D_T \;\le\; D_H \sqrt{2 - D_H^2}.$$

Proof of Lemma 10.3. With explicit densities $f := dP/dM$ and $g := dQ/dM$ it follows from $\min(a, b) \le \sqrt{ab}$ for real numbers $a, b \ge 0$ that

$$\eta_T(P, Q) \;=\; \int \min(f, g)\, dM \;\le\; \int \sqrt{fg}\, dM \;=\; \eta_H(P, Q).$$

In particular,

$$D_T(P, Q) \;=\; 1 - \eta_T(P, Q) \;\ge\; 1 - \eta_H(P, Q) \;=\; D_H(P, Q)^2.$$

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

$$\begin{aligned}
1 - \eta_T(P, Q) = D_T(P, Q) \;&=\; \frac{1}{2} \int \bigl|\sqrt{f} - \sqrt{g}\bigr| \bigl|\sqrt{f} + \sqrt{g}\bigr|\, dM \\
&\le\; \frac{1}{2} \sqrt{\int \bigl(f + g - 2\sqrt{fg}\bigr)\, dM \int \bigl(f + g + 2\sqrt{fg}\bigr)\, dM} \\
&=\; \sqrt{(1 - \eta_H(P, Q))(1 + \eta_H(P, Q))} \\
&=\; \begin{cases} \sqrt{1 - \eta_H(P, Q)^2}, \\ D_H(P, Q) \sqrt{2 - D_H(P, Q)^2}. \end{cases}
\end{aligned}$$

This proves that

$$D_T \;\le\; D_H \sqrt{2 - D_H^2} \quad \text{and} \quad 1 - \eta_T \;\le\; \sqrt{1 - \eta_H^2},$$

where the latter inequality is equivalent to $\eta_T \ge 1 - \sqrt{1 - \eta_H^2}$. □

Remark 10.4. 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

$$\begin{aligned}
\eta_T(P, Q) \;&=\; \int \min(dP, dQ), \\
D_T(P, Q) \;&=\; \frac{1}{2} \int |dP - dQ|, \\
\eta_H(P, Q) \;&=\; \int \sqrt{dP\, dQ}, \\
D_H(P, Q) \;&=\; \sqrt{\frac{1}{2} \int \bigl(\sqrt{dP} - \sqrt{dQ}\bigr)^2}.
\end{aligned}$$

Remark 10.5. Let $(\Omega', \mathcal{A}')$ be a second measurable space, and let $\tau : \Omega \to \Omega'$ be a bijection such that both τ and τ^{-1} are measurable. Then

$$\kappa(P, Q) \;=\; \kappa(P^\tau, Q^\tau) \quad \text{for } \kappa = \eta_T, D_T, \eta_H, D_H.$$

The proof of this claim is left to the reader as an exercise; see also Exercise 2.19.

Example 10.6 (Testing and Hellinger distance for univariate Gaussian shift). For real numbers μ_1, μ_2 and $\sigma > 0$,

$$\begin{aligned}\eta_T\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big) &= 2\Phi\Big(-\frac{|\mu_1-\mu_2|}{2\sigma}\Big),\\ \eta_H\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big) &= \exp\Big(-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}\Big),\end{aligned}$$

where Φ is the standard Gaussian distribution function. Hence

$$\begin{aligned}D_T\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big) &= 2\Phi\Big(\frac{|\mu_1-\mu_2|}{2\sigma}\Big) - 1,\\ D_H\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big) &= \sqrt{1-\exp\Big(-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}\Big)}.\end{aligned}$$

To verify these formulae, we first apply Remark 10.5 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 $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(\mu,1)$, where

$$\mu := \frac{|\mu_1-\mu_2|}{\sigma}.$$

But with $C = (2\pi)^{-1/2}$ and $\phi(x) := C\exp(-x^2/2)$,

$$\begin{aligned}\eta_T\big(\mathcal{N}(0,1),\mathcal{N}(\mu,1)\big) &= \int \min\big(\phi(x),\phi(x-\mu)\big)\,dx\\ &= C\int \exp\Big(-\frac{\max(x^2,(x-\mu)^2)}{2}\Big)\,dx\\ &= 2C\int_{-\infty}^{\mu/2}\exp\Big(-\frac{(x-\mu)^2}{2}\Big)\,dx\\ &= 2\int_{-\infty}^{\mu/2}\phi(x-\mu)\,dx\\ &= 2\Phi(-\mu/2).\end{aligned}$$

Moreover,

$$\begin{aligned}\eta_H\big(\mathcal{N}(0,1),\mathcal{N}(\mu,1)\big) &= \int \sqrt{\phi(x)\phi(x-\mu)}\,dx\\ &= C\int \exp\Big(-\frac{x^2+(x-\mu)^2}{4}\Big)\,dx\\ &= C\int \exp\Big(-\frac{x^2-x\mu+\mu^2/2}{2}\Big)\,dx\\ &= C\int \exp\Big(-\frac{(x-\mu/2)^2+\mu^2/4}{2}\Big)\,dx\\ &= \exp(-\mu^2/8)\int \phi(x-\mu/2)\,dx\\ &= \exp(-\mu^2/8).\end{aligned}$$

Exercise 10.7 (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

$$\begin{aligned}\eta_T\big(\mathcal{N}_d(\mu_1,\Sigma),\mathcal{N}_d(\mu_2,\Sigma)\big) &= 2\Phi\Big(-\sqrt{(\mu_2-\mu_1)^\top\Sigma^{-1}(\mu_2-\mu_1)}/2\Big),\\ \eta_H\big(\mathcal{N}_d(\mu_1,\Sigma),\mathcal{N}_d(\mu_2,\Sigma)\big) &= \exp\big(-(\mu_2-\mu_1)^\top\Sigma^{-1}(\mu_2-\mu_1)/8\big).\end{aligned}$$

Hint: Consider the transformation $\tau(x) := \Sigma^{-1/2}(x-\mu_1)$ of $x \in \mathbb{R}^d$.

Exercise 10.8 (More about the relation between testing and Hellinger distance). **(a)** Show that the inequalities

$$D_H^2 \;\le\; D_T \;\le\; D_H\sqrt{2-D_H^2}$$

are equivalent to

$$\sqrt{1-\sqrt{1-D_T^2}} \;\le\; D_H \;\le\; \sqrt{D_T}.$$

(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 $(\Omega,\mathcal{A})$ such that $D_T(P_\gamma,Q_\gamma) = D_H^2(P_\gamma,Q_\gamma) = \gamma$.

Remark 10.9 (Product measures). For $j = 1,2$, let P_j and Q_j be probability measures on a measurable space $(\Omega_j,\mathcal{A}_j)$. Then

$$\eta_H(P_1\otimes P_2, Q_1\otimes Q_2) \;=\; \eta_H(P_1,Q_1)\eta_H(P_2,Q_2).$$

For if $dP_j/dM_j = f_j$ and $dQ_j/dM_j = g_j$, then by Fubini's theorem,

$$\begin{aligned}\eta_H(P_1\otimes P_2, Q_1\otimes Q_2) &= \int_{\Omega_1\times\Omega_2}\sqrt{f_1(\omega_1)f_2(\omega_2)g_1(\omega_1)g_2(\omega_2)}\,M_1\otimes M_2(d(\omega_1,\omega_2))\\ &= \int_{\Omega_1}\sqrt{f_1g_1}\,dM_1\int_{\Omega_2}\sqrt{f_2g_2}\,dM_2\\ &= \eta_H(P_1,Q_1)\eta_H(P_2,Q_2).\end{aligned}$$

Inductively this implies that

$$\eta_H(P^{\otimes n},Q^{\otimes n}) \;=\; \eta_H(P,Q)^n$$

for arbitrary integers $n \ge 1$.

10.2 Asymptotics for Repeated Binary Experiments

Suppose we observe independent random variables $X_1,\dots,X_n$ with unknown distribution $R \in \{P,Q\}$, where P and Q are two different given probability distributions on $(\Omega,\mathcal{A})$. Then

$$D_T(P^{\otimes n},Q^{\otimes n}) \;\ge\; D_H(P^{\otimes n},Q^{\otimes n})^2 \;=\; 1-\eta_H(P,Q)^n$$

converges to 1 exponentially fast. That means, there exists a sequence of tests $\varphi_n : \Omega^n \to \{0,1\}$ 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 $(\Omega, \mathcal{A})$. 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.

$$D_T(P_n^{\otimes n}, Q_n^{\otimes n}) \;\to\; 0.$$

- They are (asymptotically) perfectly distinguishable, i.e.

$$D_T(P_n^{\otimes n}, Q_n^{\otimes n}) \;\to\; 1.$$

- They are (asymptotically) "interesting" in the sense that

$$\liminf_{n\to\infty} D_T(P_n^{\otimes n}, Q_n^{\otimes n}) \;>\; 0 \quad \text{and} \quad \limsup_{n\to\infty} D_T(P_n^{\otimes n}, Q_n^{\otimes n}) \;<\; 1.$$

It follows from Lemma 10.3 and Exercise 10.8 (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

$$\begin{aligned} D_H(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;&=\; 1 - \eta_H(P_n^{\otimes n}, Q_n^{\otimes n}) \\ &=\; 1 - \eta_H(P_n, Q_n)^n \\ &=\; 1 - \big(1 - D_H(P_n, Q_n)^2\big)^n, \end{aligned}$$

and the subsequent Lemma 10.11 shows that

$$D_H(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;=\; 1 - \exp\big(-nD_H(P_n, Q_n)^2\big) + O(n^{-1}).$$

This implies the following results:

Lemma 10.10. ***(a)*** *For $a \in \{0, 1\}$, the following two conditions are equivalent:*

$$\begin{aligned} \lim_{n\to\infty} D_T(P_n^{\otimes n}, Q_n^{\otimes n}) \;&=\; a, \\ \lim_{n\to\infty} nD_H(P_n, Q_n)^2 \;&=\; \begin{cases} \infty & \text{if } a = 1, \\ 0 & \text{if } a = 0. \end{cases} \end{aligned}$$

(b) *As $n \to \infty$, the distance $D_T(P_n^{\otimes n}, Q_n^{\otimes n})$ stays bounded away from 0 and 1 if and only if $nD_H(P_n, Q_n)^2$ stays bounded away from 0 and ∞.*

Lemma 10.11 (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 10.11. 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

$$\max_{x \in [0,1]} H_1(x) \;=\; H_1(0) \;\leq\; e^{-1}.$$

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,

$$\begin{aligned} \max_{x \in [0,1]} H_n(x) = H_n(x_n) \;&=\; e^{-nx_n} - (1-x_n)e^{-nx_n} \\ &=\; n^{-1}\, nx_n e^{-nx_n} \\ &\leq\; n^{-1} \max_{s \geq 0} se^{-s} \\ &=\; n^{-1}e^{-1}, \end{aligned}$$

because elementary calculations show that se^{-s} is maximized for $s = 1$. □

Expansions of root-densities and log-likelihood ratios. Suppose that for some measure M on $(\Omega, \mathcal{A})$ and arbitrary $n \geq 1$, the densities

$$f_n \;:=\; \frac{dP_n}{dM} \quad \text{and} \quad g_n \;:=\; \frac{dQ_n}{dM}$$

exist. Suppose that these densities satisfy the following condition:

(C1) For some probability measure P on $(\Omega, \mathcal{A})$ with density $f = dP/dM$,

$$f_n \;\to\; f \quad \text{in } L^1(M).$$

Furthermore, for some function $h \in L^2(M)$ with $\|h\|_{M,2} > 0$,

$$h_n := \sqrt{n}\big(\sqrt{g_n} - \sqrt{f_n}\big) \;\to\; h \quad \text{in } L^2(M).$$

Here and throughout the sequel, $L^r(M)$ is the space of all (equivalence classes of) measurable functions functions $h : \Omega \to \mathbb{R}$ such that $\|h\|_{M,r} < \infty$, where

$$\|h\|_{M,r} \;:=\; \Big(\int |h|^r \, dM\Big)^{1/r}$$

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 10.12. *Under Condition (C1),*

$$nD_H(P_n, Q_n)^2 \;\to\; \|h\|_{M,2}^2/2,$$

and

$$D_T(P, P_n), D_T(P, Q_n), D_T(P_n, Q_n) \;\to\; 0.$$

Proof of Lemma 10.12. 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

$$\begin{aligned}
\int |h_n^2 - h^2|\, dM \;&=\; \int |h_n - h||h_n + h|\, dM \\
&\le\; \|h_n - h\|_{M,2}\|h_n + h\|_{M,2} && \text{(Cauchy–Schwarz)} \\
&\le\; \|h_n - h\|_{M,2}\big(2\|h\|_{M,2} + \|h_n - h\|_{M,2}\big) && \text{(triangle inequality)} \\
&\to\; 0.
\end{aligned}$$

In particular,

$$nD_H(P_n, Q_n)^2 \;=\; \|h_n\|_{M,2}^2/2 \;\to\; \|h\|_{M,2}^2/2.$$

Note that $f_n \to f$ in $L^1(M)$ is equivalent to $D_T(P, P_n) \to 0$, and by Lemma 10.3, $D_T(P_n, Q_n) \le \sqrt{2}\, D_H(P_n, Q_n) \to 0$. Hence by the triangle inequality, $D_T(P, Q_n) \to 0$ as well. □

For the next result we have to augment Condition (C1) by an additional one:

(C2) The functions f and h in (C1) satisfy

$$M(h \ne 0 = f) \;=\; 0.$$

Lemma 10.13. *Suppose that Conditions (C1-2) are satisfied. Let $(A_n)_n$ be an arbitrary sequence of events $A_n \in \mathcal{A}$ 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 10.13. In terms of the densities f_n and g_n we may write

$$\begin{aligned}
n\big(Q_n(A_n) - P_n(A_n)\big) \;&=\; n\int_{A_n} (g_n - f_n)\, dM \\
&=\; n\int_{A_n} \big(\sqrt{g_n} - \sqrt{f_n}\big)\big(\sqrt{g_n} + \sqrt{f_n}\big)\, dM \\
&=\; \begin{cases} \displaystyle\int_{A_n} h_n\big(2\sqrt{nf_n} + h_n\big)\, dM, \\ \displaystyle\int_{A_n} h_n\big(2\sqrt{ng_n} - h_n\big)\, dM. \end{cases}
\end{aligned}$$

Consequently, by the Cauchy–Schwarz inequality,

$$n\big|Q_n(A_n) - P_n(A_n)\big| \;\le\; 2\sqrt{n\min\{P_n(A_n), Q_n(A_n)\}\int_{A_n} h_n^2\,dM} + \int_{A_n} h_n^2\,dM.$$

Hence it suffices to show that

$$\int_{A_n} h_n^2\,dM \;\to\; 0.$$

Since $h_n^2 \to h^2$ in $L^1(M)$, this is equivalent to

$$\int_{A_n} h^2\,dM \;\to\; 0.$$

But it follows from Lemma 10.12 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 $C > 0$,

$$\int_{A_n} h^2\,dM \;\le\; \int_{\{h^2 > Cf\}} h^2\,dM + CP(A_n) \;\to\; \int_{\{h^2 > Cf\}} h^2\,dM,$$

and

$$\lim_{C\to\infty} \int_{\{h^2 > Cf\}} h^2\,dM \;=\; \int_{\{h^2 > 0 = f\}} h^2\,dM \;=\; 0$$

by dominated convergence and our assumption (C2) 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

$$\Lambda_n \;:=\; \log\Big(\frac{dQ_n^{\otimes n}}{dM^{\otimes n}}(\boldsymbol{X}_n)\Big/\frac{dP_n^{\otimes n}}{dM^{\otimes n}}(\boldsymbol{X}_n)\Big) \;=\; \sum_{i=1}^n \log\frac{g_n}{f_n}(X_{ni}) \;\in\; [-\infty,\infty]$$

with the conventions that $\log(0) := -\infty$, $\log(\infty) := \infty$, $a/0 := \infty$ for $a > 0$ and $0/0 := 0$. 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 10.14. *Under condition (C),*

$$\Lambda_n \;\to_{\mathcal{L}}\; \begin{cases} \mathcal{N}\big(-2\|h\|_{M,2}^2, 4\|h\|_{M,2}^2\big) & \text{if } R_n = P_n \text{ for all } n, \\ \mathcal{N}\big(+2\|h\|_{M,2}^2, 4\|h\|_{M,2}^2\big) & \text{if } R_n = Q_n \text{ for all } n. \end{cases}$$

In this result "$\to_{\mathcal{L}}$" stands for convergence in distribution, meaning that for any fixed continuous function $J : [-\infty,\infty] \to \mathbb{R}$,

$$\mathbb{E}\,J(\Lambda_n) \;\to\; \begin{cases} \mathbb{E}\,J(2\|h\|_{M,2}Z - 2\|h\|_{M,2}^2) & \text{if } R_n = P_n \text{ for all } n \\ \mathbb{E}\,J(2\|h\|_{M,2}Z + 2\|h\|_{M,2}^2) & \text{if } R_n = Q_n \text{ for all } n \end{cases}$$

with a random variable $Z \sim \mathcal{N}(0,1)$. Since the limiting distributions are continuous, and since

$$\mathbb{P}\big(2\|h\|_{M,2}Z \mp 2\|h\|_{M,2}^2 \le k\big) \;=\; \Phi\Big(\frac{k \pm 2\|h\|_{M,2}^2}{2\|h\|_{M,2}}\Big),$$

Theorem 10.14 may be rephrased as follows: For arbitrary $k \in \mathbb{R}$,

$$\mathbb{P}(\Lambda_n \le k) \;\to\; \begin{cases} \Phi\Big(\dfrac{k + 2\|h\|_{M,2}^2}{2\|h\|_{M,2}}\Big) & \text{if } R_n = P_n \text{ for all } n, \\ \Phi\Big(\dfrac{k - 2\|h\|_{M,2}^2}{2\|h\|_{M,2}}\Big) & \text{if } R_n = Q_n \text{ for all } n. \end{cases}$$

This implies the following result about optimal tests of "$R_n = P_n$" versus "$R_n = Q_n$":

Corollary 10.15. *Let $\varphi_{n,\alpha} : \Omega^n \to [0,1]$ be an optimal level-α test of "$R_n = P_n$" versus "$R_n = Q_n$". Then*

$$\mathbb{E}_{Q_n}\, \varphi_{n,\alpha}(\boldsymbol{X}_n) \;\to\; \Phi\big(\Phi^{-1}(\alpha) + 2\|h\|_{M,2}\big).$$

Remark 10.16. Theorem 10.14 and Corollary 10.15 show that under Conditions (C1-2), testing $P_n^{\otimes n}$ versus $Q_n^{\otimes n}$ is asymptotically as difficult as testing

$$\mathcal{N}(0,1) \quad \text{versus} \quad \mathcal{N}(\mu,1),$$

where

$$\mu \;:=\; 2\|h\|_{M,2}.$$

Indeed,

$$D_H(P_n^{\otimes n}, Q_n^{\otimes n})^2 \;=\; 1 - \exp\big(-nD_H(P_n,Q_n)^2\big) + O(n^{-1})$$

converges to

$$1 - \exp(-\|h\|_{M,2}^2/2) \;=\; 1 - \exp\big(-\mu^2/8\big) \;=\; D_H\big(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\big)^2.$$

Proof of Corollary 10.15. 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 $k \in \mathbb{R}$,

$$\mathbb{P}_{P_n}(\Lambda_n > k) \;\to\; 1 - \Phi\Big(\frac{k + 2\|h\|_{M,2}^2}{2\|h\|_{M,2}}\Big).$$

The right hand side is strictly decreasing in k and equals α if, and only if, k is equal to

$$k_\alpha := -2\|h\|_{M,2}^2 + 2\|h\|_{M,2}\Phi^{-1}(1-\alpha).$$

Hence $k_{n,\alpha} \to k_\alpha$, and $\mathbb{E}_{Q_n}\, \varphi_{n,\alpha}(\boldsymbol{X}_n)$ converges to

$$1 - \Phi\Big(\frac{k_\alpha - 2\|h\|_{M,2}^2}{2\|h\|_{M,2}}\Big) \;=\; 1 - \Phi\big(-2\|h\|_{M,2} + \Phi^{-1}(1-\alpha)\big) \;=\; \Phi\big(\Phi^{-1}(\alpha) + 2\|h\|_{M,2}\big).$$

□

Proof of Theorem 10.14. 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 Condition (C) would still be satisfied with $-h$ in place of h.

Since $\sqrt{g_n} = \sqrt{f_n} + h_n/\sqrt{n}$, we may write

$$\Lambda_n \;=\; 2\sum_{i=1}^{n} \log \frac{\sqrt{g_n}}{\sqrt{f_n}}(X_{ni}) \;=\; 2\sum_{i=1}^{n} \log\Bigl(1 + \frac{\tilde{h}_n(X_{ni})}{\sqrt{n}}\Bigr)$$

with $\tilde{h}_n := h_n/\sqrt{f_n} \in \bigl[-\sqrt{n}, \infty\bigr]$. It follows from the wellknown 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

$$\Lambda_n \;=\; \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \tilde{h}_n(X_{ni}) - \frac{1}{n}\sum_{i=1}^{n} \tilde{h}_n(X_{ni})^2 + \mathrm{Rem}_n \tag{10.1}$$

with

$$|\mathrm{Rem}_n| \;\le\; \frac{2D_n}{3(1-D_n)^+} \cdot \frac{1}{n}\sum_{i=1}^{n} \tilde{h}_n(X_{ni})^2.$$

Now we apply the Central Limit Theorem as formulated in Corollary A.24: Suppose we can show that

$$\sqrt{n}\int \tilde{h}_n \, dP_n \;\to\; \mu, \tag{10.2}$$

$$\int \tilde{h}_n^2 \, dP_n \;\to\; \sigma^2, \tag{10.3}$$

$$\int \tilde{h}_n^2 1_{[\tilde{h}_n^2 \ge n\epsilon]}) \, dP_n \;\to\; 0 \quad \text{for any fixed } \epsilon > 0. \tag{10.4}$$

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{h}_n(X_{ni}) \;\to_{\mathcal{L}}\; \mathcal{N}(\mu,\sigma^2), \quad \frac{1}{n}\sum_{i=1}^{n} \tilde{h}_n(X_{ni})^2 \;\to_p\; \sigma^2, \quad D_n \;\to_p\; 0.$$

These facts and (10.1) imply that

$$\Lambda_n \;\to_{\mathcal{L}}\; \mathcal{N}(2\mu - \sigma^2, 4\sigma^2).$$

Consequently, it suffices to verify (10.2) with $\mu = -\|h\|_{M,2}^2/2$, (10.3) with $\sigma^2 = \|h\|_{M,2}^2$ and (10.4).

As to (10.2),

$$\begin{aligned}\sqrt{n}\int \tilde{h}_n\,dP_n &= n\int \frac{\sqrt{g_n}-\sqrt{f_n}}{\sqrt{f_n}}\,f_n\,dM\\ &= n\int\bigl(\sqrt{f_n g_n}-f_n\bigr)\,dM\\ &= n(\eta_H(P_n,Q_n)-1)\\ &= -nD_H(P_n,Q_n)^2\\ &\to -\|h\|_{M,2}^2/2,\end{aligned}$$

see Lemma 10.12. Concerning (10.3),

$$\begin{aligned}\int \tilde{h}_n^2\,dP_n &= \int \frac{h_n^2}{f_n}\,f_n\,dM\\ &= \int_{\{f_n>0\}} h_n^2\,dM\\ &= \int_{\{f_n>0\}} h^2\,dM + o(1)\\ &= \int h^2\,dM - \int_{\{f_n=0\}} h^2\,dM + o(1)\\ &\to \|h\|_{M,2}^2,\end{aligned}$$

because $A_n := \{f_n = 0\}$ satisfies $P(A_n) = P_n(A_n) + o(1) = o(1)$, see also the proof of Lemma 10.13.

It remains to verify (10.4), that means, for any fixed $\epsilon > 0$,

$$\int_{\{h_n^2 \ge \epsilon^2 n f_n\}} h_n^2\,dM \;\to\; 0.$$

Again, since $h_n^2 \to h^2$ in $L^1(M)$, it suffices to show that

$$\int_{\{h_n^2 \ge \epsilon^2 n f_n\}} h^2\,dM \;\to\; 0,$$

and the left hand side is equal to

$$\int_0^\infty M(h^2 > r \text{ and } h_n^2 \ge \epsilon^2 n f_n)\,dr \;\le\; \int_0^\infty M(h^2 > r)\,dr \;=\; \int h^2\,dM.$$

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 $\delta > 0$,

$$\begin{aligned}&M(h^2 > r \text{ and } h_n^2 \ge \epsilon^2 n f_n)\\ &\quad\le M\bigl(h^2 > r \text{ and } h^2 + \delta \ge \epsilon^2 n(f-\delta)\bigr) + M(|h_n^2 - h^2| \ge \delta) + M(|f_n - f| \ge \delta)\\ &\quad\le M\bigl(h^2 > r \text{ and } h^2 + \delta \ge \epsilon^2 n(f-\delta)\bigr) + \delta^{-1}\|h_n^2 - h^2\|_{M,1} + \delta^{-1}\|f_n - f\|_{M,1}\\ &\quad\to M(h^2 > r \text{ and } f \le \delta).\end{aligned}$$

Letting $\delta \downarrow 0$, the right hand side converges to $M(h^2 > r \text{ and } f = 0)$, and this equals 0 by assumption (C2). □

10.3 Fisher Information

Consider a statistical experiment $\mathcal{E} = (\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ 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 $(\Omega, \mathcal{A})$ such that for M-almost every $\omega \in \Omega$,

$$\Theta \ni \theta \;\mapsto\; f_\theta(\omega)$$

is continuously differentiable with gradient

$$\dot{f}_\theta(\omega) \;:=\; \Big(\frac{\partial f_\theta(\omega)}{\partial \theta_i}\Big)_{i=1}^d.$$

Assuming that for any $\theta \in \Theta$, each component of $f_\theta^{-1/2}\dot{f}_\theta$ belongs to $L^2(M)$, the matrix

$$J(\theta) \;:=\; \int \frac{\dot{f}_\theta \dot{f}_\theta^\top}{f_\theta}\, dM$$

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 M-almost every $\omega \in \Omega$ the mapping

$$\Theta \ni \theta \;\mapsto\; \lambda_\theta(\omega)$$

is continuously differentiable with gradient

$$\dot{\lambda}_\theta(\omega) \;:=\; \Big(\frac{\dot{f}_\theta}{f_\theta}(\omega)\Big)_{i=1}^d,$$

and

$$J(\theta) \;=\; \int \dot{\lambda}_\theta \dot{\lambda}_\theta^\top\, dP_\theta.$$

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

$$n\int \big(\sqrt{f_{\theta_n}} - \sqrt{f_\theta}\big)^2\, dM \;\to\; \int h_{\theta,\delta}^2\, dM, \tag{10.5}$$

then by Scheffé's theorem, Condition (C) in the previous section is 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 $\mathcal{N}(0,1)$ versus $\mathcal{N}(\mu,1)$ with

$$\mu \;=\; 2\|h\|_{M,2} \;=\; \sqrt{\delta^\top J(\theta)\delta}.$$

Condition (10.5) is satisfied, if

$$\eta_H(P_\theta, P_{\theta+\delta}) \;=\; 1 - \frac{\delta^\top J(\theta)\delta}{8} + o(\|\delta\|^2) \quad \text{as } \delta \to 0. \tag{10.6}$$

Definition 10.17 (Regular statistical experiment). A statistical experiment $\mathcal{E} = (\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ satisfying the conditions above is called *regular statistical experiment.*

Example 10.18 (Fisher information in natural exponential families). Suppose that

$$f_\theta \;=\; \exp\bigl(\theta^\top T - \kappa(\theta)\bigr)$$

for some measurable mapping $T : (\Omega, \mathcal{A}) \to \mathbb{R}^d$. Then the experiment $\mathcal{E}$ is regular: Obviously, $f_\theta > 0$. As shown in the exercises, κ is infinitely often differentiable with

$$\kappa(\theta+\delta) \;=\; \kappa(\theta) + \delta^\top \mathbb{E}_\theta(T) + \frac{1}{2}\,\delta^\top \operatorname{Var}_\theta(T)\delta + o(\|\delta\|^2)$$

as $\delta \to 0$. In particular $f_\theta(\omega)$ is a smooth function of θ with

$$\dot f_\theta \;=\; (T - \mathbb{E}_\theta(T))f_\theta, \quad \dot\lambda_\theta \;=\; T - \mathbb{E}_\theta(T),$$

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 $\mathbb{R}^d$.

Condition (10.6) is satisfied, because

$$\eta_H(P_\theta, P_{\theta+\delta}) \;=\; \exp\bigl(\kappa(\theta+\delta/2) - \kappa(\theta)/2 - \kappa(\theta+\delta)/2\bigr)$$

and

$$\begin{aligned}\kappa(\theta+\delta/2) - \kappa(\theta)/2 - \kappa(\theta+\delta)/2 \;&=\; \frac{\delta^\top J(\theta)\delta}{8} - \frac{\delta^\top J(\theta)\delta}{4} + o(\|\delta\|^2)\\ &=\; -\frac{\delta^\top J(\theta)\delta}{8} + o(\|\delta\|^2).\end{aligned}$$

Remark 10.19 (Smooth transformations of parameters). Let $\mathcal{E} = (\Omega, \mathcal{A}, (P_\theta)_{\theta\in\theta})$ be a regular statistical experiment with Fisher information $J(\cdot)$, and let $\tau : \Psi \to \Omega$ 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

$$D\tau(\psi) \;=\; \Bigl(\frac{\tau_i(\psi)}{\partial\psi_j}\Bigr)_{i,j=1}^d$$

for any $\psi \in \Psi$.

Then the experiment $\tilde{\mathcal{E}} := (\Omega, \mathcal{A}, (\tilde P_\psi)_{\psi\in\Psi})$ 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

$$\frac{\partial \tilde f_\psi(\omega)}{\partial\psi_j} \;=\; \sum_{i=1}^d \frac{\partial f_\theta(\omega)}{\partial\theta_i}\frac{\partial\tau_i(\psi)}{\partial\psi_j} \;=\; \bigl(D\tau(\psi)^\top \dot f_\theta\bigr)_j.$$

Moreover, as $\delta \to 0$,

$$\Delta \;:=\; \tau(\psi+\delta)-\tau(\psi) \;=\; D\tau(\psi)\delta + o(\|\delta\|) \;=\; O(\|\delta\|),$$

whence

$$\begin{aligned}\eta_H(\tilde{P}_{\psi+\delta},\tilde{P}_\psi) \;&=\; \eta_H(P_{\theta+\Delta},P_\theta)\\ &=\; 1-\frac{\Delta^\top J(\theta)\Delta}{8}+o(\|\Delta\|^2)\\ &=\; 1-\frac{\delta^\top D\tau(\psi)^\top J(\theta)D\tau(\psi)\delta}{8}+o(\|\delta\|^2).\end{aligned}$$

Example 10.20 (Binomial distributions). We observe $X \sim \mathrm{Bin}(n,p)$ with an unknown parameter $p \in (0,1)$. The natural parameter for the experiment $\big(\mathrm{Bin}(n,p)\big)_{p\in(0,1)}$ is given by $\tau(p) := \log(p/(1-p))$ with sufficient statistic $T = X$, 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 $\widehat{p} := \bar{X}$, which is not a coincidence as explained later.

Example 10.21 (Poisson distributions). We observe $X \sim \mathrm{Poiss}(\lambda)$ with an unknown parameter $\lambda > 0$. The natural parameter for the experiment $\big(\mathrm{Poiss}(\lambda)\big)_{\lambda>0}$ is given by $\tau(\lambda) := \log\lambda$ with sufficient statistic $T = X$, and $\tau : (0,\infty) \to \mathbb{R}$ is a diffeomrophism wit $\tau'(\lambda) = \lambda^{-1}$. 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 $\widehat{\lambda} := X$.

Implications for point estimation

With our results about testing in Section 10.2 one can prove various precision bounds for point estimators. We present one particular result:

Theorem 10.22 (Asymptotic version of the Cramér–Rao bound). *Let $\mathcal{E} = (\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ be a regular statistical experiment with Fisher information $J(\cdot)$. For each sample size $n \geq 1$ let $\widehat{\theta}_n : \Omega^n \to \mathbb{R}^d$ be an estimator such that for a fixed $\theta \in \Theta$,*

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{X}_n)-\theta_n) \;\to_{\mathcal{L}}\; \mathcal{N}_d(0,\Sigma(\theta))$$

whenever $\theta_n = \theta + O(n^{-1/2})$ and $\boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}$. If $J(\theta)$ is positive definite, then

$$\Sigma(\theta) \;\geq\; J(\theta)^{-1}$$

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 10.22. For a fixed vector $\delta \in \mathbb{R}^d \setminus \{0\}$ define

$$\theta_n \;:=\; \theta + n^{-1/2}\delta.$$

Then for any vector $\eta \in \mathbb{R}^d \setminus \{0\}$,

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{X}_n) - \theta)^\top \eta \;\to_{\mathcal{L}}\; \begin{cases} \mathcal{N}\big(0, \eta^\top \Sigma(\theta)\eta\big) & \text{if } \boldsymbol{X}_n \sim P_\theta^{\otimes n}, \\ \mathcal{N}\big(\delta^\top \eta, \eta^\top \Sigma(\theta)\eta\big) & \text{if } \boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}, \end{cases}$$

because

$$\sqrt{n}(\widehat{\theta}_n - \theta)^\top \eta \;=\; \sqrt{n}(\widehat{\theta}_n - \theta_n)^\top \eta + \delta^\top \eta.$$

Consequently,

$$\varphi_n \;:=\; \left\{\begin{matrix} 1 \\ 0 \end{matrix}\right\} \quad \text{if} \quad \frac{\sqrt{n}(\widehat{\theta}_n - \theta)^\top \eta}{\sqrt{\eta^\top \Sigma(\theta)\eta}} \left\{\begin{matrix} > \\ \leq \end{matrix}\right\} \Phi^{-1}(1-\alpha)$$

defines a test of $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$ such that

$$\mathbb{E}\,\varphi_n(\boldsymbol{X}_n) \;\to\; \begin{cases} \alpha & \text{if } \boldsymbol{X}_n \sim P_\theta^{\otimes n}. \\ \Phi\Big(\Phi^{-1}(\alpha) + \dfrac{\delta^\top \eta}{\sqrt{\eta^\top \Sigma(\theta)\eta}}\Big) & \text{if } \boldsymbol{X}_n \sim P_{\theta_n}^{\otimes n}. \end{cases}$$

If we would replace φ_n with the optimal level-α test of $P_\theta^{\otimes n}$ versus $P_{\theta_n}^{\otimes n}$, then the asymptotic power under the alternative hypothesis would be

$$\Phi\big(\Phi^{-1}(\alpha) + \sqrt{\delta J(\theta)\delta}\big).$$

Consequently,

$$\frac{\delta^\top \eta}{\sqrt{\eta^\top \Sigma(\theta)\eta}} \;\leq\; \sqrt{\delta^\top J(\theta)\delta}.$$

In other words, for arbitrary $\delta, \eta \in \mathbb{R}^d \setminus \{0\}$,

$$\delta^\top \eta \;\leq\; \sqrt{\eta^\top \Sigma(\theta)\eta}\sqrt{\delta^\top J(\theta)\delta}.$$

Setting $\delta = J(\theta)^{-1}\eta$ yields the inequality

$$\eta^\top J(\theta)^{-1}\eta \;\leq\; \eta^\top \Sigma(\theta)\eta$$

for arbitrary $\eta \in \mathbb{R}^d \setminus \{0\}$. □

Example 10.23 (Maximum-likelihood estimation in natural exponential families). Let $\mathcal{E}$ be a natural exponential family with sufficient statistic $T : (\Omega, \mathcal{A}) \to \mathbb{R}^d$ 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

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to\; \mathcal{N}_d(0, J(\theta)^{-1}) \;=\; \mathcal{N}_d(0, \mathrm{Var}_\theta(T)^{-1}).$$

To verify these claims, note first that

$$L_n(\theta) \;=\; n\big(\bar{T}_n^\top \theta - \kappa(\theta)\big)$$

with $\bar{T}_n := n^{-1}\sum_{i=1}^n T(X_{ni})$, and thus

$$\begin{aligned} \nabla L_n(\theta) \;&=\; n(\bar{T}_n - \mathbb{E}_\theta(T)), \\ D^2 L_n(\theta) \;&=\; -n\,\mathrm{Var}_\theta(T). \end{aligned}$$

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

$$Z_n \;:=\; \sqrt{n}(\bar{T}_n - \mathbb{E}_{\theta_n}(T)) \;\to_{\mathcal{L}}\; \mathcal{N}_d(0, \mathrm{Var}_\theta(T)).$$

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

$$H_n(\delta) \;=\; Z_n^\top \delta - \frac{1}{2}\,\delta^\top \mathrm{Var}_\theta(T)\delta + \mathrm{Rem}_n(\delta)$$

where

$$\sup_{\delta:\|\delta\|\le C} |\mathrm{Rem}_n(\delta)| \;\to_p\; 0$$

for any fixed $C > 0$. From this one may deduce that with probability tending to one, H_n has a unique maximizer given by

$$\mathrm{Var}_\theta(T)^{-1} Z_n + o_p(1).$$

But this is equivalent to saying that with asymptotic probability 1, the unique maximizer $\widehat{\theta}_n$ of L_n exists and satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;=\; \mathrm{Var}_\theta(T)^{-1} Z_n + o_p(1).$$

In particular,

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to_{\mathcal{L}}\; \mathcal{N}_d\big(0, \mathrm{Var}_\theta(T)^{-1}\big)$$

because $Z_n \to_{\mathcal{L}} \mathcal{N}_d(0, \mathrm{Var}_\theta(T))$.

Example 10.24 (Maximum-likelihood estimation in smoothly parametrized exponential families). Let $\tilde{\mathcal{E}} = (\Omega, \mathcal{A}, (\tilde{P}_\psi)_{\psi\in\Psi})$ be an exponential family with sufficient statistic $T : \Omega \to \mathbb{R}^d$, 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

$$\kappa(\theta) \;:=\; \log \int \exp(\theta^\top T)\, dM.$$

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

$$\sqrt{n}(\widehat{\psi}_n - \psi_n) \;\to_{\mathcal{L}}\; \mathcal{N}_d\big(0, \tilde{J}(\psi)^{-1}\big)$$

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 10.23 that with asymptotic probability 1 there exists a unique maximimum likelihood estimator $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X}_n)$ for the experiment $\mathcal{E} = (\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ with $P_{\tau(\psi)} := \tilde{P}_\psi$ such that

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \;\to_{\mathcal{L}}\; \mathcal{N}_d(0, \operatorname{Var}_\psi(T)^{-1}).$$

But then $\widehat{\psi}_n := \tau^{-1}(\widehat{\theta}_n)$ is a maximum likelihood estimator for $\tilde{\mathcal{E}}$, and elementary calculus reveals that

$$\begin{aligned}
\sqrt{n}(\widehat{\psi}_n - \psi_n) \;&=\; D\tau(\psi)^{-1}\sqrt{n}(\widehat{\theta}_n - \theta_n) + o_p(1) \\
&\to_{\mathcal{L}}\; \mathcal{N}_d\big(0, D\tau(\psi)^{-1} \operatorname{Var}_\psi(T) (D\tau(\psi)^{-1})^\top\big) \\
&=\; \mathcal{N}_d\big(0, \tilde{J}(\psi)^{-1}\big).
\end{aligned}$$

Appendix A

Auxiliary Results

A.1 Some Basics from Measure Theory

For reference in the main text, we collect some basic notions and results in this section; for a thorough account of measure theory, we refer to the monograph of Bauer (2001). Throughout this section let Ω be a nonvoid set, and let $\mathcal{P}(\Omega)$ be the family of all subsets of Ω.

Fields, σ-fields, contents and measures

Definition A.1 (Field and σ-field). A family $\mathcal{A} \subset \mathcal{P}(\Omega)$ is called a field over Ω, if it satisfies the following three conditions:

(F.1) $\emptyset, \Omega \in \mathcal{A}$.
(F.2) If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$.
(F.3) If $A, B \in \mathcal{A}$, then $A \cup B, A \cap B \in \mathcal{A}$.

A family $\mathcal{A} \subset \mathcal{P}(\Omega)$ is called a σ-field over Ω, if it satisfies (F.1-2) and the following condition:

(F.3)$_\sigma$ If $A_1, A_2, A_3, \ldots \in \mathcal{A}$, then $\bigcup_{n\geq 1} A_n, \bigcap_{n\geq 1} A_n \in \mathcal{A}$.

Remark A.2. If condition (F.2) is satisfied, then (F.1) is equivalent to $\Omega \in \mathcal{A}$, and in conditions (F.3) and (F.3)$_\sigma$ it suffices to consider only the unions or only the intersections of sets in $\mathcal{A}$.

Definition A.3 (Content and measure). Let $\mathcal{A}$ be a field over Ω. A mapping $M : \mathcal{A} \to [0, \infty]$ is called a *content*, if it satisfies the following two conditions:

(M.1) $M(\emptyset) = 0$.
(M.2) $M(A \cup B) = M(A) + M(B)$ for disjoint sets $A, B \in \mathcal{A}$.

The mapping M is called a *measure*, if it satisfies (M.1) and the following condition:

(M.2)$_\sigma$ $M\big(\bigcup_{n\geq 1} A_n\big) = \sum_{n\geq 1} M(A_n)$ for pairwise disjoint sets $A_1, A_2, A_3, \ldots$ in $\mathcal{A}$ such that $\bigcup_{n\geq 1} A_n \in \mathcal{A}$.

A content or measure M is called *finite* if $M(\Omega) < \infty$. In case of $M(\Omega) = 1$, it is called *probability content* or *probability measure*, respectively.

Any measure is obviously a content. More interesting is the question under which additional property a content is a measure.

Lemma A.4. *A content M on a field $\mathcal{A}$ is a measure if and only if it satisfies the following continuity property: If $B_1 \subset B_2 \subset B_3 \subset \cdots$ are sets in $\mathcal{A}$ such that $B := \bigcup_{n\geq 1} B_n \in \mathcal{A}$, then $M(B) = \lim_{n\to\infty} M(B_n)$.*

A finite content M on a field $\mathcal{A}$ is a measure if and only if $\lim_{n\to\infty} M(A_n) = 0$ for arbitrary sets $A_1 \supset A_2 \supset A_3 \supset \cdots$ in $\mathcal{A}$ with $\bigcap_{n\geq 1} A_n = \emptyset$.

Here is an important result about the extension of measures:

Theorem A.5 (Carathéodory). *Let M be a finite measure on a field $\mathcal{A}$ over Ω, and let $\mathcal{A}_*$ be the smallest σ-field over Ω containing $\mathcal{A}$. Then there exists a unique measure M_* on $\mathcal{A}_*$ such that $M_*(A) = M(A)$ for all $A \in \mathcal{A}$.*

Measurability and integrals

Measurability. Let $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ be measurable spaces, that means, $\mathcal{A}$ is a σ-field over Ω, and $\mathcal{B}$ is a σ-field over $\mathcal{X}$. A mapping $X : \Omega \to \mathcal{X}$ is called *$\mathcal{A}$-$\mathcal{B}$-measurable* if

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \ \in \ \mathcal{A} \quad \text{for all } B \in \mathcal{B}.$$

Sometimes one abuses notation slightly and talks about a measurable function $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$.

Suppose that $\mathcal{B}$ is generated by some family $\mathcal{E} \subset \mathcal{P}(\mathcal{X})$. That means, $\mathcal{B} = \sigma(\mathcal{E})$ is the smallest σ-field containing $\mathcal{E}$. Then X is $\mathcal{A}$-$\mathcal{B}$-measurable if and only if

$$X^{-1}(E) \ \in \ \mathcal{A} \quad \text{for all } E \in \mathcal{E}.$$

If it is clear from the context which σ-field $\mathcal{B}$ the set $\mathcal{X}$ is equipped with, one talks about a $\mathcal{A}$-measurable mapping $X : \Omega \to \mathcal{X}$ or a measurable mapping $X : (\Omega, \mathcal{A}) \to \mathcal{X}$. In particular, if $\mathcal{X}$ is some interval in $\bar{\mathbb{R}} = [-\infty, \infty]$, then $\mathcal{B}$ is tacitly understood to be the Borel-σ-field, i.e. the smallest σ-field containing all intervals in $\mathcal{X}$.

If $(f_n)_{n\geq 1}$ is a sequence of $\mathcal{A}$-measurable functions $f_n : \Omega \to \bar{\mathbb{R}}$ which converges pointwise to a function $f : \Omega \to \bar{\mathbb{R}}$, then f is $\mathcal{A}$-measurable as well.

Lebesgue integrals. Let M be a measure on a σ-field $\mathcal{A}$ over Ω. For an $\mathcal{A}$-measurable function $f : \Omega \to \bar{\mathbb{R}}$, its integral with respect to M is defined as follows:

Case 1. Let $\mathcal{G}_+$ be the set of all functions f of the form $f = \sum_{i=1}^n \lambda_i 1_{A_i}$ with $n \in \mathbb{N}$, $\lambda_i \in [0, \infty)$ and $A_i \in \mathcal{A}$. For such a function we define

$$\int f \, dM \ := \ \sum_{i=1}^n \lambda_i M(A_i).$$

One can show that this definition does not depend on the particular representation of f.

Case 2. Let $f : \Omega \to [0, \infty]$ be $\mathcal{A}$-measurable. Then

$$\int f\, dM \;:=\; \sup\Big\{ \int g\, dM : g \in \mathcal{G}_+, g \le f \Big\}.$$

Two important properties of this integral are: For $\mathcal{A}$-measurable functions $f, g : \Omega \to [0, \infty]$,

$$\int (af + bg)\, dM \;=\; a \int f\, dM + b \int g\, dM \quad \text{for arbitrary constants } a, b \in [0, \infty],$$

and

$$\int f\, dM \;\le\; \int g\, dM \quad \text{if } f \le g.$$

Case 3. Let $f : \Omega \to \bar{\mathbb{R}}$ be $\mathcal{A}$-measurable. We write $f = f^+ - f^-$ with $f^\pm := \max(\pm f, 0)$, so $|f| = f^+ + f^-$. Then

$$\int f\, dM \;:=\; \int f^+\, dM - \int f^-\, dM,$$

provided that one of the two integrals $\int f^\pm\, dM$ is finite. Otherwise the integral $\int f\, dM$ is not defined. This definition implies that

$$\Big| \int f\, dM \Big| \;\le\; \int |f|\, dM.$$

Moreover, $\int f\, dM$ is well-defined in $\mathbb{R}$ if and only if $\int |f|\, dM$ is finite.

Sometimes it is useful to indicate arguments of the functions to be integrated, so we also write

$$\int f\, dM \;=\; \int f(\omega)\, M(d\omega).$$

In addition, for sets $A \in \mathcal{A}$ one often writes

$$\int_A f\, dM \;:=\; \int 1_A f\, dM,$$

so $\int f\, dM = \int_\Omega f\, dM$.

Monotone convergence. If $(f_n)_{n \ge 1}$ is a sequence of $\mathcal{A}$-measurable functions $f_n : \Omega \to [0, \infty]$, and is $f_n \uparrow f$ pointwise as $n \to \infty$, then

$$\lim_{n \to \infty} \int f_n\, dM \;=\; \int f\, dM.$$

Dominated convergence. Let $(f_n)_{n \ge 1}$ be a sequence of $\mathcal{A}$-measurable functions $f_n : \Omega \to \bar{\mathbb{R}}$ converging pointwise to a function f. Suppose that $|f_n| \le g$ for some $\mathcal{A}$-measurable function $g : \Omega \to [0, \infty]$ such that $\int g\, dM < \infty$. Then all integrals $\int f_n\, dM$ and $\int f\, dM$ are well-defined in $\mathbb{R}$, and

$$\lim_{n \to \infty} \int |f_n - f|\, dM \;=\; 0, \quad \lim_{n \to \infty} \int f_n\, dM \;=\; \int f\, dM.$$

Scheffé's theorem. Let $(f_n)_{n\geq 1}$ be a sequence of $\mathcal{A}$-measurable functions $f_n : \Omega \to \bar{\mathbb{R}}$ converging pointwise to a function f. Suppose that

$$\limsup_{n\to\infty} \int |f_n|\, dM \;\leq\; \int |f|\, dM \;<\; \infty.$$

Then the integrals $\int f_n\, dM$ and $\int f\, dM$ are well-defined in $\mathbb{R}$ for sufficiently large n, and

$$\lim_{n\to\infty} \int |f_n - f|\, dM \;=\; 0, \quad \lim_{n\to\infty} \int f_n\, dM \;=\; \int f\, dM.$$

Dynkin systems

To verify measurability of certain functions or identity of two measures, the following type of set families is very useful.

Definition A.6 (Dynkin system). A family $\mathcal{D}$ of subsets of Ω is called a *Dynkin system over* Ω if the following three conditions are satisfied:

(D.1) $\Omega \in \mathcal{D}$.
(D.2) If $A \in \mathcal{D}$, then $\Omega \setminus A \in \mathcal{D}$.
(D.3) If $A_1, A_2, A_3, \ldots$ are pairwise disjoint sets in $\mathcal{D}$, then $\bigcup_{n\geq 1} A_n \in \mathcal{D}$.

Remark A.7. Conditions (D.1–3) are equivalent to the following three conditions:

(D'.1) $\Omega \in \mathcal{D}$.
(D'.2) If $A, B \in \mathcal{D}$ with $A \subset B$, then $B \setminus A \in \mathcal{D}$;
(D'.3) If $B_1 \subset B_2 \subset B_3 \subset \cdots$ are sets in $\mathcal{D}$, then $\bigcup_{n\geq 1} B_n \in \mathcal{D}$.

Remark A.8. Let $\mathcal{D}$ be a Dynkin system over Ω. Then $\mathcal{D}$ is a σ-field over Ω if and only if it is closed under (finite) intersections, that means, $A \cap B \in \mathcal{D}$ for arbitrary $A, B \in \mathcal{D}$.

Recall also the following well-known fact about Dynkin systems.

Theorem A.9 (Dynkin). *Let $\mathcal{E}$ be an arbitrary family of subsets of Ω. There exists a smallest Dynkin system $\mathcal{D}(\mathcal{E})$ over Ω such that $\mathcal{E} \subset \mathcal{D}(\mathcal{E})$. If $\mathcal{E}$ is closed under (finite) intersections, then the Dynkin system $\mathcal{D}(\mathcal{E})$ coincides with the smallest σ-field $\sigma(\mathcal{E})$ containing $\mathcal{E}$.*

A classical application of Dynkin systems is the following result about uniqueness of measures.

Theorem A.10 (Dynkin). *Suppose that P and Q are probability measures on $(\Omega, \mathcal{A})$, where $\mathcal{A} = \sigma(\mathcal{E})$ for some family $\mathcal{E} \subset \mathcal{P}(\Omega)$. Suppose further that $P \equiv Q$ on $\mathcal{E}$. If $\mathcal{E}$ is closed under (finite) intersections, then $Q \equiv P$ on $\mathcal{A}$.*

Proof of Theorem A.10. One can easily verify that $\mathcal{D} := \big\{A \in \mathcal{A} : P(A) = Q(A)\big\}$ defines a Dynkin system containing $\mathcal{E}$. Thus $\mathcal{D}(\mathcal{E}) \subset \mathcal{D} \subset \mathcal{A}$. But $\cap$-stability of $\mathcal{E}$ implies that $\mathcal{D}(\mathcal{E})$ coincides with $\sigma(\mathcal{E}) = \mathcal{A}$, see Theorem A.9. Hence, $\mathcal{D} = \mathcal{A}$. □

Exercise A.11. Let $\Omega = \{1, 2, 3, 4\}$, and let $\mathcal{E} = \big\{\{1, 2\}, \{1, 4\}\big\}$.

(a) Show that $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$.

(b) Determine $\mathcal{D}(\mathcal{E})$.

(c) Find two different probability measures P, Q on $\mathcal{P}(\Omega)$ such that $P \equiv Q$ on $\mathcal{D}(\mathcal{E})$.

A.2 Two Compactness Properties of Statistical Tests

Let M be a σ-finite measure on a measurable space $(\Omega, \mathcal{A})$. Further, let $\mathcal{F}$ be 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.12 (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.13. The proof of Theorem A.12 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.12. 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.12 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 Theorem 2.29, 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.14 (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.14. As in the proof of Theorem A.12, one can reduce the claim to the case of a probability measure M. 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 $\{\varphi_n \le q\}$, $n \in \mathbb{N}$, $q \in \mathbb{Q}$. 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 $n \ge 1$,

$$-\int f^- \, dM \;\le\; \int \varphi_n f \, dM \;\le\; \int f^+ \, dM,$$

and thus,

$$-\int f^- \, dM \;\le\; L(f) \;\le\; \int f^+ \, dM \quad \text{for all } f \in \mathcal{F}_*. \tag{A.1}$$

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 $\ell \geq 1$,

$$\begin{aligned}\limsup_{k,k'\to\infty}\Big|\int \varphi_{n(k)} f\,dM - \int \varphi_{n(k')} f\,dM\Big| \\ \leq\ 2\|f-f_\ell\| + \limsup_{k,k'\to\infty}\Big|\int \varphi_{n(k)} f_\ell\,dM - \int \varphi_{n(k')} f_\ell\,dM\Big| \ =\ 2\|f-f_\ell\|,\end{aligned}$$

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 $f_o := \mathbb{E}(f \mid \mathcal{A}_o)$ 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 Theorem 2.29 implies the existence of 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.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.15. *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.15. 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

$$u \ \in\ U_o^d \quad \text{with} \quad U_o \ :=\ \{z \in \mathbb{C} : -\epsilon < \operatorname{Re} z < \epsilon\}.$$

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.16. 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

$$\begin{aligned} L \equiv 0 \text{ on } (-\epsilon,\epsilon)^d \quad &\text{whence} \quad L \equiv 0 \text{ on } U_o \times (-\epsilon,\epsilon)^{d-1} \\ &\text{whence} \quad L \equiv 0 \text{ on } U_o^2 \times (-\epsilon,\epsilon)^{d-2} \\ &\ldots \\ &\text{whence} \quad L \equiv 0 \text{ on } U_o^d. \end{aligned}$$

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.16. Let $(\Omega, \mathcal{A}, M)$ be a measure space, $g : (\Omega, \mathcal{A}) \to \mathbb{C}$ and $T : (\Omega, \mathcal{A}) \to \mathbb{R}$ be measurable, and suppose that for real numbers $a < b$,

$$\int |g| \exp(cT)\, dM < \infty \qquad \text{for } c = a, b.$$

Show that

$$f(z) \;:=\; \int g \exp(zT)\, dM$$

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 8.38. 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 $K \subset \{1,\dots,n\}$ 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, $\mathbb{H} = \mathbb{H}_{\{1,\dots,n\}}$. The orthogonal projection of $\mathbb{H}$ onto $\mathbb{H}_K$ is given by Π_K with

$$\Pi_K Y \;:=\; \begin{cases} \mathbb{E}(Y) & \text{if } K=\emptyset \\ \mathbb{E}(Y \,|\, X_K) & \text{else} \end{cases}$$

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 $L = \{1,\dots,n\} \setminus K$, 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

$$\Pi_J \Pi_K \;=\; \Pi_{J\cap K} \quad \text{for arbitrary } J,K \subset \{1,\dots,n\}. \tag{A.2}$$

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 $\Pi^o_K\mathbb{H}$, $K \subset \{1,\dots,n\}$, comprise a decomposition of $\mathbb{H}$ into pairwise orthogonal subspaces.

Theorem A.17. ***(a)*** *For arbitrary sets $K \subset \{1,\dots,n\}$.*

$$\Pi_K \;=\; \sum_{J\subset K} \Pi^o_J.$$

In particular, the identity operator I may be written as

$$I \;=\; \sum_{K\subset\{1,\dots,n\}} \Pi^o_K.$$

(b) *For arbitrary sets $J,K \in \{1,\dots,n\}$,*

$$\Pi^o_J \Pi_K \;=\; \Pi_K \Pi^o_J \;=\; 1_{[J\subset K]} \Pi^o_J,$$

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 $\mathbb{H}^o_K$, $K \subset \{1,\dots,n\}$, are pairwise orthogonal.

Corollary A.18 (Hoeffding's decomposition). *Any random variable $Y \in \mathbb{H}$ can be written as*

$$Y \;=\; \sum_{K\subset\{1,\dots,n\}} \Pi_K^o Y,$$

and the random variables $\Pi_K^o Y$, $K \subset \{1,\dots,n\}$, are uncorrelated with $\Pi_\emptyset^o Y \equiv \mathbb{E}(Y)$ and $\mathbb{E}(\Pi_K^o Y) = 0$ if $K \neq \emptyset$.

For an additional random variable $Z \in \mathbb{H}$,

$$\mathbb{E}(YZ) \;=\; \sum_{K\subset\{1,\dots,n\}} \mathbb{E}(\Pi_K^o Y\, \Pi_K^o Z).$$

This corollary follows essentially from Theorem A.17, except for the statements about $\Pi_\emptyset^o Y$ and $\mathbb{E}(\Pi_K^o Y)$. But $\mathbb{H}_\emptyset^o = \mathbb{H}_\emptyset$ is the space of constants, so $\Pi_\emptyset^o Y = \Pi_\emptyset Y \equiv \mathbb{E}(Y)$, and for any nonempty set $K \subset \{1,\dots,n\}$, it follows from $\mathbb{H}_K^o \perp \mathbb{H}_\emptyset$ that $\mathbb{E}(\Pi_K^o Y) = \langle \Pi_K^o Y, 1\rangle = 0$.

Example A.19 (The case $n = 2$). Suppose that $Y = f(X_1, X_2)$. Then the Hoeffding decomposition of Y reads

$$Y - \mathbb{E}(Y) \;=\; f_1^o(X_1) + f_2^o(X_2) + f_{12}^o(X_1, X_2).$$

The three summands on the right hand side are given by

$$f_1^o(x_1) \;:=\; \mathbb{E}\, f(x_1, X_2) - \mathbb{E}(Y), \quad f_2^o(x_2) \;:=\; \mathbb{E}\, f(X_1, x_2) - \mathbb{E}(Y)$$

and

$$f_{12}^o(x_1, x_2) \;:=\; f(x_1, x_2) - \mathbb{E}\, f(x_1, X_2) - \mathbb{E}\, f(X_1, x_2) + \mathbb{E}(Y).$$

Moreover, for arbitrary random variables $g_1(X_1)$ and $g_2(X_2)$ in $L^2(\mathbb{P})$,

$$\mathbb{E}\big(f_1^o(X_1)g_2(X_2)\big) \;=\; 0 \;=\; \mathbb{E}\big(f_{12}^o(X_1, X_2)g_2(X_2)\big)$$

and

$$\mathbb{E}\big(f_2^o(X_2)g_1(X_1)\big) \;=\; 0 \;=\; \mathbb{E}\big(f_{12}^o(X_1, X_2)g_1(X_1)\big).$$

In particular, the random variables $f_1^o(X_1)$, $f_2^o(X_2)$ and $f_{12}^o(X_1, X_2)$ are centered and uncorrelated.

Proof of Theorem A.17. 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, for

$$\begin{aligned}\sum_{L\subset S} (-1)^{\#L} \;&=\; \sum_{\ell=0}^{\#S} \#\{L \subset S : \#L = \ell\}(-1)^\ell \\ &=\; \sum_{\ell=0}^{\#S} \binom{\#S}{\ell} (-1)^\ell (+1)^{\#S-\ell} \;=\; (1-1)^{\#S} \;=\; 1_{[S=\emptyset]}.\end{aligned}$$

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

$$\begin{aligned}\left.\begin{matrix}\Pi^o_J\Pi_K\\ \Pi_K\Pi^o_J\end{matrix}\right\} \;&=\; \sum_{I\subset J}(-1)^{\#(J\setminus I)}\Pi_{I\cap K}\\ &=\; \sum_{\tilde I\subset J\cap K}\;\sum_{L\subset J\setminus K}(-1)^{\#(J\setminus(\tilde I\cup L))}\Pi_{\tilde I}\\ &=\; \sum_{\tilde I\subset J\cap K}(-1)^{\#(J\setminus\tilde I)}\Big(\sum_{L\subset J\setminus K}(-1)^{-\#L}\Big)\Pi_{\tilde I}\\ &=\; \sum_{\tilde I\subset J\cap K}(-1)^{\#(J\setminus\tilde I)}1_{[J\subset K]}\Pi_{\tilde I} \;=\; 1_{[J\subset K]}\sum_{\tilde I\subset J}(-1)^{\#(J\setminus\tilde I)}\Pi_{\tilde I} \;=\; 1_{[J\subset K]}\Pi^o_J,\end{aligned}$$

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

$$\begin{aligned}\Pi^o_J\Pi^o_K \;&=\; \sum_{I\subset K}(-1)^{\#(K\setminus I)}\Pi^o_J\Pi_I\\ &=\; \sum_{I\subset K}(-1)^{\#(K\setminus I)}1_{[J\subset I]}\Pi^o_J\\ &=\; 1_{[J\subset K]}\Big(\sum_{I\subset K:J\subset I}(-1)^{\#(K\setminus I)}\Big)\Pi^o_J\\ &=\; 1_{[J\subset K]}\Big(\sum_{L\subset K\setminus J}(-1)^{\#L}\Big)\Pi^o_J \;=\; 1_{[J\subset K]}1_{[K\subset J]}\Pi^o_J \;=\; 1_{[J=K]}\Pi^o_J.\end{aligned}$$

It remains to prove part (c). As shown in Exercise A.20, 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^o_K, being a linear combination of self-adjoint operators, is self-adjoint, too. Moreover, it follows from part (b) that $\Pi^o_K\Pi^o_K=\Pi^o_K$, whence Π^o_K describes the orthogonal projection of $\mathbb{H}$ onto some linear subspace $\mathbb{H}^o_K$. The subspaces $\mathbb{H}^o_K$, $K\subset\{1,\dots,n\}$, are pairwise orthogonal, because for different index sets J,K and $Y\in\mathbb{H}^o_J$, $Z\in\mathbb{H}^o_K$,

$$\langle Y,Z\rangle \;=\; \langle\Pi^o_JY,\Pi^o_KZ\rangle \;=\; \langle Y,\Pi^o_J\Pi^o_KZ\rangle \;=\; \langle Y,0\rangle \;=\; 0.$$

Finally, by part (a), for any $K\subset\{1,\dots,n\}$,

$$\mathbb{H}_K \;=\; \sum_{J\subset K}\mathbb{H}^o_J \;=\; \mathbb{H}^o_K+\sum_{J\subsetneq K}\mathbb{H}^o_J,$$

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.20 (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.21. 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.22 (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.23 (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 $\sigma > 0$,*

$$\sum_{i=1}^{n} \mathbb{E}(Y_{ni}) \;\to\; \mu \quad \textit{and} \quad \sum_{i=1}^{n} |\,\mathbb{E}(Y_{ni})| \;=\; O(1),$$
$$\sum_{i=1}^{n} \mathbb{E}(Y_{ni}^2) \;\to\; \sigma^2,$$
$$\sum_{i=1}^{n} \mathbb{E}\big(1_{[Y_{ni}^2>\epsilon]}Y_{ni}^2\big) \;\to\; 0 \quad \textit{for any fixed } \epsilon > 0.$$

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.24. *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 $\sigma > 0$,*

$$\sqrt{n}\,\mathbb{E}(X_{n1}) \;\to\; \mu,$$
$$\mathbb{E}(X_{n1}^2) \;\to\; \sigma^2,$$
$$\mathbb{E}\big(1_{[X_{n1}^2>\epsilon n]}X_{n1}^2\big) \;\to\; 0 \quad \textit{for any fixed } \epsilon > 0.$$

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.22. 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

$$\begin{aligned}
\mathbb{E}\Big|\sum_{i=1}^{n}(Y_{ni}-\mu_{ni})\Big| &\le \mathbb{E}\Big|\sum_{i=1}^{n}(Y_{ni1}-\mu_{ni1})\Big| + \sum_{i=1}^{n}\big(\mathbb{E}\,|Y_{ni2}| + |\mu_{ni2}|\big) \\
&\le \sqrt{\mathrm{Var}\Big(\sum_{i=1}^{n} Y_{ni1}\Big)} + 2\sum_{i=1}^{n}\mathbb{E}\,|Y_{ni2}| \\
&\le \sqrt{\sum_{i=1}^{n}\mathbb{E}(Y_{ni1}^2)} + o(1) \\
&\le \sqrt{\epsilon\sum_{i=1}^{n}\mathbb{E}\,|Y_{ni}|} + o(1) \\
&\le \sqrt{\epsilon M + o(1)} + o(1) \\
&\to \sqrt{\epsilon M}.
\end{aligned}$$

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.23. One can deduce from Theorem A.22, 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:

$$\begin{aligned}
&\mathbb{E}(Z_{ni}) \;=\; 0 \quad \text{for all } n \geq 1 \text{ and } 1 \leq i \leq n, \\
&\sum_{i=1}^{n} \mathbb{E}(Z_{ni}^2) \;\to\; \sigma^2, \\
&\sum_{i=1}^{n} \mathbb{E}\big(1_{[|Z_{ni}|>\epsilon]} Z_{ni}^2\big) \;\to\; 0 \quad \text{for any fixed } \epsilon > 0.
\end{aligned}$$

These conditions imply that

$$\sum_{i=1}^{n} Z_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(0, \sigma^2),$$

and thus

$$\sum_{i=1}^{n} Y_{ni} \;=\; \mu + o(1) + \sum_{i=1}^{n} Z_{ni} \;\to_{\mathcal{L}}\; \mathcal{N}(\mu, \sigma^2).$$

□