Finite sample (exact) estimation theory

Bent Nielsen

University of Oxford

18 October 2023

Comments welcome — mistakes present with probability 1

Slightly revised version of slides written by Anders Kock 2022

Introduction

- Having introduced the measure-theoretic background, we can now begin our study statistical experiments and the ensuing estimation and testing problems.
- We start by studying the classic exact theory of point estimation and hypothesis testing in statistical experiments.
- The beauty of this theory is that it allows the construction of estimators and tests that are optimal irrespective of the sample size.
- The downside is that the theory is largely restricted to exponential families of distributions.
- The theory is thus only exact to the extent that our (restrictive) distributional assumptions are valid.

Although the exponential family type distributions may appear simple/restrictive, they are important initial test beds for inferential procedures:

1. To play around with initial ideas since explicit calculations can be made and much is known about the structure of optimal procedures (as we shall see).
2. To develop examples and counterexamples to ones ideas.
3. In particular results for the normal location model also play an important role in more complicated models where one must resort to asymptotic inference as limits in distribution are often Gaussian.
    - $\to$ When thinking of a new statistical procedure, it is natural to test its performance in the normal location model.

Remarks on history

- The exact theory that we shall start by covering was largely developed in the 1930s to 1960s.
- Highlights are:
    - UMVU estimators via complete sufficient statistics and the Rao-Blackwell-Lehmann-Scheffé theorem; and related estimators that uniformly minimize risk for all convex loss functions.
    - The Neyman-Pearson lemma and most powerful tests at level $\alpha \in (0, 1)$.

Point estimation: Further readings

This slide deck discusses the exact theory of point estimation.

- The slides should be reasonably self-contained.
- They draw on material from:
- Chapters 1–2 in Lehmann and Casella (1998).
- Chapter 7 and the appendix in Liese and Miescke (2008).
- Some (counter)examples are taken from Romano and Siegel (1986).
- Chapter 3 in Pfanzagl (1994).
- Cox and Hinkley (1974) give a less formal exposition.

Contents of this slide set

- Statistical experiments.
- Loss functions and risk.
- Exponential families.
- Sufficient statistics, regular conditional distributions.
- The factorization theorem, Rao-Blackwell theorem.
- Complete statistics.
- UMVU estimators and the Rao-Blackwell-Lehmann-Scheffé theorem.
- Cramér-Rao lower bound (will not cover this).

Statistical experiments

We shall study a *statistical experiment*

$$\mathcal{E} = \left(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\right),$$

where

- $\mathcal{X}$ is the sample space in which the observations take their values.
- $\mathcal{A}$ is a σ-algebra on $\mathcal{X}$.
- $\{P_\theta : \theta \in \Theta\}$ is a family of probability measures on $\mathcal{A}$ indexed by a parameter θ taking values in a parameter space Θ.
- For our purposes, Θ can be thought of as a subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$.
- However, non-parametric model are also encompassed by the above framework.
- Some times an experiment is also called a model, cf. p. 550 in Lehmann and Romano (2005).

- We observe the outcome of a random variable X with values in $\mathcal{X}$.
- X can be a scalar or a vector, depending on what $\mathcal{X}$ is, cf. the examples on the next slide.
- X is defined on an underlying measurable space $(\Omega, \mathcal{F})$ and is $\mathcal{F}$-$\mathcal{A}$-measurable, i.e. $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{A}$.
- For mathematical consistency we assume that there is a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability measures on $\mathcal{F}$ with the property that for all $\theta \in \Theta$ one has

  $$P_\theta = \mathbb{P}_\theta \circ X^{-1},$$

  where

  $$P_\theta(A) = (\mathbb{P}_\theta \circ X^{-1})(A) = \mathbb{P}_\theta(X^{-1}(A)) \qquad A \in \mathcal{A},$$

  that is $P_\theta = \mathbb{P}_\theta \circ X^{-1}$ is the distribution of X under $\mathbb{P}_\theta$, i.e. the image measure.

Examples of experiments

- $\mathcal{X} = \mathbb{R}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})$ and $P_\theta = \mathrm{N}(\theta, 1)$ with $\Theta = [-1, 1]$ corresponds to the family of normal distributions with means in $[-1, 1]$ and variance 1.
- $\mathcal{X} = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ and $P_{n,\theta} = \bigotimes_{i=1}^n \mathrm{N}(\theta, 1)$ with $\Theta = [-1, 1]$ corresponds to observing n independent draws from a normal distributions with mean in $[-1, 1]$ and variance 1. Here $X := (X_1, \ldots, X_n)$ is an n-dimensional vector of observations.
- $\mathcal{X} = \mathbb{N}_0$, $\mathcal{A} = \mathcal{P}(\mathbb{N}_0)$ and $P_\theta = \mathrm{B}(n, \theta)$ with $\Theta = (0, 1)$ corresponds to the family of binomial distributions with trial size n and success probability θ.

Thus, continuous as well as discrete random variables are both covered by the general setup of an experiment. So are all sample sizes $n \in \mathbb{N}$.

Inference problems

- In practice the *parameter* θ in the experiment

$$\mathcal{E} = \big(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$$

  is unknown.

- We shall thus study how one can
  1. estimate θ.
  2. test hypotheses about θ.

- We will pay particular attention to defining plausible notions of *optimality* and constructing optimal procedures.

- There is no consensus about the exact meaning of optimality. We thus study various optimality concepts in the context of estimation and testing.

The (point) estimation problem

The point estimation problem has three ingredients, which the following slides will discuss:

1. The *estimand*, that is quantity to be estimated.
2. The *estimator*, that is our educated guess of the value of the estimand.
3. The *loss function*, that is a measure of closeness of the estimator to the estimand.

The estimand

- Suppose that g is a *known* function of our choosing of the parameter $\theta \in \Theta$ that we are interested in estimating.
- That is, g is a function from Θ to some Θ'.
- Often $g(\theta)$ will simply be the identity function, that is $g(\theta) = \theta$. In this case we can choose $\Theta' = \Theta$ But it is useful to allow for more generality.
- For example, $g(\theta) = \theta^2$ in the context of $P_\theta = \mathrm{N}(\theta, 1)$.
- or if we sample n observations (independently) from a normal distribution with unknown mean μ and variance σ^2 [such that $\theta = (\mu, \sigma^2)$], we may be interested in the Sharpe-ratio $g(\mu, \sigma^2) = \mu/\sigma$. This corresponds to $g(x, y) = x/\sqrt{y}$.
- $g(\theta)$ is also referred to as the *estimand*; the unknown quantity that we wish to estimate.

The estimator

- In order to estimate $g(\theta)$ we introduce an *estimator*.
- An estimator $\delta : \mathcal{X} \to \Theta'$ of the estimand $g(\theta)$ is a (measurable) function of the data to the space Θ' in which g takes its values.
    - For our purposes, Θ' is always a subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$ and we equip it with the corresponding Borel σ-algebra $\mathcal{B}(\Theta')$. Thus, to be precise δ is $\mathcal{A}$-$\mathcal{B}(\Theta')$-measurable.
- For a particular realization $x \in \mathcal{X}$ of X, we call $\delta(x) \in \Theta'$ the *estimate* of $g(\theta)$.

The loss function

- The estimator δ is our educated guess about $g(\theta)$.
- We would like δ to be as close to $g(\theta)$ as possible.
- To measure how good our estimator is, we introduce a *loss function* $L : \Theta' \times \Theta' \to [0, \infty)$.
- The loss of δ for estimating $g(\theta)$ is then given by $L(g(\theta), \delta)$
- We shall assume throughout that

  $$L(x, y) \geq 0 \text{ for all } x, y \in \Theta' \text{ with equality if and only if } x = y,$$
  $$y \mapsto L(x, y) \text{ is } \mathcal{B}(\Theta')\text{-}\mathcal{B}(\mathbb{R})\text{-measurable for all } x \in \Theta'.$$

- Typically, $y \mapsto L(x, y)$ is continuous for all $x \in \Theta'$ and hence, a fortiori, $\mathcal{B}(\Theta')$-$\mathcal{B}(\mathbb{R})$-measurable.
- In words, we obtain a loss of zero only if the estimator δ equals the estimand $g(\theta)$.
- The measurability is imposed such that we can integrate the loss function (find its expected value).

Examples of loss functions for estimation problems

- The most commonly used loss function is probably the *quadratic loss*
$$L(x, y) = (x - y)^2.$$
That is, deviations from the truth are "penalized" through the square of the deviation.
- Larger deviations are penalized at an increasing rate.
- The absolute value loss $L(x, y) = |x - y|$ penalizes deviations from the truth at a constant rate.
- The choice of loss function is in general up to the practitioner.
- We shall see that we can construct estimators that are optimal in a suitable sense simultaneously for *all* loss functions satisfying that $y \mapsto L(x, y)$ is convex for all $x \in \Theta'$.
- Note that $L(x, y) = |x - y|^p$ for $p \geq 1$ satisfies this convexity. For $p = 1, 2$ we get the above two loss functions as special cases.

Risk

- The loss of δ of course depends on the particular realization of X since $\delta(x)$ generally varies with x.
- To eliminate this dependence, we study the expected/average loss as a function of θ.
- We call this average loss the *risk* of δ:

$$\begin{aligned} R(\theta,\delta) := \mathbb{E}_\theta L(g(\theta),\delta(X)) &= E_\theta L(g(\theta),\delta(x)) \\ &= \int_{\mathcal{X}} L(g(\theta),\delta(x)) P_\theta(dx), \end{aligned}$$

  where $\mathbb{E}_\theta$ and E_θ denote the expectations corresponding to the measures $\mathbb{P}_\theta$ (on $\mathcal{F}$) and P_θ (on $\mathcal{A}$), respectively.
  [Note also how we used the substitution rule]
- As $L \geq 0$, it is clear that $x \mapsto L(g(\theta),\delta(x)) \in \mathcal{L}(P_\theta)$ and the above integrals are well-defined.
- We now study the risk function $\theta \mapsto R(\theta,\delta)$.

- Ideally we would like our estimator δ to minimize $R(\theta, \delta)$ for *all* $\theta \in \Theta$.
- Such an estimator would be worthy of the name "best".
- However, such an estimator typically does not exist.
- In general, the estimator minimizing $R(\theta, \delta)$ depends on θ.
- Fix $\theta_0 \in \Theta$. Then $\delta \equiv g(\theta_0)$ clearly satisfies $R(\theta_0, \delta) = 0$.
- Thus, an estimator minimizing the risk for all $\theta \in \Theta$ would need to have zero risk for all θ.
- This is certainly not generally the case for the above constant estimator (not depending on the data!).
- Indeed, $\delta \equiv g(\theta_0)$ generally performs poorly for $\theta \neq \theta_0$.
  - Example: Let $\theta \in \Theta = \mathbb{R}$ with $g(\theta) = \theta$ and $L(x, y) = (x - y)^2$. Then $R(\theta, \delta) = (\theta - \theta_0)^2$. [1]

[1] It is typically not possible to have a risk of zero for all $\theta \in \Theta$. A special case where this *is* possible is the case with no variation: Let δ_θ be the Dirac measure on θ. Let $\mathcal{E} = \big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\delta_\theta : \theta \in \mathbb{R}\}\big)$. Then $\delta(X) = g(X)$ satisfies $L(g(\theta), \delta(X)) = 0$ δ_θ-a.s. for all $\theta \in \mathbb{R}$, irrespectively of g.]

In general no estimator exists that simultaneously minimizes $\theta \mapsto R(\theta, \delta)$ for all $\theta \in \Theta$.

Theorem 1 (No estimator minimizes risk at all $\theta \in \Theta$)

Consider the experiment $\mathcal{E} = \big(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$ in which the P_θ are mutually absolutely continuous [2]*, Θ has at least two elements and $g(\theta) = \theta$. Then for no loss function L does there exist an estimator $\delta : \mathcal{X} \to \Theta$ that minimizes $\theta \mapsto R(\theta, \delta)$ for all $\theta \in \Theta$.*

- Observe that in particular the family $\{\mathrm{N}(\theta, 1) : \theta \in \mathbb{R}\}$ consists of mutually absolutely continuous probability measures.[3]
- Thus, a best estimator in the sense minimizing risk at all $\theta \in \mathbb{R}$ does not exist. Not even when estimating the mean of a normal distribution.
- Recall that for L to be a loss function we impose that $L(x, y) \geq 0$ with equality if and only if $x = y$.
    - If we allowed for constant loss functions then of course every estimator would trivially minimize risk at all $\theta \in \Theta$.

[2] So that the measures P_θ have common null sets

[3] More generally, we shall see that distributions forming an exponential family are mutually absolutely continuous.

Proof

If an estimator δ minimizes risk at $\theta_0 \in \Theta$, then

$$0 = R(\theta_0, \delta) = E_{\theta_0} L(\theta_0, \delta),$$

which implies that $L(\theta_0, \delta(x)) = 0$ for P_{θ_0}-a.e. $x \in \mathcal{X}$. Thus, $\delta(x) = \theta_0$ for P_{θ_0}-a.e. $x \in \mathcal{X}$ since $L(\theta_0, \delta(x)) = 0$ if and only if $\delta(x) = \theta_0$ [by definition of loss function].

Let $\theta_1 \neq \theta_0$. Then, as $P_{\theta_1} \ll P_{\theta_0}$, it follows that $\delta(x) = \theta_0$ for P_{θ_1}-a.e. $x \in \mathcal{X}$. Hence,

$$R(\theta_1, \delta) = E_{\theta_1} L(\theta_1, \delta) = E_{\theta_1} L(\theta_1, \theta_0) > 0.$$

Hence, δ does not minimize $\theta \mapsto R(\theta, \delta)$ at θ_1 [as we know that the minimal value is 0, and the second equality used that the integral does not care about null sets.]

□

Quadratic risk

- Quadratic loss is $L(g(\theta), \delta(X)) = (\delta(X) - g(\theta))^2$.
- Quadratic risk, $R(\theta, \delta) = E_\theta(\delta(X) - g(\theta))^2$ is also called Minimum Squared Error (MSE). It satisfies

$$\begin{aligned} R(\theta, \delta) &= E_\theta((\delta(X) - E_\theta\delta(X)) - (E_\theta\delta(X) - g(\theta))^2 \\ &= E_\theta((\delta(X) - E_\theta\delta(X))^2 + (E_\theta\delta(X) - g(\theta))^2 \\ &= Var_\theta(\delta(X)) + (E_\theta\delta(X) - g(\theta))^2, \end{aligned}$$

that is the variance plus the squared bias of the estimator.

Example: Quadratic risk in $N(\mu, \sigma^2)$.

- Consider a sample $X_1, \ldots, X_n$ with $n \geq 2$ from $N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ unknown.
- Common estimators for σ^2 are $s^2/(n-1)$ and s^2/n, where $s^2 = \sum_{i=1}^n (x_i - \bar{x}_i)^2$. These are the residual variance and the maximum likelihood estimator, respectively.
- Consider the class of estimators cs^2. The quadratic risk is[4]

$$\begin{aligned} E(cs^2 - \sigma^2)^2 &= Var(cs^2) + (E(cs^2) - \sigma^2)^2 \\ &= \sigma^4((n^2-1)c^2 - 2(n-1)c + 1). \end{aligned}$$

Quadratic risk is minimized for $c = 1/(n+1)$. We find

$$R(\frac{1}{n-1}s^2, \sigma^2) > R(\frac{1}{n}s^2, \sigma^2) > R(\frac{1}{n+1}s^2, \sigma^2).$$

[4] Use that s^2 is $\sigma^2\chi^2_{n-1}$ distributed with $Var(s^2) = 2(n-1)\sigma^4$ and $E(s^2) = (n-1)\sigma^2$.

Unbiased estimators

- We have seen that it is generally too much to ask for a "best" estimator in the sense of minimizing $\theta \mapsto R(\theta, \delta)$ for all $\theta \in \Theta$.
- One way of ensuring that a "best" estimator can exist, is to restrict the class of estimators.
- A reasonable requirement of an estimator δ is that it is *unbiased* for $g(\theta)$, that is

$$\mathbb{E}_\theta \delta(X) = g(\theta) \quad \text{for all } \theta \in \Theta.$$

- Note that the constant estimator $\delta = g(\theta_0)$ from above is clearly not unbiased (unless g is constant).
- Unbiasedness can, of course, be equivalently expressed as

$$E_\theta \delta = \int_{\mathcal{X}} \delta(x) P_\theta(dx) = \int_\Omega \delta(X(\omega)) \mathbb{P}_\theta(d\omega) = g(\theta) \text{ for all } \theta \in \Theta.$$

- It turns out that for loss functions that are convex in the second argument (i.e. in the estimator), the restriction to unbiased estimators allows us to construct estimators in a (relatively) large class of experiments that *uniformly* over Θ minimize the risk.
- While the concrete estimators minimizing risk may differ across experiments, they have a common structure.
- In particular, they are functions of *complete sufficient* statistics.
- We shall now review the exponential family of distributions, a family within which unbiased estimators uniformly minimizing convex risk exist [as complete sufficient statistics are easily found].

Exponential families

A family of distributions (i.e. probability measures) $\{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{A})$ is be an m-dimensional *exponential family* if it has a density

$$\frac{dP_\theta}{d\mu}(x) = p_\theta(x) = C(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x)\right] h(x),$$

with respect to a σ-finite dominating measure μ on $\mathcal{A}$
and $T_i : \mathcal{X} \to \mathbb{R}$, $\eta_i : \Theta \to \mathbb{R}$ for $i = 1, \ldots, m$ and $h : \mathcal{X} \to [0, \infty)$ (appropriately Borel measurable).

We can equivalently write that the density can be written as

$$p_\theta(x) = \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x).$$

Clearly, $C(\theta) = \exp(-B(\theta))$.
The latter formulation is more convenient in applying the Neyman factorization theorem (to come).

Examples

Many distributions used in practice are of exponential family type.

- Normal distribution with unknown mean (and variance).
- Bernoulli distribution with unknown success probability.
- Binomial distribution with known *n* but unknown success probability.
- Poisson distribution with unknown density.
- Exponential distribution.
- Gamma distribution.
- Beta distribution.
- See Table 5.1 in Lehmann and Casella (1998) or even Wikipedia for more examples and further details.

Thus, the theory on minimum variance unbiased estimation that we will develop will apply in particular to all these distributions as it applies to exponential families.

Details for N(μ, σ^2) — a continuous example

The distribution N(μ, σ^2), $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ has density

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \theta = (\mu, \sigma^2),$$

with respect to the Lebesgue measure λ_1 on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Clearly,

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2+\mu^2-2\mu x}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\right), \end{aligned}$$

such that for μ and σ^2 unknown we have $m = 2$ with $\eta_1(\theta) = \frac{\mu}{\sigma^2}$, $\eta_2(\theta) = -\frac{1}{2\sigma^2}$, $T_1(x) = x$, $T_2(x) = x^2$, $B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{\log(\sigma^2)}{2}$ and $h(x) = \frac{1}{\sqrt{2\pi}}$.

Details for Poi(λ) — a discrete example

The Poisson distribution has density

$$p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad \lambda \in (0, \infty)$$

with respect to the counting measure τ on $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$, cf. the exercises. Clearly,

$$p_\lambda(x) = \exp\left(\log(\lambda)x - \lambda\right)\frac{1}{x!},$$

such that $m = 1$, $\eta_1(\lambda) = \log(\lambda)$, $T_1(x) = x$, $B(\lambda) = \lambda$ and $h(x) = \frac{1}{x!}$.

Canonical form and natural parameters

If we use the $\eta_i(\theta)$ in

$$p_\theta(x) = C(\theta)\exp\left[\sum_{i=1}^{m}\eta_i(\theta)T_i(x)\right]h(x)$$

as the parameters, we arrive at the *canonical form*

$$p_\theta(x) = C(\eta)\exp\left[\sum_{i=1}^{m}\eta_i T_i(x)\right]h(x).$$

We call the $\eta_i,\ i = 1,\dots,m$ the *natural parameters*.

Thus, in the above example of the normal distribution, the natural parameters are $\eta_1(\theta) = \frac{\mu}{\sigma^2}$, $\eta_2(\theta) = -\frac{1}{2\sigma^2}$. In the Poisson example the natural parameter is $\log(\lambda)$.
We call the $T_i(x)$, $i = 1,\dots,m$ the *canonical statistics*.

Natural parameter space

The *natural parameter space* is the set

$$A := \left\{ \eta \in \mathbb{R}^m : \int_{\mathcal{X}} \exp\left[\sum_{i=1}^{m} \eta_i T_i(x)\right] h(x)\mu(dx) < \infty \right\}.$$

The natural parameter space is convex, cf. exercises.
Note that for all $\eta \in A$, we have that

$$C(\eta) = 1 / \int_{\mathcal{X}} \exp\left[\sum_{i=1}^{m} \eta_i T_i(x)\right] h(x)\mu(dx)$$

since

$$1 = \int_{\mathcal{X}} p_\theta(x)\mu(dx) = C(\eta) \int_{\mathcal{X}} \exp\left[\sum_{i=1}^{m} \eta_i T_i(x)\right] h(x)\mu(dx).$$

- Observe that if $\{P_\theta : \theta \in \Theta\}$ is an exponential family on $(\mathcal{X}, \mathcal{A})$, then so is $\{P_\theta^n : \theta \in \Theta\}$ on $(\mathcal{X}^n, \mathcal{A}^n)$, where P_θ^n is the n-fold product of P_θ.
- That is, P_θ^n is the distribution of an i.i.d. sample of size n from P_θ.
- To verify this claim, note that

$$\begin{aligned}
\frac{dP_\theta^n}{d\mu^n} &= \prod_{j=1}^{n} C(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x_j)\right] h(x_j) \\
&= [C(\theta)]^n \exp\left[\sum_{i=1}^{m} \eta_i(\theta) \sum_{j=1}^{n} T_i(x_j)\right] \prod_{j=1}^{n} h(x_j) \\
&= C_n(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i^{(n)}(x_1, \ldots, x_n)\right] h_n(x_1, \ldots, x_n)
\end{aligned}$$

where $T_i^{(n)}(x_1, \ldots, x_n) = \sum_{j=1}^n T_i(x_j)$, $C_n(\theta) = [C(\theta)]^n$ and $h_n(x_1, \ldots, x_n) = \prod_{j=1}^n h(x_j)$.

Exponential families are mutually absolutely continuous

- Let us mention the fact that members of exponential families are mutually absolutely continuous.
- That is, if $\{P_\theta : \theta \in \Theta\}$ forms an exponential family, then

$$P_\theta \ll\gg P_{\theta'} \qquad \text{for all } \theta, \theta' \in \Theta.$$

Thus, whether a set is a P_θ-null set does not depend on θ.

To see this, note that for all $A \in \mathcal{A}$

$$P_\theta(A) = \int_A C(\theta) \exp\left[\sum_{i=1}^m \eta_i(\theta) T_i(x)\right] h(x)\mu(dx) = 0$$

implies $\mathbb{1}_A(x)h(x) = 0$ for μ-a.e. $x \in \mathcal{X}$. But then also

$$P_{\theta'}(A) = \int_A C(\theta) \exp\left[\sum_{i=1}^m \eta_i(\theta') T_i(x)\right] h(x)\mu(dx) = 0.$$

Uniform distribution — not an exponential family

- The continuous uniform distribution $\mathrm{U}(0,\theta)$ on $[0,\theta], \theta > 0$ has density $p_\theta(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$.
- For $0 < \theta_1 < \theta_2$ it is clear that $[\theta_1,\theta_2]$ is a $\mathrm{U}(0,\theta_1)$-null set.
- But it is *not* a $\mathrm{U}(0,\theta_2)$-null set.
- Hence, it is *false* to say $\mathrm{U}(0,\theta_1) \ll\gg \mathrm{U}(0,\theta_2)$,
- so that the statistical experiment that includes $\mathrm{U}(0,\theta_1)$ and $\mathrm{U}(0,\theta_2)$ does not form an exponential family.
- In any case, $\mathrm{U}(0,\theta_1)$ is not of exponential form since the support depends on the parameter.

Sufficient statistics

- One reason why exponential families of distribution are pleasant to work with is that they give immediate access to *sufficient statistics*.
- Sufficient statistics serve as "data reduction" devices as the sufficient statistics contain "all information" about the parameter θ.
- Furthermore, optimal estimators (and tests) are often functions of sufficient statistics.
- Thus, one does not have to search over all functions of the data to find the optimal one — it suffices to consider functions of the sufficient statistics.

To properly define sufficient statistics, we must introduce conditional distributions.

Stochastic kernels and conditional distributions

- Let us briefly abstract from the exact estimation setting studied so far.
- The following treatment and introduction of conditional distributions is taken from page 625 in Liese and Miescke (2008).

Stochastic kernels

Definition 2 (Stochastic kernels)

For two measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ a mapping $\mathsf{K} : \mathcal{B} \times \mathcal{X} \to [0, 1]$ is called a *stochastic kernel* if for every $B \in \mathcal{B}$ the function $x \mapsto \mathsf{K}(B, x)$ is $\mathcal{A}$-$\mathcal{B}([0, 1])$-measurable, and for every $x \in \mathcal{X}$, it holds that $\mathsf{K}(\cdot, \mathsf{x})$ is a probability measure on $\mathcal{B}$.

- Stochastic kernels are also referred to as Markov kernels, conditioning kernels, or simply kernels.
- We shall now see that stochastic kernels are intimately linked to conditional distributions.

Conditional distribution

Definition 3 (Conditional distribution)

Let X and Y be random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, respectively[5]. The kernel $\mathsf{K} : \mathcal{B} \times \mathcal{X} \to [0, 1]$ is called a regular conditional distribution of Y given X if

$$\mathbb{P}(X \in A, Y \in B) = \int_A \mathsf{K}(B, x) P_X(dx) \quad \text{for all} \quad A \in \mathcal{A},\ B \in \mathcal{B},$$

where $P_X = \mathbb{P} \circ X^{-1}$ is the distribution of X under $\mathbb{P}$.

We can think of $\mathsf{K}(B, x)$ as the probability of Y falling in the set B, *given/conditional on* $X = x$.

We often write $P_{Y|X}$ for the conditional distribution K.

Observe also that

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P} \circ (X, Y)^{-1}(A \times B) = P_{X,Y}(A \times B).$$

[5] Both of $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ equal $(\mathbb{R}, \mathcal{B}(R))$

Some remarks

- Observe that

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in \mathcal{X}, Y \in B) = \int_{\mathcal{X}} \mathsf{K}(B,x) P_X(dx).$$

- Thus, in accordance with intuition, the *unconditional* probability of $Y \in B$ can be found by averaging the conditional probability of $Y \in B$ (that is $\mathsf{K}(B,x)$) over all values of $x \in \mathcal{X}$.
- Note also that if $K(B,x) = \mu(B)$ for all $x \in \mathcal{X}$ and a measure μ on $\mathcal{B}$, then X and Y are independent and $P_{X,Y}(A \times B) = P_X(A) \cdot \mu(B)$ and

$$P_Y(B) = \mathbb{P}(Y \in B) = \int_{\mathcal{X}} \mathsf{K}(B,x) P_X(dx) = \mu(B).$$

"If the conditional distribution of Y given X does not depend on x then Y and X are independent".

Existence of conditional distribution

- One may ask: when does a conditional distribution of Y given X exist?
- As we will more or less exclusively be dealing with random vectors (that take value in $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)))$, the conditional distribution of Y given X will always exist, cf. page 625 in Liese and Miescke (2008), Theorem A.37.
- In fact it suffices that $\mathcal{Y}$ is a so-called complete separable metric space equipped with the corresponding Borel σ-algebra.

Finding a conditional distribution via densities

In case the random variables under consideration have a joint density, it is easy to find a conditional distribution:

- Suppose that X and Y are random variables with values in $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, respectively.
- Assume that there are σ-finite measures μ and ν on $\mathcal{A}$ and $\mathcal{B}$, respectively such that $P_{X,Y} = \mathbb{P} \circ (X, Y)^{-1} \ll \mu \otimes \nu$.
- Set $f_{X,Y} = \frac{dP_{X,Y}}{d\mu\otimes\nu}$.
- Let $\mu(A) = 0$. Then

  $$P_X(A) = P_{X,Y}(A \times \mathcal{Y}) = 0,$$

  since $(\mu \otimes \nu)(A \times \mathcal{Y}) = \mu(A) \cdot \nu(\mathcal{Y}) = 0$ and $P_{X,Y} \ll \mu \otimes \nu$.
- Thus, $P_X \ll \mu$ and similarly $P_Y \ll \nu$ when $P_{X,Y} \ll \mu \otimes \nu$.

- Define
$$f_X(x) := \frac{dP_X}{d\mu}(x) = \int f_{X,Y}(x,y)\nu(dy)$$
and
$$f_Y(y) := \frac{dP_Y}{d\nu}(y) = \int f_{X,Y}(x,y)\mu(dx)$$
which are called the *marginal densities*.

Observe that since $P_X \ll \mu$ and $P_Y \ll \nu$ (by the previous slide) $\frac{dP_X}{d\mu}$ and $\frac{dP_Y}{d\nu}$ exist by the Radon Nikodym Theorem.

You will show the two equalities (that are not definitions) in the exercises.

Definition 4 (Conditional distribution via densities)

The function

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{if } f_X(x) > 0 \\ f_Y(y) & \text{if } f_X(x) = 0 \end{cases}$$

is called the conditional density of Y, given $X = x$. The stochastic kernel $K(B,x) = \int_B f_{Y|X}(y|x)\nu(dy)$ is called the regular conditional distribution of Y, given $X = x$ based on the conditional density.

Since many common distributions, such as the ones in the exponential family, are given by their densities we can thus "easily" find conditional distributions for these.

Sufficient statistic: Definition

- Back to sufficient statistics!
- We are now in a position where we can formally define sufficient statistics in the context of the experiment

$$\mathcal{E} = \left(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\right).$$

Sufficient statistic: Definition

- Let $T : \mathcal{X} \to \mathcal{T}$ for some measurable space $(\mathcal{T}, \mathscr{T})$ (that is hopefully of lower dimension than $\mathcal{X}$).
- Denote by $P_{T,\theta} = P_\theta \circ T^{-1} = \mathbb{P}_\theta \circ (T \circ X)^{-1}$ the distribution of T under P_θ.

Definition 5 (Sufficient Statistic)

The statistic $T : \mathcal{X} \to \mathcal{T}$, $\mathcal{A}$-$\mathscr{T}$-measurable, is *sufficient* for θ if the conditional distribution of X given T does not depend on $\theta \in \Theta$, that is there exists a kernel $\mathsf{P}_{X|T} : \mathcal{A} \times \mathcal{T} \to [0, 1]$ not depending on θ such that, for all $A \in \mathcal{A}$, $C \in \mathscr{T}$,

$$\begin{aligned}\mathbb{P}_\theta(T \in C, X \in A) &= \int_C \mathsf{P}_{X|T}(A, t) P_{T,\theta}(dt) \\ &= \int_{T^{-1}(C)} \mathsf{P}_{X|T}(A, T(x)) P_\theta(dx).\end{aligned}$$

One also says that T is sufficient for the family $\{P_\theta : \theta \in \Theta\}$ or X.

The factorization theorem

- The definition of sufficiency is useful for theoretical purposes.
- However, it is not useful for finding sufficient statistics as it requires one to have a candidate T and then verify that $\mathrm{P}_{X|T}(\cdot, t)$ does not depend on θ.
- The factorization theorem gives a very practical criterion if the family $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ.

Theorem 6 (Neyman factorization theorem)

If the distributions P_θ in $\{P_\theta : \theta \in \Theta\}$ have densities $p_\theta = dP_\theta/d\mu$ with respect to a σ-finite dominating measure μ, then $T : \mathcal{X} \to \mathcal{T}$, $\mathcal{A}$-$\mathscr{T}$-measurable, is sufficient for θ if and only if there exist $\mathscr{T}$-$\mathcal{B}(\mathbb{R})$-measurable functions $g_\theta : \mathcal{T} \to [0, \infty)$, $\theta \in \Theta$ and a $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable function $h : \mathcal{X} \to [0, \infty)$ such that

$$p_\theta(x) = g_\theta(T(x))h(x) \quad \text{for } \mu\text{-almost every } x \in \mathcal{X}.$$

- A proof of the factorization theorem can be found in Section 2.6 in Lehmann and Romano (2005).

Why is it called sufficiency?

- Since $p_\theta(x) = g_\theta(T(x))h(x)$, the maximum likelihood estimator of θ maximizes

  $$\theta \mapsto g_\theta(T(x))h(x)$$

  over $\theta \in \Theta$.
- Thus, the MLE can be expressed as a function of the sufficient statistic $T(x)$.
- Two data sets that give the same value of the sufficient statistic $T(x)$ give the same value of the MLE.

Furthermore, we shall see that optimal estimators and tests are functions of sufficient statistics. Thus, in this sense, it is *sufficient* to consider statistics that are functions of these.

Examples: applying the factorization theorem

Sufficient statistics in exponential families:

- Recall that an exponential family has densities of the form

$$p_\theta(x) = \exp\left[\sum_{i=1}^{m} \eta_i(\theta)\, T_i(x) - B(\theta)\right] h(x).$$

- It thus follows immediately from the factorization theorem that $(T_1(x), \ldots, T_m(x))$ is sufficient for θ.
- Indeed, one of the attractions of exponential families is this easy access to sufficient statistics.

i.i.d. samples

- In a sample of n i.i.d observations from an exponential family it also follows that $(T_1^{(n)}, \ldots, T_m^{(n)})$ is sufficient for $\{P_\theta^n : \theta \in \Theta\}$ where we recall $T_i^{(n)}(x_1, \ldots, x_n) = \sum_{j=1}^n T_i(x_j)$.
- Thus, we easily get sufficient statistics for i.i.d. samples from exponential families.
- Observe that the dimension of the sufficient statistic is m irrespectively of n.
- This independence of m on n is particular to the exponential family of distributions.
- Indeed, if the dimension of the sufficient statistic does not depend on n and the support of the family of distributions does not depend on θ, then the family of distributions must be of exponential family type. This is known as the Pitman-Koopman-Darmois Theorem.

Sufficient statistics for the normal distribution

- For $P_\theta = \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ unknown we have already seen that $T_1(x) = x$ and $T_2(x) = x^2$.
- Hence, these are the sufficient statistics for $\theta = (\mu, \sigma^2)$.
- By the observation on the previous slide, $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is sufficient for μ and σ^2 in an i.i.d. sample of size n from $\mathrm{N}(\mu, \sigma^2)$.

Sufficient statistics for the Poisson distribution

- For $P_\lambda = \text{Poi}(\lambda)$ we have already seen that $T_1(x) = x$.
- Hence, x is sufficient for λ.
- In a sample of size n of i.i.d. observations from $\text{Poi}(\lambda)$, we thus have that $\sum_{i=1}^n x_i$ is sufficient for λ.

Exponential family vs. factorization theorem

- The factorization theorem characterizes sufficient statistics $T(x)$ as those satisfying

$$p_\theta(x) = g_\theta(T(x))h(x) \quad \text{for } \mu\text{-almost every } x \in \mathcal{X}.$$

- Exponential families impose more structure on the density p_θ, namely

$$p_\theta(x) = \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x),$$

  thus, it essentially requires that $g_\theta(T(x))$ is of the form $\exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x) - B(\theta)\right]$ in the factorization theorem. [Note that $T(x)$ can be of the form $(T_1(x), \ldots, T_m(x))$].

- Thus, there can be distributions that are not of exponential family type, but which do possess a sufficient statistic. See next slide.

Uniform distribution

- The continuous uniform distribution $\mathrm{U}(0,\theta)$ on $[0,\theta], \theta > 0$ has density $p_\theta(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$.
- We have seen that this does *not* form an exponential family.
- However, for any independent sample of size $n \in \mathbb{N}$ a sufficient statistic still exists:

  $$\prod_{i=1}^{n} p_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]}(x_{(n)}) \prod_{i=1}^{n} \mathbb{1}_{[0,\infty)}(x_i),$$

  where $x_{(n)} := \max_{1\leq i\leq n} x_i$.
- Thus, writing $\boldsymbol{x}_n = (x_1,\ldots,x_n)$, the sufficient statistic is $T(\boldsymbol{x}_n) = x_{(n)}$ as one can choose in the factorization theorem

  $$g_{n,\theta}(T(\boldsymbol{x}_n)) = \frac{1}{\theta^n}\mathbb{1}_{(-\infty,\theta]}(x_{(n)}) \quad \text{and} \quad h_n(\boldsymbol{x}_n) = \prod_{i=1}^{n} \mathbb{1}_{[0,\infty)}(x_i).$$

- Thus, $\max_{1\leq i\leq n} x_i$ is sufficient for estimating the upper bound of the support of $\mathrm{U}(0,\theta)$.

Randomized estimators

Consider the experiment $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$ in which we wish to estimate $g(\theta)$ where $g : \Theta \to \Theta'$ and Θ' is equipped with a σ-algebra $\mathcal{S}$. A *randomized estimator* is a stochastic kernel $\mathrm{D} : \mathcal{S} \times \mathcal{X} \to [0, 1]$.

D is called *randomized* because for every $x \in \mathcal{X}$ it is a probability measure on $(\Theta', \mathcal{S})$.

Intuitively, upon observing $x \in \mathcal{X}$, we make a draw from the distribution $\mathrm{D}(\cdot, x)$, i.e. we select a point in Θ' at random.

Classic, non-randomized, estimators are in a sense special cases of randomized estimators.

Setting $\mathrm{D}(\cdot, x) = \delta_{d(x)}(\cdot)$, the Dirac measure, for $d : \mathcal{X} \to \Theta'$ then for every $x \in \mathcal{X}$ one has that $\mathrm{D}(\cdot, x)$ is concentrated at $d(x)$ and is in this sense not randomized. In this case we typically write $d(x)$ rather than $\mathrm{D}(\cdot, x)$.

The risk of the randomized estimator D is defined as

$$R(\theta, \mathrm{D}) = \int \left[\int L(g(\theta), a) \mathrm{D}(da, x) \right] P_\theta(dx), \qquad \theta \in \Theta.$$

Observe that if the estimator is not randomized, that is $\mathrm{D}(\cdot, x) = \delta_{d(x)}(\cdot)$ for some $d : \mathcal{X} \to \Theta'$, then

$$\begin{aligned} R(\theta, \mathrm{D}) &= \int \left[\int L(g(\theta), a) \mathrm{D}(da, x) \right] P_\theta(dx) \\ &= \int \left[\int L(g(\theta), a) \delta_{d(x)}(da) \right] P_\theta(dx) \\ &= \int L(g(\theta), d(x)) P_\theta(dx), \end{aligned}$$

which is nothing else than our “old” definition of risk for an estimator $d : \mathcal{X} \to \Theta'$.

First consequences of sufficiency for the estimation problem

- Let δ be any non-randomized estimator of $g(\theta)$ and note that by disintegration, cf. Liese and Miescke (2008) Lemma A.40+A.41.

$$\begin{aligned} R(\theta,\delta) &= \int_{\mathcal{X}} L(g(\theta),\delta(x))P_\theta(dx) \\ &= \int_{\mathcal{T}}\int_{\mathcal{X}} L(g(\theta),\delta(x))P_{X|T}(dx,t)P_{T,\theta}(dt) \\ &= \int_{\mathcal{T}}\int_{\Theta'} L(g(\theta),u)(P_{X|T}\circ\delta^{-1})(du,t)P_{T,\theta}(dt) \\ &= \int_{\mathcal{X}}\int_{\Theta'} L(g(\theta),u)(P_{X|T}\circ\delta^{-1})(du,T(x))P_\theta(dx) \end{aligned}$$

  where $(P_{X|T}\circ\delta^{-1})(\cdot,t)$ for every $t\in\mathcal{T}$ is the distribution of δ under $P_{X|T}(\cdot,t)$ on Θ'.

- Thus, if T is sufficient for θ, then for *any* estimator δ of the estimand $g(\theta)$ the *randomized estimator* $(P_{X|T}\circ\delta^{-1})(\cdot,t)$ based on observing only t has the same risk as δ.

- It is a randomized estimator because upon observing t, we obtain our estimate by making a draw from $(P_{X|T}\circ\delta^{-1})(\cdot,t)$.

- The previous slide shows that the risk of any non-randomized estimator δ of $g(\theta)$ can be replicated by an estimator that depends on x only via $T(x)$; potentially after randomization.
- One can also show that the loss of any randomized estimator D can replicated by a potentially randomized estimator that only depends on x via $T(x)$, cf. Theorem 4.66 in Liese and Miescke (2008).
- Thus, we need only consider (potentially randomized) estimators that are functions of $T(x)$.
- This is true irrespective of the loss function used.
- If the loss function is *convex* more can be said (next slide).

Let $y \mapsto L(x, y)$ be convex for every $x \in \Theta'$.

Assume that $\int ||a||\mathrm{D}(da, x)P_\theta(dx) < \infty$. Then, by Jensen's inequality, the risk of *any* randomized estimator satisfies

$$R(\theta, \mathrm{D}) = \int \left[\int L(g(\theta), a)\mathrm{D}(da, x) \right] P_\theta(dx)$$
$$\geq \int L\left(g(\theta), d(x)\right) P_\theta(dx),$$

where $d(x) := \int a\mathrm{D}(da, x)$, a *non*-randomized estimator.

For convex loss functions we do not have to consider randomized estimators.

The Rao-Blackwell Theorem

- Consider the setting where $\Theta' = \mathbb{R}^m$ for some $m \in \mathbb{N}$.
- Often the loss function L satisfies that $y \mapsto L(x, y)$ is convex for all $x \in \mathbb{R}^m$.
- This is the case for $L(x, y) = |x - y|^p$ for $p \geq 1$.
- In this case more can be said than on the previous slides.
- Indeed, for any estimator $\delta : \mathcal{X} \to \mathbb{R}^m$ of $g(\theta)$, one has that the conditional expectation $E_\theta(\delta | T)$ has a risk that is at least as low as that of δ for all $\theta \in \Theta$.
- Importantly, we shall see that $E_\theta(\delta | T)$ does actually not depend on θ *when T is sufficient for θ* and thus write $E(\delta | T)$.
- By the factorization lemma, $E(\delta | T) = \phi(T)$ for some $\phi : \mathcal{T} \to \mathbb{R}^m$.

Preparatory observation

- Before stating and proving the Rao-Blackwell theorem let us formally establish that when T is sufficient for θ, then $E_\theta(\delta|T)$ does not depend on θ.
- What we are actually going to show is the expected fact

$$E_\theta(\delta|T=t) = \int_{\mathcal{X}} \delta(x) P_{X|T}(dx, t),$$

that is:

> "The conditional expectation of δ given $T=t$ is the expectation of δ in the conditional distribution $P_{X|T}(\cdot, t)$ of X given $T=t$."

- By definition of sufficiency this conditional distribution, and hence $E_\theta(\delta|T)$, does *not* depend on θ.
- Thus, if the following three slides are confusing, it is ok to take as given that $E_\theta(\delta|T) = \phi(T)$ for some $\phi : \mathcal{T} \to \mathbb{R}^m$ not depending on θ.

Showing that $E_\theta(\delta|T=t) = \int_{\mathcal{X}} \delta(x) P_{X|T}(dx,t)$

To show that

$$E_\theta(\delta|T=t) = \int_{\mathcal{X}} \delta(x) P_{X|T}(dx,t),$$

note that we use $(\mathcal{X}, \mathcal{A}, P_\theta)$ as the probability space upon which random variables δ and T are defined.

We must thus show that (cf. our discussion of conditional expectations)

1. $t \mapsto \int_{\mathcal{X}} \delta(x) P_{X|T}(dx,t)$ is $\mathscr{T}$-measurable. This follows from, e.g. Proposition A.39 in Liese and Miescke (2008).[6]
2. For all $B \in \mathscr{T}$

$$\int_{T^{-1}(B)} \left[\int_{\mathcal{X}} \delta(x) P_{X|T}(dx,t) \right] P_\theta(dx) = \int_{T^{-1}(B)} \delta(x) P_\theta(dx).$$

[6] Use that $t \mapsto \int_{\mathcal{X}} \mathbb{1}_A(x) P_{X|T}(dx,t) = P_{X|T}(A,t)$ is $\mathscr{T}$-measurable for all $A \in \mathcal{A}$ and the standard extension technique.

To verify that for all $B \in \mathscr{T}$

$$\mathcal{I} = \int_{T^{-1}(B)} \left[\int_{\mathcal{X}} \delta(x) P_{X|T}(dx, t) \right] P_\theta(dx) = \int_{T^{-1}(B)} \delta(x) P_\theta(dx),$$

apply a change of measure (substitution rule) to get

$$\mathcal{I} = \int_B \left[\int_{\mathcal{X}} \delta(x) P_{X|T}(dx, t) \right] P_{\theta,T}(dt)$$

note that by disintegration

$$\begin{aligned} \mathcal{I} &= \int_{\mathcal{T}} \int_{\mathcal{X}} \delta(x) \mathbb{1}_B(t) P_{X|T}(dx, t) P_{\theta,T}(dt) \\ &= \int_{\mathcal{X} \times \mathcal{T}} \delta(x) \mathbb{1}_B(t) P_{\theta,(X,T)}(dx, dt), \end{aligned}$$

where $P_{\theta,(X,T)} = P_\theta \circ (x \mapsto (x, T(x))^{-1}$ is the joint distribution of X and $T(X)$ and we used that “the integral of a function against the joint distribution can be found first integrating against the conditional distribution and then against the distribution of the conditioning variable”, cf. Proposition A.40 in Liese and Miescke (2008).

By a change of measure (substitution rule)

$$\mathcal{I} = \int_{\mathcal{X}\times\mathcal{T}} \delta(x)\mathbb{1}_B(t)P_{\theta,(X,T)}(dx,dt)$$
$$= \int_{\mathcal{X}} \delta(x)\mathbb{1}_B(T(x))P_\theta(dx) = \int_{T^{-1}(B)} \delta(x)P_\theta(dx).$$

Thus, we have shown that

$$\int_{T^{-1}(B)} \left[\int_{\mathcal{X}} \delta(x)P_{X|T}(dx,t)\right] P_\theta(dx) = \int_{T^{-1}(B)} \delta(x)P_\theta(dx),$$

and it follows that

$$\phi(t) := \int_{\mathcal{X}} \delta(x)P_{X|T}(dx,t) = E_\theta(\delta|T=t),$$

as desired. We have $E_\theta(\delta|T) = \phi(T)$.

The conditional expectation of δ given $T = t$ is just the expectation of δ in the conditional distribution of X given $T = t$.

The Rao-Blackwell Theorem

- Recall that we consider the setting where $\Theta' = \mathbb{R}^m$, $m \in \mathbb{N}$.
- For some $\phi : \mathcal{T} \to \mathbb{R}^m$, $\mathscr{T}$-$\mathcal{B}(\mathbb{R}^m)$-measurable we have $E_\theta(\delta|T) = \phi(T)$. The conditional expectation of δ given T does not depend on θ!
- For $u \in \mathbb{R}^m$, let $||u||_1 = \sum_{j=1}^m |u_j|$ denote the ℓ_1-norm of u.

Theorem 7 (Rao-Blackwell)

Suppose that $T : \mathcal{X} \to \mathcal{T}$ is sufficient for the family $\{P_\theta : \theta \in \Theta\}$ and that $\delta : \mathcal{X} \to \mathbb{R}^m$ is an estimator of $g(\theta)$ that satisfies $E_\theta||\delta||_1 < \infty$, $\theta \in \Theta$. Let $\phi(T) = E_\theta(\delta|T)$. If the loss function $L : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$ satisfies that $y \mapsto L(x, y)$ is convex for all $x \in \mathbb{R}^m$ then

$$R(\theta, \delta) \geq R(\theta, \phi(T)) \qquad \text{for all } \theta \in \Theta.$$

Furthermore, if δ is unbiased, then so is $\phi(T)$.

Proof of Rao-Blackwell Theorem

Fix $\theta \in \Theta$. By the conditional Jensen inequality,

$$\begin{aligned} R(\theta, \delta) &= E_\theta L(g(\theta), \delta) \\ &= E_\theta E_\theta \left[L(g(\theta), \delta) | T \right] \geq E_\theta L \left(g(\theta), E_\theta(\delta | T) \right). \end{aligned}$$

Since $E_\theta(\delta|T) = \phi(T)$ for all $\theta \in \Theta$ [by T being sufficient],

$$E_\theta L \left(g(\theta), E_\theta(\delta|T) \right) = E_\theta L \left(g(\theta), \phi(T) \right) = R(\theta, \phi(T)).$$

Thus,

$$R(\theta, \delta) \geq R(\theta, \phi(T)).$$

Finally, if $E_\theta \delta = g(\theta)$ then

$$E_\theta \phi(T) = E_\theta E_\theta(\delta|T) = E_\theta \delta = g(\theta).$$

□

Consequences of the Rao Blackwell Theorem

- Observe that the improved estimator $\phi(T)$ does not depend on the concrete loss function L.
- We only used that that the non-negative loss function satisfies that $y \mapsto L(x, y)$ is convex for all $x \in \mathbb{R}^m$.
- That is, $\phi(T)$ provides an improvement over δ for *all* convex loss functions.
- This is convenient since the result allows us to be reasonably agnostic about the concrete convex loss function.
- As long as the loss function is convex the above *Rao-Blackwellization* provides an improvement over any initially given estimator δ — no matter whether δ is unbiased or not.

Consequences of the Rao-Blackwell Theorem — variance reduction

- As mentioned, a popular loss function is $L(x, y) = (x - y)^2$.
- If δ is unbiased for $g(\theta)$, that is $E_\theta \delta = g(\theta)$ for all $\theta \in \Theta$, then this implies that

$$R(\theta, \delta) = E_\theta(\delta - g(\theta))^2 = E_\theta(\delta - E_\theta \delta)^2 = Var_\theta(\delta)$$

- Thus by the the Rao-Blackwell theorem

$$Var_\theta(\delta) = R(\theta, \delta) \geq R_\theta(\theta, \phi(T)) = Var_\theta(\phi(T)) \quad \text{for all } \theta \in \Theta,$$

that is the estimator $\phi(T)$ has *uniformly* no higher variance than δ.

- It is not always obvious how to find an initial unbiased estimator δ to be improved via conditioning on a sufficient statistic.
- However, if one can express $g(\theta)$ as

  $$g(\theta) = \int_{\mathcal{X}} f(x) P_\theta(dx),$$

  for some $f : \mathcal{X} \to \mathbb{R}^m$ ($\mathcal{A}$-$\mathcal{B}(\mathbb{R}^m)$-measurable), then $f(X_1)$ or $\frac{1}{n}\sum_{i=1}^n f(X_i)$ are unbiased for $\delta(\theta)$ (the latter assuming a sample of size n from P_θ at our disposal).
- This observation applies, in particular, to estimating the mean of a distribution.
- It is also not always easy to calculate $\phi(t) = E(\delta | T = t)$, cf. Example 3.2.4 in Pfanzagl (1994).
- Thus, it is not always easy to calculate the improved estimator in practice.

Towards risk optimal estimators

- The Rao-Blackwell theorem tells us, in principle, how to uniformly lower the risk of an estimator.
- We would like to find an estimator that uniformly has minimal risk.
- However, we have already argued that in general no such estimator exist.
- Let us briefly note that for any single $\theta_0 \in \Theta$ an estimator δ_0 that minimizes $R(\theta_0, \delta_0)$ is P_{θ_0}-almost surely unique if the loss function is *strictly* convex and $R(\theta_0, \delta_0) < \infty$.

Uniqueness of risk minimizer at θ_0

We continue to study the case of $\Theta' = \mathbb{R}^m$ for some $m \in \mathbb{N}$.

Theorem 8 (Uniqueness of risk minimizer)

Assume that $y \mapsto L(x, y)$ is strictly convex for all $x \in \mathbb{R}^m$ and that $\mathscr{C}$ is a convex subset of estimators. Then, for $\delta_1, \delta_2 \in \mathscr{C}$ the condition

$$R(\theta_0, \delta_1) = R(\theta_0, \delta_2) = \inf_{\delta \in \mathscr{C}} R(\theta_0, \delta) < \infty$$

implies that $\delta_1 = \delta_2$ P_{θ_0}-a.s.

"There is a unique minimizer of risk when the loss function is strictly convex and the corresponding risk is finite."

Comments

- Recall that if we do not restrict the class of estimators (a clearly convex set of estimators!), then $\delta = g(\theta_0)$ minimizes the risk as its risk is 0.
- In this case the theorem just states that $\delta = g(\theta_0)$ is the only estimator with zero risk.
- In fact this could already be gleaned from the proof of our result on general non-existence of an estimator minimizing risk at all $\theta \in \Theta$.
- We shall later restrict attention to the convex class of *unbiased* estimators for which the uniqueness result is (perhaps) slightly less trivial.

Comments on assumptions

- The *strict* convexity requirement on L can not be dropped. Think of $L(x, y) = c$ for some $c \geq 0$. Clearly all estimators have risk c (at all $\theta \in \Theta$) and hence minimize risk.
- The requirement of $\inf_{\delta \in \mathscr{C}} R(\theta_0, \delta_0) < \infty$ can not be dropped. If the minimizer of risk at θ_0 has $R(\theta_0, \delta) = \infty$ then any estimator minimizes risk at θ_0. Hence, δ_0 is not unique.

Note also that in case the members of the family $\{P_\theta : \theta \in \Theta\}$ are mutually absolutely continuous, then $\delta_1 = \delta_2$ P_{θ_0}-a.s. for some $\theta_0 \in \Theta$ is equivalent to $\delta_1 = \delta_2$ P_θ-a.s. for all $\theta \in \Theta$.

Exponential families, in particular, are mutually absolutely continuous.

Proof of Uniqueness Theorem

Let δ_1 and δ_2 minimize $R(\theta_0, \delta)$. By the assumed convexity of L the estimator $\delta_3 := \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ satisfies that

$$L(g(\theta_0), \delta_3) \leq \frac{1}{2} L(g(\theta_0), \delta_1) + \frac{1}{2} L(g(\theta_0), \delta_2). \tag{1}$$

Hence,

$$R(\theta_0, \delta_3) \leq \frac{1}{2} R(\theta_0, \delta_1) + \frac{1}{2} R(\theta_0, \delta_2) < \infty,$$

as we have assumed the minimal risk at θ_0 to be finite.
By (1)

$$\frac{1}{2} L(g(\theta_0), \delta_1) + \frac{1}{2} L(g(\theta_0), \delta_2) - L(g(\theta_0), \delta_3) \geq 0$$

such that

$$E_{\theta_0}\left(\frac{1}{2} L(g(\theta_0), \delta_1) + \frac{1}{2} L(g(\theta_0), \delta_2) - L(g(\theta_0), \delta_3)\right) \geq 0. \tag{2}$$

On the other hand, δ_1 and δ_2 minimize the risk, i.e.

$$R(\theta_0, \delta_1), R(\theta_0, \delta_2) \leq R(\theta_0, \delta_3)$$

such that

$$\begin{aligned} &E_{\theta_0}\left(\frac{1}{2}L(g(\theta_0), \delta_1) + \frac{1}{2}L(g(\theta_0), \delta_2) - L(g(\theta_0), \delta_3)\right) \\ =&\frac{1}{2}R(\theta_0, \delta_1) + \frac{1}{2}R(\theta_0, \delta_2) - R(\theta_0, \delta_3) \leq 0. \end{aligned} \tag{3}$$

(observe that we do not run into $\infty - \infty$ issues since all risks in the previous display are finite). Hence, combining (2) and (3).

$$E_{\theta_0}\left(\frac{1}{2}L(g(\theta_0), \delta_1) + \frac{1}{2}L(g(\theta_0), \delta_2) - L(g(\theta_0), \delta_3)\right) = 0.$$

The P_{θ_0}-integral of the non-negative function

$$\frac{1}{2}L(g(\theta_0),\delta_1)+\frac{1}{2}L(g(\theta_0),\delta_2)-L(g(\theta_0),\delta_3)$$

is 0.

It follows that

$$\frac{1}{2}L(g(\theta_0),\delta_1)+\frac{1}{2}L(g(\theta_0),\delta_2)-L(g(\theta_0),\delta_3)=0 \qquad P_{\theta_0}\text{-}a.s.$$

By the *strict* convexity of $y \mapsto L(x,y)$ for all $x \in \mathbb{R}^m$ this is only possible if $\delta_1 = \delta_2$. Hence, $\delta_1 = \delta_2$ P_{θ_0}-a.s.

□

Complete statistics

- In out pursuit of constructing (unbiased) estimators that uniformly minimize risk for all convex loss functions the concept of *completeness* is important.
- In the following $(\mathcal{T}, \mathscr{T})$ is a measurable space.

Definition 9 (Complete statistic)

Consider the experiment $\left(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\right)$. The statistic $T : \mathcal{X} \to \mathcal{T}$, $\mathcal{A}$-$\mathscr{T}$-measurable, is called (boundedly) complete if for every (bounded) $f : \mathcal{T} \to \mathbb{R}$ such that $\int_{\mathcal{T}} |f(t)|(P_\theta \circ T^{-1})(dt) < \infty$ and $\int_{\mathcal{T}} f(t)(P_\theta \circ T^{-1})(dt) = 0$ for every $\theta \in \Theta$ it holds that $f(T) = 0$ P_θ-almost surely for all $\theta \in \Theta$.

If T is complete, we also say that the family $\{P_\theta \circ T^{-1} : \theta \in \Theta\}$ is complete.

- In a sense completeness requires $\{P_\theta \circ T^{-1} : \theta \in \Theta\}$ to be sufficiently large since it requires for all f that if the integral is zero against all measures in the family, then f must be zero.

Examples and counterexamples

- Completeness turns out to be useful when studying estimators. Tests are mappings into $[0, 1]$ so that bounded completeness often suffices.
- We shall see that the sufficient statistics in exponential families are typically complete.
- Thus, exponential families again play a key role in terms of positive results.
- Let us first see that not all statistics complete.

A non-complete (yet sufficient!) statistic

Consider the family of uniform distributions on $[\theta, \theta+1]$, $\theta \in \mathbb{R}$. The density is

$$p_\theta(x) = \mathbb{1}_{[\theta,\theta+1]}(x) \quad \text{with respect to } \lambda_1.$$

Thus, in a sample of size one, $T(x) = x$ is *sufficient* for θ by the factorization theorem.

Let $f(x) = \sin(2\pi x)$ and observe that f has a period of one.

Clearly, $P_\theta \circ T^{-1} = P_\theta$ (recall $T(x) = x$) and so

$$\begin{aligned}\int f(t)(P_\theta \circ T^{-1})(dt) &= \int f(x) P_\theta(dx) = \int_{\mathbb{R}} f(x) p_\theta(x) \lambda_1(dx) \\ &= \int_\theta^{\theta+1} f(x) dx = \int_0^1 f(x) dx = 0,\end{aligned}$$

for all $\theta \in \mathbb{R}$. But clearly we do not have that $f(T) = 0$ P_θ-almost surely (for any $\theta \in \Theta$).

Completeness in exponential families

Let the setting be as when we introduced exponential families.

Theorem 10

Let $\{P_\theta : \theta \in \Theta\}$ *be an exponential family on* $(\mathcal{X}, \mathcal{A})$ *with density*

$$p_\theta(x) = C(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x)\right] h(x)$$

with respect to a σ*-finite dominating measure* μ *on* $\mathcal{A}$*. If* $\{(\eta_1(\theta), \dots, \eta_m(\theta)) : \theta \in \Theta\} \subseteq \mathbb{R}^m$ *has non-empty interior, then* $(T_1, \dots, T_m)$ *is complete.*

- We equivalently say that $\{P_\theta \circ [T_1, \dots, T_m]^{-1} : \theta \in \Theta\}$ is complete.
- If $\{(\eta_1(\theta), \dots, \eta_m(\theta)) : \theta \in \Theta\}$ contains an m-dimensional rectangle, then it of course has non-empty interior.

- Thus, the sufficient statistics in exponential families are often complete.
- A proof of the above theorem can be found in, e.g., Pfanzagl (1994), Theorem 1.6.10.
- If the parameter space Θ for the exponential family is the natural parameter space and it is open, we say that the exponential family is *regular*. Thus, the sufficient statistic in regular exponential families is complete.
- If the parameter space has an empty interior, we say that the exponential family is *curved*.

Examples: Normal distribution

- Recall that $\eta_1(\theta) = \frac{\mu}{\sigma^2}$, $\eta_2(\theta) = -\frac{1}{2\sigma^2}$, $\theta = (\mu, \sigma^2)$.
- Clearly, as soon as (μ, σ^2) is allowed to vary in an open rectangle R, then the image of R under $\theta \mapsto (\eta_1(\theta), \eta_2(\theta))$ contains an open rectangle.
  - Consider the center $\theta_c := (\mu_c, \sigma_c^2)$ of R and note that $(\eta_1(\theta_c), \eta_2(\theta_c))$ is in the interior of $\{\eta_1(\theta), \eta_2(\theta) : \theta \in R\}$. We can change θ a little bit and remain within the latter set.
- Hence, irrespectively of sample of size n, $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is complete.
- We have already seen that the same statistic is also sufficient.
- With the non-linear restriction $\sigma^2 = \mu^2$, the statistic $T = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)'$ remains sufficient, but it is not complete.[7]

[7] Non-completeness: Let $f(T) = 2(\sum_{i=1}^n x_i)^2 - (n+1)\sum_{i=1}^n x_i^2$ and show $E(f(T)) = 0$. Use that $\sum_{i=1}^n x_i$ is $N(n\theta, n\theta^2)$ so that $E(\sum_{i=1}^n x_i)^2 = \theta^2 n(n+1)$, while $E\sum_{i=1}^n x_i^2 = nEx_1^2 = 2n\theta^2$.

Examples: Poisson distribution

- Recall that $\eta_1(\lambda) = \log(\lambda)$.
- Thus, as soon as λ is allowed to vary in an interval, then so does $\eta_1(\lambda)$.
- Hence, in an independent sample of size n, $\sum_{i=1}^{n} x_i$ is complete.
- We have already seen that the same statistic is also sufficient.

A complete statistic outside exponential families

Although complete statistics are most easily found in exponential families, they can exist outside these:

- The continuous uniform distribution $\mathrm{U}(0,\theta)$ on $[0,\theta], \theta > 0$ has density $p_\theta(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$.
- This does *not* form an exponential family. The problem, we have seen, is that the support depends on θ.
- We have seen that $T(\boldsymbol{x}_n) = \max_{1\le i\le n} x_i$ is sufficient for θ in an i.i.d. sample of size n.
- It turns out that T is also complete.
- To see this, note that for all $0 < u < \theta$

$$P_\theta\left(T(\boldsymbol{x}_n) \le u\right) = P_\theta\left(\bigcap_{i=1}^{n}\{x_i \le u\}\right) = \left(\frac{u}{\theta}\right)^n$$

.

- Thus, the density of T is $p_{\theta,T}(t) = n\frac{t^{n-1}}{\theta^n}$ for $0 < t < \theta$ and 0 otherwise with respect to λ_1.

- Hence, let $f : \mathbb{R} \to \mathbb{R}$ and assume that for all $\theta \in \Theta$

$$E_\theta f[T(x)] = \int_0^\theta f(t)(P_\theta \circ T^{-1})(dt) = \frac{n}{\theta^n} \int_0^\theta f(t) t^{n-1} dt = 0.$$

- Thus, denoting by f^+ and f^- the positive and negative part of f and using that $f = f^+ - f^-$, it follows that

$$\int_0^\theta f^+(t) t^{n-1} dt = \int_0^\theta f^-(t) t^{n-1} dt \qquad \text{for all } \theta \in [0, \infty).$$

- Thus, the two measures

$$\nu^+(A) := \int_A f^+(t) t^{n-1} dt \quad \text{and} \quad \nu^-(A) := \int_A f^-(t) t^{n-1} dt$$

  on the Borel sets of $[0, \infty)$ agree on all intervals and thus $\nu^+ = \nu^-$.

- Here we used that if two measures agree on intervals in the real line, then they agree on $\mathcal{B}(\mathbb{R})$.

- Since $\nu^+ = \nu^-$ and these measures are both absolutely continuous with respect to λ_1, then they are σ-finite. Thus, it follows by the Radon-Nikodym theorem that their densities are equal λ_1-a.e., that is

$$f^+(t)t^{n-1} = f^-(t)t^{n-1} \qquad \text{for } \lambda_1\text{-}a.e.\ t \in [0,\infty),$$

  and so

$$f^+(t) = f^-(t) \qquad \text{for } \lambda_1\text{-}a.e.\ t \in [0,\infty)$$

- Since $(P_\theta \circ T^{-1}) \ll \lambda_1$ we also have $f^+ = f^-$ $(P_\theta \circ T^{-1})$-almost surely
- Hence, $f = f^+ - f^- = 0$ $(P_\theta \circ T^{-1})$-almost surely.
- It follows that T is complete.

Unbiased Estimation

- We have already seen that without restricting the class of estimators, we can not generally find an estimator that is best for all $\theta \in \Theta$.
- One reasonable way of restricting the class of estimators is to require an estimator $\delta : \mathcal{X} \to \Theta' = \mathbb{R}^m$ to be unbiased for the estimand $g(\theta)$, that is

$$E_\theta \delta(x) = \int_{\mathcal{X}} \delta(x) P_\theta(dx) = g(\theta) \quad \text{for all } \theta \in \Theta.$$

- By how much do we reduce the class of estimators by imposing unbiasedness?
- → Potentially by a lot!

No unbiased estimator may exist

- Let X be distributed according to the binomial distribution $\text{Binom}(n, p)$ with $n \in \mathbb{N}$ known and $0 < p < 1$ unknown.
- That is, our family of probability measures on (the power set of) $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ is $\{\text{Binom}(n, p) : p \in (0, 1)\}$.
- We know that $E_p X = np$ such that $\frac{X}{n}$ is an unbiased estimator of p.
- Thus, if $g(p) = p$, there exists an unbiased estimator of $g(p)$.

- Now consider $g(p) = \frac{1}{p}$.
- For an estimator $\delta : \mathbb{N}_0 \to [0,1]$ to be unbiased, we need

$$E_p\delta = \sum_{k=0}^{n} \delta(k) \binom{n}{k} p^k (1-p)^{n-k} = g(p) = \frac{1}{p} \text{ for } p \in (0,1).$$

- That no such δ exists can be seen by noting that $\frac{1}{p} \to \infty$ as $p \to 0$ but $E_p\delta = \sum_{k=0}^{n} \delta(k)\binom{n}{k} p^k (1-p)^{n-k} \to 0$ as $p \to 0$.
- Thus, no unbiased estimator of $g(p) = \frac{1}{p}$ exists.
- The example shows that the existence of unbiased estimators may depend on the concrete choice of g.
- Recall that if one can express $g(\theta)$ as

$$g(\theta) = \int_{\mathcal{X}} f(x) P_\theta(dx),$$

for some $f : \mathcal{X} \to \mathbb{R}^m$, then unbiased estimators exist since $f(X_1)$ or $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ are unbiased for $g(\theta)$ (the latter assuming a sample of size n from P_θ at our disposal).

Uniform minimum variance unbiased estimator (UMVUE)

Definition 11 (General optimality definition)

Fix an estimand $g : \Theta \to \mathbb{R}^m$ and a loss function L. An unbiased estimator δ of $g(\theta)$ is said to uniformly minimize the estimation risk if $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Theta$ and all unbiased estimators $\delta' : \mathcal{X} \to \mathbb{R}^m$.

- For $m = 1$, a loss function of particular interest is $L(x, y) = (x - y)^2$.
- We have seen that in this case the risk of an unbiased estimator is its variance.
- Without unbiasedness we have already seen that no best estimator in the above sense may exist.

Definition 12 (Uniform minimum variance unbiased estimator)

Fix an estimand $g : \Theta \to \mathbb{R}$. An unbiased estimator δ of $g(\theta)$ is said to uniformly minimize the variance if $Var_\theta(\delta) \leq Var_\theta(\delta')$ for all $\theta \in \Theta$ and all unbiased estimators $\delta' : \mathcal{X} \to \mathbb{R}$. We call δ the *uniform minimum variance unbiased estimator* or UMVUE.

Observations on uniqueness

- Recall that if $y \mapsto L(x, y)$ is strictly convex for all $x \in \mathbb{R}^m$ then the minimizer of risk δ at any θ is unique P_θ-almost surely if also $R(\theta, \delta) < \infty$.
- Typically the measures $\{P_\theta : \theta \in \Theta\}$ share the same null sets and it thus makes sense to talk about *the* uniform minimizer of strictly convex risk.
    - Thus, the optimality of δ at a single point $\theta_0 \in \Theta$ identifies δ P_{θ_0}-a.s. uniquely and hence P_θ-a.s. uniquely for all $\theta \in \Theta$ by the mutual absolute continuity of the members of $\{P_\theta : \theta \in \Theta\}$.
- In particular $L(x, y) = (x - y)^2$ is strictly convex in y for all $x \in \mathbb{R}$ and it thus makes sense to talk about *the* UMVU estimator if this has finite variance.

Rao-Blackwell-Lehmann-Scheffé theorem

- Having introduced sufficient and complete statistics, we are now in a position to state the classic *Rao-Blackwell-Lehmann-Scheffé theorem*.
- This is one of the core results in the theory of risk optimal unbiased estimators.
- Frequently, the result is just called the Lehmann-Scheffé theorem.
- However, its proof builds on the Rao-Blackwell theorem and hence their name is also sometimes attached to the theorem.

- We know that for strictly convex loss functions any unbiased estimator minimizing risk at a $\theta \in \Theta$ is P_θ-a.s. uniquely determined if it has finite risk.
- We now show that this estimator is a function of a complete sufficient statistic whenever such a statistic exists.

We continue to work with the case in which $\Theta' = \mathbb{R}^m$ for some $m \in \mathbb{N}$.

Theorem 13 (Rao-Blackwell-Lehmann-Scheffé)

Suppose that $T : \mathcal{X} \to \mathcal{T}$ is sufficient and complete in the experiment $\big(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$. Suppose, furthermore, that there exists at least one unbiased estimator $\delta : \mathcal{X} \to \mathbb{R}^m$ of the estimand $g(\theta)$ and that the loss function $L(x, y)$ is convex in y for all $x \in \mathbb{R}^m$.

1. *There exists an unbiased estimator of $g(\theta)$ of the form $\phi(T)$ where $\phi : \mathcal{T} \to \mathbb{R}^m$, $\mathscr{T}$-$\mathcal{B}(\mathbb{R}^m)$-measurable.*
2. *If $\phi : \mathcal{T} \to \mathbb{R}^m$ is $\mathscr{T}$-$\mathcal{B}(\mathbb{R}^m)$-measurable and $\phi(T)$ is unbiased for $g(\theta)$, then $\phi(T)$ uniformly minimizes risk among unbiased estimators. That is for any other unbiased estimator $\delta' : \mathcal{X} \to \mathbb{R}^m$ of $g(\theta)$ it holds that*

   $$R(\theta, \phi(T)) \leq R(\theta, \delta') \qquad \text{for all } \theta \in \Theta.$$

   $\phi(T)$ is P_θ-a.s., $\theta \in \Theta$, uniquely determined in the class of all unbiased estimators that are functions of T. If, furthermore, $L(x, y)$ is strictly convex in y for all $x \in \mathbb{R}^m$ and $R(\theta, \phi(T)) < \infty$ then $\phi(T)$ is the P_θ-a.s. unique unbiased estimator of minimal risk.

Main conclusion

- (1) in the Rao-Blackwell-Lehmann-Scheffé Theorem asserts the *existence* of an unbiased estimator that is a function of the complete sufficient statistic.
    - This actually only uses the sufficiency but not the completeness of T as can be seen from the proof below.
- (2) asserts that the unbiased estimator that is a function of T is *unique* [in the class of all unbiased estimators that are functions of T] and uniformly minimizes risk.

Main take-away

A function ϕ of a complete sufficient statistic T uniformly minimizes the convex risk among all unbiased estimators of its expectation.

Uniqueness of the best estimator — a little less important

$\phi(T)$ is unique in the class of unbiased estimators that are functions of T [in terms of convex risk nothing is lost by restricting attention to this class since the risk is weakly decreasing under Rao-Blackwellization].

If, furthermore, the risk of $\phi(T)$ is finite and the loss function is strictly convex, then $\phi(T)$ uniquely minimizes risk in the class of all unbiased estimators

Proof of Rao-Blackwell-Lehmann-Scheffé theorem

1) Let δ be unbiased for θ (such δ exists by assumption). Since T is sufficient for θ there exists a $\phi : \mathcal{T} \to \mathbb{R}^m$ that is $\mathscr{T}$-$\mathcal{B}(\mathbb{R}^m)$-measurable such that $\phi(T) = E_\theta(\delta|T)$ P_θ-a.s. for all $\theta \in \Theta$.

By iterated expectations

$$E_\theta \phi(T) = E_\theta[E_\theta(\delta|T)] = E_\theta \delta = g(\theta) \qquad \text{for all } \theta \in \Theta,$$

that is $\phi(T)$ is unbiased.

[Note that the above argument basically repeats the proof of the second part of the Rao-Blackwell Theorem]

2) Let δ' be another unbiased estimator of $g(\theta)$. Then, by the Rao-Blackwell theorem, there exists a $\psi : \mathcal{T} \to \mathbb{R}^m$, $\mathscr{T}$-$\mathcal{B}(\mathbb{R}^m)$-measurable, such that

1. $E_\theta \psi(T) = g(\theta)$ for all $\theta \in \Theta$.
2. $R(\theta, \psi(T)) \leq R(\theta, \delta')$ for all $\theta \in \Theta$.

Hence, for all $\theta \in \Theta$,

$$0 = E_\theta \phi(T) - E_\theta \psi(T) = \int_{\mathcal{T}} [\phi(t) - \psi(t)](P_\theta \circ T^{-1})(dt).$$

By completeness of T it follows that $\phi(T) = \psi(T)$ P_θ-almost surely for all $\theta \in \Theta$.
The previous equation display also establishes that there there is at most one unbiased estimator that is a function of T up to a P_θ-null set.

[Here we see how the definition of completeness is the "right" one. It is exactly what is needed to make the proof work.]

Hence,

$$R(\theta, \phi(T)) = R(\theta, \psi(T)) \leq R(\theta, \delta') \qquad \text{for all } \theta \in \Theta.$$

Thus, since δ' was arbitrary, $\phi(T)$ has uniformly no larger risk than *any* other unbiased estimator of $g(\theta)$.

If furthermore, $y \mapsto L(x, y)$ is strictly convex for all $x \in \mathbb{R}^m$ and $R(\theta, \phi(T)) < \infty$ then it follows by the uniqueness theorem (page 69) that $\phi(T)$ is unique P_θ-a.s.

□

Comments on Rao-Blackwell-Lehmann-Scheffé theorem

- Observe that the estimator $\phi(T)$ does *not* depend on the particular convex loss function used.
- Thus, $\phi(T)$ uniformly minimizes the risk in the class of unbiased estimators for *any* convex loss function.
- This is useful since even if two researchers disagree on the exact convex loss function to be used, the same estimator is optimal for both of them.

UMVU estimators

- Let $m = 1$.
- A particular (strictly) convex loss function is the square loss.
- Thus, $\phi(T)$ is in particular *a* UMVU estimator of $g(\theta)$!
- That is for any unbiased estimator $\delta : \mathcal{X} \to \mathbb{R}$

$$Var_\theta(\phi(T)) \leq Var_\theta(\delta) \qquad \text{for all } \theta \in \Theta.$$

- If $Var_\theta(\phi(T)) < \infty$ then $\phi(T)$ is *the* UMVUE, i.e. it is unique.

Some practical consequences of the Rao-Blackwell-Lehmann-Scheffé theorem

- Let δ and δ' be unbiased estimators of $g(\theta)$.
- One could then obtain $\phi(T)$ by calculating $E_\theta(\delta|T) =: \phi(T)$ or $E_\theta(\delta'|T) =: \psi(T)$. Which of these to use?
- Well, being unbiased, it follows by the Rao-Blackwell-Lehmann-Scheffé theorem $\phi(T)$ and $\psi(T)$ are identical P_θ-almost surely for all $\theta \in \Theta$.
- Thus, it does not matter which one we use.
- Hence, one can — in principle — use *any* unbiased estimator for which the conditional expectation given T is easy to calculate.
- Recall that we have already remarked that these conditional expectations can be tricky to calculate.
- In any case, the Rao-Blackwell-Lehmann-Scheffé theorem tells us that we need only consider functions of the complete sufficient statistic T in our search for an unbiased estimator uniformly minimizing the risk.

Comments on uniqueness

- If $R(\theta, \delta) = \infty$ for all unbiased estimators and $\theta \in \Theta$ then $\phi(T)$ need not be the unique minimal convex risk unbiased estimator [as every unbiased estimator has the same risk].
- That is why we generally only claim uniqueness within the class of unbiased estimators that are functions of the complete sufficient statistic T.
- We know that as soon as risk is finite and the loss function is strictly convex, then the uniqueness extends to the full class of unbiased estimators [be they functions of T or not].

Example: UMVU Estimation in $\mathrm{N}(\mu, \sigma^2)$

- Consider a sample $X_1, \ldots, X_n$ with $n \geq 2$ from $\mathrm{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ unknown.
- We have already seen that $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is complete sufficient for $\theta = (\mu, \sigma^2)$.
- Since
$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i = \phi_n \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right),$$
for $\phi_n : \mathbb{R}^2 \to \mathbb{R}$ defined via $(u, v) \mapsto \frac{u}{n}$ is an unbiased estimator of μ that is a function of a complete sufficient statistic it follows from the Rao-Blackwell-Lehmann-Scheffé Theorem that $\bar{x}$ is a UMVU estimator of μ. [use $g(\mu, \sigma^2) = \mu$]
- It is the *unique* UMVUE since the loss is strictly convex and $Var_\theta(\bar{x}_n) = \sigma^2/n < \infty$.
- Of course $\bar{x}$ is actually the unique unbiased estimator uniformly minimizing risk for *any* strictly convex loss functions satisfying the conditions of the Rao-Blackwell-Lehmann-Scheffé theorem.

Estimating σ^2

- Define $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.
- s^2 is clearly a function of the complete sufficient statistic $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$.
- Since $\frac{s^2}{\sigma^2} \sim \chi^2(n-1)$ for all $\theta = (\mu, \sigma^2)$, it follows that

$$E_\theta\left(\frac{s^r}{\sigma^r}\right) = E_\theta\left(\frac{s^2}{\sigma^2}\right)^{r/2} = \frac{2^{r/2}\Gamma([n-1+r]/2)}{\Gamma([n-1]/2)} =: \frac{1}{K_{n,r}}.$$

- Hence, by the Rao-Blackwell-Lehmann-Scheffé theorem, $s^r K_{n,r}$ is a UMVU estimator of σ^r since it is unbiased and a function of a complete sufficient statistic [assuming $n-1+r>0$].
- Hence, we can estimate any power of σ unbiased and with minimal variance.
- By properties of the Γ function, $K_{n,2} = 1/(n-1)$ such that, as expected, $\frac{s^2}{n-1}$ is UMVU for σ^2 (choose $r=2$).
- Note, the quadratic risk minimizer is $\frac{s^2}{n+1}$ (page 21).

Further estimands

- Since $(\bar{x}, s^2/(n-1))$ is a function of a complete sufficient statistic that is unbiased for (μ, σ^2) it also follows that it is UMVU for this estimand.
- By linearity of expectations, any linear function of $l(\mu, \sigma^2)$ is UMVU estimable by $l\left(\bar{x}, s^2/(n-1)\right)$.
- Another estimand of interest μ/σ. This is called the Sharpe ratio in finance.
- Since $\bar{x}$ and s are independent, it follows that $\bar{x} \cdot s^{-1} K_{n,-1}$ is unbiased for μ/σ.
- Being a function of the complete sufficient statistic $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$, it follows by the Rao-Blackwell-Lehmann-Scheffé theorem that $\bar{x} \cdot s^{-1} K_{n,-1}$ is a UMVUE for $g(\mu, \sigma^2) = \mu/\sigma$.

Poisson distribution

- We have seen that in a sample of size n from a Poisson distribution with intensity $\lambda > 0$, one has that $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$ is a complete sufficient statistic.
- Since $E_\lambda \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \lambda$, it follows that $\frac{1}{n} \sum_{i=1}^n x_i$ is a UMVUE of λ since it is unbiased and a function of a complete sufficient statistic.
- Since $Var_\lambda(\bar{x}_n) = \frac{\lambda}{n} < \infty$, it follows that $\bar{x}_n$ is the $\otimes_{i=1}^n \mathrm{Poi}_\lambda$-a.s. unique UMVUE.

Exponential families

- Above we considered minimum risk unbiased estimation within the Normal and Poisson distribution.
- These are of course special members of the class of exponential families.
- Within exponential families we have easy access to complete sufficient statistics.
- Hence, the Rao-Blackwell-Lehmann-Scheffé Theorem is particularly useful within exponential families.
- It gives us a recipe for constructing unbiased estimators that uniformly minimize convex risk.
- These must be functions of $(T_1, \dots, T_m)$ in

$$p_\theta(x) = C(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x)\right] h(x).$$

The following corollary merely combines the sufficient condition for completeness of the sufficient statistics in exponential families with the Rao-Blackwell-Lehmann-Scheffé theorem.

Corollary 14

Consider the experiment $\left(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\right)$ *where* $\{P_\theta : \theta \in \Theta\}$ *forms an exponential family, that is the densities of the* P_θ *are of the form*

$$p_\theta(x) = C(\theta) \exp\left[\sum_{i=1}^{m} \eta_i(\theta) T_i(x)\right] h(x),$$

with respect to a σ*-finite dominating measure* μ *on* $\mathcal{A}$*. If* $\{(\eta_1(\theta), \ldots, \eta_m(\theta)) : \theta \in \Theta\} \subseteq \mathbb{R}^m$ *has non-empty interior and for some* $g : \Theta \to \mathbb{R}^k$ *there exists an unbiased estimator, then there exists an unbiased estimator that depends on* x *only via* $(T_1(x), \ldots, T_m(x))$*. This estimator uniformly minimizes the estimation risk for any loss function* $L(x, y)$ *that is convex in* y *for every* $x \in \mathbb{R}^k$*.*

A “ridiculous” UMVU estimator

- Suppose X has a Poisson distribution with $\lambda > 0$, that is its density with respect to the counting measure τ (probability mass function) is

$$p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

- Suppose the estimand is $g(\lambda) = e^{-3\lambda}$.
- Consider estimating $e^{-3\lambda}$ by the following function of the complete sufficient statistic x: $\delta(x) = (-2)^x$. Now

$$\begin{aligned} E_\lambda \delta(x) &= \sum_{k=0}^{\infty} (-2)^k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \cdot \frac{(-2\lambda)^k}{k!} = e^{-\lambda} e^{-2\lambda} \\ &= e^{-3\lambda}, \end{aligned}$$

where we used that $\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$ for all $x \in \mathbb{R}$.

- $\delta(x) = (-2)^x$ is in fact the unique unbiased estimator of $e^{-3\lambda}$.
- This follows from the completeness of $T(x) = x$. Assume there exists another unbiased estimator δ' of $e^{-3\lambda}$. Then,

$$0 = E_\lambda \delta - E_\lambda \delta' = E_\lambda[\delta - \delta'] \qquad \text{for all } \lambda > 0,$$

  which implies that $\delta(x) = \delta'(x)$ for P_λ-almost all x.
- Being the unique unbiased estimator, δ is also the UMVUE.

- Even though $\delta(x) = (-2)^x$ is the unique unbiased estimator, it is “ridiculous” in the following sense.
    1. $\delta(x)$ is negative for all x odd although $e^{-3\lambda} > 0$. The biased estimator $\delta'(x) = \max(\delta(x), 0)$ does not have this deficiency.
    2. In fact $\delta'(x)$ is closer to $e^{-3\lambda}$ for all x odd and equal to $\delta(x)$ for x even. Thus, δ' is never has higher, but often strictly lower loss than δ.
    3. Thus, δ' has a smaller risk than δ when using quadratic loss and hence a lower mean square error.
- The lesson is that a UMVU estimator need not be good.
- In this example restricting ourselves to unbiased estimators removed δ' from consideration.
- Outside the class of unbiased estimators we could find an estimator δ' that many would agree is better than δ. But of course δ' is not unbiased.

Further examples of UMVU estimators

- Many more examples of applications of the Rao-Blackwell-Lehmann-Scheffé theorem can be found on pages 91–113 in Lehmann and Casella (1998).
- Romano and Siegel (1986) also contains many insightful (counter)examples. The above example of a “ridiculous” UMVUE is taken from there.

A summary

- In a sense the theory of UMVU estimation is a theory of exponential families.
- The reason is that the Rao-Blackwell-Lehmann-Scheffé Theorem relies crucially on access to a complete sufficient statistic.
- Such complete sufficient statistics can be found in a systematic way within exponential families.
- Outside exponential families one typically needs more ad hoc analysis.

Main take-away

A function of a complete sufficient statistic uniformly minimizes the convex risk among all unbiased estimators of its expectation. That is, an unbiased estimator that is a function of a complete sufficient statistic uniformly minimizes convex risk.

The Cramér-Rao lower bound

- The theory of UMVU estimators is in some sense a theory confined to exponential families as it relies on access to a complete sufficient statistic.
- We have seen that such complete sufficient statistics are readily available in many exponential families.
- In case we can't assert that an estimator is UMVU by means of the Rao-Blackwell-Lehmann-Scheffé theorem, we would still like to be able to judge how "good" a given estimator is.
- To this end, we shall now establish the *Cramér-Rao lower bound* on the variance of an (unbiased) estimator.
- This lower bound is also called the (Fisher) Information Inequality.

- With the Cramér-Rao lower bound at hand, one can then compare the variance of a given estimator to the lower bound in order to judge how good the estimator is.
- It turns out that the lower bound is typically not sharp, i.e. it is lower than the lowest attainable variance of an unbiased estimator.
- However, it may still some insights into how good an estimator is.

For simplicity, we shall develop our theory for θ being a one-dimensional parameter [and $g(\theta)$ too]. See Chapter 2.6 in Lehmann and Casella (1998) for multivariate extensions.

Fisher Information

Definition 15

Let $\mathcal{E} = \big(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$ be an experiment where Θ is an open subset of $\mathbb{R}$. If the elements of $\{P_\theta : \theta \in \Theta\}$ have densities $\{p_\theta : \theta \in \Theta\}$ with respect to a σ-finite dominating measure μ such that $\frac{d}{d\theta} p_\theta(x)$ exists for μ-almost all $x \in \mathcal{X}$ then we define the *Fisher Information* $I(\theta)$ at θ as

$$I(\theta) := E_\theta \left[\frac{d}{d\theta} \log(p_\theta) \right]^2 = E_\theta \left[\frac{p'_\theta}{p_\theta} \right]^2 = \int_{\mathcal{X}} \left[\frac{p'_\theta(x)}{p_\theta(x)} \right]^2 p_\theta(x) \mu(dx).$$

Note the information $I(\theta)$ is large when the expected value of the squared derivative of $\theta \mapsto \log(p_\theta)$ is large. That is when the log-likelihood function is steep, i.e. small changes in θ induce large changes in $\log(p_\theta)$. Put differently, there is a lot if *information* about θ in the log-likelihood.

Some heuristics on information

If integration and differentiation can be interchanged then, using that

$$\int_{\mathcal{X}} p_\theta(x)\mu(dx) = 1, \tag{4}$$

one obtains that

$$0 = \frac{d}{d\theta}\int_{\mathcal{X}} p_\theta(x)\mu(dx) = \int_{\mathcal{X}} \frac{d}{d\theta} p_\theta(x)\mu(dx).$$

In addition,

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \frac{d}{d\theta} p_\theta(x)\mu(dx) = \int_{\mathcal{X}} \left[\frac{d}{d\theta}\log(p_\theta(x))\right] p_\theta(x)\mu(dx) \\ &= E_\theta\left[\frac{d}{d\theta}\log(p_\theta)\right], \end{aligned}$$

from which it follows that

$$I(\theta) = Var_\theta\left(\frac{d}{d\theta}\log(p_\theta)\right). \tag{5}$$

If, furthermore, $\theta \mapsto p_\theta(x)$ is twice differentiable for μ-almost all $x \in \mathcal{X}$ and if the second derivative in (4) can be found by differentiating twice under the integral, then

$$\int_{\mathcal{X}} \frac{d^2}{d\theta^2} p_\theta(x)\mu(dx) = \frac{d}{d\theta} \int_{\mathcal{X}} \frac{d}{d\theta} p_\theta(x)\mu(dx) = 0. \tag{6}$$

Furthermore,

$$\frac{d^2}{d\theta^2} \log(p_\theta) = \frac{d}{d\theta} \frac{\frac{d}{d\theta} p_\theta}{p_\theta} = \frac{\left(\frac{d^2}{d\theta^2} p_\theta\right) p_\theta - \left(\frac{d}{d\theta} p_\theta\right)^2}{p_\theta^2}.$$

Thus, taking expectations in the previous display and using (6)

$$\begin{aligned}
&E_\theta\left(\frac{d^2}{d\theta^2}\log(p_\theta)\right)\\
=&\int_{\mathcal{X}}\left(\frac{d^2}{d\theta^2}\log(p_\theta(x))\right)p_\theta(x)\mu(dx)\\
=&\int_{\mathcal{X}}\frac{d^2}{d\theta^2}p_\theta(x)\mu(dx)-\int_{\mathcal{X}}\left(\frac{p'_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)\mu(dx)\\
=&-I(\theta).
\end{aligned}$$

Put concisely,

$$I(\theta)=-E_\theta\left(\frac{d^2}{d\theta^2}\log(p_\theta)\right). \tag{7}$$

The information identity arises when combining (5) and (7):

$$I(\theta)=-E_\theta\left(\frac{d^2}{d\theta^2}\log(p_\theta)\right)=Var_\theta\left(\frac{d}{d\theta}\log(p_\theta)\right).$$

Information in i.i.d. samples

- If we observe n independent observations from P_θ with density p_θ, then P_θ^n has density

$$p_{\theta,n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\theta(x_i)$$

- Thus, the information at sample size n at $\theta \in \Theta$ is

$$I_n(\theta) = E_{n,\theta}\left[\frac{d}{d\theta}\log\left(\prod_{i=1}^{n} p_\theta\right)\right]^2$$

$$= \sum_{i=1}^{n} E_\theta\left[\frac{d}{d\theta}\log(p_\theta)\right]^2 = nI(\theta),$$

where we used the independence assumption and that $E_\theta \log(p_\theta) = 0$.

- Thus, information accumulates additively.

Information in normal distribution

- Consider the case of $\mathrm{N}(\mu, \sigma_0^2)$ where $\sigma_0^2 > 0$ is known.
- Then, recalling that $p_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}$

$$\log p_\mu(x) = -\log(\sqrt{2\pi}\sigma_0) - \frac{(x-\mu)^2}{2\sigma_0^2}$$

such that the μ-derivative

$$[p'_\mu(x)]^2 = \left[\frac{\partial}{\partial\mu} p_\mu(x)\right]^2 = \frac{(x-\mu)^2}{\sigma_0^4},$$

and so

$$I(\mu) = E_\mu[p'_\mu(x)]^2 = \frac{E_\mu(x-\mu)^2}{\sigma_0^4} = \frac{1}{\sigma_0^2}.$$

- "The smaller σ_0^2, the larger the information about μ."
- Clearly, $I_n(\mu) = \frac{n}{\sigma_0^2}$.

Information in Poisson distribution

- Consider the case of $\mathrm{Poi}(\lambda)$ with $\lambda > 0$ unknown.
- Then, recalling that $p_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$,

$$\log p_\lambda(x) = -\lambda + x\log(\lambda) - \sum_{i=1}^{x}\log(i)$$

and so

$$\log p'_\lambda(x) = -1 + \frac{x}{\lambda}.$$

- It follows that since $E_\lambda x^2 = Var_\lambda(x) + (E_\lambda x)^2 = \lambda + \lambda^2$

$$\begin{aligned} I(\lambda) &= E_\lambda[\log p'_\lambda(x)]^2 \\ &= E_\lambda\left(1 + \frac{x^2}{\lambda^2} - 2\frac{x}{\lambda}\right) = 1 + \frac{\lambda+\lambda^2}{\lambda^2} - 2 = \frac{1}{\lambda}. \end{aligned}$$

- Clearly, $I_n(\lambda) = \frac{n}{\lambda}$

Heuristics of the Cramér-Rao lower bound

- As the name indicates, the information bound (Cramér-Rao lower bound) bound provides a lower bound on an estimator in terms of the (Fisher) information.
- It is thus not surprising that the function $q_\theta(x) := \frac{d}{d\theta}\log(p_\theta(x))$ and its expectation, that is the information, play a crucial rule.
- Let δ be an estimator with finite variance of $g(\theta)$ satisfying $E_\theta\delta = g(\theta) + b(\theta)$. Thus, $b(\theta)$ is the bias.
- We shall show that

$$Var_\theta(\delta) \geq \frac{[g'(\theta) + b'(\theta)]^2}{I(\theta)}.$$

- This is the *Cramér-Rao lower bound* on the variance of the estimator δ.

Remarks on Cramér-Rao inequality

- In case δ is unbiased for $g(\theta)$, the inequality reduces to

$$Var_\theta(\delta) \geq \frac{[g'(\theta)]^2}{I(\theta)}.$$

- If, furthermore, the estimand is $g(\theta) = \theta$, then

$$Var_\theta(\delta) \geq \frac{1}{I(\theta)}.$$

- And in an i.i.d. sample if size n

$$Var_\theta(\delta) \geq \frac{1}{nI(\theta)}.$$

Example: Normal distribution

- Consider the case of $\mathrm{N}(\mu, \sigma_0^2)$ where $\sigma_0^2 > 0$ is known.
- In a sample of size n we have seen that $I_n(\mu) = \frac{n}{\sigma_0^2}$.
- Thus, we get that $Var_\mu(\delta) \geq \frac{\sigma_0^2}{n}$ for "any" unbiased estimator δ [regularity conditions to come].
- Since $E_\mu(\bar{x}) = \mu$ and $Var_\mu(\bar{x}) = \frac{\sigma_0^2}{n}$ we conclude that $\bar{x}$ has minimal variance at all μ in the class of unbiased estimators.
- That is, $\bar{x}$ is UMVU.
- Of course we knew this already since $T(\boldsymbol{x}) = \bar{x}$ is a complete sufficient statistic that unbiased for μ.

Example: Poisson distribution

- Consider the case of Poi(λ) with $\lambda > 0$ unknown.
- In a sample of size n we have seen that $I_n(\lambda) = \frac{n}{\lambda}$.
- Thus, we get that $Var_\mu(\delta) \geq \frac{\lambda}{n}$ for "any" unbiased estimator δ [regularity conditions to come].
- Since $E_\lambda(\bar{x}) = \lambda$ and $Var_\lambda(\bar{x}) = \frac{\lambda}{n}$ we conclude that $\bar{x}$ has minimal variance at all $\lambda > 0$ in the class of unbiased estimators.
- That is, $\bar{x}$ is UMVU.
- Of course we knew this already since $T(\boldsymbol{x}) = \bar{x}$ is a complete sufficient statistic that unbiased for λ.

Sharpness of the Cramér-Rao bound

- It turns out to be no coincidence that the Cramér-Rao bound is sharp/attainable in the above two examples.
- In fact, one can show that if there exists an unbiased estimator δ of $g(\theta)$ that achieves the lower bound for all $\theta \in \Theta$, then $\{P_\theta : \theta \in \Theta\}$ is an exponential family with δ as a linear function of the canonical statistic (page 28. See Bickel and Doksum (2015) Theorem 3.4.2 for a precise statement [see also Theorem 2.5.12 in Lehmann and Casella (1998)]

An example where the Cramér-Rao bound is not sharp

- Consider the case of Poi(λ) with $\lambda > 0$ unknown.
- We wish to estimate $g(\lambda) = e^{-\lambda}$.
- Observe that $\delta(x) = \mathbb{1}_{\{0\}}(x)$ is unbiased since

$$E_\lambda \delta = P_\lambda(x = 0) = e^{-\lambda} \qquad \text{for all } \lambda > 0.$$

- Being a function of the complete sufficient statistic x, it follows that δ is UMVU.
- Furthermore, since $\delta^2 = \delta$

$$Var_\lambda(\delta) = E_\lambda \delta^2 - (E_\lambda \delta)^2 = E_\lambda \delta - (E_\lambda \delta)^2 = e^{-\lambda}(1 - e^{-\lambda})$$

- But the Cramér-Rao bound only guarantees

$$Var_\lambda(\delta) \geq \frac{[g'(\lambda)]^2}{I(\lambda)} = \lambda e^{-2\lambda},$$

which can be shown to be smaller than $e^{-\lambda}(1 - e^{-\lambda})$.

Cramér-Rao bound in $N(\mu, \sigma^2)$

- Consider a sample $X_1, \ldots, X_n$ with $n \geq 2$ from $N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ unknown.
- The joint density is

$$p_\theta(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} X_i - \frac{n\mu^2}{2\sigma^2}\Big).$$

  Thus, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is canonical statistic (page 28).
- For any $a, b \in \mathbb{R}$ then $E(a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2) = an\mu + b((n-1)\sigma^2 + n\mu^2)$.
- The estimator $\bar{X}$ for μ is a linear function of the canonical statistic. It is unbiased and attains the Cramér-Rao bound.
- No linear function has expectation σ^2 for arbitrary μ. Thus, no estimator for σ^2 attains the Cramér-Rao bound.

Basic idea of proof of Cramér-Rao lower bound

The basic idea is the proof of the Cramér-Rao lower bound is simple:

1. Use the Cauchy-Schwarz inequality to conclude

$$Cov_\theta(\delta, q_\theta) \leq \left[Var_\theta(\delta) Var_\theta(q_\theta) \right]^{1/2}.$$

2. Hence,

$$Var_\theta(\delta) \geq \frac{Cov_\theta^2(\delta, q_\theta)}{Var_\theta(q_\theta)} = \frac{Cov_\theta^2(\delta, q_\theta)}{I(\theta)},$$

   the equality following from (5).

3. Show that $Cov_\theta^2(\delta, q_\theta) = [g'(\theta) + b'(\theta)]^2$.

Thus, let us establish (3).

We have seen that, provided integration and differentiation can be interchanged in

$$\int_{\mathcal{X}} p_\theta(x)\mu(dx) = 1$$

then

$$E_\theta q_\theta = E_\theta \left(\frac{d}{d\theta} \log(p_\theta) \right) = 0.$$

If, in addition, integration and differentiation can be interchanged in

$$\int_{\mathcal{X}} \delta(x) p_\theta(x) \mu(dx) = E_\theta \delta = g(\theta) + b(\theta),$$

then

$$\begin{aligned} Cov_\theta(\delta, q_\theta) &= E_\theta(\delta q_\theta) \\ &= \int_{\mathcal{X}} \delta(x) \left(\frac{d}{d\theta} \log(p_\theta(x)) \right) p_\theta(x) \mu(dx) \\ &= \int_{\mathcal{X}} \delta(x) p'_\theta(x) \mu(dx) \\ &= \frac{d}{d\theta} \int_{\mathcal{X}} \delta(x) p_\theta(x) \mu(dx) \\ &= g'(\theta) + b'(\theta), \end{aligned}$$

and therefore, as claimed,

$$Cov^2_\theta(\delta, q_\theta) = [g'(\theta) + b'(\theta)]^2.$$

Thus, all that we need to do is to provide conditions under which we can interchange the order of integration and differentiation in

1. $\int_{\mathcal{X}} p_\theta(x)\mu(dx) = 1.$
2. $\int_{\mathcal{X}} \delta(x)p_\theta(x)\mu(dx) = g(\theta) + b(\theta).$

Such conditions can be found as consequences of the dominated convergence theorem, cf. Theorem 6.28 in Klenke (2020).

Remark — Exponential families

These conditions are satisfied for exponential families as we may have time to show!

- We consider the setting of $g : \mathbb{R} \to \mathbb{R}$. That is the parameter and the function of it that we wish to estimate are one-dimensional.
- Recall that $E_\theta \delta = g(\theta) + b(\theta)$ for some $b : \mathbb{R} \to \mathbb{R}$.
- Thus, b is the bias function of δ.

A set of sufficient conditions for the Cramér-Rao bound

Let $\mathcal{E} = \big(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$ be an experiment. We give sufficient conditions for the Cramér-Rao bound to hold at an interior point $\theta_0 \in \Theta$. There exists an $\varepsilon > 0$ such that $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subseteq \Theta$ and

A1 The elements of $\{P_\theta : \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)\}$ have densities $\{p_\theta : \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)\}$ with respect to a σ-finite dominating measure μ.

A2 $\frac{d}{d\theta} p_\theta(x)$ exists for μ-almost all $x \in \mathcal{X}$ at all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$. Furthermore, $A := \big\{x \in \mathcal{X} : p_\theta(x) > 0\big\}$ does not depend on $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$.

A3 $I(\theta_0) > 0$.

A4 There exists a measurable function $h : \mathcal{X} \to [0, \infty)$ such that $|\frac{d}{d\theta} p_\theta(x)| \leq h(x)$ for all $x \in A$ and all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ and $\int_A \frac{h^2(x)}{p_{\theta_0}(x)} \mu(dx) < \infty$.

Comments on assumptions

- Assumption A1 imposes the existence of densities in a neighborhood of θ_0.
- Assumption A2 imposes that these densities are diiferentiable.
- The information at θ_0 is positive
- A4 is the most tricky assumptions. This is what allows us to interchange differentiation and integration as desired.

OBS: If Θ is an open interval, there is typically nothing lost in imposing A1–A3 to hold for al $\theta \in \Theta$. However, it is essential that we only require the domination in A4 to hold in a neighborhood of θ_0.

Theorem 16 (Cramér-Rao lower bound on variance)

Let Assumptions A1–A4 be satisfied. Then, for any estimator $\delta : \mathcal{X} \to \mathbb{R}$ *satisfying* $\int_{\mathcal{X}} \delta^2(x) p_{\theta_0}(x) \mu(dx) < \infty$, *with*

$$E_\theta \delta = g(\theta) + b(\theta) \qquad \text{for all } \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon),$$

with g *and* b *differentiable at* θ_0, *it holds that*

$$Var_{\theta_0}(\delta) \geq \frac{[g'(\theta_0) + b'(\theta_0)]^2}{I(\theta_0)}.$$

Proof

We must show that at θ_0

$$\frac{d}{d\theta}\int_{\mathcal{X}} p_\theta(x)\mu(dx) = \int_{\mathcal{X}} \frac{d}{d\theta} p_\theta(x)\mu(dx)$$

and

$$\frac{d}{d\theta}\int_{\mathcal{X}} \delta(x)p_\theta(x)\mu(dx) = \int_{\mathcal{X}} \frac{d}{d\theta}\left[\delta(x)p_\theta(x)\right]\mu(dx), \tag{8}$$

for any δ satisfying $E_\theta \delta^2 < \infty$. Since $\delta \equiv 1$ clearly satisfies this, we need only establish (8). Furthermore, since $A := \left\{x \in \mathcal{X} : p_\theta(x) > 0\right\}$

$$\int_{\mathcal{X}} \delta(x)p_\theta(x)\mu(dx) = \int_{A} \delta(x)p_\theta(x)\mu(dx)$$

such that it suffices to show that

$$\frac{d}{d\theta}\int_{A} \delta(x)p_\theta(x)\mu(dx) = \int_{A} \frac{d}{d\theta}\left[\delta(x)p_\theta(x)\right]\mu(dx), \tag{9}$$

at θ_0.

To allow interchanging differentiation and integration in (9), we must exhibit[8] a function $g \geq 0$ such that $|\delta(x)\frac{d}{d\theta}p_\theta(x)| \leq g(x)$ for μ-almost all $x \in A$ for all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ and such that

$$\int_A g(x)\mu(dx) < \infty.$$

To this end recall that, by assumption, there exists a function $h \geq 0$ with the property $|\frac{d}{d\theta}p_\theta(x)| \leq h(x)$ for all $x \in A$ and all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$. Hence,

$$\left|\delta(x)\frac{d}{d\theta}p_\theta(x)\right| \leq |\delta(x)|h(x) \qquad \text{for all } x \in A,$$

and all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$.

[8] This result follows from the dominated convergence theorem applied to the difference quotient, and noting that the condition to the dominating function can be localized (Billingsley, 1986, Theorem 16.8).

By the Cauchy-Schwarz inequality

$$\begin{aligned}\int_A |\delta(x)|h(x)\mu(dx) &= \int_A |\delta(x)|p_{\theta_0}^{1/2}(x)\frac{h(x)}{p_{\theta_0}^{1/2}(x)}\mu(dx) \\ &= \int_{\mathcal{X}} |\delta(x)|p_{\theta_0}^{1/2}(x)\mathbb{1}_A(x) \cdot \frac{h(x)}{p_{\theta_0}^{1/2}(x)}\mathbb{1}_A(x)\mu(dx) \\ &\leq \left(\int_A \delta^2(x)p_{\theta_0}(x)\mu(dx) \cdot \int_A \frac{h^2(x)}{p_{\theta_0}(x)}\mu(dx)\right)^{1/2} \\ &< \infty.\end{aligned}$$

Thus, we can use $g = |\delta|h$.

We may skip the remaining slides. They basically establish that the regularity conditions A1–A4 are satisfied for one-parameter exponential families.

The slides also contain some material on how to calculate the moments of the sufficient statistic in exponential families.

Cramér-Rao and Exponential families

- In order for the Cramér-Rao bound to apply we need that

$$\int_{\mathcal{X}} \frac{h^2(x)}{p_{\theta_0}(x)} \mu(dx) < \infty,$$

- for h as specified in the theorem.
- This may not look helpful.
- Let us see, however, that for exponential families this condition is satisfied.

Lemma 17 (Differentiability)

Given measurable functions $f : \mathcal{X} \to \mathbb{R}$, $h : \mathcal{X} \mapsto \mathbb{R}$ *and* $T : \mathcal{X} \mapsto \mathbb{R}$ *denote by*

$$A_f = \left\{ \eta \in \mathbb{R} : \int_{\mathcal{X}} |f(x)| e^{\eta T(x)} h(x) \mu(dx) < \infty \right\}.$$

Then

$$\eta \mapsto \int_{\mathcal{X}} f(x) e^{\eta T(x)} h(x) \mu(dx)$$

is infinitely differentiable on the interior of A_f *and*

$$\frac{d^m}{d\eta^m} \int_{\mathcal{X}} f(x) e^{\eta T(x)} h(x) \mu(dx) = \int_{\mathcal{X}} T^m(x) f(x) e^{\eta T(x)} h(x) \mu(dx).$$

Proof

Define $F(\eta) := \int_{\mathcal{X}} f(x)e^{\eta T(x)}h(x)\mu(dx)$ and let $\eta_0 \in A_f^0$ (the interior of A_f). There exists an $\varepsilon > 0$ such that $[\eta_0 - \varepsilon, \eta_0 + \varepsilon] \subseteq A_f$ and hence $F(\eta)$ exists and is finite for all $\eta \in [\eta_0 - \varepsilon, \eta_0 + \varepsilon]$. Let $c_n \to 0$. Then, eventually $\eta_0 + c_n \in [\eta_0 - \varepsilon, \eta_0 + \varepsilon]$ and

$$\frac{F(\eta_0 + c_n) - F(\eta_0)}{c_n} = \int_{\mathcal{X}} \frac{e^{c_n T(x)} - 1}{c_n} f(x)e^{\eta_0 T(x)}h(x)\mu(dx).$$

Clearly $\frac{e^{c_n T(x)}-1}{c_n} \to T(x)$ for all $x \in \mathcal{X}$ and, by the dominated convergence theorem, it thus suffices to find an integrable majorant of $\left|\frac{e^{c_n T(x)}-1}{c_n} f(x)e^{\eta_0 T(x)}h(x)\right|$ not depending on c_n in order to verify that

$$F'(\eta_0) = \int_{\mathcal{X}} T(x)f(x)e^{\eta T(x)}h(x)\mu(dx).$$

To this end, observe that (since eventually $|c_n| \leq \varepsilon$)

$$\begin{aligned}
\left|\frac{e^{c_n T(x)} - 1}{c_n}\right| &= \left|\sum_{i=1}^{\infty} \frac{(c_n T(x))^i}{c_n \cdot i!}\right| \\
&\leq \sum_{i=1}^{\infty} \frac{|c_n|^{i-1} |T(x)|^i}{i!} \\
&\leq \sum_{i=1}^{\infty} \frac{\varepsilon^i |T(x)|^i}{\varepsilon \cdot i!} \\
&= \frac{e^{\varepsilon |T(x)|}}{\varepsilon} \\
&\leq \frac{e^{\varepsilon T(x)} + e^{-\varepsilon T(x)}}{\varepsilon}.
\end{aligned}$$

It follows that

$$\left|\frac{e^{c_n T(x)}-1}{c_n}f(x)e^{\eta_0 T(x)}h(x)\right| \le \frac{e^{\varepsilon T(x)}+e^{-\varepsilon T(x)}}{\varepsilon}|f(x)|e^{\eta_0 T(x)}h(x)$$

is nor greater than

$$\frac{e^{(\eta_0+\varepsilon)T(x)}}{\varepsilon}|f(x)|h(x)+\frac{e^{(\eta_0-\varepsilon)T(x)}}{\varepsilon}|f(x)|h(x),$$

which is integrable since $(\eta_0+\varepsilon)$ and $(\eta_0-\varepsilon)$ are elements of A_f. Hence, since $\eta_0 \in A_f^0$ was arbitrary

$$F'(\eta)=\int_{\mathcal{X}} T(x)f(x)e^{\eta T(x)}h(x)\mu(dx) \qquad \text{for all } \eta \in A_f^0. \quad (10)$$

Using $f(x)T(x)$ in lieu of $f(x)$ we can now successively repeat the above arguments starting with (10) to get the desired result. Alternatively, use induction. □

Using $f \equiv 1$ in the above differentiability lemma one has that in the exponential family (in natural form) with μ-densities

$$p_\eta(x) = e^{\eta T(x) - B(\eta)} h(x),$$

it holds that

$$B(\eta) = \log \left(\int_{\mathcal{X}} e^{\eta T(x)} h(x) \mu(dx) \right)$$

is infinitely differentiable on the interior of the natural parameter space. In particular, $|B(\eta)|$ is bounded on any compact subset of the natural parameter space.

Moments of sufficient statistic in the exponential family

Since

$$e^{B(\eta)} = \int_{\mathcal{X}} e^{\eta T(x)} h(x) \mu(dx)$$

for any η in the natural parameter space, one has that

$$\begin{aligned} \frac{d^m}{d\eta^m} e^{B(\eta)} &= \frac{d^m}{d\eta^m} \int_{\mathcal{X}} e^{\eta T(x)} h(x) \mu(dx) \\ &= \int_{\mathcal{X}} T^m(x) e^{\eta T(x)} h(x) \mu(dx), \end{aligned}$$

in the interior of the natural parameter space. Hence, for $m \in \mathbb{N}$,

$$e^{-B(\eta)} \frac{d^m}{d\eta^m} e^{B(\eta)} = E_\eta T^m.$$

In particular, for $m = 1, 2$, respectively

$$E_\eta T = B'(\eta) \quad \text{and} \quad E_\eta T^2 = B''(\eta) + [B'(\eta)]^2,$$

such that

$$Var_\eta T = B''(\eta).$$

We can now (finally) verify that exponential families satisfy Assumptions A1–A4 at any θ_0 such that $\eta_0 := \eta(\theta_0) \neq 0$ is in the interior, A^0, of the natural parameter space $A := \left\{\eta \in \mathbb{R} : \int_{\mathcal{X}} e^{\eta T(x)} h(x)\mu(dx) < \infty\right\}$. Let η be continuously differentiable and write

$$p_{\eta(\theta)}(x) = e^{\eta(\theta)T(x) - B(\eta(\theta))} h(x)$$

A1 and A2 are clearly satisfied. Since

$$\frac{d}{d\theta} \log(p_\eta(\theta)) = [T(x) - B'(\eta(\theta))] \cdot \eta'(\theta)$$

it follows that

$$I(\theta_0) = E_{\theta_0}[T(x) - B'(\eta(\theta_0))]^2 \cdot [\eta'(\theta)]^2 = Var_{\eta(\theta_0)}(T) \cdot [\eta'(\theta_0)]^2 > 0.$$

and so A3 is satisfied.

- It follows by the chain rule and the fact that $\theta \mapsto \eta'(\theta)$ is bounded on compact sets [by its continuity] that it is enough to verify A4 in the natural parameter space A.
- That is we show that there exists a measurable function $h : \mathcal{X} \to [0, \infty)$ and $\varepsilon > 0$ such that $|\frac{d}{d\theta} p_\eta(x)| \leq h(x)$ for all $x \in A$ and all $\eta \in (\eta_0 - \varepsilon, \eta_0 + \varepsilon)$ and $\int_A \frac{h^2(x)}{p_{\eta_0}(x)} \mu(dx) < \infty$.

Note that

$$\begin{aligned} |p'_\eta(x)| &= \left| [T(x) - B(\eta)] e^{\eta T(x) - B(\eta)} h(x) \right| \\ &= |T(x)| e^{\eta T(x) - B(\eta)} h(x) + |B'(\eta)| e^{\eta T(x) - B(\eta)} h(x) \end{aligned}$$

Since $\eta_0 \in A^0$, there exists an $\varepsilon > 0$ such that $[\eta_0 - 2\varepsilon, \eta_0 + 2\varepsilon] \subseteq A^0$.

Furthermore, since $B(\eta)$ is bounded on $[\eta_0 - \varepsilon, \eta_0 + \varepsilon]$, there exists a $c > 0$ such that for all

$$|T(x)|e^{\eta T(x) - B(\eta)} h(x) \leq c|T(x)| \left(e^{(\eta_0 - \varepsilon)T(x)} + e^{(\eta_0 + \varepsilon)T(x)}\right) h(x),$$

and similarly

$$|B'(\eta)|e^{\eta T(x) - B(\eta)} h(x) \leq c \left(e^{(\eta_0 - \varepsilon)T(x)} + e^{(\eta_0 + \varepsilon)T(x)}\right) h(x).$$

Thus, noting that the right-hand sides of the previous two displays do not depend on η, we can use the sum of these as our candidate for h.

We now verify the integrability condition:

Using that $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$,

$$\begin{aligned} \frac{h^2(x)}{p_{\eta_0}} &\leq 2c^2 e^{B(\eta_0)} T^2(x) \left(e^{(\eta_0 - 2\varepsilon) T(x)} + e^{(\eta_0 + 2\varepsilon) T(x)} \right) h(x) \\ &\quad + 2c^2 e^{B(\eta_0)} \left(e^{(\eta_0 - 2\varepsilon) T(x)} + e^{(\eta_0 + 2\varepsilon) T(x)} \right) h(x), \end{aligned}$$

which is integrable since $[\eta_0 - 2\varepsilon, \eta_0 + 2\varepsilon] \subseteq A^0$.

References

Bickel, P. and K. Doksum (2015): *Mathematical Statistics: basic ideas and selected topics*, vol. 1, CRC Press.

Billingsley, P. (1986): *Probability and Measure*, Wiley, 2nd ed.

Cox, D. R. and D. V. Hinkley (1974): *Theoretical Statistics*, Chapman & Hall.

Klenke, A. (2020): *Probability Theory: a comprehensive course*, Springer Science & Business Media.

Lehmann, E. and G. Casella (1998): *Theory of Point Estimation*, Springer.

Lehmann, E. and J. Romano (2005): *Testing Statistical Hypotheses*, Springer.

Liese, F. and K.-J. Miescke (2008): *Statistical Decision Theory*, Springer.

Pfanzagl, J. (1994): *Parametric Statistical Theory*, Walter de Gruyter.

Romano, J. and A. Siegel (1986): *Counterexamples in Probability and Statistics*, Wadsworth & Brooks Cole.