Introduction to Mathematical Statistics

Pavel Chigansky

Department of Statistics, The Hebrew University, Mount Scopus, Jerusalem 91905, Israel

E-mail address: pchiga@mscc.huji.ac.il

Preface

These are the lecture notes for the courses "Statistical Inference and Applications A" (52303)/ "Theory of Statistics" (52314) I have taught at the Statistics Department of the Hebrew University during the fall semesters of 2009-2012 .

The course is divided into two parts: the introduction to multivariate probability theory and introduction to mathematical statistics. The probability is served on an elementary (non-measure theoretic) level and is similar in the spirit to the corresponding part of Casella & Berger text [**3**] ([**14**] is my choice for a deeper insight into probability theory).

The statistical part is in the spirit of Bickel & Doksum [**1**], Casella & Berger [**3**]. For an in depth reading the comprehensive classical texts are recommended: Lehmann & Casella [**8**], Lehmann & Romano [**9**], Borovkov[**2**]. The text of Shao [**12**] (comes with the solutions guide [**13**] to the exercises) is highly recommended, if your probability is already measure theoretic. The book of Ibragimov and Khasminskii [**6**] is an advanced treatment of the asymptotic theory of estimation (both parametric and non-parametric). A.Tsybakov's [**15**] is an excellent text, focusing on the nonparametric estimation. Finally, the papers [**16**] and [**7**] contain the proofs of some facts, mentioned in the text.

Please do not hesitate to e-mail your comments/bug reports to the author.

P.Ch., 21/01/2013

Contents

Part 1

Probability

CHAPTER 1

Probabilistic description of a single random variable

a. Probability space

In probability theory the outcomes of experiments are identified with points in a set Ω, called the *sampling space*. Subsets of Ω are called events and an event A is said to have occurred, if the realized $\omega \in \Omega$ belongs to A. Probability measure $\mathbb{P}$ is a function which assigns numbers in $[0,1]$ to events. It is required to satisfy the normalization property $\mathbb{P}(\Omega) = 1$ and additivity[1]:

$$A \cap B = \emptyset \quad \Longrightarrow \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

In words, this means that if A and B are mutually exclusive, the probability of either A or B to occur is the sum of their individual probabilities.

For technical reasons, beyond our scope, one cannot define many natural probability measures on all the subsets of Ω, if it is uncountable, e.g., $\Omega = \mathbb{R}$. Luckily it can be defined on a rich enough collection of subsets, which is denoted by $\mathcal{F}$ and called the σ-algebra[2] of events. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called *probability space*. In general construction of probability measures on $(\Omega, \mathcal{F})$ can be a challenging mathematical problem, depending on the complexity of Ω (think e.g. about Ω consisting of all continuous functions).

Probability measures on $\Omega := \mathbb{R}^d$ can always be defined by the *cumulative distribution functions* (or c.d.f. in short). Hence in this course, a probability measure will be always identified with the corresponding c.d.f. For $d = 1$, a function $F : \mathbb{R} \mapsto [0,1]$ is a legitimate c.d.f. if and only if it satisfies the following properties

(i) F is a nondecreasing function

(ii) $\lim_{x\to\infty} F(x) = 1$

(iii) $\lim_{x\to-\infty} F(x) = 0$

(iv) F is right continuous

The probability measure of semi-infinite intervals is defined by the formulas

$$\mathbb{P}\big((-\infty, x]\big) := F(x), \quad \text{and} \quad \mathbb{P}\big((-\infty, x)\big) := \lim_{\varepsilon \searrow 0} F(x-\varepsilon) =: F(x-)$$

and is extended to other types of intervals or their finite (or countable) unions through additivity. For example, since $(-\infty, a] \cup (a, b] = (-\infty, b]$ for $a < b$, by additivity

$$\mathbb{P}\big((-\infty, a]\big) + \mathbb{P}\big((a, b]\big) = \mathbb{P}\big((-\infty, b]\big)$$

[1] in fact, a stronger property of σ-additivity is required, but again this is way beyond our scope.

[2] the term σ-algebra comes from the fact that $\mathcal{F}$ is to be closed under taking compliments and *countable* unions or intersections

and hence

$$\mathbb{P}\big((a,b]\big) = F(b) - F(a). \tag{1a1}$$

Similarly, probabilities of other intervals, e.g. $[a,b)$, $[a,b]$, etc. or unions of intervals are defined. In fact, this assignment defines $\mathbb{P}$ on subsets (events), much more general than intervals, but we shall rarely need to calculate such probabilities explicitly. These definitions explain the need for the conditions (i)-(iv).

The c.d.f. is said to be purely discrete (atomic) if it is a piecewise constant function with jumps at $x_k \in \mathbb{R}$, $k \in \mathbb{N}$ with the corresponding sizes $\{p_k\}$:

$$F(x) = \sum_{k:x_k \le x} p_k, \quad x \in \mathbb{R}. \tag{1a2}$$

In this case,

$$\mathbb{P}(\{x_k\}) = F(x_k) - F(x_k-) = p_k,$$

i.e. each value x_k is assigned a positive probability p_k (check that (i)-(iv) imply $p_k > 0$ and $\sum_k p_k = 1$). Often $\{x_k\} = \mathbb{N}$, in which case $\{p_k\}$ is referred to as *probability mass function* (p.m.f.) Here are some familiar examples

Example 1a1 (Bernoulli distribution). Bernoulli distribution with parameter $p \in (0,1)$, denoted $\mathrm{Ber}(p)$, has the c.d.f.

$$F(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ 1-p & x \in [0,1) \\ 1 & x \in [1,\infty) \end{cases}$$

This function satisfies (i)-(iv) (sketch a plot and check). We have

$$\mathbb{P}(X=0) = F(0) - F(0-) = 1-p, \quad \text{and} \quad \mathbb{P}(X=1) = F(1) - F(1-) = 1-(1-p) = p.$$

For any $y \notin \{0,1\}$,

$$\mathbb{P}(X=y) = F(y) - F(y-) = 0,$$

since $F(y)$ is continuous at those y's. ■

Example 1a2 (Poisson distriution). The Poisson distribution with rate parameter $\lambda > 0$ has piecewise constant c.d.f with jumps at $\{0,1,...\} = \{0\} \cup \mathbb{N} =: \mathbb{Z}_+$ and its p.m.f. is given by

$$\mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{Z}_+$$

Hence e.g.

$$P(X \in [3, 17.5]) = F(17.5) - F(3-) = F(17) - F(3-) = \sum_{k=3}^{17} \frac{e^{-\lambda}\lambda^k}{k!}.$$

The latter expression can be evaluated numerically if λ is given as a number (rather as a symbolic variable). Also

$$F(x) = \sum_{k:k \le x} \frac{e^{-\lambda}\lambda^k}{k!}.$$

■

Other examples of discrete distributions are Geometric, Hypergeometric, Binomial, etc.

If F has the form

$$F(x) = \int_{-\infty}^{x} f(u)du \tag{1a3}$$

for some function f, it is said to have density f (check that (i)-(iv) imply $\int_{\mathbb{R}} f(u)du = 1$ and that $f(u) \geq 0$, if it is continuous). Of course, in general c.d.f may increase in other ways, e.g. to have both jumps and continuous parts.

Note that a continuous c.d.f. assigns zero probability to each point in Ω. Does this imply that we cannot get any particular outcome? Certainly not, since $\mathbb{P}(\Omega) = 1$. This seemingly paradoxical situation is quite intuitive: after all, drawing $1/2$ or any other number from the interval $[0, 1]$ with the uniform distribution on it feels like impossible. Mathematically, this is resolved by means of *the measure theory*, which is the way probability is treated rigorously[3].

EXAMPLE 1a3. Normal (Gaussian) c.d.f. $N(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ has the density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad x \in \mathbb{R}.$$

■

Other frequently encountered p.d.f's are Exponential, Cauchy, Gamma, etc.

b. Random variable

Functions on the sampling space are called *random variables*. We shall denote random variables by capital letters to distinguish them from the values they take. A random variable $X : \Omega \mapsto \mathbb{R}$ generates events of the form $\{\omega \in \Omega : X(\omega) \in A\}$, where A is a subset of $\mathbb{R}$, e.g. an interval. The function

$$F_X(x) := \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \in (-\infty, x]\}\big)$$

is called the c.d.f. of the random variable X. It defines the probabilities of the events, generated by X, similarly to (1a1). In fact, for the *coordinate* random variables $X(\omega) := \omega$, the c.d.f. of X coincides with the c.d.f., which defines $\mathbb{P}$:

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(\{\omega \in \Omega : \omega \leq x\}) = F(x). \tag{1b1}$$

For other random variables, F_X is always a c.d.f., but the connection between F_X and F can be more complicated.

EXAMPLE 1b1. One simple yet important example of r.v. is the *indicator* of an event A:

$$I(A) := \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^c. \end{cases}$$

Since $I(A) \in \{0, 1\}$, it is in fact a Bernoulli r.v. with parameter $\mathbb{P}(I(A) = 1) = \mathbb{P}(A)$.

It is customary in probability theory not to elaborate the structure of the function $X(\omega)$ and even omit ω from the notations, but just specify the corresponding c.d.f. F_X. This is sufficient in many problems, since F_X defines probabilities of all the events, generated by X, regardless of the underlying probability space.

[3] consult books, if you are curious. My favorite text on the subject is [**14**]

A random variable X is said to be discrete (atomic) or to have a density according to the type of its c.d.f. Also random variables inherit names from their distributions, e.g., $X \sim \mathrm{Poi}(\lambda)$ means that the c.d.f. of X is $\mathrm{Poi}(\lambda)$.

Here is a simple example of a random variable, whose c.d.f have both discrete and continuous parts:

EXAMPLE 1b2. Let $X \sim U([0,1])$ and set $Y = \max(1/2, X)$. Clearly Y takes values in $[1/2, 1]$. Hence for $x < 1/2$,

$$F_Y(x) = \mathbb{P}(Y \le x) = 0.$$

Further,

$$F_Y(1/2) = \mathbb{P}(Y \le 1/2) = \mathbb{P}(X \le 1/2) = 1/2$$

and for $x > 1/2$,

$$F_Y(x) = \mathbb{P}(Y \in [1/2, x]) = \mathbb{P}(X \in [0, x]) = x.$$

To summarize,

$$F_Y(x) = \begin{cases} 0, & x < 1/2 \\ 1/2, & x = 1/2 \\ x, & x \in (1/2, 1] \\ 1, & x > 1 \end{cases}$$

Hence Y has an atom at $\{1/2\}$ and a continuous nontrivial part. ■

c. Expectation

Expectation of a random variable is averaging over all its possible realizations. More precisely, given a function g, defined on the range of the r.v. X

$$\mathbb{E}g(X) := \sum_i g(x_i)\mathbb{P}(X = x_i) = \sum_i g(x_i)\big(F_X(x_i) - F_X(x_i-)\big) =$$

$$\sum_i g(x_i)\Delta F_X(x_i) =: \int_{\mathbb{R}} g(x)dF_X(x),$$

if X is discrete, and

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)f_X(x)dx =: \int_{\mathbb{R}} g(x)dF_X(x),$$

if X has p.d.f. $f_X(x)$. Note that in the two cases the notation $\int_{\mathbb{R}} g(x)dF_X(x)$ is interpreted differently, depending on the context[4].

For particular functions $g(\cdot)$, expectation has special names: e.g. $\mathbb{E}X$ is the *mean* of X, $\mathbb{E}X^p$ is the p-th moment of X, $\mathbb{E}(X - \mathbb{E}X)^2$ is the *variance* of X, etc.

EXAMPLE 1c1. For[5] $X \sim \mathrm{Poi}(\lambda)$,

$$\mathbb{E}X = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = ... = \lambda,$$

[4] you can think of $dF_X(x) = f_X(x)dx$ if $F_X'(x) = f_X(x)$ and $dF(x) = \Delta F_X(x)$ if $F_X(x)$ has a jump at x. However, do not think that this is just a caprice of notations: in fact, $\int g(x)dF(x)$ makes perfect sense as the Lebesgue integral.

[5] "= ... =" means that an obvious calculation is omitted

and for $X \sim N(\mu, \sigma^2)$

$$\mathbb{E}X = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = ... = \mu,$$

and

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X-\mu)^2 = \int_{\mathbb{R}} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = ... = \sigma^2$$

■

Being an integral (or series), expectation of $g(X)$ does not have to be finite or even well defined:

EXAMPLE 1c2. Recall that X is a Cauchy r.v. if it has the density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad x \in \mathbb{R}.$$

Consider $g(x) := |x|^{1/2}$, then

$$\mathbb{E}|X|^{1/2} = \int_{\mathbb{R}} \frac{|x|^{1/2}}{\pi}\frac{1}{1+x^2} dx := \lim_{N\to\infty} \int_{-N}^{N} \frac{|x|^{1/2}}{\pi}\frac{1}{1+x^2} dx = ... = \pi/\sqrt{2}.$$

Now let $g(x) = |x|$:

$$\mathbb{E}|X| = \int_{\mathbb{R}} \frac{|x|}{\pi}\frac{1}{1+x^2} dx := \lim_{N\to\infty} \int_{-N}^{N} \frac{|x|}{\pi}\frac{1}{1+x^2} dx = \infty,$$

since the function $x/(1+x^2)$ behaves as $1/x$ for large x, whose integral behaves as $\log N$. Notice $\mathbb{E}|X|$ would be the same if we used a different limiting procedure in the above improper integral, e.g. if instead of $-N$ we would have taken $-2N$.

Something different happens, if we try to calculate $\mathbb{E}X$: the choice of the upper and lower integration limits before going to ∞ changes the value of the integral completely. For example,

$$\lim_{N\to\infty} \int_{-N}^{N} \frac{x}{\pi}\frac{1}{1+x^2} dx = 0,$$

since the function is antisymmetric and

$$\lim_{N\to\infty} \int_{-2N}^{N} \frac{x}{\pi}\frac{1}{1+x^2} dx = \lim_{N\to\infty} \int_{-2N}^{-N} \frac{x}{\pi}\frac{1}{1+x^2} dx = ... = -\infty.$$

Hence $\mathbb{E}X$ cannot be *defined* in this case unambiguously. If we still insist on defining $\mathbb{E}X$ e.g. by choosing the symmetric case as above, the emerging object may not satisfy the usual properties of expectations (see below), which limits severely its usefulness. ■

For non-negative (non-positive) r.v. the expectation is always defined, possibly taking infinite value. Furthermore, the expectation of a r.v. X is well defined and finite, if $\mathbb{E}|X| < \infty$.

Expectation satisfies a number of useful and important properties, including

(1) "*monotonicity*"[6]

$$X \geq 0 \quad \Longrightarrow \quad \mathbb{E}X \geq 0.$$

(2) $\mathbb{E}c = c$ for a constant c (note that a constant can be viewed as a r.v. as well)

[6]make sure you understand what $X \geq 0$ means when X is a r.v.

We shall recall more properties as the story unfolds.

d. Moment generating function (m.g.f.)

As we have recalled above, probability distribution of a r.v. is determined by c.d.f. or, equivalently, by p.d.f. or by p.m.f., when they exist. It turns out that there is a useful alternative characterization of a distribution [7]:

DEFINITION 1d1. *Let X be a r.v. such that $\mathbb{E}e^{\delta|X|} < \infty$ for some $\delta > 0$, then the m.g.f. of X is*

$$M_X(t) := \mathbb{E}e^{tX}, \quad t \in (-\delta, \delta).$$

EXAMPLE 1d2. Let $X \sim \text{Exp}(\lambda)$. Then

$$\mathbb{E}e^{\delta|X|} = \mathbb{E}e^{\delta X} = \int_0^\infty e^{\delta x}\lambda e^{-\lambda x}dx < \infty,$$

if $\delta < \lambda$. Hence

$$M_X(t) = \mathbb{E}e^{tX} = \int_0^\infty \lambda e^{(t-\lambda)x}dx = ... = (1 - t/\lambda)^{-1}, \quad t \in (-\lambda, \lambda).$$

■

EXAMPLE 1d3. For the Cauchy r.v. the m.g.f. is not defined: for all $\delta > 0$, $\mathbb{E}e^{\delta|X|} = \infty$. ■

Note that if m.g.f. is well defined on an open interval $(-\delta, \delta)$ near the origin, then X has finite absolute moments (and hence the moments) of all orders:

$$\mathbb{E}|X|^p = \int_{\mathbb{R}} |x|^p dF(x) \le \int_{\mathbb{R}} Ce^{\delta/2|x|}dF(x) = C\mathbb{E}e^{\delta/2|X|} < \infty,$$

since $|x|^p \le Ce^{\delta|x|}$ for all $x \in \mathbb{R}$ and sufficiently large C.

One can show that the function $M_X(t)$ is smooth (i.e. differentiable any number of times). Moreover[8],

$$M'_X(0) := \frac{d}{dt}M_X(t)\Big|_{t=0} = \frac{d}{dt}\int_{\mathbb{R}} e^{tx}dF_X(x)\Big|_{t=0} \overset{\dagger}{=} \int_{\mathbb{R}} \frac{d}{dt}e^{tx}dF_X(x)\Big|_{t=0} =$$
$$\int_{\mathbb{R}} xe^{tx}dF_X(x)\Big|_{t=0} = \int_{\mathbb{R}} xdF_X(x) = \mathbb{E}X,$$

where the interchanging the derivative and integral in † should and can be justified (think how).

Similarly, for $k \ge 1$,

$$M_X^{(k)}(0) = \mathbb{E}X^k,$$

which is where the name of $M_X(t)$ comes from: if one knows $M_X(t)$, one can reconstruct (generate) all the moments.

---

[7] another characterization of the probability distribution of r.v. X is the *characteristic function* $\varphi_X(t) := \mathbb{E}e^{itX}$, where i is the imaginary unit (i.e. the Fourier transform of its c.d.f.) The advantage of c.f. is that it is *always* well defined (unlike the m.g.f.). It is a powerful tool in many applications (e.g. limit theorems, etc.). You will need to know some complex analysis to discover more.

[8] recall that we interpret $\int h(x)dF_X$ differently, depending on the context - this is just a convenient unifying notation in our course

EXAMPLE 1d4. Let $X \sim \text{Exp}(\lambda)$, then

$$M'_X(0) = \frac{d}{dt}(1 - t/\lambda)^{-1}_{\Big|t=0} = ... = 1/\lambda,$$

which agrees with the calculations in the preceding example. ■

It is then very plausible that knowing all the moments is enough to be able to reconstruct $F_X(x)$. Indeed the m.g.f., when it exists, completely determines [9] $F_X(x)$ and consequently the probability law of X.

While in principle $F_X(x)$ can be reconstructed from $M_X(t)$, the reconstruction formula is quite complicated and is beyond our scope. However in many cases, $M_X(t)$ can be recognized to have a particular form, which identifies the corresponding $F_X(x)$. This simple observation, as we shall see, is quite powerful.

While m.g.f. is defined for r.v. of any type, it is more convenient to deal with *probability generating function* (p.g.f.) if X is integer valued:

DEFINITION 1d5. *The p.g.f. of a discrete r.v. X with values in $\mathbb{Z}_+$ is*

$$G_X(t) := \mathbb{E}t^X = \sum_k t^k \mathbb{P}(X = k), \quad t \in D,$$

where D is the domain of convergence of the series.

Note that $G_X(t)$ is always well defined as the series are absolutely summable:

$$\Big|\sum_k t^k \mathbb{P}(X = k)\Big| \le \sum_k |t|^k \le \frac{1}{1 - |t|} < \infty, \quad \forall t \in (-1, 1),$$

and clearly $G_X(1) = 1$.

Note that $G_X(0) = \mathbb{P}(X = 0)$ and

$$G'_X(0) := \frac{d}{dt} G_X(t)_{|t=0} = \sum_{k=1}^{\infty} k t^{k-1} \mathbb{P}(X = k)_{|t=0} = \mathbb{P}(X = 1)$$

and similarly

$$G_X^{(m)}(0) = m!\mathbb{P}(X = m) = m! p_m,$$

i.e. p.m.f. is in one-to-one correspondence with p.g.f. (and hence with c.d.f.) - p.g.f. "generates" p.m.f.

EXAMPLE 1d6. Let $X \sim \text{Geo}(p)$, then

$$G_X(t) = \sum_{k=1}^{\infty} t^k p(1 - p)^{k-1} = ... = \frac{pt}{1 - tq},$$

[9] though plausible, the moments do not determine F_X in general. The fact that there is a one-to-one correspondence between $M_X(t)$ and $F_X(x)$ stems from the theory of Fourier-Laplace transform from functional analysis

where $q = 1 - p$. Then

$$\mathbb{P}(X = 0) = G_X(0) = 0$$
$$\mathbb{P}(X = 1) = G'_X(0) = \frac{p(1 - tq) + qpt}{(1 - tq)^2}\Big|_{t=0} = p$$
etc.

■

Note that $M_X(\ln t) = G_X(t)$ and hence p.g.f. is m.g.f. in disguise.

Exercises

PROBLEM 1.1. Let X be a r.v. with

(a) p.d.f.
$$f_X(x) = c(1 - x^2)I(x \in (-1, 1)).$$
(b) p.d.f.
$$f_X(x) = (1 - |x|)I(|x| \leq 1)$$
(c) p.m.f.
$$p_X(k) = \begin{cases} \frac{1}{N}, & x \in \{0, ..., N - 1\} \\ 0 & \text{otherwise} \end{cases}$$
(d) Poisson distribution

Answer the following questions for each one of the cases above:

(1) Find the normalization constant c
(2) Find the c.d.f.
(3) Find $\mathbb{E}X$, $\text{var}(X)$
(4) Calculate $\mathbb{P}(X \leq 1/2)$, $\mathbb{P}(1/3 \leq X \leq 1/2)$, $\mathbb{P}(-1/4 \leq X \leq 1/3)$, $\mathbb{P}(X \geq -1)$
(5) Find the m.g.f./p.g.f (and specify the domain). Use the latter to calculate $\mathbb{E}X$ and $\text{var}(X)$

PROBLEM 1.2. Check the following properties of the indicator function:

(1) $I_A \sim \text{Ber}(p)$, with $p = \mathbb{P}(A)$. In particular, $\mathbb{E}I_A = \mathbb{P}(A)$
(2) $I_{\cap_i A_i} = \prod_i I_{A_i}$
(3) $I_{\cup_i A_i} = \max_i I_{A_i}$
(4) $I_A = I_A^2$
(5) $I_{A^c} = 1 - I_A$
(6) $A \subseteq B \implies I_A \leq I_B$

PROBLEM 1.3. Let X be a discrete r.v. with integer values with the p.g.f. $G_X(s)$ and the m.g.f. $M_X(t)$. Show that

$$\frac{d}{ds}G_X(s)_{|s=1} = \mathbb{E}X \frac{d^2}{ds^2}G_X(s)_{|s=1} = \mathbb{E}X^2 - \mathbb{E}X$$

using

(1) the definition of $G_X(s)$;
(2) the relation $G_X(s) = M_X(\ln s)$

PROBLEM 1.4.

(1) Find the m.g.f. of $N(\mu, \sigma^2)$ and use it to check that $\mathbb{E}X = \mu$ and $\mathrm{var}(X) = \sigma^2$.
(2) For $X \sim N(\mu, \sigma^2)$ prove

$$\mathbb{E}(X-\mu)^p = \begin{cases} 0 & p \text{ odd} \\ \sigma^p \frac{p!}{2^{p/2}(p/2)!} & p \text{ even} \end{cases}$$

PROBLEM 1.5. Find the c.d.f. (or p.d.f./p.m.f. if appropriate) of the r.v. with the following m.g.f.'s

(1) $M_X(t) = e^{t(t-1)}$, $t \in \mathbb{R}$
(2) $M_X(t) = 1/(1-t)$, $|t| < 1$
(3) $M_X(t) = pe^t + 1 - p$, $t \in \mathbb{R}$
(4) $M_x(t) = e^{tC}$, $t, C \in \mathbb{R}$
(5) $M_X(t) = \exp(e^t - 1)$, $t \in \mathbb{R}$

PROBLEM 1.6. Show that $X \geq 0 \implies \mathbb{E}X \geq 0$

CHAPTER 2

Probabilistic description of several random variables

In majority of situations, there is a need to consider probabilities on $\mathbb{R}^d$ for $d > 1$ or, alternatively, a number of random variables *simultaneously*. The familiar example, is the usual mantra: "let $X_1, ..., X_n$ be i.i.d. random variables" or the same, with a more statistical flavor: "let $X_1, ..., X_n$ be a sample from the p.d.f. $f(x)$". What is actually meant here, is that we obtain values (realizations) of the r.v.'s $(X_1, ..., X_n)$, which are, in this case, independent (whatever it means at this point) and have the same probability law. But how do we describe several r.v.'s simultaneously which are dependent or/and not identically distributed, etc.?

a. A pair of random variables

A probability measure $\mathbb{P}$ on $\Omega := \mathbb{R}^2$ is assigned by a two-dimensional c.d.f F. Probabilistic description of a pair of random variables (X, Y) on Ω is accomplished by their *joint* c.d.f.

$$F_{XY}(u, v) := \mathbb{P}\big(X \le x, Y \le y\big).$$

Analogously to the one-dimensional case, the c.d.f. of the coordinate random variables

$$\big(X(\omega), Y(\omega)\big) = (\omega_1, \omega_2) = \omega$$

coincides with the c.d.f., defining the probability measure $\mathbb{P}$. Hence defining a c.d.f. on $\mathbb{R}^2$ and construction of a random vector (X, Y) with values in $\mathbb{R}^2$ and a given joint c.d.f. is essentially the same problem. Below we shall refer to random variables, when exploring the properties of two-dimensional c.d.f's.

Clearly a c.d.f. must satisfy[1] $F_{XY}(-\infty, v) = F_{XY}(u, -\infty) = 0$ for all $u, v \in \mathbb{R}$ and $F_{XY}(\infty, \infty) = 1$. These properties are parallel to the familiar requirements $F_X(-\infty) = 0$ and $F_X(\infty) = 1$ in dimension one. $F_{XY}(u, v)$ must be right-continuous in each one of the coordinates, hence e.g.

$$\mathbb{P}(X \le x, Y < y) = F_{XY}(x, y-).$$

The monotonicity is also valid in $\mathbb{R}^2$, but should be interpreted correctly. In particular, the condition $\mathbb{P}(X \in (a_1, b_1], Y \in (a_2, b_2]) \ge 0$, implies (think how)

$$\Delta_{a_1,b_1}\Delta_{a_2,b_2}F_{XY} \ge 0, \quad \forall a_1 \le b_1, a_2 \le b_2, \tag{2a1}$$

where for a function h of n variables (in this case $n = 2$),

$$\Delta_{a_i,b_i}h(x_1, ..., x_n) := h(x_1, ..., b_i, ..., x_n) - h(x_1, ..., a_i, ..., x_n).$$

Hence a c.d.f should satisfy (2a1) and it turns to be not only necessary, but also the sufficient condition.

[1] $F_{XY}(-\infty, v) := \lim_{u\to-\infty} F_{XY}(u, v)$, etc.

It is already clear at this point that individual c.d.f.'s of each component can be found from F_{XY}, e.g.:

$$F_X(u) = \mathbb{P}(X \le u) = \mathbb{P}(X \le u, Y \in \mathbb{R}) = F_{XY}(u, \infty).$$

However, in general one cannot restore F_{XY} from its *marginals* F_X and F_Y (see Example 2a6 below). Thus F_{XY} is a more complete probabilistic description of (X, Y).

All of these properties are automatically satisfied in the two particular cases: the *jointly* discrete and *jointly* continuous.

***.** Jointly discrete random variables

Definition 2a1. *A random vector (X, Y) is (jointly) discrete if there is a countable (or finite) number of points $(x_k, y_m) \in \mathbb{R}^2$, $k \ge 1$, $m \ge 1$ and positive numbers $p_{k,m} > 0$, such that $\sum_{k,m} p_{k,m} = 1$ and*

$$\mathbb{P}\Big((X, Y) = (x_k, y_m)\Big) = \mathbb{P}\big(X = x_k, Y = y_m\big) := p_{k,m}.$$

The numbers $p_{k,m}$ (along with the points (x_k, y_m)) are called *joint* p.m.f. of (X, Y). Note that this definition yields an explicit expression for the j.c.d.f:

$$F_{XY}(u, v) = \sum_{k: x_k \le u} \sum_{m: y_m \le v} p_{k,m},$$

and for a subset $A \in \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in A) = \sum_{k,m: (x_k, y_m) \in A} p_{k,m}.$$

In particular,

$$F_{XY}(u, \infty) = \mathbb{P}\Big((X, Y) \in (-\infty, u] \times \mathbb{R}\Big) = \sum_{k: x_k \le u} \sum_{m: y_m \in \mathbb{R}} p_{k,m} = \sum_{k: x_k \le u} p_X(k) = F_X(u),$$

where $p_X(k) := \sum_m p_{k,m}$ is the p.m.f. of X. Similarly, we get $F_{XY}(\infty, v) = F_Y(v)$. This means that the c.d.f.'s of both entries of the vector (which are one dimensional random variables) are recovered from the joint c.d.f., i.e. the probabilistic description of X and Y as individual r.v., is in fact incorporated in their joint probabilistic characterization. From this point of view, $F_X(u)$ and $F_Y(v)$ are the *marginal* c.d.f's of the j.c.d.f $F_{XY}(u, v)$. Similarly, $p_X(k)$ and $p_Y(m)$ are the *marginal* p.m.f's of the j.p.m.f. $p_{k,m}$ (or in other popular notation, $p_{XY}(k, m)$).

Example 2a2. Consider the random vector (X, Y) with values in $\mathbb{N}^2$ and j.p.m.f.

$$p_{k,m} = (1-p)^{k-1} p (1-r)^{m-1} r, \quad k, m \in \mathbb{N},$$

where $p, r \in (0, 1)$. Let's check that it is indeed a legitimate j.p.m.f.: clearly $p_{k,m} \ge 0$ and

$$\sum_{k,m} p_{k,m} = \sum_k (1-p)^{k-1} p \sum_m (1-r)^{m-1} r = \ldots = 1.$$

The p.m.f. of X (or X-marginal p.m.f. of $p_{k,m}$) is

$$p_k = \sum_m p_{k,m} = (1-p)^{k-1} p \sum_m (1-r)^{m-1} r = (1-p)^{k-1} p, \quad k \in \mathbb{N},$$

i.e. $X \sim \mathrm{Geo}(p)$. Similarly, $Y \sim \mathrm{Geo}(r)$. ■

Note that two different j.c.d.f's may have identical marginals (see also the Example 2a6 below)

EXAMPLE 2a3. Let (X, Y) be a r.v. taking values in

$$\{(0,0),(0,1),(1,0),(1,1)\}$$

with the j.p.m.f.:

$$\begin{aligned} p_{XY}(0,0) &= 1/4 - \varepsilon \\ p_{XY}(0,1) &= 1/4 + \varepsilon \\ p_{XY}(1,0) &= 1/4 + \varepsilon \\ p_{XY}(1,1) &= 1/4 - \varepsilon \end{aligned}$$

where $\varepsilon \in (0, 1/4)$ is an arbitrary constant. Clearly, j.p.m.f depends on ε, while the marginals do not: check that $X \sim \text{Ber}(1/2)$ and $Y \sim \text{Ber}(1/2)$ irrespectively of ε. ■

If (X, Y) is a discrete random vector, then both X and Y are discrete random variables, e.g.

$$\sum_k \mathbb{P}(X = x_k) = \sum_k \sum_m \mathbb{P}(X = x_k, Y = y_m) = 1.$$

The converse is also true, i.e. if X is discrete and Y is discrete, then (X, Y) is discrete. [2]

Given a function $g : \mathbb{R}^2 \mapsto \mathbb{R}$, the expectation of $g(X, Y)$ is defined:

$$\mathbb{E}g(X,Y) := \sum_{k,m} g(x_k, y_m) p_{k,m}.$$

As in the one dimensional case, the expectation $\mathbb{E}g(X, Y)$ may take values in $\mathbb{R} \cup \{\pm\infty\}$ or may not be defined at all (recall the Example 1c2).

The expectation is linear: for $a, b \in \mathbb{R}$,

$$\mathbb{E}(aX + bY) = \sum_{k,m} (ax_k + by_m) p_{k,m} = a \sum_{k,m} x_k p_{k,m} + b \sum_{k,m} y_m p_{k,m} =$$
$$a \sum_k x_k p_X(k) + b \sum_m y_m p_Y(m) = a\mathbb{E}X + b\mathbb{E}Y. \quad (2a2)$$

[2] Proof: suppose that X takes values in $\{x_1, x_2, ...\}$ and Y takes values in $\{y_1, y_2, ...\}$, then

$$1 = \sum_k \mathbb{P}(X = x_k) = \sum_k \mathbb{P}(X = x_k, Y \in \mathbb{R}) = \mathbb{P}(X \in \{x_1, x_2, ...\}, Y \in \mathbb{R}),$$

which implies $\mathbb{P}(X \in \mathbb{R} \setminus \{x_1, x_2, ...\}, Y \in \mathbb{R}) = 0$. Similarly, $\mathbb{P}(X \in \mathbb{R}, Y \in \mathbb{R} \setminus \{y_1, y_2, ...\}) = 0$. Recall that if $\mathbb{P}(A) = 0$ and $\mathbb{P}(B) = 0$, then $\mathbb{P}(A \cap B) = 0$, hence we conclude

$$\mathbb{P}(\mathbb{R} \setminus \{x_1, x_2, ...\}, Y \in \mathbb{R} \setminus \{y_1, y_2, ...\}) = 0,$$

which implies $\sum_{k,m} \mathbb{P}(X = x_k, Y = y_k) = \mathbb{P}(X \in \{x_1, x_2, ...\}, Y \in \{y_1, y_2, ...\}) = 1$, i.e. (X, Y) is a discrete random vector.

EXAMPLE 2a4. Let (X,Y) be as in the Example 2a2 and $g(x,y)=x^2y$, then

$$\mathbb{E}g(X,Y)=\sum_{k,m}k^2m(1-p)^{k-1}p(1-r)^{m-1}r=$$
$$\sum_{k=1}^{\infty}k^2(1-p)^{k-1}p\sum_{m=1}^{\infty}m(1-r)^{m-1}r=...=\frac{2-p}{p^2}\frac{1}{r}$$

■

The expectations of monomials $\mathbb{E}X^\ell Y^n$, $\ell,n\in\mathbb{N}$ are called the *joint moments*.

The *joint* p.g.f (or j.m.g.f.) is (here we assume that $x_k=k$ and $y_m=m$, i.e. the r.v. are integer valued):

$$G_{XY}(s,t):=\mathbb{E}s^Xt^Y=\sum_{k,m}s^kt^mp_{k,m},$$

which is well defined in e.g. the rectangular $|s|<1,|t|<1$. Analogously to one dimension, the j.p.m.f. can be generated from j.p.g.f.:

$$\frac{\partial^k}{\partial s^k}\frac{\partial^m}{\partial t^m}G_{XY}(0,0)=k!m!p_{k,m}.$$

Jointly continuous random variables. A pair of random variables (X,Y) are called jointly continuous (or, equivalently, the vector (X,Y) is continuous), if the j.c.d.f. $F_{XY}(u,v)$ is a continuous function, jointly[3] in (u,v). We shall deal exclusively with a subclass of continuous random vectors [4], whose c.d.f. is defined by:

$$F_{XY}(u,v)=\int_{-\infty}^{u}\int_{-\infty}^{v}f_{XY}(x,y)dxdy,$$

where $f_{XY}(x,y)$ is the *joint* p.d.f., i.e. a non-negative integrable function satisfying

$$\iint_{\mathbb{R}^2}f_{XY}(x,y)dxdy=1.$$

Note that e.g. $F_{XY}(u,-\infty)=0$ and

$$F_{XY}(u,\infty)=\int_{-\infty}^{u}\Big(\int_{\mathbb{R}}f_{XY}(u,v)dv\Big)du=\int_{-\infty}^{u}f_X(u)du=F_X(u)=:\mathbb{P}(X\le u),$$

where $f_X(u):=\int_{\mathbb{R}}f_{XY}(u,v)dv$ is the p.d.f. of X (indeed it is non-negative and integrates to 1). Hence as in the discrete case, the one dimensional p.d.f.'s and c.d.f's are *marginals* of j.p.d.f. and j.c.d.f.

EXAMPLE 2a5. Let (X,Y) be a r.v. with j.p.d.f.

$$f_{XY}(x,y)=\frac{1}{2\pi}e^{-x^2/2-y^2/2},\quad x,y\in\mathbb{R}.$$

[3]recall that a function $h(x,y)$ is jointly continuous in (x,y), if for *any* sequence (x_n,y_n), which converges to (x,y), $\lim_n h(x_n,y_n)=h(x,y)$

[4]and call them just "continuous", abusing the notations

It is easy to see that f_{XY} integrates to 1 and hence is a legitimate j.p.d.f.

$$f_X(x) = \int_{\mathbb{R}} \frac{1}{2\pi} e^{-x^2/2-y^2/2} dy = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R},$$

i.e. $X \sim N(0,1)$ (and similarly $Y \sim N(0,1)$). ■

j.c.d.f is not uniquely determined by its marginals, i.e. two different j.c.d.f's may have identical marginals:

EXAMPLE 2a6. Suppose X and Y have p.d.f.'s f_X and f_Y. Let α be a number in $[-1,1]$ and define

$$f_{XY}(x,y;\alpha) = f_X(x) f_Y(y) \Big(1 + \alpha\big(2F_X(x) - 1\big)\big(2F_Y(y) - 1\big)\Big).$$

Since $2F_X(x) - 1 \in [-1,1]$ for all $x \in \mathbb{R}$, f_{XY} is a non-negative function. Further, since $\frac{d}{dx} F_X^2(x) = 2F_X(x) f_X(x)$

$$\int_{\mathbb{R}} 2F_X(x) f_X(x) dx = \int_{\mathbb{R}} \frac{d}{dx} F_X^2(x) dx = 1,$$

we have $\iint_{\mathbb{R}^2} f_{XY}(x,y;\alpha) dx dy = 1$ and thus f_{XY} is a legitimate two dimensional p.d.f. Also a direct calculation shows that its marginals are f_X and f_Y, regardless of α. However, f_{XY} is a different function for different α's! ■

Expectation of $g(X,Y)$ for an appropriate function $g : \mathbb{R}^2 \mapsto \mathbb{R}$ is defined

$$\mathbb{E}g(X,Y) = \iint_{\mathbb{R}^2} g(u,v) f_{XY}(u,v) du dv.$$

Again it may be finite or infinite or may not exist at all. The linearity property as in (2a2) is readily checked.

The joint moments and the j.m.g.f. are defined similar to the discrete case, e.g.

$$M_{XY}(s,t) = \mathbb{E}e^{sX+tY},$$

if the expectation is finite in an open vicinity of the origin. In this case,

$$\frac{\partial^k}{\partial s^k} \frac{\partial^m}{\partial t^m} M_{XY}(s,t)\Big|_{(s,t)=(0,0)} = \mathbb{E}X^k Y^m.$$

The components of continuous random vector are continuous r.v. as we have already seen. However, the r.v. X and Y can be continuous individually, but not jointly:

EXAMPLE 2a7. Consider $X \sim N(0,1)$ and $Y := X$, then each X and Y are $N(0,1)$, i.e. continuous, while not jointly continuous. Indeed, suppose that (X,Y) has a density $f_{XY}(x,y)$. Since $\mathbb{P}(X \neq Y) = 0$,

$$1 = \iint_{\mathbb{R}^2} f_{XY}(u,v) du dv = \iint_{\mathbb{R}^2 \setminus \{(x,y): x=y\}} f_{XY}(u,v) du dv = \mathbb{P}(X \neq Y) = 0,$$

where the first equality is a property of the integral over $\mathbb{R}^2$. The contradiction indicates that j.p.d.f. doesn't exist in this case.

REMARK 2a8. Of course, random vectors may be neither discrete nor continuous in various ways. In particular, their entries may be r.v. of the "mixed" type, i.e. contain both continuous and discrete components, or the vector may contain both discrete and continuous entries. Calculation of the expectations in this case follow the same pattern, but can be more involved technically.

b. Independence

From here on we shall consider both discrete and continuous random vectors within the same framework, emphasizing differences only when essential. All of the following properties are equivalent[5] and can be taken as the *definition* of independence of X and Y:

(1) the j.c.d.f. factors into the product of individual c.d.f's:

$$F_{XY}(x,y) = F_X(x)F_Y(y), \quad \forall x,y \in \mathbb{R}.$$

(2) the j.m.g.f. factors into the product of individual m.g.f's:

$$M_{XY}(s,t) = M_X(s)M_Y(t), \quad \forall s,t \in D,$$

where D is the relevant domain;

(3) the j.p.d.f. factors into the product of individual p.d.f's in the continuous case:

$$f_{XY}(x,y) = f_X(x)f_Y(y), \quad \forall x,y \in \mathbb{R}.$$

or the j.p.m.f. factors into the product of individual p.m.f's in the discrete case:

$$p_{XY}(k,m) = p_X(k)p_Y(m), \quad \forall k,m \in \mathbb{N}.$$

(4) for *all* bounded functions g,h,

$$\mathbb{E}g(X)h(Y) = \mathbb{E}g(X)\mathbb{E}h(Y).$$

The equivalence between some of the characterizations (e.g. 3$\Leftrightarrow$ 1) is not hard to see directly from the definitions, while equivalence of the others is technically (but not conceptually!) more involved (e.g. 4 $\Leftrightarrow$ 1).

The intuition behind independence can be gained from the familiar notion of conditional probabilities: notice that (4) with $g(x) = I(x \in A)$ and $h(x) = I(x \in B)$, where A and B are subsets (e.g. intervals) of $\mathbb{R}$, yields:

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

If $\mathbb{P}(X \in B) > 0$, then conditional probability of the event $\{X \in A\}$, given $\{X \in B\}$ has occurred, equals the a priori probability of $\{X \in A\}$:

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

Hence independent r.v. generate independent events.

EXAMPLE 2b1. X and Y are independent in both Examples 2a2 and 2a5 above, as is readily verified by applying e.g. (3) above. ■

[5] of course, assuming that all of them apply (e.g. $M_{XY}(s,t)$ might not be defined)

EXAMPLE 2b2. Let $X \sim U([0,1])$ and $Y = I(X > 1/2)$. Note that (X, Y) does not categorize as discrete or continuous random vector. X and Y are not independent, since e.g.

$$\mathbb{E}XY = \mathbb{E}XI(X \geq 1/2) = \int_{1/2}^{1} x dx = 3/8,$$

while

$$\mathbb{E}X\mathbb{E}I(X \geq 1/2) = 1/2 \cdot 1/2 = 1/4,$$

contradicting (4) above. ■

The familiar notion of correlation between two r.v. is a weaker notion of independence.

DEFINITION 2b3. *Let X and Y be a pair of r.v. with finite second moments, $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$. Then the correlation coefficient between X and Y is*

$$\rho_{XY} := \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}},$$

where $\mathrm{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$, *etc. and* $\mathrm{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$.

To see that ρ_{XY} is a number in $[-1, 1]$, we shall need

LEMMA 2b4 (Cauchy–Schwarz inequality). *Let Z_1 and Z_2 be r.v. with finite second moments, then*

$$\left(\mathbb{E}Z_1Z_2\right)^2 \leq \mathbb{E}Z_1^2\mathbb{E}Z_2^2.$$

The equality is attained if and only if Z_1 and Z_2 are proportional, i.e. $Z_1 = cZ_2$ for a constant $c \neq 0$.

PROOF. Since $(Z_1 - \alpha Z_2)^2 \geq 0$ for any $\alpha \in \mathbb{R}$, it follows that $\mathbb{E}(Z_1 - \alpha Z_2)^2 \geq 0$ or expanding

$$\mathbb{E}(Z_1)^2 - 2\alpha\mathbb{E}Z_1Z_2 + \alpha^2\mathbb{E}(Z_2)^2 \geq 0.$$

The C-S inequality is obtained with $\alpha := \mathbb{E}Z_1Z_2/\mathbb{E}(Z_2)^2$. If $Z_1 = cZ_2$, then the inequality obviously holds with equality (by direct calculation). Conversely, if the equality holds, then for α as above, $(Z_1 - \alpha Z_2)^2 = 0$, which implies that $Z_1 = \alpha Z_2$ (at least with probability 1) as claimed. □

Taking $Z_1 = X - \mathbb{E}X$ and $Z_2 = Y - \mathbb{E}Y$ we deduce that $\rho_{XY} \in [-1, 1]$. If $|\rho_{XY}| = 1$, then $Y = aX + b$ for some constants a and b, which means that Y is a linear function of X (and hence there is a strong dependence between X and Y).

Independence implies $\rho_{XY} = 0$, since $\mathrm{cov}(X, Y) = 0$ for independent X and Y (check!). The converse does not have to be true in general, i.e. uncorrelated r.v. may be dependent:

EXAMPLE 2b5. Let $X \sim N(0, 1)$ and $Y := \xi X$, where ξ is a r.v. taking values $\{1, -1\}$ with probabilities $1/2$ and independent of X. Then

$$\mathrm{cov}(X, Y) = \mathbb{E}XY = \mathbb{E}\xi X^2 = \mathbb{E}\xi\mathbb{E}X^2 = 0.$$

However, X and Y are not independent:

$$\mathbb{E}X^2Y^2 = \mathbb{E}X^4\xi^2 = \mathbb{E}X^4 = 3\mathbb{E}X^2 = 3,$$

while

$$\mathbb{E}X^2\mathbb{E}Y^2 = \mathbb{E}X^2\mathbb{E}X^2 = 1.$$

■

The *covariance* matrix of the vector $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is by definition[6]

$$\mathrm{Cov}(X, X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) \end{pmatrix}.$$

A particularly important class of random vectors is Gaussian.

DEFINITION 2b6. *A random vector $X = (X_1, X_2)$ is Gaussian with mean $\mu = (\mu_1, \mu_2)$ and the covariance matrix*

$$S = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

if it has a j.p.d.f. of the form

$$f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\Big\{ -\frac{1}{2}\frac{1}{1-\rho^2}\Big(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \Big) \Big\}, \quad (2b1)$$

for $x \in \mathbb{R}^2$.

By a direct (but tedious) calculation, one can verify that indeed $\mathbb{E}X = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\mathrm{Cov}(X, X) = S$ as above. ρ is the correlation coefficient between X_1 and X_2. The marginal p.d.f.'s are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Remarkably, for Gaussian vectors lack of correlation implies independence

PROPOSITION 2b7. *Let $X = (X_1, X_2)$ be a Gaussian random vector in $\mathbb{R}^2$. Then X_1 and X_2 are independent if and only if they are uncorrelated.*

PROOF. Independence implies lack of correlation in general. Suppose that X_1 and X_2 are uncorrelated, i.e. $\rho = 0$. Then the j.p.d.f. factors into product of the individual p.d.f.'s as can be readily seen from (2b1). □

REMARK 2b8. The Example 2b5 also demonstrates that (X, Y) with Gaussian marginals[7] and uncorrelated entries may not be a Gaussian vector!

Try to imagine how various ingredients of the Gaussian j.p.d.f. affect its shape (look at its counters, sections etc.).

A calculation yields the formula for Gaussian j.m.g.f:

LEMMA 2b9. *Let X be a Gaussian vector as above, then*

$$M_X(t) = \exp\Big\{ \mu^\top t + \frac{1}{2} t^\top S t \Big\}, \quad t \in \mathbb{R}^2.$$

PROOF. direct (tedious) calculation □

[6]Note that $\mathrm{Cov}(X, X)$ is a symmetric matrix.

[7]Note that $M_Y(t) = \mathbb{E}e^{t\xi X} = \frac{1}{2}\mathbb{E}e^{tX} + \frac{1}{2}\mathbb{E}e^{-tX} = \frac{1}{2}e^{t^2/2} + \frac{1}{2}e^{t^2/2} = e^{t^2/2}$, which is an m.g.f. of $N(0, 1)$

REMARK 2b10. In fact, the above expression for m.g.f. is well defined even if $\rho = \pm 1$ (i.e. the covariance matrix S is singular). In these cases, the j.p.d.f does not exist and the Gaussian vector is referred to as *degenerate*. Do not think, however, that this is a pathology: e.g. if $X \sim N(0,1)$, the vector (X, X) is a degenerate Gaussian vector.

Gaussian distribution is stable under linear transformations, namely

LEMMA 2b11. *Let X be a Gaussian vector as above, A be a 2×2 matrix and b a vector in $\mathbb{R}^2$. Then $AX + b$ is a Gaussian vector with mean $A\mu + b$ and covariance matrix $ASA^\top$.*

PROOF. By definition,

$$M_{AX+b}(t) = \mathbb{E} \exp\left\{t^\top (AX + b)\right\} = \exp\left\{t^\top b\right\} \mathbb{E} \exp\left\{(A^\top t)^\top X\right\} =$$
$$\exp\left\{t^\top b\right\} \exp\left\{\mu^\top A^\top t + \frac{1}{2}(A^\top t)^\top S A t\right\} = \exp\left\{(A\mu + b)^\top t + \frac{1}{2} t^\top A S A^\top t\right\}.$$

Since the j.m.g.f. uniquely determines the probability law, $X \sim N(A\mu + b, ASA^\top)$. □

REMARK 2b12. Note that we haven't required that A is a nonsingular matrix. If A is in fact nonsingular (i.e. $\det(A) \neq 0$) then the vector $Y = AX + b$ has a j.p.d.f. Otherwise, it is a degenerate vector (see the Remark 2b10 above). For example, if X_1 and X_2 are i.i.d. $N(0,1)$ and

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

then by the Lamma 2b11, the vector $Y = AX + b$ is Gaussian with mean

$$\mathbb{E}Y = A\mathbb{E}X + b = b,$$

and covariance matrix

$$S_Y = A S_X A^\top = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}.$$

The latter means that $Y_1 \sim N(0,2)$ and $Y_2 \sim N(1,2)$ and $\rho(Y_1, Y_2) = -1$ (think why). Hence the j.p.d.f. is not well defined (as S_Y^{-1} doesn't exist). Hence Y is a degenerate Gaussian vector. This, of course, should be anticipated since $Y_1 = X_1 - X_2$ and $Y_2 = -X_1 + X_2 + 1$ and hence $Y_2 = -Y_1 + 1$.

EXAMPLE 2b13. Let X be a Gaussian random vector in $\mathbb{R}^2$ with i.i.d. standard components and let θ be a deterministic (non-random) angle in $[0, 2\pi]$. Recall that

$$U(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is the rotation matrix, i.e.[8] the vector $y = U(\theta)x$ has the same length as x and the angle between x and y is θ. The vector $Y = U(\theta)X$, i.e. the rotation of X by angle θ, is Gaussian with zero mean and covariance matrix $U(\theta)U^\top(\theta) = I$. This means that standard Gaussian distribution in $\mathbb{R}^2$ is invariant under deterministic rotations. ■

[8] convince yourself by a planar plot

c. Several random variables

The probabilistic description of a random vector $X = (X_1, ..., X_n)$ in $\mathbb{R}^n$ for $n > 2$ is completely analogous to the two dimensional case. The probabilities are assigned by the j.c.d.f

$$\mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n) := F_X(x_1, ..., x_n) = F_X(x), \quad x = (x_1, ..., x_n) \in \mathbb{R}^n.$$

The j.c.d.f $F_X(x)$, $x \in \mathbb{R}^n$, should satisfy the usual properties. The random vector X can be either discrete, in which case its j.c.d.f is determined by the j.p.m.f $p_X(k)$, $k \in \mathbb{N}^n$, or continuous, when j.c.d.f is given by an n-fold integral of j.p.d.f. $f_X(x)$, $x \in \mathbb{R}^n$. j.p.m.f./j.p.d.f. should sum/integrate to 1.

The one-dimensional marginals are defined for all the quantities as in the two-dimensional case, e.g.

$$f_{X_2}(x_2) = \int_{\mathbb{R}^{n-1}} f_X(x_1, x_2, ..., x_n) dx_1 dx_3 ... dx_n.$$

Hence once again, the probability law of each individual component is completely determined by the probability law of the vector. In addition, one can recover the joint probability law of any subset of r.v.: for example, the j.p.d.f. of (X_1, X_n) is given by

$$f_{X_1 X_n}(x_1, x_n) = \int_{\mathbb{R}^{n-2}} f_X(x_1, x_2, ..., x_{n-1}, x_n) dx_2 ... dx_{n-1}.$$

or, equivalently,

$$F_{X_1 X_n}(x_1, x_2) = F_X(x_1, \infty, ..., \infty, x_n).$$

Similarly, the *k-dimensional marginals* are found. Expectation of a function $g : \mathbb{R}^n \mapsto \mathbb{R}$ of a continuous random vector is

$$\mathbb{E}g(X) = \int_{\mathbb{R}^n} g(x) f_X(x) dx_1 ... dx_n$$

and of a discrete r.v. (with e.g. integer values components)

$$\mathbb{E}g(X) = \sum_{x \in \mathbb{N}^n} g(x) p_X(x).$$

The j.m.g.f is

$$M_X(t) = \mathbb{E}e^{\sum_{i=1}^n t_i X_i}, \quad t \in D \subseteq \mathbb{R}^n,$$

where D is an open vicinity of the origin in $\mathbb{R}^n$. The joint moments can be extracted from j.m.g.f. similarly to the two dimensional case. The covariance matrix is defined if all the entries of X have finite second moments:

$$\mathrm{Cov}(X, X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top.$$

The r.v. $(X_1, ..., X_n)$ are *independent* if their j.c.d.f. (j.p.d.f., etc.) factors into the product of individual c.d.f's (p.d.f.'s, etc.) analogously to the corresponding properties (1)-(4) above. Independence of all the components of the vector clearly implies independence of any subset of its components. However, a random vector may have dependent components, which are e.g. pairwise (or triplewise, etc.) independent.

In statistics, n independent samples from p.d.f. f means a realization of the random vector $X = (X_1, ..., X_n)$, with the j.p.d.f. of the form:

$$f_X(x) = \prod_{i=1}^{n} f(x_i), \quad x \in \mathbb{R}^n.$$

EXAMPLE 2c1. Suppose $X = (X_1, ..., X_n)$ is a vector of i.i.d. random variables, where X_1 takes values in $\{x_1, ..., x_k\}$ with positive probabilities $p_1, ..., p_k$. Let Y_i be the number of times x_i appeared in the vector X, i.e. $Y_i = \sum_{m=1}^{n} I(X_m = x_i)$ and consider the random vector $Y = (Y_1, ..., Y_k)$. Clearly each Y_i takes values in $\{0, ..., n\}$, i.e. it is a discrete r.v. Hence the random vector Y is also discrete. The j.p.m.f. of Y is given by

$$p_Y(n_1, ..., n_k) = \begin{cases} 0 & n_1 + ... + n_k \neq n \\ \frac{n!}{n_1! ... n_k!} p_1^{n_1} ... p_k^{n_k} & n_1 + ... + n_k = n \end{cases}.$$

The corresponding distribution is called *multinomial*. The particular case $k = 2$, is readily recognized as the familiar binomial distribution. One dimensional marginals are binomial, since each Y_i is a sum of n Bernoulli i.i.d. random variables $I(X_m = x_i)$, $m = 1, ..., n$. (this can also be seen by the direct and more tedious calculation of the marginal). Similarly, the two dimensional marginals are trinomial, etc.

Let's check dependence between the components. Clearly $Y_1, ..., Y_k$ are not independent, since $Y_1 = n - \sum_{i=2}^{k} Y_k$ and hence e.g. Y_1 and $\sum_{i=2}^{k} Y_k$ are fully correlated, which would be impossible, were $Y_1, ..., Y_n$ independent. What about pairwise independence: e.g. are Y_1 and Y_2 independent ...? Intuitively, we feel that they are not: after all if $Y_1 = n$, then $Y_2 = 0$! This is indeed the case:

$$\mathbb{P}(Y_1 = n, Y_2 = 0) = \mathbb{P}(Y_1 = n) = p_1^n,$$

while

$$\mathbb{P}(Y_1 = n)\mathbb{P}(Y_2 = 0) = p_1^n (1 - p_2)^n.$$

■

Gaussian vectors in $\mathbb{R}^n$ play an important role in probability theory.

DEFINITION 2c2. *X is a Gaussian random vector with mean μ and nonsingular[9] covariance matrix S, if it has j.p.d.f. of the form:*

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(S)}} \exp\Big\{ -\frac{1}{2}(x-\mu)^\top S^{-1}(x-\mu)\Big\}, \quad x \in \mathbb{R}^n.$$

It is not hard to see that the two-dimensional density from Definition 2b6 is obtained directly from the latter general formula. If S is a diagonal matrix, i.e. all the entries of X are uncorrelated, $X_1, ..., X_n$ are independent. Furthermore, if $\mu = 0$ and S is an identity matrix, then $X_1, ..., X_n$ are i.i.d. standard Gaussian r.v.'s.

Both Lemmas 2b9 and 2b11 remain valid in the general multivariate case, i.e. the j.m.g.f. of a Gaussian vector is an exponential of a quadratic form and Gaussian multivariate distribution is stable under linear transformations.

[9] positive definite

Exercises

PROBLEM 2.1. A pair of fair dice is thrown. Let X be the sum of outcomes and Y be the Bernoulli r.v. taking value 1 if the first dice comes up even.

(1) Show that (X, Y) is discrete and find its j.p.m.f. (present your answer as a table)
(2) Calculate the marginal p.m.f.'s
(3) Calculate the moments $\mathbb{E}X$, $\mathbb{E}Y$ and $\mathbb{E}XY$

PROBLEM 2.2. Answer the questions from the previous problem for the r.v. (X, Y), defined by the following experiment. Two Hanukkah tops are spined: X and Y are the outcomes of the first and the second tops respectively.

PROBLEM 2.3. Let (X, Y) be a random vector with the j.p.d.f

$$f_{XY}(x, y) = \frac{1}{2} xy I(y \in [0, x]) I(x \in [0, 2]).$$

(1) Verify that f_{XY} is a legitimate j.p.d.f.
(2) Find the support of f_{XY}, i.e. $D := \{(x, y) \in \mathbb{R}^2 : f_{XY}(x, y) > 0\}$
(3) Find the j.c.d.f of (X, Y)
(4) Find the marginals of j.p.d.f
(5) Are X and Y independent ?
(6) Find the marginals of j.c.d.f.

PROBLEM 2.4. Let (X, Y) be a random vector distributed uniformly on the triangle with the corners at $(-2, 0), (2, 0), (0, 2)$.

(1) Find the j.p.d.f.
(2) Find the marginal p.d.f.'s
(3) Are X and Y independent ?
(4) Calculate the correlation coefficient between X and Y

PROBLEM 2.5. Let (X, Y) be a random vector with j.p.d.f $f_{XY}(x, y) = 12(x-y)^2 I((x, y) \in A)$ where $A = \{(x, y) : 0 \le x \le y \le 1\}$.

(1) find the marginal p.d.f.'s
(2) Calculate $\mathbb{P}(1/2 \le X + Y \le 1)$, $\mathbb{P}(X + Y \le 1/2)$

PROBLEM 2.6. Let F be a one dimensional c.d.f.. Are the following functions legitimate two-dimensional j.c.d.f's ? Prove, if your answer is positive and give a counterexample if negative.

(1) $F(x, y) := F(x) + F(y)$

(2) $F(x,y) := F(x)F(y)$
(3) $F(x,y) := \max\{F(x), F(y)\}$
(4) $F(x,y) := \min\{F(x), F(y)\}$

PROBLEM 2.7. Let X and Y be Bernoulli r.v. with parameter $p \in (0,1)$. Show that if X and Y are uncorrelated, they are also independent.

Hint: the j.p.m.f. of the vector (X,Y) is supported on four vectors: $(0,0)$, $(0,1)$ $(1,0)$ and $(1,1)$. Denote the corresponding probabilities by p_{00}, p_{01}, p_{10} and p_{11}. Analyze the constraints imposed on j.p.m.f. by the lack of correlation between X and Y and deduce that these constraints also imply independence.

PROBLEM 2.8. Let X and Y be independent random variables. Find $\mathbb{P}(X = Y)$ when
(1) $X, Y \sim \text{Geo}(p)$
(2) (X,Y) is a non-degenerate Gaussian vector

PROBLEM 2.9. Let (X,Y) be a Gaussian vector with parameters μ_x, μ_y, σ_x and σ_y and ρ.
(1) Find the p.d.f. of the average $(X+Y)/2$
(2) Find the p.d.f. of $(X-Y)/2$
(3) Find the values of ρ so that the r.v. in (1) and (2) are independent

PROBLEM 2.10. Let (X,Y) be a random vector, whose components have finite nonzero second moments and correlation coefficient ρ. Show that for $a, b, c, d \in \mathbb{R}$

$$\rho(aX + b, cY + d) = \frac{ac}{|ac|}\rho(X,Y).$$

PROBLEM 2.11. Show that for a discrete random vector (X,Y), the properties

$$\mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(X), \quad \forall g, f \text{ bounded}$$

and

$$p_{XY}(k,m) = p_X(k)p_Y(m), \quad \forall k, m \in \mathbb{N}$$

are equivalent.

PROBLEM 2.12. Show that the general formula for the Gaussian density from Definition 2c2 in two dimensions reduces to the one from Definition 2b6

PROBLEM 2.13. Let X and Y be independent random vectors in $\mathbb{R}^k$ and $\mathbb{R}^m$ respectively and let $g : \mathbb{R}^k \mapsto \mathbb{R}$ and $h : \mathbb{R}^m \mapsto \mathbb{R}$ be some functions. Show that $\xi = g(X)$ and $\eta = h(Y)$ are independent r.v.

PROBLEM 2.14. Let X be a standard Gaussian vector in $\mathbb{R}^n$ (i.e. with i.i.d. $N(0,1)$ entries)

(1) Show that the empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is a Gaussian random variable and find its mean and variance
(2) Argue that the vector $(X_1, \bar{X})$ is a Gaussian vector in $\mathbb{R}^2$, find its mean and covariance matrix
(3) Argue that the vector $R := (X_1 - \bar{X}, ..., X_n - \bar{X})$ is a Gaussian vector, find its mean and covariance matrix
(4) Show that $\bar{X}$ and R are independent

PROBLEM 2.15. Let X and Y be i.i.d. $N(0,1)$ r.v.'s Show that $X^2 - Y^2$ and $2XY$ have the same probability law.

Hint: Note that $X^2 - Y^2 = 2\frac{X-Y}{\sqrt{2}}\frac{X+Y}{\sqrt{2}}$

PROBLEM 2.16. Let $X = (X_1, X_2, X_3)$ be a random vector with j.p.d.f.

$$f_X(x) = \frac{1}{(2\pi)^{3/2}} \exp\Big(-\frac{1}{2}x_1^2 - \frac{1}{2}(x_2 - x_1)^2 - \frac{1}{2}(x_3 + x_2)^2\Big), \quad x \in \mathbb{R}^3$$

(1) Check that f_X is indeed a legitimate j.p.d.f
(2) Find all the one dimensional marginal p.d.f's
(3) Find all the two dimensional marginal j.p.d.f's. Are (X_1, X_2), (X_2, X_3), (X_1, X_3) Gaussian vectors in $\mathbb{R}^2$? If yes, find the corresponding parameters.
(4) Verify that f_X can be put in the following form:

$$f_X(x) = \frac{1}{(2\pi)^{3/2}\sqrt{det(S)}} \exp\Big(-\frac{1}{2}(x-\mu)^\top S^{-1}(x-\mu)\Big), \quad x \in \mathbb{R}^3,$$

where

$$S = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -1 & -2 & 3 \end{pmatrix}$$

PROBLEM 2.17. Consider $f : \mathbb{R}^n \mapsto \mathbb{R}$

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\prod_{i=1}^n \sigma_i^2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2}\Big), \quad x \in \mathbb{R}^n$$

(1) Prove that f is a j.p.d.f
(2) Let $X = (X_1, ..., X_n)$ be a random vector with j.p.d.f. f. Show that its entries are independent and specify the conditions for μ_i and σ_i's such that $X_1, ..., X_n$ are i.i.d
(3) Find the probability law of $S = \sum_{i=1}^n X_n$
(4) Suppose that n is even and find the probability law of

$$S_e = \sum_{i=1}^{n/2} X_{2i} \quad \text{and} \quad S_o = \sum_{i=1}^{n/2} X_{2i-1}.$$

Are S_e and S_o independent ? Identically distributed ?

(5) Suppose that $n \geq 2$ and find the probability law of $Y = \sum_{i=1}^{n-1} X_i$ and $Z = \sum_{i=2}^{n} X_i$. Are Z and Y independent ? Identically distributed ?

PROBLEM 2.18. Let $X = (X_1, ..., X_n)$ be i.i.d. $N(0,1)$ r.v. and $\xi = (\xi_1, ..., \xi_n)$ i.i.d. Ber$(1/2)$. Assuming that X and ξ are independent, find the probability law of

$$Z = \sum_{i=1}^{n} (1 - 2\xi_i) X_i.$$

CHAPTER 3

Conditional expectation

a. The definition and the characterization via orthogonality

Recall the definition of the *conditional probability* of an event A, given event B:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \tag{3a1}$$

where $\mathbb{P}(B) > 0$. If we know that B occurred, i.e. the experiment outcome ω is known to belong to B, the event A occurs, i.e. $\omega \in A$, if and only if $\omega \in A \cap B$. Hence the conditional probability in (3a1) is proportional to $\mathbb{P}(A \cap B)$ and is normalized so that $\mathbb{P}(A|B)$ is a legitimate probability measure with respect to A, for any fixed B.

Consider a pair of random variables (X, Y) and suppose we want to calculate the conditional probability of an event A, generated by X, say of $A := \{X \leq 0\}$, having observed the realization of the random variable Y. If Y takes a countable number of values, $Y \in \{y_1, y_2, ...\}$, and $\mathbb{P}(Y = y_i) > 0$, then we can use the formula (3a1):

$$\mathbb{P}(A|Y = y_i) = \frac{\mathbb{P}(A, Y = y_i)}{\mathbb{P}(Y = y_i)}.$$

However, how do we define the probability $\mathbb{P}(A|Y = y)$, if Y is a continuous random variable and hence $\mathbb{P}(Y = y) = 0$ for all y's ...? An intuitively appealing generalization is the "infinitesimal" definition

$$\mathbb{P}(A|Y = y) := \lim_{\varepsilon \to 0} \frac{\mathbb{P}(A, |Y - y| \leq \varepsilon)}{\mathbb{P}(|Y - y| \leq \varepsilon)}.$$

While conceptually simple (and, in fact, correct - see Remark 3d8 below), this definition turns to have a somewhat limited scope and does not easily reveal some of the important sides of conditioning with respect to random variables[1]. Below we shall take a different approach to conditional probabilities, based on its variational characterization.

Let $X = (X_1, ..., X_n)$ be a random vector and suppose that we have sampled (i.e. obtained the realizations of) $X_2, ..., X_n$ and would like to guess the realization of X_1. If for example, X had multinomial distribution, this would be simple and precise, namely $n - \sum_{i=2}^{n} X_i$ would be the only possible realization of X_1. On the other hand, if the components of X were all independent, then, intuitively, we feel that any guess, based on $X_2, ..., X_n$ would be as bad as a guess, not utilizing any information, gained from observing $X_2, ..., X_n$. The problem we want to address in this chapter is the following: for a random vector $X = (X_1, ..., X_n)$ with known distribution,

[1] at its ultimate modern form, the basic conditioning is with respect to σ-algebras of events. Hence conditioning with respect to random variables is defined as conditioning with respect to the σ-algebra of events, they generate. Again, this truth is beyond the scope of these lecture notes

how do we guess the realization of X_1, given the realizations of $X_2, ..., X_n$ - preferably in the best possible way ?

To formulate this problem of *prediction* in rigorous terms, we have to define what we mean by "guess" and what is "best" for us. It is natural to consider the guesses of X_1 of the form $g(X_2, ..., X_n)$, where g is an $\mathbb{R}^{n-1} \mapsto \mathbb{R}$ function: this means that each time we get a realization of $X_2, ..., X_n$, we plug it into g and thus generate a real number, which we interpret as the guess of X_1. How do we compare different guesses ? Clearly, looking at the error $X_1 - g(X_2, ..., X_n)$ would be meaningless, since it varies randomly from realization to realization (or experiment to experiment). Instead we shall measure the quality of the guess g by the *mean square error* (MSE)

$$\mathbb{E}\big(X_1 - g(X_2, ..., X_n)\big)^2.$$

The best guess for us will be the one which minimizes the MSE over all guesses, i.e. we aim to find g^* so that

$$\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)^2 \leq \mathbb{E}\big(X_1 - g(X_2, ..., X_n)\big)^2 \quad \text{for all } g\text{'s} \tag{3a2}$$

Finding such g^* does not seem to be an easy problem, since the space of potential candidates for g - all[2] functions on $\mathbb{R}^{n-1}$ - is vast and apparently it doesn't have any convenient structure to perform the search[3]. However, things are not as hopeless as they may seem, due to the following simple observation:

LEMMA 3a1 (Orthogonality property). *g^* satisfies* (3a2) *if and only if*

$$\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)h(X_2, ..., X_n) = 0, \quad \textit{for all functions } h. \tag{3a3}$$

PROOF. Assume that (3a3) holds, then

$$\begin{aligned}
&\mathbb{E}\big(X_1 - g(X_2, ..., X_n)\big)^2 = \\
&\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n) + g^*(X_2, ..., X_n) - g(X_2, ..., X_n)\big)^2 = \\
&\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)^2 + \mathbb{E}\big(g^*(X_2, ..., X_n) - g(X_2, ..., X_n)\big)^2 + \\
&2\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)\big(g^*(X_2, ..., X_n) - g(X_2, ..., X_n)\big) \stackrel{\dagger}{=} \\
&\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)^2 + \mathbb{E}\big(g^*(X_2, ..., X_n) - g(X_2, ..., X_n)\big)^2 \geq \\
&\mathbb{E}\big(X_1 - g^*(X_2, ..., X_n)\big)^2,
\end{aligned}$$

where we used (3a3) in †. This verifies (3a2). Conversely, assume that g^* satisfies (3a2), then in particular it holds for $g := g^* + \varepsilon h$, where $\varepsilon > 0$ is a constant and h is an arbitrary function of $X_2, ..., X_n$:

$$\mathbb{E}\big(X_1 - g^*\big)^2 \leq \mathbb{E}\big(X_1 - g^* - \varepsilon h\big)^2,$$

[2]of course, the functions g must be such that the expectations are well defined

[3]if for example, we would have to find the optimal g^* from a finite collection of candidate functions $\{g_1, ..., g_\ell\}$, the task would be simple: just calculate all the corresponding MSE's and choose the minimal one. Also it would be simple if we were looking for g^* within certain family of candidates: e.g. of the form $g(X_2, ..., X_n) = \sum_{i=2}^n a_i X_i + b$, where a_i's and b are real numbers. Then we could derive an expression for MSE, which depends only on a_i's and b. In this case we could find the minimal MSE by the tools from calculus: derivatives, etc. But this is not what we want: we are searching for g^* among all (not too weird) functions !

or equivalently

$$\mathbb{E}\big(X_1 - g^*\big)^2 - \mathbb{E}\big(X_1 - g^* - \varepsilon h\big)^2 = 2\varepsilon\mathbb{E}(X_1 - g^*)h - \varepsilon^2\mathbb{E}h^2 \le 0.$$

Dividing by ε and passing to the limit $\varepsilon \to 0$, we get $\mathbb{E}(X_1 - g^*)h \le 0$. If we replace h with $-h$, the converse conclusion $\mathbb{E}(X_1 - g^*)h \ge 0$ is obtained, which implies $\mathbb{E}(X_1 - g^*)h = 0$ for all h. □

REMARK 3a2. (3a3) means that the prediction error $X_1 - g^*(X_2, ..., X_n)$ should be *orthogonal*[4] to any function of the conditioning random vector $(X_2, ..., X_n)$.

The optimal predictor $g^*(X_1, ..., X_n)$ is called *the conditional expectation* of X_1 given $X_2, ..., X_n$ and is denoted by $\mathbb{E}(X_1|X_2, ..., X_n)$. Do not be tricked by the name: the conditional expectation is a random variable by itself! It is customary to denote the optimal function $g^* : \mathbb{R}^{n-1} \mapsto \mathbb{R}$ by $\mathbb{E}(X_1|X_2 = x_2, ..., X_n = x_n) := g^*(x_2, ..., x_n)$, which will be referred to as the *function, realizing the conditional expectation.*

Since the conditions (3a3) and (3a2) are equivalent, both can be used to define the conditional expectation: for example, any r.v. which is a function of $X_2, ..., X_n$ and satisfies the orthogonality property (3a3) is *by definition* the conditional expectation. Simple as it is, this definition is quite powerful as we shall see shortly. In particular, it can be actually used to find the conditional expectations: one immediate way is to suggest a candidate and to check that it satisfies (3a3)! Here are some simple examples:

EXAMPLE 3a3. Let X be a random variable and let $Y = X^3$. What is $\mathbb{E}(X|Y)$? Let's try $g^*(Y) = \sqrt[3]{Y}$. This is certainly a function of the condition, and (3a3) holds trivially: $\mathbb{E}(X - \sqrt[3]{Y})h(Y) = \mathbb{E}(X - \sqrt[3]{X^3})h(Y) = \mathbb{E}(X - X)h(Y) = 0$ for all h. Hence $\mathbb{E}(X|Y) = \sqrt[3]{Y}$. Of course, this is also clear from the definition of g^* as the minimizer: the MSE corresponding to the predictor $\sqrt[3]{Y}$ is zero and hence is minimal. Note that along the same lines $\mathbb{E}(X|X) = X$.■

EXAMPLE 3a4. Suppose that X_1 and X_2 are independent. What would be $\mathbb{E}(X_1|X_2)$? The natural candidate is $\mathbb{E}(X_1|X_2) = \mathbb{E}X_1$. $\mathbb{E}X_1$ is a constant and hence is a (rather simple) function of X_2. Moreover, by independence, $\mathbb{E}h(X_2)(X_1 - \mathbb{E}X_1) = \mathbb{E}h(X_2)\mathbb{E}(X_1 - \mathbb{E}X_1) = 0$, which verifies (3a3). Let's see now that e.g. $g(x) = \mathbb{E}X_1 - 1$ is a bad candidate: let's check whether

$$\mathbb{E}\big(X_1 - g(X_2)\big)h(X_2) = 0, \quad \forall h \quad ...?$$

To this end, we have:

$$\mathbb{E}\big(X_1 - \mathbb{E}X_1 + 1\big)h(X_2) = \mathbb{E}\big(X_1 - \mathbb{E}X_1 + 1\big)\mathbb{E}h(X_2) = \mathbb{E}h(X_2).$$

Clearly, the latter does not vanish for *all* h: e.g. not for $h(x) = x^2$. ■

But how do we find the conditional expectation in less trivial situations, when finding *the right* candidate is less obvious ? In fact, it is even not clear whether or when the conditional expectation exists! Before addressing these questions, let us note that if it exists, it is essentially unique: i.e. if one finds a right candidate it is *the* right candidate !

LEMMA 3a5. *If g^* and $\tilde{g}^*$ satisfy* (3a3)*, then* $\mathbb{P}(g^* = \tilde{g}^*) = 1$.

PROOF. Subtracting the equalities $\mathbb{E}(X_1 - g^*)(g^* - \tilde{g}^*) = 0$ and $\mathbb{E}(X_1 - \tilde{g}^*)(g^* - \tilde{g}^*) = 0$, we get $\mathbb{E}(g^* - \tilde{g}^*)^2 = 0$ and the claim (think why?). □

[4]two random variables ξ and η are said to be orthogonal, if $\mathbb{E}\xi\eta = 0$

b. The Bayes formulae

There is no general formula for conditional expectation and in certain situations the computations are challenging. For discrete or continuous random vectors, the conditional expectation can be calculated by means of the Bayes formula. Let's demonstrate the idea on the familiar grounds

EXAMPLE 3b1. Let (X, Y) be a random vector in $\mathbb{N}^2$ with the j.p.m.f. $p_{XY}(k, m)$, $k, m \in \mathbb{N}$. The Bayes formula from the basic course of probability tells that the conditional distribution (or more precisely, the conditional p.m.f.) of X given Y is

$$p_{X|Y}(k; m) = \mathbb{P}(X = k|Y = m) = \frac{p_{XY}(k, m)}{p_Y(m)}$$

and

$$\mathbb{E}(X|Y = m) = \sum_k k p_{X|Y}(k; m) = \frac{\sum_k k p_{XY}(k, m)}{p_Y(m)}.$$

The latter is nothing but the optimal function $g^*(m)$ and the corresponding conditional expectation is

$$\mathbb{E}(X|Y) = \sum_k k p_{X|Y}(k; Y) = \frac{\sum_k k p_{XY}(k, Y)}{p_Y(Y)}.$$

Let's see how these formulae are obtained from the orthogonality property: we are looking for a function g^*, such that

$$\mathbb{E}\big(X - g^*(Y)\big)h(Y) = 0, \quad \forall h.$$

On one hand,

$$\mathbb{E}Xh(Y) = \sum_{k,m} kh(m)p_{XY}(k, m)$$

and on the other hand,

$$\mathbb{E}g^*(Y)h(Y) = \sum_{k,m} g^*(m)h(m)p_{XY}(k, m).$$

Hence

$$\mathbb{E}\big(X - g^*(Y)\big)h(Y) = \sum_{k,m} kh(m)p_{XY}(k, m) - \sum_{k,m} g^*(m)h(m)p_{XY}(k, m) = \sum_m h(m)\Big(\sum_k k p_{XY}(k, m) - g^*(m)\sum_k p_{XY}(k, m)\Big).$$

The latter expression should equal zero for *any* function h: this would be the case if we choose

$$\sum_k k p_{XY}(k, m) - g^*(m)\sum_k p_{XY}(k, m) = 0, \quad \forall m \in \mathbb{N},$$

which is precisely the Bayes formula as above:

$$g^*(m) = \frac{\sum_k k p_{XY}(k, m)}{\sum_k p_{XY}(k, m)} = \frac{\sum_k k p_{XY}(k, m)}{p_Y(m)}.$$

Note that any other essentially different choice of g^* such that (3a3) holds, is impossible, by uniqueness of the conditional expectation. ■

Following the very same steps we can derive various other Bayes formulae:

PROPOSITION 3b2 (The discrete Bayes formula). *Let X be a discrete random vector and let $p_X(k)$, $k \in \mathbb{N}^n$ be its j.p.m.f. Then for a function $\varphi : \mathbb{R}^m \mapsto \mathbb{R}$*

$$\mathbb{E}\big(\varphi(X_1, ..., X_m)|X_{m+1} = k_{m+1}, ..., X_n = k_n\big) = \frac{\sum_{k_1,...,k_m} \varphi(k_1, ..., k_m)p_X(k_1, ..., k_n)}{\sum_{k_1,...,k_m} p_X(k_1, ..., k_n)}, \quad (k_{m+1}, ..., k_n) \in \mathbb{N}^{n-m}. \quad (3b1)$$

REMARK 3b3. The conditional expectation is obtained by evaluating the function

$$(k_{m+1}, ..., k_n) \mapsto \mathbb{E}\big(\varphi(X_1, ..., X_m)|X_{m+1} = k_{m+1}, ..., X_n = k_n\big),$$

at $k_{m+1} := X_{m+1}$, ..., $k_n := X_n$.

PROOF. The claim is proved by verifying the characterizing properties of the conditional expectation. First, note that the suggested candidate (3b1) is clearly a function of the coordinates appearing in the condition. Thus it is left to establish the orthogonality property (3a3): for an arbitrary h we should check

$$\mathbb{E}\Big(\varphi(X_1, ..., X_m) - g^*(X_{m+1}, ..., X_n)\Big)h(X_{m+1}, ..., X_n) = 0$$

To this end:

$$\begin{aligned}
&\mathbb{E}g^*(X_{m+1}, ..., X_n)h(X_{m+1}, ..., X_n) = \\
&\sum_{x_1,...,x_n} g^*(x_{m+1}, ..., x_n)h(x_{m+1}, ..., x_n)p_X(x_1, ..., x_n) = \\
&\sum_{x_1,...,x_n} \frac{\sum_{k_1,...,k_m} \varphi(k_1, ..., k_m)p_X(k_1, ...k_m, x_{m+1}, ..., x_n)}{\sum_{\ell_1,...,\ell_m} p_X(\ell_1, ..., \ell_m, x_{m+1}, ..., x_n)} \times \\
&\qquad\qquad h(x_{m+1}, ..., x_n)p_X(x_1, ..., x_n) = \\
&\sum_{x_{m+1},...,x_n} \frac{\sum_{k_1,...,k_m} \varphi(k_1, ..., k_m)p_X(k_1, ...k_m, x_{m+1}, ..., x_n)}{\sum_{\ell_1,...,\ell_m} p_X(\ell_1, ..., \ell_m, x_{m+1}, ..., x_n)} \times \\
&\qquad\qquad h(x_{m+1}, ..., x_n) \sum_{x_1,...,x_m} p_X(x_1, ..., x_n) = \\
&\sum_{x_{m+1},...,x_n} \sum_{k_1,...,k_m} \varphi(k_1, ..., k_m)p_X(k_1, ...k_m, x_{m+1}, ..., x_n)h(x_{m+1}, ..., x_n) = \\
&\sum_{x_1,...,x_n} \varphi(x_1, ..., x_m)h(x_{m+1}, ..., x_n)p_X(x_1, ..., x_n) = \\
&\mathbb{E}\varphi(X_1, ..., X_m)h(X_{m+1}, ..., X_n),
\end{aligned}$$

which verifies the claim. □

Note that

$$\mathbb{E}\big(\varphi(X_1,...,X_m)|X_{m+1}=k_{m+1},...,X_n=k_n\big)= \sum_{x_1,...,x_m}\varphi(x_1,...,x_m)p_{X_1,...,X_m|X_{m+1},...,X_n}(x_1,...,x_m;k_{m+1},...,k_n),$$

where

$$p_{X_1,...,X_m|X_{m+1},...,X_n}(x_1,...,x_m;x_{m+1},...,x_n):=\frac{p_X(x_1,...,x_n)}{\sum_{x_1,...,x_m}p_X(x_1,...,x_n)}$$

is the *conditional* j.p.m.f. of $X_1,...,X_m$, given $X_{m+1},...,X_n$. The corresponding conditional j.m.g.f. and j.p.g.f are defined in the usual way.

REMARK 3b4. In particular, for $\varphi(x)=I(X_1=j)$, we recover the familiar expression for conditional probabilities: e.g.

$$\mathbb{E}\big(I(X_1=j)|X_2=x_2,X_3=x_3\big)=\frac{p_X(j,x_2,x_3)}{\sum_i p_X(i,x_2,x_3)}=\mathbb{P}(X_1=j|X_2=x_2,X_3=x_3).$$

EXAMPLE 3b5. Consider $Y_1,...,Y_k$ from Example 2c1 with multinomial distribution. The conditional p.m.f. of Y_1, given Y_k is found by means of the Bayes formula

$$p_{Y_1|Y_k}(x;y)=\frac{p_{Y_1Y_k}(x;y)}{\sum_j p_{Y_1Y_k}(j,y)}=\frac{p_{Y_1Y_k}(x;y)}{p_{Y_k}(y)}= \frac{\frac{n!}{x!y!(n-x-y)!}p_1^x p_k^y(1-p_1-p_k)^{n-x-y}}{\frac{n!}{y!(n-y)!}p_k^y(1-p_k)^{n-y}}= \frac{(n-y)!}{x!(n-x-y)!}\frac{p_1^x(1-p_1-p_k)^{n-x-y}}{(1-p_k)^{n-y}}= \binom{n-y}{(n-y)-x}\Big(\frac{p_1}{1-p_k}\Big)^x\Big(1-\frac{p_1}{1-p_k}\Big)^{(n-y)-x}$$

for $x=0,...,n-y$ and zero otherwise. This is easily recognized as the Binomial distribution $\mathrm{Bin}(p_1/(1-p_k),n-y)$. Hence e.g. the conditional expectation of Y_1, given Y_k is

$$\mathbb{E}(Y_1|Y_k)=(n-Y_k)\frac{p_1}{1-p_k}.$$

Again, note that the latter is a random variable ! ■

Similarly, the conditional expectation is calculated in the continuous case:

PROPOSITION 3b6 (The continuous Bayes formula). *Let X be a continuous random vector taking values in $\mathbb{R}^n$ with j.p.d.f. $f_X(x)$, $x\in\mathbb{R}^n$. Then for any $\varphi:\mathbb{R}^m\mapsto\mathbb{R}$*

$$\mathbb{E}\big(\varphi(X_1,...,X_m)|X_{m+1}=x_{m+1},...,X_n=x_n\big)= \int_{\mathbb{R}^m}\varphi(x_1,...,x_m)f_{X_1,...,X_m|X_{m+1},...,X_n}(x_1,...,x_m;x_{m+1},...,x_n)dx_1...dx_m,$$

where

$$f_{X_1,...,X_m|X_{m+1},...,X_n}(x_1,...,x_m;x_{m+1},...,x_n)=\frac{f_X(x)}{\int_{\mathbb{R}^m}f_X(x)dx_1...dx_m},$$

is the conditional j.p.d.f. of $X_1, ..., X_m$, given $X_{m+1}, ..., X_n$. The conditional expectation is obtained by plugging $X_{m+1}, ..., X_n$ into the function $\mathbb{E}\big(\varphi(X_1, ..., X_m)|X_{m+1} = x_{m+1}, ..., X_n = x_n\big)$.

PROOF. Similar to the discrete case □

For $n = 2$, the latter reads

COROLLARY 3b7. *Let X_1 and X_2 be jointly continuous r.v.'s with the j.p.d.f. $f_{X_1X_2}(u, v)$. Then for a function $\phi : \mathbb{R} \mapsto \mathbb{R}$*

$$\mathbb{E}(\phi(X_1)|X_2) = \int_{\mathbb{R}} \phi(u) f_{X_1|X_2}(u; X_2) du,$$

where

$$f_{X_1|X_2}(u; v) = \frac{f_{X_1X_2}(u, v)}{f_{X_2}(v)} = \frac{f_{X_1X_2}(u, v)}{\int_{\mathbb{R}} f_{X_1X_2}(s, v) ds}, \tag{3b2}$$

is the conditional p.d.f. of X_1, given X_2.

EXAMPLE 3b8. Let (X, Y) be a random vector with the p.d.f.

$$f_{XY}(x, y) = (x + y) I(x \in (0, 1)) I(y \in (0, 1)).$$

Let's calculate $\mathbb{E}(X|Y)$. To apply the Bayes formula we need the marginal p.d.f. f_Y:

$$f_Y(y) = \int_{\mathbb{R}} (x + y) I(x \in (0, 1)) I(y \in (0, 1)) dx = I(y \in (0, 1))(1/2 + y).$$

Hence for $x, y \in (0, 1)$, the conditional p.d.f. of X given Y is

$$f_{X|Y}(x; y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{x + y}{y + 1/2},$$

and zero otherwise. The conditional c.d.f. of X given Y is

$$F_{X|Y}(x; y) = \int_0^x f_{X|Y}(s; y) ds = \int_0^x \frac{s + y}{y + 1/2} ds = \frac{x^2/2 + yx}{y + 1/2}, \quad x, y \in (0, 1),$$

$F_{X|Y}(0-; y) = 0$ and $F_{X|Y}(1; y) = 1$. Further, for $y \in (0, 1)$

$$\mathbb{E}(X|Y = y) = \int_0^1 x f_{X|Y}(x; y) dx = \int_0^1 x \frac{x + y}{y + 1/2} dx = \frac{1/3 + y/2}{y + 1/2}.$$

Finally, the best MSE predictor of X given Y is given by the conditional expectation:

$$\mathbb{E}(X|Y) = \frac{1/3 + Y/2}{Y + 1/2}.$$

What if X is neither discrete nor continuous, but e.g. of a mixed type? Pay attention, that so far we haven't discussed the existence of the conditional expectation in general: perhaps, in certain situations there is no g^* which satisfies (3a3) ...? It turns out that the conditional expectation exists under very mild conditions [5] and, moreover, is often given by an abstract Bayes formula. However, the actual calculation of conditional expectations can be quite involved. In

[5] the conditional expectation of X given Y exists, if e.g. $\mathbb{E}|X| < \infty$. Note that nothing is mentioned of the conditioning r.v. Y.

some cases, which do not fit neither of the propositions above, the conditional expectation can be found directly from the property (3a3). Here is an example:

EXAMPLE 3b9. Let $X \sim \exp(\lambda)$ and $Y = \xi X$, where $\xi \sim \mathrm{Ber}(1/2)$, independent of X. We are interested to calculate $\mathbb{E}(X|Y)$. Note that (X, Y) is neither jointly continuous (since Y is not continuous) nor jointly discrete vector (since X is continuous) and hence formally all the Bayes formulae derived above are not applicable. In fact, using (3a3) to find $\mathbb{E}(X|Y)$ can be viewed as deriving the Bayes formula for this particular situation.

We are looking for a function $g^* : \mathbb{R} \mapsto \mathbb{R}$, such that

$$\mathbb{E}(X - g^*(Y))h(Y) = 0, \quad \forall h.$$

We have

$$\mathbb{E}Xh(Y) = \mathbb{E}Xh(\xi X) = \mathbb{E}Xh(\xi X)I(\xi = 0) + \mathbb{E}Xh(\xi X)I(\xi = 1) =$$
$$\mathbb{E}Xh(0)I(\xi = 0) + \mathbb{E}Xh(X)I(\xi = 1) = h(0)\frac{1}{2}\mathbb{E}X + \frac{1}{2}\mathbb{E}Xh(X) =$$
$$h(0)\frac{1}{2}\frac{1}{\lambda} + \frac{1}{2}\int_{\mathbb{R}} uh(u)f_X(u)du$$

Similarly,

$$\mathbb{E}g^*(Y)h(Y) = \mathbb{E}g^*(\xi X)h(\xi X) = \frac{1}{2}g^*(0)h(0) + \frac{1}{2}\int_{\mathbb{R}} g^*(u)h(u)f_X(u)du.$$

Hence

$$\mathbb{E}(X - g^*(Y))h(Y) = h(0)\frac{1}{2}\frac{1}{\lambda} + \frac{1}{2}\int_{\mathbb{R}} uh(u)f_X(u)du$$
$$- \frac{1}{2}g^*(0)h(0) - \frac{1}{2}\int_{\mathbb{R}} g^*(u)h(u)f_X(u)du =$$
$$\frac{1}{2}h(0)(1/\lambda - g^*(0)) + \frac{1}{2}\int_{\mathbb{R}} h(u)(u - g^*(u))f_X(u)du$$

The latter equals zero for *any* h if we choose $g^*(0) = 1/\lambda$ and $g^*(u) = u$ for $u \neq 0$. Hence

$$\mathbb{E}(X|Y = u) = \begin{cases} u & u \neq 0 \\ \frac{1}{\lambda} & u = 0 \end{cases}$$

and

$$\mathbb{E}(X|Y) = YI(Y \neq 0) + \frac{1}{\lambda}I(Y = 0).$$

The latter confirms the intuition: it is optimal predict X by the value of Y, whenever it is not 0, and predict $\mathbb{E}X = \frac{1}{\lambda}$ when $Y = 0$ (i.e. when Y doesn't tell anything about X). Think, however, how could you possibly get this answer by means of the Bayes formulae derived above ...?! ■

c. Conditioning of Gaussian vectors

Of a particular importance and simplicity are the formulae for conditional expectation for Gaussian random vectors.

PROPOSITION 3c1 (Normal Correlation Theorem). *Let X be a Gaussian r.v. in $\mathbb{R}^2$ as in Definition 2b6. Then $f_{X_1|X_2}(x_1;x_2)$ is Gaussian with the (conditional) mean:*

$$\mathbb{E}(X_1|X_2=x_2)=\mu_1+\frac{\rho\sigma_1}{\sigma_2}(x_2-\mu_2)$$

and the (conditional) variance:

$$\operatorname{var}(X_1|X_2=x_2)=\sigma_1^2(1-\rho^2).$$

REMARK 3c2. In the Gaussian case the optimal predictor is a *linear* function of the condition and the conditional variance does not depend on the condition at all. These remarkable properties make the Gaussian vectors very special in statistics.

PROOF. Direct application of the Bayes formula (3b2) to the Gaussian j.p.d.f. $f_X(x)$. □

Here is how the formulae look like in the general multivariate case:

PROPOSITION 3c3. *Let (X,Y) be a Gaussian vector with values in $\mathbb{R}^{k+m}$, where the first k coordinates are denoted by X and the rest $n-m$ coordinates are denoted by Y. Let $\mu=(\mu_X,\mu_Y)$ be the expectation vector and the covariance matrix*

$$S=\begin{pmatrix} S_X & S_{XY} \\ S_{YX} & S_Y \end{pmatrix}.$$

Assume that S_Y is nonsingular. Then the conditional distribution of X given Y is Gaussian with the mean

$$\mathbb{E}(X|Y)=\mu_X+S_{XY}S_Y^{-1}(Y-\mu_Y)$$

and the covariance

$$\operatorname{cov}(X|Y)=S_X-S_{XY}S_Y^{-1}S_{YX}.$$

PROOF. (a direct tedious calculation) □

d. Properties

Not less important than the computational techniques of the conditional expectation, are its properties listed below.

PROPOSITION 3d1. *Let X,Y,Z be random vectors, a,b be real constants and ψ,ϕ functions with appropriate domains of definition*

(1) *linearity:*
$$\mathbb{E}(aX+bY|Z)=a\mathbb{E}(X|Z)+b\mathbb{E}(Y|Z).$$
(2) $X\geq Y \quad\Longrightarrow\quad \mathbb{E}(X|Z)\geq\mathbb{E}(Y|Z)$
(3) $\mathbb{E}\Big(\mathbb{E}(X|Y,Z)\big|Z\Big)=\mathbb{E}(X|Z)$
(4) $\mathbb{E}\Big(\mathbb{E}(X|Z)\big|Y,Z\Big)=\mathbb{E}(X|Z)$
(5) $\mathbb{E}(X|a)=\mathbb{E}X$
(6) $\mathbb{E}\mathbb{E}(X|Y)=\mathbb{E}X$
(7) $\mathbb{E}(c|Y)=c$
(8) $\mathbb{E}(\phi(X)\psi(Y)|Y)=\psi(Y)\mathbb{E}(\phi(X)|Y)$

(9) *if X and Y are independent, then*

$$\mathbb{E}(\psi(X,Y)|Y) = \int \psi(x,Y)dF_X(x)$$

PROOF. All the properties are conveniently checked by applying the characterization by (3a3).

(1) Clearly $a\mathbb{E}(X|Z)+b\mathbb{E}(Y|Z)$ is a legitimate candidate for $\mathbb{E}(aX+bY|Z)$ as it is a function of Z. Let h be an arbitrary function, then

$$\mathbb{E}\Big(aX+bY-a\mathbb{E}(X|Z)-b\mathbb{E}(Y|Z)\Big)h(Z) = \\ a\mathbb{E}\Big(X-\mathbb{E}(X|Z)\Big)h(Z)+b\mathbb{E}\Big(Y-\mathbb{E}(Y|Z)\Big)h(Z)=0,$$

where we used linearity of the expectation 2a2 in the first equality and (3a3), applied to $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$ individually.

(2) by linearity, it is enough to show that $\xi \geq 0$ implies $\mathbb{E}(\xi|Z) \geq 0$ (with $\xi = X-Y$). Note that

$$\big(\xi-\mathbb{E}(\xi|Z)\big)I\big(\mathbb{E}(\xi|Z)\leq 0\big) = \\ \xi I\big(\mathbb{E}(\xi|Z)\leq 0\big)-\mathbb{E}(\xi|Z)I\big(\mathbb{E}(\xi|Z)\leq 0\big) \geq \xi I\big(\mathbb{E}(\xi|Z)\leq 0\big)\geq 0,$$

where we used $\xi \geq 0$ in the latter inequality. On the other hand, by the orthogonality property (3a3),

$$\mathbb{E}\big(\xi-\mathbb{E}(\xi|Z)\big)I\big(\mathbb{E}(\xi|Z)\leq 0\big)=0,$$

and hence with probability one

$$\big(\xi-\mathbb{E}(\xi|Z)\big)I\big(\mathbb{E}(\xi|Z)\leq 0\big)=0,$$

i.e.

$$\mathbb{E}(\xi|Z)I\big(\mathbb{E}(\xi|Z)\leq 0\big)=\xi I\big(\mathbb{E}(\xi|Z)\leq 0\big)\geq 0.$$

But then

$$\mathbb{E}(\xi|Z)=\mathbb{E}(\xi|Z)I\big(\mathbb{E}(\xi|Z)>0\big)+\mathbb{E}(\xi|Z)I\big(\mathbb{E}(\xi|Z)\leq 0\big)\geq 0,$$

as claimed.

(3) we have to check that

$$\mathbb{E}\Big(\mathbb{E}(X|Y,Z)-\mathbb{E}(X|Z)\Big)h(Z)=0$$

for an arbitrary h. Indeed,

$$\mathbb{E}\Big(\mathbb{E}(X|Y,Z)-X+X-\mathbb{E}(X|Z)\Big)h(Z) = \\ \mathbb{E}\Big(\mathbb{E}(X|Y,Z)-X\Big)h(Z)+\mathbb{E}\Big(X-\mathbb{E}(X|Z)\Big)h(Z)=0,$$

where we applied (3a3) to each term separately (note that $h(Z)$ can be seen as a function of (Z,Y)).

(4) holds since $\mathbb{E}(X|Z)$ is a function of (X,Y) and certainly

$$\mathbb{E}\Big(\mathbb{E}(X|Z)-\mathbb{E}(X|Z)\Big)h(Z,Y)=0.$$

(5) holds since
$$\mathbb{E}(X - \mathbb{E}X)h(a) = 0$$
for an arbitrary h.
(6) follows from $\mathbb{E}(X - \mathbb{E}(X|Y))h(Y) = 0$ for $h(y) := 1$
(7) $\mathbb{E}(c - c)h(Y) = 0$ for all h
(8) By (3a3)
$$\mathbb{E}\Big(\phi(X)\psi(Y) - \psi(Y)\mathbb{E}(\phi(X)|Y)\Big)h(Y) = \mathbb{E}\Big(\phi(X) - \mathbb{E}(\phi(X)|Y)\Big)\psi(Y)h(Y) = 0,$$
and the claim follows from (3a3) by arbitrariness of h.
(9) (straightforward)

□

Remark 3d2. Some of the properties are intuitive, in view of the optimality property (3a2) of the conditional expectation. For example, the best guess of a deterministic constant given anything is the constant itself (this choice yields zero MSE), which is the claim in (7).

Remark 3d3. Verifying the above properties by means of the Bayes formulae is more cumbersome (if at all possible): try e.g. to check (3).

Remark 3d4. As follows from the proofs, conditioning with respect to an arbitrary number of random variables enjoys the same properties.

These properties play the central role in calculations involving the conditional expectations. Here is one cute application:

Example 3d5. Let $X_1, ..., X_n$ be i.i.d. r.v. (not necessarily continuous or discrete!) and $S_n = \sum_{i=1}^n X_i$. We would like to calculate $\mathbb{E}(X_1|S_n)$. To this end, note that
$$S_n = \mathbb{E}(S_n|S_n) = \sum_{i=1}^n \mathbb{E}(X_i|S_n).$$
Set $S_{n\setminus i} = \sum_{j\neq i} X_j$ and notice that X_i and $S_{n\setminus i}$ are independent (why?) and $S_{n\setminus i}$ and $S_{n\setminus j}$ have the same distribution (as X_i's are identically distributed). Then for a bounded function h and any i,
$$\mathbb{E}X_i h(S_n) = \mathbb{E}X_i h(S_{n\setminus i} + X_i) = \int_{\mathbb{R}} xh(s+x)dF_{X_i}(x)dF_{S_{n\setminus i}}(s) = \int_{\mathbb{R}} xh(s+x)dF_{X_1}(x)dF_{S_{n\setminus 1}}(s) = \mathbb{E}X_1 h(S_n)$$
and, consequently,
$$\mathbb{E}\big(X_i - \mathbb{E}(X_1|S_n)\big)h(S_n) = \mathbb{E}\big(X_1 - \mathbb{E}(X_1|S_n)\big)h(S_n) = 0$$
which implies that $\mathbb{E}(X_i|S_n) = \mathbb{E}(X_1|S_n)$ by (3a3). Hence
$$S_n = \sum_{i=1}^n \mathbb{E}(X_i|S_n) = n\mathbb{E}(X_1|S_n)$$
and $\mathbb{E}(X_1|S_n) = S_n/n$. ■

EXAMPLE 3d6. Let X_1, X_2 and X_3 be i.i.d $N(0,1)$ r.v's and let

$$Y := \frac{X_1 + X_2 X_3}{\sqrt{1+X_3^2}}.$$

We would like to find the distribution of Y.

Since (X_1, X_2, X_3) is a Gaussian vector, the conditional distribution of (X_1, X_2), given X_3 is also Gaussian with the (conditional) mean

$$\begin{pmatrix} \mathbb{E}(X_1|X_3) \\ \mathbb{E}(X_2|X_3) \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where we used the independence of X_1 and X_3, and X_2 and X_3. The conditional covariance matrix of (X_1, X_2) given X_3 is

$$\begin{pmatrix} \mathrm{var}(X_1|X_3) & \mathrm{cov}(X_1, X_2|X_3) \\ \mathrm{cov}(X_1, X_2|X_3) & \mathrm{var}(X_2|X_3) \end{pmatrix} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2|X_3) \\ \mathrm{cov}(X_1, X_2|X_3) & \mathrm{var}(X_2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where we again used independence of X_1 and X_3, and X_2 and X_3 to conclude that e.g.

$$\mathrm{var}(X_1|X_3) = \mathbb{E}(X_1^2|X_3) - \big(\mathbb{E}(X_1|X_3)\big)^2 = \mathbb{E}(X_1^2) - \big(\mathbb{E}X_1\big)^2 = \mathrm{var}(X_1)$$

and independence of (X_1, X_2, X_3) to get

$$\mathrm{cov}(X_1, X_2|X_3) = \mathbb{E}(X_1X_2|X_3) - \mathbb{E}(X_1|X_3)\mathbb{E}(X_2|X_3) = \\ \mathbb{E}X_1X_2 - \mathbb{E}X_1\mathbb{E}X_2 = \mathrm{cov}(X_1, X_2) = 0.$$

To recap, the conditional distribution of (X_1, X_2) given X_3 is Gaussian with zero mean and unit covariance matrix. Hence

$$\begin{aligned} M_Y(t) =& \mathbb{E}\exp(tY) = \mathbb{E}\exp\Big(t\frac{X_1 + X_2X_3}{\sqrt{1+X_3^2}}\Big) = \\ & \mathbb{E}\mathbb{E}\Big(\exp\Big(t\frac{X_1 + X_2X_3}{\sqrt{1+X_3^2}}\Big)\Big|X_3\Big) = \\ & \mathbb{E}\mathbb{E}\Big(\exp\Big(tX_1\frac{1}{\sqrt{1+X_3^2}}\Big)\Big|X_3\Big)\mathbb{E}\Big(\exp\Big(tX_2\frac{X_3}{\sqrt{1+X_3^2}}\Big)\Big|X_3\Big) = \\ & \mathbb{E}\exp\Big(\frac{1}{2}t^2\frac{1}{1+X_3^2}\Big)\exp\Big(\frac{1}{2}t^2\frac{X_3^2}{1+X_3^2}\Big) = \\ & \mathbb{E}\exp\Big(\frac{1}{2}t^2\frac{1}{1+X_3^2} + \frac{1}{2}t^2\frac{X_3^2}{1+X_3^2}\Big) = \mathbb{E}\exp\Big(\frac{1}{2}t^2\Big) = \exp\Big(\frac{1}{2}t^2\Big). \end{aligned}$$

The latter is identified as the m.g.f. of an $N(0,1)$ r.v. ■

EXAMPLE 3d7. Consider X from Example 2b13 and suppose that θ is a random variable on $[0, 2\pi]$ (rather than a deterministic angle) with arbitrary distribution, independent of X. For $t \in \mathbb{R}^2$,

$$\mathbb{E}\exp\Big\{tU(\theta)X\Big\} = \mathbb{E}\mathbb{E}\Big(\exp\Big\{tU(\theta)X\Big\}|\theta\Big).$$

By the property 9,

$$\mathbb{E}\Big(\exp\Big\{tU(\theta)X\Big\}|\theta\Big) = \exp\Big\{t^\top U(\theta)U^\top(\theta)t\Big\} = \exp(\|t\|^2),$$

and hence

$$\mathbb{E}\exp\Big\{tU(\theta)X\Big\} = \exp(\|t\|^2),$$

which means that the standard Gaussian distribution is invariant under independent *random* rotations. ■

REMARK 3d8. Let (X, Y) be a continuous random vector. What do we mean by $\mathbb{P}(X \in [1,2]|Y = 5)$ ...? Certainly this cannot be interpreted as the conditional probability of the event $A = \{X \in [1,2]\}$, given the event $B = \{Y = 5\}$, since $\mathbb{P}(B) = 0$ and thus $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = 0/0$ is not well defined ! One natural way to deal with this is to define $\mathbb{P}(X \in [1,2]|Y = 5)$ by the limit

$$\mathbb{P}(X \in [1,2]|Y = 5) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(X \in [1,2], |Y-5| \le \varepsilon)}{\mathbb{P}(|Y-5| \le \varepsilon)}.$$

This approach actually works when X and Y are random variable taking real values, and yielding the aforementioned Bayes formulae. This approach, however, has serious disadvantages: (1) it doesn't work in a greater generality [6] and, more importantly for this course, it is not easy to derive the properties of the obtained object.

The standard modern approach is through conditional expectation: by definition, $\mathbb{P}(X \in [1,2]|Y) = \mathbb{E}\big(I(X \in [1,2])\big|Y\big)$ and the latter makes perfect sense: it is the essentially unique random variable, which is defined either by (3a2) or equivalently by (3a3). Moreover, $\mathbb{P}(X \in [1,2]|Y = u)$, $u \in \mathbb{R}$ is the function (called $g^*(u)$ above), which realizes the conditional expectation.

The following example is an illuminating manifestation[7]

EXAMPLE 3d9. Let U and V be i.i.d. r.v.'s with uniform distribution over the interval $(0,1)$. Define $D := U - V$ and $R := U/V$. We shall calculate $\mathbb{E}(U|D)$ and $\mathbb{E}(U|R)$ by the Bayes formulae. To this end, that the random vector (U, D) has the joint density

$$f_{UD}(x,y) = f_U(x)f_V(x-y) = I(x \in [0,1])I(x-y \in [0,1]), \quad x \in (0,1),\ y \in (-1,1)$$

and the p.d.f. of D is given by

$$f_D(y) = \int_0^1 f_{UD}(x,y)dx = \int_0^1 I(x-y \in [0,1])dx = 1 - |y|.$$

Hence by the Bayes formula

$$\mathbb{E}(U|D) = \frac{\int_0^1 x f_{UD}(x,D)dx}{f_D(D)} = \frac{\int_0^1 xI(x-D \in [0,1])dx}{1-|D|} =$$

$$\frac{1}{1-|D|}\begin{cases} \frac{1}{2}(D+1)^2 & D \in (-1,0] \\ \frac{1}{2} - \frac{1}{2}D^2 & D \in [0,1) \end{cases} = \frac{1}{2}(1+D).$$

[6] particularly, for random processes, etc.

[7] a folklore example, communicated to the author by Y.Ritov

To find $\mathbb{E}(U|R)$ we shall use the orthogonality characterization. To this end, for a bounded function h

$$\mathbb{E}Uh(R) = \int_0^1 x\Big(\int_0^1 h(x/y)dy\Big)dx = \int_0^1 x\Big(\int_x^\infty h(z)\frac{x}{z^2}dz\Big)dx = \int_0^\infty h(z)\frac{1}{z^2}\Big(\int_0^{1\wedge z} x^2dx\Big)dz = \int_0^\infty h(z)\frac{1}{z^2}\frac{1}{3}(1\wedge z)^3dz.$$

Similarly,

$$\mathbb{E}g^*(R)h(R) = \int_0^1\int_0^1 g^*(x/y)h(x/y)dydx = \int_0^\infty g^*(z)h(z)\frac{1}{z^2}\Big(\int_0^{1\wedge z} xdx\Big)dz = \int_0^\infty g^*(z)h(z)\frac{1}{z^2}\frac{1}{2}(1\wedge z)^2dz.$$

Hence the requirement

$$\mathbb{E}\big(U - g^*(R)\big)h(R) = 0, \quad \forall h$$

is met by the choice

$$g^*(z) = \frac{\frac{1}{3}(1\wedge z)^3}{\frac{1}{2}(1\wedge z)^2} = \frac{2}{3}(1\wedge z),$$

i.e.

$$\mathbb{E}(U|R) = \frac{2}{3}(1\wedge R).$$

Consequently,

$$\mathbb{E}(U|D=0) = 1/2, \quad \text{and} \quad \mathbb{E}(U|R=1) = \frac{2}{3}.$$

This may seem counterintuitive, since $\{R=1\} = \{D=0\} = \{U=V\}$, but we shall predict U differently on $\{R=1\}$ and $\{D=0\}$! This is no paradox, since we measure the quality of the prediction of U by the *mean* square error and not the individual errors for particular realization of R or D. In fact, for an arbitrary number $a\in\mathbb{R}$,

$$\mathbb{E}(U|D) = \begin{cases}\frac{1}{2}(1+D), & D\neq 0\\ a, & D=0\end{cases}, \quad \text{with prob. } 1,$$

and hence comparing $\mathbb{E}(U|D)$ and $\mathbb{E}(U|R)$ for individual realizations of U and R is meaningless.

To get an additional insight into the mysterious numbers $1/2$ and $2/3$, explore the limits

$$\lim_{\varepsilon\to 0}\frac{\mathbb{E}UI(|D|\le\varepsilon)}{\mathbb{P}(|D|\le\varepsilon)} = \frac{1}{2}, \quad \text{and} \quad \lim_{\varepsilon\to 0}\frac{\mathbb{E}UI(|R-1|\le\varepsilon)}{\mathbb{P}(|R-1|\le\varepsilon)} = \frac{2}{3}.$$

Geometrically, the first limit suggests to calculate the area of the linear strip

$$\{(x,y)\in(0,1)\times(0,1) : |x-y|\le\varepsilon\},$$

whose "center of mass" is at its mid point, corresponding to $1/2$, while the second limit has to do with the area of the sector

$$\{(x,y)\in(0,1)\times(0,1) : |x/y-1|\le\varepsilon\},$$

whose "center of mass" is shifted towards the "thick" end, yielding $2/3$.

Exercises

PROBLEM 3.1. Find $F_{X|Y}(x;y)$, $f_{X|Y}(x;y)$, $F_{Y|X}(y;x)$ and $f_{Y|X}(y;x)$ for the random vector from Problem 2.3

PROBLEM 3.2. Let (X,Y) be a random vector, such that $\mathbb{E}(Y|X=x)$ is in fact not a function of x. Show that $\mathrm{var}(Y) = \mathbb{E}\mathrm{var}(Y|X)$.

PROBLEM 3.3 (The law of total variance). For a pair of random variables X and Y, show that

$$\mathrm{var}(X) = \mathbb{E}\mathrm{var}(X|Y) + \mathrm{var}(\mathbb{E}(X|Y)),$$

if $\mathbb{E}X^2 < \infty$.

PROBLEM 3.4. Let $X \sim U([0,1])$ and suppose that the conditional law of Y given X is binomial $\mathrm{Bin}(n,X)$.

(1) Calculate $\mathbb{E}Y$
(2) Find the p.m.f. of Y for $n=2$
(3) Generalize to $n>2$ (Leave your answer in terms of the so called β-function:

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx,$$

which reduces to $\beta(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ for integer a and b)

PROBLEM 3.5. Let X,Y be a random variables with finite second moments and $\mathrm{var}(Y)>0$. Show that

$$\mathbb{E}(Y-aX-b)^2 \geq \mathrm{var}(X) - \frac{\mathrm{cov}^2(X,Y)}{\mathrm{var}(Y)}, \quad a,b \in \mathbb{R},$$

and that the minimum is attained by the optimal linear predictor:

$$a^* = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}, \quad b^* = \mathbb{E}X - a^*\mathbb{E}Y.$$

Explain how the latter can be used to predict X, given Y.

PROBLEM 3.6. Alice wants to transmit a message to Bob via a noisy communication channel. Let X be Alice's message and assume that it is a symmetric Bernoulli r.v. Bob gets $Y = X + Z$ at the output of the channel, where $Z \sim N(0,1)$, independent of X.

(1) Find the optimal linear predictor $\hat{X}$ of X given Y. Calculate the expectation of the error $X - \hat{X}$ and the MSE $\mathbb{E}(X-\hat{X})^2$. Suppose that Alice sent 1 and Bob obtained 1/2: what guess would Bob generate by the linear predictor ?
(2) Find the joint probability law of (X,Y). Is (X,Y) a Gaussian r.v. ?

(3) Find the optimal predictor $\tilde{X}$ of X given Y, among the nonlinear predictors of the form $\phi(Y)$ for some function ϕ. Calculate the expectation of the error $X - \tilde{X}$ and the MSE $\mathbb{E}(X - \tilde{X})^2$. Suppose that Alice sent 1 and Bob obtained 1/2: what guess would Bob generate by the linear predictor ?

PROBLEM 3.7. Answer the question from the previous problem, assuming $X \sim N(1/2, 1/4)$ (instead of $X \sim \mathrm{Ber}(1/2)$).

PROBLEM 3.8. Let (X, Y) be a Gaussian vector in $\mathbb{R}^2$ with the parameters $\mu_X = 5$, $\mu_Y = 10$, $\sigma_X = 1$, $\sigma_Y = 5$.

(1) Find $\rho(X, Y)$ if $\rho(X, Y) > 0$ and $\mathbb{P}(4 < Y < 16|X = 5) = 0.954...$
(2) If $\rho(X, Y) = 0$, find $\mathbb{P}(X + Y < 16)$

PROBLEM 3.9. Let (X, Y) be a Gaussian vector with the parameters μ_x,μ_Y,σ_X^2,σ_Y^2,ρ

(1) Is $\mathbb{E}(X|Y)$ a Gaussian r.v. ? Find its mean and variance
(2) Is $(X, \mathbb{E}(X|Y)$ a Gaussian vector ? Find its mean and covariance matrix.
(3) Calculate $\mathbb{E}\big(\mathbb{E}(X|Y)|X\big)$
(4) Calculate $\mathbb{E}\big(X|\mathbb{E}(X|Y)\big)$

Hint: remember that $\mathbb{E}(X|Y)$ is a linear function of Y

PROBLEM 3.10. Let $X \sim U([0, 1])$ and $Y = XI(X \geq 1/2)$. Find $\mathbb{E}(X|Y)$.

PROBLEM 3.11. n i.i.d. experiments with the success probability $p > 0$ are performed. Let X be the number of successes in the n experiments and Y be the number of successes in the first m experiments (of course, $m \leq n$).

(1) Find the j.p.m.f of (X, Y)
(2) Calculate the conditional p.m.f of Y, given X and identify it as one of the standard p.m.f.'s

PROBLEM 3.12. (*) Let X be a r.v. with the p.d.f. $f(x)$ and let $Y := g(X)$, where g is a piecewise strictly monotonous differentiable function (so that the set of roots $g^{-1}(y) = \{x : g(x) = y\}$ is a finite or countable set). Prove that the conditional law of X, given Y, is discrete and its p.m.f. is given by:

$$\mathbb{P}(X = x|Y) = \frac{f(x)/|g'(x)|}{\sum_{s \in g^{-1}(Y)} f(s)/|g'(s)|}, \quad x \in g^{-1}(Y),$$

where g' is the derivative of g. Think what changes if g has zero derivative on a nonempty interval ...?

PROBLEM 3.13. Let $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$ and set $Z = I(X \geq s)$ for some $s > 0$.

(1) Find the conditional law of X, given Z
(2) Calculate $\mathbb{E}(X|Z)$

PROBLEM 3.14. Consider the j.p.d.f.

$$f_{X,Y}(x,y) = \frac{1}{y} e^{-x/y} e^{-y} I(x \in (0,\infty)) I(y \in (0,\infty)).$$

(1) Find $f_Y(y)$ and identify it with one of the standard p.d.f.'s (**Hint:** no calculation is required)
(2) Find $f_{X|Y}(x;y)$ and identify it with one of the standard p.d.f.'s (**Hint:** no calculation is required) Find $\mathbb{E}(X|Y)$
(3) Verify that f_{XY} is indeed a j.p.d.f.
(4) Are X and Y independent ?

PROBLEM 3.15. Consider the j.p.d.f.

$$f_{XY}(x,y) = 2xy I\big((x,y) \in A\big), \quad A = \{(x,y) \in \mathbb{R}^2 : 0 \le x \le 2y \le 2\}$$

(1) Find f_X,f_Y, $f_{X|Y}$ and $f_{Y|X}$
(2) Find $\mathbb{E}(Y|X)$ and $\mathbb{E}(X|Y)$
(3) Find $\mathrm{var}(X)$, $\mathbb{E}(X^2|Y)$ and $\mathrm{var}(X|Y) := \mathbb{E}\Big(\big(X - \mathbb{E}(X|Y)\big)^2 \big| Y\Big)$
(4) Calculate $\mathrm{cov}(X,Y)$

PROBLEM 3.16. Let X and Y be real valued r.v.'s and ϕ a $\mathbb{R} \mapsto \mathbb{R}$ function.
(1) Show that $\mathbb{E}(X|Y) = \mathbb{E}(X|\phi(Y))$, if ϕ is one-to-one
(2) Give an example of X,Y and ϕ so that the claim in (1) fails to hold

CHAPTER 4

Transformations of random vectors

Let X be a random vector in $\mathbb{R}^n$ with known distribution. Then for a function $g : \mathbb{R}^n \mapsto \mathbb{R}^m$, $Y := g(X)$ is a random vector. How can we deduce the distribution of Y (in any convenient form, e.g. p.d.f., p.m.f. etc.) ? Such question arises for example if we are able to sample from one distribution and would like to obtain a sample from another. Another typical application is when we form an estimate of a parameter from a sample of large dimension and would like to have a compact probabilistic description of our estimator.

The answer to this question in general can be quite complicated, and we shall focus on a number of simple, but frequently encountered situations.

a. $\mathbb{R} \mapsto \mathbb{R}$ transformations

Let's explore the possible situations through a number of simple examples

EXAMPLE 4a1. Let $X \sim U([0,1])$ and $Y = I(X \leq 1/2)$ (i.e. $g(x) = I(x \leq 1/2)$). Clearly $Y \sim \text{Ber}(p)$ where

$$p = \mathbb{P}(Y = 1) = \mathbb{P}(X \leq 1/2) = 1/2.$$

■

This example demonstrates how a discrete r.v. is obtained from a continuous r.v. if g takes a countable (or finite) number of values. Here is a generalization:

PROPOSITION 4a2. *Let $p(k)$ be a p.m.f (supported on integers) and $X \sim U([0,1])$, then*

$$Y := \min\Big\{k \geq 0 : \sum_{i=0}^{k} p_i \geq X\Big\}$$

has p.m.f. $p(k)$.

PROOF.

$$\mathbb{P}(Y = j) = \mathbb{P}\Big(\sum_{i=0}^{j} p(i) \geq X, \sum_{i=0}^{j-1} < X\Big) = \int_{\sum_{i=0}^{j-1} p(i)}^{\sum_{i=0}^{j} p(i)} dx = \sum_{i=0}^{j} p(i) - \sum_{i=0}^{j-1} p(i) = p(j).$$

□

Can we get a continuous r.v. from a discrete one ? Obviously not:

PROPOSITION 4a3. *Let X be a r.v. with p.m.f. p_X and $g : \mathbb{R} \mapsto \mathbb{R}$, then $Y = g(X)$ is discrete, and*

$$p_Y(i) = \sum_{j : g(x_j) = y_i} p_X(j).$$

PROOF. Let $\{x_1, x_2, ...\}$ be the set of values of X, then Y takes values in the discrete set $\{g(x_1), g(x_2), ...\}$. Moreover,

$$p_Y(i) = \mathbb{P}(Y = y_i) = \mathbb{P}(X \in \{x_j : g(x_j) = y_i\}) = \mathbb{P}\left(\bigcup_{x_j : g(x_j) = y_i} \{X = x_j\}\right) = \sum_{j : g(x_j) = y_i} p_X(j)$$

□

Here is a particular, but important, example of a different flavor:

PROPOSITION 4a4. *Let F and G be continuous and strictly increasing c.d.f's and let X be a r.v. sampled from F, then $U = F(X)$ has uniform distribution on $[0, 1]$ and $Y := G_Y^{-1}(F_X(X))$ is a sample from G.*

PROOF. If X is a r.v. with a strictly increasing c.d.f. $F(x)$, then for $U := F(X)$

$$\mathbb{P}(U \le x) = \mathbb{P}(F(X) \le x) = \mathbb{P}(X \le F^{-1}(x)) = F(F^{-1}(x)) = x, \quad x \in [0, 1]$$

and $F(0-) = 0$ and $F(x) = 1$, $x \ge 1$.

Further, if U is a r.v. with uniform distribution on $[0, 1]$ and $G(x)$ is a strictly increasing continuous c.d.f., then

$$\mathbb{P}(Y \le v) = \mathbb{P}(G^{-1}(U) \le v) = \mathbb{P}(U \le G(v)) = \int_0^{G(v)} ds = G(v).$$

□

REMARK 4a5. The typical application of this Proposition 4a4 is the following: suppose we can generate r.v. X with a particular distribution F_X (e.g. uniform or normal etc.) and we want to generate r.v. Y with a different distribution F_Y. If F_X and F_Y satisfy the above conditions, this can be done by setting $Y := F_Y^{-1}(F_X(X))$.

Let's now consider the setting, when a r.v. with p.d.f. is mapped by a differentiable one-to-one function g:

PROPOSITION 4a6. *Let X be a r.v. with p.d.f. $f_X(x)$ and let $Y = g(X)$, where g is a differentiable and strictly monotonous function on the interior[1] of the support of f_X. Then Y has the p.d.f.*

$$f_Y(v) = \frac{f_X\big(g^{-1}(v)\big)}{\big|g'\big(g^{-1}(v)\big)\big|}, \quad v \in \mathbb{R}.$$

PROOF. Suppose g increases, then

$$F_Y(v) := \mathbb{P}(Y \le v) = \mathbb{P}(g(X) \le v) = \mathbb{P}(X \le g^{-1}(v)) = F_X(g^{-1}(v)), \tag{4a1}$$

[1]Recall that the interior of a subset $A \subseteq \mathbb{R}$, denoted by A°, is the set of all internal points of A, whereas a point $x \in A$ is internal, if there is an open interval V_x, such that $x \in V_x \subseteq A$. For example, 0.5 is an interval point of $[0, 1)$, while 0 and 1 are not.

where the last equality follows from the strict monotonicity of g. Since g is invertible, i.e. $g(g^{-1}(x)) = x$, $x \in \mathbb{R}$ and as g is differentiable and strictly increasing, taking derivative of the latter identity we get $g'\big(g^{-1}(x)\big)\frac{d}{dx}g^{-1}(x) = 1$ and hence

$$\frac{d}{dx}g^{-1}(x) = \big[g'\big(g^{-1}(x)\big)\big]^{-1}.$$

Now differentiating (4a1) w.r.t. v, we get

$$f_Y(v) := \frac{d}{dv}F_Y(v) = f_X\big(g^{-1}(v)\big)\big[g'\big(g^{-1}(v)\big)\big]^{-1}.$$

Notice that the latter expression is indeed positive as it should be. For strictly decreasing g we get the same answer with minus (so that the obtained expression is again positive). □

EXAMPLE 4a7. Let $X \sim U([0,1])$, i.e. $f_X(x) = I(x \in [0,1])$ and define $Y = 1/X$. Note that the function $g(x) = 1/x$ is strictly decreasing and differentiable on $(0,1)$. Moreover, $g'(x) = (1/x)' = -1/x^2$ and $g^{-1}(v) = 1/v$, $v \in (1,\infty)$, hence

$$f_y(v) = \frac{1}{v^2}I(g^{-1}(v) \in [0,1]) = \frac{1}{v^2}I(v > 1).$$

Check e.g. that $\int_1^\infty f_Y(v)dv = 1$ and $\mathbb{E}Y = \infty$. ■

The following generalizes the latter proposition to the setting, when the transformation g is only piecewise monotone

PROPOSITION 4a8. *Let X be a r.v. with p.d.f $f_X(x)$ and g is a function of the form*

$$g(x) = \sum_{i=1}^{m} I(x \in A_i)g_i(x), \quad m \geq 1$$

where A_i, $i = 1, ..., m$ are pairwise disjoint intervals partitioning the support of f_X, and where $g_i(x)$ are differentiable and strictly monotonous[2] functions on A_i° (the interiors of A_i) respectively. Then $Y = g(X)$ has the p.d.f.

$$f_Y(v) = \sum_{i=1}^{m} \frac{f_X\big(g_i^{-1}(v)\big)}{\big|g_i'\big(g_i^{-1}(v)\big)\big|} I(g_i^{-1}(v) \in A_i^\circ), \quad v \in \bigcup_{i=1}^{m} A_i^\circ. \tag{4a2}$$

REMARK 4a9. Note that Proposition 4a6 is a particular case of the latter proposition, corresponding to $m = 1$, i.e when g is monotonous on all the support.

REMARK 4a10. Note that f_Y in (4a2) remains undefined outside the open set $\cup_{j=1}^m A_j^\circ$, which consists of a finite number of points. At these points f_Y can be defined arbitrarily without affecting any probability calculations, it is involved in (why?)

PROOF. Since g_i is a monotonous function and A_i is an interval, say $A_i := [a_i, b_i]$, the image of A_i under g_i is the interval $[g_i(a_i), g_i(b_i)]$, if g_i increases and $[g(b_i), g(a_i)]$, if g decreases. For a fixed i and $v \in \mathbb{R}$, consider the quantity

$$\mathbb{P}(Y \leq v, X \in A_i) = \mathbb{P}(g_i(X) \leq v, X \in A_i),$$

[2]for simplicity, we shall assume that $g'(x) > 0$, if g is increasing and $g'(x) < 0$ if it is decreasing, for all $x \in \mathbb{R}$

and suppose for definiteness that g_i increases on A_i. Then

$$\mathbb{P}(Y \le v, X \in A_i) = \begin{cases} 0 & v \le g(a_i) \\ \int_{a_i}^{g_i^{-1}(v)} f_X(x)dx & v \in \big(g(a_i), g(b_i)\big) \\ \mathbb{P}(X \in A_i) & v \ge g(b_i) \end{cases}$$

and hence

$$\frac{d}{dv}\mathbb{P}(Y \le v, X \in A_i) = \begin{cases} f_X(g_i^{-1}(v))\frac{d}{dv}g_i^{-1}(v) & v \in \big(g(a_i), g(b_i)\big) \\ 0 & \text{otherwise} \end{cases} =$$

$$\begin{cases} f_X(g_i^{-1}(v))\Big[g_i'\big(g_i^{-1}(v)\big)\Big]^{-1} & v \in \big(g(a_i), g(b_i)\big) \\ 0 & \text{otherwise} \end{cases}$$

Similar non-negative expression is obtained, if g_i decreases on A_i. The claimed formula now follows from the total probability decomposition:

$$\mathbb{P}(Y \le v) = \sum_i \mathbb{P}(Y \le v, X \in A_i)$$

□

Example 4a11. In many cases it is convenient to bypass the general formula and act directly. Let $X \sim N(0,1)$ and let $Y = X^2$. Then obviously $\mathbb{P}(Y \le v) = 0$ if $v \le 0$ and for $v > 0$

$$F_Y(v) = \mathbb{P}(Y \le v) = \mathbb{P}(X^2 \le v) = \mathbb{P}(X \ge -\sqrt{v}, X \le \sqrt{v}) = F_X(\sqrt{v}) - F_X(-\sqrt{v}).$$

Differentiating w.r.t v we get

$$f_Y(v) = \frac{d}{dv}F_Y(v) = f_x(\sqrt{v})\frac{1}{2}\frac{1}{\sqrt{v}} + f_X(-\sqrt{v})\frac{1}{2}\frac{1}{\sqrt{v}}.$$

Now with $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ we obtain:

$$f_Y(v) = \frac{1}{\sqrt{v}}f_X(\sqrt{v}) = \frac{1}{\sqrt{v}}\frac{1}{\sqrt{2\pi}}e^{-v/2}I(v > 0). \tag{4a3}$$

Now let's get the same answer by applying the general recipe from the last proposition. The function $g(x) = x^2$ is decreasing on $(-\infty, 0]$ and increasing on $(0, \infty)$ (note that inclusion of $\{0\}$ to either intervals is not essential, since we deal only with the interiors of A_i's). We have $g_1^{-1}(v) = -\sqrt{v}$ and $g_2^{-1}(v) = \sqrt{v}$ and $g_i'(x) = 2x$. For $v < 0$, $f_Y(v) = 0$ since v does not belong neither to $g(A_1) = [0, \infty)$ nor to $g(A_2) = (0, \infty)$. For $v > 0$ the formula yields

$$f_Y(v) = f_X(-\sqrt{v})\frac{1}{2\sqrt{v}}I(v \in (0,\infty)) + f_X(\sqrt{v})\frac{1}{2\sqrt{v}}I(v \in (0,\infty)),$$

which is the same p.d.f. as in (4a3). ■

b. Some special $\mathbb{R}^n \mapsto \mathbb{R}$ transformations

min **and** max**.**

PROPOSITION 4b1. *Let $(X_1, ..., X_n)$ be i.i.d. r.v. with common c.d.f F and let $X_{(1)} := \min(X_1, ..., X_n)$ and $X_{(n)} := \max(X_1, ..., X_n)$. Then*

$$F_{X_{(n)}}(x) = F^n(x), \quad F_{X_{(1)}}(x) = 1 - \Big(1 - F(x)\Big)^n.$$

If F has p.d.f. f, then

$$f_{X_{(n)}}(x) = nf(x)F^{n-1}(x), \quad f_{X_{(1)}}(x) = nf(x)\Big(1 - F(x)\Big)^{n-1}.$$

PROOF. We have

$$F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \le x) = \mathbb{P}(X_1 \le x, ..., X_n \le x) = \prod_{i=1}^{n} \mathbb{P}(X_i \le x) = \mathbb{P}^n(X_1 \le x) = F^n(x),$$

and

$$1 - F_{X_{(1)}}(x) = \mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, ..., X_n > x) = \prod_{i} \mathbb{P}(X_i > x) = \Big(1 - F(x)\Big)^n.$$

The expressions for the corresponding p.d.f.'s are obtained by differentiating. □

EXAMPLE 4b2. Let $X_1 \sim U([0,1])$, i.e. $f(x) = I(x \in [0,1])$. Then by the above formulae

$$f_{X_{(n)}}(x) = nx^{n-1} I(x \in [0,1])$$

and

$$f_{X_{(1)}}(x) = n\big(1 - x\big)^{n-1} I(x \in [0,1]).$$

Note that the p.d.f. of min concentrates around 0, while p.d.f. of max is more concentrated around 1 (think why).

Sum.

PROPOSITION 4b3. *For a pair of real valued random variables X and Y with the joint p.d.f. $f_{XY}(x, y)$, the sum $S = X + Y$ has the p.d.f., given by the convolution integral*

$$f_S(u) = \int_{\mathbb{R}} f_{XY}(x, u - x)dx = \int_{\mathbb{R}} f_{XY}(u - x, x)dx.$$

PROOF. For $u \in \mathbb{R}$

$$\mathbb{P}(S \le u) = \mathbb{P}(X + Y \le u) = \iint_{\mathbb{R}^2} I(s + t \le u) f_{XY}(s, t)dsdt =$$

$$\int_{\mathbb{R}} \Big(\int_{\mathbb{R}} I(s \le u - t) f_{XY}(s, t)ds \Big)dt = \int_{\mathbb{R}} \Big(\int_{-\infty}^{u-t} f_{XY}(s, t)ds \Big)dt =$$

$$\int_{\mathbb{R}} \Big(\int_{-\infty}^{u} f_{XY}(s' - t, t)ds' \Big)dt = \int_{-\infty}^{u} \Big(\int_{\mathbb{R}} f_{XY}(s' - t, t)dt \Big)ds'.$$

Taking the derivative w.r.t. u, we obtain the claimed formula. □

REMARK 4b4. If X and Y take integer values, $S = X + Y$ is an integer valued r.v. with the p.m.f. given by the convolution

$$p_S(k) = \sum_m p_{XY}(k - m, m).$$

REMARK 4b5. If X and Y are i.i.d. with common p.d.f. f (or p.m.f), then $f_S(u) = \int_{\mathbb{R}} f(u-x, x)dx =: f * f$. By induction, if $X_1, ..., X_n$ are i.i.d. $f_S = f^{*n}$ is the n-fold convolution of f.

EXAMPLE 4b6. Let X and Y be i.i.d. r.v. with common uniform distribution $U([0, 1])$. Then

$$f_S(u) = \int_{\mathbb{R}} I(u - x \in (0, 1))I(x \in (0, 1))dx = \int_0^1 I(u - x \in (0, 1))dx = \int_0^1 I(x \in (u - 1, u))dx.$$

If $u < 0$, then $(u-1, u)\cap(0, 1) = \emptyset$ and hence the integral is zero. If $u \in [0, 1)$, $(u-1, u)\cap(0, 1) = (0, u)$ and the integral yields u. If $u \in [1, 2)$, then $(u - 1, u) \cap (0, 1) = (u - 1, 1)$ and the integral gives $2 - u$. Finally, for $u \geq 2$, $(u - 1, u) \cap (0, 1) = \emptyset$ and the integral is zero again. Hence we get:

$$f_S(u) = (1 - |u - 1|)I(u \in [0, 2]).$$

Calculating the convolution integral (sum) can be tedious. Sometimes it is easier to tackle the problem by means of the m.g.f.:

PROPOSITION 4b7. *Let $X_1, ..., X_n$ be i.i.d. r.v. with the common m.g.f $M(t)$. Then $S = \sum_{i=1}^n X_i$ has the m.g.f.*

$$M_S(t) = M^n(t).$$

PROOF.

$$M_S(t) = \mathbb{E}e^{St} = \mathbb{E}e^{\sum_{i=1}^n X_i t} = \mathbb{E}\prod_{i=1}^n e^{X_i t} = \prod_{i=1}^n \mathbb{E}e^{X_i t} = \big(\mathbb{E}e^{X_1 t}\big)^n = M^n(t),$$

where we used independence and identical distribution of X_i's. □

REMARK 4b8. Similarly, if X_1 is a discrete r.v. with p.g.f. $G_X(s)$, the sum S is also discrete with the p.g.f. $G_S(s) = G_X^n(s)$.

In many cases $M^n(t)$ can be identified with some standard known m.g.f. Here are some examples:

EXAMPLE 4b9. Let $X_1 \sim \text{Ber}(p)$. Then

$$G_S(s) = \big(ps + 1 - p\big)^n.$$

Hence

$$p_S(0) = G_S(0) = (1 - p)^n$$
$$p_S(1) = G'_S(0) = np\big(ps + 1 - p\big)^{n-1}_{|s=0} = np(1 - p)^{n-1}$$
$$...$$

which recovers the familiar $\text{Bin}(n,p)$.

EXAMPLE 4b10. Let $X_1 \sim \text{Poi}(\lambda)$. The corresponding m.g.f. is

$$M(t) = \exp\big\{\lambda(e^t - 1)\big\}.$$

Hence

$$M_S(t) = M^n(t) = \exp\big\{n\lambda(e^t - 1)\big\},$$

which is recognized as the m.g.f. of the Poisson distribution $\text{Poi}(n\lambda)$. Hence $S \sim \text{Poi}(n\lambda)$. Similarly, if $X_1, ..., X_n$ are independent r.v. with $X_i \sim \text{Poi}(\lambda_i)$, then $S \sim \text{Poi}\Big(\sum_i \lambda_i\Big)$.

EXAMPLE 4b11. If $X_1, ..., X_n$ are independent Gaussian random variables $X_i \sim N(\mu_i, \sigma_i^2)$, then $S \sim N(\sum \mu_i, \sum_i \sigma_i^2)$ (which can also be deduced from the fact that Gaussian distribution in $\mathbb{R}^n$ is stable under linear transformations).

c. Differentiable $\mathbb{R}^n \mapsto \mathbb{R}^n$ transformations

Let $X = (X_1, ..., X_n)$ be a random vector with j.p.d.f $f_X(x)$, $x \in \mathbb{R}^n$ and let $g : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a given function. The following proposition gives the formula for the j.p.d.f. of $Y = g(X)$ for appropriate g's.

PROPOSITION 4c1. *Let $X_1, ..., X_n$ be a random vector with continuous joint p.d.f. $f_X(x)$, $x \in \mathbb{R}^n$. Denote by D the support of f_X in $\mathbb{R}^n$:*

$$D := \text{cl}\big\{x \in \mathbb{R}^n : f_X(x) > 0\big\},$$

and consider a function[3] $g : D \mapsto \mathbb{R}^n$. Suppose that there are pairwise disjoint subsets D_i, $i = 1, ..., m$, such that the set $D \setminus \cup_{i=1}^m D_i$ has probability zero:

$$\int ... \int_{D \setminus \cup_{i=1}^m D_i} f_X(x)dx = 0$$

and the function g is one-to-one on all D_i's. Let g_i^{-1} be the inverse of g on D_i and define the corresponding Jacobians

$$J_i(y) = \det \begin{pmatrix} \frac{\partial}{\partial y_1} g_{i1}^{-1}(y) & \frac{\partial}{\partial y_2} g_{i1}^{-1}(y) & ... & \frac{\partial}{\partial y_n} g_{i1}^{-1}(y) \\ \frac{\partial}{\partial y_1} g_{i2}^{-1}(y) & \frac{\partial}{\partial y_2} g_{i2}^{-1}(y) & ... & \frac{\partial}{\partial y_n} g_{i2}^{-1}(y) \\ ... & ... & ... & ... \\ \frac{\partial}{\partial y_1} g_{in}^{-1}(y) & \frac{\partial}{\partial y_2} g_{in}^{-1}(y) & ... & \frac{\partial}{\partial y_n} g_{in}^{-1}(y) \end{pmatrix}, \quad y \in D_i, \quad i = 1, ..., m.$$

Assume that all the partial derivatives above are continuous on[4] $g(D_i)$'s and $J_i(y) \neq 0$ for all $y \in g(D_i)$, for all $i = 1, ..., m$. Then

$$f_Y(y) = \sum_{i=1}^m |J_i(y)| f_X(g_i^{-1}(y)), \quad y \in \cup_i g(D_i).$$

PROOF. A change of variable for n-fold integrals (tools from n-variate calculus). □

[3] the domain of g might differ from D by a set of points in $\mathbb{R}^n$ with zero probability - see the examples below.

[4] for a subset $C \subseteq \mathbb{R}^n$ and a function $h : \mathbb{R}^n \mapsto \mathbb{R}$, $g(C)$ denotes the image of C under g, i.e. $g(C) := \{g(x), x \in C\}$

REMARK 4c2. Pay attention that for $n = 1$, the latter reduces to the claim of Proposition 4a8.

EXAMPLE 4c3. Let X_1 and X_2 be i.i.d. r.v. with $N(0,1)$ distribution. We would like to find the j.p.d.f. of $Y_1 = X_1 + X_2$ and $Y_2 = X_1/X_2$. In terms of the ingredients of the proposition, $f_X(x) = \frac{1}{\pi} e^{-x_1^2/2 - x_2^2/2}$, $x \in \mathbb{R}^2$, $D = \mathbb{R}^2$ and $g(x) = (x_1 + x_2, x_1/x_2) : \mathbb{R}^2 \mapsto \mathbb{R}^2$. Note that the domain of g is $\mathbb{R}^2 \setminus \{(x_1, x_2) \in \mathbb{R} : x_2 = 0\}$, i.e. g is defined on the plane off the line $\alpha := \{x_2 = 0\}$ (which has probability zero). The inverse function can be found from the system of equations

$$y_1 = x_1 + x_2$$
$$y_2 = x_1/x_2$$

These yield $x_1 = y_2 x_2$ and $y_1 = (y_2 + 1)x_2$. If $y_2 \neq -1$, then

$$x_2 = y_1/(y_2 + 1)$$
$$x_1 = y_2 y_2/(y_2 + 1).$$

If $y_2 = -1$, i.e. $x_1 = -x_2$, then $y_1 = 0$: the whole line $\ell := \{(x_1, x_2) : x_1 = -x_2\}$ is mapped to a single point $(0, -1)$. Hence g is invertible on $\mathbb{R}^2 \setminus (\ell \cup \alpha)$ with

$$g^{-1}(y) = \Big(y_1 y_2/(y_2 + 1), y_1/(y_2 + 1)\Big).$$

Since $\int_\ell f_X(x)dx = 0$, the natural choice of the partition is just $D_1 = \mathbb{R}^2 \setminus (\ell \cup \alpha)$ and $g_1(y) := g(y)$. The range of D_1 under g_1 is $\mathbb{R} \setminus \{(0, -1)\}$. Let's calculate the Jacobian:

$$J = \det \begin{pmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(y) & \frac{\partial}{\partial y_2} g_1^{-1}(y) \\ \frac{\partial}{\partial y_1} g_2^{-1}(y) & \frac{\partial}{\partial y_2} g_2^{-1}(y) \end{pmatrix} = \det \begin{pmatrix} \frac{\partial}{\partial y_1} y_1 y_2/(y_2+1) & \frac{\partial}{\partial y_2} y_1 y_2/(y_2+1) \\ \frac{\partial}{\partial y_1} y_1/(y_2+1) & \frac{\partial}{\partial y_2} y_1/(y_2+1) \end{pmatrix} =$$
$$\det \begin{pmatrix} y_2/(y_2+1) & y_1/(1+y_2)^2 \\ 1/(y_2+1) & -y_1/(y_2+1)^2 \end{pmatrix} = -y_2 y_1/(y_2+1)^3 - y_1/(1+y_2)^3 = \frac{-y_1}{(1+y_2)^2}.$$

Now we are prepared to apply the formula: for $y \in \mathbb{R} \setminus (0, -1)$

$$f_Y(y) = |J(y)| f_X(g^{-1}(y)) =$$
$$\frac{|y_1|}{(1+y_2)^2} \frac{1}{2\pi} \exp\Big\{ -\frac{1}{2}\Big(y_1 y_2/(y_2+1)\Big)^2 - \frac{1}{2}\Big(y_1/(y_2+1)\Big)^2 \Big\} =$$
$$\frac{|y_1|}{2\pi(1+y_2)^2} \exp\Big\{ -\frac{1}{2} \frac{y_1^2(y_2^2+1)}{(y_2+1)^2} \Big\}.$$

One can check that the Y_1 marginal is $N(0, 2)$. Let's calculate the Y_2 marginal p.d.f.:

$$f_{Y_2}(y_2) = \int_{\mathbb{R}} f_{Y_1 Y_2}(y_1, y_2) dy_1 = \int_{\mathbb{R}} \frac{|y_1|}{2\pi(1+y_2)^2} \exp\Big\{ -\frac{1}{2} \frac{y_1^2(y_2^2+1)}{(y_2+1)^2} \Big\} dy_1 =$$
$$\frac{1}{2\pi(1+y_2)^2} 2 \int_0^\infty y_1 \exp\Big\{ -\frac{1}{2} \frac{y_1^2(y_2^2+1)}{(y_2+1)^2} \Big\} dy_1 = \ldots = \frac{1}{\pi} \frac{1}{1+y_2^2},$$

i.e. Y_2 has standard Cauchy distribution. ■

Exercises

PROBLEM 4.1. Let $F_{XY}(u,v)$, $u,v \in \mathbb{R}^2$ be a continuous j.c.d.f., such that both $v \mapsto F_Y(v)$ and $u \mapsto F_{X|Y}(u;v)$, $v \in \mathbb{R}$ are strictly increasing. Suggest a way to produce a sample from F_{XY}, given a sample of i.i.d. r.v's U and V with uniform distribution on $[0,1]$.

PROBLEM 4.2. Let $X_1, ..., X_n$ be i.i.d. random variables with the common distribution F. Let $X_{(1)}, ..., X_{(n)}$ be the permutation of $X_1, ..., X_n$, such that $X_{(1)} \leq ... \leq X_{(n)}$. $X_{(i)}$ is called the i-th *order statistic* of $X_1, ..., X_n$.

(1) Show that the c.d.f. of $X_{(i)}$ is given by

$$F_{X_{(i)}}(x) = \sum_{j=i}^{n} F(x)^j \big(1 - F(x)\big)^{n-j}.$$

(2) Show that if F has the p.d.f. f, then $X_{(i)}$ has the p.d.f.

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}(1 - F(x))^{n-i} f(x).$$

(3) Discuss the cases $i = 1$ and $i = n$.

(4) Assuming that F has the p.d.f. f, show that $X' := (X_{(1)}, ..., X_{(n)})$ has the j.p.d.f. given by

$$f_{X'}(x) = \begin{cases} n! \prod_{i=1}^{n} f(x_i), & x_1 < ... < x_n \\ 0 & \text{otherwise} \end{cases}$$

Hint: define the sets S_i, $i = 1, ..., n!$,

$$S_1 := \{x \in \mathbb{R}^n : x_1 < x_2 < ... < x_n\}$$
$$S_2 = \{x \in \mathbb{R}^n : x_2 < x_1 < ... < x_n\}$$

and so on. Use the Jacobian formula.

PROBLEM 4.3. Let X and Y be r.v. and define $Z = X - Y$.

(1) Find the p.d.f. of X, assuming that (X, Y) has j.p.d.f.

(2) Find the p.m.f. of X, assuming that (X, Y) has j.p.m.f.

PROBLEM 4.4. Find the p.d.f. of $Z = X/Y$, if X and Y are r.v. with j.p.d.f. $f(x,y)$, $(x,y) \in \mathbb{R}^2$

PROBLEM 4.5. Let $X \sim N(0, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$ be independent r.v. Show that the p.d.f. of $Z = X/Y$ is Cauchy:

$$f_Z(x) = \frac{1}{\pi\gamma(1 + (t/\gamma)^2)}, \quad x \in \mathbb{R},$$

and find the corresponding scaling parameter γ.

Hint: Introduce $\tilde{X} = X/\sigma_1$ and $\tilde{Y} = Y/\sigma_2$, show that $\tilde{X}/\tilde{Y}$ has standard Cauchy density (i.e. with $\gamma = 1$) and deduce the claim

PROBLEM 4.6. Let $X_1, ..., X_n$ be i.i.d. r.v. with the common exponential distribution of rate $\lambda > 0$.

(1) Show that $X_{(1)} = \min(X_1, ..., X_n)$ has exponential distribution and find its rate.
(2) Show that $S = \sum_{i=1}^n X_i$ has Γ distribution and find its parameters

PROBLEM 4.7. Let X and Y be i.i.d. standard Gaussian r.v. Define $R := \sqrt{X^2 + Y^2}$, the distance of the random point (X, Y) from the origin and $\phi := \arctan(Y/X)$, the angle formed by the vector (X, Y) and the x-axis.

(1) Prove that R and ϕ are independent, $\phi \sim U([0, 2\pi])$ and R has the Rayleigh p.d.f.:

$$f_R(x) = re^{-r^2/2} I(r \geq 0).$$

Hint: the function $g(x, y) = \left(\sqrt{x^2 + y^2}, \arctan(y/x)\right)$ is invertible and its inverse is given by

$$g^{-1}(r, \phi) = (r\cos\phi, r\sin\phi), \quad (r, \phi) \in \mathbb{R}_+ \times [0, 2\pi)$$

(2) Let R and ϕ be independent r.v. with Rayleigh and $U([0, \pi])$ distributions. Show that (X, Y) are i.i.d. standard Gaussian r.v.

Hint: the transformation is one-to-one and onto.

PROBLEM 4.8. Let U and V be i.i.d. r.v. with the common distribution $U, V \sim U([0, 1])$. Show that

$$Y = \sqrt{-2\ln U}\sin(2\pi V), \quad X = \sqrt{-2\ln U}\cos(2\pi V),$$

are i.i.d. standard Gaussian r.v.[5]

Hint: define $R = \sqrt{X^2 + Y^2}$ and $\phi = \arctan(X/Y)$ and show that R and ϕ are independent and have the Rayleigh and $U([0, 2\pi])$ distributions. Refer to the previous problem.

PROBLEM 4.9. Let $X \sim U([0, 1])$. Find an appropriate function g, so that $Y = g(X)$ has each of the following distributions:

(1) $U([a, b])$, $b > a \in \mathbb{R}$.
(2) uniform distribution on the points $\{x_1, ..., x_n\} \subseteq \mathbb{R}$
(3) Poi(λ)

[5]this suggests a way to generate a pair of i.i.d. Gaussian r.v. from a pair of i.i.d. uniformly distributed r.v.

(4) exponential distribution with rate λ
(5) standard Cauchy distribution
(6) $N(\mu, \sigma^2)$ (express your answer in terms of the standard Gaussian c.d.f. $\Phi(x)$)

PROBLEM 4.10. Let X be a r.v. with c.d.f. F. Express the c.d.f. of $Y = \max(X, 0)$ in terms of F

PROBLEM 4.11. Let X and Y be r.v. with finite expectations. Prove or disprove[6]:
(1) $\mathbb{E}\max(X, Y) \geq \max(\mathbb{E}X, \mathbb{E}Y)$
Hint: note that $\max(X, Y) \geq X$ and $\max(X, Y) \geq Y$
(2) $\mathbb{E}\max(X, Y) + \mathbb{E}\min(X, Y) = \mathbb{E}(X + Y)$

PROBLEM 4.12. Let X and Y be i.i.d. standard Gaussian r.v. Show that $2XY$ and $X^2 - Y^2$ have the same probability laws.
Hint: $X^2 - Y^2 = 2\frac{X-Y}{\sqrt{2}}\frac{X+Y}{\sqrt{2}}$

PROBLEM 4.13. Let $X_1, ..., X_n$ be independent r.v., $X_i \sim \text{Poi}(\lambda_i)$. Show that the conditional p.m.f. of X_1, given $S = \sum_{i=1}^n X_i$ is Binomial and find the corresponding parameters.

[6]this problem demonstrates that sometimes a trick is required to avoid heavy calculations ;)

CHAPTER 5

A preview: first applications to Statistics

a. Normal sample

Consider the following classic example of statistical inference: suppose a statistician observes the realizations of i.i.d. Gaussian r.v. $X_1, ..., X_n$ with the common law $N(\mu, \sigma^2)$ and would like to estimate μ, while σ^2 is also unknown, but is not of immediate interest (such parameters are called *nuisance* parameters). A natural estimator of μ is the empirical mean

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This type of estimator is known as *point estimator*, as it generates a point in $\mathbb{R}$, viewed as the guess (estimate) of the value of μ. We shall study such estimators in depth in the next chapter. Often one would also like to have some quantitative measure of *confidence* about the obtained estimate. This is achieved by the *confidence* interval estimator $[a(X), b(X)]$, where $a(X)$ and $b(X)$ are functions of the sample. The interval estimator is said to attain confidence of $1-\alpha \in [0,1]$, if the actual value of the parameter μ belongs to it with probability not less than $1-\alpha$:

$$\mathbb{P}\big(\mu \in [a(X), b(X)]\big) \geq 1-\alpha.$$

For example, if you buy a lamp in the supermarket next door, typically the mean time to failure will be specified on the package: this is the point estimator of the mean lifetime of the lamp, produced by the manufacturer in the lab, prior to sending the lamps for sale. If you want to build a radio and need to buy an electronic component (say, resistor), you would typically find its specification in the form $10 \pm 1\%$ [ohm], which means that its nominal value is estimated to be 10 [ohm] (this is again the point estimate) and it is very likely to be somewhere in the interval $[9.9 : 10.1]$, more precisely with confidence level 0.96 (which is a common standard in electronics). The price of the resistor increases with the precision of the confidence interval, which is controlled by the confidence probability or its length (per same confidence).

As we shall see shortly (Proposition 5a1 below), in the Gaussian setting as above the quantity

$$\sqrt{n-1}\frac{\bar{X}_n - \mu}{\hat{\sigma}_n(X)},$$

where $\hat{\sigma}_n(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$, has certain p.d.f., call it f for the moment, which depends only on n - but not on μ or σ^2! This provides a convenient way to construct a confidence interval with any required confidence $1-\alpha \in (0,1)$. Namely, let

$$C(X; z) := \big[\bar{X}_n - z\hat{\sigma}_n, \bar{X}_n + z\hat{\sigma}_n\big],$$

where $z > 0$ is chosen to fit the required confidence level $1 - \alpha$:

$$\mathbb{P}(\mu \in C(X; z)) = \mathbb{P}\Big(\bar{X}_n - z\hat{\sigma}_n \le \mu \le \bar{X}_n + z\hat{\sigma}_n\Big) = \mathbb{P}\Big(\Big|\sqrt{n-1}\frac{\bar{X}_n - \mu}{\hat{\sigma}_n}\Big| \le \sqrt{n-1}z\Big) = \int_{-z\sqrt{n-1}}^{z\sqrt{n-1}} f(x)dx.$$

Now if we require the confidence level of $1-\alpha$, we have to solve the equation $\int_{-z\sqrt{n-1}}^{z\sqrt{n-1}} f(x)dx = 1 - \alpha$ for z. It is clear that this equation has a unique solution, $z^*(\alpha)$, which can be found numerically. Thus we have constructed an interval estimator for μ with confidence $1-\alpha$. Notice also that for a given confidence level $1 - \alpha$, smaller z would emerge for large n, i.e. the $1 - \alpha$ confidence interval, so constructed, shrinks as $n \to \infty$. This is plausible, since the uncertainty about the location of μ should decrease as the number of observations grow.

Of course, the latter procedure is possible and makes sense only if f does not depend on the unknown quantities. Unfortunately, in general this would be rarely the case. One famous and practically very popular exception is the Gaussian i.i.d. setting as above:

PROPOSITION 5a1. *Let $X = (X_1, ..., X_n)$ be i.i.d. r.v. with the common distribution $N(\mu, \sigma^2)$, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}_n^2(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ be the empirical mean and variance. Then for any $n \ge 2$,*

(1) $\bar{X}_n \sim N(\mu, \sigma^2/n)$
(2) *$\bar{X}_n$ and the vector of residuals $R = (X_1 - \bar{X}_n, ..., X_n - \bar{X}_n)$ are independent; in particular $\bar{X}_n$ and $\hat{\sigma}_n^2(X)$ are independent r.v.*
(3) *$n\hat{\sigma}_n^2(X)/\sigma^2 \sim \chi^2(n-1)$, where $\chi^2(k)$ is the χ-square distribution with k degrees of freedom, which has the p.d.f.*

$$f_k^{\chi^2}(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x \in [0, \infty), \tag{5a1}$$

where $\Gamma(x)$ is the Γ-function (generalization of the factorial to non-integer values).

(4)

$$\sqrt{n-1}\frac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}^2(X)}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2}} \sim \mathrm{Stt}(n-1)$$

where $\mathrm{Stt}(k)$ is the Student t-distribution with k degrees of freedom, which has the p.d.f.

$$f_k^S(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}, \quad x \in \mathbb{R}. \tag{5a2}$$

(5) *Let $X' = (X'_1, ..., X'_m)$ be i.i.d. r.v.'s with the common distribution $N(\mu', \sigma'^2)$, independent of X introduced above. Let $\bar{X}'_m = \frac{1}{m}\sum_{i=1}^m X'_i$ and $\hat{\sigma}_m^2(X') = \frac{1}{m}\sum_{i=1}^m (X'_i - \bar{X}'_m)^2$. Then*

$$\frac{\hat{\sigma}_m^2(X')/\sigma'^2}{\hat{\sigma}_n^2(X)/\sigma^2} \sim \mathrm{Fis}(m-1, n-1),$$

where $\mathrm{Fis}(k,\ell)$ *is the Fisher F-distribution, which has the p.d.f.*

$$f^{Fis}(x)=\frac{\Gamma\Big(\frac{k+\ell}{2}\Big)}{\Gamma\Big(\frac{k}{2}\Big)\Gamma\Big(\frac{\ell}{2}\Big)}\Big(\frac{k}{\ell}\Big)^{k/2}\frac{x^{k/2-1}}{\Big(1+k/\ell\ x\Big)^{(k+\ell)/2}},\quad x\in(0,\infty).$$

Remark 5a2. The confidence interval construction for i.i.d. Gaussian sample described above is based on (4). The property (5) is useful in other statistical applications (some to be explored later).

Remark 5a3. The Student distribution has a history:

`http://en.wikipedia.org/wiki/William_Sealy_Gosset`

Proof.

(1) the claim holds, since $\bar X$ is a linear transformation of the Gaussian vector and $\mathbb{E}\bar X_n=\frac{1}{n}\sum_i\mathbb{E}X_i=\mu$ and

$$\mathrm{var}(\bar X_n)=\mathbb{E}(\bar X_n-\mu)^2=\mathbb{E}\Big(\frac{1}{n}\sum_i(X_i-\mu)\Big)^2=$$
$$\frac{1}{n^2}\sum_i\sum_j\mathbb{E}(X_i-\mu)(X_j-\mu)\overset{\dagger}{=}\frac{1}{n^2}n\sigma^2=\sigma^2/n,$$

where † holds by independence of X_i's.

(2) Note that the vector $(\bar X_n,R)$ is a linear map of $(X_1,...,X_n)$ and hence is a Gaussian vector as well (in $\mathbb{R}^{n+1}$). Hence to show that $\bar X_n$ is independent of R, it is enough to check that $\bar X_n$ and $X_i-\bar X_n$ are uncorrelated for all $i=1,...,n$. $\mathbb{E}X_n=\mu$ and $\mathbb{E}(X_i-\bar X_n)=0$ and hence

$$\mathrm{cov}(\bar X_n,X_i-\bar X_n)=\mathbb{E}(\bar X_n-\mu)(X_i-\bar X_n)=\sigma^2\mathbb{E}\bar Z_n(Z_i-\bar Z_n),$$

where we have defined $Z_i:=(X_i-\mu)/\sigma$, which are i.i.d. $N(0,1)$ r.v. Since $\mathbb{E}Z_iZ_j=0$, $\mathbb{E}\bar Z_nZ_i=1/n$ and, as we have already seen, $\mathbb{E}(\bar Z_n)^2=\mathrm{var}(\bar Z_n)=1/n$, which implies

$$\mathrm{cov}(\bar X_n,X_i-\bar X_n)=0,\quad\forall i$$

and in turn that $\bar X_n$ is independent of R. Since $\hat\sigma_n^2(X)$ is in fact a function of R, $\bar X_n$ and $\hat\sigma_n^2(X)$ are independent as well. Note that the Gaussian property played the crucial role in establishing independence!

(3) First note that $\hat\sigma_n^2(X)/\sigma^2=\hat\sigma_n^2(Z)$ (where Z_i's are defined above)

$$\hat\sigma_n^2(X)/\sigma^2=\frac{1}{n}\sum_i(X_i-\bar X_n)^2/\sigma^2=$$
$$\frac{1}{n}\sum_i\Big((X_i-\mu)/\sigma-\frac{1}{n}\sum_j(X_j-\mu)/\sigma\Big)^2=\frac{1}{n}\sum_i\big(Z_i-\bar Z_n\big)^2=\hat\sigma_n^2(Z),$$

hence it is enough to check that $n\hat{\sigma}_n^2(Z) = \sum_i (Z_i - \bar{Z}_n)^2 \sim \chi^2(n-1)$. Note that $\sum_i (Z_i - \bar{Z}_n)^2 = \sum_i Z_i^2 - n(\bar{Z}_n)^2$ and hence

$$\mathbb{E}\exp\Big\{t\sum_i Z_i^2\Big\} = \mathbb{E}\exp\Big\{t\sum_i (Z_i-\bar{Z}_n)^2 + tn(\bar{Z}_n)^2\Big\} \stackrel{\dagger}{=} \mathbb{E}\exp\Big\{t\sum_i (Z_i-\bar{Z}_n)^2\Big\}\mathbb{E}\exp\Big\{tn(\bar{Z}_n)^2\Big\},$$

where † holds since $\bar{Z}_n$ and $Z_i - \bar{Z}_n$, $i = 1, ..., n$ are independent. So the m.g.f. of $n\hat{\sigma}_n^2(Z)$ is given by

$$\mathbb{E}\exp\Big\{t\sum_i (Z_i-\bar{Z}_n)^2\Big\} = \frac{\mathbb{E}\exp\Big\{t\sum_i Z_i^2\Big\}}{\mathbb{E}\exp\Big\{tn(\bar{Z}_n)^2\Big\}} = \frac{\Big(\mathbb{E}\exp\Big\{tZ_1^2\Big\}\Big)^n}{\mathbb{E}\exp\Big\{tn(\bar{Z}_n)^2\Big\}},$$

where the latter equality holds by independence of Z_i's. Now we shall need the following fact: for $\xi \sim N(0,1)$ and $t \in (-1/2, 1/2)$,

$$\mathbb{E}e^{t\xi^2} = \int_{\mathbb{R}} e^{tx^2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2(1-2t)}dx = \frac{1}{\sqrt{1-2t}}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}\sqrt{(1-2t)^{-1}}}e^{-\frac{x^2}{2(1-2t)^{-1}}}dx = \frac{1}{\sqrt{1-2t}}.$$

Recall that $\bar{Z}_n \sim N(0, 1/n)$ and thus $\xi := \sqrt{n}\bar{Z}_n \sim N(0,1)$, so

$$\mathbb{E}\exp\Big\{tn(\bar{Z}_n)^2\Big\} = \mathbb{E}\exp\Big\{t\xi^2\Big\} = \frac{1}{\sqrt{1-2t}}, \quad |t| < 1/2.$$

Similarly

$$\mathbb{E}\exp\Big\{tZ_1^2\Big\} = \frac{1}{\sqrt{1-2t}}, \quad |t| < 1/2.$$

Assembling all parts together we obtain:

$$\mathbb{E}\exp\Big\{t\sum_i (Z_i-\bar{Z}_n)^2\Big\} = \frac{1}{(1-2t)^{(n-1)/2}}, \qquad |t| < 1/2.$$

The latter expression is the m.g.f. of the density given in (5a1) with $k := n-1$ degrees of freedom, as can be verified by a direct (tedious) calculation.

(4) Once again, it is enough to verify the claim for Z_i's:

$$\frac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}_n^2(X)}} = \frac{(\bar{X}_n-\mu)/\sigma}{\sqrt{\hat{\sigma}_n^2(X)/\sigma^2}} = \frac{\bar{Z}_n}{\sqrt{\hat{\sigma}_n^2(Z)}}.$$

Note that

$$\sqrt{n-1}\frac{\bar{Z}_n}{\sqrt{\hat{\sigma}_n^2(Z)}} = \sqrt{n-1}\frac{\bar{Z}_n}{\sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i-\bar{Z}_n)^2}} = \frac{\sqrt{n}\bar{Z}_n}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i-\bar{Z}_n)^2}} =: \frac{V}{\sqrt{U/(n-1)}}.$$

By the preceding calculations, $V \sim N(0,1)$ and $U \sim \chi^2_{n-1}$ and hence the claim holds by the following lemma:

LEMMA 5a4. *Let $U \sim \chi^2(k)$ and $V \sim N(0,1)$ be independent r.v. Then $V/\sqrt{U/k} \sim$ Stt(k).*

PROOF. Define $\tilde{U} := U$ and $\tilde{V} := V/\sqrt{U/k}$, i.e. $\tilde{U} = g_1(V,U)$ and $\tilde{V} = g_2(V,U)$ with $g_1(x,y) = x$ and $g_2(x,y) = y/\sqrt{x/k}$ Obviously, the j.p.d.f. of (U,V) is supported on $(\mathbb{R}_+, \mathbb{R})$, on which g is invertible:

$$g^{-1}(\tilde{x},\tilde{y}) = (\tilde{x}, \tilde{y}\sqrt{\tilde{x}/k}),$$

whose Jacobian is

$$J = \begin{pmatrix} 1 & 0 \\ \frac{1}{2}\tilde{y}/\sqrt{\tilde{x}k} & \sqrt{\tilde{x}/k} \end{pmatrix}, \quad \Longrightarrow \quad \det J = \sqrt{\tilde{x}/k}.$$

Hence the j.p.d.f. of $(\tilde{U},\tilde{V})$ is given by:

$$f_{\tilde{U}\tilde{V}}(\tilde{x},\tilde{y}) = \sqrt{\tilde{x}/k}\frac{(1/2)^{k/2}}{\Gamma(k/2)}\tilde{x}^{k/2-1}e^{-\tilde{x}/2}\frac{1}{\sqrt{2\pi}}e^{-(\tilde{y}\sqrt{\tilde{x}/k})^2/2} =$$
$$\frac{1}{\sqrt{k}\sqrt{2\pi}}\frac{(1/2)^{k/2}}{\Gamma(k/2)}\tilde{x}^{(k-1)/2}e^{-\tilde{x}/2\left(1+\tilde{y}^2/k\right)}$$

for $(\tilde{x},\tilde{y}) \in (\mathbb{R}_+, \mathbb{R})$ and zero otherwise. Now the distribution of $V = \tilde{V}/\sqrt{U/k}$ is obtained as the marginal:

$$f_{\tilde{V}}(\tilde{y}) = \int_{\mathbb{R}_+} f_{\tilde{U}\tilde{V}}(\tilde{x},\tilde{y})d\tilde{x} = \frac{1}{\sqrt{k}\sqrt{2\pi}}\frac{(1/2)^{k/2}}{\Gamma(k/2)}\int_{\mathbb{R}_+}\tilde{x}^{(k-1)/2}e^{-\tilde{x}/2\left(1+\tilde{y}^2/k\right)}d\tilde{x} =$$
$$... = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-(k+1)/2}$$

□

(5) The claim follows from the following fact: if $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$, then

$$\frac{U/m}{V/n} \sim \text{Fis}(m,n).$$

The proof is a straightforward (tedious) calculation, which we shall omit.

□

Exercises

PROBLEM 5.1. Let X_1 and X_2 be i.i.d. standard Gaussian r.v.'s. Find the probability laws of the following r.v.'s:

(1) $(X_1 - X_2)/\sqrt{2}$
(2) $(X_1 + X_2)^2/(X_1 - X_2)^2$

(3) $(X_1 + X_2)/|X_1 - X_2| = (X_1 + X_2)/\sqrt{(X_1 - X_2)^2}$
(4) X_1^2/X_2^2

Hint: make use of the Fisher F-distribution and the Student t-distribution

PROBLEM 5.2. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$ distribution. Find the mean[1] and and the variance of $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$

Hint: make use of the χ^2 distribution

PROBLEM 5.3. Let $X_1, ..., X_n$ be independent r.v., $X_i \sim N(\mu, \sigma_i^2)$. Define:

$$U := \frac{\sum_i X_i/\sigma_i^2}{\sum_j 1/\sigma_j^2}, \quad V := \sum_i (X_i - U)^2/\sigma_i^2.$$

(1) Show that U and V are independent
(2) Show that U is Gaussian and find its mean and variance
(3) Show that $V \sim \chi^2_{n-1}$

Hint: Recall the proof in the special case $\sigma_i^2 = \sigma^2$, $i = 1, ..., n$

PROBLEM 5.4.

(1) Show that if $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then $U + V \sim \chi^2_{n+m}$
    Hint: use the connection between the Gaussian distribution and χ^2
(2) Show that if $X_1, ..., X_n$ are i.i.d. Exp(λ) r.v.'s then $T := 2\lambda \sum_i X_i$ has χ^2_{2n} distribution
    Hint: Prove for $n = 1$ and use the answer in (1)

[1] the expression for the mean should explain the often use of the normalizing factor $1/(n-1)$ instead of $1/n$

Part 2

Statistical inference

CHAPTER 6

Statistical model

a. Basic concepts

Generally and roughly speaking the *statistical inference* deals with drawing conclusions about objects which cannot be observed directly, but are only known to have influence on apparently random observable phenomena. It is convenient to view statistical inference as consisting of three steps:

(Step 1) *Modeling*: postulating the statistical model, i.e. the random mechanism which presumably had produced the observed data. The statistical model is specified up to an unknown *parameter* to be inferred from the data. Once a model is postulated, the statistician defines the scope of inference, i.e. poses the questions of interest. The three canonical problems are:

* *point estimation*: guessing the value of the parameter
* *interval estimation*: constructing an interval, to which the value of the parameter belongs with high confidence
* *hypothesis testing*: deciding whether the value of the parameter belongs to a specific subset of possible values

The main tool in statistical modeling is probability theory.

(Step 2) *Synthesis/Analysis*: deriving a statistical procedure (an algorithm), based on the postulated model, which takes the observed data as its input and generates the relevant conclusions. This is typically done by methodologies, which rely on the optimization theory and numerical methods. Once a procedure is chosen, its quality is to be assessed or/and compared with alternative procedures, if such are available.

(Step 3) *Application*: predicting/or taking decisions on the basis of the derived conclusions

REMARK 6a1. The order is by no means canonical. For example, often the choice of the model is motivated by the question under consideration or the methodological or computational constraints.

REMARK 6a2. The above scheme corresponds to the *inferential* statistics, distinguished from the *descriptive* statistics. The latter is concerned with data exploration by means of various computational tools (e.g. empirical means, histograms, etc.) without presuming or modeling randomness of the data. The quality assessment in descriptive statistics is often subjective and non-rigorous.

Here is an (almost) real life example:

EXAMPLE 6a3. Suppose that you are offered the following game in a casino: a coin is tossed and you either lose your bet, if it comes up tails, or double it, if it comes up heads. You play n games and would like to decide whether to stop or continue playing. The data available to you is the record of outcomes of the n games, i.e. a binary string $x = (x_1, ..., x_n) \in \{0, 1\}^n$.

One reasonable way to model the data is to assume that the tosses are independent in the probabilistic sense and the probability of getting heads in each toss is a number $\theta \in [0, 1]$, which is left unspecified. Our hope now is that on the basis of the data x, we can produce an accurate guess of θ and then base our decision of whether to stop or to continue playing. The suggested model is one of many alternatives: for example, why not to assume that each toss has its own probability of success ? Or furthermore, also discard the assumption of independence, etc. Making the model more flexible (sophisticated, detailed) we potentially allow more accurate predictions regarding the outcomes of the future games. On the other hand, we feel that we might not be able to produce accurate estimates of the parameters in the detailed model on the basis of just n observations. Yet on the third hand, simpler inference algorithms are anticipated to emerge for simpler models. These are just examples of the reasoning for choosing a statistical model. To a large extent, this process is subjective, being based on the experience with the data.

Suppose that we decided to stick to the simple i.i.d. model: we presume that the obtained data is an i.i.d. sample from $X \sim \mathrm{Ber}(\theta)$. What kind of conclusions would we like to derive ? One obvious and natural goal would be to guess the value of θ, i.e. to estimate the true value of the unknown parameter. Perhaps, having an estimated value is not enough and we would like to get a whole range of values to which the true value belongs with a high probability. Or in view of our further intentions, we might pose a more modest question: is it true that $\theta \in [0, 1/2)$? If the answer is yes, then we shall not want to continue playing as in the long run we are going to lose.

Suppose we nevertheless chose to estimate the value of θ. A natural way to proceed would be to calculate the empirical frequency of wins $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and to accept it as our guess $\hat{\theta}_n := \bar{x}_n$. This choice is intuitively plausible, since if we indeed believe that the data has been produced by the presumed model, then by the Law of Large Numbers (to be explored in details below), $\bar{x}_n$ should be close to the true value of θ, at least for large n's. While this simple algorithm is the first thing, which comes to mind, other options are possible: for example, the estimator[1]

$$\tilde{\theta}_n := \sqrt{\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} x_{2i} x_{2i-1}} \tag{6a1}$$

will also yield an estimate close to the true value of θ for large n (think why). Clearly none of the procedures (and in fact no procedure) would yield the exact value of the parameter with probability 1. How do we measure precision then ? Which algorithm is more precise ? Is there an algorithm which gives the best possible precision ? How large n should be to guarantee the desired precision ? These and many more interesting questions is the main subject of mathematical statistics and of this course.

[1] $\lfloor x \rfloor$ is the largest integer smaller or equal to x

Finally, when we generate an estimate of θ and e.g. decide to continue playing, we may use $\hat{\theta}_n$ to calculate various predictions regarding the future games: e.g. how much time it should take on average till we either bankrupt or win certain sum, etc. ■

Now that we have a rough idea of what the statistical inference problem is, let's give it an exact mathematical formulation.

DEFINITION 6a4. *A statistical model (or an experiment) is a collection of probabilities* $\mathcal{P} = (\mathbb{P}_\theta)$*, parameterized by* $\theta \in \Theta$*, where* Θ *is the parameter space.*

If the available data is a vector of real numbers, the probabilities $\mathbb{P}_\theta$ can be defined by means of j.c.d.f's (or j.p.d.f's, j.p.m.f.'s if exist) and the data is thought of as a realization of a random vector X with the particular j.c.d.f., corresponding to the actual value θ_0 of the parameter. This value is unknown to the statistician and is to be inferred on the basis of the observed realization of X and the postulated model.

EXAMPLE 6a3 (continued) In this case $\mathbb{P}_\theta$ is the j.p.m.f. of i.i.d. $X = (X_1, ..., X_n)$ with $X_1 \sim \text{Ber}(\theta)$:

$$\mathbb{P}_\theta(X = x) = p_X(x; \theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \quad x \in \{0,1\}^n.$$

The parameter space is $\Theta = [0, 1]$. ■

EXAMPLE 6a5. A plant produces lamps and its home statistician believes that the lifetime of a lamp has exponential distribution. To estimate the mean lifetime she chooses n lamps sporadically, puts them on test and records the corresponding lifetimes. In this setting, $\mathbb{P}_\theta$ is given by the j.p.d.f.

$$f_X(x; \theta) = \theta^n \prod_{i=1}^n e^{-\theta x_i} I(x_i \geq 0), \quad x \in \mathbb{R}^n.$$

The parameter space $\Theta = \mathbb{R}_+$.

EXAMPLE 6a6. Suppose that we want to receive a shipment of oranges and suspect that part of them rot off. To check the shipment, we draw (sample) oranges from it at random without replacements. Denote by N the number of oranges in the shipment and by $n \leq N$ the size of the sample (the number of oranges drawn). Suppose that a percentage θ of all the oranges rot off. A combinatorial calculation reveals that the number of rotten oranges in the sample has Hyper Geometric p.m.f.

$$p_\theta(k) = \frac{\binom{\theta N}{k}\binom{(1-\theta)N}{n-k}}{\binom{N}{n}}, \quad k \in \Big\{ \max\big(0, n-(1-\theta)N\big), ..., \min(n, \theta N)\Big\}.$$

Since θN should be an integer, the natural choice of the parametric space is $\Theta = \Big\{0, \frac{1}{N}, ..., \frac{N}{N}\Big\}$. ■

In these examples, the parameter space was one-dimensional. Here are examples with higher dimensional parameter spaces.

EXAMPLE 6a7. Suppose we produce a coffee machine, which accepts coins of all values as payment. The machine recognizes different coins by their weight. Hence prior to installing the machine in the campus, we have to tune it. To this end, we shall need to estimate the typical (mean) weight of each type of coin and the standard deviation from this typical weight. For each type, this can be done by e.g. n weighings. A reasonable [2] statistical model would be e.g. to assume that the measured weights X_i's are i.i.d. and $X_1 \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Hence $\mathbb{P}_\theta$ is given by the j.p.d.f.:

$$f_X(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\Big\}, \quad x \in \mathbb{R}^n. \tag{6a2}$$

The unknown parameter is two dimensional $\theta = (\mu, \sigma^2)$ and the natural choice of the parameter space is $\Theta = \mathbb{R} \times \mathbb{R}_+$. ■

EXAMPLE 6a8. Suppose that instead of assuming i.i.d. model in the Example 6a3, there is a reason to believe that the tosses are independent, but not identically distributed (e.g. tosses are done by different people to avoid fraud, but each person actually cheats in his/her own special way). This corresponds to the statistical model

$$p_X(x;\theta) = \prod_{i=1}^{n} \theta_i^{x_i}(1-\theta_i)^{1-x_i}, \quad x \in \{0,1\}^n, \quad \theta = (\theta_1, ..., \theta_n) \in [0,1]^n.$$

Hence the parameter space Θ is n-dimensional. ■

Roughly speaking, models with parameter space of a finite dimension are called *parametric*. Here is a natural example of a nonparametric model:

EXAMPLE 6a5 (continued) Instead of presuming exponential distribution, one can assume that the data is still i.i.d. but with completely unknown p.d.f. In this case the probability laws $\mathbb{P}_\theta$ are parameterized by the space of all functions, which can serve as legitimate p.d.f.'s:

$$\Theta = \left\{ u \mapsto \theta(u) : \theta(u) \geq 0, \int_{\mathbb{R}} \theta(u)du = 1 \right\}.$$

Θ is an infinite dimensional space, in the sense that each element in it - a function of u - is specified by its values at an infinite number of points (all points in $\mathbb{R}$!). ■

REMARK 6a9. In many situations the choice of the parameter space Θ is based on some a priori knowledge of the unknown parameter. For example, if you are pretty sure that the heads probability of the coin does not deviate from 1/2 by more than $\pm\varepsilon$ (known to you at the outset), then $\Theta = (1/2 - \varepsilon, 1/2 + \varepsilon)$ would be the natural choice.

In this course we shall mainly be concerned with parametric models and, moreover, assume that $\Theta \subseteq \mathbb{R}^d$ for some $d < \infty$. Let's start with some statistical slang:

[2] Strictly speaking the Gaussian model is inappropriate, since it allows coins with negative weight with, perhaps very small, but nonzero probability. Nevertheless, we expect that σ^2 is very small, compared to μ and hence the tiny probabilities of absurd events would not alter much the conclusions derived on the basis of Gaussian model. This is an example of practical statistical thinking. As we do not go into modeling step in this course we shall ignore such aspects bravely and thoughtlessly: we shall just assume that the model is already given to us and focus on its analysis, etc.

DEFINITION 6a10. *An arbitrary function of the data (but not of the unknown parameter!) is called* statistic.

In Example 6a3, both $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T(X) := \sqrt{\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2 \rfloor} X_{2i}X_{2i-1}}$ are statistics.

REMARK 6a11. If $\mathbb{P}_\theta$ is a probability on $\mathbb{R}^n$, $X \sim \mathbb{P}_\theta$ and $T(x)$ is a function on $\mathbb{R}^n$, we shall refer both to $T(x)$, $x \in \mathbb{R}^n$ (i.e. to the function $x \mapsto T(x)$) and to $T(X)$ (i.e. to the random variable $T(X)$, obtained by plugging X into T) as statistic. The precise intention should be clear from the context.

In our course, typically we shall be given a statistical model and will be mostly concerned with two questions: how to construct a statistical procedure and how to assess its performance (accuracy). Hence we focus on Step 2 in the above program, assuming that Step 1 is already done and it is known how to carry out Step 3, after we come up with the inference results.

b. The likelihood function

In what follows we shall impose more structure on the statistical models, namely we shall consider models which satisfy one of the following conditions

(R1) $\mathbb{P}_\theta$ is defined by a j.p.d.f. $f(x;\theta)$ for all $\theta \in \Theta$

(R2) $\mathbb{P}_\theta$ is defined by a j.p.m.f. $p(x;\theta)$, such that $\sum_i p(x_i;\theta) = 1$ for a set $\{x_1, x_2, ...\}$ which does not depend on θ.

We shall refer to these assumptions as *regularity*[3]. It will allow us to define [4] the *likelihood* function

DEFINITION 6b1. *Let $\mathbb{P}_\theta$, $\theta \in \Theta$ be a model satisfying either* (R1) *or* (R2). *The function*

$$L(x;\theta) := \begin{cases} f_X(x;\theta), & \text{if } \mathbb{P}_\theta \text{ satisfies (R1)} \\ p_X(x;\theta), & \text{if } \mathbb{P}_\theta \text{ satisfies (R2)} \end{cases},$$

is called likelihood.

All the models considered above satisfy either (R1) or (R2).

EXAMPLE 6a7 (continued) $\mathbb{P}_\theta$ is defined by the Gaussian j.p.d.f. (6a2) for any $\theta \in \Theta$ and hence satisfies (R1). ■

EXAMPLE 6a3 (continued) $\mathbb{P}_\theta$ is defined by the j.p.m.f:

$$p_X(x;\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \quad x \in \{0,1\}^n,$$

and

$$\sum_{x\in\Omega} p_X(x;\theta) = 1,$$

for the set of all 2^n binary strings, $\Omega = \{0,1\}^n$ (which does not depend on θ), i.e. (R2) is satisfied. ■

[3] In statistics, the term "regularity" is not rigid and may have completely different meanings, depending on the context, sometimes, even on subdiscipline, author, book, etc.

[4] without going into more involved probability theory, usually required to define the likelihood

EXAMPLE 6b2. Let X be a random variable distributed uniformly on the set $\{1, ..., \theta\}$, $\theta \in \Theta = \mathbb{N}$ (think of a real life experiment supported by this model):

$$p(k;\theta) = \begin{cases} \frac{1}{\theta} & k = 1, ..., \theta \\ 0 & \text{otherwise} \end{cases}.$$

Since $\sum_{i\in\mathbb{N}} p(i;\theta) = 1$ for all $\theta \in \Theta$, (R2) holds. ■

Here are models which do not fit our framework:

EXAMPLE 6b3. Consider a sample $X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$ and set $Y = \max(0, X)$. Let $\mathbb{P}_\theta$ be the probability law of Y. The model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ doesn't satisfy neither (R1) nor (R2), since the c.d.f. which defines $\mathbb{P}_\theta$, i.e. the c.d.f. of Y, has both continuous and discrete parts. ■

EXAMPLE 6b4. Let X be a binary random variable which takes two values $\{0, \theta\}$, with probabilities $\mathbb{P}_\theta(X = 0) = 1 - \theta$ and $\mathbb{P}_\theta(X = \theta) = \theta$, $\theta \in \Theta = (0, 1)$. Clearly, (R1) does not hold. (R2) does not hold, since the p.m.f. is supported on $\{0, \theta\}$, which depends[5] on θ. Note that if we observe the event $X > 0$, we can determine the value of θ exactly: $\theta = X$. ■

Do not think that the statistical models which do not fit our framework, are of no interest. On the contrary, they are often even more fun, but typically require different mathematical tools.

c. Identifiability of statistical models

Intuitively, we feel that $T(X) = \bar{X}$ in the Example 6a3 is a reasonable guess of θ, since it is likely to be close to the actual value at least when n is large. On the other hand, if $X_i = \theta Z_i$ with unknown $\theta \in \Theta = \mathbb{R}$ and i.i.d. $N(0, 1)$ r.v.'s Z_i, any guess of the sign of θ, based on the observation of $X_1, ..., X_n$ will be as bad as deciding it by tossing an independent coin, discarding all the data. On the third hand, if we were not interested in the signed value of θ, but only in its absolute value, e.g. $\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$ would be a decent guess (again, intuitively).

This simple consideration leads us to the following notion

DEFINITION 6c1. *A model $\mathbb{P}_\theta, \theta \in \Theta$ is identifiable if*

$$\theta, \theta' \in \Theta,\ \theta \neq \theta' \quad \Longrightarrow \quad \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}.$$

This definition requires an elaboration: what do we mean by $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$, i.e. one probability is different from another probability? This means that there is an event A, such that $\mathbb{P}_\theta(A) \neq \mathbb{P}_{\theta'}(A)$. When $\mathbb{P}_\theta$ has a discrete support, this is equivalent to the corresponding p.m.f.'s being different (think why?):

$$p(x;\theta) \neq p(x;\theta'), \quad \text{for some } x.$$

For $\mathbb{P}_\theta, \theta \in \Theta$ define by a p.d.f. identifiability amounts to existence of an open ball B, such that[6]

$$f(x;\theta) \neq f(x;\theta'), \quad \forall x \in B.$$

Here is a handy shortcut:

[5] more precisely is not a subset of any countable set for all $\theta \in (0, 1)$

[6] in the continuous case, requiring e.g. $f(x;\theta) \neq f(x;\theta')$ at an isolated point $x \in \mathbb{R}$ whenever $\theta \neq \theta'$ is clearly not enough.

PROPOSITION 6c2. *The model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is identifiable if there exists a statistic $T(X)$, $X\sim\mathbb{P}_\theta$, whose expectation is a one-to-one function of $\theta\in\Theta$, i.e. such that*

$$\theta\neq\theta' \quad\Longrightarrow\quad \mathbb{E}_\theta T(X)\neq\mathbb{E}_{\theta'}T(X).$$

PROOF. Suppose that the model is not identifiable, i.e. there is a pair of parameters $\theta\neq\theta'$, for which $\mathbb{P}_\theta=\mathbb{P}_{\theta'}$. Then, in particular, $\mathbb{E}_\theta T(X)=\mathbb{E}_{\theta'}T(X)$, which is a contradiction and hence the model is identifiable. □

REMARK 6c3. Estimation of parameters for non identifiable models is meaningless [7]. If the constructed model turns to be nonidentifiable, a different parametrization (model) is to be found to turn it into an identifiable one.

EXAMPLE 6a3 (continued) Recall that $\mathbb{P}_\theta$ is the probability law of i.i.d. r.v. $X_1,...,X_n$ with $X_1\sim\mathrm{Ber}(\theta)$, $\theta\in[0,1]$. Since

$$\mathbb{E}_\theta X_1=\theta$$

is a one-to-one function of $\theta\in\Theta$, the model is identifiable.

An alternative way to get to the same conclusion is to consider the p.m.f. $p(x;\theta)$ at e.g. x with $x_i=1$, $i=1,...,n$:

$$p\big((1...1),\theta\big)=\theta^n\neq\theta'^n=p\big((1...1),\theta'\big),\quad\forall\theta\neq\theta'\in\Theta.$$

■

EXAMPLE 6a7 (continued) If $X_1,...,X_n$ are i.i.d. $N(\mu,\sigma^2)$ and $\theta=(\theta_1,\theta_2)=(\mu,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+$, the model is identifiable:

$$\mathbb{E}_\theta X_1=\theta_1=\mu,\quad \mathbb{E}_\theta X_1^2=\theta_2+\theta_1^2=\sigma^2+\mu^2.$$

The function $g(\theta)=(\theta_1,\theta_2+\theta_1^2)$ is one-to-one on Θ: indeed, suppose that for $\theta\neq\theta'$, $g(\theta)=g(\theta')$ which means $\theta_1=\theta_1'$ and $\theta_2+\theta_1^2=\theta_2'+\theta_1'^2$. The latter implies $\theta_1=\theta_1'$ and $\theta_2=\theta_2'$, which is a contradiction. Hence the model is identifiable.

Now let's check what happens if we would have chosen a different parametrization, namely $\theta=(\theta_1,\theta_2)=(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}$. Let $\theta=(0,1)$ and $\theta'=(0,-1)$. Since $\theta_2'=\sigma$ appears in the Gaussian density only with an absolute value, it follows that $\mathbb{P}_\theta=\mathbb{P}_{\theta'}$ for the specific choice of $\theta\neq\theta'$ and hence the model with such parametrization is not identifiable, confirming our premonitions above. ■

Here is a less obvious example:

EXAMPLE 6c4. Let $X=(X_1,...,X_n)$ be an i.i.d. sample from $N(\mu,\sigma^2)$ and suppose that we observe $Y=(Y_1,...,Y_n)$, where $Y_i=X_i^2$. Let $\mathbb{P}_\theta$ be the probability law of Y, where $\theta=(\mu,\sigma)\in\mathbb{R}_+\times\mathbb{R}_+$ (note that μ is known to be nonnegative). Is this model identifiable[8] ? Note that

$$\mathbb{E}_\theta Y_1=\mathbb{E}_\theta X_1^2=\mathrm{var}_\theta(X_1)+(\mathbb{E}_\theta X_1)^2=\sigma^2+\mu^2$$

---

[7] Running a bit ahead, suppose that we have a statistic $T(X)$, which we use as a point estimator of θ. A good estimator of θ should be close to the values of θ for *all* (!) $\theta\in\Theta$. For non-identifiable models this is impossible. Suppose that $\theta\neq\theta'$ (think of $\theta=\theta'+d$, where d is a large number) and $\mathbb{P}_\theta=\mathbb{P}_{\theta'}$, i.e. the model is not identifiable. If $T(X)$ is a good estimator of θ, then its probability distribution should be highly concentrated around θ, when $X\sim\mathbb{P}_\theta$. But $\mathbb{P}_\theta=\mathbb{P}_{\theta'}$ and hence it is also highly concentrated around θ, when $X\sim\mathbb{P}_{\theta'}$. But since the distance between θ and θ' is large, the latter means that the distribution of $T(X)$ is poorly concentrated around θ', when $X\sim\mathbb{P}_{\theta'}$ - i.e. it is a bad estimator of θ' !

[8] we have seen already, in Example 6a7, that it is identifiable with a different parameter space. Think why this implies identifiability for the new parameter space under consideration.

and

$$\mathbb{E}_\theta Y_1^2 = \mathbb{E}_\theta X_1^4 = \mathbb{E}_\theta (X_1 - \mu + \mu)^4 = \\ \mathbb{E}_\theta (X_1-\mu)^4 + 4\mathbb{E}_\theta (X_1-\mu)^3\mu + 6\mathbb{E}_\theta (X_1-\mu)^2\mu^2 + 4\mathbb{E}_\theta (X_1-\mu)\mu^3 + \mu^4 = \\ 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4 = 3(\sigma^2+\mu^2)^2 - 2\mu^4.$$

The function $g(\theta) := (\sigma^2+\mu^2, 3(\sigma^2+\mu^2)^2 - 2\mu^4)$ is invertible on $(\mu,\sigma) \in \mathbb{R}_+ \times \mathbb{R}_+$ (check!) and hence the model is identifiable.

Now suppose we observe only the signs of X_i's, i.e. $\xi = (\xi_1, ..., \xi_n)$ with

$$\xi_i = \operatorname{sign}(X_i) := \begin{cases} 1 & X_i \geq 0 \\ -1 & X_i < 0 \end{cases}.$$

Let $\mathbb{P}_\theta$ be the law of ξ with the same parametrization as before. Is this model identifiable ...? In this case, $\mathbb{P}_\theta$ is given by its j.p.m.f., namely for $u \in \{1,-1\}^n$ and θ as above

$$p_\xi(u;\theta) = \prod_{i=1}^n \Big\{ I(u_i = 1)\mathbb{P}_\theta(\xi_i = 1) + I(u_i = -1)\mathbb{P}_\theta(\xi_i = -1) \Big\} = \\ \prod_{i=1}^n \Big\{ I(u_i = 1)\mathbb{P}_\theta(X_i \geq 0) + I(u_i = -1)\mathbb{P}_\theta(X_i < 0) \Big\}$$

Further,

$$\mathbb{P}_\theta(X_1 < 0) = \mathbb{P}_\theta\Big(\frac{X_1 - \mu}{\sigma} < -\mu/\sigma\Big) = \Phi(-\mu/\sigma),$$

and hence $p_\xi(u;\theta)$ depends on $\theta = (\mu,\sigma)$ only through the ratio μ/σ. Clearly this model is not identifiable: for example, $\theta = (1,1)$ and $\theta' = (2,2)$ yield the same distribution of the data: $p_\xi(u;\theta) = p_\xi(u;\theta')$ for all $u \in \{1,-1\}^n$. This means that the observation of ξ cannot be used to construct a reasonable estimate of (μ,σ).

■

d. Sufficient statistic

Consider the following simple model: we observe the realizations of $X_1 = \theta + Z_1$ and $X_2 = Z_2$, where Z_1, Z_2 are i.i.d. $N(0,1)$ r.v.'s and would like to infer $\theta \in \mathbb{R}$. It is intuitively clear that X_2 is irrelevant as far as inference of θ is concerned, since it is just noise, not affected in any way by the parameter value. In particular, the statistic $T(X) = X_1$ is "sufficient" for the purpose of e.g. guessing the value of θ. On the other hand, if Z_1 and Z_2 were dependent, then such "sufficiency" would be less apparent (and in fact false as we shall be able to see shortly).

Definition 6d1. *A statistic T is sufficient for the model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ if the conditional distribution of $X \sim \mathbb{P}_\theta$, given $T(X)$, does not depend on θ.*

The meaning of the latter definition is made particularly transparent through the following two-stage procedure. Suppose that we sample from $\mathbb{P}_\theta$ once, call this sample X, and calculate $T(X)$. Discard X and keep only $T(X)$. Since T is sufficient, by definition the conditional distribution of X, given $T(X)$, does not depend on the unknown value of θ. Hence we are able to sample from the conditional distribution of X, given $T(X)$, without knowing the value of θ!

Let X' be a sample from this conditional distribution[9]. Typically, the obtained realizations of X and X' will not be the same and hence we would not be able to restore the original discarded realization of X. However X and X' will have the same probability distribution $\mathbb{P}_\theta$ and hence bear the very same statistical "information" about θ. Indeed, by the very definition of X'

$$\mathbb{P}_\theta\big(X'=x|T(X)\big)=\mathbb{P}_\theta\big(X=x|T(X)\big)$$

where we assumed for definiteness that all the random vectors involved are discrete. Since $T(X)$ is sufficient, the latter doesn't in fact depend on θ and

$$\mathbb{P}_\theta\big(X'=x\big)=\mathbb{E}_\theta\mathbb{P}_\theta\big(X'=x|T(X)\big)=\mathbb{E}_\theta\mathbb{P}_\theta\big(X=x|T(X)\big)=$$
$$\mathbb{E}_\theta p_{X|T}\big(x;T(X)\big)=\mathbb{E}_\theta\mathbb{P}_\theta\big(X=x|T(X)\big)=\mathbb{P}_\theta(X=x),\quad\forall x.$$

To recap, no matter what kind of inference we are going to carry out on the basis of the sample X, the value of a sufficient statistic $T(X)$ is all we need to keep, to be able to sample from the original distribution without knowing the parameter and hence to attain the very same accuracy in the statistical analysis, we would be able to attain should we have kept the original sample X! In this sense, the sufficient statistic is a "summary" statistic.

EXAMPLE 6a3(continued) Let's show that $T(X)=\sum_{i=1}^n X_i$ is a sufficient statistic for i.i.d $X=(X_1,...,X_n)$, $X_1\sim\text{Ber}(\theta)$, $\theta\in\Theta=(0,1)$ r.v.'s The conditional j.p.m.f. of X given $T(X)$ is given by the Bayes formula: let $x\in\{0,1\}^n$ and $t\in\{0,...,n\}$; if $t\neq\sum_{i=1}^n x_i$, then $\mathbb{P}(X=x,T(X)=t)=0$ and hence $\mathbb{P}(X=x|T(X)=t)=0$. Otherwise, i.e. for $t=\sum_i x_i$,

$$\mathbb{P}_\theta\big(X=x|T(X)=t\big)=\frac{\mathbb{P}_\theta\big(X=x,T(X)=t\big)}{\mathbb{P}_\theta\big(T(X)=t\big)}=\frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}=\frac{1}{\binom{n}{t}},$$

where we used the fact that $T(X)\sim\text{Bin}(n,\theta)$. Hence X conditionally on $T(X)$ is distributed uniformly on the set of binary vectors

$$\Big\{x\in\{0,1\}^n:\sum_{i=1}^n x_i=t\Big\},$$

and thus the conditional distribution X given $T(X)$ doesn't depend on θ. Hence $T(X)$ is indeed a sufficient statistic.

Let's see how the aforementioned hypothetic experiment works out in this particular case. Suppose we tossed the coin $n=5$ times and obtained $\{X=(11001)\}$, for which $T(X)=3$. Now sample from the uniform distribution on all the strings which have precisely 3 ones (you can actually easily list all of them). Notice that this is feasible without the knowledge of θ. The obtained sample, say $\{X'=(00111)\}$, is clearly different from X, but is as relevant to any statistical question regarding θ as the original data X, since X' and X are samples from the very same distribution.

The statistic $S(X)=\sum_{i=1}^{n-1}X_i$ is intuitively not sufficient, since it ignores X_n, which might be useful for inference of θ. Let's verify the intuition by a contradiction. Suppose it is sufficient. Then the conditional distribution of X given $S(X)$ doesn't depend on θ. On the other hand, X_n

[9]you might need to enlarge the probability space to do this, but we keep the same notations for brevity

and $S(X)$ are independent and thus $\mathbb{E}_\theta(X_n|S(X)) = \mathbb{E}X_n = \theta$, which contradicts the assumption of sufficiency. Hence $S(X)$ is not sufficient as expected. ■

EXAMPLE 6d2. Suppose $X = (X_1, ..., X_n)$ are i.i.d. $N(\theta, 1)$ r.v., with $\theta \in \mathbb{R}$. Let's check that $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient. Note that X and $\bar{X}$ are jointly Gaussian (since $\bar{X}$ is a linear transformation of a Gaussian vector X). Hence the conditional distribution of X given $\bar{X}$ is Gaussian as well and it is enough to check that $\mathbb{E}_\theta(X_i|\bar{X})$ and $\mathrm{cov}_\theta\big(X_i, X_j|\bar{X}\big)$ do not depend on θ for any i and j (think why!). Using the explicit formulae for conditional expectations in the Gaussian case, we find

$$\mathbb{E}_\theta(X_i|\bar{X}) = \mathbb{E}_\theta X_i + \frac{\mathrm{cov}_\theta(X_i, \bar{X})}{\mathrm{var}_\theta(\bar{X})}(\bar{X} - \mathbb{E}_\theta\bar{X}) = \theta + \frac{1/n}{1/n}(\bar{X} - \theta) = \bar{X}$$

and

$$\mathrm{cov}_\theta(X_i, X_i|\bar{X}) = \mathrm{var}_\theta(X_i|\bar{X}) \overset{\dagger}{=} \mathrm{var}_\theta(X_i) - \frac{\mathrm{cov}_\theta^2(X_i, \bar{X})}{\mathrm{var}_\theta(\bar{X})} = 1 - \frac{1/n^2}{1/n} = 1 - 1/n,$$

where the equality † is the formula from the Normal Correlation Theorem 3c1. To calculate $\mathrm{cov}_\theta(X_i, X_j|\bar{X})$ when $i \neq j$,

$$\mathrm{cov}_\theta(X_i, X_j|\bar{X}) = \mathbb{E}_\theta\big(X_i - \bar{X}\big)\big(X_j - \bar{X}\big)\big|\bar{X}\big) = \mathbb{E}_\theta(X_iX_j|\bar{X}) - \bar{X}^2,$$

we shall use the following trick. Notice that

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}_\theta(X_iX_j|\bar{X}) = \mathbb{E}_\theta(X_i\bar{X}|\bar{X}) = \bar{X}\mathbb{E}_\theta(X_i|\bar{X}) = \bar{X}^2, \tag{6d1}$$

and on the other hand

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}_\theta(X_iX_j|\bar{X}) = \frac{1}{n}\mathbb{E}_\theta(X_i^2|\bar{X}) + \frac{1}{n}\sum_{j\neq i} \mathbb{E}_\theta(X_iX_j|\bar{X}).$$

Moreover, by the i.i.d. property (similarly to Example 3d5),

$$\mathbb{E}_\theta(X_iX_j|\bar{X}) = \mathbb{E}_\theta(X_iX_\ell|\bar{X}), \quad \forall j \neq \ell,$$

and, also $\mathbb{E}_\theta(X_i^2|\bar{X}) = \mathrm{var}_\theta(X_i|\bar{X}) + \bar{X}^2 = 1 - 1/n + \bar{X}^2$, and hence

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}_\theta(X_iX_j|\bar{X}) = \frac{1}{n}\Big(1 - \frac{1}{n}\Big) + \frac{1}{n}\bar{X}^2 + \frac{n-1}{n}\mathbb{E}_\theta(X_iX_j|\bar{X}). \tag{6d2}$$

Combining (6d1) and (6d2) we get

$$\frac{1}{n}\Big(1 - \frac{1}{n}\Big) + \frac{1}{n}\bar{X}^2 + \frac{n-1}{n}\mathbb{E}_\theta(X_iX_j|\bar{X}) = \bar{X}^2$$

or

$$\mathbb{E}_\theta(X_iX_j|\bar{X}) = \bar{X}^2 - \frac{1}{n}$$

and finally

$$\mathrm{cov}_\theta(X_i, X_j|\bar{X}) = \mathbb{E}_\theta(X_iX_j|\bar{X}) - \bar{X}^2 = \bar{X}^2 - \frac{1}{n} - \bar{X}^2 = -\frac{1}{n}.$$

Thus the conditional probability distribution of X given $\bar{X}$ is the same for all θ, verifying the sufficiency of $\bar{X}$.

Let's simulate the procedure, demonstrating sufficiency in action: suppose we obtained an i.i.d. $X_1, ..., X_n$ sample from $N(\theta, 1)$ and calculated $\bar{X}_n$. Discard X and keep only $\bar{X}_n$. Sample from the Gaussian distribution with the mean vector

$$\mu(X) := \begin{pmatrix} \bar{X}_n \\ \vdots \\ \bar{X}_n \end{pmatrix}$$

and covariance matrix with the entries

$$S_{ij} = \operatorname{cov}(X_i, X_j) = \begin{cases} 1 - \frac{1}{n} & i = j \\ -\frac{1}{n} & i \neq j \end{cases}.$$

The obtained new sample, call it X', is a realization of a random vector with n i.i.d. $N(\theta, 1)$ components. Note that in this case the event $\{X = X'\}$ has zero probability, i.e. we shall never get the very same realization back. However, as far as inference of θ is concerned, it is enough to keep just one real number $\bar{X}$ instead of n real numbers $X_1, ..., X_n$!

■

As you can see, verifying that a given statistic is sufficient for a given model may be quite involved computationally. Moreover, it is not apparent neither from the definition nor from the examples, how a nontrivial sufficient statistic can be found for a given model. The main tool which makes both of these tasks straightforward is

THEOREM 6d3 (Fisher-Neyman factorization theorem). *Let $\mathbb{P}_\theta, \theta \in \Theta$ be a model with likelihood $L(x;\theta)$ and let $X \sim \mathbb{P}_\theta$. Statistic $T(X)$ is sufficient if and only if there exist functions $g(u,t)$ and $h(x)$ (with appropriate domains), so that*

$$L(x;\theta) = g(\theta, T(x))h(x) \quad \forall x, \theta. \tag{6d3}$$

PROOF. We shall give the proof for the discrete case, leaving out the more technically involved (but similar in spirit) continuous case. When X is discrete, the likelihood equals the p.m.f., call it $p_X(x;\theta)$, $x \in \{x_1, x_2, ...\}$. Suppose (6d3) holds, i.e.

$$p_X(x;\theta) = g(\theta, T(x))h(x)$$

for some functions g and h. We shall show that T is sufficient, by checking that the conditional law of X given $T(X)$ doesn't depend on θ. For any x and $t \neq T(x)$,

$$p_{X|T(X)}(x;t,\theta) = \mathbb{P}_\theta(X = x | T(X) = t) = 0,$$

which certainly doesn't depend on θ. Further, for $t = T(x)$, by the Bayes formula

$$p_{X|T(X)}(x;t,\theta) = \mathbb{P}_\theta(X = x | T(X) = t) = \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(x) = t)} = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T(x) = t)}$$
$$\frac{g(\theta,t)h(x)}{\sum_{x'} g(\theta,t)h(x')} = \frac{g(\theta,t)h(x)}{g(\theta,t)\sum_{x'} h(x')} = \frac{h(x)}{\sum_{x'} h(x')},$$

which does not depend on θ as well.

Conversely, suppose that $T(X)$ is a sufficient statistic. To prove the claim we shall exhibit functions g and h such that (6d3) holds. To this end, note that since $T(X)$ is a function of X,

$$p_{X,T(X)}(x,t;\theta) = \mathbb{P}_\theta(X=x, T(X)=t) = \mathbb{P}_\theta(X=x, T(x)=t) = \begin{cases} p_X(x;\theta) & t = T(x) \\ 0 & t \neq T(x) \end{cases},$$

and hence $p_{X|T(X)}(x;t) = 0$ for $t \neq T(x)$. Then

$$p_X(x;\theta) = \sum_t p_{X|T(X)}(x;t) p_{T(X)}(t;\theta) = p_{X|T(X)}(x;T(x)) p_{T(X)}(T(x);\theta)$$

and (6d3) holds with

$$g(\theta, T(x)) := p_{T(X)}(T(x);\theta) \quad \text{and} \quad h(x) := p_{X|T(X)}(x;T(x)),$$

where $h(x)$ does not depend on θ by sufficiency of T. □

Let's apply the F-N theorem to the preceding examples:

Example 6a3 (continued) The likelihood function is

$$L(x;\theta) = p_X(x;\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^{T(x)}(1-\theta)^{n-T(x)}, \quad x \in \{0,1\}^n, \theta \in \Theta,$$

where $T(x) = \sum_{i=1}^n x_i$. Hence (6d3) holds with $h(x) = 1$ and $g(\theta, t) = \theta^t(1-\theta)^{n-t}$. ■

Example 6d2 (continued) The likelihood is the j.p.d.f. in this case:

$$\begin{aligned} L(x;\theta) = f_X(x;\theta) &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\Big) = \\ &\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n x_i^2 + \theta\sum_{i=1}^n x_i - \frac{n}{2}\theta^2\Big) = \\ &\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n x_i^2\Big) \exp\Big(\theta n\bar{x} - \frac{n}{2}\theta^2\Big). \end{aligned} \tag{6d4}$$

By F-N theorem, $T(X) = \bar{X}$ is sufficient, since (6d3) is satisfied with

$$h(x) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n x_i^2\Big)$$

and $g(\theta, t) = \exp\Big(\theta n t - \frac{n}{2}\theta^2\Big)$. Compare this to the calculations, required to check that $\bar{X}$ is sufficient directly from the definition! ■

Let's demonstrate the power of F-N theorem by slightly modifying the latter example:

Example 6a7 (continued)

Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$ r.v.'s. If σ^2 is known, then we are in the same situation as in the previous example. If however both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$

$$\begin{aligned} L(x;\theta) =& \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\Big) = \\ & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2}\frac{1}{\sigma^2} n\mu)^2\Big) =: \\ & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2}\frac{1}{\sigma^2} n\overline{x^2} + \frac{\mu}{\sigma^2} n\bar{x} - \frac{1}{2}\frac{1}{\sigma^2} n\mu^2\Big). \end{aligned} \tag{6d5}$$

The statistic $\bar{X}$ is no longer sufficient, since it is impossible to factorize the likelihood to match (6d3), so that h will be only a function of x: the first term in the latter exponent depends both on x and on σ^2, which is now the unknown parameter. A sufficient statistic here is $T(X) = (\bar{X}, \overline{X^2})$. Note that it is a two dimensional vector[10].

Suppose now that $\mu = \theta$ and $\sigma = \sqrt{\theta}$, where $\theta \in \mathbb{R}_+$ is the unknown parameter. Then by F-N theorem and (6d5), the statistic $T(X) = \overline{X^2}$ is sufficient. ■

Example 6d4. Let $X_1 \sim U([0,\theta])$, $\theta \in \Theta = (0,\infty)$ and $X_1, ..., X_n$ be i.i.d. r.v.'s. The likelihood is the j.p.d.f. :

$$L(x;\theta) = \prod_{i=1}^n \frac{1}{\theta} I(x_i \in [0,\theta]) = \theta^{-n} I(\max_i x_i \le \theta), \quad \forall x \in \mathbb{R}_+^n.$$

Hence by the F-N theorem $\max_i X_i$ is a sufficient statistic. ■

Note that neither the definition nor F-N Theorem do not say anything on the uniqueness of the sufficient statistic: the factorization of the likelihood can be done in many different ways to yield different sufficient statistics. In fact a typical statistical model has many quite different sufficient statistics. In particular, the original data, i.e. X sampled from $\mathbb{P}_\theta$, is trivially a sufficient statistic: indeed, $\mathbb{P}_\theta(X \le u | X) = I(X \le u)$ for any $u \in \mathbb{R}$ and $I(X \le u)$ doesn't depend on θ. Of course, this is also very intuitive: after all the original data is all we have!

This suggests the following relation between statistics:

Definition 6d5. *$T(X)$ is coarser[11] than $T'(X)$ if $T(X) = f(T'(X))$ for some function f.*

"Coarser" in this definition means "revealing less details on the original data": one can calculate $T(X)$ from $T'(X)$ but won't be able to calculate $T'(X)$ from $T(X)$ (unless f is one-to-one). Hence some "information" will be possibly lost. In the Example 6a7, the statistic $T(X) = (\bar{X}, \overline{X^2})$ is coarser than X itself: clearly, one can calculate $T(X)$ from X, but not vise versa (if $n \ge 2$). The trivial statistic $T'(X) \equiv 17$ is coarser than both T and X: it is so coarse that it is useless for any inference (and of course not sufficient).

Definition 6d6. *Two statistics T and T' are equivalent, if there is a one-to-one function f such that $T(X) = f(T'(X))$.*

[10] more precisely, $(\bar{x}, \overline{x^2})$ is two dimensional since, $\bar{x}$ does not determine $\overline{x^2}$ and vise versa (check!). Compare e.g. with $(\bar{x}, \bar{x})$ which takes values in $\mathbb{R}^2$, but in fact takes values only in a one dimensional manifold, namely on the diagonal $\{(x,y) \in \mathbb{R}^2 : x = y\}$

[11] a bit more precisely, $\mathbb{P}_\theta\big(T(X) = f(T'(X))\big) = 1$ for all θ is enough.

The statistics are equivalent if they reveal the same details about the data.

REMARK 6d7. If S is coarser than (or equivalent to) T and S is sufficient, then T is sufficient as well (convince yourself, using the F-N factorization).

EXAMPLE 6a7 (continued) The statistics $T(X) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right) =: (T_1(X), T_2(X))$ and $S(X) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right) =: (S_1(X), S_2(X))$ are equivalent. Indeed S can be recovered from T:

$$S_1(X) = T_1(X)$$

and

$$S_2(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = T_2(X) - T_1^2(X).$$

and vise versa. Clearly T and X are not equivalent: T is coarser than X as previously mentioned. ■

This discussion leads us to the question: is there coarsest ("minimal") sufficient statistic ? And if there is, how it can be found ?

DEFINITION 6d8. *The sufficient statistic T is minimal if it is coarser than any other sufficient statistic.*

It can be shown that the minimal sufficient statistic exists (at least for our type of models). The proof is beyond the scope of our course. Finding minimal statistic does not appear at the outset an easy problem: in principle, if one tries to find it via the definition, she would have to perform a search among all sufficient statistics (or at least all nonequivalent statistics), which is practically impossible. Remarkably, checking that a particular sufficient statistic is minimal is easier, as suggested by the following lemma. Note that in practice candidates for the minimal sufficient statistic are offered by the F-N factorization theorem.

LEMMA 6d9. *A sufficient statistic S is minimal sufficient if*

$$\frac{L(x;\theta)}{L(y;\theta)} \text{ doesn't depend on } \theta \quad \Longrightarrow \quad S(x) = S(y). \tag{6d6}$$

REMARK 6d10. Note that by F-N factorization theorem, $S(x) = S(y)$ implies that $\dfrac{L(x;\theta)}{L(y;\theta)}$ does not depend on θ.

PROOF. Suppose that (6d6) holds and let $T(X)$ be a sufficient statistic. Then by the F-N theorem

$$L(x;\theta) = g(\theta, T(x))h(x)$$

for some g and h and

$$\frac{L(x;\theta)}{L(y;\theta)} = \frac{g(\theta, T(x))h(x)}{g(\theta, T(y))h(y)}.$$

Let x and y be such that $T(x) = T(y)$, then the latter equals $h(x)/h(y)$, which doesn't depend on θ. But then by (6d6) it follows that $S(x) = S(y)$. Hence $T(x) = T(y)$ implies $S(x) = S(y)$, which means[12]that $S(x) = f(T(x))$ for some f. Since T was an arbitrary sufficient statistic and S is sufficient, S is by definition minimal sufficient. □

EXAMPLE 6a7 (continued) Suppose that σ^2 is known (say $\sigma^2 = 1$) and we would like to infer $\mu \in \mathbb{R}$. Applying the F-N factorization theorem to (6d4), we see that $\bar{X}$ is a sufficient statistic. Is it minimal ?

$$\frac{L(x;\theta)}{L(y;\theta)} = \frac{\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n x_i^2\Big) \exp\Big(\theta n\bar{x} - \frac{n}{2}\theta^2\Big)}{\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n y_i^2\Big) \exp\Big(\theta n\bar{y} - \frac{n}{2}\theta^2\Big)} =$$

$$\exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i^2 - y_i^2)\Big) \exp\Big(\theta n(\bar{x} - \bar{y})\Big).$$

The latter is independent of θ if and only if $\bar{x} = \bar{y}$ (check!). Hence $\bar{x}$ is minimal sufficient by the preceding Lemma. Remarkably, keeping just $\bar{X}$ and discarding all the observations is enough for all the purposes of inference of θ and any coarser statistic won't be sufficient!

The whole data X is, of course, sufficient but is not minimal, as it is finer than $\bar{X}$ and we should expect that the conditions of the Lemma cannot be satisfied. Indeed, if $\frac{L(x;\theta)}{L(y;\theta)}$ doesn't depend on θ, it is still possible to have $x \neq y$.

■

EXAMPLE 6d4 (continued) In this case,

$$\frac{L(x;\theta)}{L(y;\theta)} = \frac{I(\max_i x_i \le \theta)}{I(\max_i y_i \le \theta)}.$$

Using the conventions[13] $0/0 = 1$, $1/0 = \infty$ we see that if $S(x) := \max_i x_i < \max_i y_i =: S(y)$, then

$$\frac{L(x;\theta)}{L(y;\theta)} = \begin{cases} 1, & \theta \le S(x) \\ 0, & \theta \in (S(x), S(y)] \\ 1, & \theta > S(y) \end{cases},$$

which is a nontrivial (nonconstant) function of θ. Similarly, it is a nonconstant function of θ in the case $S(x) > S(y)$:

$$\frac{L(x;\theta)}{L(y;\theta)} = \begin{cases} 1, & \theta \le S(y) \\ \infty, & \theta \in (S(y), S(x)] \\ 1, & \theta > S(x) \end{cases},$$

Hence $S(x) \neq S(y)$ implies that $\frac{L(x;\theta)}{L(y;\theta)}$ is a nonconstant function of θ, which confirms (6d6) by negation. Consequently, $S(X) = \max_i X_i$ is the minimal sufficient statistic. ■

---

[12]**Lemma:** *Let ϕ and ψ be real functions on $\mathbb{R}^d$. There is a real function f, such that $\phi(x) = f\big(\psi(x)\big)$ if and only if for all $x, y \in \mathbb{R}^d$, $\psi(x) = \psi(y)$ implies $\phi(x) = \phi(y)$.*

Proof: Suppose $\psi(x) = \psi(y)$ implies $\phi(x) = \phi(y)$. Then $f(z) := \phi\big(\psi^{-1}(z)\big)$ is uniquely defined (why?) and $f(\psi(x)) = \phi(x)$ by construction. The other direction is obvious.

[13]if you use other conventions, you will still deduce that $\max_i x_i$ is minimal sufficient.

Exercises

PROBLEM 6.1 (Problem 2.1.1, page 80 [**1**]). Give a formal statement of the following models identifying the probability laws of the data and the parameter space. State whether the model in question is parametric or nonparametric.

(1) A geologist measures the diameters of a large number n of pebbles in an old stream bed. Theoretical considerations lead him to believe that the logarithm of pebble diameter is normally distributed with mean μ and variance σ^2. He wishes to use his observations to obtain some information about μ and σ^2, but has in advance no knowledge of the magnitudes of the two parameters.
(2) A measuring instrument is being used to obtain n independent determinations of a physical constant μ. Suppose that the measuring instrument is known to be biased to the positive side by 0.1 units. Assume that the errors are otherwise identically distributed normal random variables with known variance.
(3) In part (2) suppose that the amount of bias is positive but unknown. Can you perceive any difficulties in making statements about μ for this model?
(4) The number of eggs laid by an insect follows a Poisson distribution with unknown mean λ. Once laid, each egg has an unknown chance p of hatching and the hatching of one egg is independent of the hatching of the others. An entomologist studies a set of n such insects observing both the number of eggs laid and the number of eggs hatching for each nest.

PROBLEM 6.2 (Problem 2.1.2, page 80 [**1**]). Are the following parametrizations identifiable? (Prove or disprove.)

(1) The parametrization of Problem 6.1 (3).
(2) The parametrization of Problem 6.1 (4).
(3) The parametrization of Problem 6.1 (4) if the entomologist observes only the number of eggs hatching but not the number of eggs laid in each case.

PROBLEM 6.3 (Problem 2.1.3, page 80 [**1**]). Which of the following parametrizations are identifiable ? (Prove or disprove.)

(1) $X_1, ..., X_p$ are independent r.v. with $X_i \sim N(\alpha_i + \nu, \sigma^2)$. $\theta = \left(\alpha_1, ..., \alpha_p, \nu, \sigma^2\right)$ and $\mathbb{P}_\theta$ is the distribution of $X = (X_1, ..., X_n)$
(2) Same as (1) above with $\alpha = (\alpha_1, ..., \alpha_p)$ restricted to $\{\alpha : \sum_{i=1}^p \alpha_i = 0\}$.
(3) X and Y are independent $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, $\theta = (\mu_1, \mu_2)$ and we observe $Y - X$
(4) $X_{ij}, i = 1, ..., p; j = 1, ..., b$ are independent with $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ where $\mu_{ij} = \nu + \alpha_i + \lambda_j$, $\theta = (\alpha_1, ..., \alpha_p, \lambda_1, ..., \lambda_b, \nu, \sigma^2)$ and $\mathbb{P}_\theta$ is the distribution of $X_{11}, ..., X_{pb}$.
(5) Same as (4) with $(\alpha_1, ..., \alpha_p)$ and $(\lambda_1, ..., \lambda_b)$ restricted to sets where $\sum_{i=1}^p \alpha_i = 0$ and $\sum_{j=1}^b \lambda_j = 0$.

PROBLEM 6.4 (Problem 2.1.4, page 81 [**1**]). 5. The number n of graduate students entering a certain department is recorded. In each of k subsequent years the number of students graduating and of students dropping out is recorded. Let N_i be the number dropping out and M_i the number graduating by the end of year i, $i = 1, ..., k$. The following model is proposed.

$$\mathbb{P}_\theta\big(N_1 = n_1, M_1 = m_1, ..., N_k = n_k, M_k = m_k\big) = \frac{n!}{n_1!...n_k!m_1!...m_k!r!}\mu_1^{n_1}...\mu_k^{n_k}\nu_1^{m_1}...\nu_k^{m_k}\rho^r$$

where

$$\sum_{i=1}^{k}\mu_i + \sum_{j=1}^{k}\nu_j + \rho = 1, \mu_i, \nu_\in(0,1), \quad i = 1, ..., k$$
$$n_1 + ... + n_k + m_1 + ... + m_k + r = n$$

and $\theta = (\mu_1, ..., \mu_k, \nu_1, ..., \nu_k)$ is unknown.

(1) What are the assumptions underlying the model ?
(2) θ is very difficult to estimate here if k is large. The simplification $\mu_i = \pi(1-\mu)^{i-1}\mu$, $\nu_i = (1-\pi)(1-\nu)^{i-1}\nu$ for $i = 1, .., k$ is proposed where $0 < \pi < 1$, $0 < \mu < 1$, $0 < \nu < 1$ are unknown. What assumptions underline the simplification ?

PROBLEM 6.5 (Problem 2.1.6 page 81, [**1**]). Which of the following models are regular ? (Prove or disprove)

(1) $\mathbb{P}_\theta$ is the distribution of X, when X is uniform on $(0, \theta)$, $\Theta = (0, \infty)$
(2) $\mathbb{P}_\theta$ is the distribution of X when X is uniform on $\{0, 1, ..., \theta\}$, $\Theta = \{1, 2, ...\}$
(3) Suppose $X \sim N(\mu, \sigma^2)$. Let $Y = 1$ if $X \leq 1$ and $Y = X$ if $X > 1$. $\theta = (\mu, \sigma^2)$ and $\mathbb{P}_\theta$ is the distribution of Y
(4) Suppose the possible control responses in an experiment are 0.1 , ... , 0.9 and they occur with frequencies $p(0.1)$,...,$p(0.9)$. Suppose the effect of a treatment is to increase the control response by a fixed amount θ. Let $\mathbb{P}_\theta$ be the distribution of a treatment response.

PROBLEM 6.6 (based on Problem 2.2.1, page 82, [**1**]). Let $X_1, ..., X_n$ be a sample from Poi(θ) population with $\theta > 0$.

(1) Show directly that $\sum_{i=1}^n X_i$ is a sufficient statistic
(2) Establish the same result by the F-N theorem
(3) Which one of the following statistics is sufficient:

$$T_1(X) = \Big(\sum_{i=1}^{n} X_i\Big)^2$$
$$T_2(X) = \big(X_1, ..., X_{n-1}\big)$$
$$T_3(X) = \big(T_1(X), T(X)\big)$$
$$T_4(X) = \big(T_1(X), T_2(X)\big)$$

(4) Order the statistics above according to coarseness relation
(5) Which statistic is minimal among the statistics mentioned above ?
(6) Show that $T(X)$ is minimal sufficient (among *all* sufficient statistics)

PROBLEM 6.7 (based on Problem 2.2.2, page 82, [**1**]). Let n items be drawn in order without replacement from a shipment of N items of which θN are bad. Let $X_i = 1$ if the i-th item drawn is bad, and $X_i = 0$ otherwise.

(1) Show directly that $T(X) = \sum_{i=1}^{n} X_i$ is sufficient
(2) Show that $T(X)$ is sufficient applying the F-N theorem
(3) Which of the following statistics is sufficient ?

$$T_1(X) = \sqrt{\sum_{i=1}^{n} X_i}$$
$$T_2(X) = \sum_{i=1}^{n-1} X_i$$
$$T_3(X) = \big(T_1(X), \min_i X_i\big)$$
$$T_4(X) = \big(T_1(X), T_2(X)\big)$$

(4) Order the statistics above according to coarseness relation
(5) Which statistic is minimal among the statistics mentioned above ?
(6) Show that $T(X)$ is minimal sufficient (among *all* sufficient statistics)

PROBLEM 6.8 (Problem 2.2.3, page 82, [**1**]). Let $X_1, ..., X_n$ be an i.i.d. sample from one of the following p.d.f.'s

(1)
$$f(x;\theta) = \theta x^{\theta-1}, \quad x \in (0,1), \theta > 0$$
(2) the Weibull density
$$f(x;\theta) = \theta a x^{a-1} \exp(-\theta x^a), \quad x, a, \theta \in (0,\infty)$$
(3) the Pareto density
$$f(x;\theta) = \theta a^{\theta}/x^{\theta+1}, \quad x, a, \theta \in (0,\infty)$$

where a is a fixed constant and θ is the unknown parameter. Find a real valued sufficient statistic for θ.

PROBLEM 6.9 (Problem 2.2.6 page 82 [**1**]). Let X take the specified values $v_1, ..., v_{k+1}$ with probabilities $\theta_1, ..., \theta_{k+1}$ respectively. Suppose that $X_1, ..., X_n$ are i.i.d. with the same distribution as X. Suppose that $\theta = (\theta_1, ..., \theta_{k+1})$ is unknown and may range over the set $\Theta = \{(\theta_1, ..., \theta_{k+1}) : \theta_i \geq 0, \sum_{i=1}^{k+1} \theta_i = 1\}$. Let N_j be the number of X_i which equal v_j

(1) Show that $N = (N_1, ..., N_k)$ is sufficient for θ
(2) What is the distribution of $(N_1, ..., N_{k+1})$?

PROBLEM 6.10 (Problem 2.2.7 page 83 [**1**]). Let $X_1, ..., X_n$ be an i.i.d. sample from the p.d.f.

$$f(x;\theta) = \frac{1}{\sigma} \exp\Big\{ - \frac{x-\mu}{\sigma} \Big\} I(x - \mu \geq 0).$$

Let $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$

(1) Show that $\min_i X_i$ is sufficient for μ when σ is fixed (known)
(2) Find a one dimensional sufficient statistic for σ when μ is fixed
(3) Exhibit a two dimensional sufficient statistic for θ

PROBLEM 6.11 (Problem 2.2.9 page 83 [**1**]). Let $X_1, ..., X_n$ be an i.i.d. sample from the p.d.f.

$$f(x;\theta) = a(\theta) h(x) I(x \in [\theta_1, \theta_2]),$$

where $h(x) \geq 0$, $\int_{\mathbb{R}} h(x) < \infty$, $\theta = (\theta_1, \theta_2)$ with $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_1 < \theta_2\}$ and $a(\theta) = \Big(\int_{\theta_1}^{\theta_2} h(x) dx \Big)^{-1}$. Find the two dimensional statistic for this problem and apply your result to the family of uniform distributions on $[\theta_1, \theta_2]$.

CHAPTER 7

Point estimation

Point estimation deals with estimating *the value* of the unknown parameter or a quantity, which depends on the unknown parameter (in a known way). More precisely, given a statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$, the observed data $X \sim \mathbb{P}_\theta$ and a function $q : \Theta \mapsto \mathbb{R}$, an *estimator* of $q(\theta)$ is a statistic $T(X)$, taking values in $q(\Theta) := \{q(\theta), q \in \Theta\}$. Estimating the value of the parameter itself, fits this framework with $q(\theta) := \theta$. The realization of the estimator for a particular set of data is called the *estimate*.

a. Methods of point estimation

In this section we shall introduce the basic methods of point estimation. Typically different methods would give different estimators. The choice of the particular method in a given problem depends on various factors, such as the complexity of the emerging estimation procedure, the amount of available data, etc. It should be stressed that all of the methods in this section, originated on a heuristic basis and none of them *guarantees* to produce good or even reasonable estimators at the outset: an additional effort is usually required to assess the quality of the obtained estimators and thus to refute or justify their practical applicability.

Remark 7a1. If $\hat\theta$ is a point estimator of θ, the quantity $q(\theta)$ can be estimated by the "plug-in" $\hat q(X) := q\big(\hat\theta(X)\big)$. Usually this yields reasonable results (being well justified in some situations - as e.g. in Lemma 7a15 below). Below we shall mainly focus on estimating θ itself, keeping in mind this remark.

Substitution principles. The methods, known collectively as *substitution principles*, are typically (but not exclusively) used for large i.i.d. samples $X_i \sim \mathbb{P}_\theta$, $i = 1, ..., n$ and, in the simplest form, are based on the following heuristics. Suppose that for a given statistical model $\mathbb{P}_\theta$, $\theta \in \Theta$, we manage to find a function $\phi : \mathbb{R} \mapsto \mathbb{R}$, such that $\psi(\theta) := \mathbb{E}_\theta\phi(X_1)$ is a one-to-one function of $\theta \in \Theta$. When the number of observed data points is large, motivated by the law of large numbers, we anticipate that the empirical average of ϕ would be close to its expectation. This suggests the following estimator of θ:

$$\hat\theta_n(X) := \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)\right).$$

Frequently (but not always), polynomials or indicator functions ϕ are used in practice, in which cases, this approach is known as the *method of moments* or *frequency substitution* respectively.

Example 6a3 (continued) For $X_1 \sim \mathrm{Ber}(\theta)$,

$$\mathbb{E}_\theta X_1 = \theta, \quad \forall\theta \in \Theta.$$

In this case the method of moments can be formally applied with $\phi(x) = x$ and $\psi(\theta) = \theta$, suggesting the statistic $\bar{X}_n$ as the estimator of θ.

Here is another possibility: note that $\mathbb{E}_\theta X_1 X_2 = \mathbb{E}_\theta X_1 \mathbb{E}_\theta X_2 = \theta^2$, which is invertible on $\theta \in [0,1]$. Hence by the method of moments

$$\tilde{\theta}_n := \sqrt{\frac{1}{[n/2]} \sum_{i=1}^{[n/2]} X_{2i-1} X_{2i}},$$

which is the estimator mentioned in (6a1).

The method of moments is well suited for estimation of functions of θ. For example, the empirical variance of $X_1, ..., X_n$, i.e. $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the method of moment estimator of the function $q(\theta) = \theta(1-\theta)$. ■

Example 7a2 (Hardy-Weinberg proportions). Consider (first generation of) a population in which the alleles[1] A and a are encountered with probabilities θ and $1-\theta$ respectively, $\theta \in (0,1)$. If the alleles are chosen at random and independently for each individual in the next generation, then the probability of having the AA genotype is θ^2, the aa genotype is $(1-\theta)^2$ and Aa genotype $2\theta(1-\theta)$. Note that the probabilities of alleles A and a in the second generation is the same as in the first: $\mathbb{P}(A) = \mathbb{P}(AA) + \frac{1}{2}\mathbb{P}(Aa) = \theta$ and $\mathbb{P}(a) = \mathbb{P}(aa) + \frac{1}{2}\mathbb{P}(Aa) = 1-\theta$. This property is known in genetics as equilibrium.

Suppose we sample n individuals from the population, observe their genotypes and would like to estimate the probability (proportion) of A allele in the population. The corresponding statistical model is an i.i.d. sample $X_1, ..., X_n$, where X_1 takes values in $\{AA, Aa, AA\}$ with probabilities θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$ respectively. Define the empirical frequencies $N_\ell := \sum_{i=1}^n I(X_i = \ell)$, $\ell \in \{AA, Aa, aa\}$ and note that

$$\mathbb{E}_\theta N_{AA} = \theta^2$$
$$\mathbb{E}_\theta N_{Aa} = 2\theta(1-\theta)$$
$$\mathbb{E}_\theta N_{aa} = (1-\theta)^2$$

The *frequency substitution* estimators, based N_{AA} and N_{aa}: are

$$\hat{\theta} = \sqrt{N_{AA}/n}$$
$$\bar{\theta} = 1 - \sqrt{N_{aa}/n}$$

Since $\mathbb{E}_\theta N_{Aa}$ is not a one-to-one function of $\theta \in \Theta$, it doesn't fit the frequency substitution method as is. Having several different estimators of θ can be practically handy, since some genotypes can be harder to observe than the others.

Here is another alternative: since $\mathbb{E}\big(I(X_1 = AA) + \frac{1}{2}I(X_1 = Aa)\big) = \theta^2 + \theta(1-\theta) = \theta$, the frequency substitution suggests the estimator:

$$\tilde{\theta}(x) = \frac{N_{AA}}{n} + \frac{1}{2}\frac{N_{Aa}}{n}.$$

■

The substitution principle is applicable to parameter spaces with higher dimension:

[1] look up for "HardyWeinberg principle" in wikipedia for more details about this model

EXAMPLE 7a3. Let $X_1, ..., X_n$ be an i.i.d. sample from the $\Gamma(\alpha, \beta)$ distribution with the p.d.f.

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} I(x \geq 0),$$

where the unknown parameter is $\theta := (\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+ =: \Theta$.

A calculation reveals that

$$\mathbb{E}_\theta X_1 = ... = \alpha\beta$$

and

$$\mathbb{E}_\theta X_1^2 = ... = \beta^2 \alpha(\alpha + 1).$$

Denote the empirical moments by

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i, \quad \overline{X_n^2} := \frac{1}{n} \sum_{i=1}^n X_i^2.$$

and

$$\hat{\sigma}_n^2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \overline{X_n^2} - \bar{X}_n^2.$$

The method of moments estimator of θ is given by the solution of the equations:

$$\bar{X}_n = \alpha\beta$$
$$\overline{X_n^2} = \beta^2 \alpha(\alpha + 1)$$

which gives

$$\hat{\alpha}(X) := \frac{\bar{X}_n^2}{\overline{X_n^2} - \bar{X}_n^2} = \frac{\bar{X}_n^2}{\hat{\sigma}_n^2(X)}, \quad \hat{\beta}(X) := \frac{\overline{X_n^2} - \bar{X}_n^2}{\bar{X}_n} = \frac{\hat{\sigma}_n^2(X)}{\bar{X}_n}$$

Note that $\hat{\alpha}$ is well defined since $\hat{\sigma}_n^2(X) \geq 0$ and the equality holds with zero probability. ■

REMARK 7a4. The method of moments (and other substitution principles) does not require precise knowledge of the distribution of the sample, but only of the dependence of moments on the parameter. This can be a practical advantage, when the former is uncertain.

Least squares estimation. In many practical situations the observed quantities are known to satisfy a noisy functional dependence, which itself is specified up to unknown parameters. More precisely, one observes the pairs (X_i, Y_i), $i = 1, ..., n$ which are presumed to satisfy the equations

$$Y_i = g_i(X_i, \theta) + \varepsilon_i, \quad i = 1, ..., n,$$

where g_i's are known functions, ε_i's are random variables and θ is the unknown parameter. The *least squares* estimator of θ is defined as

$$\hat{\theta}(X, Y) := \text{argmin}_{\theta \in \Theta} \sum_{i=1}^n \big(Y_i - g_i(X_i, \theta)\big)^2, \tag{7a1}$$

where any of the minimizers is chosen, when the minimum is not unique. In other words the least squares estimator is the best fit of a known curve to noisy measurements.

One classical example is the linear regression: one observes the pairs (X_i, Y_i) $i = 1, ..., n$, presumes that $Y_i = \theta_1 X_i + \theta_2 + \varepsilon_i$ and would like to estimate $\theta = (\theta_1, \theta_2)$ given the observations. Working out the minimization in (7a1) yields the familiar regression formulae.

Here is an example in the same spirit:

EXAMPLE 7a5. An asteroid is detected by a radar and its position is measured at times $0 = t_0 <= t_1 < ... < t_n$. Let $v = (v_x, v_y, v_z)$ be the velocity vector of the asteroid at time 0 (the first moment of detection). Assuming that the only force acting on the asteroid is gravitation[2], the position at any time $t \geq 0$ is given by

$$\begin{aligned} x(t) &= x_0 + v_x t \\ y(t) &= y_0 + v_y t \\ z(t) &= z_0 + v_z t - gt^2/2, \end{aligned}$$

where $p(0) = (x_0, y_0, z_0)$ is the asteroid position at time $t = 0$ and g is the Earth gravity constant $g = 9.78033...$ [m/s^2]. To predict the location and the time of the impact we have to estimate the vector of initial velocities v. Once the estimator $\hat{v}$ is calculated the impact point is predicted as the intersection of the curve

$$\begin{aligned} \hat{x}(t) &= x_0 + \hat{v}_x t \\ \hat{y}(t) &= y_0 + \hat{v}_y t \\ \hat{z}(t) &= z_0 + \hat{v}_z t - gt^2/2, \end{aligned}$$

with the equation describing the surface of the Earth (an ellipsoid, to the first order of approximation). Statistics is relevant in this problem for at least two reasons: (1) the position cannot be measured without errors, (2) many forces (weaker than Earth gravity) acting on the asteroid cannot be taken into account (and thus are modeled as noise). The relevant statistical model is

$$\begin{aligned} x_i := x(t_i) &= x_0 + v_x t_i + \varepsilon_x(i) \\ y_i := y(t_i) &= y_0 + v_y t_i + \varepsilon_y(i) \\ z_i := z(t_i) &= z_0 + v_z t_i - gt_i^2/2 + \varepsilon_z(i), \end{aligned}$$

where $\varepsilon_x, \varepsilon_y, \varepsilon_z$ are random variables, which model the errors. Notice that we don't have to assume any particular distribution of ε's to apply the LS method. Hence at this stage we don't really need the full description of statistical model (we shall need one to analyze the accuracy of the emerging estimators).

The data in this problem is (x_i, y_i, z_i, t_i) and the corresponding least squares estimator is given by:

$$\hat{v} = \text{argmin}_{v \in \mathbb{R}^3} \left(\sum_{i=1}^{n} \left(x_i - x_0 - v_x t_i \right)^2 + \sum_{i=1}^{n} \left(y_i - y_0 - v_y t_i \right)^2 + \sum_{i=1}^{n} \left(z_i - z_0 - v_z t_i + \frac{1}{2} g t_i^2 \right)^2 \right).$$

[2]this is an oversimplification: the trajectory of the asteroid is significantly affected by the drag force applied by the atmosphere, which itself varies depending on the mass being evaporated, the effective area of asteroid body and the material it is composed of. Moreover, the gravity of the Earth depends both on the instantaneous height and longitude/latitude coordinates, etc.

The latter is a quadratic function, whose minimum is found in the usual way by means of differention:

$$\hat{v}_x = \frac{\sum_i x_i t_i - x_0 \sum_i t_i}{\sum_i t_i^2}$$
$$\hat{v}_y = \frac{\sum_i y_i t_i - y_0 \sum_i t_i}{\sum_i t_i^2}$$
$$\hat{v}_z = \frac{\sum_i z_i t_i - z_0 \sum_i t_i + \frac{1}{2} g \sum_i t_i^3}{\sum_i t_i^2}$$

■

REMARK 7a6. This example demonstrates that in some situations it is easier to postulate the statistical model in the form, different from the canonical one, i.e. postulating a family of probability distributions. The definition of $\mathbb{P}_\theta$ in these cases is implicit: in the last example, $\mathbb{P}_\theta$ will be completely specified if we make the additional assumption that ε_i's are i.i.d. and sampled from[3] $N(0, \sigma^2)$ with known or unknown σ^2.

Maximum likelihood estimation.

DEFINITION 7a7. *For a regular model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and a sample $X \sim \mathbb{P}_\theta$, the Maximum Likelihood estimator (MLE) is*

$$\hat{\theta}(X) := \operatorname{argmax}_{\theta\in\Theta} L(x;\theta) = \operatorname{argmax}_{\theta\in\Theta} \log L(x;\theta),$$

assuming[4] argmax *exists and taking any of the maximizers, when it is not unique.*

The heuristical basis behind the ML method is to choose the parameter, for which the corresponding $\mathbb{P}_\theta$ assigns maximal probability to the observed realization of X.

REMARK 7a8. For many models, the likelihood function is a product of similar terms (e.g. when $\mathbb{P}_\theta$ corresponds to an i.i.d. sample), hence considering log-likelihood is more convenient. Clearly, this does not affect the estimator itself (since log is a strictly increasing function). Also, typically, the maximal value of the likelihood is not of immediate interest.

EXAMPLE 6a3 (continued) For n independent tosses of a coin, the log-likelihood is ($x \in \{0,1\}^n$, $\theta \in \Theta = [0,1]$):

$$\log L_n(x;\theta) = S_n(x) \log\theta + \big(n - S_n(x)\big) \log(1-\theta),$$

with $S_n(x) = \sum_{i=1}^n x_i$. If $S_n(x) \notin \{0, n\}$, then

$$\lim_{\theta\to 0} \log L_n(x;\theta) = \lim_{\theta\to 1} \log L_n(x;\theta) = -\infty$$

[3] the choice of $N(\mu, \sigma^2)$ with unknown μ leads to a non-identifiable model - think why

[4] This assumption is sometimes shortly incorporated into the definition of the MLE:

$$\hat{\theta}(X) \in \operatorname{argmax}_{\theta\in\Theta} L(x;\theta),$$

i.e. $\hat{\theta}(X)$ is any point in the set of maximizers.

and hence the maximum is attained in the interior of Θ. As $\log L_n(x;\theta)$ is a smooth function of θ on $(0,1)$, all local maximizers are found by differentiation:

$$S_n(x)\frac{1}{\theta} - \big(n - S_n(x)\big)\frac{1}{1-\theta} = 0,$$

which gives

$$\hat\theta_n(X) = \frac{1}{n}S_n(X) = \bar X_n, \tag{7a2}$$

which is the familiar intuitive estimator. Since the local maximum is unique it is also the global one and hence the MLE.

If $S_n(x) = 0$, $\log L_n(x;\theta) = n\log(1-\theta)$, which is a decreasing function of θ. Hence in this case, the maximum is attained at $\theta = 0$. Similarly, for $S_n(x) = n$, the maximum is attained at $\theta = 1$. Note that these two solutions are included in the general formula (7a2). ■

EXAMPLE 7a2 (continued) The log-likelihood is

$$\log L_n(x;\theta) = N_{AA}(x)\log\theta^2 + N_{Aa}(x)\log 2\theta(1-\theta) + N_{aa}(x)\log(1-\theta)^2 = \\ \big(2N_{AA}(x) + N_{Aa}(x)\big)\log\theta + N_{Aa}(x)\log 2 + \big(2N_{aa}(x) + N_{Aa}(x)\big)\log(1-\theta).$$

The maximization, done as in the previous example, gives the following intuitive estimator:

$$\hat\theta(x) = \frac{2N_{AA}(x) + N_{Aa}(x)}{2n} = \frac{N_{AA}(x)}{n} + \frac{1}{2}\frac{N_{Aa}(x)}{n}.$$

Note that it coincides with one of the frequency substitution estimators obtained before (but not the others). ■

EXAMPLE 7a9. Let $X_1, ..., X_n$ be an i.i.d. sample from $N(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2) \in \Theta = \mathbb{R}\times\mathbb{R}_+$. The log-likelihood is

$$\log L_n(x;\theta) = \log\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\theta_2 - \sum_{i=1}^n\frac{(x_i-\theta_1)^2}{2\theta_2}.$$

This function is differentiable on $\mathbb{R}\times\mathbb{R}_+$ and its gradient vanishes at all the local maximizers:

$$\frac{\partial}{\partial\theta_1}\log L_n(x;\theta) = \sum_{i=1}^n\frac{(x_i-\theta_1)}{\theta_2} = 0$$

$$\frac{\partial}{\partial\theta_2}\log L_n(x;\theta) = -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^n(x_i-\theta_1)^2 = 0.$$

Solving these equations for θ_1 and θ_2 gives the unique solution:

$$\hat\theta_1(x) = \frac{1}{n}\sum_{i=1}^n x_i = \bar x_n$$

$$\hat\theta_2(x) = \frac{1}{n}\sum_{i=1}^n(x_i-\bar x_n)^2$$

To see that the obtained extremum is indeed a local maximum we shall examine the Hessian matrix (of second order derivatives) and verify that it's eigenvalues are negative at $(\hat\theta_1, \hat\theta_2)$. A calculation shows that it is indeed the case.

To see that the found local maximum is also global, we have to check the value of the log-likelihood on the boundary, namely as θ_2 approaches to zero. The case $x_1 = x_2 = ... = x_n$ can be excluded from the consideration, since probability of getting such sample is zero. Hence $\sum_i (x_i - \theta_1)^2$ is strictly positive uniformly over $\theta_1 \in \mathbb{R}$. This in turn implies $\lim_{\theta_2 \to 0} \log L_n(x;\theta) = -\infty$ uniformly over $\theta_1 \in \mathbb{R}$. Hence the found maximizer is the MLE of θ. ■

EXAMPLE 7a10. Let $X_1, ..., X_n$ be a sample from the uniform density on $[0, \theta]$, with the unknown parameter $\theta \in \Theta = (0, \infty)$. The likelihood is

$$L_n(x;\theta) = \prod_{i=1}^n \frac{1}{\theta} I(x_i \in [0,\theta]) = (1/\theta)^n I(\max_i x_i \le \theta) = \begin{cases} 0 & \theta < \max_i x_i \\ \frac{1}{\theta^n} & \theta \ge \max_i x_i \end{cases}, \quad x \in \mathbb{R}^n_+.$$

Since $1/\theta^n$ is a decreasing function, the unique maximum is attained at $\hat\theta(x) = \max_i x_i$. ■

The computation of the MLE amounts to solving an optimization problem and hence neither existence nor uniqueness of the MLE is clear in advance. Moreover, even when MLE exists and is unique, its actual calculation may be quite challenging and in many practical problems is done numerically (which is typically not a serious drawback in view of the computational power of the modern hard/software).

Below are some examples, which demonstrate the existence and uniqueness issues.

EXAMPLE 7a11. Let $Y_1, ..., Y_n$ be an i.i.d. sample from $U([0,1])$ and let $X_i = Y_i + \theta$, where $\theta \in \Theta = \mathbb{R}$. We would like to estimate θ given $X_1, ..., X_n$. The corresponding likelihood is

$$L_n(x;\theta) = \prod_{i=1}^n I(\theta \le x_i \le \theta + 1) = I(\min_i x_i \ge \theta, \max_i x_i \le \theta + 1) = I(\max_i x_i - 1 \le \theta \le \min_i x_i).$$

Note that for any fixed n, $\mathbb{P}_\theta\Big(\min_i X_i - \big(\max_i X_i - 1\big) = 0\Big) = \mathbb{P}_\theta\Big(\max_i X_i - \min_i X_i = 1\Big) = 0$ and hence with probability 1, the maximizer, i.e. the MLE, is not unique: any point in the interval $[\max_i X_i - 1, \min_i X_i]$ can be taken as MLE. ■

Here is a natural and simple example when MLE does not exist:

EXAMPLE 7a12. Let $X_1, ..., X_n$ be sampled from Poi(θ), with unknown $\theta \in \Theta = \mathbb{R}_+$. The log-likelihood is

$$\log L_n(x;\theta) = -n\theta + \Big(\sum_{i=1}^n x_i\Big) \log\theta - \sum_{i=1}^n \log x_i!, \quad x \in \mathbb{N}^n.$$

If $S_n(x) = \sum_{i=1}^n x_i > 0$, this expression maximized by $\hat\theta(x) = S_n(x)/n = \bar x_n$. If, however, $S_n(x) = 0$, i.e. the sample was all zeros, $\log L_n(x;\theta) = -n\theta$, which does not have a maximum on the *open* interval $(0, \infty)$ (it's supremum is 0, which does not belong to the parametric space). ■

Here is a more vivid demonstration

EXAMPLE 7a13 (Kifer-Wolfovitz). Let $\xi_1 \sim \text{Ber}(1/2)$ and define

$$X_1 = Z_1(\sigma\xi_1 + 1 - \xi_1) + \mu,$$

where $Z_1 \sim N(0,1)$, independent of ξ_1. Hence, conditioned on $\{\xi_1 = 0\}$, $X_1 \sim N(\mu, 1)$ and conditioned on $\{\xi = 1\}$, $X_1 \sim N(\mu, \sigma^2)$.

Suppose we observe[5] the i.i.d. r.v.'s $X_1, ..., X_n$ and would like to estimate $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. The likelihood is ($\varphi$ is the standard Gaussian p.d.f. and $x \in \mathbb{R}^n$)

$$L_n(x;\theta) = \prod_{i=1}^n \left(\frac{1}{2}\frac{1}{\sigma}\varphi\Big(\frac{x_i - \mu}{\sigma}\Big) + \frac{1}{2}\varphi\Big(x_i - \mu\Big)\right) \geq \frac{1}{2}\frac{1}{\sigma}\varphi\Big(\frac{x_1 - \mu}{\sigma}\Big) \prod_{i=2}^n \frac{1}{2}\varphi\Big(x_i - \mu\Big),$$

where the inequality is obtained by removing all the missing positive terms. The obtaned lower bound can be made arbitrarily large by choosing $\mu := x_1$ and taking $\sigma \to 0$. Hence the MLE does not exist. ■

REMARK 7a14. Notice that MLE must be a function of a sufficient statistic, which immediately follows from the F-N factorization theorem. In particular, MLE can be constructed using the minimal sufficient statistic. This is appealing from the practical point of view, since the minimal sufficient statistic is all we need for the inference purpose in general (see also Remark 7d11).

Here is another useful property of MLE:

LEMMA 7a15. *MLE is invariant under reparametrization. More precisely, suppose $\hat\theta$ is the MLE for the model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and let h be a one-to-one function, mapping the parameter space onto a set E. Define $(\mathbb{P}_\eta)_{\eta\in E}$ by $\mathbb{P}_\eta := \mathbb{P}_{h^{-1}(\eta)}$, $\eta \in E$. Then the MLE of η is given by $\hat\eta = h(\hat\theta)$.*

PROOF. Let $\tilde L(x;\eta)$ be the likelihood function for the reparameterized model:

$$\tilde L(x;\eta) = L(x; h^{-1}(\eta)).$$

By the definition of MLE $L(x;\theta) \leq L(x;\hat\theta)$ and hence

$$\tilde L(x;\eta) = L(x; h^{-1}(\eta)) \leq L(x;\hat\theta) = \tilde L(x; h(\hat\theta)),$$

which means that $h(\hat\theta)$ maximizes the likelihood $\tilde L(x;\eta)$, i.e. it is the MLE of η. □

EXAMPLE 7a16. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$. It is required to calculate the MLE of the first two moments, i.e. $m_1(\mu) = \mu$ and $m_2(\mu, \sigma^2) = \sigma^2 + \mu^2$. The function (m_1, m_2) is invertible and hence by the Lemma the MLE's are

$$\hat m_1(X) = \hat\mu(X) = \bar X_n,$$

and

$$\hat m_2(X) = \hat\sigma^2(X) + \hat\mu(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar X_n)^2 + \bar X_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

The direct calculation is more cumbersome (try!) ■

[5] such sample is obtained by independent tossings of a fair coin, and sampling from $N(\mu, \sigma^2)$ if it comes out heads and from $N(\mu, 1)$ otherwise.

In summary let us revisit:

EXAMPLE 6a6 (continued) From a shipment of N oranges, in which θN oranges rot off, we sample without replacement $n \leq N$ oranges and would like to estimate θ, the proportion of wasted fruit. The available data is the number of rotten fruit in the sample X, which has the Hyper Geometric distribution:

$$p_X(k;\theta) = \frac{\binom{\theta N}{k}\binom{(1-\theta)N}{n-k}}{\binom{N}{n}}, \quad \max(0, n-(1-\theta)N) \leq k \leq \min(n, \theta N),$$

where $\theta \in \Theta = \left\{0, \frac{1}{N}, ..., \frac{N}{N}\right\}$.

Let's apply the method of moments to find an estimator of θ: to this end, we shall use the first moment[6]of X:

$$\mathbb{E}_\theta X = n\theta.$$

The statistic X/n, suggested by the method of moments, may take non-integer values and e.g. can be rounded to yield a valid estimator:

$$\tilde{\theta}(X) = \frac{1}{N}\lfloor NX/n \rfloor.$$

Now let's calculate the MLE of θ. For convenience we shall reparameterize the model by defining $M := \theta N$ (the number of rotten oranges in the shipment). By Lemma 7a15, $\hat{\theta}(X) = \hat{M}(X)/N$. The likelihood for the model with the new parametrization is

$$L(k, M) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad M \in \{0, ..., N\}.$$

---

[6]Here is how $\mathbb{E}_\theta X$ can be calculated: let ξ_i be the indicator of the i-th orange in the sample being rotten. Thus $X = \sum_{i=1}^n \xi_i$ and $\mathbb{E}_\theta X = \sum_{i=1}^n \mathbb{E}_\theta \xi_i$. For $j = 2, ..., n$,

$$\mathbb{E}_\theta \xi_j = \mathbb{E}_\theta \mathbb{P}_\theta(\xi_j | \xi_1, ..., \xi_{j-1}) = \mathbb{E}_\theta \frac{\theta N - \sum_{\ell=1}^{j-1} \xi_\ell}{N-(j-1)} = \frac{\theta N - \mathbb{E}_\theta \sum_{\ell=1}^{j-1} \xi_\ell}{N-(j-1)}.$$

Hence the quantities $r_m = \sum_{i=1}^m \mathbb{E}_\theta \xi_i$, $m = 1, ..., n$ satisfy the recursion,

$$r_m = \sum_{i=1}^m \frac{\theta N - r_{j-1}}{N-(j-1)} = \frac{\theta N - r_{m-1}}{N-(m-1)} + \sum_{i=1}^{m-1} \frac{\theta N - r_{j-1}}{N-(j-1)} = \frac{\theta N - r_{m-1}}{N-(m-1)} + r_{m-1}$$

subject to $r_0 = 0$. Hence $r_1 = \theta$ and by induction $r_m = m\theta$. Hence $\mathbb{E}_\theta X = r_n = n\theta$.

Let's study the monotonicity of $L(k, M)$, as a sequence in M for a fixed k. For $M \geq k$

$$\frac{L(k,M+1)}{L(k,M)} = \frac{\binom{M+1}{k}\binom{N-M-1}{n-k}}{\binom{N}{n}} \frac{\binom{N}{n}}{\binom{M}{k}\binom{N-M}{n-k}} =$$

$$\frac{(M+1)!}{k!(M+1-k)!}\frac{(N-M-1)!}{(n-k)!(N-M-1-(n-k)!}\frac{k!(M-k)!}{M!}\frac{(n-k)!(N-M-(n-k))!}{(N-M)!} =$$

$$\frac{(M+1)}{(M+1-k)}\frac{(N-M-(n-k))}{(N-M)} = \frac{1-\frac{n-k}{N-M}}{1-\frac{k}{M+1}}.$$

Let's see for which values of M the latter expression is less than 1:

$$\frac{1-\frac{n-k}{N-M}}{1-\frac{k}{M+1}} \leq 1$$

$$\Updownarrow$$

$$1-\frac{n-k}{N-M} \leq 1-\frac{k}{M+1}$$

$$\Updownarrow$$

$$\frac{n-k}{N-M} \geq \frac{k}{M+1}$$

$$\Updownarrow$$

$$M \geq \frac{k}{n}(N+1) - 1.$$

Hence the sequence $M \mapsto L(k, M)$ increases for all M's less than $M^* := \frac{k}{n}(N+1) - 1$ and decreases otherwise. If M^* is an integer, then $L(k, M^*) = L(k, M^*+1)$ and $L(k, M^*) > L(k, M)$ for all $M \notin \{M^*, M^*+1\}$, hence the maximizer is not unique:

$$\hat{M}(k) \begin{cases} = 0, & k = 0 \\ \in \left\{M^*+1, M^*\right\}, & k \notin \{0, n\}, \\ = N, & k = n \end{cases}$$

and $\hat{\theta}(X) = \hat{M}(X)/N$.

If k/n is non-integer, then $\hat{M}(k) = \lfloor k/n \rfloor (N+1)$ and $\hat{\theta}(X) = \lfloor X/n \rfloor (1 + 1/N)$, which is only slightly different from the method of moments estimator. ■

b. Elements of the statistical decision theory

In the previous section we have seen several techniques, which can be used to construct point estimators. Different methods produce different estimators (and hence estimates) in general. How do we compare the performance of point estimators? For example, suppose we obtain an i.i.d. sample $X_1, ..., X_n$ from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. As we saw, the MLE estimator of σ is given by:

$$\hat{\sigma}_n(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}. \tag{7b1}$$

However, another estimator[7]

$$\tilde{\sigma}_n(X) = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_n|, \tag{7b2}$$

appears as plausible as $\hat{\sigma}_n(X)$. Which one is better? As we shall see, in a certain sense, this question can be resolved in favor of $\hat{\sigma}_n(X)$; however, the answer is far from being obvious at the outset. The questions of comparison and optimality of *decision rules*, such as point estimators or statistical tests, etc. are addressed by the *statistical decision theory* [8]. Below we will present the essentials in the context of point estimators, deferring consideration of statistical hypothesis testing to the next chapter.

It is clear that comparing the realizations of estimators, i.e. the corresponding estimates, is meaningless, since the conclusion will depend on the particular outcome of the experiment. Hence we shall compare the *expected* performances of the point estimators.

Definition 7b1. *For a statistical model* $(\mathbb{P}_\theta)_{\theta\in\Theta}$ *and a* loss function $\ell : \Theta \times \Theta \mapsto \mathbb{R}_+$*, the* ℓ*-risk of the point estimator* T *is*

$$R_\ell(\theta; T) := \mathbb{E}_\theta \ell\big(\theta, T(X)\big),$$

where $X \sim \mathbb{P}_\theta$.

The *loss* function in this definition measures the 'loss' incurred by estimating θ by $T(X)$ (for each particular realization) and the *risk* is the expected loss. Intuitively, the loss should increase with the magnitude of deviation of the estimate from the true value of the parameter θ, i.e. larger errors should be assigned greater loss. Here are some popular loss functions for $\Theta \subseteq \mathbb{R}^d$:

$$\begin{aligned}
\ell_p(\theta, \eta) &= \sum_{i=1}^{d}|\theta_i - \eta_i|^p =: \|\theta - \eta\|_p^p \\
\ell_0(\theta, \eta) &= I(\theta \neq \eta) \\
\ell_\infty(\theta, \eta) &= \max_i |\theta_i - \eta_i|
\end{aligned}$$

[7]The comparison between $\hat{\sigma}_n(X)$ and $\tilde{\sigma}_n(X)$ was the subject of the dispute between Sir. R.Fihser, who advocated for the MLE, and the physicist A.Eddington, who favored the other estimator.

[8]a concise account can be found in e.g. [**2**]

These losses can be associated with distances between θ and η: the larger distance the greater loss is suffered. The corresponding risks of an estimator T are given by

$$R_p(\theta,T)=\mathbb{E}_\theta\sum_{i=1}^d|\theta_i-T_i(X)|^p=\mathbb{E}_\theta\|\theta-\eta\|_p^p$$
$$R_0(\theta,T)=\mathbb{E}_\theta I(\theta\neq T(X))=\mathbb{P}_\theta\big(\theta\neq T(X)\big)$$
$$R_\infty(\theta,T)=\mathbb{E}_\theta\max_i|\theta_i-T_i(X)|.$$

The choice of the loss function in a particular problem depends on the context, but sometimes is motivated by the simplicity of the performance analysis. In this course we shall consider almost exclusively the quadratic (MSE) risk $R_2(\theta,T)$ (assuming that $T(X)$ is such that the risk is finite and omitting 2 from the notations):

$$R(\theta,T)=\mathbb{E}_\theta\big(\theta-T(X)\big)^2.$$

This risk makes sense in many models and, moreover, turns to be somewhat more convenient to deal with technically.

Remark 7b2. If the objective is to estimate $q(\theta)$, for a known function q, the MSE risk is defined similarly

$$R(q(\theta),T)=\mathbb{E}_\theta(q(\theta)-T(X))^2,$$

and the theory we shall develop below translates to this more general case with minor obvious adjustments.

In view of the above definitions, we are tempted to compare two estimators T_1 and T_2 by comparing their risks, i.e. $R(\theta,T_1)$ and $R(\theta,T_2)$. Sometimes, this is indeed possible:

Example 7b3. Let $X_1,...,X_n$ be an i.i.d. sample from $U([0,\theta])$, $\theta>0$. As we saw, the MLE of θ is

$$\hat\theta_n(X)=\max_i X_i=:M_n(X)$$

Another reasonable estimator of θ can be suggested by the method of moments:

$$\tilde\theta_n(X)=2\bar X_n.$$

Recall that M_n has the p.d.f.

$$f_{M_n}(x)=\frac{n}{\theta^n}x^{n-1}I(x\in(0,\theta)).$$

Hence

$$\mathbb{E}_\theta M_n=\int_0^\theta\frac{n}{\theta^n}x^n dx=\theta\frac{n}{n+1}\tag{7b3}$$

and

$$\mathbb{E}_\theta M_n^2=\int_0^\theta\frac{n}{\theta^n}x^{n+1}dx=\theta^2\frac{n}{n+2}.$$

Consequently,

$$R(\theta,\hat\theta_n)=\mathbb{E}_\theta(\theta-M_n)^2=\theta^2-2\theta\mathbb{E}_\theta M_n+\mathbb{E}_\theta M_n^2=$$
$$\theta^2\Big(1-\frac{2n}{n+1}+\frac{n}{n+2}\Big)=\theta^2\frac{2}{(n+1)(n+2)}.\tag{7b4}$$

The risk of $\tilde{\theta}_n$ is given by

$$R(\theta, \tilde{\theta}_n) = \mathbb{E}_\theta(\theta - 2\bar{X}_n)^2 = \frac{4}{n}\mathrm{var}_\theta(X_1) = \frac{4}{n}\frac{\theta^2}{12} = \theta^2\frac{1}{3n}. \tag{7b5}$$

Since

$$\frac{1}{3n} - \frac{2}{(n+1)(n+2)} = \frac{(n-1)(n-2)}{3n(n+1)(n+2)} \geq 0,$$

it follows

$$R(\theta, \hat{\theta}_n) \leq R(\theta, \tilde{\theta}_n), \quad \forall\theta \in \Theta = (0, \infty),$$

where the inequality is strict for $n \geq 3$. Hence $\hat{\theta}_n$ yields better (smaller) risk than $\tilde{\theta}_n$, for all values of the parameter. ■

This example motivates the following notion

DEFINITION 7b4. *An estimator $\tilde{\theta}$ is inadmissible, if there exists an estimator $\tilde{\theta}'$, such that*

$$R(\theta, \tilde{\theta}) \geq R(\theta, \tilde{\theta}') \quad \forall\theta \in \Theta,$$

and the inequality is strict at least for some θ.

In other words, an estimator is inadmissible if there is an estimator with better risk. Let us stress once again that better risk means that the risk function is smaller or equal *for all* $\theta \in \Theta$.

The notion of the *admissible estimator* is obtained by negation: $\tilde{\theta}$ is admissible if no other estimator has better risk. Notice that this does not exclude the possibility of having an estimator which yields smaller risk only on a part of the parameter space (in fact, we shall see shortly,this is always the case).

In the preceding example, $\tilde{\theta}_n(X)$ is inadmissible for $n \geq 3$. Is $\hat{\theta}_n(X)$ admissible ...? This cannot be concluded on the basis of the preceding calculations: to establish admissibility of $\hat{\theta}_n(X)$ we have to prove that there doesn't exist an estimator which improves the risk of $\hat{\theta}_n(X)$ for all values of the parameter.

REMARK 7b5. Practically one may prefer an inadmissible estimator due to its simpler structure, subjective appeal, etc.

REMARK 7b6. Admissibility is a very weak property. For example, for $X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$, the constant estimator $\hat{\theta} \equiv c$, which doesn't depend on X is admissible ! Suppose the contrary holds, i.e. $\hat{\theta} \equiv c$ is inadmissible, then there is an estimator $\hat{\theta}'(X)$ such that $\mathbb{E}_\theta(\theta - \hat{\theta}'(X))^2 \leq (\theta - c)^2$ for all $\theta \in \mathbb{R}$ and for some θ this inequality is strict. In particular, for $\theta := c$ we get $\mathbb{E}_c(c - \hat{\theta}'(X))^2 = 0$, i.e. $\hat{\theta}'(X) = c$, $\mathbb{P}_c$-a.s. Let φ be the standard normal density, then for any $\theta \in \mathbb{R}$

$$\mathbb{E}_\theta(\theta - \hat{\theta}'(X))^2 = \int(\theta - \hat{\theta}'(x))^2\varphi(x-\theta)dx = \int(\theta - \hat{\theta}'(x))^2\frac{\varphi(x-\theta)}{\varphi(x-c)}\varphi(x-c)dx =$$

$$\mathbb{E}_c(\theta - \hat{\theta}'(X))^2\frac{\varphi(X-\theta)}{\varphi(X-c)} = (\theta-c)^2\mathbb{E}_c\frac{\varphi(X-\theta)}{\varphi(X-c)} = (\theta-c)^2,$$

i.e. the risk of $\hat{\theta}'$ coincides with the risk of $\hat{\theta} \equiv c$. The obtained contradiction shows that $\hat{\theta} \equiv c$ is admissible.

REMARK 7b7. Establishing inadmissibility of an estimator amounts to finding another estimator with better risk. Establishing admissibility of an estimator appears to be a much harder task: we have to check that better estimators don't exist! It turns out that *constructing* an admissible estimator is sometimes an easier objective. In particular, the Bayes estimators are admissible (see Lemma 7c8 below).

The following celebrated example, whose earlier version was suggested by C.Stein, demonstrates that estimators can be inadmissible in a surprising way

EXAMPLE 7b8 (W.James and C.Stein, 1961). Let X be a normal vector in $\mathbb{R}^p$ with independent entries $X_i \sim N(\theta_i, 1)$. Given a realization of X, it is required to estimate the vector θ. Since X_i's are independent, X_i doesn't seem to bear any relevance to estimating the values of θ_j, for $j \neq i$. Hence it makes sense to estimate θ_i by X_i, i.e. $\hat{\theta} = X$. This estimator is reasonable from a number of perspectives: it is the ML estimator and, as we shall see below, the optimal unbiased estimator of θ and also the minimax estimator. The quadratic risk of $\hat{\theta} = X$ is constant

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta \|X - \theta\|^2 = \mathbb{E}_\theta \sum_{i=1}^{p} (X_i - \theta_i)^2 = p, \quad \theta \in \mathbb{R}^p.$$

This 'natural' estimator $\hat{\theta} = X$ can also be shown admissible for $p = 1$ and $p = 2$, but, surprisingly, not for $p \geq 3$! The following estimator, constructed by W.James and C.Stein, has a strictly better risk for all $\theta \in \mathbb{R}^p$:

$$\hat{\theta}^{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X.$$

Note that this estimator, unlike $\hat{\theta} = X$, uses all the components of X to estimate each individual component of θ and the values of X with smaller norms are pushed further towards the origin. The latter property is often referred to as *shrinkage* and the estimators with similar property are called *shrinking* estimators. Some further details can be found in the concise article [**11**].

To compute the risk of $\hat{\theta}^{JS}$, we shall need the following simple identity:

LEMMA 7b9 (Stein's lemma). *For $\zeta \sim N(0,1)$ and a continuously differentiable function $h : \mathbb{R} \mapsto \mathbb{R}$,*

$$\mathbb{E}\zeta h(\zeta) = \mathbb{E}h'(\zeta),$$

whenever the expectations are well defined and finite.

PROOF. Note that the standard normal density φ satisfies $\varphi'(x) = -x\varphi(x)$ and under our integrability assumptions, the claim follows by integration by parts

$$\mathbb{E}h'(\zeta) = \int h'(x)\varphi(x)dx = -\int h(x)\varphi'(x)dx = \int h(x)x\varphi(x)dx = \mathbb{E}\zeta h(\zeta).$$

□

Also we shall check that for $p \geq 3$,

$$\mathbb{E}_\theta \frac{1}{\|X\|^2} < \infty.$$

To this end, let e_1 be the vector with the entries $e_{1,1} = 1$ and $e_{1,i} = 0$ for $i = 2, ..., p$ and note that the vectors θ and $e_1\|\theta\|$ have the same norms. Hence there exists an orthonormal (rotation)

matrix U_θ, such that $U_\theta\theta = e_1\|\theta\|$. Since $X = Z + \theta$, where Z is the vector with i.i.d. $N(0,1)$ entries,

$$\|X\|^2 = \|U_\theta X\|^2 = \left\|U_\theta Z + e_1\|\theta\|\right\|^2.$$

But as U_θ is orthonormal, $U_\theta Z$ has the same distribution as Z and

$$\mathbb{E}_\theta \frac{1}{\|X\|^2} = \int_{\mathbb{R}^p} \frac{1}{\left\|z + e_1\|\theta\|\right\|^2} \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\|z\|^2} dz =$$
$$\int_{\mathbb{R}^p} \frac{1}{\|v\|^2} \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(v_1 - \|\theta\|)^2 - \frac{1}{2}\sum_{i\geq 2} v_i^2\right) dv \leq$$
$$\int_{\mathbb{R}^p} \frac{1}{\|v\|^2} \exp\left(v_1\|\theta\| - \frac{1}{2}\sum_{i=1}^p v_i^2\right) dv \leq \int_{\mathbb{R}^p} \frac{1}{\|v\|^2} \exp\left(v_1\|\theta\| - \frac{1}{4}v_1^2 - \frac{1}{4}\sum_{i=1}^p v_i^2\right) dv \overset{\dagger}{\leq}$$
$$e^{\|\theta\|^2} \int_{\mathbb{R}^p} \frac{1}{\|v\|^2} \exp\left(-\frac{1}{4}\sum_{i=1}^p v_i^2\right) dv \overset{\ddagger}{=} e^{\|\theta\|^2} \int_0^\infty \frac{1}{r^2} \exp\left(-\frac{1}{4}r^2\right) r^{p-1} dr < \infty,$$

where in † we used the elementary inequality $xa - \frac{1}{4}x^2 \leq a^2$ for all $x \in \mathbb{R}$ and ‡ holds by the change of variables to polar coordinates (the term r^{p-1} is the Jacobian).

Now let us get back to calculating the risk of the James-Stein estimator:

$$\|\hat{\theta}^{JS} - \theta\|^2 = \left\|X - \frac{p-2}{\|X\|^2}X - \theta\right\|^2 = \|Z\|^2 - 2\left\langle Z, \frac{p-2}{\|X\|^2}X\right\rangle + \left\|\frac{p-2}{\|X\|^2}X\right\|^2 =$$
$$\|Z\|^2 - 2\left\langle Z, \frac{p-2}{\|X\|^2}X\right\rangle + \frac{(p-2)^2}{\|X\|^2},$$

where

$$\left\langle Z, \frac{p-2}{\|X\|^2}X\right\rangle = \sum_{i=1}^p Z_i X_i \frac{p-2}{\|X\|^2} = (p-2)\sum_{i=1}^p Z_i \frac{Z_i + \theta_i}{\|Z+\theta\|^2}.$$

Consider $h(x_i) := \frac{x_i + \theta_i}{\|x+\theta\|^2}$ as a function of $x_i \in \mathbb{R}$ with all other coordinates x_j, $j \neq i$ fixed, so that $x_j \neq \theta_j$ at least for some j. In this case, $h(x_i)$ is smooth in x_i and

$$\partial_i h(x_i) = \frac{\|x+\theta\|^2 - 2(x_i+\theta_i)^2}{\|x+\theta\|^4} = \frac{1}{\|x+\theta\|^2} - 2\frac{(x_i+\theta_i)^2}{\|x+\theta\|^4}.$$

Applying Stein's lemma and using independence of Z_i's, we get

$$\mathbb{E}_\theta Z_i \frac{Z_i + \theta_i}{\|Z+\theta\|^2} = \mathbb{E}_\theta \mathbb{E}_\theta\left(Z_i h(Z_i)\middle| Z_j, j \neq i\right) = \mathbb{E}_\theta \mathbb{E}_\theta\left(\partial_i h(Z_i)\middle| Z_j, j \neq i\right) =$$
$$\mathbb{E}_\theta \mathbb{E}_\theta\left(\frac{1}{\|Z+\theta\|^2} - 2\frac{(Z_i+\theta_i)^2}{\|Z+\theta\|^4}\middle| Z_j, j \neq i\right) = \mathbb{E}_\theta \frac{1}{\|X\|^2} - 2\mathbb{E}_\theta \frac{X_i^2}{\|X\|^4}$$

and hence

$$\mathbb{E}_\theta \left\langle Z, \frac{p-2}{\|X\|^2} X \right\rangle = (p-2) \sum_{i=1}^{p} \left(\mathbb{E}_\theta \frac{1}{\|X\|^2} - 2\mathbb{E}_\theta \frac{X_i^2}{\|X\|^4} \right) =$$

$$(p-2)\mathbb{E}_\theta \left(\frac{p}{\|X\|^2} - 2\frac{1}{\|X\|^2} \right) = (p-2)^2 \mathbb{E}_\theta \frac{1}{\|X\|^2}$$

Plugging this back, we obtain

$$R(\hat{\theta}^{JS}, \theta) = \mathbb{E}_\theta \|\hat{\theta}^{JS} - \theta\|^2 = \mathbb{E}_\theta \left(\|Z\|^2 - 2 \left\langle Z, \frac{p-2}{\|X\|^2} X \right\rangle + \frac{(p-2)^2}{\|X\|^2} \right) =$$

$$p - 2(p-2)^2 \mathbb{E}_\theta \frac{1}{\|X\|^2} + \mathbb{E}_\theta \frac{(p-2)^2}{\|X\|^2} = p - (p-2)^2 \mathbb{E}_\theta \frac{1}{\|X\|^2} < p = R(\hat{\theta}, \theta),$$

for all $\theta \in \mathbb{R}^p$. ■

However, risks do not *have to be* comparable. Here is a simple example:

EXAMPLE 7b10. Suppose we toss a coin twice, i.e. obtain an i.i.d. sample $X = (X_1, X_2)$ from $\mathrm{Ber}(\theta)$, $\theta \in [0,1]$. As we saw, the MLE of θ is given by $\hat{\theta}(X) = \frac{X_1+X_2}{2}$. Another estimator, suggested earlier by the frequency substitution, is $\tilde{\theta}(X) = \sqrt{X_1 X_2}$. Let's calculate the corresponding risks:

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \Big(\theta - \frac{X_1 + X_2}{2} \Big)^2 = \frac{1}{4} \big(\mathrm{var}_\theta(X_1) + \mathrm{var}_\theta(X_2) \big) = \frac{1}{2}\theta(1-\theta),$$

and

$$R(\theta, \tilde{\theta}) = \mathbb{E}_\theta \Big(\theta - \sqrt{X_2 X_2} \Big)^2 = \theta^2 - 2\theta \mathbb{E}_\theta \sqrt{X_2 X_2} + \mathbb{E}_\theta X_2 X_2 = \theta^2 - 2\theta^3 + \theta^2 = 2\theta^2(1-\theta).$$

The obtained risks are plotted at Figure 1: $R(\theta, \hat{\theta})$ is worse (greater) than $R(\theta, \tilde{\theta})$ for $\theta \in (0, 1/4)$ and vise versa for $\theta \in (1/4, 1)$ ■

This example shows that the risks of two estimators may satisfy opposite inequalities on different regions of Θ: since we do not know in advance to which of the regions the *unknown* parameter belongs, preferring one estimator to another in this situation does not make sense.

If we cannot compare some estimators, maybe we can still find the *best* estimator $\hat{\theta}^*$ for which

$$R(\theta, \hat{\theta}^*) \le R(\theta, \tilde{\theta}), \ \forall \theta \in \Theta \quad \text{for } \textit{all} \text{ estimators } \tilde{\theta} \ ?$$

A simple argument shows that the latter is also impossible! Suppose that the best estimator exists. Take the trivial estimator $\tilde{\theta}(X) \equiv \theta_0$. The corresponding risk

$$R(\theta, \tilde{\theta}) = \mathbb{E}_\theta (\theta - \theta_0)^2 = (\theta - \theta_0)^2,$$

vanishes at $\theta := \theta_0$ and hence the risk of the best estimator must vanish at θ_0 as well:

$$R(\theta_0, \hat{\theta}^*) \le R(\theta_0, \tilde{\theta}) = 0, \quad \Longrightarrow \quad R(\theta_0, \hat{\theta}^*) = 0.$$

But θ_0 was arbitrary, hence in fact $R(\theta, \hat{\theta}^*) = 0$ for all $\theta \in \Theta$, i.e. errorless estimation is possible, which is obviously a nonsense. This contradiction shows that the best estimator does not exist.

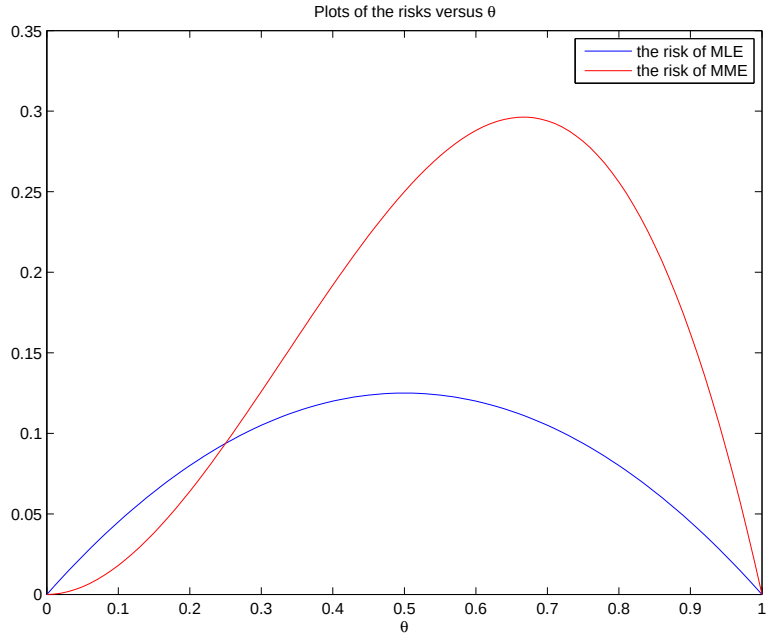

FIGURE 1. the risks $R(\theta, \hat{\theta})$ and $R(\theta, \tilde{\theta})$ as functions of $\theta \in [0, 1]$

REMARK 7b11. Note that nonexistence of the best estimator does not contradict the existence of the admissible estimators:

$$R(\theta, \hat{\theta}^a) \leq R(\theta, \tilde{\theta}), \quad \forall \theta \qquad \text{for all } \textit{comparable} \text{ estimators } \tilde{\theta}.$$

Indeed, in many problems an admissible estimator can be found explicitly.

In view of this discussion, we are faced with the two basic questions:

(1) How *any* two estimators can nevertheless be compared ?
(2) If we find a way to make all estimators comparable, does then the best estimator exist? If yes, how can it be actually found?

Answers can be provided by a number of approaches, which we shall first survey below and shall discuss some of them in details.

Reduction of the risk to scalar. The basic problem with comparing estimators by their risk functions is being unable to compare functions in general. The natural idea to cope with this problem is to reduce the risk function to a nonnegative real number (scalar) by means of some transformation. This way all estimators are assigned positive numbers and thus become comparable.

Minimax estimation. For an estimator $\hat{\theta}$, the maximal (supremum) risk is

$$r(\hat{\theta}) := \sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$

In words, $r(\hat\theta)$ is the worst value of the risk. Any two estimators $\hat\theta$ and $\tilde\theta$ are now compared by comparing the corresponding maximal risks $r(\hat\theta)$ and $r(\tilde\theta)$ and the estimator with smaller one is preferred.

The best estimator $\hat\theta^\circ$, i.e. the one which yields the smallest maximal risk is called *minimax*:

$$r(\hat\theta^\circ) = \sup_{\theta\in\Theta} R(\theta,\hat\theta^\circ) \le \sup_{\theta\in\Theta} R(\theta,\hat\theta) = r(\hat\theta), \quad \text{for any other estimator } \hat\theta.$$

In words, the minimax estimator optimizes the worst case performance. The principle disadvantage of this approach is that the estimator which yields the best performance in the worst case (which maybe quite bad on its own), may perform quite poorly for other values of the parameters.

Bayes estimation. In the Bayesian approach, one chooses a *prior* probability distribution (e.g. in the form of p.d.f. if Θ is a continuum) $\pi(\theta)$, $\theta\in\Theta$ and defines the Bayes risk of the estimator $\hat\theta$ as

$$r(\pi,\hat\theta) = \int_\Theta R(\theta,\hat\theta)\pi(\theta)d\theta.$$

Any two estimators are comparable by their Bayes risks. The estimator, which minimizes the Bayes risk, is called the *Bayes estimator* and is usually given by an explicit (in fact Bayes, what else ?) formula (see more details in Section c below).

The prior distribution π, from which the value of the parameter is sampled, is interpreted as the subjective a priori belief regarding θ. For example, if $\Theta=[0,1]$ (as in the coin tossing), the prior $\pi = U([0,1])$ makes all θ's equiprobable. If for some reason we believe that θ cannot deviate from $1/2$ too much, we may express this belief by choosing a prior, concentrated more around $\theta=1/2$, e.g. $\pi(\theta) = C_r\big(\theta(1-\theta)\big)^r$, where r is a positive number and C_r is the corresponding normalizing constant. The Bayes estimator is then viewed as the optimal fusion of the a priori belief about the unknown parameter and the data obtained in the experiment.

On one hand, introducing a prior can be viewed as an advantage, since it allows to incorporate prior information about the parameter into the solution. On the other hand, if such information is not available or, more frequently, cannot be translated to a particular prior distribution, the choice of the prior is left arbitrary and thus problematic (do not think that a uniform prior is not 'informative'!). On the third hand, the Bayes estimator enjoys many good decision theoretic properties and turns to yield the best possible behavior asymptotically, as the sample size increases. We shall discuss Bayes estimators in more details below.

Restricting estimators to a class. While the best estimator does not exist, remarkably it may exist and even be explicitly computable, if we consider a restricted class of estimators satisfying certain additional properties.

Equivariant estimators. Let $X=(X_1,...,X_n)$ be an i.i.d. sample from the density $f(x-\theta)$, where θ is the unknown *location* parameter[9]. An estimator $\hat\theta(X)$ of θ is called equivariant[10] if $\hat\theta(x+s) = s+\hat\theta(x)$ for any shift[11] $s\in\mathbb{R}$. Remarkably, this extra structure of the estimators guarantees existence of the best estimator, which can be explicitly computed.

[9] in other words, $X_i=\theta+Y_i$, where Y_i's are sampled from the p.d.f. f

[10] more generally, equivariance is defined with respect to a group of operations

[11] for a vector $x\in\mathbb{R}^n$ and a scalar $c\in\mathbb{R}$, $x-c$ stands for the vector with entries x_i-c

Unbiased estimators.

DEFINITION 7b12. *Let $X \sim \mathbb{P}_\theta$. The bias function of an estimator*[12] *$\hat{\theta}$ is*

$$b(\theta, \hat{\theta}) := \mathbb{E}_\theta \hat{\theta}(X) - \theta.$$

The estimator $\hat{\theta}$ is unbiased if $b(\theta, \hat{\theta}) = 0$, $\forall \theta \in \Theta$, i.e. $\mathbb{E}_\theta \hat{\theta}(X) = \theta$.

REMARK 7b13. If the objective is to estimate $q(\theta)$ for some known q, the bias of an estimator $\hat{q}(X)$ is defined similarly

$$b(q(\theta), \hat{q}) = \mathbb{E}_\theta \hat{q}(X) - q(\theta).$$

Unbiased estimators produce estimates which are dispersed around the true value of the parameter. This may be practically appealing in some problems. Moreover, it turns out that often it is possible to find the best estimator within the class of all unbiased estimators[13]. The related theory is very elegant and frequently useful. We shall dwell on it in details in Section d.

Asymptotic (large sample) theory. In many practical situations the amount of the available data is as large as we want. After all, repeating a statistical experiment is often only a matter of resources (time, money, etc.) and if the precision is important, one may be prepared to pay for it. In these cases, it makes sense to consider sequences of estimators and to compare them asymptotically, as the number of samples increases.

The typical example is estimation of the mean θ from the i.i.d. sample $X_1, ..., X_n$ from $N(\theta, 1)$ distribution. As we saw, the corresponding MLE is $\hat{\theta}_n(X) = \bar{X}_n$, which can be viewed as a sequence of estimators, indexed by n. Another reasonable estimator of the mean is the empirical median, i.e. $\tilde{\theta}_n(X) = \text{med}(X_1, ..., X_n) = X_{(\lfloor n/2 \rfloor)}$, where $X_{(i)}$ stands for the i-th order statistic.

It can be shown, that both estimators become more and more precise as n increases: in fact, both converge in a certain sense to the true value of the parameter as $n \to \infty$. Hence one can compare these estimators asymptotically as $n \to \infty$ by e.g. the rate of convergence or asymptotic variance, etc.

Asymptotic estimation has a deep and beautiful theory with many practical applications. We shall touch upon the basic notions of it in Section g below.

c. Bayes estimator

As mentioned above, the Bayes estimator, i.e. the estimator which minimizes the Bayes risk with respect to a prior π:

$$r(\hat{\theta}) = \int_\Theta R(\theta, \hat{\theta}) \pi(d\theta),$$

over all estimators $\hat{\theta}$ is computable explicitly:

PROPOSITION 7c1. *The Bayes estimator with respect to the MSE risk and the prior π is given by the Bayes formula:*

$$\hat{\theta}^*_\pi(X) = \frac{\int_\Theta sL(X; s)\pi(s)ds}{\int_\Theta L(X; r)\pi(r)dr}, \tag{7c1}$$

[12] Pay attention that definition of the bias implicitly requires the expectation to be well defined

[13] Note that the trivial estimators, such as used in the discussion preceding Remark 7b11 are biased and thus excluded from consideration.

where $L(x;\theta)$ is the likelihood of the model $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

PROOF. Suppose for definiteness[14] that $X \sim \mathbb{P}_\theta$ is a random vector in $\mathbb{R}^n$ with j.p.d.f. $f(x;\theta)$. Then the Bayes risk of an estimator $\hat{\theta}$ with respect to the prior π is

$$r_\pi(\hat{\theta}) = \int_\Theta R(s,\hat{\theta})\pi(s)ds = \int_\Theta\int_{\mathbb{R}^n} (s-\hat{\theta}(x))^2 f(x;s)\pi(s)dxds.$$

Note that $f(x;s)\pi(s)$ is a j.p.d.f. on $\mathbb{R}^n\times\Theta$ (since it is nonnegative and integrates to one) and let $\mathbb{E}$ denote the corresponding expectation. Let θ be a sample from π. Then

$$r_\pi(\hat{\theta}) = \mathbb{E}\big(\theta-\hat{\theta}(X)\big)^2,$$

which is recognized as the MSE of the prediction error of θ, from the observation of X. The minimal MSE is attained by the conditional expectation, which in this case is given by the Bayes formula:

$$\mathbb{E}(\theta|X) = \frac{\int_\Theta sf(X;s)\pi(s)ds}{\int_\Theta f(X;r)\pi(r)dr}.$$

This is nothing but (7c1), as $L(x;s)=f(x;s)$ by definition. □

The Bayes estimator with respect to risks, corresponding to other losses are computable similarly. Let $\theta\sim\pi$ and $X\sim\mathbb{P}_\theta$, then for a given loss function and any statistic $T(x)$, the Bayes risk is given by[15]

$$r_\ell(\pi,T) = \int_\Theta \mathbb{E}_\eta \ell\big(\eta,T(X)\big)\pi(\eta)d\eta = \int_\Theta\int_{\mathbb{R}^n} \ell\big(\eta,T(x)\big) f_{X|\theta}(x;\eta)\pi(\eta)dxd\eta =$$
$$\int_{\mathbb{R}^n}\bigg(\int_\Theta \ell(\eta,T(x))f_{\theta|X}(\eta;x)d\eta\bigg)f_X(x)dx \geq \int_{\mathbb{R}^n}\inf_{z\in\Theta}\bigg(\int_\Theta \ell(\eta,z)f_{\theta|X}(\eta;x)d\eta\bigg)f_X(x)dx.$$

If the infimum is actually attained for each $x\in\mathbb{R}^n$, which is typically the case[16], then the Bayes estimator is given by:

$$z^*(X)\in \operatorname{argmin}_{z\in\Theta}\int_\Theta \ell(\eta,z)f_{\theta|X}(\eta;X)d\eta.$$

REMARK 7c2. Note that for any loss function, $z^*(X)$ depends on the data X only through the *posterior* distribution $f_{\theta|X}$. The posterior distribution can be thought as an update of the prior distribution on the basis of the obtained observations. In other words, the data updates our prior belief about the parameter.

Let's see how this recovers the formula (7c1) for ℓ_2 loss:

$$\int_\Theta \ell(\eta,z)f_{\theta|X}(\eta;X)d\eta = \int_\Theta(\eta-z)^2 f_{\theta|X}(\eta;X)d\eta =$$
$$\int_\Theta \eta^2 f_{\theta|X}(\eta;X)d\eta - 2z\int_\Theta \eta f_{\theta|X}(\eta;X)d\eta + z^2.$$

[14]The discrete case is treated similarly

[15]we shall assume that all the p.d.f.'s involved exists; the discrete or even a more abstract case is treated along the same lines

[16]think of a simple condition for existence of the Bayes estimator

The latter is a parabola in z, which has a unique minimum at

$$z^*(X) := \int_\Theta \eta f_{\theta|X}(\eta; X)d\eta = \mathbb{E}(\theta|X),$$

yielding (7c1). For the ℓ_1 loss, we get:

$$\int_\Theta \ell(\eta, z) f_{\theta|X}(\eta; X)d\eta = \int_\Theta |\eta - z| f_{\theta|X}(\eta; X)d\eta.$$

If e.g. $\Theta = \mathbb{R}$, then

$$\int_\Theta |\eta - z| f_{\theta|X}(\eta; X)d\eta = \int_{-\infty}^z (z - \eta) f_{\theta|X}(\eta; X)d\eta + \int_z^\infty (\eta - z) f_{\theta|X}(\eta; X)d\eta.$$

The latter is a differentiable function in z, if the integrands are continuous, and the derivative is given by[17]:

$$\frac{d}{dz}\int_\Theta |\eta - z| f_{\theta|X}(\eta; X)d\eta = \int_{-\infty}^z f_{\theta|X}(\eta; X)d\eta - \int_z^\infty f_{\theta|X}(\eta; X)d\eta = \\ F_{\theta|X}(z; X) - \big(1 - F_{\theta|X}(z; X)\big) = 2F_{\theta|X}(z; X) - 1.$$

Equating the latter to 0 we find that the extremum is attained at $z^*(X)$ which is the solution of

$$F_{\theta|X}(z; X) = 1/2.$$

Differentiating one more time w.r.t. z, we find that the extremum is in fact a minimum. Hence $z^*(X)$ is the median of the posterior in this case.

Example 7c3. Suppose that we use a device, which is known to be quite precise in terms of the random errors, but may have a significant constant error, which we would like to estimate (and compensate from the device measurements in the future). More precisely, we perform n measurements and obtain an i.i.d. sample from $N(\theta, 1)$, where $\theta \in \mathbb{R}$ is the unknown parameter. Suppose we do not believe that the constant error term is too large, which we express by assuming that θ is itself a random variable sampled from $N(0, \sigma^2)$, where σ^2 is known to us and it expresses our belief regarding the dispersion of the error. Hence $X_i = \theta + Z_i$, where Z_i's are i.i.d. $N(0, 1)$ r.v.

In this setting, our prior is $\pi = N(0, \sigma^2)$ and hence $\theta, X_1, ..., X_n$ are jointly Gaussian. So the general formula for the Bayes estimator can be bypassed, using its particular form in the Gaussian case:

$$\mathbb{E}(\theta|X_1, ..., X_n) = \mathbb{E}\theta + \mathrm{Cov}(\theta, X)\mathrm{Cov}^{-1}(X, X)(X - \mathbb{E}X).$$

We have $\mathbb{E}\theta = 0$, $\mathbb{E}X = \big(0, \quad ... \quad 0\big)^\top$ and

$$\mathrm{cov}(\theta, X_i) = \sigma^2, \quad \mathrm{cov}(X_i, X_j) = \begin{cases} \sigma^2 + 1 & i = j \\ \sigma^2 & i \neq j \end{cases}$$

Let $\mathbf{1}$ denote the column vector of ones, then

$$\mathrm{Cov}(\theta, X) = \mathbb{E}\theta X^\top = \sigma^2 \mathbf{1}^\top$$

[17]Recall that $\frac{d}{dz}\phi(z, z) = \frac{\partial}{\partial x}\phi(x, z)_{x:=z} + \frac{\partial}{\partial y}\phi(z, y)_{y:=z}$

and

$$\mathrm{Cov}(X,X) = \mathbb{E}XX^\top = \sigma^2\mathbf{1}\mathbf{1}^\top + I,$$

where I is the n-by-n identity matrix. Let's check that[18] $\mathrm{Cov}^{-1}(X,X) = I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top$:

$$\Big(\sigma^2\mathbf{1}\mathbf{1}^\top + I\Big)\Big(I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top\Big) = \sigma^2\mathbf{1}\mathbf{1}^\top - \frac{\sigma^2}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top\mathbf{1}\mathbf{1}^\top + I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top =$$
$$\sigma^2\mathbf{1}\mathbf{1}^\top - \frac{\sigma^2 n}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top + I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top = \sigma^2\mathbf{1}\mathbf{1}^\top + I - \mathbf{1}\mathbf{1}^\top\Big(\frac{\sigma^2 n}{\sigma^{-2}+n} + \frac{1}{\sigma^{-2}+n}\Big) =$$
$$\sigma^2\mathbf{1}\mathbf{1}^\top + I - \sigma^2\mathbf{1}\mathbf{1}^\top = I,$$

where we used the fact $\mathbf{1}^\top\mathbf{1} = n$. Hence

$$\mathbb{E}(\theta|X) = \sigma^2\mathbf{1}^\top\Big(I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top\Big)X = \sigma^2\Big(\mathbf{1}^\top - \frac{n}{\sigma^{-2}+n}\mathbf{1}^\top\Big)X =$$
$$\sigma^2\Big(1 - \frac{n}{\sigma^{-2}+n}\Big)\mathbf{1}^\top X = \frac{1}{\sigma^{-2}+n}\sum_{i=1}^n X_i.$$

The estimator depends explicitly on σ^2: if σ is large, i.e. the uncertainty about θ is big, the estimator is close to the empirical mean $\bar{X}_n$ (i.e. the a priori belief is essentially ignored and the estimator relies on the data alone). If σ^2 is small, the prior is concentrated around zero and the estimator is close to zero, virtually regardless of the observations (for moderate n). If we continue to increase n, we shall again get back to the empirical mean. This behavior is intuitively appealing, since for large sample the empirical mean is close to the actual value of the parameter by virtue of the law of large numbers (regardless of our prior beliefs!).

■

Example 7c4. Consider the n i.i.d. tosses of a coin with unknown parameter θ. Suppose we tend to believe that θ should not be too far from $1/2$ and express our belief by assuming the prior density

$$\pi_r(\eta) = C_r\eta^r(1-\eta)^r, \quad \eta\in[0,1],$$

where C_r is the normalizing constant (whose value as we shall see won't play any role) and $r \geq 0$ is a parameter, which measures the strength of our faith: for large r, $\pi_r(\eta)$ is strongly concentrated around $1/2$ and for small r, it is close to the uniform distribution (which itself corresponds to $r=0$). The Bayes estimator under the quadratic loss is the posterior mean given by (7c1):

$$\hat{\theta}^*(X) = \frac{\int_0^1 \eta\eta^{S_n(X)}(1-\eta)^{n-S_n(X)}\pi_r(\eta)d\eta}{\int_0^1 \eta^{S_n(X)}(1-\eta)^{n-S_n(X)}\pi_r(\eta)d\eta} = \frac{\int_0^1 \eta^{S_n(X)+1+r}(1-\eta)^{n-S_n(X)+r}d\eta}{\int_0^1 \eta^{S_n(X)+r}(1-\eta)^{n-S_n(X)+r}d\eta} =$$
$$\frac{\Gamma(S_n(X)+2+r)/\Gamma(3+2r+n)}{\Gamma(S_n(X)+1+r)/\Gamma(2+2r+n)}.$$

[18] this is an application of the well known matrix Woodbury identity, known also as the matrix inversion lemma

For integer values[19] of r the latter reads:

$$\hat{\theta}^*(X) = \frac{(S_n(X)+1+r)!/(2+2r+n)!}{(S_n(X)+r)!/(1+2r+n)!} = \frac{S_n(X)+1+r}{2+2r+n}.$$

For large r (and small n), the Bayes estimator essentially ignores the data and yields $1/2$, "trusting" more the prior knowledge about θ. As n grows, the difference between the Bayes estimator and the MLE disappears, no matter how large r was: the prior knowledge is irrelevant when the data is abundant.

Note that for $r = 0$, the Bayes estimator $\hat{\theta}^*(X) = (S_n(X)+1)/(2+n)$ is different from the MLE $S_n(X)/n$, contrary to a tempting interpretation of the uniform prior as "non-informative". ■

Computation of the Bayes estimator as e.g. in (7c1) can be quite involved in general, depending on both prior and the likelihood. In this connection, the *conjugate priors* may be handy.

Definition 7c5. *A family of prior probability distributions π is said to be conjugate to a family of likelihood functions $L(x;\theta)$ if the resulting posterior distributions are in the same family as prior; the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood.*

Example 7c6. Gaussian priors are conjugate to the Gaussian likelihoods, corresponding to the i.i.d. sample of size n from $N(\theta,\sigma^2)$, $\theta \in \Theta$ with known σ^2. To check this we shall show that the posterior is Gaussian as well. To this end, note that a sample X from the model $\mathbb{P}_\theta$, corresponding to such likelihood can be written as

$$X_i = \theta + \sigma Z_i, \quad i = 1, ..., n$$

where Z_i's are i.i.d. $N(0,1)$ r.v.'s Let $\theta \sim \pi = N(\mu,\tau^2)$ (an arbitrary Gaussian density), then $\theta, X_1, ..., X_n$ are jointly Gaussian (i.e. form a Gaussian vector) and hence the conditional distribution of θ given X is Gaussian with the parameters, which can be computed explicitly as in the Example 7c3. ■

Here is an example, popular in computer science

Example 7c7. Let's show that Dirichlet distribution is conjugate to multinomial likelihood and find the corresponding parameters of the posterior. The p.m.f. of the multinomial distribution with parameters

$$\theta = (\theta_1, ..., \theta_k) \in \Theta = \mathcal{S}^{k-1} := \{u \in \mathbb{R}^k : u_i \geq 0, \sum_{i=1}^k u_i = 1\},$$

is given by

$$p(x;\theta) = L(x;\theta) = \frac{n!}{x_1!...x_k!}\theta_1^{x_1}...\theta_k^{x_k}, \quad x \in \mathbb{N}^k, \sum_{i=1}^k x_i = n.$$

[19] recall that $\Gamma(m) = (m-1)!$ for integer $m \geq 1$

Hence the conjugate prior should be a distribution on the simplex $\mathcal{S}^{k-1}$. The Dirichlet distribution $D(\alpha)$ is defined by the the p.d.f.

$$f(\eta;\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} \eta_i^{\alpha_i - 1}, \quad \eta \in \mathcal{S}^{k-1}$$

where $\alpha = (\alpha_1, ..., \alpha_k) \in \mathbb{R}_+^k$, the normalizing constant is given by

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\Big(\sum_{i=1}^{k} \alpha_i\Big)},$$

and where $\eta_k := 1 - \sum_{j=1}^{k-1} \eta_j$ is defined. Notice that the p.d.f. actually depends only on the first $k-1$ coordinates of η, i.e. it is a p.d.f. on $\mathbb{R}^{k-1}$ (and not on $\mathbb{R}^k$, think why).

To check the claim we have to verify that the posterior is Dirichlet and find its parameters. Let X be a sample from the above multinomial distribution with $\theta \sim \pi = D(\alpha)$. Then by an appropriate Bayes formula we have

$$f_{\theta|X}(\eta; x) \propto \frac{n!}{x_1!...x_k!} \eta_1^{x_1} ... \eta_k^{x_k} \prod_{i=1}^{k} \eta_i^{\alpha_i - 1} \propto \prod_{i=1}^{k} \eta_i^{(\alpha_i + x_i) - 1}$$

where the symbol $\propto$ stands for equality up to the multiplicative normalizing constant which depends only on x (and not on η). Hence the resulting posterior is identified as Dirichlet with parameter $\alpha + x$, where $x = (x_1, ..., x_k)$.

Let's see how these conjugates are typically used. Suppose we roll a dice n times independently and would like to estimate the probabilities to get each one of the six sides. Hence we obtain a realization of n i.i.d. r.v.'s $\xi_1, ..., \xi_n$ each taking value in $\{1, ..., 6\}$ with probabilities $\theta_1, ..., \theta_6$. Let X_i be the number of times the i-th side came up, i.e. $X_i = \sum_{m=1}^{n} I(\xi_m = i)$. As we saw, $X = (X_1, ..., X_6)$ is a sufficient statistic and hence nothing is lost if we regard X as our observation. If we put a prior Dirichlet distribution on Θ, then the emerging Bayes estimator (under the quadratic risk) has a very simple form.

A direct calculation shows that for $V \sim D(\alpha)$,

$$\mathbb{E}V_i = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_k}.$$

Since we *already know* that the posterior is $D(\alpha + X)$, we conclude that

$$\mathbb{E}(\theta_i | X) = \frac{\alpha_i + X_i}{\sum_{j=1}^{6} (\alpha_j + X_j)}, \quad i = 1, ..., 6.$$

Note that before obtaining any observations, our best guess of θ_i is just $\alpha_i / \sum_{j=1}^{6} \alpha_j$. As n grows, the estimator becomes more influenced by X and for large n becomes close to the usual empirical frequency estimator.

If we would have chosen a different prior, the Bayes estimator might be quite nasty, requiring complicated numerical calculations. ■

A list of conjugate priors and the corresponding likelihoods can be looked up in the literature. When dealing with a particular model $\mathbb{P}_\theta$, one may be lucky to find a family of conjugate priors,

in which case a computationally simple and usually very reasonable estimator [20] automatically emerges.

Some properties of the Bayes estimator. The Bayes estimator satisfies a number of interesting properties.

LEMMA 7c8. *The Bayes estimator $\hat{\theta}^*$ is admissible, if the corresponding prior density is positive and the risk of any estimator is continuous in θ.*

REMARK 7c9. The continuity assumption is quite weak: e.g. it is satisfied, if the loss function is continuous in both variables and $\theta \mapsto \mathbb{P}_\theta$ is continuous (in an appropriate sense).

PROOF. Let $\hat{\theta}^*$ be the Bayes estimator with respect to the loss function ℓ and the prior π. Suppose $\hat{\theta}^*$ is inadmissible and let $\hat{\theta}$ be an estimator such that

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \ell(\theta, \hat{\theta}) \leq \mathbb{E}_\theta \ell(\theta, \hat{\theta}^*) = R(\theta, \hat{\theta}^*), \quad \forall \theta \in \Theta,$$

where the inequality is strict for some $\theta' \in \Theta$. The continuity of risks imply that in fact the latter inequality is strict on an open neighborhood of θ' and hence

$$r_B(\hat{\theta}) = \int_\Theta R(\theta, \hat{\theta}) \pi(\theta) d\theta < \int_\Theta R(\theta, \hat{\theta}^*) \pi(\theta) d\theta = r_B(\hat{\theta}^*),$$

which contradicts optimality of the Bayes estimator. □

The following lemma establishes the important connection between the Bayes and the minimax estimators.

LEMMA 7c10. *A Bayes estimator with constant risk is minimax.*

PROOF. Let $\hat{\theta}^*$ be the Bayes estimator w.r.t. to the loss function ℓ and the prior π, such that

$$R(\theta, \hat{\theta}^*) = \mathbb{E}_\theta \ell(\theta, \hat{\theta}^*) \equiv C,$$

for some constant $C > 0$. Then for any estimator $\hat{\theta}$,

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}^*) = C = \int_\Theta R(\theta, \hat{\theta}^*) \pi(\theta) d\theta \leq \int_\Theta R(\theta, \hat{\theta}) \pi(\theta) d\theta \leq \sup_{\theta \in \Theta} R(\theta, \hat{\theta}),$$

where the inequality holds by the Bayes optimality of $\hat{\theta}^*$. □

The preceding lemma is only rarely useful, since the Bayes estimators with constant risks are hard to find. The following variation is more applicable:

LEMMA 7c11. *Let (π_i) be a sequence of priors and $\hat{\theta}^{*i}$ be the corresponding Bayes estimators. If the Bayes risks converge to a constant*

$$\lim_i \int_\Theta R(\theta, \hat{\theta}^{*i}) \pi_i(\theta) d\theta = C, \tag{7c2}$$

then for any estimator $\hat{\theta}$

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \geq C.$$

[20] in particular, when the number of the sample grows any Bayes estimator will typically forget its prior

PROOF. For any estimator $\hat{\theta}$,

$$\sup_{\theta\in\Theta} R(\theta,\hat{\theta}) \geq \int_\Theta R(\theta,\hat{\theta})\pi_i(\theta)d\theta \geq \int_\Theta R(\theta,\hat{\theta}^{*i})\pi_i(\theta)d\theta \xrightarrow{i\to\infty} C.$$

□

EXAMPLE 7c12. In Example 7c3 we saw that the Bayes estimator of θ from the sample $X_1,...,X_n$ with $X_i \sim N(\theta,1)$ and the Gaussian prior $N(0,\sigma^2)$ is given by

$$\hat{\theta}^{*\sigma} = \mathbb{E}(\theta|X) = \frac{1}{\sigma^{-2}+n}\sum_{i=1}^n X_i.$$

The corresponding Bayes risk is

$$r(\hat{\theta}^*) = \mathbb{E}(\theta-\hat{\theta}^*(X))^2 = \mathbb{E}\mathrm{var}(\theta|X) = \mathrm{var}(\theta) - \mathrm{cov}(\theta,X)\mathrm{cov}(X,X)^{-1}\mathrm{cov}(X,\theta) =$$
$$\sigma^2 - \sigma^2\mathbf{1}^\top\big(I - \frac{1}{\sigma^{-2}+n}\mathbf{1}\mathbf{1}^\top\big)\mathbf{1}\sigma^2 = \frac{\sigma^2/n}{1/n+\sigma^2}.$$

Since $\lim_{\sigma\to\infty} r(\hat{\theta}^*) = 1/n$, for any estimator $\hat{\theta}$,

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta(\theta-\hat{\theta})^2 \geq 1/n.$$

This bound is attained by the estimator $\bar{X}$, which is therefore minimax.

d. UMVU estimator

In this section we shall explore theory of unbiased estimators in more details. Let us start with a number of examples, recalling the definition of unbiased estimators.

EXAMPLE 7d1. Let $X_1,...,X_n$ be the i.i.d. sample from $N(\mu,\sigma^2)$, where $\theta = (\theta_1,\theta_2) = (\mu,\sigma^2) \in \Theta = \mathbb{R}\times\mathbb{R}_+$ is the unknown paremeter. We saw that the MLE of μ is given by the empirical mean:

$$\hat{\mu}_n(X) = \bar{X}_n.$$

Since $\mathbb{E}_\theta\hat{\mu}_n = \mathbb{E}_\theta\bar{X}_n = \mu$, the estimator $\hat{\mu}_n$ is unbiased. What about the MLE of σ^2:

$$\hat{\sigma}^2_n(X) = \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2 \quad ?$$

Recall that $Z_i := (X_i-\mu)/\sigma \sim N(0,1)$ and hence

$$\mathbb{E}_\theta\hat{\sigma}^2_n(X) = \mathbb{E}_\theta\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2 = \sigma^2\mathbb{E}_\theta\frac{1}{n}\sum_{i=1}^n (Z_i-\bar{Z})^2 =$$
$$\sigma^2\Big(\mathbb{E}_\theta\frac{1}{n}\sum_{i=1}^n Z_i^2 - \mathbb{E}_\theta(\bar{Z}_n)^2\Big) = \sigma^2\Big(1-1/n\Big).$$

Since $\mathbb{E}_\theta\hat{\sigma}^2_n(X) \neq \sigma^2$, it is a biased estimator of σ^2 with the bias

$$b(\sigma^2,\hat{\sigma}^2_n) = \sigma^2\Big(1-1/n\Big) - \sigma^2 = -\sigma^2/n,$$

which depends only on σ^2. A slight modification yields an unbiased estimator of σ^2:

$$\tilde{\sigma}_n^2(X) := \hat{\sigma}_n^2(X)/\Big(1 - 1/n\Big) = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

■

EXAMPLE 7d2. For an i.i.d. sample $X_1, ..., X_n$ from $U([0,\theta])$ with the unknown parameter $\theta \in \Theta = (0,\infty)$, the MLE of θ is $\hat{\theta}_n(X) = \max_i X_i$. By (7b3), $\mathbb{E}_\theta \hat{\theta}_n(X) = \theta\frac{n}{n+1}$ and hence $\hat{\theta}_n$ is biased:

$$b(\theta, \hat{\theta}_n) = \theta\frac{n}{n+1} - \theta = -\theta/(n+1).$$

Again, slightly modifying $\hat{\theta}_n$ we obtain an unbiased estimator

$$\tilde{\theta}_n(X) = \frac{n+1}{n}\hat{\theta}_n(X) = \frac{n+1}{n}\max_i X_i,$$

■

While unbiased estimators are often practically appealing, do not think they are always available! The unbiased estimator may not exist, as the following example demonstrates:

EXAMPLE 7d3. Let $X_1, ..., X_n$ be i.i.d. coin tosses with probability of heads $\theta \in (0,1)$. We want to estimate the odds, i.e. the ratio $q(\theta) = \theta/(1-\theta)$. Suppose T is an estimator of $q(\theta)$, then

$$\mathbb{E}_\theta T(X) = \sum_{x\in\{0,1\}^n} T(x)\theta^{S(x)}(1-\theta)^{n-S(x)},$$

where $S(x) = \sum_i x_i$. The latter is a polynomial of a *finite order* in θ. On the other hand, the function $q(\theta)$ can be written as the series:

$$q(\theta) = \frac{\theta}{1-\theta} = \theta\sum_{i=0}^\infty \theta^i, \quad \forall \theta \in (0,1),$$

where we used the formula for the sum of geometric progression. Since polynomials of different orders cannot be equal, we conclude that $\mathbb{E}_\theta T(X) \neq \theta/(1-\theta)$ for some $\theta \in \mathbb{R}_+$. But as T was an arbitrary statistic, we conclude that no unbiased estimator of $q(\theta)$ exists. ■

It is convenient to decompose the quadratic risk into the variance and bias terms:

$$R(\theta,\hat{\theta}) = \mathbb{E}_\theta\big(\theta - \mathbb{E}_\theta\hat{\theta}(X) + \mathbb{E}_\theta\hat{\theta}(X) - \hat{\theta}(X)\big)^2 = \\ \mathbb{E}_\theta\big(\theta - \mathbb{E}_\theta\hat{\theta}(X)\big)^2 + \mathbb{E}_\theta\big(\mathbb{E}_\theta\hat{\theta}(X) - \hat{\theta}(X)\big)^2 = \mathrm{var}_\theta(\hat{\theta}) + b^2(\theta,\hat{\theta}) \quad (7d1)$$

where we used the property

$$2\mathbb{E}_\theta\big(\theta - \mathbb{E}_\theta\hat{\theta}(X)\big)\big(\mathbb{E}_\theta\hat{\theta}(X) - \hat{\theta}(X)\big) = 2\big(\theta - \mathbb{E}_\theta\hat{\theta}(X)\big)\mathbb{E}_\theta\big(\mathbb{E}_\theta\hat{\theta}(X) - \hat{\theta}(X)\big) = \\ 2\big(\theta - \mathbb{E}_\theta\hat{\theta}(X)\big)\underbrace{\big(\mathbb{E}_\theta\hat{\theta}(X) - \mathbb{E}_\theta\hat{\theta}(X)\big)}_{=0} = 0.$$

In particular, the quadratic risk for unbiased estimators consists only of the variance term.

DEFINITION 7d4. *The Uniformly Minimal Variance Unbiased Estimator (UMVUE) $\hat{\theta}^*$ of θ is an unbiased estimator satisfying:*

$$R(\theta, \hat{\theta}^*) = \mathrm{var}_\theta(\hat{\theta}^*) \leq \mathrm{var}_\theta(\hat{\theta}) = R(\theta, \hat{\theta}), \quad \forall \theta \in \Theta,$$

for all unbiased estimators $\hat{\theta}$.

In words, the UMVUE is the optimal unbiased estimator.

REMARK 7d5. Note that the unbiased estimators may not be comparable. Moreover, the UMVUE may be inadmissible[21], i.e. its risk can be improvable by another estimator (which, of course, must be biased).

Of course, the UMVUE may not exist. For instance, unbiased estimators do not exist at all in the Example 7d3. Even if unbiased estimators do exist, it may be still impossible to choose UMVUE:

EXAMPLE 7d6. (Problem 19 page 130 from [**8**], solution by David Azriel)

Suppose X is a r.v. taking values $\{\theta-1, \theta, \theta+1\}$ with equal probabilities, where $\theta \in \Theta = \mathbb{Z}$ (signed integers) is the unknown parameter. We shall show that for any $j \in \mathbb{Z}$, there is an unbiased estimator $T_j(X)$, such that $\mathrm{var}_\theta(T_j) = 0$ for $\theta := j$. Suppose the UMVUE exists and denote it by T. Then $\mathrm{var}_\theta(T) \leq \mathrm{var}_\theta(T_j)$ for all $\theta \in \Theta$ and since j is arbitrary, it follows that $\mathrm{var}_\theta(T) = 0$ for all $\theta \in \Theta$, i.e., the parameter can be estimated precisely. Since this is impossible (prove!), the UMVUE does not exist. This is the same argument we used to show that there is no estimator, which minimizers the risk uniformly over all θ's.

To this end, note that $\mathbb{E}_\theta X = \theta$, i.e. X is an unbiased estimator of θ. For a fixed $j \in \mathbb{Z}$ and $x \in \{j-1, j, j+1\}$, define

$$\delta_j(x) = \begin{cases} -1 & x = j+1 \\ 0 & x = j \\ 1 & x = j-1 \end{cases}$$

and extend the definition to other $x \in \mathbb{Z}$ by periodicity, i.e. $\delta_j(x) := \delta_j(x+3)$ for all $x \in \mathbb{Z}$ (sketch the plot). Note that for any $\theta \in \mathbb{Z}$,

$$\mathbb{E}_\theta \delta_j(X) = \frac{1}{3}\Big(\delta_j(\theta-1) + \delta_j(\theta) + \delta_j(\theta+1)\Big) = 0,$$

since the average over any three neighboring values of $\delta_j(x)$ equals zero (look at your plot). Now set $T_j(X) = X + \delta_j(X)$. The estimator $T_j(X)$ is unbiased:

$$\mathbb{E}_\theta T_j(X) = \theta + \mathbb{E}_\theta \delta_j(X) = \theta.$$

Under distribution $\mathbb{P}_j$, X takes the values in $\{j-1, j, j+1\}$ and $T_j(X) \equiv j$. Hence $\mathrm{var}_j(T_j(X)) = 0$ as claimed.

Note that the crux of the proof is construction of a family of nontrivial unbiased estimators of zero $\delta_j(X)$, $j \in \mathbb{Z}$. This is precisely, what the notion of *completeness*, to be defined below, excludes. ■

[21] If $X_1, ..., X_n$ is a sample from $U([0,1])$, $T^*(X) = \frac{n+1}{n} \max_i X_i$ is UMVUE (as we shall see in the sequel). However, T^* is inadmissible: for example, the estimator $T'(X) = \frac{n+2}{n+1} \max_i X_i$ has better risk, i.e. $R(\theta, T') \leq R(\theta, T^*)$ for all $\theta \in \Theta$ (check!)

However, if UMVUE exists, it is essentially unique:

LEMMA 7d7. *If $T_1(X)$ and $T_2(X)$ are both UMVUEs, then $T_1(X) = T_2(X)$, $\mathbb{P}_\theta$-a.s*

PROOF. Suppose T_1 and T_2 are UMVUEs. Then

$$\mathrm{var}_\theta\Big(\frac{1}{2}T_1 + \frac{1}{2}T_2\Big) = \frac{1}{4}\mathrm{var}_\theta(T_1) + \frac{1}{2}\mathrm{cov}_\theta(T_1, T_2) + \frac{1}{4}\mathrm{var}_\theta(T_2) = \frac{1}{2}\mathrm{var}_\theta(T_1) + \frac{1}{2}\mathrm{cov}_\theta(T_1, T_2)$$

where the last equality follows from $\mathrm{var}_\theta(T_1) = \mathrm{var}_\theta(T_2)$ (as both T_1 and T_2 are UMVUE and thus in particular have the same variance). By the Cauchy–Schwarz inequality, $\mathrm{cov}_\theta(T_1, T_2)^2 \le \mathrm{var}_\theta(T_1)\mathrm{var}(T_2) = \mathrm{var}_\theta(T_1)^2$ and hence we obtain:

$$\mathrm{var}_\theta\Big(\frac{1}{2}T_1 + \frac{1}{2}T_2\Big) \le \mathrm{var}_\theta(T_1).$$

But $\mathbb{E}_\theta\Big(\frac{1}{2}T_1 + \frac{1}{2}T_2\Big) = \frac{1}{2}\theta + \frac{1}{2}\theta = \theta$, i.e. $\frac{1}{2}T_2 + \frac{1}{2}T_2$ is by itself an unbiased estimator and hence the latter inequality can hold only with equality (since T_1 is UMVUE):

$$\mathrm{var}_\theta\Big(\frac{1}{2}T_1 + \frac{1}{2}T_2\Big) = \mathrm{var}_\theta(T_1).$$

Again opening the brackets, this implies $\mathrm{cov}_\theta(T_1, T_2) = \mathrm{var}_\theta(T_1)$. But then

$$\mathrm{var}_\theta(T_1 - T_2) = \mathrm{var}(T_1) - 2\mathrm{cov}(T_1, T_2) + \mathrm{var}(T_2) = 0,$$

which means that $T_1 = T_2$, $\mathbb{P}_\theta$-a.s.

□

One of the main tools in analysis of the unbiased estimators is the Rao–Blackwell theorem, which states that conditioning on a sufficient statistic improves the MSE risk:

THEOREM 7d8 (Rao-Blackwell). *Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a statistical model, $X \sim \mathbb{P}_\theta$, $S(X)$ be a sufficient statistic and $T(X)$ an estimator of $q(\theta)$ for some known function q, with $\mathbb{E}_\theta T^2(X) < \infty$. Then the statistic*

$$T^*(X) := \mathbb{E}_\theta(T(X)|S(X))$$

is an estimator of $q(\theta)$ with the same bias as $T(X)$ and smaller MSE risk:

$$R(q(\theta), T^*) \le R(q(\theta), T), \quad \forall\theta \in \Theta, \tag{7d2}$$

where the inequality is strict, unless $T(X) = T^*(X)$ $\mathbb{P}_\theta$-a.s.

PROOF. Note that since $S(X)$ is sufficient, the conditional law of X given $S(X)$ does not depend on θ and in particular, $\mathbb{E}_\theta(T(X)|S(X))$ is only a function of X, i.e. a statistic indeed. Moreover,

$$\mathbb{E}_\theta T^*(X) = \mathbb{E}_\theta\mathbb{E}_\theta(T(X)|S(X)) = \mathbb{E}_\theta T(X),$$

i.e. T and T^* have the same bias function and, since MSE is a sum of the variance and the squared bias, (7d2) follows if we prove that $\mathrm{var}_\theta(T^*) \le \mathrm{var}_\theta(T)$:

$$\begin{aligned}
&\mathrm{var}_\theta(T) = \mathbb{E}_\theta\big(T(X) - \mathbb{E}_\theta T(X)\big)^2 = \mathbb{E}_\theta\big(T(X) - \mathbb{E}_\theta T^*(X)\big)^2 = \\
&\mathbb{E}_\theta\big(T(X) - T^*(X) + T^*(X) - \mathbb{E}_\theta T^*(X)\big)^2 = \\
&\mathbb{E}_\theta\big(T(X) - T^*(X)\big)^2 + 2\mathbb{E}_\theta\big(T^*(X) - \mathbb{E}_\theta T^*(X)\big)\big(T(X) - T^*(X)\big) + \\
&\mathbb{E}_\theta\big(T^*(X) - \mathbb{E}_\theta T^*(X)\big)^2.
\end{aligned}$$

Note that $T^*(X) = \mathbb{E}_\theta(T(X)|S(X))$ is a function of $S(X)$ and hence by orthogonality property

$$\mathbb{E}_\theta\big(T^*(X) - \mathbb{E}_\theta T^*(X)\big)\big(T(X) - T^*(X)\big) = \mathbb{E}_\theta \underbrace{\big(T^*(X) - \mathbb{E}_\theta T^*(X)\big)}_{\text{a function of } S(X)} \big(T(X) - \mathbb{E}_\theta(T(X)|S(X))\big) = 0.$$

The two equalities imply

$$\mathrm{var}_\theta(T) = \mathbb{E}_\theta\big(T(X) - T^*(X)\big)^2 + \mathrm{var}_\theta(T^*) \geq \mathrm{var}_\theta(T^*),$$

where the equality holds if and only if $T(X) = T^*(X)$, $\mathbb{P}_\theta$-a.s. □

Corollary 7d9. *For unbiased T, the estimator T^* obtained by the R-B procedure, is unbiased with improved risk.*

The R-B theorem suggests yet another way to construct estimators: come up with any unbiased estimator and improve it by applying the R-B procedure. Here is a typical application

Example 7d10. Let X_i be the number of customers, who come at a bank branch at the i-th day. Suppose we measure $X_1, ..., X_n$ and would like to estimate the probability of having no customers during a day. Let's accept the statistical model, which assumes that $X_1, ..., X_n$ are i.i.d. r.v.'s and $X_1 \sim \mathrm{Poi}(\theta)$, $\theta > 0$. We would like to estimate $e^{-\theta}$, which is the probability of having no clients under this statistical model.

An obvious[22] unbiased estimator of $e^{-\theta}$ is $T(X) = I(X_1 = 0)$: indeed, $\mathbb{E}_\theta T(X) = \mathbb{P}_\theta(X_1 = 0) = e^{-\theta}$. As we saw before, $S(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for this model. B-R theorem requires calculation of

$$T^*(X) = \mathbb{E}_\theta\big(T(X)|S(X)\big) = \mathbb{E}_\theta\big(I(X_1 = 0)|S(X)\big) = \mathbb{P}_\theta\big(X_1 = 0|S(X)\big).$$

To this end, for $k \in \mathbb{N}$,

$$\mathbb{P}_\theta\big(X_1 = 0|S(X) = k\big) = \frac{\mathbb{P}_\theta\big(X_1 = 0, S(X) = k\big)}{\mathbb{P}_\theta(S(X) = k)} = \frac{\mathbb{P}_\theta\Big(X_1 = 0, \sum_{j=2}^n X_j = k\Big)}{\mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big)} =$$

$$\frac{\mathbb{P}_\theta\big(X_1 = 0\big)\mathbb{P}_\theta\Big(\sum_{j=2}^n X_j = k\Big)}{\mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big)} = \frac{e^{-\theta}e^{-(n-1)\theta}\big((n-1)\theta\big)^k/k!}{e^{-n\theta}\big(n\theta\big)^k/k!} = \frac{(n-1)^k}{n^k} = \Big(1 - \frac{1}{n}\Big)^k,$$

where we used the fact that a sum of independent Poisson r.v.'s is Poisson. Hence the statistic

$$T^*(X) = \Big(1 - \frac{1}{n}\Big)^{S(X)}, \tag{7d3}$$

is an unbiased estimator of θ, whose risk improves the risk of the original estimator $T(X)$. Note that if n is large, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \approx \theta$ (by the law of large numbers) and [23]

$$T^*(X) = \Big(1 - \frac{1}{n}\Big)^{S(X)} = \Big(1 - \frac{1}{n}\Big)^{n\bar{X}_n} \approx e^{-\theta}.$$

[22] note that strictly speaking, $I(X_1 = 0)$ cannot be accepted as a point estimator of $e^{-\theta}$, since none of its values, i.e. $\{0, 1\}$, is in the range of $e^{-\theta}$, $\theta \in \mathbb{R}_+$. However, the claim of R-B theorem (and many other results above and below) do not depend on this assumption

[23] recall that for $\alpha_n := (1 - 1/n)^n$, $\log \alpha_n = n \log(1 - 1/n) \approx -1 + o(1/n)$, i.e. $\alpha_n \approx e^{-1}$

More precisely, $T^*(X)$ converges to $e^{-\theta}$ as $n \to \infty$ in an appropriate sense. In fact, we shall see below that $T^*(X)$ is UMVUE ! ■

REMARK 7d11. Applying the R-B procedure requires calculation of the conditional expectation, which is often a serious practical drawback. Moreover, such standard estimators as the MLE and the Bayes estimator cannot be actually improved using the R-B lemma, since both are functions of the minimal sufficient statistic (see the next paragraph).

Clearly, for any fixed estimator $T(X)$, the actual quality of the R-B improved estimator $T^*(X)$ will heavily depend on the choice of the sufficient statistic. For example, the whole sample X, which is a trivial sufficient statistic, obviously won't yield any improvement. On the other hand, if $S(X)$ is the minimal sufficient statistic, then $\mathbb{E}_\theta(T(X)|S(X))$ cannot be improved any further by no other sufficient statistic (why?). However, in general this will not give the UMVUE, since $\mathbb{E}_\theta(T'(X)|S(X))$ for a different estimator $T'(X)$, may still give a better estimator, even if $S(X)$ is the minimal sufficient. Can the R-B procedure produce the optimal (UMVU) estimator...? An answer to this question can be given in terms of the following notion of *completeness*:

DEFINITION 7d12. *Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a statistical model and $X \sim \mathbb{P}_\theta$. A statistic $T(X)$ is complete if*

$$\mathbb{E}_\theta g\big(T(X)\big) = 0,\ \forall\theta \in \Theta \quad \Longrightarrow \quad g\big(T(X)\big) = 0,\ \mathbb{P}_\theta - a.s.\ \forall\theta \in \Theta$$

REMARK 7d13. If $T(X)$ and $T'(X)$ are equivalent statistics and $T(X)$ is complete, so is $T'(X)$. Indeed, let g be a function such that $\mathbb{E}_\theta g(T'(X)) = 0$ for all $\theta \in \Theta$. Since $T'(X)$ and $T(X)$ are equivalent, there is a one-to-one function ψ such that $T'(X) = \psi(T(X))$. Let $\tilde{g}(u) := g(\psi(u))$, then

$$\mathbb{E}_\theta\tilde{g}(T(X)) = \mathbb{E}_\theta g\big(\psi(T(X)\big) = \mathbb{E}_\theta g(T'(X)) = 0.$$

Since T is complete, we conclude that $\tilde{g}(T(X)) = 0$ and thus $g(T'(X)) = 0$, $\mathbb{P}_\theta$-a.s. which means that $T'(X)$ is complete.

Let's see how this notion helps in finding the UMVUE.

LEMMA 7d14 (Lehmann-Scheffé). *If the sufficient statistic $S(X)$ is complete, then there is at most one unbiased estimator, coarser than $S(X)$.*

PROOF. Suppose that $T(X) = g(S(X))$ and $T'(X) = g'(S(X))$ are both unbiased, then

$$\mathbb{E}_\theta\big(g(S(X)) - g'(S(X))\big) = \mathbb{E}_\theta\big(T(X) - T'(X)\big) = 0,$$

and since $S(X)$ is complete, $g(S(X)) = g'(S(X))$, i.e. $T(X) = T'(X)$, $\mathbb{P}_\theta$-a.s. □

This implies that R-B procedure applied to any unbiased estimator, using a complete sufficient statistic, yields the unique and hence optimal unbiased estimator, i.e.

COROLLARY 7d15. *If $S(X)$ is the complete sufficient statistic, the R-B procedure yields the UMVUE.*

This suggests at least two ways of searching for the UMVUE:

(1) Rao-Blackwellize an unbiased estimator, using the complete sufficient statistic.

(2) Find an unbiased estimator, which is a function of the complete sufficient statistic.

As we shall shortly see below, the sufficient statistic $S(X)$ in Example 7d10 is complete and hence the estimator (7d3) is UMVUE. The second approach is demonstrated in the following example:

EXAMPLE 7d16. Suppose that $X_1, ..., X_n$ are i.i.d. $U([0,\theta])$ r.v. with the unknown parameter $\theta \in \mathbb{R}_+$. The minimal sufficient statistic $M = \max_i X_i$ has the density $f_M(x) = nx^{n-1}/\theta^n$, supported on $[0,\theta]$. Suppose that we know that M is complete and would like to find the UMVUE of $q(\theta)$, where q is a given differentiable function. Consider the estimator of the form $\psi(M)$, where ψ is a continuous function. If $\psi(M)$ is an unbiased estimator of $q(\theta)$

$$q(s) = \mathbb{E}_s\psi(M) = \frac{1}{s^n}\int_0^s nx^{n-1}\psi(x)dx, \quad s > 0.$$

Taking the derivative, we obtain

$$q'(s) = -ns^{-n-1}\int_0^s nx^{n-1}\psi(x)dx + ns^{-1}\psi(s) = -ns^{-1}q(s) + ns^{-1}\psi(s),$$

and

$$\psi(s) = \frac{1}{n}sq'(s) + q(s).$$

Hence taking $q(\theta) := \theta$, we obtain the UMVUE of θ:

$$\hat{\theta}(X) = \frac{1}{n}M + M = \frac{n+1}{n}\max_i X_i,$$

which we already encountered in Example 7d2. For $q(\theta) = \sin(\theta)$ we obtain the UMVUE

$$\widehat{q}(X) := \frac{1}{n}M\cos(M) + \sin(M).$$

■

The main question remains: how to establish completeness of a sufficient statistic? The first helpful observation is the following:

LEMMA 7d17. *A complete sufficient statistic is minimal.*

We have already seen (recall the discussion following the Example 7d10 on page 122), that if R-B is carried out with a sufficient statistic, which is not minimal, UMVUE cannot be obtained in general. In view of the L-S theorem, this implies that a complete sufficient statistic is necessarily minimal. Here is a direct proof:

PROOF. Let $T(X)$ be a complete sufficient statistic and $S(X)$ be the minimal sufficient statistic. Since $S(X)$ is sufficient, $\mathbb{E}_\theta(T(X)|S(X))$ does not depend on θ, i.e. $\mathbb{E}_\theta(T(X)|S(X)) = \phi(S(X))$ for some function ϕ. But as $S(X)$ is the minimal sufficient statistic, it is, by definition, coarser than any other sufficient statistic, and in particular, $S(X) = \psi(T(X))$ for some function

ψ. Hence $\mathbb{E}_\theta(T(X)|S(X)) = \phi\Big(\psi\big(T(X)\big)\Big)$. Let $g(u) := u - \phi(\psi(u))$, then

$$\mathbb{E}_\theta g(T(X)) = \mathbb{E}_\theta\Big(T(X) - \phi(\psi(T(X)))\Big) = \mathbb{E}_\theta\Big(T(X) - \phi(S(X))\Big) =$$
$$\mathbb{E}_\theta\Big(T(X) - \mathbb{E}_\theta(T(X)|S(X))\Big) = \mathbb{E}_\theta T(X) - \mathbb{E}_\theta\mathbb{E}_\theta(T(X)|S(X)) =$$
$$\mathbb{E}_\theta T(X) - \mathbb{E}_\theta T(X) = 0, \quad \forall\theta \in \Theta.$$

But $T(X)$ is complete and hence $g(T(X)) \equiv 0$ with probability one, i.e.

$$\mathbb{E}_\theta\big(T(X)|S(X)\big) = T(X). \tag{7d4}$$

By definition of the conditional expectation, $\mathbb{E}_\theta\big(T(X)|S(X)\big)$ is a function of $S(X)$ and hence (7d4) implies that $T(X)$ is coarser than $S(X)$. But on the other hand, $S(X)$ is minimal sufficient and hence is coarser than $T(X)$. Hence $S(X)$ and $T(X)$ are equivalent and thus $T(X)$ is minimal sufficient. □

Remark 7d18. A minimal sufficient statistic does not have to be complete and hence R-B conditioning on the minimal sufficient statistic is not enough to get the UMVUE (revisit Example 7d6).

The above lemma shows that the only candidate for a complete sufficient statistic is the minimal sufficient. How do we check that the minimal sufficient statistic is complete? Sometimes, this can be done directly by the definition, as the following examples demonstrate.

Example 6a3 (continued) We saw that for n independent tosses of a coin with probability of heads θ, $S(X) = \sum_{i=1}^n X_i$ is the minimal sufficient statistic. Let us check whether it is complete. Let g be a function such that $\mathbb{E}_\theta g(S(X)) = 0$, $\theta \in \Theta$. Recall that $S(X) \sim \mathrm{Bin}(n, \theta)$ and hence

$$\mathbb{E}_\theta g(S(X)) = \sum_{i=0}^n g(i)\binom{n}{i}\theta^i(1-\theta)^{n-i} = (1-\theta)^n\sum_{i=0}^n g(i)\binom{n}{i}\left(\frac{\theta}{1-\theta}\right)^i$$

Hence

$$\sum_{i=0}^n g(i)\binom{n}{i}\left(\frac{\theta}{1-\theta}\right)^i \equiv 0, \quad \forall\theta \in (0,1)$$

or, equivalently, for any $\theta/(1-\theta) \in \mathbb{R}_+$. But since the left hand side is a polynomial in $\theta/(1-\theta)$, we conclude that all its coefficients equal zero, i.e. $g(i) = 0$, and thus $g(S(X)) = 0$. Hence $S(X)$ is complete.

Now recall that $\bar{X}_n = S(X)/n$ is an unbiased estimator of θ. The R-B procedure does not change $\bar{X}_n$ and hence by completeness of $S(X)$, it is the UMVUE.

Let us see, e.g. that the trivial sufficient statistic $X = (X_1, ..., X_n)$ is not complete. To this end, we should exhibit a function g, such that $\mathbb{E}_\theta g(X) = 0$, but $g(X)$ is not identically zero (with probability one). A simple choice is $g(X) = X_1 - X_2$: clearly $g(X)$ is not identically zero

$$\mathbb{P}_\theta(g(X) \neq 0) = 1 - \mathbb{P}_\theta(X_1 = X_2) = 1 - \theta^2 - (1-\theta)^2$$

However $\mathbb{E}_\theta g(X) = \theta - \theta = 0$. Hence X is not complete. ■

EXAMPLE 7d19. Let $X_1, ..., X_n$ be a sample from $U([0,\theta])$, $\theta > 0$ distribution. As we saw, $M(X) = \max_{i=1,...,n} X_i$ is the sufficient statistic. To check completeness let g be a function such that $\mathbb{E}_\theta g(M) = 0$, for all $\theta \in \Theta$. Recall that M has the density (Example 7b3)

$$f_M(x) = \frac{n}{\theta^n} x^{n-1} I(x \in (0,\theta)),$$

and hence

$$\mathbb{E}_\theta g(M) = \int_0^\infty g(s) \frac{n}{\theta^n} s^{n-1} I(s \in (0,\theta)) ds = \frac{n}{\theta^n} \int_0^\theta g(s) s^{n-1} ds.$$

So $\mathbb{E}_\theta g(M) = 0$ reads

$$\int_0^\theta g(s) s^{n-1} ds = 0, \quad \forall \theta \in \Theta,$$

which implies that $g(s) = 0$ for almost all $s \in [0,1]$. Hence $g(M) = 0$, $\mathbb{P}_\theta$-a.s., which verifies completeness of M.

Recall that $(1 + 1/n)M(X)$ is an unbiased estimator of θ; since $M(X)$ is complete, it is in fact the UMVUE by the L-S theorem. ■

These examples indicate that checking completeness can be quite involved technically. Luckily, it can be established at a significant level of generality for the so called *exponential family* of distributions.

DEFINITION 7d20. *The probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$ on $\mathbb{R}^m$ belong to the k-parameter exponential family if the corresponding likelihood (i.e. either p.d.f. or p.m.f.) is of the form*

$$L(x;\theta) = \exp\Big\{ \sum_{i=1}^k c_i(\theta) T_i(x) + d(\theta) + S(x) \Big\} I(x \in A), \quad x \in \mathbb{R}^m \tag{7d5}$$

where c_i's and d are $\Theta \mapsto \mathbb{R}$ functions, T_i's and S are statistics taking values in $\mathbb{R}$ and $A \subseteq \mathbb{R}^m$ does not depend on θ.

REMARK 7d21. Clearly, if the probability distribution of X_1 belongs to an exponential family, the probability distribution of $X_1, ..., X_n$ also belongs to the same exponential family (check!).

REMARK 7d22. By the F-N factorization theorem, $T = (T_1(x), ..., T_k(x))$ is a sufficient statistic.

EXAMPLE 7d23. The Bernoulli distribution belongs to one parameter exponential family:

$$p(x;\theta) = \theta^x (1-\theta)^{1-x} I(x \in \{0,1\}) = \exp\Big\{ x \log\theta + (1-x)\log(1-\theta) \Big\} I(x \in \{0,1\}),$$

which is of the form (7d5) with $T_1(x) = x$, $c(\theta) = \log\frac{\theta}{1-\theta}$, $d(\theta) = \log(1-\theta)$, $S(x) = 0$ and $A = \{0,1\}$.

By the preceding remark, the distribution of an i.i.d. sample $X_1, ..., X_n$ from $\mathrm{Ber}(\theta)$, also belongs to one parameter exponential family. ■

EXAMPLE 7d24. Let $X_1, ..., X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ with unknown parameter $\theta = (\theta_1, \theta_1) = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. The right hand side of (6d5) tells that $\mathbb{P}_\theta$ belongs to 2-parameters exponential family with $c_1(\theta) = n\theta_1/\theta_2$, $c_2(\theta) = -1/2n1/\theta_2$ and $T_1(x) = \bar{x}_n$, $T_2(x) = \overline{x_n^2}$, $d(\theta) = -1/2n\theta_1^2/\theta_2 - n\log\theta_2$ and $S(x) = -\frac{n}{2}\log(2\pi)$ and $A = \mathbb{R}^n$. ■

EXAMPLE 7d25. Uniform distribution $U([0,\theta])$, $\theta \in \mathbb{R}_+$ does not belong to the exponential family, since its support depends on θ (no appropriate A can be identified). ■

The following fact, whose proof is beyond the scope of our course, is often handy:

THEOREM 7d26. *Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ belong to k-parameter exponential family of distributions with the likelihood function* (7d5). *The canonical statistic*

$$T(X) = \big(T_1(X), ..., T_k(X)\big)$$

is complete, if the interior [24] *of the range $\{c(\theta), \theta \in \Theta\}$ of*

$$c(\theta) = \big(c_1(\theta), ..., c_k(\theta)\big)$$

is not empty.

EXAMPLE 6a3 (continued) Let's see how this theorem is applied to deduce completeness of the statistic $S(X) = \sum_{i=1}^n X_i$ (which we have already checked directly). The likelihood in this case is the j.p.m.f.:

$$L(x;\theta) = \theta^{S(x)}(1-\theta)^{n-S(x)} = \exp\Big\{S(x)\log\frac{\theta}{1-\theta} + n\log(1-\theta)\Big\}, \quad x \in \{0,1\}^n, \theta \in [0,1].$$

Hence $L(x;\theta)$ belongs to the one parameter exponential family with $c(\theta) := \log\frac{\theta}{1-\theta}$. When θ increases from 0 to 1, the function $c(\theta)$ moves from $-\infty$ to $+\infty$, i.e. $\{c(\theta), \theta \in [0,1]\} = \mathbb{R}$, which trivially has a non-empty interior (in fact any open ball, i.e. interval in this case, is contained in $\mathbb{R}$). Hence the statistic $S(X)$ is complete. ■

EXAMPLE 7d24 (continued) The sample $X_1, ..., X_n$ from $N(\mu, \sigma^2)$ with the unknown parameter $\theta := (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, has the likelihood:

$$L(x;\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\Big\{-\frac{1}{2}\sum_{i=1}^n\frac{(x_i-\mu)^2}{\sigma^2}\Big\} =$$

$$\exp\Big\{-n/2\log(2\pi) - n/2\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{n}{2\sigma^2}\mu^2\Big\}, \quad x \in \mathbb{R}^n.$$

which belongs to the 2-parameter exponential family with:

$$T(x) = (T_1(x), T_2(x)) = \Big(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\Big)$$

and

$$c(\theta) = \big(c_1(\theta), c_2(\theta)\big) = \Big(\theta_1/\theta_2, -1/(2\theta_2)\Big).$$

---

[24] Consider the Euclidian space $\mathbb{R}^d$ with the natural distance metric, i.e. $\|x-y\| = \sqrt{\sum_{i=1}^d (x_i-y_i)^2}$. An open ball with radius $\rho > 0$ and center x, is defined as $B_\rho(x) = \{y \in \mathbb{R}^d : \|x-y\| < \rho\}$, i.e. all the points which are not further away from x than ρ. Let A be a set in $\mathbb{R}^d$. A point $x \in A$ is an internal point, if there exists an open ball with sufficiently small radius $\varepsilon > 0$, such that $B_\varepsilon(x) \subset A$. The interior of A, denoted by A° is the set of all internal points. A point x is a boundary point of A, if it is not internal and any open ball around x contains an internal point of A. The set of the boundary points is called the boundary of A and is denoted by ∂A. A set A is closed if it contains all its boundary points.

As (θ_1, θ_2) varies in $\mathbb{R} \times \mathbb{R}_+$, $c(\theta)$ takes values in $\mathbb{R} \times \mathbb{R}_-$, which obviously has a non-empty interior (why ?). Hence $T(X)$ is a complete statistic.

By Remark 7d13, the sufficient statistic $T'(X) := \left(\bar{X}_n, \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right)$, being equivalent to $T(X)$, is also complete. But we have already seen, that $T'(X)$ is the unbiased estimator of (μ, σ^2) and in view of completeness, the L-S theorem implies that $T'(X)$ is the UMVUE. ■

Now we can explain why the estimator (7b1) is better than the estimator (7b2) and in which sense. First of all, note that the estimator

$$\tilde{\sigma}_n(X) = \frac{1}{n}\sum_{i=1}^n |X_i - \bar{X}_n|$$

is not a function of the minimal sufficient statistic $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ and hence is inadmissible, being strictly improvable by means of the Rao-Blackwell lemma. The estimator

$$\hat{\sigma}_n(X) = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$

is the optimal one among the estimators with the same bias, since it is coarser than the minimal sufficient statistic, which, as we saw, is complete.

Further,

$$\hat{\sigma}_n(X) = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \sigma\sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2},$$

with $Z_i := (X_i - \mu)/\sigma$. Let $C_2(n) := \mathbb{E}_\theta \sqrt{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2}$ and note that it does not depend on θ, as Z_i's are i.i.d. $N(0,1)$ r.v.'s. Hence $\hat{\sigma}_n(X)/C_2(n)$ is an unbiased estimator of σ. Similarly $\tilde{\sigma}_n(X)/C_1(n)$, where $C_1(n) := \mathbb{E}_\theta \frac{1}{n}\sum_{i=1}^n |Z_i - \bar{Z}_n|$ (again independent of θ), is also an unbiased estimator of σ. However, $\hat{\sigma}_n(X)/C_2(n)$ is a function of the complete sufficient statistic and hence is UMVUE. Thus the normalized unbiased versions of $\hat{\sigma}_n(X)$ and $\tilde{\sigma}_n(X)$ are comparable and the former has better risk.

Example 7d10 (continued) Let's see that the unbiased estimator (7d3) is in fact UMVUE of $e^{-\theta}$. Since $T^*(X)$ is equivalent (why?) to $S(X)$, by Remark 7d13 and L-S theorem it is enough to check that $S(X)$ is complete. To this end, note the likelihood for this statistical model:

$$L(x;\theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \exp\left\{-\theta n + \log\theta S(x) - \sum_{i=1}^n \log(x_i!)\right\},$$

belongs to the one parameter exponential family with $c(\theta) = \log\theta$. The range of $c(\theta)$ over $\theta > 0$ is $\mathbb{R}$ and hence $S(X)$ is complete by the Lemma 7d26. ■

Finally, let us demonstrate that the "empty interior" part of the Lemma 7d26 cannot be in general omitted:

Example 7d27. Consider a sample $X_1, ..., X_n$ from $N(\theta, \theta^2)$, where $\theta \in \mathbb{R}_+$ is the unknown parameter. Repeating the calculations from the preceding example, we see that $L(x;\theta)$ still

belongs to the 2-parameter exponential family, and

$$c(\theta) = \Big(1/\theta, -1/(2\theta^2)\Big), \quad \theta \in \mathbb{R}_+.$$

The latter is a one dimensional curve in $\mathbb{R} \times \mathbb{R}_-$ and hence its interior is empty (sketch the plot of $c(\theta)$). Hence L-S theorem cannot be applied to check completeness of $T(X)$ (as defined in the preceding example).

In fact, the statistic is easily seen to be incomplete: take e.g.

$$g(T(X)) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \frac{n}{n+1}(\bar{X}_n)^2,$$

which is a function of $T(X) = \big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\big)$ after a rearrangement of terms. We have

$$\mathbb{E}_\theta g(T(X)) = \mathbb{E}_\theta \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \frac{n}{n+1}\mathbb{E}_\theta(\bar{X}_n)^2 =$$

$$\theta^2 - \frac{n}{n+1}\Big(\theta^2 + \frac{1}{n}\theta^2\Big) = 0, \quad \forall \theta \in \mathbb{R}_+.$$

However, it is obvious that $g\big(T(X)\big)$ is a non-degenerate random variable (e.g. its variance is nonzero). ■

The notions of completeness and sufficiency are related to ancillarity:

Definition 7d28. *A statistic T is ancillary if its probability distribution does not depend on $\theta \in \Theta$.*

While an ancillary statistic does not contain any information about the unknown parameter on its own, it nevertheless may be very much relevant for the purpose of inference in conjunction with other statistics. An ancillary statistic T' is an *ancillary complement* of an insufficient statistic T, if (T, T') is sufficient. For example, if $X_i = \theta + U_i$, where $U_i \sim U([0,1])$ are independent, the statistic $T'(X) = X_1 - X_2$ is an ancillary complement of the statistic $T(X) = X_1 + X_2$.

A statistic is called first order ancillary if its expectation is constant w.r.t. θ. Hence by definition, a statistic is complete if any coarser first order ancillary statistic is trivial. In particular, it is complete if any coarser ancillary statistic is trivial. A deeper relation to completeness is revealed by the following theorem

Theorem 7d29 (D. Basu). *A complete sufficient statistic and an ancillary statistic are independent.*

Proof. Let S be complete sufficient and T ancillary. Then for a fixed $x \in \mathbb{R}$, $\psi(x) := \mathbb{P}_\theta(T \le x)$ doesn't depend on θ. On the other hand, by sufficiency of S

$$\phi(x; S) := \mathbb{P}_\theta(T \le x | S)$$

also doesn't depend on θ. Moreover,

$$\mathbb{E}_\theta\big(\psi(x) - \phi(x;S)\big) = \mathbb{E}_\theta\Big(\mathbb{P}_\theta(T \le x) - \mathbb{P}_\theta(T \le x|S)\Big) = 0$$

and by completeness of S, it follows that $\phi(x;S)=\psi(x)$ or

$$\mathbb{P}_\theta(T\le x|S)=\mathbb{P}_\theta(T\le x),$$

and the claim follows by arbitrariness of $x\in\mathbb{R}$. □

The Basu theorem is typically used to establish independence of statistics. For examples, if X_i's are i.i.d. $N(\theta,1)$, then, as we saw, $\bar X$ is sufficient and complete and $\hat\sigma^2=\frac{1}{n}\sum_{i=1}^n(X_i-\bar X)^2$ is ancillary. Hence by the Basu theorem, they are independent. Similarly, e.g., $\bar X$ and $\max_i X_i-\min_i X_i$ are independent. Various applications of the Basu theorem are surveyed in [**5**].

e. The Cramer-Rao information bound

In many practical situations, the optimal (in appropriate sense) estimator may not exist or be too cumbersome to compute. Then typically a simpler ad-hoc estimator is used and it is desirable to assess how good it is. Remarkably, it turns out that the risk of *any* estimator can often be lower bounded by a quantity, depending only on the statistical model under consideration (and not a particular estimator at hand).

THEOREM 7e1 (Cramer-Rao information bound). *Consider the statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ with one dimensional parameter space $\Theta\subseteq\mathbb{R}$, given by the family of p.d.f.'s $f(x;\theta)$, $\theta\in\Theta$, $x\in\mathbb{R}^n$. Let $X\sim\mathbb{P}_\theta$ and $T(X)$ be an estimator of θ. Assume that $f(x;\theta)$ is differentiable in θ and that the following derivatives and integrations are interchangeable:*

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}f(x;\theta)dx=\int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}f(x;\theta)dx \tag{7e1}$$

and

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}T(x)f(x;\theta)dx=\int_{\mathbb{R}^n}T(x)\frac{\partial}{\partial\theta}f(x;\theta)dx. \tag{7e2}$$

Then

$$\mathrm{var}_\theta(T)\ge\frac{\big(\psi'(\theta)\big)^2}{I(\theta)}, \tag{7e3}$$

where $\psi(\theta):=\mathbb{E}_\theta T(X)$ and

$$I(\theta):=\mathbb{E}_\theta\bigg(\frac{\partial}{\partial\theta}\log f(X;\theta)\bigg)^2=\int_{\mathbb{R}^n}\bigg(\frac{\partial}{\partial\theta}\log f(x;\theta)\bigg)^2f(x;\theta)dx$$

is called the Fisher information, contained in the sample X.

PROOF. Since $\int_{\mathbb{R}^n}f(x;\theta)dx=1$, by (7e1)

$$\int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}f(x;\theta)dx=0. \tag{7e4}$$

Moreover, (7e2) reads:

$$\int_{\mathbb{R}^n}T(x)\frac{\partial}{\partial\theta}f(x;\theta)dx=\frac{\partial}{\partial\theta}\psi(\theta)=:\psi'(\theta). \tag{7e5}$$

Multiplying (7e4) by $\psi(\theta)$ and subtracting it from (7e5), we obtain:

$$\int_{\mathbb{R}^n}\big(T(x)-\psi(\theta)\big)\frac{\partial}{\partial\theta}f(x;\theta)dx=\psi'(\theta).$$

Hence, by the Cauchy - Schwarz inequality,

$$\big(\psi'(\theta)\big)^2 = \bigg(\int_{\mathbb{R}^n} \big(T(x)-\psi(\theta)\big)\frac{\partial}{\partial\theta}f(x;\theta)dx\bigg)^2 = \bigg(\int_{\mathbb{R}^n} \big(T(x)-\psi(\theta)\big)\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta)dx\bigg)^2 \leq$$

$$\bigg(\int_{\mathbb{R}^n} \big(T(x)-\psi(\theta)\big)^2 f(x;\theta)dx\bigg)\bigg(\int_{\mathbb{R}^n} \bigg(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\bigg)^2 f(x;\theta)dx\bigg) = \mathrm{var}_\theta(T)I(\theta), \tag{7e6}$$

which is the claimed inequality. □

Corollary 7e2. *Under the assumptions of Theorem 7e1, the MSE risk of an unbiased estimator $T(X)$ satisfies*

$$R(\theta,T) = \mathrm{var}_\theta(T) \geq \frac{1}{I(\theta)}. \tag{7e7}$$

Proof. Follows from (7e3) with $\psi(\theta)=\theta$. □

Remark 7e3. The Cramer-Rao bound is valid under similar assumptions for the discrete models, for which the p.d.f. in the definition of the Fisher information in replaced by p.m.f. For definiteness, we shall state all the results in the continuous setting hereafter, but all of them translate to the discrete setting as is.

The multivariate version of the C-R bound, i.e. when $\Theta \subseteq \mathbb{R}^d$, with $d>1$, is derived along the same lines: in this case, the Fisher information is a matrix and the inequality is understood as comparison of nonnegative definite matrices.

Remark 7e4. The estimator, whose risk attains the C-R lower bound, is called *efficient*. Hence the efficient unbiased estimator is the UMVUE. However, it is possible that UMVUE does not attain the C-R bound (see Example 7e10 below) and hence is not necessarily efficient. In fact, the Cauchy-Schwarz inequality in (7e6) saturates if and only if

$$\Big(T(x)-\psi(\theta)\Big)C(\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta)$$

for some function $C(\theta)$, independent of x. Integrating both parts, we see that equality is possible if $f(x;\theta)$ belongs to the exponential family with the canonical sufficient statistic $T(x)$ (the precise details can be found in [**16**]).

The assumptions (7e1) and (7e2) are not as innocent as they might seem at the first glance. Here is a simpler sufficient condition:

Lemma 7e5. *Assume that the support of $f(x;\theta)$ does not depend on θ and for some $\delta>0$,*

$$\int_{\mathbb{R}^n} |h(x)| \sup_{u\in[\theta-\delta,\theta+\delta]} \Big|\frac{\partial}{\partial\theta}f(x;u)\Big|dx < \infty, \quad \forall\theta\in\Theta \tag{7e8}$$

with both $h(x)\equiv 1$ and $h(x)=T(x)$. Then (7e1) *and* (7e2) *hold.*

The proof of this lemma relies on the classical results from real analysis, which are beyond the scope of this course. Note, for instance, that the model corresponding to $U([0,\theta])$, $\theta>0$ does not satisfy the above assumptions (as its support depends on θ).

Before working out a number of examples, let us prove some useful properties of the Fisher information.

LEMMA 7e6. *Let X and Y be independent r.v. with the Fisher informations $I_X(\theta)$ and $I_Y(\theta)$ respectively. Then the Fisher information contained in the vector (X,Y) is the sum of individual informations:*

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta).$$

PROOF. By independence, $f_{X,Y}(u,v;\theta) = f_X(u;\theta)f_Y(v;\theta)$ and

$$\frac{\partial}{\partial\theta}\log f_{X,Y}(u,v;\theta) = \frac{\partial}{\partial\theta}\log f_X(u;\theta) + \frac{\partial}{\partial\theta}\log f_Y(v;\theta).$$

Recall that $\mathbb{E}_\theta \frac{\partial}{\partial\theta}\log f_X(X;\theta) = 0$ and hence, again, using the independence,

$$I_{X,Y}(\theta) = \mathbb{E}_\theta\bigg(\frac{\partial}{\partial\theta}\log f_{X,Y}(X,Y;\theta)\bigg)^2 =$$
$$\mathbb{E}_\theta\bigg(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\bigg)^2 + 2\mathbb{E}_\theta\frac{\partial}{\partial\theta}\log f_X(X;\theta)\mathbb{E}_\theta\frac{\partial}{\partial\theta}\log f_Y(Y;\theta) +$$
$$\mathbb{E}_\theta\bigg(\frac{\partial}{\partial\theta}\log f_Y(Y;\theta)\bigg)^2 = I_X(\theta) + I_Y(\theta).$$

□

In particular,

COROLLARY 7e7. *Let $X = (X_1,...,X_n)$ be an i.i.d. sample from the p.d.f $f(u;\theta)$ with the Fisher information $I(\theta)$, then*

$$I_X(\theta) = nI(\theta).$$

LEMMA 7e8. *Assume that $f(x;\theta)$ is twice differentiable and*

$$\frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}^n} f(x;\theta)dx = \int_{\mathbb{R}^n}\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx, \tag{7e9}$$

then

$$I(\theta) = -\mathbb{E}_\theta\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = -\int_{\mathbb{R}^n} f(x;\theta)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)dx.$$

PROOF. Denote by $f'(x;\theta)$ and $f''(x;\theta)$ the first and second partial derivatives of $f(x;\theta)$ with respect to θ. Then

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\partial}{\partial\theta}\frac{f'(x;\theta)}{f(x;\theta)} = \frac{f''(x;\theta)f(x;\theta) - \big(f'(x;\theta)\big)^2}{f^2(x;\theta)}.$$

The claim holds if $\mathbb{E}_\theta\frac{f''(X;\theta)}{f(X;\theta)} = 0$. Indeed, by (7e9)

$$\mathbb{E}_\theta\frac{f''(X;\theta)}{f(X;\theta)} = \int_{\mathbb{R}^n} f''(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\underbrace{\int_{\mathbb{R}^n} f(x;\theta)dx}_{=1} = 0.$$

□

EXAMPLE 7e9. Let $X = (X_1, ..., X_n)$ be a sample from $N(\theta, \sigma^2)$, where σ^2 is known. Let's calculate the Fisher information for this model. By Corollary 7e7, $I_X(\theta) = nI(\theta)$, with

$$I(\theta) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\theta)^2/(2\sigma^2)} = \frac{1}{2\sigma^2} \mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} (X-\theta)^2 = 1/\sigma^2.$$

It is not hard to check the condition (7e8) in this case[25] and hence for an unbiased estimator $T(X)$

$$\mathrm{var}_\theta(T) \geq \frac{\sigma^2}{n}.$$

Since $\bar{X}_n$ is an unbiased estimator of θ and $\mathrm{var}_\theta(\bar{X}_n) = \sigma^2/n$, the risk of $\bar{X}_n$ attains the C-R bound, which confirms our previous conclusion that $\bar{X}_n$ is the UMVUE. ■

EXAMPLE 7e10. Let $X = (X_1, ..., X_n)$ be a sample from $\mathrm{Ber}(\theta)$, $\theta \in [0, 1]$. The likelihood for this model is the j.p.m.f. of X. The Fisher information of $\mathrm{Ber}(\theta)$ p.m.f. is

$$I(\theta) = \mathbb{E}_\theta \bigg(\frac{\partial}{\partial\theta} \log \Big(\theta^{X_1} (1-\theta)^{1-X_1} \Big) \bigg)^2 = \mathbb{E}_\theta \bigg(X_1 \frac{1}{\theta} - (1-X_1) \frac{1}{1-\theta} \bigg)^2 =$$
$$\frac{1}{\theta^2}\theta + \frac{1}{(1-\theta)^2}(1-\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

Since $\mathrm{Ber}(\theta)$ random variable takes a finite number of values, the conditions analogous to (7e1) and (7e2) (with integrals replaced by sums) obviously hold (for any T) and hence the risk of all unbiased estimators is lower bounded by $\frac{1}{n}\theta(1-\theta)$. But this is precisely the risk of the empirical mean $\bar{X}_n$, which is therefore UMVUE. ■

Here is a simple example, in which UMVUE does not attain the C-R bound:

EXAMPLE 7e11. Let $X \sim \mathrm{Poi}(\theta)$, $\theta > 0$. The Fisher information of $\mathrm{Poi}(\theta)$ is

$$I(\theta) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} \log \Big(e^{-\theta} \frac{\theta^X}{X!} \Big) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} \Big(-\theta + X \log\theta \Big) = \mathbb{E}_\theta \frac{X}{\theta}^2 = \frac{1}{\theta},$$

and the C-R bound for an unbiased estimator $T(X)$ of $e^{-\theta}$ is

$$\mathrm{var}_\theta(T) \geq \frac{\big(\psi'(\theta)\big)^2}{I(\theta)} = \theta e^{-2\theta}.$$

On the other hand, the statistic $\hat{\theta}(X) = I(X = 0)$ is an unbiased estimator of $e^{-\theta}$. Since the Poisson distribution belongs to the one parameter exponential family, by Lemma 7d26 and L-S theorem, $\hat{\theta}(X)$ is in fact the UMVUE of $e^{-\theta}$. The corresponding risk is readily found:

$$\mathrm{var}_\theta(\hat{\theta}) = \mathrm{var}_\theta\big(I(X=0)\big) = e^{-\theta}(1 - e^{-\theta}).$$

Hence the best attainable variance is strictly greater than the Cramer-Rao lower bound (the two curves are plotted on Figure 2). Note that this does not contradict Remark 7e4, since $T(X) = I(X = 0)$ is not the canonical sufficient statistic of $\mathrm{Poi}(\theta)$.

■

[25] in fact, checking (7e8) or other alternative conditions can be quite involved in general and we shall not dwell on this (important!) technicality. Note also that (7e8) involves $T(X)$ as well.

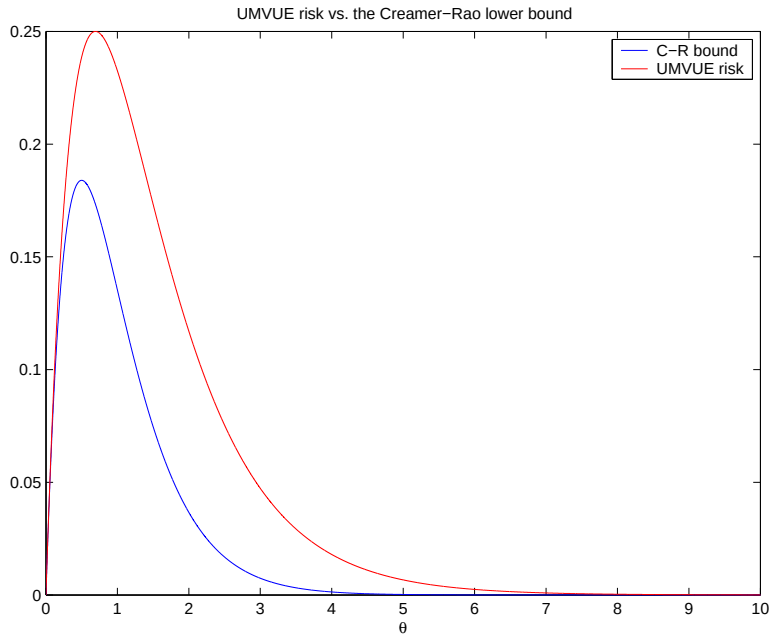

FIGURE 2. The risk of UMVUE versus the Cramer-Rao lower bound

The quantity $I(\theta)$ is the "information" contained in the sample: if $I(\theta)$ is small the C-R bound is large, which means that high precision estimation in this model is impossible. It turns out that the information, contained in the original sample, is preserved by sufficient statistic:

LEMMA 7e12. *Consider a statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$, given by a p.d.f. $f_X(x;\theta)$ and let $I_X(\theta)$ be the Fisher information contained in the sample $X \sim \mathbb{P}_\theta$. Let $T(X)$ be a statistic with the p.d.f. $f_T(t;\theta)$ and Fisher information $I_T(\theta)$. Then*

$$I_X(\theta) \geq I_T(\theta),$$

where the equality holds if $T(X)$ is sufficient.

PROOF. The proof hinges on the following key observation:

$$\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta)|T(X)\Big) = \frac{\partial}{\partial\theta}\log f_T(T(X);\theta). \tag{7e10}$$

Indeed, for an arbitrary test function[26] h:

$$\mathbb{E}_\theta \frac{\partial}{\partial\theta}\log f_X(X;\theta)h\big(T(X)\big) = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}\log f_X(x;\theta)h\big(T(x)\big)f_X(x;\theta)dx =$$
$$\int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} f_X(x;\theta)h\big(T(x)\big)dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} f_X(x;\theta)h\big(T(x)\big)dx =$$
$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta h\big(T(X)\big) = \frac{\partial}{\partial\theta}\int_{T(\mathbb{R}^n)} h(t)f_T(t;\theta)dt = \int_{T(\mathbb{R}^n)} h(t)\frac{\partial}{\partial\theta}f_T(t;\theta)dt =$$
$$\int_{T(\mathbb{R}^n)} h(t)\Big(\frac{\partial}{\partial\theta}\log f_T(t;\theta)\Big)f_T(t;\theta)dt = \mathbb{E}_\theta\frac{\partial}{\partial\theta}\log f_T(T(X);\theta)h\big(T(X)\big),$$

which, by the orthogonality property (3a3) of conditional expectation, yields (7e10). Further,

$$\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta) - \frac{\partial}{\partial\theta}\log f_T(T(X);\theta)\Big)^2 = \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\Big)^2 -$$
$$2\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\frac{\partial}{\partial\theta}\log f_T(T(X);\theta)\Big) + \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_T(T(X);\theta)\Big)^2 =$$
$$I_X(\theta) + I_T(\theta) - 2\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_T(T(X);\theta)\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\Big|T(X)\Big)\Big) =$$
$$I_X(\theta) + I_T(\theta) - 2\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_T(T(X);\theta)\Big)^2 = I_X(\theta) - I_T(\theta).$$

As the left hand side is nonnegative, we conclude that $I_X(\theta) \geq I_T(\theta)$.

If $T(X)$ is sufficient, then by the F-N factorization theorem $f_X(x;\theta) = g(\theta, T(x))h(x)$ for some g and h and hence

$$\mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log f_X(X;\theta)|T(X)\Big) = \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log g(\theta,T(X))|T(X)\Big) = \frac{\partial}{\partial\theta}\log g(\theta,T(X))$$

and hence by (7e10),

$$\frac{\partial}{\partial\theta}\log f_T(T(X);\theta) = \frac{\partial}{\partial\theta}\log g(\theta,T(X)) = \frac{\partial}{\partial\theta}\log\Big(g(\theta,T(X))h(X)\Big) = \frac{\partial}{\partial\theta}\log f_X(X;\theta).$$

This implies $I_X(\theta) = I_T(\theta)$, which completes the proof. □

When the unknown parameter is *location*, i.e. the model is given by the p.d.f. $f(x+\theta)$, $\theta \in \mathbb{R}$, the Fisher information is independent of θ:

$$I = \int_{\mathbb{R}}\Big(\frac{f'(x+\theta)}{f(x+\theta)}\Big)^2 f(x+\theta)dx = \int_{\mathbb{R}}\Big(\frac{f'(u)}{f(u)}\Big)^2 f(u)du.$$

In this case, the Fisher information satisfies various elegant inequalities, such as

LEMMA 7e13. *(A.J.Stam) If X and Y are independent random variables with finite Fisher informations I_X and I_Y, then*

$$\frac{1}{I_{X+Y}} \geq \frac{1}{I_X} + \frac{1}{I_Y},$$

where the equality is attained if and only if X and Y are Gaussian.

[26] we exchange derivative and integral fearlessly, assuming that the required conditions are satisfied

PROOF. (see e.g. [**7**]) □

f. Equivariant estimators

Let $X = (X_1, ..., X_n)$ be an i.i.d. sample from the density $f(x-\theta)$, where $\theta \in \mathbb{R}$ is the unknown *location* parameter. An estimator $\hat{\theta}(X)$ is equivariant if $\hat{\theta}(X+s) = s + \hat{\theta}(X)$ for any shift $s \in \mathbb{R}$.

DEFINITION 7f1. *The estimator, minimizing the quadratic risk among all estimators, equivariant with respect to location shifts, is called the Pitman estimator.*

PROPOSITION 7f2. *The Pitman estimator is*

$$\hat{\theta}(X) = \bar{X} - \mathbb{E}_0\big(\bar{X}|X-\bar{X}\big) = \frac{\int_{\mathbb{R}} u \prod_{i=1}^n f(X_j - u)du}{\int_{\mathbb{R}} \prod_{j=1}^n f(X_i - v)dv}. \tag{7f1}$$

PROOF. Suppose that $g(X)$ is an equivariant statistic, i.e. for all $x \in \mathbb{R}^n$ and $c \in \mathbb{R}$

$$g(x+c) - g(x) = c, \tag{7f2}$$

which in particular implies,

$$g(x) = \bar{x} + g(x - \bar{x}), \quad \forall x \in \mathbb{R}^d. \tag{7f3}$$

Conversely, any g, solving this functional equation, satisfies (7f2). Let S be the subset of functions of the form $\bar{x} + \phi(x-\bar{x})$, $x \in \mathbb{R}^n$ for some real valued $\phi(\cdot)$. Clearly any solution of the above equation belongs to S. Conversely, a direct check shows that any function in S is a solution of (7f3). Hence all the functions satisfying (7f3) (and thus also (7f2)) are of the form

$$g(x) = \bar{x} + \phi(x - \bar{x}).$$

For such statistics

$$R(\theta, g(X)) = \mathbb{E}_\theta\big(\theta - \bar{X} - \phi(X-\bar{X})\big)^2 = \mathbb{E}_0\big(\bar{X} + \phi(X-\bar{X})\big)^2 \geq \mathbb{E}_0\Big(\bar{X} - \mathbb{E}_0\big(\bar{X}|X-\bar{X}\big)\Big)^2,$$

where the equality is attained with $\varphi(x) = -\mathbb{E}_0\big(\bar{X}|X - \bar{X} = x - \bar{x}\big)$, which verifies the first equality in (7f1).

Further, note that $X - \bar{X}$ and $(Z_1, ..., Z_{n-1}) := (X_1 - X_n, ..., X_{n-1} - X_n)$ are equivalent statistics: $Z_i = X_i - \bar{X} - (X_n - \bar{X})$, $i = 1, ..., n-1$ and, conversely, $X_i - \bar{X} = \Big(1 - \frac{1}{n}\Big)Z_i - \frac{1}{n}\sum_{k \neq i} Z_k$, $i = 1, ..., n-1$. Hence

$$\bar{X} - \mathbb{E}\big(\bar{X}|X-\bar{X}\big) = X_n - (X_n - \bar{X}) - \mathbb{E}\big(\bar{X}|X-\bar{X}\big) = \\ X_n - \mathbb{E}\big(\bar{X} + X_n - \bar{X}|X-\bar{X}\big) = X_n - \mathbb{E}\big(X_n|X_1 - X_n, ..., X_{n-1} - X_n\big).$$

The vector $V = (X_1 - X_n, ..., X_{n-1} - X_n, X_n)$ has the p.d.f (under $\mathbb{P}_0$):

$$f_V(v_1, ..., v_n) = f(v_n) \prod_{i=1}^{n-1} f(v_i + v_n)$$

and hence

$$\mathbb{E}\big(X_n|X_1-X_n,...,X_{n-1}-X_n\big)=\frac{\int_{\mathbb{R}} vf(v)\prod_{i=1}^{n-1}f(X_i-X_n+v)dv}{\int_{\mathbb{R}} f(u)\prod_{i=1}^{n-1}f(X_i-X_n+u)du}=$$
$$\frac{\int_{\mathbb{R}}(X_n-v)f(X_n-v)\prod_{i=1}^{n-1}f(X_i-v)dv}{\int_{\mathbb{R}} f(X_n-u)\prod_{i=1}^{n-1}f(X_i-u)du}=X_n-\frac{\int_{\mathbb{R}} v\prod_{i=1}^{n}f(X_i-v)dv}{\int_{\mathbb{R}}\prod_{i=1}^{n}f(X_i-u)du},$$

which verifies the second equality in (7f1).

□

EXAMPLE 7f3. For $X_1\sim N(\theta,1)$, $\bar X$ is independent of $X-\bar X$ and hence the Pitman estimator is $\bar X$. For $X_1\sim U([\theta,\theta+1])$, the Pitman estimator is $(\max_j X_j+\min_j X_j-1)/2$ and for $X_1\sim\exp(1)$, the Pitman estimator is $\min_j X_j-1/n$. ■

g. Asymptotic theory of estimation

In many practical situations the amount of the available data can be as abundant as we wish and it is reasonable to require that the estimator produces more precise guesses of the parameter as the number of observations grows. The main object studied by the asymptotic theory of estimation is a *sequence* of estimators and its goal is to compare different sequences of estimators in the asymptotic regime, i.e. when the number of observations[27] tends to infinity.

For example, we already mentioned several reasons for preferring the estimator (7b1):

$$\hat\sigma_n(X)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar X_n)^2}.$$

over the estimator (7b2)

$$\tilde\sigma_n(X)=\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar X_n|.$$

In particular, after appropriate normalization (7b1) becomes the optimal unbiased estimator. However, even the optimal estimator may perform quite poorly if we don't have enough data (i.e. n is small, say 10 or so). Hence, the natural question is whether $\hat\sigma_n(X)$ is getting closer to the actual value of σ, when n is large. The intuition, based on the law of large numbers, tells us that this indeed will be the case. But, perhaps $\tilde\sigma_n(X)$ is as good as $\hat\sigma_n(X)$, when n is large ? How do we compare the two in the asymptotic regime ?

These and many more questions are in the heart of the asymptotic theory and it turns out that the answers are often surprising and counterintuitive at the first glance. In this section we shall introduce the basic notions and explore some simple examples (which will hopefully induce appetite for a deeper dive).

Let's start with a simulation: a coin[28] was tossed 500 times and $\bar X_n$ is calculated for $n=1,2,...,500$. This was repeated three times independently (think of them as three different realizations of $\bar X_n$), and Figure 3 depicts the values of $\bar X_n$ as a function of n, in the three experiments (of 500 tosses each). The actual value of the heads probability is $\theta_0=2/3$.

[27] other asymptotic regimes are possible: small noise asymptotic, etc.

[28] well, a digital coin ;)

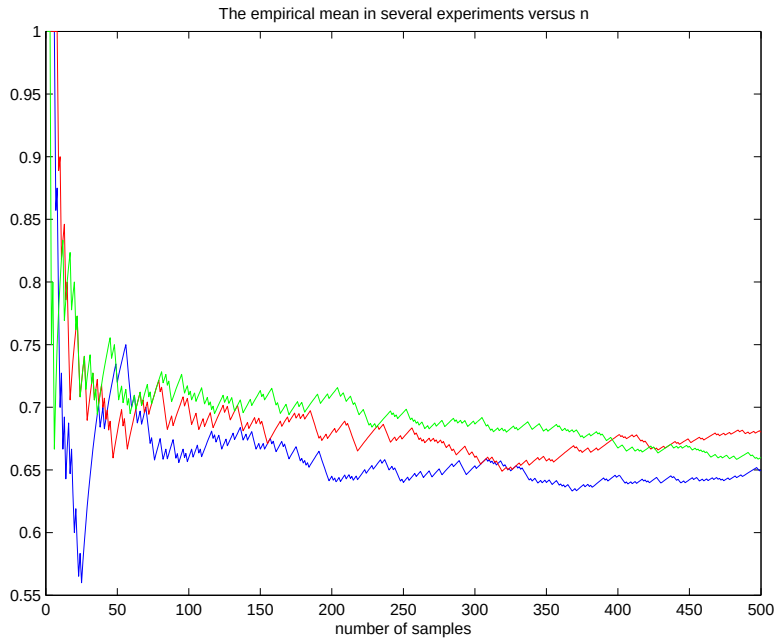

FIGURE 3. The empirical mean $\bar{X}_n$, calculated in three experiments, is plotted versus the sample size n. The actual value of the parameter (probability of heads) is $\theta_0 = 2/3 = 0.666...$

Observe that for small n, the estimator $\bar{X}_n$ is quite unreliable: say at $n = 50$, the red experiment produced an estimate, reasonably close to the actual value of the parameter $\theta_0 = 2/3$; however, this appears to be purely 'good luck', as the other two experiments generated very bad estimates. On the other hand, all three experiments generated relatively good estimates at $n = 500$ and it seems that increasing n would have improved the estimates furthermore. Hence on the intuitive level we feel that for large values of n all three experiments will produce arbitrarily precise estimates.

The picture is hardly surprising: after all $\bar{X}_n$ is a random variable and in different experiments we obtain its different realizations. Note that $\mathrm{var}_\theta(\bar{X}_n) = \frac{1}{n}\theta(1-\theta)$, which decreases with n: for small n's, $\bar{X}_n$ has a large variance and hence we obtained realizations, which are quite dispersed; for larger n, its variance decreases and hence the obtained realizations were close to each other.

Recall that $\mathbb{E}_\theta \bar{X}_n = \theta$ for all θ and all n, and hence as n increases, the distribution[29] of the random variable $\bar{X}_n$ concentrates around the true value of the parameter. In particular,

$$\lim_{n\to\infty} \mathrm{var}_\theta(\bar{X}_n) = 0, \quad \forall \theta \in \Theta = [0,1], \tag{7g1}$$

which means that the sequences of estimators $\bar{X}_n$, viewed as random variables, *converges* to the true value of θ.

[29] which in this case is just the normalized $\mathrm{Bin}(n,\theta)$, but this is not really important

Note however, that the convergence in this probabilistic setting is very different from the convergence of deterministic sequences of numbers. Recall that a sequence of numbers a_n converges to a number a, called the limit of $(a_n)_{n\geq 1}$ and denoted $a := \lim_{n\to\infty} a_n$, if for any $\varepsilon > 0$, there exists an integer $N(\varepsilon)$, possibly dependent on ε, so that $|a_n - a| \leq \varepsilon$ for all $n \geq N(\varepsilon)$. That is, the entries of the sequence a_n, starting from some $N(\varepsilon)$ are all in a small (2ε-wide) interval around a.

Hence, for an arbitrarily small $\varepsilon > 0$, (7g1) implies that there is an integer $N(\varepsilon)$, such that $\mathrm{var}_\theta(\bar{X}_n)$ (which is a number) is not further away from zero than ε. However, this does not mean that $\bar{X}_n$ itself must be within the distance of ε from θ: in fact, sometimes (i.e. for some realizations) it won't! After all, a random variable with small variance is not prohibited from generating large values[30]. Hence the convergence of random variables to a limit, either random or deterministic, indicates that large deviations from the limits are improbable (though possible!) for large n's.

On the technical level, the difference between convergence of sequences of numbers and random variables is even deeper. Convergence of sequences of real numbers, or more generally of real vectors in $\mathbb{R}^d$, does not depend on the norm we choose to measure the distance between the vectors. For example, if we measure the distance between two vectors $x, y \in \mathbb{R}^d$, by either $d_2(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ or $d_\infty(x,y) = \max_{i=1,\dots,d} |x_i - y_i|$, a sequence $x_n \in \mathbb{R}^d$ either converges or diverges in the two metrics *simultaneously*. One can show that for *finite dimensional* linear spaces any two norms are equivalent in this sense.

The situation is radically different for sequences with entries in the infinite dimensional spaces, such as e.g. functions. Since random variables are functions on Ω, many nonequivalent modes of convergence emerge, some of which to be discussed below.

Convergence of sequences of random variables. How do we define convergence for a sequence of random variables $(\xi_n)_{n\geq 1}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$? To tackle this question, it will be convenient to think of a realizations of $(\xi_n)_{n\geq 1}$ as deterministic sequences. Then a natural obvious way to defined convergence is to require that all the *realizations* of $(\xi_n)_{n\geq 1}$ converge to a limit. This limit itself may and will in general depend on the particular realization, i.e. will be a random variable on its own:

Definition 7g1. *A sequence (ξ_n) converges pointwise to a random variable ξ if*

$$\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega), \quad \forall \omega \in \Omega.$$

It turns out, however, that this definition is too strong to be satisfied in a majority of cases. Here is a simple demonstration:

Example 7g2. Consider a sequence of i.i.d. random variables $(X_n)_{n\geq 1}$ with $X_n \sim N(0,1)$ and let $Y_n = X_n/n$. Does the sequence $(Y_n)_{n\geq 1}$ converge to a limit in the latter sense ? Obviously not: for example, the sequence $(1, -2, 3, +4, ...)$ is a legitimate realization of $(X_n)_{n\geq 1}$ and the corresponding realization of $(Y_n)_{n\geq 1}$ is the sequence $(1, -1, 1, -1, ...)$, which does not converge. Thus $(Y_n)_{n\geq 1}$ does not converge pointwise. On the other hand, it is intuitively clear that X_n/n will typically be very small for large n and hence in a certain weaker sense must converge to 0. ■

[30] e.g. $N(0, 0.001)$ can generate the value in $[1000 : 1001]$ with small, but *non-zero* probability, which means that we may observe this in an experiment

DEFINITION 7g3. *The sequence of r.v.'s* $(\xi_n)_{n\geq 1}$ *converges*[31] *in* L^p, $p \geq 1$ *to a random variable* ξ, *if* $\mathbb{E}|\xi|^p < \infty$ *and* $\mathbb{E}|\xi|_n^p < \infty$ *for all* n *and*

$$\lim_{n\to\infty} \mathbb{E}\big|\xi_n - \xi\big|^p = 0.$$

EXAMPLE 7g2 (continued) Let's check that the sequence $(Y_n)_{n\geq 1}$ does converge to $Y \equiv 0$ (viewed as a constant random variable) in L_2. Indeed,

$$\mathbb{E}(Y_n - Y)^2 = \mathbb{E}(X_n/n - 0)^2 = \frac{1}{n^2}\mathbb{E}X_n^2 = 1/n^2 \xrightarrow{n\to\infty} 0,$$

and hence, by definition $Y_n \xrightarrow[n\to\infty]{L_2} 0$.

But what if X_n has the Cauchy distribution, rather than the Gaussian one ...? In this case, $\mathbb{E}(Y_n - 0)^2 = \mathbb{E}(X_n/n)^2 = \infty$ for any $n \geq 1$ and hence the convergence in L_2 is lost. This means that the L_2 convergence is too strong to capture the intuition that X_n/n still goes to zero as $n \to \infty$: after all, X_n is sampled from a probability density which is centered at zero and hence we expect that for large n, X_n/n will be typically small.

■

DEFINITION 7g4. *The sequence* $(\xi_n)_{n\geq 1}$ *converges in probability to a random variable* ξ, *if for any* $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\big(|\xi_n - \xi| \geq \varepsilon\big) = 0.$$

EXAMPLE 7g2 (continued)

Let $(X_n)_{n\geq 1}$ be i.i.d. standard Cauchy random variables and again define the sequence of random variables $Y_n := X_n/n$. The sequence Y_n converges in probability to zero. Indeed, for an $\varepsilon > 0$,

$$\mathbb{P}\big(|Y_n - 0| \geq \varepsilon\big) = \mathbb{P}(|X_n| \geq \varepsilon n) = \mathbb{P}(|X_1| \geq \varepsilon n) = 1 - F_X(\varepsilon n) + F_X(-\varepsilon n).$$

Since $\varepsilon > 0$, $\varepsilon n \to \infty$ as $n \to \infty$ and $\lim_{n\to\infty} F_X(\varepsilon n) = 1$ and

$$\lim_{n\to\infty} F_X(-\varepsilon n) = \lim_{n\to-\infty} F_X(\varepsilon n) = 0,$$

which implies:

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n| \geq \varepsilon\big) = 0,$$

as claimed.

■

This example shows that convergence in probability is generally weaker than in L_2, i.e. a sequence may converge in probability, but not in L_2. The converse, however, is impossible:

LEMMA 7g5. *Convergence in* L_2 *implies convergence in probability.*

[31] i.e. (ξ_n) converges as a sequence in the space of functions (random variables) with finite p-norm

PROOF. Suppose that $(\xi_n)_{n\geq 1}$ converges to ξ in L_2, then by the Chebyshev inequality[32] for any $\varepsilon > 0$

$$\mathbb{P}(|\xi_n - \xi| > \varepsilon) \leq \varepsilon^{-2}\mathbb{E}(\xi_n - \xi)^2 \xrightarrow{n\to\infty} 0,$$

which verifies the convergence in probability. □

Of course, a sequence may not converge at all:

EXAMPLE 7g6. Let ξ_n a sequence of i.i.d. r.v.'s with the common p.d.f. f. Suppose that $\xi_n \to \xi$ in probability. Then for any $\varepsilon > 0$

$$\mathbb{P}\big(|\xi_n - \xi_{n+1}| \geq \varepsilon\big) = \mathbb{P}\big(|\xi_n - \xi + \xi - \xi_{n+1}| \geq \varepsilon\big) \leq$$
$$\mathbb{P}\big(|\xi_n - \xi| \geq \varepsilon/2\big) + \mathbb{P}\big(|\xi - \xi_{n+1}| \geq \varepsilon/2\big) \xrightarrow{n\to\infty} 0.$$

On the other hand,

$$\mathbb{P}\big(|\xi_n - \xi_{n+1}| \geq \varepsilon\big) = \int_{\mathbb{R}^2} I(|s-t| \geq \varepsilon) f(s)f(t)dsdt = 1 - \int_{\mathbb{R}^2} I(|s-t| < \varepsilon) f(s)f(t)dsdt,$$

which is arbitrarily close to 1 for small $\varepsilon > 0$. The obtained contradiction shows that ξ_n does not converge in probability. ■

Here are some useful facts about the convergence in probability:

LEMMA 7g7. *If $\xi_n \xrightarrow{\mathbb{P}} \xi$ and g is a continuous function, then $g(\xi_n) \xrightarrow{\mathbb{P}} g(\xi)$ in probability.*

PROOF. Recall that $g : \mathbb{R} \mapsto \mathbb{R}$ is continuous if for any $x \in \mathbb{R}$ and any $\varepsilon > 0$, there is a $\delta_x > 0$, such that $|x - y| \leq \delta_x$ implies $|g(x) - g(y)| \leq \varepsilon$. Note that δ_x may depend on x (and on ε). However, for any $x \in [-C, C]$, for any $\varepsilon > 0$, one can choose[33] a $\delta_C > 0$, independent of x (!) and such that $x, y \in [-C, C]$ and $|x - y| \leq \delta$ imply $|g(x) - g(y)| \leq \varepsilon$. Hence[34]

$$\mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon\big) = \mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon, |\xi_n| \vee |\xi| \leq C\big) +$$
$$\mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon, |\xi_n| \vee |\xi| > C\big). \quad (7g2)$$

Let δ_C be as mentioned before, then

$$\mathbb{P}\Big(|\xi_n - \xi| \leq \delta_C, |\xi_n| \vee |\xi| \leq C\Big) \leq \mathbb{P}\big(|g(\xi_n) - g(\xi)| \leq \varepsilon, |\xi_n| \vee |\xi| \leq C\big)$$

and hence for any $C > 0$,

$$\mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon, |\xi_n| \vee |\xi| \leq C\big) \leq$$
$$\mathbb{P}\Big(|\xi_n - \xi| \geq \delta_C, |\xi_n| \vee |\xi| \leq C\Big) \leq \mathbb{P}\Big(|\xi_n - \xi| \geq \delta_C\Big) \xrightarrow{n\to\infty} 0, \quad (7g3)$$

---

[32]The Chebyshev inequality states that for a nonnegative r.v. η and a positive constant a, $\mathbb{P}(\eta > a) \leq \mathbb{E}\eta^p/a^p$ for any $p > 0$ (if $\mathbb{E}\eta^p = \infty$, then the inequality is trivial). Proof: suppose $\mathbb{E}\eta^p < \infty$, then

$$\mathbb{E}\eta^p = \mathbb{E}\eta^p I(\eta > a) + \underbrace{\mathbb{E}\eta^p I(\eta \leq a)}_{\geq 0} \geq \mathbb{E}\eta^p I(\eta > a) \geq a^p I(\eta > a) = a^p \mathbb{P}(\eta > a),$$

which is the claimed inequality after rearrangement. □

[33]**Theorem:** *continuous functions are uniformly continuous on compacts.*

[34]recall the notation $a \vee b = \max(a, b)$

where the convergence holds, since $\xi_n \to \xi$ in probability. On the other hand,

$$\mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon, |\xi_n| \vee |\xi| > C\big) \leq \mathbb{P}\big(|\xi_n| \vee |\xi| > C\big) = \mathbb{P}\big(\{|\xi_n| > C\} \cup \{|\xi| > C\}\big) \leq \mathbb{P}\big(|\xi_n| > C\big) + \mathbb{P}\big(|\xi| > C\big). \quad (7g4)$$

Moreover, since $|\xi_n| \leq |\xi_n - \xi| + |\xi|$, it follows $\{|\xi_n| > C\} \subseteq \{|\xi_n - \xi| > C/2\} \cup \{|\xi| > C/2\}$ and, consequently,

$$\mathbb{P}\big(|\xi_n| > C\big) \leq \mathbb{P}(|\xi_n - \xi| > C/2) + \mathbb{P}(|\xi| > C/2). \quad (7g5)$$

Since $\xi_n \to \xi$ in probability, $\lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| > C/2) = 0$ and hence, combining (7g4) and (7g5), we get:

$$\overline{\lim_{n\to\infty}}\, \mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon, |\xi_n| \vee |\xi| \leq C\big) \leq \mathbb{P}(|\xi| > C/2) + \mathbb{P}\big(|\xi| > C\big) \leq 2\mathbb{P}(|\xi| > C/2). \quad (7g6)$$

Substitution of (7g3) and (7g6) into (7g2) gives:

$$\overline{\lim_{n\to\infty}}\, \mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon\big) \leq 2\mathbb{P}(|\xi| > C/2).$$

Taking $C \to \infty$ yields the claim, i.e. $\lim_n \mathbb{P}\big(|g(\xi_n) - g(\xi)| \geq \varepsilon\big) = 0$.

□

LEMMA 7g8. *If $\xi_n \xrightarrow{\mathbb{P}} \xi$ and $\eta_n \xrightarrow{\mathbb{P}} \eta$, then*

(1) $\xi_n + \eta_n \xrightarrow{\mathbb{P}} \xi + \eta$
(2) $c\xi_n \xrightarrow{\mathbb{P}} c\xi$ *for any* $c \in \mathbb{R}$
(3) $\xi_n\eta_n \xrightarrow{\mathbb{P}} \xi\eta$

PROOF. To check (1), let $\varepsilon > 0$, then

$$\mathbb{P}(|\xi_n + \eta_n - \xi - \eta| \geq \varepsilon) \leq \mathbb{P}(|\xi_n - \xi| \geq \varepsilon/2) + \mathbb{P}(|\eta_n - \eta| \geq \varepsilon/2) \xrightarrow{n\to\infty} 0.$$

Other claims are proved similarly. □

REMARK 7g9. Be warned: various facts, familiar from convergence of real sequences, may or may not hold for various types of convergence of random variables. For example, if $\xi_n \to \xi$ in L_2, then, depending on the function g, $g(\xi_n)$ may not converge to $g(\xi)$ in L_2, even if g is continuous (if e.g. $\xi_n = X_n/n$ with i.i.d. $N(0,1)$ sequence $(X_n)_{n\geq 1}$, then for $g(x) = e^{x^4}$, $\mathbb{E}g^2(\xi_n) = \infty$ and hence the convergence in L_2 fails.)

A completely different kind of convergence is the convergence in distribution (or *weak* convergence[35]):

DEFINITION 7g10. *A sequence of random variables $(\xi_n)_{n\geq 1}$ converges in distribution to a random variable ξ, if*

$$\lim_{n\to\infty} \mathbb{P}(\xi_n \leq x) = \mathbb{P}(\xi \leq x),$$

[35]the term "weak" comes from analysis

for all $x \in \mathbb{R}$, at which $F_\xi(x) = \mathbb{P}(\xi \le x)$ is continuous[36].

REMARK 7g11. Note that the convergence in distribution is convergence of the c.d.f.'s of ξ_n's, rather than ξ_n's themselves[37].

EXAMPLE 7g12. Let $(\xi_n)_{n\ge1}$ be an i.i.d sequence. As we have seen, (ξ_n) does not converge in probability. However, (ξ_n) trivially converges in distribution:

$$F_{\xi_n}(x) \equiv F_{\xi_1}(x), \quad \forall x \in \mathbb{R}.$$

for all $n \ge 1$. ■

Before considering much more convincing examples of the weak convergence, let's explore some of its useful properties.

LEMMA 7g13. *If $\xi_n \to \xi$ weakly and α_n is a sequence of numbers, $\alpha_n \to \infty$, then $\xi_n/\alpha_n \xrightarrow{\mathbb{P}} 0$.*

PROOF. Let $F_n(x)$ and $F(x)$ be the c.d.f.'s of ξ_n and ξ respectively, then (w.l.o.g. $\alpha_n > 0$ is assumed)

$$\mathbb{P}(|\xi_n/\alpha_n| \ge \varepsilon) = \mathbb{P}(\xi_n \ge \varepsilon\alpha_n) + \mathbb{P}(\xi_n \le -\varepsilon\alpha_n).$$

Let $c > 0$ be a continuity point of F and let n be large enough so that $-\varepsilon\alpha_n < -c$, then

$$\mathbb{P}(\xi_n \le -\varepsilon\alpha_n) = F_n(-\varepsilon\alpha_n) \le F_n(-c) \le \big|F_n(-c) - F(-c)\big| + F(-c) \xrightarrow{n\to\infty} F(-c).$$

Since F can have only countable number of discontinuities, c can be chosen to make the right hand side arbitrarily small. Hence $\lim_n \mathbb{P}(\xi_n \le -\varepsilon\alpha_n) = 0$. Similarly, $\lim_n \mathbb{P}(\xi_n \ge \varepsilon\alpha_n) = 0$ is shown and the claim follows. □

The following lemma, whose proof we shall omit, give an alternative characterization of the weak convergence

THEOREM 7g14. *[part of the Portmanteau theorem] The following are equivalent*

(i) $\xi_n \xrightarrow{d} \xi$
(ii) $\mathbb{E}g(\xi_n) \to \mathbb{E}g(\xi)$ *for any bounded continuous function g*
(iii) $\mathbb{E}e^{it\xi_n} \to \mathbb{E}e^{it\xi}$ *for all*[38] $t \in \mathbb{R}$

---

[36] the convergence is allowed to fail at the points at which the target c.d.f. F is discontinuous, simply because requiring otherwise may yield an unreasonably strong notion of convergence. For example, by the law of large numbers $\bar{X}_n := \frac{1}{n}\sum_{j=1}^n X_i$ converges in probability to $X \equiv 0$ if X_i's are i.i.d. $N(0,1)$ r.v.'s. However, the c.d.f. of $\bar{X}_n$ fails to converge to the (degenerate) c.d.f of 0 at $x = 0$:

$$\mathbb{P}(\bar{X}_n \le 0) = \frac{1}{2} \ne 1 = \mathbb{P}(X \le 0).$$

It would be embarrassing not to say that $\bar{X}_n \to 0$ is distribution, if $\bar{X}_n \to 0$ in probability, as the latter should be expected to be stronger. Hence by definition, the convergence in distribution excludes convergence at the discontinuity point 0 from consideration.

[37] ξ_n may be even defined on different probability spaces

[38] $\varphi_\xi(t) := \mathbb{E}e^{it\xi}$, where i is the imaginary unit, is called the characteristic function of the r.v. ξ. It resembles the definition of the moment generating function with t replaced by it. This seemingly minor detail is in fact major, since the characteristic function is always defined (unlike the m.g.f. which requires existence of all moments at least). As the m.g.f. function, the characteristic function also determins the distribution (being it's Fourier transform). If you're not familiar with complex analysis, just replace characteristic function in any appearance in the text with m.g.f., and pay the price, losing the generality.

This theorem has a number of useful corollaries.

COROLLARY 7g15. *If $\xi_n \xrightarrow{d} \xi$, then $g(\xi_n) \xrightarrow{d} g(\xi)$ for any continuous function.*

PROOF. By the preceding lemma, we have to show that $\mathbb{E}\phi(g(\xi_n)) \to \mathbb{E}\phi(g(\xi))$ for any bounded continuous function ϕ. This holds, since $x \mapsto \phi(g(x))$ is continuous and bounded and $\xi_n \xrightarrow{d} \xi$. □

Example 7g12 shows that (ξ_n) which converges weakly, may not converge in probability. The converse, however, is true:

LEMMA 7g16. *$\xi_n \xrightarrow{\mathbb{P}} \xi$ implies $\xi_n \xrightarrow{d} \xi$.*

PROOF. Let ϕ be a bounded continuous function, then by Lemma 7g7, $\phi(\xi_n) \to \phi(\xi)$ is probability and since ϕ is bounded [39] $\mathbb{E}\phi(\xi_n) \to \mathbb{E}\phi(\xi)$, which by Theorem 7g14 implies $\xi_n \xrightarrow{d} \xi$. □

LEMMA 7g17 (Slutsky's theorem). *If $\xi_n \xrightarrow{\mathbb{P}} c$ for a constant $c \in \mathbb{R}$ and $\eta_n \xrightarrow{d} \eta$, then* [40] *$(\xi_n, \eta_n) \xrightarrow{d} (c, \eta_n)$ and in particular:*

(1) $\xi_n + \eta_n \xrightarrow{d} c + \eta$
(2) $\xi_n \eta_n \xrightarrow{d} c\eta$
(3) $\eta_n / \xi_n \xrightarrow{d} \eta / c$, *if* $c \neq 0$

PROOF. To prove $(\xi_n, \eta_n) \xrightarrow{d} (c, \eta_n)$ we shall check that[41] for arbitrary bounded continuous functions $\psi : \mathbb{R} \mapsto \mathbb{R}$ and $\phi : \mathbb{R} \mapsto \mathbb{R}$

$$\mathbb{E}\psi(\xi_n)\phi(\eta_n) \xrightarrow{n\to\infty} \psi(c)\mathbb{E}\phi(\eta).$$

To this end,

$$\begin{aligned}
&\big|\mathbb{E}\psi(\xi_n)\phi(\eta_n) - \psi(c)\mathbb{E}\phi(\eta)\big| \leq \\
&\big|\mathbb{E}\psi(\xi_n)\phi(\eta_n) - \psi(c)\mathbb{E}\phi(\eta_n)\big| + \big|\psi(c)\mathbb{E}\phi(\eta_n) - \psi(c)\mathbb{E}\phi(\eta)\big| = \\
&\mathbb{E}|\phi(\eta_n)|\big|\psi(\xi_n) - \psi(c)\big| + \psi(c)\big|\mathbb{E}\phi(\eta_n) - \mathbb{E}\phi(\eta)\big| \xrightarrow{n\to\infty} 0,
\end{aligned}$$

where the first term on the right hand side converges to zero by Lemma 7g7 as $\xi_n \xrightarrow{\mathbb{P}} c$ (think why) and the second term converges to zero by Theorem 7g14. Now the claims (1)-(3) follow, since $\xi_n \xrightarrow{\mathbb{P}} c$ implies $\xi_n \xrightarrow{d} c$ (by Lemma 7g16) and $(x, y) \mapsto x + y$, $(x, y) \mapsto xy$ are continuous functions on $\mathbb{R}^2$ and $(x, y) \mapsto x/y$ is continuous on $\mathbb{R}^2 \setminus \{(x, y) : y \neq 0\}$. □

[39] $\xi_n \to \xi$ in probability does not necessarily implies $\mathbb{E}\xi_n \to \mathbb{E}\xi$. For example, this fails if $\xi_n = X/n$, where X is Cauchy r.v. However, Lebesgue's Dominated Convergence theorem states that if there exists a r.v. ζ with $\mathbb{E}\zeta < \infty$, such that $|\xi_n| \leq \zeta$ for all n and $\xi_n \xrightarrow{\mathbb{P}} \xi$, then $\xi_n \xrightarrow{\mathbb{L}_1} \xi$ and $\mathbb{E}\xi_n \to \mathbb{E}\xi < \infty$. In particular, the latter holds if $|\xi_n| \leq M$ for some constant $M > 0$.

[40] convergence in distribution of random vectors is defined similarly, as convergence of c.d.f's (think, how the discontinuity sets may look like)

[41] you may suspect that taking bounded continuous functions of the 'product' form $h(s, t) = \psi(s)\phi(t)$ is not enough and we shall consider the whole class of bounded continuous functions $h : \mathbb{R}^2 \mapsto \mathbb{R}$. In fact, the former is sufficient, but of course requires a justification (which we omit).

REMARK 7g18. Beware: the claims of the preceding lemma may fail, if ξ_n converges in probability to a non-degenerate r.v., rather than a constant (can you give an example ?)

Here is another useful fact:

LEMMA 7g19. *Suppose that ξ_n's are defined on the same probability space and $\xi_n \xrightarrow{d} c$, where c is a constant. Then $\xi_n \xrightarrow{\mathbb{P}} c$.*

PROOF. Since c is constant, we can consider $c = 0$ without loss of generality (think why). Let F_n be the c.d.f.'s of ξ_n, then for $\varepsilon > 0$

$$\mathbb{P}(|\xi_n| \geq \varepsilon) = F_n(-\varepsilon) + 1 - F_n(\varepsilon) \xrightarrow{n\to\infty} 0,$$

since the $F_n(x) \to F(x) = \mathbf{1}_{\{x\geq 0\}}$ for all $x \in \mathbb{R} \setminus \{0\}$. □

Limit theorems. One of the classical *limit theorems* in probability is the Law of Large Numbers (LLN), which can be stated with different types of convergence under appropriate assumptions. Here is a particularly primitive version:

THEOREM 7g20 (an LLN). *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. r.v. with $\mathbb{E}|X_1|^2 < \infty$. Then the sequence $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $n \geq 1$ converges to $\mu := \mathbb{E}X_1$ in L_2 (and hence also in probability).*

PROOF. By the i.i.d. property:

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n X_i - \mu\Big)^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(X_i - \mu)^2 = \frac{\mathrm{var}(X_1)}{n} \xrightarrow{n\to\infty} 0,$$

which means that $\frac{1}{n}\sum_{i=1}^n X_i$ converges to μ in L_2. □

The latter, perhaps, is the most naive version of the LLN: it makes a number of strong assumptions, among which the most restrictive is boundedness of the second moment. The more classical version of LLN is

THEOREM 7g21 (The weak[42] LLN). *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. r.v. with $\mathbb{E}|X_1| < \infty$. Then the sequence $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $n \geq 1$ converges to $\mu := \mathbb{E}X_1$ in probability.*

PROOF. Let us show that $\bar{X}_n \xrightarrow{d} \mu$, which implies the claim by Lemma 7g19. Since $\mathbb{E}|X_1| < \infty$, the characteristic function $\psi(t) = \mathbb{E}e^{itX_1}$ is continuously differentiable (check). Hence by Taylor's theorem

$$\mathbb{E}e^{it\bar{X}_n} = \psi^n(t/n) = \Big(\psi(0) + \frac{t}{n}\psi'(\tilde{t}_n)\Big)^n = \Big(1 + \frac{t}{n}\psi'(0) + \frac{t}{n}\big(\psi'(\tilde{t}_n) - \psi'(0)\big)\Big)^n = \Big(1 + \frac{t}{n}i\mu + r_n\Big)^n$$

where $|\tilde{t}_n| \leq t$ and

$$|nr_n| = |t|\big|\psi'(\tilde{t}_n) - \psi'(0)\big| \xrightarrow{n\to\infty} 0,$$

[42] in fact under the assumptions of the weak LLN, one can check that the empirical mean converges to $\mathbb{E}X_1$ in a much stronger sense, namely with probability one. This stronger result is called the strong LLN

by continuity of ψ'. Hence

$$\mathbb{E}e^{it\bar{X}_n} \xrightarrow{n\to\infty} e^{it\mu}$$

and the claim follows by Theorem 7g14. □

Do not think, however, that the empirical mean always converges:

EXAMPLE 7g22. Let $X_1, ..., X_n$ be i.i.d. standard Cauchy r.v.'s. Since $\mathbb{E}|X_1| = \infty$, the expectation of X_1 doesn't exist and the LLN doesn't apply. This does not necessarily excludes convergence of $\bar{X}_n$ to some limit. However, it can be shown[43] that $\bar{X}_n$ also has standard Cauchy distribution. Hence $\bar{X}_n$ at least does not converge to a constant in any sense (in fact, it doesn't converge to a random variable either, think why) ■

The weak LLN states that the empirical mean $\bar{X}_n - \mu$ converges to zero. Can we quantify the speed of convergence? If $\bar{X}_n - \mu$ converges to zero, then there might be a sequence α_n increasing in n, such that $\alpha_n(\bar{X}_n - \mu)$ ceases to converge to zero. For example, let's try $\alpha_n = n^{1/3}$:

$$\operatorname{var}\Big(n^{1/3}(\bar{X}_n - \mu)\Big) = n^{2/3}\frac{1}{n}\operatorname{var}(X_1) \xrightarrow{n\to\infty} 0,$$

i.e. $\bar{X}_n - \mu$ converges to zero, faster than $n^{1/3}$. The above hints that $n^{1/2}$ might be just the right rate:

$$\operatorname{var}\Big(n^{1/2}(\bar{X}_n - \mu)\Big) = n\frac{1}{n}\operatorname{var}(X_1) = \operatorname{var}(X_1),$$

which means that $n^{1/2}(\bar{X}_n - \mu)$ does not converge to zero in L_2 any more (and, on the other hand, has a bounded variance uniformly in n). Remarkably, much more can be claimed:

THEOREM 7g23 (Central Limit Theorem). *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. r.v.'s with $\mu := \mathbb{E}X_1$ and $\sigma^2 := \operatorname{var}(X_1) < \infty$. Then $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma}$ converges weakly to a standard Gaussian random variable, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq x\Big) = \Phi(x), \quad \forall x \in \mathbb{R}, \tag{7g7}$$

where Φ is the c.d.f. of $N(0,1)$.

PROOF. The r.v.'s $Y_i := (X_i - \mu)/\sigma$ are i.i.d. with zero mean and unit variance and (7g7) is equivalent to

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \leq x\Big) = \Phi(x), \quad \forall x \in \mathbb{R}. \tag{7g8}$$

Let $\psi(t) = \mathbb{E}e^{itY_1}$ and $\psi_n(t) = \mathbb{E}\exp\Big\{it\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\Big\}$ By the i.i.d. property

$$\psi_n(t) = \psi^n(t/\sqrt{n}), \quad t \in \mathbb{R}.$$

Since $\mathbb{E}Y_1^2 < \infty$, the characteristic function $\psi(t)$ is twice continuously differentiable (think why) and hence by the Taylor theorem

$$\psi_n(t) = \Big(\psi(0) + \frac{t}{\sqrt{n}}\psi'(0) + \frac{1}{2}\frac{t^2}{n}\psi''(\tilde{t}_n)\Big)^n = \Big(\psi(0) - \frac{t^2}{2}\frac{1}{n} + \frac{t^2}{2}\frac{1}{n}\big(\psi''(\tilde{t}_n) - \psi''(0)\big)\Big)^n$$

[43] the easiest way is by means of characteristic functions (note that m.g.f.'s are useless in this case)

where $|\tilde{t}_n| \le |t|$ and $\psi'(0) = i\mathbb{E}Y_1 = 0$ and $\psi''(0) = -1$. By continuity of ψ'', it follows

$$\lim_n \psi_n(t) = e^{-\frac{1}{2}t^2}, \quad t \in \mathbb{R},$$

which by Theorem 7g14 proves the claim.

□

Back to statistics. Now we are ready to introduce the basic notions of the asymptotic theory of point estimation:

DEFINITION 7g24. *A sequence of estimators $(\hat{\theta}_n)$, $n \ge 1$ is called consistent, if $\hat{\theta}_n \to \theta$ in $\mathbb{P}_\theta$-probability, i.e. for any $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}_\theta\Big(\big|\hat{\theta}_n - \theta\big| \ge \varepsilon\Big) = 0, \quad \forall \theta \in \Theta.$$

In words, this definition means that large estimation errors are getting less probable as the sampling size increases. More qualitative information on the convergence is provided by the following notion:

DEFINITION 7g25. *A consistent sequence of estimators $(\hat{\theta}_n)$ is said to have an asymptotic (error) distribution F, if there exists a sequence of numbers $\alpha_n \nearrow \infty$ such that the scaled estimation error converges to F in distribution: for any $\theta \in \Theta$,*

$$\lim_{n\to\infty} \mathbb{P}_\theta\Big(\alpha_n\big(\hat{\theta}_n - \theta\big) \le x\Big) = F(x;\theta),$$

for all x, at which $x \mapsto F(x;\theta)$ is continuous.

REMARK 7g26. If $F(x;\theta)$ in the latter definition is Gaussian, the sequence $(\hat{\theta}_n)$ is called *asymptotically normal* (AN).

EXAMPLE 7g27. Let $X_1, ..., X_n$ be a sample with unknown mean $\theta = \mathbb{E}_\theta X_1 \in \mathbb{R}$ and finite variance $\mathbb{E}_\theta X_1^2 = \sigma^2$. Then by the LLN the sequence of estimators $\hat{\theta}_n = \bar{X}_n$ is consistent and by the CLT

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2), \quad \theta \in \Theta.$$

This means that the sequence of estimators $(\hat{\theta}_n)$ is consistent and asymptotically normal with rate $\sqrt{n}$ and variance σ^2. Remarkably, this result holds without any further assumptions on the c.d.f. of X_1, other than finiteness of the second moment. ■

Though the rate $\sqrt{n}$ and the asymptotic normality frequently emerge in statistical models, do not think that this is always the case:

EXAMPLE 7b3 (continued) Let us revisit the problem of estimating θ from the sample $X_1, ..., X_n$, where $X_1 \sim U([0,\theta])$, $\theta > 0$. We considered the MLE $\hat{\theta}_n(X) = \max_i X_i = M_n$ and the estimator $\tilde{\theta}_n(X) = 2\bar{X}_n$ and calculated the corresponding risks in (7b4) and (7b5). As we saw, the risks are comparable for any $n \ge 1$ and the estimator $\hat{\theta}_n(X)$ has smaller risk than the estimator $\tilde{\theta}_n(X)$ (and hence the latter is inadmissible).

Since everything is explicitly computable in this example, there is no need to appeal to the asymptotic theory. Still it will be instructive to analyze this example asymptotically. First of all, note that both sequences of estimators are consistent, since $R(\theta, \hat{\theta}_n) = \mathbb{E}_\theta(\theta - \hat{\theta}_n(X))^2 \to 0$

and $R(\theta, \tilde{\theta}_n) = \mathbb{E}_\theta(\theta - \tilde{\theta}_n(X))^2 \to 0$ as $n \to \infty$, i.e. both $\hat{\theta}_n$ and $\tilde{\theta}_n$ converge to θ in L_2 and hence also in $\mathbb{P}_\theta$-probability for all $\theta \in \Theta = \mathbb{R}_+$.

Now let's find the corresponding asymptotic error distributions. By the CLT, $\sqrt{n}(2\bar{X}_n - \theta)$ converges in distribution to $N\big(0, 4\mathrm{var}_\theta(X_1)\big)$, i.e. to $N(0, \theta^2/3)$. Hence $\tilde{\theta}_n$ is asymptotically normal with rate $\sqrt{n}$ and variance $\theta^2/3$. What about the MLEs $\hat{\theta}_n$?

Note that $X_1 - \theta$ has uniform distribution on $[-\theta, 0]$ and hence $M_n - \theta$ has the c.d.f.

$$F_n(x) = \begin{cases} 0 & x < -\theta \\ (x/\theta + 1)^n & x \in [-\theta, 0) \\ 1 & x \geq 0 \end{cases}$$

Note that

$$\lim_n F_n(x) = \mathbf{1}_{\{x \geq 0\}}, \quad x \in \mathbb{R}$$

which means that $\hat{\theta}_n - \theta \xrightarrow{d} 0$ under $\mathbb{P}_\theta$ and hence, by Lemma 7g19, $\hat{\theta}_n - \theta \xrightarrow{\mathbb{P}_\theta} 0$. This verifies consistency of $(\hat{\theta}_n)$.

Further, let (α_n) be an increasing sequence of positive numbers, then for $x \in \mathbb{R}$,

$$\mathbb{P}_\theta\big(\alpha_n(M_n - \theta) \leq x\big) = \mathbb{P}_\theta\big(M_n - \theta \leq x/\alpha_n\big) = F_n(x/\alpha_n) =$$
$$\begin{cases} 0 & x/\alpha_n < -\theta \\ \left(\frac{1}{\theta}\frac{x}{\alpha_n} + 1\right)^n & x/\alpha_n \in [-\theta, 0) \\ 1 & x/\alpha_n \geq 0 \end{cases} = \begin{cases} 0 & x < -\theta\alpha_n \\ \left(\frac{1}{\theta}\frac{x}{\alpha_n} + 1\right)^n & x \in [-\theta\alpha_n, 0) \\ 1 & x \geq 0 \end{cases}.$$

A nontrivial limit is obtained if we choose $\alpha_n := n$

$$\lim_n F_n(x) = \begin{cases} e^{x/\theta} & x \in (-\infty, 0) \\ 1 & x \geq 0 \end{cases} =: F(x).$$

Since $F(x)$ is a continuous non-decreasing function with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$, we conclude that $n(M_n - \theta)$ converges in distribution to a random variable with c.d.f. $F(x)$, which has the density

$$f(x;\theta) = \frac{1}{\theta} e^{x/\theta} \mathbf{1}_{\{x<0\}}, \quad x \in \mathbb{R}.$$

This is readily recognized as the density of $-\eta$, where η is exponential r.v. with mean θ. Hence the sequence $(\hat{\theta}_n)$ is consistent with rate n, which is much faster than the consistency rate of $\sqrt{n}$ offered by the sequence of the estimators $(\tilde{\theta}_n)$: roughly speaking, the accuracy attained by $\tilde{\theta}_n$ for $n = 900$, can be attained by $\hat{\theta}_n$ with only $n = 30$ samples! ■

Asymptotic error distribution of point estimators is often used in practice for construction of *approximate* confidence intervals. In this regard, the following lemma proves handy.

LEMMA 7g28 (the Delta method). *Let $\hat{\theta}_n$ be a consistent sequence of estimators, which is asymptotically normal with the rate $\sqrt{n}$ and variance $V(\theta)$. Then for a continuously differentiable function g, the sequence of plug-in estimators $g(\hat{\theta}_n)$ is consistent for $g(\theta)$ and is asymptotically normal with the variance $\big(g'(\theta)\big)^2 V(\theta)$.*

PROOF. Being differentiable, g is continuous and $g(\hat\theta_n)$ is a consistent sequence of estimators for $g(\theta)$ by Lemma 7g7. Further, the Lagrange formula for the reminder in the Taylor expansion yields:

$$g(\hat\theta_n) - g(\theta) = g'(\theta)\big(\hat\theta_n - \theta\big) + \Big(g'(\eta_n) - g'(\theta)\Big)\big(\hat\theta_n - \theta\big),$$

where $|\eta_n - \theta| \le |\theta - \hat\theta_n|$. Multiplying both sides by $\sqrt{n}$, we get

$$\sqrt{n}\big(g(\hat\theta_n) - g(\theta)\big) = g'(\theta)\sqrt{n}\big(\hat\theta_n - \theta\big) + \Big(g'(\eta_n) - g'(\theta)\Big)\sqrt{n}\big(\hat\theta_n - \theta\big) \tag{7g9}$$

The first term on the right hand side converges weakly to the Gaussian r.v. with zero mean and variance $\big(g'(\theta)\big)^2 V(\theta)$ by (2) of Slutsky's Lemma 7g17. Similarly the second term converges to 0 in probability by continuity of the derivative g' and the claim follows from Slutsky's Lemma 7g17. □

The following example demonstrates a number of approaches to construction of confidence intervals (recall Section 5, page 65), based on the limit theorems and the Delta method.

EXAMPLE 7g29. Let $X_1, ..., X_n$ be a sample from the distribution $\mathrm{Ber}(\theta)$ where $\theta \in \Theta = (0,1)$ is the unknown parameter. We would like to construct a confidence interval (shortly c.i.) with the given confidence level $1-\alpha$ (e.g. $1-\alpha = 0.95$), i.e. find an interval of the form $I(X) := [a_-(X), a_+(X)]$, such that

$$\mathbb{P}_\theta\Big(\theta \in [a_-(X), a_+(X)]\Big) \ge 1-\alpha, \quad \forall \theta \in \Theta. \tag{7g10}$$

The exact solution of this problem is computationally demanding, though possible (think how). When n is large, the limit theorems can be used to suggest approximate solutions in a number of ways.

The pivot method

Consider the sequence of random variables:

$$Y_n := \sqrt{n}\frac{\bar X_n - \theta}{\sqrt{\bar X_n(1-\bar X_n)} + \mathbf{1}_{\{\bar X_n \in \{0,1\}\}}}.$$

By the law of large numbers $\lim_n \bar X_n(1-\bar X_n) = \theta(1-\theta)$ and $\lim_n \mathbf{1}_{\{\bar X_n = 1\}} = \lim_n \mathbf{1}_{\{\bar X_n = 0\}} = 0$ in $\mathbb{P}_\theta$-probability and hence, by the CLT and Slutsky's lemma $Y_n \xrightarrow{d} N(0,1)$ under $\mathbb{P}_\theta$. The weak limit of the *pivot* random variables (Y_n) suggests the confidence interval:

$$\begin{aligned} a_-(X) &= \bar X_n - z_{1-\alpha/2} n^{-1/2}\Big(\sqrt{\bar X_n(1-\bar X_n)} + \mathbf{1}_{\{\bar X_n \in \{0,1\}\}}\Big) \\ a_+(X) &= \bar X_n + z_{1-\alpha/2} n^{-1/2}\Big(\sqrt{\bar X_n(1-\bar X_n)} + \mathbf{1}_{\{\bar X_n \in \{0,1\}\}}\Big) \end{aligned} \tag{7g11}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard Gaussian distribution, i.e. $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$ and hence

$$\mathbb{P}_\theta\Big(\theta \in [a_-(X), a_+(X)]\Big) = \mathbb{P}_\theta\big(Y_n \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\big) \approx 1 - 2\big(1 - \Phi(z_{1-\alpha/2})\big) = 1-\alpha,$$

as required in (7g10), where the approximate equality 'replaces' the convergence if n is large enough [44]. Note that for smaller α's, larger $z_{1-\alpha/2}$ emerges, i.e. wider c.i.'s correspond to more stringent requirements. On the other hand, for a fixed α, and large n, we get narrower c.i.'s, as should be expected.

Variance stabilizing transformation

Alternatively, the confidence interval can be constructed by means of the *variance stabilizing* transformation as follows. Let $g : (0,1) \mapsto \mathbb{R}$ be a continuously differentiable function, then applying the Delta method of Lemma (7g28) we have

$$\sqrt{n}\big(g(\bar{X}_n) - g(\theta)\big) \xrightarrow[n\to\infty]{d} N\Big(0, \big(g'(\theta)\big)^2\theta(1-\theta)\Big), \quad \forall\theta \in \Theta.$$

Now, if we choose g so that $g'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}$, the asymptotic variance will equal 1, irrespectively of θ. If in addition, g is increasing and invertible on $(0,1)$, the following c.i. can be suggested:

$$I(X) := \Big[g^{-1}\Big(g(\bar{X}_n) - z_{1-\alpha/2}/\sqrt{n}\Big), g^{-1}\Big(g(\bar{X}_n) + z_{1-\alpha/2}/\sqrt{n}\Big)\Big]. \tag{7g12}$$

Indeed,

$$\begin{aligned}
&\mathbb{P}_\theta\big(\theta \in I(X)\big) = \mathbb{P}_\theta\Big(g(\theta) \in g(I(X))\Big) = \\
&\mathbb{P}_\theta\Big(g(\theta) \in [g(\bar{X}_n) - z_{1-\alpha/2}/\sqrt{n}, g(\bar{X}_n) + z_{1-\alpha/2}/\sqrt{n}]\Big) = \\
&\mathbb{P}_\theta\Big(\sqrt{n}\big(g(\bar{X}_n) - g(\theta)\big) \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\Big) \xrightarrow{n\to\infty} 1 - 2\Phi(-z_{1-\alpha/2}) = 1-\alpha.
\end{aligned}$$

Of course, the question now is whether we can find an appropriate function g, referred to as the *variance stabilizing* transformation. In our case

$$g(\theta) = \int_0^\theta \frac{1}{\sqrt{s(1-s)}} ds$$

does the job, since the integral is well defined. Clearly, it is continuously differentiable on $(0,1)$, increasing and invertible. A calculation yields $g(\theta) = \frac{1}{2}\pi + \arcsin(2\,\theta - 1)$. Since we required only a particular form of g', we may omit the constant and just take $g(\theta) := \arcsin(2\,\theta - 1)$.

Wilson's method

As we saw above, by the CLT

$$\mathbb{P}_\theta\left(\left|\sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta)}}\right| \le z_{1-\alpha/2}\right) \approx 1-\alpha.$$

[44] How large n should be to support firmly such a belief ...? Clearly, to answer this question we must know more about convergence in the CLT (e.g. the rate of convergence , etc.). Applied statisticians usually use various rules of thumbs, which often can be justified by a considerable mathematical effort (see Example 8a4 below)

Note that

$$\Big\{\Big|\sqrt{n}\frac{\bar X_n-\theta}{\sqrt{\theta(1-\theta)}}\Big|\le z_{1-\alpha/2}\Big\}=\Big\{n\frac{(\bar X_n-\theta)^2}{\theta(1-\theta)}\le z^2_{1-\alpha/2}\Big\}=$$
$$\Big\{(\bar X_n-\theta)^2\le n^{-1}z^2_{1-\alpha/2}\theta(1-\theta)\Big\}=$$
$$\Big\{\theta^2(1+z^2_{1-\alpha/2}/n)-\theta(2\bar X_n+z^2_{1-\alpha/2}/n)+\bar X_n^2\le 0\Big\}=\big\{a_-(X)\le\theta\le a_+(X)\big\},$$

where

$$a_\pm(X):=\frac{\bar X_n+\frac12\frac1n z^2_\alpha\pm z_{1-\alpha/2}\sqrt{\frac1n\bar X_n(1-\bar X_n)+\frac{1}{4n^2}z^2_\alpha}}{1+\frac1n z^2_\alpha},$$

and hence the confidence interval $I(X)=[a_-(X),a_+(X)]$ has the coverage probability close to $1-\alpha$ when n is large. ■

The asymptotic risk of estimators is often easier to compute than their exact risk for a fixed sample size. In this regard, the Delta-method is the main tool:

EXAMPLE 7g30. Let $X_1,...,X_n$ be a sample from $\mathrm{Ber}(\theta)$ with $\theta\in(0,1)$ and consider the MLE $\hat\theta_n(X)=\bar X_n$. Since $\mathbb{E}_\theta|X_1|=\theta<\infty$, by the weak LLN, the sequence $\hat\theta_n(X)$ converges to θ in $\mathbb{P}_\theta$-probability for any $\theta\in\Theta$. Hence $\hat\theta_n(X)$ is a consistent sequence of estimators. Furthermore, as $\mathrm{var}_\theta(X_1)=\theta(1-\theta)<\infty$, by the CLT

$$\lim_{n\to\infty}\mathbb{P}\Big(\sqrt{n}\frac{\bar X_n-\theta}{\sqrt{\theta(1-\theta)}}\le x\Big)=\Phi(x),\quad\forall x\in\mathbb{R},$$

or equivalently (replace x by $x/\sqrt{\theta(1-\theta)}$),

$$\lim_{n\to\infty}\mathbb{P}\Big(\sqrt{n}(\bar X_n-\theta)\le x\Big)=\Phi(x/\sqrt{\theta(1-\theta)}),\quad\forall x\in\mathbb{R},$$

where $\Phi(x)$ is the standard Gaussian c.d.f. Hence the sequence $\hat\theta_n(X)=\bar X_n$ is asymptotically normal with the limit variance $\theta(1-\theta)$.

Now consider another sequence of estimators:

$$\tilde\theta_n(X)=\sqrt{\frac{1}{[n/2]}\sum_{i=1}^{[n/2]}X_{2i-1}X_{2i}}.$$

$Y_i:=X_{2i}X_{2i-1}$, $i=1,...,[n/2]$ are i.i.d. $\mathrm{Ber}(\theta^2)$ r.v's and hence again by the weak LLN

$$\frac{1}{[n/2]}\sum_{i=1}^{[n/2]}X_{2i-1}X_{2i}\overset{\mathbb{P}_\theta}{\to}\theta^2,$$

and since $u\mapsto\sqrt{u}$ is a continuous function, $\tilde\theta_n(X)\overset{\mathbb{P}_\theta}{\to}\sqrt{\theta^2}=\theta$, i.e. the sequence $\tilde\theta_n$ is also consistent.

Note that by the CLT,

$$\sqrt{[n/2]}\left(\tilde\theta_n^2(X)-\theta^2\right)\overset{d}{\to}N\big(0,\theta^2(1-\theta^2)\big),\quad n\to\infty$$

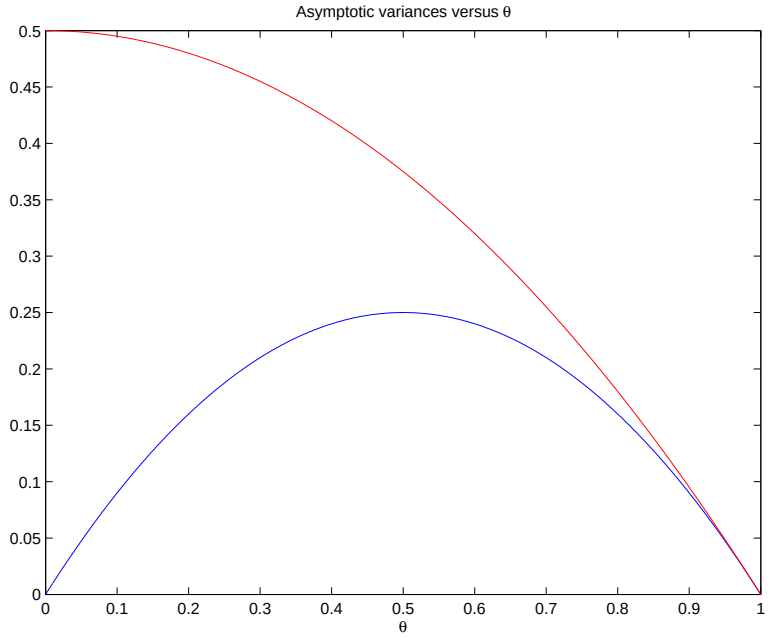

FIGURE 4. asymptotic variances of $\hat{\theta}_n$ and $\tilde{\theta}_n$ versus θ

Now we can apply the Delta method with $g(u) = \sqrt{u}$ to the sequence $\tilde{\theta}_n^2(X)$:

$$\sqrt{[n/2]}\left(g\big(\tilde{\theta}_n(X)\big) - g(\theta^2)\right) = \sqrt{[n/2]}\left(\tilde{\theta}_n(X) - \theta\right) \xrightarrow{d} N\Big(0, \big(g'(\theta^2)\big)^2\theta^2(1-\theta^2)\Big) = N(0, (1-\theta^2)/4).$$

Finally, $[n/2]/n \to 1/2$, we conclude (why?)

$$\sqrt{n}\left(\tilde{\theta}_n(X) - \theta\right) \xrightarrow{d} N(0, 1-\theta^2)$$

In this example, one can calculate the risks of $\hat{\theta}_n$ and $\tilde{\theta}_n$ for each fixed n and check whether they are comparable. However, computing the risk of $\tilde{\theta}_n$ does not appear to be an easy calculation. Hence considering the estimators in the asymptotic regime, i.e. through the asymptotic risks, often leads to a more tractable problem.

In this example, the competing sequences of estimators have the same rates, and moreover, are both asymptotically normal: hence, we can try to compare them by the asymptotic variance. Indeed, we see that the asymptotic variance of $\hat{\theta}_n$ is uniformly smaller than the variance of $\tilde{\theta}_n$ (see Figure 4). Hence we conclude that $\tilde{\theta}_n$ is asymptotically inferior to $\hat{\theta}_n$. Of course, just as in the non-asymptotic case two sequences of estimators may not be comparable by their asymptotic variances. ■

Two sequences of estimators do not have to be comparable, even if they have the same rate and the same limit type of distribution (e.g. asymptotically normal): it is possible that the asymptotic variances, being functions of θ, satisfy opposite inequalities on different regions of

Θ (similarly to Example 7b10). Note that a sequence of consistent estimators $(\hat{\theta}_n)$, which also converges in L_1, must be asymptotically unbiased [45]:

$$\lim_n \mathbb{E}_\theta \hat{\theta}_n = \theta, \quad \forall \theta \in \Theta.$$

Thus there still may exist[46] a sequence of estimators $(\hat{\theta}_n^*)$, which attains the Cramer-Rao information bound for unbiased estimators, asymptotically as $n \to \infty$:

$$\lim_n n\mathbb{E}_\theta(\hat{\theta}_n^* - \theta)^2 = \frac{1}{I(\theta)}, \quad \theta \in \Theta. \tag{7g13}$$

Guided by the intuition from the finite sample setting, we may think that no sequence of estimators can yield asymptotic risk, smaller than this bound and hence regard $(\hat{\theta}_n^*)$ *asymptotically efficient* (optimal). The following example shows that asymptotic optimality is a more delicate matter!

Example 7g31 (J.Hodges). Consider an i.i.d. sample $X_1, X_2, ...$ from $N(\theta, 1)$ distribution, where $\theta \in \mathbb{R}$ is the unknown parameter. As we already saw, the empirical mean $\bar{X}_n$ is the estimator, which has various optimality properties: it is the UMVUE, attaining the Cramer-Rao bound for unbiased estimators, as well as the minimax estimator for each fixed $n \geq 1$. Hence it is tempting to think that $\bar{X}_n$ is asymptotically optimal, in the sense that if $(\tilde{\theta}_n)$ is a sequence of asymptotically unbiased estimators, then

$$\underline{\lim}_n n^{-1}\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 \geq \lim_n n^{-1}\mathbb{E}_\theta(\bar{X}_n - \theta) = 1, \quad \forall \theta \in \mathbb{R}. \tag{7g14}$$

Remarkably, this is not the case! Consider the sequence of estimators

$$\tilde{\theta}_n(X) := \begin{cases} \bar{X}_n & |\bar{X}_n| \geq n^{-1/4} \\ 0 & |\bar{X}_n| < n^{-1/4} \end{cases}$$

Then under $\mathbb{P}_\theta$,

$$\begin{aligned} &\sqrt{n}\big(\tilde{\theta}_n - \theta\big) = \sqrt{n}\big(\bar{X}_n - \theta\big)\mathbf{1}_{\{|\bar{X}_n| \geq n^{-1/4}\}} - \sqrt{n}\theta\mathbf{1}_{\{|\bar{X}_n| < n^{-1/4}\}} = \\ &= \sqrt{n}\big(\bar{X}_n - \theta\big)\mathbf{1}_{\{|\sqrt{n}(\bar{X}_n - \theta) + \sqrt{n}\theta| \geq n^{1/4}\}} - \sqrt{n}\theta\mathbf{1}_{\{|\sqrt{n}(\bar{X}_n - \theta) + \sqrt{n}\theta| < n^{1/4}\}} = \\ &\stackrel{d}{=} Z\mathbf{1}_{\{|Z + \sqrt{n}\theta| \geq n^{1/4}\}} - \sqrt{n}\theta\mathbf{1}_{\{|Z + \sqrt{n}\theta| < n^{1/4}\}} \end{aligned}$$

where $\stackrel{d}{=}$ stands for equality in distribution and $Z \sim N(0,1)$. Hence for $\theta = 0$,

$$\mathbb{E}_\theta n\big(\tilde{\theta}_n - \theta\big)^2 = \mathbb{E}_\theta Z^2 \mathbf{1}_{\{|Z| \geq n^{1/4}\}} \leq \sqrt{\mathbb{E}_\theta Z^4}\sqrt{\mathbb{P}_\theta(|Z| \geq n^{1/4})} \xrightarrow{n\to\infty} 0,$$

where we used the Cauchy–Schwarz inequality. For $\theta \neq 0$,

$$\begin{aligned} \mathbb{E}_\theta n\big(\tilde{\theta}_n - \theta\big)^2 =& \mathbb{E}_\theta\Big(Z\mathbf{1}_{\{|Z + \sqrt{n}\theta| \geq n^{1/4}\}} - \sqrt{n}\theta\mathbf{1}_{\{|Z + \sqrt{n}\theta| < n^{1/4}\}}\Big)^2 = \\ & \mathbb{E}_\theta\Big(Z - (Z + \sqrt{n}\theta)\mathbf{1}_{\{|Z + \sqrt{n}\theta| < n^{1/4}\}}\Big)^2 = 1 + r_n, \end{aligned}$$

[45] sometimes, this is referred as approximately unbiased, while the term "asymptotically unbiased" is reserved for a slightly different notion

[46] just like in the UMVUE theory: not all unbiased estimators are comparable, while the optimal unbiased estimator can be found through R-B if the minimal sufficient statistic is complete!

with the residual satisfying

$$|r_n| = \Big|\mathbb{E}_\theta\Big(2Z(Z+\sqrt{n}\theta) + (Z+\sqrt{n}\theta)^2\Big)\mathbf{1}_{\{|Z+\sqrt{n}\theta|<n^{1/4}\}}\Big| \le$$
$$3\mathbb{E}_\theta\big(|Z| + \sqrt{n}|\theta|\big)^2\mathbf{1}_{\{|Z+\sqrt{n}\theta|<n^{1/4}\}} \le 6\sqrt{\mathbb{E}_\theta\big(|Z|^2+n\theta^2\big)^2}\sqrt{\mathbb{P}_\theta\big(|Z+\sqrt{n}\theta|<n^{1/4}\big)}.$$

For $\theta > 0$ and all $n \ge (2/\theta)^4$

$$\mathbb{P}_\theta\big(|Z+\sqrt{n}\theta|<n^{1/4}\big) \le \mathbb{P}_\theta\big(Z<n^{1/4}-\sqrt{n}\theta\big) \le \mathbb{P}_\theta\Big(Z<-\frac{\theta}{2}\sqrt{n}\Big) = \Phi\Big(-\frac{\theta}{2}\sqrt{n}\Big),$$

where $\Phi(x)$ is the c.d.f. of $N(0,1)$ distribution, which for $x<-1$ satisfies

$$\Phi(x) = \int_{-\infty}^x \frac{1}{2\pi}e^{-y^2/2}dy \le -\int_{-\infty}^x ye^{-y^2/2}dy = e^{-x^2/2}.$$

Plugging these bounds back, we see that $r_n \to 0$ for $\theta > 0$ and, similarly, for $\theta < 0$. To recap,

$$\lim_n n\mathbb{E}_\theta(\tilde\theta_n(X)-\theta)^2 = \begin{cases} 1 & \theta \ne 0 \\ 0 & \theta = 0 \end{cases} \tag{7g15}$$

which shows that (7g14) fails for the sequence $(\tilde\theta_n)$ at the point $\theta = 0$. Hodges' estimator is *superefficient*, e.g., in the sense that its asymptotic variance is better than the asymptotic Cramer-Rao bound for unbiased estimators!

This is of course no paradox: Hodges' estimator is biased for each fixed n and hence doesn't have to satisfy the Cramer-Rao bound for *unbiased* estimators. The relevant Cramer-Rao bound for $\tilde\theta_n$ is

$$R(\theta,\tilde\theta_n) = \mathrm{var}_\theta(T) + b^2(\theta,\tilde\theta_n) \ge \frac{\Big(\frac{\partial}{\partial\theta}b(\theta,\tilde\theta_n)+1\Big)^2}{I_n(\theta)} + b^2(\theta,\tilde\theta_n) =: \mathrm{CR}_n(\theta),$$

where $b(\theta,\hat\theta_n) = \mathbb{E}_\theta\hat\theta_n - \theta$ is the bias and $I_n(\theta) = n$ is the Fisher information in the sample. A calculation reveals that $nb^2(\theta,\tilde\theta_n) \to 0$ for any $\theta \in \Theta$, but $\lim_n \frac{\partial}{\partial_\theta}b(\theta,\tilde\theta_n)\big|_{\theta=0} = -1$ and hence

$$n\mathrm{CR}_n(\theta)\big|_{\theta=0} \to 0$$

which agrees with (7g15). ■

This example shows that a sequence of consistent and asymptotically normal estimators, satisfying (7g13), can be outperformed. Hodges' example can be modified so that the set of points in Θ, at which the estimator is superefficient, is infinitely countable. Remarkably, a theorem of L.Le Cam shows that the set of such points cannot be essentially larger: more precisely, it has zero Lebesgue measure ('length'). This allows to define the following local minimax notion of asymptotic efficiency:

Definition 7g32. *A sequence of estimators $(\hat\theta_n^*)$ is asymptotically efficient with rate $\sqrt{n}$ if*

$$\varliminf_{\delta\searrow 0}\varliminf_{n\to\infty}\sup_{|\theta-\theta_0|\le\delta} n\mathbb{E}_\theta(\hat\theta_n-\theta)^2 \ge \varliminf_{\delta\searrow 0}\varliminf_{n\to\infty}\sup_{|\theta-\theta_0|\le\delta} n\mathbb{E}_\theta(\hat\theta_n^*-\theta)^2, \quad \forall\theta_0\in\Theta,$$

for any other sequence of estimators (θ_n).

Note that in this sense, Hodges' estimator is not better than $\bar{X}_n$ anymore, since $n\mathbb{E}_\theta(\bar{X}_n - \theta)^2 = 1$ for all n and $\theta \in \mathbb{R}$ and

$$\varliminf_{\delta \searrow 0} \varliminf_{n\to\infty} \sup_{|\theta-\theta_0|\le\delta} n\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 \ge 1.$$

On the other hand, we saw in Example 7c12 that $\bar{X}_n$ is minimax for any fixed n if $\Theta = \mathbb{R}$. However, $\bar{X}_n$ is no longer minimax on small bounded intervals: for example, for the trivial estimator $\hat{\theta} \equiv 0$

$$\sup_{\theta\in[-\delta,\delta]} \mathbb{E}_\theta(\hat{\theta} - \theta) = \delta^2$$

which is smaller than

$$\sup_{\theta\in[-\delta,\delta]} \mathbb{E}_\theta(\bar{X}_n - \theta) = 1/n$$

if δ is small enough. Nevertheless, it can be shown that $\bar{X}_n$ is asymptotically efficient in the sense of definition (7g32) (see Theorem 7g34 below).

REMARK 7g33. We already saw that Hodges' estimator is superefficient: it performs as well as the asymptotically efficient estimator $\bar{X}_n$ at all points of Θ and outperforms it at the *superefficiency* point 0. Hence it is tempting to think that Hodges' estimator is asymptotically efficient as well. Again, counterintuitively, it is not and, moreover, its local minimax risk is infinite! To see why, let $\theta_n := C/\sqrt{n}$ with a positive constant $C > 0$. Then

$$\mathbb{E}_{\theta_n} n\big(\tilde{\theta}_n - \theta_n\big)^2 = \mathbb{E}_{\theta_n}\Big(Z\mathbf{1}_{\{|Z+\sqrt{n}\theta_n|\ge n^{1/4}\}} - \sqrt{n}\theta_n\mathbf{1}_{\{|Z+\sqrt{n}\theta_n|<n^{1/4}\}}\Big)^2 =$$

$$\mathbb{E}_{\theta_n}\Big(Z\mathbf{1}_{\{|Z+C|\ge n^{1/4}\}} - C\mathbf{1}_{\{|Z+C|<n^{1/4}\}}\Big)^2 \xrightarrow{n\to\infty} C^2.$$

Since for all n large enough,

$$\sup_{|\theta|\le\delta} n\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 \ge \mathbb{E}_{\theta_n} n\big(\tilde{\theta}_n - \theta_n\big)^2$$

we conclude that

$$\varliminf_{n} \sup_{|\theta|\le\delta} n\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 \ge C^2$$

and, since C is arbitrary,

$$\varliminf_{\delta \searrow 0} \varliminf_{n\to\infty} \sup_{|\theta-\theta_0|\le\delta} n\mathbb{E}_\theta(\tilde{\theta}_n - \theta)^2 = \infty.$$

Roughly speaking, this means that Hodges' estimator may perform very poorly at the points close to the superefficiency point for large n.

Finding asymptotically efficient estimators is an interesting problem, which is beyond the scope of our course. Remarkably, the sequence of the MLE's often turns to be consistent, asymptotically normal and asymptotically efficient:

THEOREM 7g34. [47] *Let $X_1, ..., X_n$ be a sample from the probability density*[48] *$f(x;\theta)$, $\theta \in \Theta$ and let $(\hat{\theta}_n)$ be the sequence of MLE estimators:*

$$\hat{\theta}_n(X) \in \operatorname{argmax}_{\theta\in\Theta} L(\theta; X),$$

[47] adapted from [**2**]

[48] the theorem holds also for the discrete case with appropriate adjustments

where $L(X;\theta)=\prod_{i=1}^n f(X_i,\theta)$ *is the likelihood of the model.*

Assume that

(A_c) $\Theta=(a,b)$ *for some* $-\infty<a<b<\infty$

(A_0) *the model is identifiable, i.e. for any* $\theta_1\neq\theta_2$, $f(x;\theta_1)\neq f(x;\theta_2)$ *for all* $x\in J$, *where* J *is an open interval in* $\mathbb{R}$.

(R) $\sqrt{f(x;\theta)}$ *is continuously differentiable and the Fisher information*

$$I(\theta)=\int_{\mathbb{R}}\frac{\big(f'(x;\theta)\big)^2}{f(x;\theta)}dx$$

is finite and strictly positive on Θ.

Then $(\hat\theta_n)$ *is consistent* [49]. *If in addition to the above assumptions,*

(RR_m) *the function* $\ell(x;\theta)=\log f(x;\theta)$ *is twice continuously differentiable in* θ. *The function* $|\ell''(x;\theta)|$ *is majorized by a function* $h(x)$ *independent of* θ, *i.e.* $|\ell''(x;\theta)|\le h(x)$, *for which*

$$\int_{\mathbb{R}}h(x)f(x;\theta)dx<\infty,\quad\forall\theta\in\Theta$$

(RR_d) *the differentiation under the integral is valid:*

$$\int_{\mathbb{R}}f'(x;\theta)dx=0,\quad\forall\theta\in\Theta$$

then $(\hat\theta_n)$ *is asymptotically normal*

$$\sqrt{n}\big(\hat\theta_n-\theta\big)\xrightarrow{d}N\big(0,1/I(\theta)\big),\quad\theta\in\Theta^\circ$$

and asymptotically efficient, moreover[50], *the moments of all orders converge.*

Example 7g35. Let $X_1,...,X_n$ be a sample from Cauchy density with the location parameter $\theta\in\Theta=[-c,c]$ (for some constant $c>0$):

$$f(x;\theta)=\frac{1/\pi}{1+(x-\theta)^2}.$$

Note that the empirical mean $\bar X_n$ is not consistent for θ in this case (see Problem 7.58). The MLE $\hat\theta_n$ cannot be found in an explicit form beyond $n\ge2$ and hence the maximization of the likelihood is to be done numerically[51]. The assumptions (A_c) and (A_0) obviously hold. The function $\sqrt{f(x;\theta)}$ is continuously differentiable and

$$I(\theta)=\frac1\pi\int\left(\frac{2(x-\theta)}{\big(1+(x-\theta)^2\big)^2}\right)^2\frac{1}{1+(x-\theta)^2}dx=\frac1\pi\int\frac{4(x-\theta)^2}{\big(1+(x-\theta)^2\big)^3}dx<\infty$$

---

[49] in fact, even stronger result holds under these assumptions, namely $\frac{\sqrt{n}}{a_n}(\hat\theta_n-\theta)\xrightarrow{\mathbb{P}_\theta}0$ for any unbounded sequence of numbers (a_n).

[50] recall that convergence in distribution guarantees convergence of expectations for continuous and bounded functions, but not necessarily for polynomials, i.e. moments

[51] Since the likelihood $L_n(x;\theta)$ decreases as $|\theta|\to\infty$, the MLE exists and is one of the roots of the score function $\frac{\partial}{\partial\theta}L(X;\theta)$. However, the number of roots of the score function grows with n and hence deciding which root is the MLE may be a challenge (remarkably, the set of roots converges to a Poisson process on $\mathbb{R}$ after an appropriate rescaling, see the paper [**10**])

is positive and bounded for any $\theta \in \Theta$. Hence by Theorem 7g34, the sequence of MLE's is consistent for θ. By checking further assumptions, we may be lucky to be able to conclude that $(\hat{\theta}_n)$ is also asymptotically normal (try!). ■

The mathematical tools for proving, or even, sketching the ideas, leading to this remarkable conclusion are far beyond the scope of our course, but we shall briefly demonstrate what kind of difficulties, arise in establishing e.g. consistency of MLE.

The asymptotic analysis is usually straightforward if the MLE is given by an explicit formula (as in Examples 7b3 and 7g30) or as the unique root of some equation (see Problem 7.57). However, in the majority of practical situations, the MLE is found numerically for the concrete sample at hand and its asymptotic analysis should be based directly on its definition as a maximizer. A number of different approaches have been developed for this purpose, and we shall sketch one of them below.

Suppose that we sample $X = (X_1, ..., X_n)$ from a p.d.f. $f(x;\theta)$. Then the corresponding log-likelihood is given by

$$\log L_n(x;\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\theta), \quad x \in \mathbb{R}^n, \quad \theta \in \Theta.$$

The MLE is defined as a maximizer of $\log L_n(X;\theta)$:

$$\hat{\theta}_n(X) = \operatorname{argmax}_{\theta\in\Theta} \log L_n(X;\theta).$$

Denote by θ_0 the actual value of the parameter (unknown to us) and note that the very same $\hat{\theta}_n(X)$ is obtained if we maximize a different quantity, more convenient for the purposes of analysis:

$$\hat{\theta}_n(X) = \operatorname{argmax}_{\theta\in\Theta} \frac{1}{n} \log \Big(L_n(X;\theta)/L_n(X;\theta_0) \Big).$$

Now if $\mathbb{E}_{\theta_0} |\log f(X_1;\theta)| < \infty$ for all $\theta_0, \theta \in \Theta$, then by the strong LLN

$$\frac{1}{n} \log \Big(L_n(X;\theta)/L_n(X;\theta_0) \Big) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\theta_0)} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta_0}-a.s.}$$

$$\mathbb{E}_{\theta_0} \log \frac{f(X_1;\theta)}{f(X_1;\theta_0)} = \int_{\mathbb{R}} \log \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx =: H(\theta,\theta_0).$$

The quantity $-H(\theta,\theta_0)$ is called the Kullback-Leibler relative entropy (or divergence) and it is not hard to see [52] that it is nonnegative and has the properties, similar[53] to those of a distance between the p.d.f.'s $f(x;\theta)$ and $f(x;\theta_0)$. If the statistical model is identifiable, then $\theta \mapsto H(\theta,\theta_0)$ has a unique maximum at θ_0, i.e. $H(\theta,\theta_0) < H(\theta_0,\theta_0) = 0$ for all $\theta \neq \theta_0$.

Thus we have a sequence of (random) functions

$$H_n(\theta;\theta_0) := \frac{1}{n} \log \Big(L_n(X;\theta)/L_n(X;\theta_0) \Big)$$

converging to $H(\theta,\theta_0)$, which has a unique maximum at $\theta := \theta_0$. It is tempting to conclude that for each fixed θ_0, the maximizer of $H_n(\theta;\theta_0)$ over θ converges to the maximizer of $H(\theta,\theta_0)$, that

[52] using the Jensen inequality

[53] The K-L divergence is not a true distance, since it is not symmetric

is $\hat{\theta}_n \to \theta_0$ and thus $\hat{\theta}_n$ is consistent. Unfortunately, this does not have to be the case[54] and for the latter to hold, a stronger - uniform over Θ - convergence is required in general. Establishing such convergence is harder and can be done under appropriate assumption. Once consistency is established, one can study the asymptotic normality, essentially using Taylor's expansion.

Exercises

Methods of point estimation.

PROBLEM 7.1. Consider a population made up of three different types of individuals occurring in the Hardy-Weinberg proportions θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$, respectively, where $0 < \theta < 1$.

(1) Show that $T_3 = N_1/n + N_2/2n$ is a frequency substitution estimator of θ
(2) Using the estimator of (1), what is a frequency substitution estimator of the odds ratio $\theta/(1-\theta)$?
(3) Suppose X takes the values $1, 0, 1$ with respective probabilities p_1, p_2, p_3 given by the Hardy-Weinberg proportions. By considering the first moment of X, show that T_3 is a method of moment estimator of θ.
(4) Find the MLE 0f θ and compare to the other estimators, obtained so far.

PROBLEM 7.2. Consider n systems with failure times $X_1, ..., X_n$ assumed to be independent and identically distributed with exponential $\exp(\lambda)$ distributions.

(1) Find the method of moments estimator of λ based on the first moment.
(2) Find the method of moments estimator of λ based on the second moment.
(3) Combine your answers to (1) and (2) to get a method of moment estimator of λ based on the first two moments.
(4) Find the method of moments estimator of the probability $P(X_1 > 1)$ that one system will last at least a month.
(5) Find the MLE of λ

PROBLEM 7.3. Let $X_1, ..., X_n$ be the indicators of n Bernoulli trials with probability of success 0.

(1) Show that $\bar{X}_n$ is a method of moments estimator of θ.
(2) Exhibit method of moments estimators for $\mathrm{var}_\theta(X_1) = \theta(1-\theta)$ first using only the first moment and then using only the second moment of the population. Show that these estimators coincide.
(3) Argue that in this case all frequency substitution estimators of $q(\theta)$ must agree with $q(\bar{X}_n)$.
(4) Find the MLE of θ

PROBLEM 7.4. Suppose $X = (X_1, ..., X_n)$ where the X_i are independent $N(0, \sigma^2)$

(1) Find an estimator of σ^2 based on the second moment.

[54] think of a counterexample: a sequence of (even deterministic) functions $f_n(x) \to f(x)$ for all x, where f_n and f have unique maxima x_n^* and x^*, but $x_n^* \not\to x^*$

(2) Construct an estimator of σ using the estimator of part (1) and the equation $\sigma = \sqrt{\sigma^2}$
(3) Use the empirical substitution principle to construct an estimator of σ using the relation $\mathbb{E}|X_1| = \sigma\sqrt{2\pi}$.

PROBLEM 7.5. An object of unit mass is placed in a force field of unknown constant intensity θ. Readings $Y_1, ..., Y_n$ are taken at times $t_1, ..., t_n$ on the position of the object. The reading Y_i differs from the true position $(\theta/2)t_i^2$ by a random error ε_i. We suppose the ε_i to have mean 0 and be uncorrelated with constant variance.

(1) Find the least square estimator (LSE) of θ.
(2) Can you compute the MLE of θ without additional assumptions ? The method of moments estimator ?

PROBLEM 7.6. Let $Y_1, ..., Y_n$ be independent random variables with equal variances such that $\mathbb{E}Y_i = \alpha z_i$ where the z_i are known constants. Find the least squares estimator of α.

PROBLEM 7.7. Suppose $Y_1, ..., Y_{n_1+n_2}$ are given by

$$Y_i = \begin{cases} \theta_1 + \varepsilon_i, & i = 1, ..., n_1 \\ \theta_2 + \varepsilon_i, & i = n_1 + 1, ..., n_2, \end{cases}$$

where $\varepsilon_1, ..., \varepsilon_{n_1+n_2}$ are independent $N(0, \sigma^2)$ variables.

(1) Find the LSE of $\theta = (\theta_1, \theta_2)$
(2) Find the MLE of θ, when σ^2 is known
(3) Find the MLE of θ, when σ^2 is unknown

PROBLEM 7.8. Let $X_1, ..., X_n$ be a sample from one of the following distributions. Find the MLE of θ.

(1) $f(x;\theta) = \theta e^{-\theta x}, x \geq 0, \theta > 0$ (exponential p.d.f.)
(2) $f(x;\theta) = \theta c^\theta x^{-(\theta+1)}, x \geq c$, $c > 0$ and $\theta > 0$ (Pareto p.d.f.)
(3) $f(x;\theta) = c\theta^c x^{-(c+1)}, x \geq \theta$, $c > 0$, $\theta > 0$ (Pareto p.d.f.)
(4) $f(x;\theta) = \sqrt{\theta} x^{\sqrt{\theta}-1}$, $x \in [0, 1]$, $\theta > 0$ ($\beta(\sqrt{\theta}, 1)$ p.d.f.)
(5) $f(x;\theta) = (x/\theta^2) \exp\{-x^2/2\theta^2\}$, $x > 0$, $\theta > 0$ (Rayleigh p.d.f.)
(6) $f(x;\theta) = \theta c x^{c-1} \exp\{-\theta x^c\}$, $x \geq 0$, $c > 0$, $\theta > 0$ (Weibull p.d.f.)

PROBLEM 7.9. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$. Find the MLE of $\theta = (\mu, \sigma^2)$ under the assumption that $\mu \geq 0$.

PROBLEM 7.10. Let $X_1, ..., X_n$ be a sample from the p.d.f.

$$f(x;\theta) = \begin{cases} \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, & x \geq \mu \\ 0 & x < \mu \end{cases}$$

where $\sigma > 0$ and $\mu \in \mathbb{R}$.

(1) Find the MLE of $\theta = (\mu, \sigma^2)$
(2) Find the MLE of $\mathbb{P}_\theta(X_1 \geq t)$ for $t > \mu$.

PROBLEM 7.11. Let $X_1 \sim N(\mu, \sigma^2)$. Show that the MLE of $\theta = (\mu, \sigma^2)$ doesn't exist.

PROBLEM 7.12. Let $X_1, ..., X_n$ be independent r.v. with $X_i \sim N(\theta_i, 1)$, $\theta_i \in \mathbb{R}$

(1) Find the MLE for $\theta = (\theta_1, ..., \theta_n)$
(2) For $n = 2$, find the MLE of θ, under the constraint $\theta_1 \leq \theta_2$

PROBLEM 7.13. Let $X_1, ..., X_k$ be a sample from Geo$(1-\theta)$. The observations are given by $Y_i = \min(X_i, r+1)$, where r is a positive integer. If X_i's are interpreted as times, then they can be observed only till certain known time r (this is known in statistics as censored observations). The p.m.f. of Y_1 is[55]

$$p(k;\theta) = \begin{cases} \theta^{k-1}(1-\theta), & k = 1, ..., r \\ \theta^r, & k = r+1 \end{cases}.$$

Let M be the number of Y_i's, such that $Y_i = r+1$. Show that the MLE of θ is given by:

$$\hat{\theta}(Y) = \frac{\sum_{i=1}^n Y_i - n}{\sum_{i=1}^n Y_i - M}.$$

PROBLEM 7.14. Let $X_1, ..., X_n$ be a sample form the Cauchy distribution with the location parameter θ, i.e.

$$f(x;\theta) = \frac{1/\pi}{1 + (x-\theta)^2}, \quad x \in \mathbb{R}, \theta \in \mathbb{R}.$$

(1) Show that if $n = 1$, then the MLE is $\hat{\theta} = X_1$
(2) Show that for $n = 2$, the equation $\frac{\partial}{\partial \theta} f(x_1, x_2; \theta) = 0$ has several roots and find the MLE

PROBLEM 7.15. Let $X_1, ..., X_n$ be a random sample from p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} I(x \in (0,1)), \quad \theta > 0.$$

(1) Show that $\mathbb{E}_\theta X_1 = \theta/(\theta+1)$
(2) Find the method of moments estimator θ
(3) Compute the maximum likelihood estimator of θ
(4) Find the maximum likelihood estimator of $\mathbb{E}_\theta$

[55] note that $\mathbb{P}_\theta(X_1 > r) = \theta^r$

Optimality.

PROBLEM 7.16. Let X be the p.m.f. of a Poisson r.v., conditioned to be positive[56]:

$$p(k;\theta) = \frac{e^{-\theta}\theta^k/k!}{1-e^{-\theta}}, \quad k \in \{1,2,...\},\ \theta > 0.$$

It is required to estimate $q(\theta) = 1 - e^{-\theta}$ and for this purpose the estimator

$$T^*(X) = \begin{cases} 0 & X \text{ is odd} \\ 2 & X \text{ is even} \end{cases}.$$

Show that T^* is an unbiased estimator of $q(\theta)$. Is it a good estimator ...?

PROBLEM 7.17. Let $X \sim N(\theta, 1)$ and consider the estimators of θ of the form $T_{a,b}(X) = aX + b$, where $a, b \in \mathbb{R}$.

(1) Calculate the MSE risk of $T_{a,b}$
(2) Compare the MSE risks of $T_{1/2,0}$ and $T_{1,0} = X$ (the "natural" estimator) as a function of θ. Show that none of these estimators is better than the other for al $\theta \in \Theta$.
(3) Is there an estimator of the form $T_{a,b}$ with the MSE risk, better than that of $T_{1,0} = X$ for all $\theta \in \Theta$?
(4) Show that $T_{1,0}$ is the only unbiased estimator among $T_{a,b}$, $a, b \in \mathbb{R}$.

PROBLEM 7.18. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$. It is required to estimate σ^2.

(1) Calculate the MSE risk of the estimators

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{and} \quad \tilde{S}^2 := \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

(2) Show that the MSE risk of S^2 is greater than of $\tilde{S}^2$ for all $\theta \in \Theta$ (i.e. S^2 is inadmissible).

Bayes estimation.

PROBLEM 7.19. Prove Lemma 7c8 in the case of countable parametric space Θ.

PROBLEM 7.20 (The sunrise problem of Laplace). "What is the probability that the sun will rise tomorrow?" is the question which P-S. Laplace addressed in the 18-th century. He suggested that at the beginning of the world (he has literally taken the date from the Bible), the probability of the sun to rise was completely uncertain, which he expressed by assuming that p was sampled form $U([0,1])$. Further, he assumed that the sun rises each day independently with the same conditional probability of success p. More precisely, if X_i takes value 1 if the sun rises on the i-th morning, then X_n, $n \geq 1$ forms a sequence of i.i.d. Ber(p) r.v.'s, conditioned on p. Prove the *rule of succession* :

$$\mathbb{P}(X_{n+1} = 1 | X_1 + ... + X_n = s) = \frac{s+1}{n+2}.$$

[56] i.e. the conditional p.m.f of Y given $\{Y > 0\}$, where $Y \sim \text{Poi}(\theta)$.

PROBLEM 7.21. Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a statistical model with Θ being a discrete set of points. Show that the Bayes estimator of θ with respect to the prior π and the loss function:

$$\ell_0(\theta,\eta) = I(\theta \neq \eta),$$

is given by the *maximum a posteriori probability* (MAP)

$$\hat{\theta}^*(X) = \text{argmax}_{\eta\in\Theta}\mathbb{P}(\theta = \eta|X).$$

Explain, why the MAP estimator minimizes the probability of error in guessing the value of θ, given the observation X.

PROBLEM 7.22. Show that β distribution is conjugate to the likelihood of n i.i.d. Ber(θ), $\theta \in [0,1]$ r.v.'s Find the corresponding posterior parameters and deduce the Bayes estimator under MSE.

PROBLEM 7.23. Prove that Pareto distribution with the p.d.f.

$$f(x;c,\alpha) = \frac{\alpha c^\alpha}{x^{\alpha+1}} I(x \geq c), \quad c > 0, \alpha > 0$$

is conjugate to the likelihood of the sample $X = (X_1, ..., X_n)$ of size n from $U([0,\theta])$, $\theta \in \mathbb{R}_+$. Show that the posterior distribution has the Pareto density with parameters $\max(c, \max_i X_i)$ and $\alpha + n$. Check that the Bayes estimator under MSE is given by:

$$\hat{\theta}^*(X) = \frac{(\alpha+n)\max(c, X_1, ..., X_n)}{\alpha+n-1}.$$

PROBLEM 7.24. Let $X_1, ..., X_n$ be a sample from Poi(θ), $\theta > 0$. Assume the prior $\Gamma(\alpha,\beta)$.

(1) Find the Bayes estimator of θ with respect to the loss function $\ell(\theta,\eta) = (\theta-\eta)^2/\theta$
(2) Find the Bayes risk of the Bayes estimator

PROBLEM 7.25. Let $X = (X_1, ..., X_n)$ be a sample from the exponential distribution exp(θ), $\theta > 0$. Assume the prior $\Gamma(\alpha,\beta)$.

(1) Find the conditional distribution of θ given X. Is the prior conjugate to the likelihood ?
(2) Find the Bayes estimator of θ with respect to the loss function $\ell(\theta,\eta) = (\theta-\eta)^2$.
(3) Find the MLE of θ
(4) Calculate the risk functions of the estimators in (3) and (2) and compare
(5) Calculate the Bayes risk of the estimaets in (3) and (2) and compare

PROBLEM 7.26. Show that Bayesian estimator with respect to quadratic loss is biased, if the posterior is non-degenerate (i.e. the posterior variance is positive with positive probability)

PROBLEM 7.27. Find the minimax estimator for $\theta \in \mathbb{R}_+$, given the sample $X_1, ..., X_n \sim$ Poi(θ)

PROBLEM 7.28. Show that the naive estimator $\hat{\theta} = X$ is minimax in Stein's example 7b8.

Unbiased estimation.

PROBLEM 7.29. Argue that an estimator, which is not a function of the minimal sufficient statistic, is inadmissible with respect to the quadratic risk.

PROBLEM 7.30. Let $X_1, ..., X_n$ be a sample from a p.d.f. with unknown mean μ and variance σ^2. Define $T(X) = \sum_{i=1}^n c_i X_i$.

(1) Show that T is an unbiased estimator of μ if and only if $\sum_{i=1}^n c_i = 1$.
(2) Show that $\bar{X}_n$ is the UMVUE among all estimators of the above form.

PROBLEM 7.31. Let $X_1, ..., X_n$ be a sample from $N(\mu, 1)$. It is required to estimate $\mathbb{P}_\mu(X_1 \geq 0) = \Phi(\mu)$

(1) Show that $T(X) = I(X_1 \geq 0)$ is an unbiased estimator
(2) Apply the R-B theorem with the sufficient statistic $\bar{X}_n$ to obtain an improved estimator.
Hint: note that $(X_1, \bar{X})$ is a Gaussian vector.
(3) Show that the obtained estimator is UMVUE

PROBLEM 7.32. Let $X_1, ..., X_n$ be a sample from $U([0, \theta])$, $\theta > 0$ and let $M_n(X) = \max_i X_i$.

(1) Show that $T^*(X) = \frac{n+1}{n} M_n(X)$ is an unbiased estimator of θ.
(2) Let $T(X) = \frac{n+2}{n+1} M_n(X)$. Show that

$$R(\theta, T) < R(\theta, M_n), \quad \text{and} \quad R(\theta, T) < R(\theta, T^*), \quad \forall \theta \in \Theta$$

and conclude that both M_n and T^* (which is the UMVUE!) are inadmissible.

PROBLEM 7.33. In each one of the following cases, show that ϕ is an unbiased estimator of the parameter of interest, find a sufficient statistic and improve ϕ by means of the R-B procedure.

(1) X_1, X_2 is a sample from Geo(θ), $\phi(X) = I(X_1 = 1)$ is an estimator of θ
(2) $X_1, ..., X_n$ is a sample form Ber(θ), $\phi(X) = \prod_{i=1}^n X_i$ is an estimator of θ^n
(3) $X_1, ..., X_n$ is a sample form Ber(θ), $\phi(X) = X_1 - X_1 X_2$ is an estimator of $\theta(1 - \theta)$

PROBLEM 7.34. Let $X_1 \sim$ Geo(θ). It is required to estimate $\theta/(1 + \theta)$ and the estimator $T(X) = e^{-X}$ is considered.

(1) Is $T(X)$ unbiased ?
(2) Let X_2 be an additional sample from Geo(θ), independent of X_1 and define $S = X_1 + X_2$. Find the conditional distribution of X_1, given S.
(3) Apply the R-B procedure with[57] S from (2), to improve the estimator from (1)

PROBLEM 7.35. Let $X_1, ..., X_n$ be a sample from Ber(θ). Show that $S(X) = \sum_{i=1}^n X_i$ is a complete statistic, if Θ contains more than n points. Argue that $\bar{X}_n$ is the UMVUE of θ.

[57] check that S is sufficient

PROBLEM 7.36. Let $X_1, ..., X_n$ be a sample from $U([\theta_1, \theta_2])$, where $\theta = (\theta_1, \theta_2)$ is the unknown parameter.

(1) Specify the parametric space Θ.
(2) Show that $T(X) = (\min_i X_i, \max_i X_i)$ is a sufficient statistic.
(3) Assuming that T is complete, argue that $(T_1 + T_2)/2$ is the UMVUE of the mean $(\theta_1 + \theta_2)/2$

PROBLEM 7.37. Consider $N = (N_1, ..., N_k) \sim \text{Mult}(n; \theta)$, where $\theta = (\theta_1, ..., \theta_k) \in \mathcal{S}^{k-1} = \{x \in \mathbb{R}^{k-1} : x_i \geq 0, \sum_{i=1}^k x_i = 1\}$ is the unknown parameter. Show that N is a complete sufficient statistic and find the UMVUE of $\theta_2 - \theta_1$.

PROBLEM 7.38. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$.

(1) Show that if μ in unknown and σ^2 is known, then $\bar{X}_n$ is the UMVUE of μ.
(2) Show that if μ is known and σ^2 is unknown, then $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is the UMVUE of σ^2.

PROBLEM 7.39. Let $X_1, ..., X_n$ be a sample from $\Gamma(p, \lambda)$, where $\theta = (p, \lambda)$ is the unknown parameter. Find the UMVUE of p/λ.

PROBLEM 7.40. Let $X \sim \text{Bin}_t(n, \theta)$ be a sample from *truncated* Binomial distribution:

$$p(k; \theta) = \frac{\binom{n}{k} \theta^k (1-\theta)^{n-k}}{1 - (1-\theta)^n}, \quad k \in \{1, ..., n\}.$$

(1) Show that X is a complete sufficient statistic for θ.
(2) Show that $\mathbb{E}_\theta X = n\frac{\theta}{1-(1-\theta)^n}$. Conclude that X/n is the UMVUE of $q(\theta) = \mathbb{E}_\theta X$.

PROBLEM 7.41. Let $X \sim \text{Bin}(n, \theta)$. Show that $\frac{X(n-X)}{n(n-1)}$ is the UMVUE of $\theta(1-\theta)$.

PROBLEM 7.42. Let $(\mathbb{P}_\theta)_{\theta \in \mathbb{N}}$ be the family of uniform distributions on the first θ integers $U(\{1, ..., \theta\})$.

(1) Show that $X \sim \mathbb{P}_\theta$ is a complete sufficient statistic
(2) Show that $2X - 1$ is the UMVUE of θ
(3) Let $(\mathbb{P}'_\theta)_{\theta \in \mathbb{N} \setminus k} = \{\mathbb{P}_{\theta \in \mathbb{N}}\} \setminus \mathbb{P}_k$, where k is a fixed integer. Show that $X' \sim \mathbb{P}'_\theta$ is not complete. **Hint:** Calculate $\mathbb{E}'_\theta g(X')$ for

$$g(i) = \begin{cases} 0, & i \neq k, k+1 \\ 1, & i = k \\ -1, & i = k+1 \end{cases}$$

(4) Show that $2X' - 1$ is not UMVUE of θ for $\mathbb{P}'_\theta$. **Hint:** Consider the estimator $T(X') = \begin{cases} 2X' - 1, & X' \notin \{k, k+1\} \\ 2k, & X' \in \{k, k+1\} \end{cases}$

PROBLEM 7.43. Let $X = (X_1, ..., X_n)$ be a sample from the $\beta(\theta, 1)$ p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} I\big(x \in (0,1)\big), \quad \theta > 0$$

and consider the estimator $T(X) = -\frac{1}{n}\sum_{i=1}^n \log X_i$.

(1) Show that $\mathbb{E}_\theta T(X) = 1/\theta$ and $\mathrm{var}_\theta(T) = \frac{1}{n\theta^2}$.
(2) Show that the Fisher information is given by $I(\theta) = 1/\theta^2$ and that $T(X)$ is the UMVUE of $1/\theta$.

PROBLEM 7.44. Solve the previous problem for the Weibull p.d.f.

$$f(x;\theta) = cx^{c-1}\theta e^{-\theta x^c} I(x > 0),$$

$\theta > 0, c > 0$ and the estimator $T(X) = \frac{1}{n}\sum_{i=1}^n X_i^c$.

PROBLEM 7.45 (Best Linear Unbiased Estimator). Let $X = A\theta + \varepsilon$, where $\theta \in \mathbb{R}^d$ is the unknown parameter, A is a known $n \times d$ matrix and ε is a random vector in $\mathbb{R}^n$ with $\mathbb{E}_\theta \varepsilon = 0$ and the covariance matrix $\Gamma := \mathrm{cov}(\varepsilon, \varepsilon) > 0$.

An estimator $T(X)$ is linear if $T(X) := BX + v$ for some $d \times n$ matrix B and a vector v.

(1) Show that any unbiased linear estimator of θ is of the form $T(X) = BX$, where $BA = I$
(2) Show that the MSE risk of an unbiased estimator is given by

$$\mathbb{E}_\theta \|\theta - T(X)\|^2 = \mathrm{tr}\left(B\Gamma B^\top\right)$$

(3) Assuming that Γ is an identity matrix, show that the linear estimator with the minimal MSE risk (the BLUE) is given by

$$T^*(X) := (A^\top A)^{-1} A^\top X$$

Hint: Note that for any B, satisfying $BA = I$,

$$BB^\top = (B - B^*)(B - B^*)^\top + (A^\top A)^{-1},$$

where $B^* = (A^\top A)^{-1} A^\top$.

(4) Derive the formula for the BLUE, when $\Gamma > 0$ is a diagonal matrix.

Asymptotic theory.

PROBLEM 7.46. Prove that if (ξ_n) converges to a constant c weakly, then it converges to c is probability.

PROBLEM 7.47. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d r.v. with the common distribution $U([0,1])$ and let $M_n = \max_{i\leq n} X_i$. Introduce the sequences

$$Y_n = \sqrt{n}(1 - M_n)$$
$$Z_n = n(1 - M_n)$$
$$W_n = n^2(1 - M_n).$$

Check whether each one of these sequences converges in probability ? in distribution ?

PROBLEM 7.48. Let $(X_n)_{n\geq 1}$ be a sequence or r.v. with the p.m.f.

$$\mathbb{P}(X_n = k) = \begin{cases} \frac{1}{2n} & k = n \\ \frac{1}{2n} & k = 2 \\ 1 - \frac{1}{n} & k = 3 \\ 0 & \text{otherwise} \end{cases}$$

(1) Does $(X_n)_{n\geq 1}$ converge in probability ? If yes, what is the limit ?
(2) Does $\mathbb{E}X_n$ converge ? Compare to your answer in (1)
(3) Answer (1) and (2) if "$k = n$" in the first line of definition of X_n is replaced with "$k = 1$"

PROBLEM 7.49. Prove that $X_n \sim \text{Bin}(n, p_n)$ with p_n satisfying

$$\lim_{n\to\infty} np_n = \lambda > 0$$

converges weakly to $X \sim \text{Poi}(\lambda)$.

Hint: check that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k and argue that this implies the result.

PROBLEM 7.50. Let $X_n \sim \text{Poi}(n)$, show that $(X_n - n)/\sqrt{n}$ converges weakly to $N(0, 1)$.

PROBLEM 7.51. Let $X_1, ..., X_n$ be a sample from $U([\theta - 1/2, \theta + 1/2])$, where $\theta \in \mathbb{R}$.

(1) Show that the MLE $\hat{\theta}_n$ of θ is not unique
(2) Show that any choice of $\hat{\theta}_n$ yields a consistent sequence of estimators

PROBLEM 7.52. A beuro of statistics wants to choose a sample, so that the empirical proportion of voters for a particular candidate will be less than 50% with probability 0.01, when the actual proportion is 52%. Suggest how to choose the sample size on the basis of CLT ?

PROBLEM 7.53. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Show that $S_n^2(X) := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is a consistent and asymptotically normal estimator of σ^2. Calculate its limit variance.

Hint: use the particular properties of the distribution of $S_n^2(X)$ and apply CLT.

PROBLEM 7.54. Let $X \sim \mathrm{HG}(N, N\theta, n)$, where $\theta \in \Theta = (0,1)$. Suppose that $n = \phi(N)$ for an increasing function ϕ (i.e. the size of the sample n grows with the population (shipment, etc.) size N). Find conditions on ϕ so that the sequence of estimators $\hat{\theta}_N(X) = X/n = X/\phi(N)$ is consistent.

PROBLEM 7.55. For the sample $X_1, ..., X_n$ from each one of the following distributions, find the sequence of MLEs as $n \to \infty$, show that it is consistent and find the asymptotic error distribution (under appropriate scaling)

(1) $\mathrm{Poi}(\theta)$, $\theta > 0$
(2) $\mathrm{Ber}(\theta)$, $\theta \in [0,1]$
(3) $\mathrm{Geo}(1/\theta)$, $\theta > 0$

PROBLEM 7.56. Construct confidence interval estimator for the parameter λ of the i.i.d. $\mathrm{Poi}(\lambda)$ sample $X_1, ..., X_n$

(1) Using the CLT and Slutsky's theorem
(2) Using the corresponding variance stabilizing transformation

PROBLEM 7.57 (MLE in exponential families). Let $X_1, ..., X_n$ be an i.i.d. sample from the density, which belongs to the 1-exponential family

$$f(x;\theta) = \exp\big(c(\theta)T(X) + d(\theta) + S(x)\big), \quad x \in \mathbb{R},\ \theta \in \Theta.$$

Assuming that $c(\theta)$ is a one-to-one function, the natural parametrization of the exponential family is defined by $\eta := c(\theta)$ with $\eta \in c(\Theta)$:

$$\tilde{f}(x;\eta) = \exp\big(\eta T(X) + \tilde{d}(\eta) + S(x)\big), \quad x \in \mathbb{R},\ \eta \in c(\Theta),$$

where $\tilde{d}(\eta) := d(c^{-1}(\eta))$.

(1) Show that $\tilde{d}'(\eta) = -\mathbb{E}_\eta T(X_1)$ and $\tilde{d}''(\theta) = -\mathrm{var}_\eta(T(X_1))$.
(2) Show that the MLE $\hat{\eta}_n$ is the unique root of the equation

$$d'(\eta) = -\frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

(3) Using the LLN, conclude that the MLE $(\hat{\eta}_n)$ is consistent
(4) Show that the MLE $\hat{\theta}_n$ of θ is consistent

Hint: recall that the MLE is invariant under reparametrization

(5) Apply the Δ-method to show that the MLE $(\hat{\eta}_n)$ is asymptotically normal with the rate $\sqrt{n}$ and find the corresponding asymptotic variance (compare with the Fisher information).
(6) Apply the Δ-method once again to derive the asymptotic variance of the MLE $(\hat{\theta}_n)$.

PROBLEM 7.58. Let $X^n = (X_1, ..., X_n)$ be a sample from the Laplace density

$$f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad x \in \mathbb{R}$$

(1) Argue that the statistics $\bar{X}_n$ form a consistent sequence of estimators for the location parameter θ

(2) Show that the MLE of θ can be taken to be the sample median $\text{med}(X^n) := X_{(\lceil n/2 \rceil)}$ where $\lceil n/2 \rceil$ is the smallest integer greater or equal $n/2$ and $X_{(1)}, ..., X_{(n)}$ is the order statistic of X^n.

Hint: note that the MLE is not unique for even n.

(3) Prove consistency of $\text{med}(X^n)$

Hint: note that e.g.

$$\left\{\text{med}(X^n) \geq \theta + \varepsilon\right\} = \left\{\sum_{i=1}^{n} \mathbf{1}_{\{X_i - \theta \geq \varepsilon\}} \geq n/2\right\}.$$

The Cauchy density with the location parameter $\theta \in \mathbb{R}$ and the scaling parameter $\gamma > 0$ is

$$f(x; \theta, \gamma) = \frac{1}{\gamma \pi} \frac{1}{1 + \left(\frac{x-\theta}{\gamma}\right)^2}$$

(4) Let $X^n = (X_1, ..., X_n)$ be a sample from $\text{Cau}(\theta, \gamma)$ density with the unknown location parameter θ and $\gamma = 1$. Is $\bar{X}_n$ consistent ? Is $\text{med}(X^n)$ consistent ?

Hint: It can be shown that the characteristic function of the Cauchy random variable is given by:

$$\varphi(t; \theta, \gamma) = \int_{\mathbb{R}} e^{itx} f(x; \theta, \gamma) dx = e^{it\theta - \gamma|t|}.$$

(5) Suggest yet another consistent estimator of θ, using the substitution principle method.

CHAPTER 8

Hypothesis testing

In contrast to parameter estimation (either point or interval), which deals with guessing the *value* of the unknown parameter, we are often interested to know only whether or not the parameter lies in a specific region of interest in the parametric space. A typical instance of such a problem is *signal detection*, frequently arising in electrical engineering. A radar transmitter sends a signal in a particular direction: if the signal encounters an object on its way, the echo of the signal is returned to the radar receiver, otherwise the receiver picks up only noise. Hence the receiver has to decide whether the received transmission contains a signal or it consists only of the background noise. How to formalize this problem mathematically ? How to construct reasonable test procedures ? How to compare different test procedures ? Is there a best procedure ? All these questions are addressed within the statistical framework of hypothesis testing, which we shall explore below.

a. The setting and terminology

Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a statistical model and suppose $\Theta = \Theta_0 \cup \Theta_1$, where the subsets Θ_0 and Θ_1 do not intersect. The value of the parameter θ is unknown and our goal is to decide whether θ belongs to Θ_0 or to Θ_1, given a sample $X \sim \mathbb{P}_\theta$. Using the statistical language, we want to test the *null hypothesis* $H_0 : \theta \in \Theta_0$ against the *alternative* $H_1 : \theta \in \Theta_1$, based on the sample $X \sim \mathbb{P}_\theta$. If Θ_0 (or/and Θ_1) consists of a single point, the null hypothesis (or/and the alternative) is called *simple*, otherwise it is referred as *composite*.

EXAMPLE 8a1. In the signal detection problem[1], a reasonable statistical model is an i.i.d. sample $X = (X_1, ..., X_n)$ from $N(\theta, \sigma^2)$, where σ^2 is the known intensity (variance) of the noise and θ is the unknown parameter. The null hypothesis is then $H_0 : \theta = 0$, i.e. Θ_0 consists of a single point $\{0\}$ and we want to test it against the alternative $H_1 : \theta \neq 0$, i.e. $\Theta_1 = \Theta \setminus \{0\}$. Thus we want to test a simple hypothesis against composite alternative.

If we know that the signal may have only positive (but still unknown) value, then we may consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. If, in addition, this value is known, say θ_0, then the problem is to test $H_0 : \theta = 0$ against $H_1 : \theta = \theta_0$, i.e. testing a simple hypothesis against a simple alternative. ■

REMARK 8a2. While technically the roles of the null hypothesis and the alternative are symmetric, it is customary to think about H_0 as some "usual" theory/state/etc., which we want to reject in favor of the alternative theory/state.

[1] it is convenient to think of this problem to memorize the terminology: e.g. null hypothesis can be thought of as absence of the signal, etc.

Given a sample $X \sim \mathbb{P}_\theta$ we have to *reject* or *accept* the null hypothesis. Suppose for definiteness that the sample X takes values in $\mathbb{R}^n$. Then any[2] test can be defined by specifying the *region of rejection* (or *critical* region), i.e. a set $C \subset \mathbb{R}^n$ such that the null hypothesis H_0 is rejected if and only if the event $\{X \in C\}$ occurs. The complement $\mathbb{R}^n \setminus C$ is called the *region of acceptance* (of the null hypothesis). This can be equivalently reformulated in terms of the *critical function* (or *test function*) $\delta(X) := I(X \in C)$, i.e H_0 is rejected if and only if $\{\delta(X) = 1\}$ occurs. Usually the test function can be put in the form $\delta(X) = I\big(T(X) \geq c\big)$, where $T(X)$ is the *test statistic*, i.e. a function from $\mathbb{R}^n$ to $\mathbb{R}$, depending only on the sample, and c is a real constant, called the *critical value*.

EXAMPLE 8a1 (continued) If we want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, intuitively we feel that we shall reject the null hypothesis H_0 if the empirical mean is far from zero: in terms of the objects defined above we use the test statistic $T(X) = \bar{X}_n$ and reject H_0 if and only if $\{|\bar{X}_n| \geq c\}$, where c is the critical value to be chosen (see below). This kind of tests are called *two-sided* (or *two-tailed*) as they reject if the test statistic takes values on "both sides" with respect to the null hypothesis.

Now suppose that we know that the signal value is positive, i.e. $H_0 : \theta = 0$ is to be tested against $H_1 : \theta > 0$. In this case, the *one-sided* (*one-tailed*) test $\{\bar{X}_n \geq c\}$ appears as a reasonable choice, as it rejects H_0 for large enough values of $\bar{X}_n$. ■

How do we measure performance of a test? Note that a test can produce two types of errors, namely, we may either reject H_0, when actually H_0 is true: this is the so called *type I error* (or *α-error* or the *false alarm* error or *false positive* error); or we may accept H_0, when H_1 is true: this is the *type II error* (or *β-error* or *detection* error or *false negative* error). Both types of errors are conveniently expressed in terms of the *power* function:

$$\pi(\theta, \delta) := \mathbb{E}_\theta \delta(X) = \mathbb{P}_\theta(T(X) \geq c).$$

Note that for $\theta \in \Theta_0$, the power function $\pi(\theta, \delta)$ is the α-error of the test (at a specific θ) and for $\theta \in \Theta_1$ is the probability of correct acceptance of the particular alternative θ. The α-error of a test is defined as the highest probability over Θ_0 of erroneously rejecting H_0:

$$\alpha := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq c) = \sup_{\theta \in \Theta_0} \pi(\theta, \delta), \tag{8a1}$$

and the β-error is the highest probability over Θ_1 of erroneously accepting H_0:

$$\beta := \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(T(X) < c) = 1 - \inf_{\theta \in \Theta_1} \pi(\theta, \delta).$$

Let's try to get a feeling of how c affects these three quantities (it may be helpful to think of the signal detection again). For large critical values c, we shall rarely reject H_0 and, in particular, shall rarely reject erroneously, and consequently shall get small α-error. On the other hand, for exactly the same reason we shall get little power at $\theta \in \Theta_1$ (and hence large β-error). Conversely, small c will cause frequent rejections and hence higher false alarms, however will yield higher

[2] in fact, we may also consider randomized tests, i.e. those in which the decision is taken not only on the basis of the sample realization, but also on an auxiliary randomness. More specifically, $\delta(X)$ is allowed to take values in the interval $[0, 1]$ (rather than in the set of two points $\{0, 1\}$) and a coin with probability of heads $\delta(X)$ is tossed: H_0 is rejected if the coin comes up heads. It turns out that this general framework is more flexible and convenient in certain situations (see Remark 8b6)

values of power at $\theta \in \Theta_1$. This tradeoff is the key to the optimality theory of tests: we would like to make small false alarm errors and simultaneously have large power at the alternatives (or, equivalently, to make small errors of both types simultaneously).

If a test erroneously rejects H_0 with probability less than $\alpha > 0$ for all $\theta \in \Theta_0$, then it is said to have the *level of significance* α. Clearly a test of level α is also of level α' is $\alpha' > \alpha$. The smallest level of the test is referred to as its *size* (which is nothing but the definition (8a1)).

EXAMPLE 8a1 (continued) Let's calculate the power function of the test $\delta(X) = \{\bar{X}_n \geq c\}$ of $H_0 : \theta = 0$ against $H_1 : \theta > 0$:

$$\pi(\theta, \delta) = \mathbb{P}_\theta\big(\bar{X}_n \geq c\big) = \mathbb{P}_\theta\Big(\sqrt{n}(\bar{X}_n - \theta)/\sigma \geq \sqrt{n}(c-\theta)/\sigma\Big) = 1 - \Phi\Big(\sqrt{n}(c-\theta)/\sigma\Big),$$

where Φ is the c.d.f of a standard Gaussian r.v. and we used the fact that $\xi := \sqrt{n}(\bar{X}_n - \theta)/\sigma \sim N(0,1)$. Note that $\pi(\theta, \delta)$ is an increasing function of θ (why ?). The size of the test is

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta, \delta) = \pi(0, \delta) = 1 - \Phi\big(\sqrt{n}c/\sigma\big).$$

Hence if we want our test to have the size α, we shall choose

$$c(\alpha) = \sigma \Phi^{-1}(1-\alpha)/\sqrt{n}.$$

For larger values of n, $c(\alpha)$ gets smaller, which agrees with the fact that $\bar{X}_n$ is closer to 0 under H_0. For the critical value corresponding to the size α, the power function reads

$$\pi(\theta, \delta) = 1 - \Phi\Big(\sqrt{n}(\sigma\Phi^{-1}(1-\alpha)/\sqrt{n} - \theta)/\sigma\Big) = 1 - \Phi\Big(\Phi^{-1}(1-\alpha) - \sqrt{n}\theta/\sigma)\Big), \qquad (8a2)$$

Pay attention (see Figure 1) that for small values of the alternative $\theta \in \Theta_1$, the power is close to α (which is of course anticipated as the power function is continuous in this case). On the other hand, for any $\theta \in \Theta_1$, the power tends to 1 as $n \to \infty$, i.e. the test makes small errors of both types. In particular, we can choose n large enough to force arbitrarily small β-errors at any value of θ from the alternative. ■

REMARK 8a3. This example shows that the power of the test is close to α at θ's close to the null hypothesis, which of course confirms the fact that close values of the hypothesis and alternatives are hard to distinguish. Hence often the practical requirement is formulated in terms of the size of the test α and the minimal power of the test outside some *indifference* region of length Δ, where good testing quality cannot be expected. For instance, in the previous example one can require level $\alpha = 0.01$ and minimal power 0.9 outside the indifference region $[0, 1]$ (i.e. $\Delta = 1$). The corresponding n is easily found, using monotonicity of $\pi(\theta, \delta)$ and the formula (8a2).

Another practically useful notion is *the p-value* of the test. Suppose a simple null hypotheses is tested by means of a test statistic T. The p value is defined as the probability of getting the values of the test statistic more extreme than the actually observed one:

$$p(t) := \mathbb{P}_{\theta_0}(T \geq t), \quad t \in \mathbb{R}.$$

Large p-value indicates that the observed value of the test statistic is typical under the null hypothesis, which is thus to be accepted. If T has a continuous increasing c.d.f., then the random variable $p(T)$ has uniform distribution, irrespectively of $\mathbb{P}_{\theta_0}$ (think why). Hence the

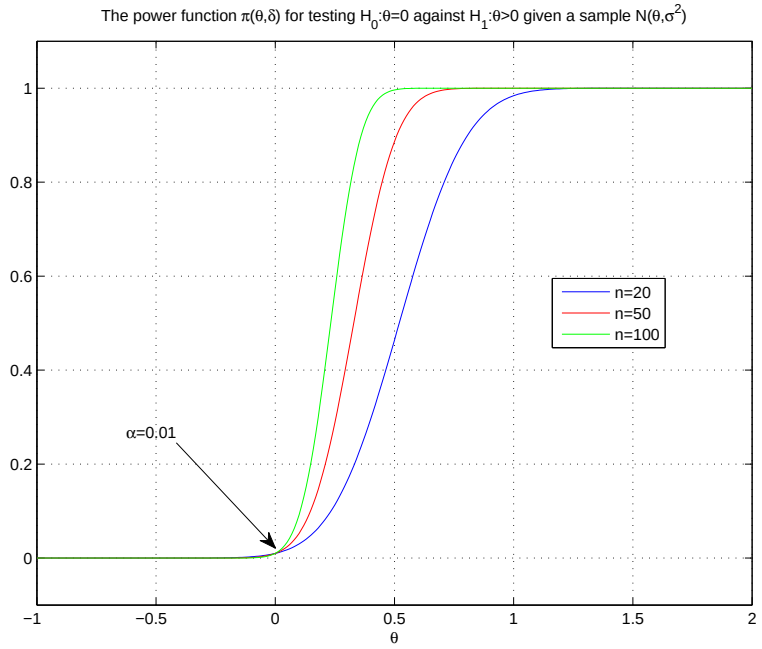

FIGURE 1. $\pi(\theta, \delta)$ for the test in Example 8a1

test, which rejects the null hypotheses if and only if the p-value is less than $\alpha \in (0, 1)$, has the significance level α.

The p-value formalism of the statistical hypothesis testing is preferred by some practitioners, since in addition to acceptance/rejection decision, it provides a quantitative information of how firmly the null hypothesis is supported.

Here is an example with a more practical flavor:

EXAMPLE 8a4. Suppose we toss a coin n times and would like to test the hypothesis that the coin is fair. Assuming the i.i.d. Ber(θ) model, we are faced with the problem of testing $H_0 : \theta = 1/2$ against $H_1 : \theta \neq 1/2$ on the basis of the outcome of n tosses $X = (X_1, ..., X_n)$. Since the empirical mean $\bar{X}_n$ is close to the actual value of θ (at least for large n's), the two-sided test $\delta(X) = \{|\bar{X}_n - 1/2| \geq c\}$ appears reasonable. The power function of this test is given by:

$$\pi(\theta, \delta) = \mathbb{P}_\theta\big(|\bar{X}_n - 1/2| \geq c\big) = \sum_{k:|k/n-1/2|\geq c} \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

The critical value of the test of level α is the minimal c satisfying the inequality

$$\pi(1/2, \delta) \leq \alpha \quad \Longrightarrow \quad \sum_{k:|k/n-1/2|\geq c} \binom{n}{k} \leq 2^n \alpha,$$

which can be found numerically. For large n, such calculation can be difficult and, alternatively, one can use the CLT to construct an approximate test. Assuming that n is large enough (to justify the approximation) and adding the subscript n to c to emphasize its dependence on n,

we get:

$$\pi(1/2,\delta)=\mathbb{P}_{1/2}\big(|\bar X_n-1/2|\ge c_n\big)=\mathbb{P}_{1/2}\Big(\Big|\sqrt{n}\frac{\bar X_n-1/2}{1/2}\Big|\ge 2c_n\sqrt{n}\Big)\xrightarrow{n\to\infty}2\Phi\big(-2z\big),$$

where the convergence holds, if $c_n:=z/\sqrt{n}$ is chosen with a constant z. Now, we shall choose z to meet the size requirement, i.e. $2\Phi\big(-2z\big)=\alpha$ or $z:=-\frac{1}{2}\Phi^{-1}(\alpha/2)$. To recap, the critical value

$$c_n(\alpha)=-\frac{1}{2\sqrt{n}}\Phi^{-1}(\alpha/2),$$

yields a test, whose size is approximately α. The approximation is justified on the basis of CLT, if we believe that n is large enough to regard the limit as a valid approximation for a fixed n.

In fact, the approximation can be treated more rigorously using the following refinement of the CLT:

THEOREM 8a5 (Berry-Esseen). *Let (ξ_i) be a sequence of i.i.d. random variables with zero mean and unit variance. Then the c.d.f. $F_n(x)=\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\le x\right)$ satisfies*

$$\big|F_n(x)-\Phi(x)\big|\le\frac{Cm_3}{\sqrt{n}},\quad x\in\mathbb{R},\ n\ge 1,$$

where $m_3:=\mathbb{E}|\xi_1|^3$ and C is an absolute constant with value less than 0.4784

For the problem at hand $\xi_i:=\dfrac{X_i-\mathbb{E}_{1/2}X_i}{\sqrt{\mathrm{var}_{1/2}(X_i)}}=2(X_i-1/2)$ and with $c_n:=z/\sqrt{n}$,

$$\mathbb{P}_{1/2}\Big(\Big|\sqrt{n}\frac{\bar X_n-1/2}{1/2}\Big|\ge 2c_n\sqrt{n}\Big)=\mathbb{P}_{1/2}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\Big|\ge 2z\Big)=1-F_n(2z)+F_n(-2z).$$

By the B-E theorem,

$$|F_n(\pm 2z)-\Phi(\pm 2z)|\le\frac{C\mathbb{E}|\xi_1|^3}{\sqrt{n}}=\frac{C}{\sqrt{n}},$$

and for $z^*:=-\frac{1}{2}\Phi^{-1}(\alpha/2)$

$$\begin{aligned}&\Big|\mathbb{P}_{1/2}\Big(\Big|\sqrt{n}\frac{\bar X_n-1/2}{1/2}\Big|\ge 2z^*\Big)-\alpha\Big|\le\\&\Big|1-F_n(2z^*)-\frac{1}{2}\alpha\Big|+\Big|F_n(-2z^*)-\frac{1}{2}\alpha\Big|=\\&\Big|1-F_n(2z^*)-\Big(1-\Phi(2z^*)\Big)\Big|+\Big|F_n(-2z^*)-\Phi\big(-2z^*\big)\Big|\le 2Cn^{-1/2}\le\frac{1}{\sqrt{n}}.\end{aligned}$$

Let us stress that the latter inequality holds for any n and not just asymptotically. Hence for a given α, one can choose an n, which yields the approximation of the required quality.

Application of the limit theorems must be done with care. For example, suppose that we want to calculate the power of the obtained test at a particular value of alternative (e.g. at an end point of an indifference region, say at $\theta_1:=3/4$):

$$\pi(3/4,\delta)=\mathbb{P}_{3/4}\big(|\bar X_n-1/2|\ge z^*/\sqrt{n}\big).$$

Note that this time the critical value is already fixed. Let's try to implement the same approach as above:

$$\mathbb{P}_{3/4}\big(|\bar{X}_n - 1/2| \geq z^*/\sqrt{n}\big) = \mathbb{P}_{3/4}\left(\sqrt{n}\frac{\bar{X}_n - 3/4}{\sqrt{3}/4} \geq \sqrt{n}\frac{z^*/\sqrt{n} - 1/4}{\sqrt{3}/4}\right) +$$
$$\mathbb{P}_{3/4}\left(\sqrt{n}\frac{\bar{X}_n - 3/4}{\sqrt{3}/4} \leq -\sqrt{n}\frac{z^*/\sqrt{n} + 1/4}{\sqrt{3}/4}\right) = \mathbb{P}_{3/4}\left(\sqrt{n}\frac{\bar{X}_n - 3/4}{\sqrt{3}/4} \geq \frac{z^* - 1/4\sqrt{n}}{\sqrt{3}/4}\right) +$$
$$\mathbb{P}_{3/4}\left(\sqrt{n}\frac{\bar{X}_n - 3/4}{\sqrt{3}/4} \leq -\frac{z^* + 1/4\sqrt{n}}{\sqrt{3}/4}\right) = 1 - F_n\left(\frac{z^* - 1/4\sqrt{n}}{\sqrt{3}/4}\right) + F_n\left(-\frac{z^* + 1/4\sqrt{n}}{\sqrt{3}/4}\right).$$

Note that $F_n(\cdot)$ in the latter expression is evaluated at points, which themselves depend on n and hence CLT is not informative (or applicable) anymore: this expression converges[3] to 1 as $n \to \infty$. In particular, pretending that e.g. $F_n\left(\frac{z^*-1/4\sqrt{n}}{\sqrt{3}/4}\right) \approx \Phi\left(\frac{z^*-1/4\sqrt{n}}{\sqrt{3}/4}\right)$ cannot be easily justified (though often done in practice).

Finally, let's demonstrate how p-value is calculated for this problem. Suppose that we toss the coin 50 times and obtain 35 heads. The p-value is then given by:

$$\mathbb{P}_{1/2}(|\bar{X}_{50} - 1/2| \geq |35/50 - 1/2|) = \mathbb{P}_{1/2}(|S_{50} - 25| \geq 10) = 2\sum_{k \geq 35}\binom{50}{k}(1/2)^{50} = 0.0066...$$

which indicates that the obtained sample poorly supports the hypothesis H_0 (or in other words this outcome is not typical for a fair coin). ■

b. Comparison of tests and optimality

From the decision theoretic point of view, it makes sense to compare tests of the same size:

Definition 8b1. *Let δ and $\tilde{\delta}$ be tests of size α. The test δ is more powerful than $\tilde{\delta}$, if*

$$\pi(\theta, \delta) \geq \pi(\theta, \tilde{\delta}), \quad \forall \theta \in \Theta_1.$$

Definition 8b2. *A test δ^* of size α is uniformly most powerful (UMP) if it is more powerful than any other test δ of size α.*

As in the case of the point estimators, two tests of size α do not have to be comparable in the sense of the latter definition and the UMP test does not have to exist. When both the hypothesis and the alternative are simple, the UMP[4] test can be found explicitly:

Theorem 8b3 (Neyman-Pearson lemma). *Consider the statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ with $\Theta = \{\theta_0, \theta_1\}$, $X \sim \mathbb{P}_\theta$ and let $L(x;\theta_0)$ and $L(x;\theta_1)$ be the corresponding likelihood functions. Then the likelihood ratio test[5]:*

$$\delta^*(X) = I\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} \geq c(\alpha)\right),$$

of size $\alpha < 1$ is MP.

---

[3] again, by the Berry-Esseen theorem (check!)

[4] in fact it is MP (most powerful) test, since the alternative is simple

[5] A better way to define the test is

$$\delta^*(X) = I\Big(L(X;\theta_1) \geq c(\alpha)L(X;\theta_0)\Big),$$

REMARK 8b4. The test is not counterintuitive: if the sample X comes from $\mathbb{P}_{\theta_0}$, then $L(X;\theta_0)$ will be typically greater than $L(X;\theta_1)$ and hence the test statistic in δ^* is small, i.e. H_0 is rarely rejected. This also provides the heuristic grounds for the so called generalized likelihood ratio test, which is the main tool in more general hypothesis testing problems (see Section c below).

REMARK 8b5. Note that if $S(X)$ is the minimal sufficient statistic for $\theta \in \Theta$, then by the F-N factorization theorem the likelihood ratio test depends on the sample only through $S(X)$.

PROOF. We shall give the proof for the continuous case, when the likelihoods are in fact p.d.f.'s on $\mathbb{R}^n$ (the discrete case is treated similarly). We have to show that if δ is a test of level α, i.e. $\mathbb{E}_{\theta_0}\delta(X) \le \alpha$, then

$$\mathbb{E}_{\theta_1}\delta^*(X) \ge \mathbb{E}_{\theta_1}\delta(X). \tag{8b1}$$

To this end

$$\begin{aligned}
&\mathbb{E}_{\theta_1}\delta(X) = \int_{\mathbb{R}^n} \delta(x) f(x;\theta_1)dx = \\
&c(\alpha)\int_{\mathbb{R}^n} \delta(x) f(x;\theta_0)dx + \int_{\mathbb{R}^n} \delta(x)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx = \\
&c(\alpha)\mathbb{E}_{\theta_0}\delta(X) + \int_{\mathbb{R}^n} \delta(x)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx.
\end{aligned} \tag{8b2}$$

Further,

$$\begin{aligned}
&\int_{\mathbb{R}^n} \delta(x)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx \overset{\dagger}{\le} \\
&\int_{\mathbb{R}^n} \delta(x) I\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0) \ge 0\Big)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx \overset{\ddagger}{\le} \\
&\int_{\mathbb{R}^n} I\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0) \ge 0\Big)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx = \\
&\int_{\mathbb{R}^n} \delta^*(x)\Big(f(x;\theta_1) - c(\alpha)f(x;\theta_0)\Big)dx = \mathbb{E}_{\theta_1}\delta^*(X) - c(\alpha)\mathbb{E}_{\theta_0}\delta^*(X) = \\
&\mathbb{E}_{\theta_1}\delta^*(X) - c(\alpha)\alpha,
\end{aligned}$$

where † holds since $\delta(x) \ge 0$ and ‡ is true as $\delta(x) \le 1$. Plugging the latter inequality into (8b2) yields

$$\mathbb{E}_{\theta_1}\delta(X) \le c(\alpha)\Big(\mathbb{E}_{\theta_0}\delta(X) - \alpha\Big) + \mathbb{E}_{\theta_1}\delta^*(X) \le \mathbb{E}_{\theta_1}\delta^*(X),$$

where the last inequality holds, since $c(\alpha) \ge 0$ (otherwise $\alpha = 1$) and $\mathbb{E}_{\theta_0}\delta(X) \le \alpha$. □

REMARK 8b6. If the likelihood ratio does not have a continuous distribution, as will typically be the case if X is discrete, then it might be impossible to find the critical value c, so that the N-P test has the precise level α. In this case, the UMP test exists in the more general class of randomized test. In practice, if a randomized test is undesirable, one can switch to the largest achievable size less than α.

---

which does not run into the problem of division by zero (in case, the likelihoods do not have the same support)

EXAMPLE 8b7. Suppose that we observe the outcome of n independent tosses of one of two coins with the head probabilities $0 < \theta_0 < \theta_1 < 1$ respectively. We want to decide which coin was actually tossed, i.e. test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ given the sample $X = (X_1, ..., X_n)$.

The likelihood ratio statistic is

$$\frac{L(X;\theta_1)}{L(X;\theta_0)} = \left(\frac{\theta_1}{\theta_0}\right)^{S_n(x)} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-S_n(x)}$$

where $S_n(X) = \sum_{i=1}^n X_i$, and the N-P test rejects H_0 if and only if

$$\left(\frac{\theta_1}{\theta_0}\right)^{S_n(x)} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-S_n(x)} \geq c,$$

or, equivalently,

$$\left(\frac{1/\theta_0 - 1}{1/\theta_1 - 1}\right)^{S_n(x)} \geq c \left(\frac{1-\theta_0}{1-\theta_1}\right)^n.$$

Since $\theta_0 < \theta_1$, the N-P test is

$$\delta^*(X) = \big\{S_n(X) \geq c'\big\},$$

where the critical value c' is to be chosen to match the desired level α. The power function of this test is

$$\pi(\theta, \delta^*) = \sum_{k \geq c'} \binom{n}{k} \theta^k (1-\theta)^{n-k}, \quad \theta \in \{\theta_0, \theta_1\}.$$

If α is in the range of $\pi(\theta_0, \delta^*)$ considered as a function of c', then the corresponding critical value is the unique root of the equation

$$\sum_{k \geq c^*(\alpha)} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k} = \alpha,$$

which can be found either numerically. The MP test of level α is now

$$\delta^*(X) = \big\{S_n(X) \geq c^*(\alpha)\big\}. \tag{8b3}$$

■

EXAMPLE 8b8. Consider the signal detection problem, when we know the exact value of the signal to be $\theta_1 > 0$. We observe n i.i.d. samples from $N(\theta, \sigma^2)$, $\theta \in \Theta = \{0, \theta_1\}$ where $\sigma^2 > 0$ is known and would like to decide whether the signal is present. Put into the above framework, we are faced with the problem of testing $H_0 : \theta = 0$ against the alternative $H_1 : \theta = \theta_1$. The corresponding likelihood is

$$L_n(x;\theta) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\Big), \quad x \in \mathbb{R}^n, \theta \in \Theta \tag{8b4}$$

and the likelihood ratio test statistic is

$$\frac{L_n(X;\theta_1)}{L_n(X;0)} = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right) = \exp\left(\frac{\theta_1}{\sigma^2}\sum_{i=1}^n X_i - n/2\right),$$

and the N-P test rejects H_0 (i.e. decides that there is a signal) if and only if

$$\exp\left(\frac{\theta_1}{\sigma^2}\sum_{i=1}^{n} X_i - n/2\right) \geq c,$$

for a critical value c to be chosen to meet the level specification. Note that the above is equivalent to the test suggested in Example 8a1 on the intuitive grounds:

$$\bar{X}_n \geq c',$$

where c' is related to c by a one-to-one correspondence, which is not of immediate interest to us, since we shall choose c' directly to get a test of level α. To this end, note that under H_0, $\bar{X}_n \sim N(0, \sigma^2/n)$ and hence the level of the test is given by

$$\mathbb{P}_0\Big(\bar{X}_n \geq c'\Big) = \mathbb{P}_0\Big(\sqrt{n}\bar{X}_n/\sigma \geq \sqrt{n}c'/\sigma\Big) = 1 - \Phi\big(\sqrt{n}c'/\sigma\big) = \alpha,$$

which gives $c' = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)$. Thus the MP test of level α is then

$$\delta^*(X) = \left\{\bar{X}_n \geq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)\right\}. \tag{8b5}$$

The power of this test at θ_1 is given by

$$\begin{aligned}\pi(\theta_1, \delta^*) =& \mathbb{P}_{\theta_1}\Big(\bar{X}_n \geq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)\Big) = \mathbb{P}_{\theta_1}\Big(\underbrace{\sqrt{n}\frac{\bar{X}_n - \theta_1}{\sigma}}_{\sim N(0,1)} \geq \Phi^{-1}(1-\alpha) - \frac{\theta_1}{\sigma/\sqrt{n}}\Big) = \\ & 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\theta_1}{\sigma/\sqrt{n}}\right).\end{aligned} \tag{8b6}$$

As n increases, the power at θ_1 approaches 1, while the level remains α. ■

Note that the test (8b5) is most powerful against *any* simple alternative $\theta_1 > 0$ and does not itself depend on θ_1. Hence we can use δ^* to test $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1$, without knowing the value of θ_1! Let $\delta(X)$ be an arbitrary level α test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$, i.e. $\mathbb{E}_0\delta(X) \leq \alpha$. Then by the N-P lemma for any fixed $\theta_1 \in \Theta_1$,

$$\pi(\theta_1, \delta^*) \geq \pi(\theta_1, \delta),$$

and since θ_1 is arbitrary alternative, we conclude that δ^* is UMP:

$$\pi(\theta, \delta^*) \geq \pi(\theta, \delta), \quad \forall \theta \in \Theta_1. \tag{8b7}$$

Of course, independence of the test δ^* of the alternative parameter value θ_1 was crucial and, in fact, in general will not be the case.

Furthermore, note that the power function (8b6) of δ^* is continuous and monotonically increasing in θ_1. Consider δ^* as a test for the problem

$$\begin{aligned} H_0 &: \theta \leq 0 \\ H_1 &: \theta > 0 \end{aligned} \tag{8b8}$$

Clearly, δ^* has size α in this problem

$$\sup_{\theta \leq 0} \pi(\theta, \delta^*) = \pi(0, \delta^*) = \alpha.$$

Now let δ be another test of size α, i.e. $\sup_{\theta\leq 0}\pi(\theta,\delta)\leq\alpha$, and suppose that

$$\pi(\theta',\delta)>\pi(\theta',\delta^*) \tag{8b9}$$

for some $\theta'\in\Theta_1$. Consider the tests δ and δ^* for testing the simple problem

$$\begin{aligned} H_0&:\theta=0\\ H_1&:\theta=\theta'. \end{aligned}$$

Again by N-P lemma δ^* is MP in this problem, which contradicts (8b9), since $\pi(0,\delta)\leq\alpha$. Thus we conclude that δ^* is the UMP test for the more complex problem (8b8).

To recap, remarkably the N-P likelihood ratio test, which, at the outset, is optimal only for testing simple hypothesis versus simple alternative, is in fact UMP test in the two latter examples. It turns out that this is a somewhat more general phenomena:

Definition 8b9. *A family of probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$, $\Theta\subseteq\mathbb{R}$ with the likelihood $L(x;\theta)$ is said to be a monotone likelihood ratio family in statistic $T(X)$, if for any $\theta_0<\theta_1$, $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_1}$ are distinct and $L(x;\theta_1)/L(x;\theta_0)$ is a strictly increasing function of $T(x)$.*

Theorem 8b10 (Karlin-Rubin). *If $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a monotone likelihood ratio family with respect to $T(X)$ and $T(X)$ has a continuous distribution under $\mathbb{P}_{\theta_0}$, then $T(X)$ is the optimal test statistic*[6] *for testing*

$$\begin{aligned} H_0&:\theta\leq\theta_0\\ H_1&:\theta>\theta_0 \end{aligned} \tag{8b10}$$

and the corresponding power function is increasing.

Remark 8b11. 'Strictly increasing' in the above definition can be replaced with nondecreasing and continuity of the distribution in the Theorem 8b10 can be omitted, if randomized tests are allowed, or alternatively, some levels are forbidden.

Proof. By the assumption

$$R(x;\theta_1,\theta_0):=L(x;\theta_1)/L(x;\theta_0)=\phi(T(x);\theta_1,\theta_0),$$

where $t\mapsto\phi(t,\theta_1,\theta_0)$ is a strictly increasing function for any $\theta_1>\theta_0$. Hence for any c in the range of ϕ

$$\left\{\frac{L(x;\theta_1)}{L(x;\theta_0)}\geq c\right\}=\Big\{T(X)\geq\phi^{-1}(c;\theta_1,\theta_0)\Big\}.$$

In particular, this is true for the critical value $c_\alpha(\theta_0,\theta_1)$, which yields the level α at θ_0, i.e.

$$\mathbb{P}_{\theta_0}\left(\frac{L(x;\theta_1)}{L(x;\theta_0)}\geq c_\alpha(\theta_1,\theta_0)\right)=\alpha,$$

and it follows that

$$\mathbb{P}_{\theta_0}\Big(T(X)\geq\phi^{-1}\big(c_\alpha(\theta_1,\theta_0);\theta_1,\theta_0\big)\Big)=\alpha.$$

[6] i.e. the α level $\{T(X)\geq c(\alpha)\}$ is UMP

Since $T(X)$ has a continuous distribution under $\mathbb{P}_{\theta_0}$, the latter can be solved for $c'_\alpha(\theta_0) := \phi^{-1}\big(c_\alpha(\theta_1,\theta_0);\theta_1,\theta_0\big)$, which does not depend on θ_1. Hence we conclude that the level α likelihood ratio test for the simple problem

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

with $\theta_1 > \theta_0$ has the form $\{T(X) \geq c'_\alpha\}$, where c'_α does not depend on θ_1.

We shall prove shortly that the power function of this test $\pi(\theta,\delta^*) = \mathbb{E}_\theta\delta^*(X)$ is strictly increasing in θ. Then $\sup_{\theta\leq\theta_0}\pi(\theta,\delta^*) = \pi(\theta_0,\delta^*) = \alpha$, i.e. δ^* is a level α test for the problem (8b10).

Let's repeat the arguments from the discussion, preceding Definition 8b9, to show that δ^* is in fact UMP for the problem (8b10). Suppose that this is not the case and there is a test δ with $\sup_{\theta\leq\theta_0}\pi(\theta,\delta) \leq \alpha$, such that

$$\pi(\theta',\delta) > \pi(\theta',\delta^*), \tag{8b11}$$

for some $\theta' > \theta_0$. Now consider the problem of testing two simple hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta'.$$

Note that $\pi(\theta_0,\delta) \leq \alpha$ and hence (8b11) contradicts the statement of the N-P lemma, which shows that δ^* is UMP for the problem (8b10), as claimed.

To complete the proof, it remains to check that the power function $\pi(\theta,\delta^*)$ is strictly increasing in θ. To this end, we shall show that for an nondecreasing function ψ, the function $\theta \mapsto \mathbb{E}_\theta\psi\big(T(X)\big)$ is increasing: in particular, with[7] $\psi(u) := I(u \geq c(\alpha))$ this implies that the power $\pi(\theta,\delta^*) = \mathbb{E}_\theta I(T(x) \geq c(\alpha))$ increases in θ.

[7] $I(u \geq c(\alpha))$ is a nondecreasing function of u: it jumps from 0 to 1 as u increases

Recall that $L(x;\theta) = f(x;\theta)$ and let $\theta' > \theta$, then

$$\mathbb{E}_{\theta'}\psi(T(X)) - \mathbb{E}_{\theta}\psi(T(X)) = \int_{\mathbb{R}^n} \psi(T(x))\big(f(x;\theta') - f(x;\theta)\big)dx =$$

$$\int_{\mathbb{R}^n} \psi(T(x))\Big(\underbrace{\frac{f(x;\theta')}{f(x;\theta)}}_{=R(x;\theta',\theta)} - 1\Big) f(x;\theta)dx =$$

$$\int_{R(x;\theta',\theta)>1} \psi(T(x))\underbrace{\big(R(x;\theta',\theta)-1\big)}_{>0} f(x;\theta)dx +$$

$$\int_{R(x;\theta',\theta)<1} \psi(T(x))\underbrace{\big(R(x;\theta',\theta)-1\big)}_{<0} f(x;\theta)dx \geq$$

$$\inf_{x:R(x;\theta',\theta)>1} \psi(T(x)) \int_{R(x;\theta',\theta)>1} \big(R(x;\theta',\theta)-1\big) f(x;\theta)dx +$$

$$\sup_{x:R(x;\theta',\theta)<1} \psi(T(x)) \int_{R(x;\theta',\theta)<1} \big(R(x;\theta',\theta)-1\big) f(x;\theta)dx =$$

$$\Big(\inf_{x:R(x;\theta',\theta)>1} \psi(T(x)) - \sup_{x:R(x;\theta',\theta)<1} \psi(T(x))\Big) \int_{R(x;\theta',\theta)>1} \big(f(x;\theta') - f(x;\theta)\big)dx,$$

where in the last equality we used the identity:

$$0 = \int_{\mathbb{R}^n} \big(f(x;\theta') - f(x;\theta)\big)dx = \int_{\mathbb{R}^n} \big(R(x;\theta',\theta)-1\big) f(x;\theta)dx =$$

$$\int_{R(x;\theta',\theta)<1} \big(R(x;\theta',\theta)-1\big) f(x;\theta)dx + \int_{R(x;\theta',\theta)>1} \big(R(x;\theta',\theta)-1\big) f(x;\theta)dx.$$

Note that

$$\int_{R(x;\theta',\theta)>1} \big(f(x;\theta') - f(x;\theta)\big)dx > 0$$

and it is left to show that

$$\inf_{x:R(x;\theta',\theta)>1} \psi(T(x)) \geq \sup_{x:R(x;\theta',\theta)<1} \psi(T(x)).$$

The latter holds, since ψ is an nondecreasing function and $R(x;\theta',\theta)$ is increasing in $T(x)$. □

Here is a frequently encountered instance of the monotone likelihood ratio family

LEMMA 8b12. *One-parameter exponential family (see* (7d5)*)* $(\mathbb{P}_\theta)_{\theta\in\Theta}$ *with*

$$L(x;\theta) = \exp\Big(c(\theta)T(X) + d(\theta) + S(x)\Big) I(x \in A),$$

where $c(\theta)$ *is a strictly increasing function, is monotone likelihood ratio family in* $T(X)$*.*

PROOF. Since $c(\theta)$ is an increasing function, for $\theta_1 > \theta_0$, $c(\theta_1) - c(\theta_0) > 0$ and thus the likelihood ratio

$$\frac{L(x;\theta_1)}{L(x;\theta_0)} = \exp\Big(\big(c(\theta_1) - c(\theta_0)\big)T(X) + d(\theta_1) - d(\theta_0)\Big) I(x \in A),$$

increases in $T(x)$. □

COROLLARY 8b13. *The conclusion of Theorem 8b10 holds for one-parameter exponential family as in Lemma 8b12 (if $T(X)$ has continuous distribution).*

EXAMPLE 8b8 (continued) The likelihood function (8b4) belongs to the one parameter exponential family:

$$L_n(x;\theta) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2\Big) = \\ \exp\Big(-n/2\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\theta}{\sigma^2}\sum_{i=1}^n x_i - \frac{n}{2\sigma^2}\theta^2\Big),$$

with $c(\theta) = \theta/\sigma^2$, which is an increasing function. Hence by Corollary 8b13, it is also monotonic likelihood ratio family and hence by K-R Theorem 8b10, the level α test (8b5) is UMP in testing $H_0 : \theta \le 0$ against $H_1 : \theta > 0$. ■

EXAMPLE 8b7 (continued) The likelihood

$$L(x;\theta) = \theta^{S_n(x)}(1-\theta)^{n-S_n(x)} = \exp\Big\{S_n(x)\log\frac{\theta}{1-\theta} + n\log(1-\theta)\Big\}$$

belong to the exponential family with $c(\theta) = \log\theta/(1-\theta)$, which is strictly increasing. Hence the test (8b3) is UMP in testing $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$. ■

Unfortunately, K-R theorem does not easily extend to hypothesis testing problems with a more complicated structure: e.g. it doesn't appear to have a natural analog in the case of multivariate parameter space or when the alternative is two-sided. In fact, UMP test may fail to exist at all. To see why we shall need a result, converse to the Neyman-Pearson lemma:

PROPOSITION 8b14. *Let δ be a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, with both types of errors less than those of the N-P test. Then the two tests coincide, except perhaps on the set on which the likelihood ratio equals the critical value of the N-P test.*

PROOF. Let δ^* be the N-P test, then under the assumptions of the proposition

$$\mathbb{E}_{\theta_0}\delta(X) \le \mathbb{E}_{\theta_0}\delta^*(X)$$

and

$$\mathbb{E}_{\theta_1}\big(1-\delta(X)\big) \le \mathbb{E}_{\theta_1}\big(1-\delta^*(X)\big).$$

The first inequality reads

$$\int_{\mathbb{R}^n} \delta(x) f_X(x;\theta_0)dx \le \int_{\mathbb{R}^n} \delta^*(x) f_X(x;\theta_0)dx$$

and the second one:

$$\int_{\mathbb{R}^n} \delta(x) f_X(x;\theta_1)dx \ge \int_{\mathbb{R}^n} \delta^*(x) f_X(x;\theta_1)dx.$$

Let c be the critical value of the N-P test (w.l.o.g. $c \geq 0$), then multiplying the former inequality by $-c$ and adding it to the latter, we get:

$$\int_{\mathbb{R}^n} \delta(x)\Big(f_X(x;\theta_1) - cf_X(x;\theta_0)\Big)dx \geq \int_{\mathbb{R}^n} \delta^*(x)\Big(f_X(x;\theta_1) - cf_X(x;\theta_0)\Big)dx,$$

and

$$\int_{\mathbb{R}^n} \big(\delta(x) - \delta^*(x)\big)\Big(f_X(x;\theta_1) - cf_X(x;\theta_0)\Big)dx \geq 0$$

Let $D := \{x \in \mathbb{R}^n : f_X(x;\theta_1) - cf_X(x;\theta_0) \geq 0\}$, then the latter reads:

$$\int_D \underbrace{\big(\delta(x) - 1\big)}_{\leq 0}\underbrace{\Big(f_X(x;\theta_1) - cf_X(x;\theta_0)\Big)}_{\geq 0}dx + \int_{D^c} \underbrace{\big(\delta(x) - 0\big)}_{\geq 0}\underbrace{\Big(f_X(x;\theta_1) - cf_X(x;\theta_0)\Big)}_{\leq 0}dx \geq 0.$$

The latter inequality is possible if and only if

$$\delta(X) = \begin{cases} 1 & f_X(x;\theta_1) - cf_X(x;\theta_0) > 0 \\ 0 & f_X(x;\theta_1) - cf_X(x;\theta_0) < 0, \end{cases}$$

i.e. $\delta(x)$ coincides with $\delta^*(x)$ as claimed.

□

Here is a typical example, in which UMP doesn't exist:

Example 8b15. Let $X_1, ..., X_n$ be a sample from $N(\theta, 1)$, $\theta \in \mathbb{R}$ distribution and suppose we want test the problem

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0,$$

where θ_0 is known. Suppose δ^* is the UMP test of level α. In particular, this implies that δ^* is MP for the problem

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta',$$

for some $\theta' > \theta_0$. By Proposition 8b14, δ^* coincides with the level α N-P test for the latter problem, i.e. $\delta^*(X) = \{\bar{X} \geq c_\alpha\}$ for an appropriate critical value c_α. But on the other hand, δ^* is MP for the problem

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = -\theta',$$

and hence must have the form $\delta^*(X) = \{\bar{X} \leq c'_\alpha\}$. This contradiction shows that the UMP test does not exist for the two-sided problem. ■

The UMP test may not exist even for the one-sided problem of testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, as the following example demonstrates.

EXAMPLE 8b16. Let X be a single sample from the Cauchy density with the location parameter $\theta \in \Theta = \mathbb{R}$:

$$f(x;\theta) = \frac{1/\pi}{1+(x-\theta)^2}, \quad x \in \mathbb{R}.$$

The likelihood ratio is

$$R(x;\theta_1,\theta_0) := \frac{L(x;\theta_1)}{L(x;\theta_0)} = \frac{1+(x-\theta_0)^2}{1+(x-\theta_1)^2}.$$

For fixed $\theta_1 > \theta_0$, this ratio is not monotonous in x: take e.g. $\theta_0 = 0$ and $\theta_1 = 1$, for which

$$\lim_{x\to\pm\infty} R(x;1,0) = 1 \quad \text{and} \quad R(1;1,0) = 2.$$

Hence this model is not monotonic likelihood ratio family.

Does the UMP level α test exist for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$? Suppose it does and call it δ^*. Then δ^* is the MP test for

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1,$$

for any $\theta_1 > \theta_0$. Since the likelihood ratio statistic has continuous distribution, by Proposition 8b14 δ^* must coincide with the N-P test for the latter problem. But since θ_1 is arbitrary and δ^* does not depend on θ_1, this is possible only if the level α N-P test does not depend on the value of the alternative θ_1.

Let $\theta_0 = 0$ for definiteness, then the N-P test is given by

$$\left\{\frac{L(x;\theta_0)}{L(x;\theta_1)} \geq c\right\} = \left\{\frac{1+x^2}{1+(x-\theta_1)^2} \geq c\right\}.$$

It is easy to get convinced[8] that for different θ_1, the α level test has different critical regions. ■

c. Generalized likelihood ratio test

While the N-P test is not applicable in the general hypothesis testing problem, it motivates the following heuristic approach. Suppose we sample X from $\mathbb{P}_\theta$, and we would like to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0 \uplus \Theta_1 = \Theta$. Let $\hat\theta_0(X)$ be the MLE estimator of θ, under hypothesis H_0

$$\hat\theta_0(X) = \operatorname{argmax}_{\theta\in\Theta_0} L(X;\theta)$$

and similarly

$$\hat\theta_1(X) = \operatorname{argmax}_{\theta\in\Theta_1} L(X;\theta).$$

With enough data at hand (e.g. in the large sample asymptotic regime), we expect that $\hat\theta_1$ will be close to the actual value of the parameter, when H_1 is true, while $\hat\theta_0$ will be far from it (as it is forced to be in Θ_0). Hence the test statistic

$$\lambda(X) := \frac{L\big(X;\hat\theta_1(X)\big)}{L\big(X;\hat\theta_0(X)\big)} = \frac{\sup_{\theta\in\Theta_1} L\big(X;\theta\big)}{\sup_{\theta\in\Theta_0} L\big(X;\theta\big)} \tag{8c1}$$

[8] e.g. a calculation reveals that the likelihood ratio has a global maximum at $\theta_1/2 + \sqrt{(\theta_1/2)^2+1}$, whose value increases with θ_1. Hence for small α we shall get even nonintersecting critical regions

will yield large values and hence the test $\{\lambda(X) \geq c\}$, with an appropriate critical value c will tend to reject the null hypothesis correctly. Conversely, when H_0 is true, the null hypothesis will be correctly accepted with large probability.

This test is called *generalized* likelihood ratio test (GLRT). While the test statistic (8c1) is not guaranteed to produce optimal tests, it nevertheless typically does lead to good tests in many problems of interest and thus is used as the main test design tool in practice. Moreover, the critical value of GLRT can be efficiently approximated[9]for large n (see Theorem 8c6 below).

REMARK 8c1. If the likelihood $L(x;\theta)$ is continuous in θ and Θ_0 has a smaller dimension than Θ_1 (which itself has the dimension of Θ), then (8c1) reads:

$$\lambda(X) = \frac{\sup_{\theta\in\Theta} L\big(X;\theta\big)}{\sup_{\theta\in\Theta_0} L\big(X;\theta\big)}.$$

Such situations are typical in applications and hence this latter form is sometimes called GLRT. In this context, the maximizers over Θ_0 and Θ_1 are referred to as *restricted* and *unrestricted* MLEs respectively.

Below we shall explore a number of classical examples.

EXAMPLE 8b7 (continued) A coin is tossed and we want to decide whether it is fair or not, i.e. test $H_0 : \theta_0 = 1/2$ against $H_1 : \theta \neq 1/2$. The GLRT statistic in this case is given by

$$\lambda(X) = \frac{\sup_{\theta\in(0,1)\setminus 1/2} L\big(X;\theta\big)}{\sup_{\theta\in\{1/2\}} L\big(X;\theta\big)} = \frac{\sup_{\theta\in(0,1)} \theta^{S_n(X)}(1-\theta)^{n-S_n(X)}}{(1/2)^n} =$$
$$2^n(\bar{X}_n)^{S_n(X)}(1-\bar{X}_n)^{n-S_n(X)} = 2^n(\bar{X}_n)^{n\bar{X}_n}(1-\bar{X}_n)^{n(1-\bar{X}_n)},$$

and thus

$$\log\lambda(X) = n\log 2 + n\bar{X}_n\log\bar{X}_n + n(1-\bar{X}_n)\log(1-\bar{X}_n) =: n\Big(h(1/2) - h\big(\bar{X}_n\big)\Big),$$

where

$$h(p) = -p\log p - (1-p)\log(1-p),$$

is the Shannon *entropy* of the Ber(p) r.v. It is easy to see that $h(p)$ is maximal at $p = 1/2$ and symmetric around it. The GLRT rejects H_0 if and only if

$$\delta(X) = \big\{h\big(\bar{X}_n\big) \leq c\big\},$$

i.e. if the 'empirical' entropy is small. Note that this is a two-sided test and, in fact, it is equivalent to our original suggestion $\{|\bar{X}_n - 1/2| \geq c\}$ (why?). To get an α test, we have to choose an appropriate c, which for large n can be done approximately, using normal approximation. ■

EXAMPLE 8c2. (matched pairs experiment)

Suppose we want to establish effectiveness of a medicine, whose effect varies, depending on the weight, sleeping regime, diet, etc. of the patient. To neutralize the effect of different habits, a pair of patients with the similar habits is chosen, one of them is given a placebo and the other

[9]One of the major difficulties in constructing practical tests is calculation of the critical value for the given size α. This basically reduces to finding the distribution of the the test statistic, which can be a complicated function of the sample.

one is treated by the medicine. The responses are recorded and the experiment is repeated independently n times with other pairs. In this setting it is not unnatural to assume that the differences in responses $X_1, ..., X_n$ is a sample from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. In this problem, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. Under hypothesis H_0, the value of μ is known to be 0 and the MLE of σ^2 is given by

$$\hat{\sigma}_0^2(X) = \frac{1}{2}\sum_{i=1}^{n} X_i^2.$$

Under H_1, the MLE of μ and σ^2 are given by[10]

$$\hat{\mu}_1 = \bar{X}_n$$
$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

The GLRT statistic is:

$$\lambda(X) = \frac{L(X;\hat{\theta}_1)}{L(X;\hat{\theta}_0)} =$$
$$\frac{\left(\hat{\sigma}_1^2(X)\right)^{-n/2}}{\left(\hat{\sigma}_0^2(X)\right)^{-n/2}} \exp\left\{ -\frac{1}{2}\frac{1}{\hat{\sigma}_1^2(X)} \underbrace{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}_{=n\hat{\sigma}_1^2(X)} + \frac{1}{2}\frac{1}{\hat{\sigma}_0^2(X)} \underbrace{\sum_{i=1}^{n} X_i^2}_{=n\hat{\sigma}_0^2(X)} \right\} = \left(\frac{\hat{\sigma}_0^2(X)}{\hat{\sigma}_1^2(X)}\right)^{n/2}.$$

Note that

$$\hat{\sigma}_0^2(X)/\hat{\sigma}_1^2(X) = \frac{\hat{\sigma}_1^2(X) + \bar{X}_n^2}{\hat{\sigma}_1^2(X)} = 1 + \frac{\bar{X}_n^2}{\hat{\sigma}_1^2(X)} = 1 + \left(\frac{\bar{X}_n}{\sqrt{\hat{\sigma}_1^2(X)}}\right)^2,$$

and hence the GLRT test rejects H_0 if and only if

$$\left\{ \left| \sqrt{n-1}\frac{\bar{X}_n}{\sqrt{\hat{\sigma}_n^2(X)}} \right| \geq c \right\},$$

which is the familiar two sided test. By Proposition 5a1,

$$\sqrt{n-1}\frac{\bar{X}_n}{\sqrt{\hat{\sigma}_n^2(X)}} \sim \mathrm{Stt}(n-1),$$

under H_0 and α level test is obtained with the critical value c solving the equation

$$\alpha = 2F_{\mathrm{Stt}(n-1)}(-c) \quad \Longrightarrow \quad c(\alpha) = -F^{-1}_{\mathrm{Stt}(n-1)}(\alpha/2).$$

The obtained test may not be UMP, but is still convenient and powerful enough for practical purposes. ■

Example 8c3. If the medicine in the previous example does not depend too much on the patients' habits, the experiment can be performed in a simpler manner: a group of m patients is given a placebo and another group of n patients is treated with the drug. Assuming that the responses $X_1, ..., X_m$ and $Y_1, ..., Y_n$ of the first and the second groups are independent, with

[10] pay attention that we are maximizing with respect to μ over $\mathbb{R} \setminus \{0\}$, but this leads to the usual estimator, since the likelihood is continuous in θ - recall Remark 8c1.

$X_i \sim N(\mu_0, \sigma^2)$ and $Y_i \sim N(\mu_1, \sigma^2)$, where μ_0, μ_1 and σ are unknown, we would like to decide whether the medicine has had any effect, i.e. to test the hypothesis $H_0 : \mu_0 = \mu_1$ against $H_1 : \mu_0 \neq \mu_1$.

In this setting, $\theta = (\mu_0, \mu_1, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ and Θ_0 is the hyperplane $\mu_0 = \mu_1$. The likelihood is

$$L(x, y; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu_0)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_1)^2 \Big).$$

Under H_0, the MLE of $\mu = \mu_0 = \mu_1$ is

$$\hat{\mu} = \frac{1}{n+m}\Big(\sum_{i=1}^{n} Y_i + \sum_{j=1}^{m} X_j \Big) = \frac{n}{n+m} \bar{Y}_n + \frac{m}{n+m} \bar{X}_m$$

and the MLE of σ^2 is

$$\hat{\sigma}_0^2 = \frac{1}{n+m}\Big(\sum_{i=1}^{n} (Y_i - \hat{\mu})^2 + \sum_{j=1}^{m} (X_j - \hat{\mu})^2 \Big).$$

Under H_1, the MLE's are $\hat{\mu}_0 = \bar{X}_m$, $\hat{\mu}_1 = \bar{Y}_n$ and

$$\hat{\sigma}_1^2 = \frac{1}{n+m}\Big(\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 + \sum_{j=1}^{m} (X_i - \bar{X}_m)^2 \Big).$$

The corresponding GLRT statistic is

$$\lambda(X, Y) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{n+m} \exp\Big(- \frac{1}{2\hat{\sigma}_1^2}\Big(\sum_{i=1}^{m} (X_i - \bar{X}_m)^2 + \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \Big) + \frac{1}{2\hat{\sigma}_0^2}\Big(\sum_{i=1}^{m} (X_i - \hat{\mu})^2 + (Y_i - \hat{\mu})^2 \Big) \Big) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{n+m}.$$

This statistic is equivalent to

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} = \frac{\sum_{i=1}^{n} \Big(Y_i - \frac{n}{n+m}\bar{Y}_n - \frac{m}{n+m}\bar{X}_m \Big)^2 + \sum_{j=1}^{m} \Big(X_j - \frac{n}{n+m}\bar{Y}_n - \frac{m}{n+m}\bar{X}_m \Big)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 + \sum_{j=1}^{m} (X_i - \bar{X}_m)^2}.$$

Note that

$$\sum_{i=1}^{n} \Big(Y_i - \frac{n}{n+m}\bar{Y}_n - \frac{m}{n+m}\bar{X}_m \Big)^2 = \sum_{i=1}^{n} \Big(Y_i - \bar{Y}_n + \bar{Y}_n - \frac{n}{n+m}\bar{Y}_n - \frac{m}{n+m}\bar{X}_m \Big)^2 =$$

$$\sum_{i=1}^{n} \Big(Y_i - \bar{Y}_n + \frac{m}{n+m}\big(\bar{Y}_n - \bar{X}_m\big) \Big)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 + n\Big(\frac{m}{n+m}\Big)^2 \big(\bar{Y}_n - \bar{X}_m\big)^2$$

and similarly

$$\sum_{j=1}^{m} \Big(X_j - \frac{n}{n+m}\bar{Y}_n - \frac{m}{n+m}\bar{X}_m \Big)^2 = \sum_{j=1}^{m} (X_i - \bar{X}_m)^2 + m\Big(\frac{n}{n+m}\Big)^2 \big(\bar{Y}_n - \bar{X}_m\big)^2.$$

Then

$$\frac{\hat\sigma_0^2}{\hat\sigma_1^2} = 1 + \frac{\Big(n\Big(\frac{m}{n+m}\Big)^2 + m\Big(\frac{n}{n+m}\Big)^2\Big)\big(\bar Y_n - \bar X_m\big)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2 + \sum_{j=1}^m (X_i - \bar X_m)^2} = 1 + \frac{\frac{nm}{n+m}\big(\bar Y_n - \bar X_m\big)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2 + \sum_{j=1}^m (X_i - \bar X_m)^2}.$$

Hence the GLRT rejects H_0 if and only if

$$\left\{\frac{\big(\bar Y_n - \bar X_m\big)^2}{\sum_{i=1}^n (Y_i - \bar Y_n)^2 + \sum_{j=1}^m (X_i - \bar X_m)^2} \ge c\right\}.$$

Recall that by Proposition 5a1 the r.v.'s $\bar X_m$, $\sum_{j=1}^m (X_i - \bar X_m)^2$, $\bar Y_n$ and $\sum_{i=1}^n (Y_i - \bar Y_n)^2$ are independent. Moreover, under H_0, $\bar X_m \sim N(\mu_0, \sigma^2/m)$ and $\bar Y_n \sim N(\mu_1, \sigma^2/n)$ with $\mu_0 = \mu_1$ and hence $\bar X_m - \bar Y_n \sim N\big(0, \sigma^2(1/m + 1/n)\big)$. The denominator has χ^2 distribution, and hence the test statistic has Student distribution with explicitly computable parameters (see Proposition 5a1 (4)). The level α test is now readily obtained, by choosing the appropriate critical value c, expressed via the Student distribution quintile. ■

REMARK 8c4. If the unknown variances of the two populations in the preceding problem are not assumed equal, the MLEs of the parameters cannot be found explicitly anymore and consequently the GLRT statistic does not admit a closed form. In fact, even its numerical computation turns to be quite challenging. This variant, often referred in the literature to as the Behrens-Fisher problem, generated much research and a number of alternative solutions have been proposed over the years. A classical approach due to Behrens and Fisher is to use the standardized difference of the empirical means as the test statistic:

$$T(X,Y) = \frac{\bar X - \bar Y}{\sqrt{\hat\sigma^2(X) + \hat\sigma^2(Y)}},$$

where $\bar X = \frac{1}{m}\sum_{i=1}^m X_i$, $\bar Y = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat\sigma^2(X)$ and $\hat\sigma^2(Y)$ are the corresponding estimators of the variances. A nontrivial issue is to find or approximate the distribution of $T(X,Y)$ under H_0 to be able to find an appropriate critical value, etc.

EXAMPLE 8c5 (Test of independence). Given two samples $X_1, ..., X_n$ and $Y_1, ..., Y_n$ we would like to test whether they are independent or not. If the samples are Gaussian, the question of independence translates to the question of lack of correlation. Assume that the pairs (X_i, Y_i) are i.i.d. and have bivariate Gaussian j.p.d.f. (2b1):

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{1}{1-\rho^2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\},$$

where all the parameters $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ are unknown and vary in the corresponding subspaces. We want to test the hypothesis $H_0 : \rho = 0$ against $H_1 : \rho \ne 0$. The corresponding likelihood function is

$$L(x,y;\theta) = \left(\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\right)^n \exp\left\{-\frac{1}{1-\rho^2}\left(\sum_i \frac{(x_i-\mu_1)^2}{2\sigma_1^2} - \sum_i \frac{\rho(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2} + \sum_i \frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)\right\}.$$

Under both H_0 and H_1, the MLE estimators are

$$\hat{\mu}_1 = \bar{X}_n$$
$$\hat{\mu}_2 = \bar{Y}_n$$
$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$
$$\hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2.$$

Under H_1, the MLE of ρ is

$$\hat{\rho} = \frac{1}{n\hat{\sigma}_1\hat{\sigma}_2}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

A direct calculation reveals the GLRT statistic:

$$\log \lambda(X, Y) = -\frac{n}{2}\log(1 - \hat{\rho}^2).$$

Hence the GLRT rejects H_0 if and only if

$$\big\{|\hat{\rho}| \geq c\big\}.$$

Recall that $\tilde{X}_i = (X_i - \mu_1)/\sigma_1$ and $\tilde{Y}_i = (Y_i - \mu_2)/\sigma_2$ are $N(0,1)$ r.v.'s and thus it is easy to see that the distribution of $\hat{\rho}$ depends only on ρ and not on the rest of the parameters. Further, it can be shown that

$$T_n(X) = \frac{\sqrt{n-2}\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$$

has Student distribution with $n-2$ degrees of freedom. Now the level α test can be readily obtained, by choosing an appropriate critical value. ■

Potentially we may encounter two difficulties when trying to apply the GLRT: first, as calculating the GLRT statistic reduces to an optimization problem, it can easily be quite involved technically and be challenging even numerically; second, even if a closed form statistic can be found, its distribution under H_0 might be hard to find and, consequently, it is not clear how to choose the critical value to achieve the required significance level.

Wilks' theorem resolves the second difficulty asymptotically as $n \to \infty$:

THEOREM 8c6. *Let $X_1, ..., X_n$ be a sample from the density $f(x;\theta)$, satisfying the assumptions of Theorem 7g34 (consistency and asymptotic normality of MLE, page 155) hold. Let $\lambda_n(X)$ be the GLRT statistic*

$$\lambda_n(X) := \frac{\sup_{\theta\in\Theta} L_n(X;\theta)}{L_n(X;\theta_0)},$$

for testing

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0,$$

where θ_0 is an interior point of the closed interval $\Theta \subset \mathbb{R}$. Then[11] *under $\mathbb{P}_{\theta_0}$*

$$2\log\lambda_n(X) \xrightarrow{d} \chi^2_1.$$

PROOF. (sketch) Since Θ is closed and the likelihood is continuous in θ, $\sup_{\theta\in\Theta} L_n(X;\theta) = L_n(X;\hat\theta_n)$, where $\hat\theta_n$ is the MLE of θ. Expanding into powers of $\hat\theta_n - \theta_0$ around $\hat\theta_n$ and using the continuity of the second derivative, we get

$$2\log\lambda_n(X) = 2\log L_n(X;\hat\theta_n) - 2\log L_n(X;\theta_0) = \\ 2\frac{\partial}{\partial\theta}\log L_n(X;\hat\theta_n)(\hat\theta_n-\theta_0) - \frac{\partial^2}{\partial\theta^2}\log L_n(X;\theta_0)(\hat\theta_n-\theta_0)^2 + r_n(X), \quad (8c2)$$

where $r_n(X)$ is the reminder term, converging to zero in $\mathbb{P}_{\theta_0}$-probability as $n\to\infty$ (fill in the details!).

If $\hat\theta_n(X) \in \Theta^\circ$ (the interior of Θ), then $\frac{\partial}{\partial\theta}L_n(X;\hat\theta_n) = 0$, and hence

$$\frac{\partial}{\partial\theta}L_n(X;\hat\theta_n) = \frac{\partial}{\partial\theta}L_n(X;\hat\theta_n)\mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}.$$

Consequently,

$$\frac{\partial}{\partial\theta}\log L_n(X;\hat\theta_n)(\hat\theta_n-\theta_0) = \mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}(\hat\theta_n-\theta_0)\left(\sum_{i=1}^n \ell'(X_i;\hat\theta_n)\right) = \\ \mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}(\hat\theta_n-\theta_0)\left(\sum_{i=1}^n \ell'(X_i;\theta_0) + \sum_{i=1}^n \ell''(X_i;\tilde\theta_n)\big(\hat\theta_n-\theta_0\big)\right)$$

where $\ell(x;\theta) = \log f(x;\theta)$ and where $|\tilde\theta_n - \theta_0| \le |\hat\theta_n - \theta_0|$. Recall that $|\ell''(x,\theta)| \le h(x)$ for some $h(x)$, satisfying $\mathbb{E}_{\theta_0}h(X_1) < \infty$, hence by Slutsky's theorem

$$\mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}\left|\sum_{i=1}^n \ell''(X_i;\tilde\theta_n)\big(\hat\theta_n-\theta_0\big)^2\right| \le \mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}\left(\frac{1}{n}\sum_{i=1}^n h(X_i)\right)\big(\sqrt{n}(\hat\theta_n-\theta_0)\big)^2 \xrightarrow{n\to\infty} 0,$$

where the convergence holds, since the second term converges in probability to a constant by LLN, the third term converges weakly since $(\hat\theta_n)$ is asymptotically normal with rate $\sqrt{n}$ and the first term converges to zero in probability, since $\hat\theta_n$ converges to θ_0, which is in the interior of Θ.

Similarly,

$$\left|\mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}(\hat\theta_n-\theta_0)\sum_{i=1}^n \ell'(X_i;\theta_0)\right| = \mathbf{1}_{\{\hat\theta_n\notin\Theta^\circ\}}\sqrt{n}\big|\hat\theta_n-\theta_0\big|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell'(X_i;\theta_0)\right| \xrightarrow{\mathbb{P}_{\theta_0}} 0,$$

where the convergence holds by Slutsky's theorem (note that $\mathbb{E}_{\theta_0}\ell'(X_i;\theta_0) = 0$ and the CLT is applicable as $I(\theta_0) < \infty$). To recap, the first term on the right hand side of (8c2) converges to zero in $\mathbb{P}_{\theta_0}$-probability.

Finally, we have

$$-\frac{\partial^2}{\partial\theta^2}\log L_n(X;\theta_0)(\hat\theta_n-\theta_0)^2 = -\left(\frac{1}{I(\theta_0)}\frac{1}{n}\sum_{i=1}^n \ell''(X_i;\theta_0)\right)\left(\sqrt{n}\sqrt{I(\theta_0)}(\hat\theta_n-\theta_0)\right)^2 \xrightarrow{d} \chi^2_1,$$

[11] χ^2_1 is the χ-square distribution with one degree of freedom

where the convergence holds, since the first term converges to 1 by the LLN and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1/I(\theta_0))$. □

Wilks' theorem extends to the multivariate case as follows. Suppose that we are given a sample from the p.d.f. $f(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$, $\dim(\Theta) = k$ and we want to test

$$\begin{aligned} H_0 &: R(\theta) = 0 \\ H_1 &: R(\theta) \neq 0 \end{aligned} \tag{8c3}$$

where $R : \Theta \mapsto \mathbb{R}^r$ with $r \leq k$ is a continuously differentiable function of full rank, so that the relation $R(\theta) = 0$ defines $(k-r)$-dimensional manifold Θ_0. The GLRT statistic is

$$\lambda_n(X) = \frac{\sup_{\theta\in\Theta} L_n(X;\theta)}{\sup_{\theta\in\Theta_0} L_n(X;\theta)},$$

where $L_n(X;\theta) = \prod_{i=1}^n f(X_i;\theta)$. Under appropriate technical conditions, under $\mathbb{P}_{\theta_0}$

$$2\log\lambda_n(X) \xrightarrow{d} \chi^2_r,$$

where χ^2_r is the χ-square distribution with r degrees of freedom. Hence the number of degrees of freedom of the limit distribution is the dimension of the parameter space minus the dimension of the constrained space under H_0.

Wald's test and Rao's score test. Consider the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against two-sided composite alternative $H_1 : \theta \neq \theta_0$, given the the i.i.d. sample $X^n = (X_1, ..., X_n)$. Wald's test statistic is

$$W_n(X^n) := nI(\hat{\theta}_n)\big(\hat{\theta}_n - \theta_0\big)^2$$

where $\hat{\theta}_n$ is the MLE of θ and $I(\theta)$, $\theta \in \Theta$ is the Fisher information in X_1. Since under appropriate conditions the MLE is asymptotically normal with variance $1/I(\theta)$, $W_n(X^n)$ converges in distribution to the χ^2_1 random variable as $n \to \infty$, if $I(\theta)$ is a continuous function of θ.

An alternative is Rao's score test statistic

$$S_n(X^n) = \frac{\big(\partial_\theta \log L(X^n;\theta_0)\big)^2}{nI(\theta_0)},$$

where $L_n(X^n;\theta)$ is the likelihood function of the sample. Under appropriate conditions on the density $f(x;\theta)$, $\mathbb{E}_\theta \partial \log f(X_1,\theta) = 0$ and $\mathbb{E}_\theta\big(\partial \log f(X_1,\theta)\big)^2 = I(\theta)$ and hence by the CLT

$$S_n(X^n) = \left(\frac{1}{\sqrt{I(\theta_0)}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta \log f(X_i;\theta)\right)^2 \xrightarrow[n\to\infty]{d} \chi^2_1.$$

Both tests generalize to the setting (8c3). Wald's test statistic is

$$W_n(X^n) = R^\top(\hat{\theta}_n)\Big(\nabla^\top R(\hat{\theta}_n) I_n^{-1}(\hat{\theta}_n) \nabla R^\top(\hat{\theta}_n)\Big)^{-1} R(\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the *unrestricted* MLE of θ (i.e. the maximizer of the likelihood over Θ), $\nabla R(\theta)$ is the gradient matrix of R and $I_n(\theta) = n\mathbb{E}_\theta \nabla \log f(X_1;\theta) \nabla^\top \log f(X_1;\theta)$ is the Fisher information matrix.

Rao's score test statistic is

$$S_n(X) := \nabla^\top \log L_n\big(X^n; \hat\theta_n^{(0)}\big) I_n^{-1}(\hat\theta_n^{(0)}) \nabla \log L_n\big(X^n; \hat\theta_n^{(0)}\big),$$

where $\hat\theta_n^{(0)}$ is the *restricted* MLE of θ (i.e., over Θ_0).

It can be shown that both statistics converge in distribution to χ^2_r random variable under H_0, just as the GLRT statistic. Note that Wald's and Rao's tests require calculation of only unrestricted and restricted MLEs respectively, while the GLRT test requires both.

d. Some classical tests

Pearson's χ^2 test. Suppose we observe the outcome of $S_n \sim \text{mult}(n, p)$, where $p = (p_1, ..., p_k)$ is the vector of probabilities and would like to test

$$\begin{aligned} H_0 &: p = q \\ H_1 &: p \neq q, \end{aligned} \tag{8d1}$$

for a known vector q.

The GLRT statistic is

$$\lambda_n(X) = \frac{\sup_{p \in \mathcal{S}^{k-1}} p_1^{S_n(1)} ... p_k^{S_n(k)}}{q_1^{S_n(1)} ... q_k^{S_n(k)}}.$$

The supremum can be found by the Lagrange multipliers: the Lagrangian is

$$\Lambda(p, \lambda) = \sum_{i=1}^k S_n(i) \log p_i + \lambda \left(1 - \sum_{i=1}^k p_i\right),$$

and taking the derivatives w.r.t. p_i and equating them to zero gives:

$$S_n(i) = p_i \lambda.$$

Summing up over i gives $\lambda = n$ and hence the MLEs are given by (check the conditions for maximum)

$$\hat p_i = S_n(i)/n = \bar X_n(i)$$

Hence (on the event $\cap_{i=1}^k \{\bar X_n(i) > 0\}$)

$$\log \lambda_n(X) = -n \sum_{i=1}^k \bar X_n(i) \left(\log q_i - \log \bar X_n(i)\right).$$

By the LLN, $\bar X_n \overset{\mathbb{P}_{\theta_0}}{\to} q$ under H_0 and expanding the latter expression into power series of $\bar X_n - q$ around $\bar X_n$ we get

$$2 \log \lambda_n(X) = -2n \sum_{i=1}^k \bar X_n(i) \frac{1}{\bar X_n(i)} \left(q_i - \bar X_n(i)\right) + n \sum_{i=1}^k \bar X_n(i) \frac{1}{\bar X_n^2(i)} \left(q_i - \bar X_n(i)\right)^2 + r_n.$$

The first term on the right vanishes, since $\sum_i \bar X_n(i) = \sum_i q_i = 1$, and the residual term can be seen to converge to zero in probability under H_0. Since

$$\Theta = \mathcal{S}^{k-1} := \left\{ q \in \mathbb{R}^k : q_i \geq 0, \sum_{i=1}^k q_i = 1 \right\},$$

$\dim(\Theta) = k-1$ and hence by the multivariate Wilks' theorem the second term converges weakly to χ^2_{k-1} distribution. For large n, the second term is the dominant one and hence it makes sense to use it as the test statistic for the problem, ignoring other terms.

To recap, the so called Pearson's χ^2 statistic

$$\chi^2(S_n) := \sum_{i=1}^{k} \frac{(S_n(i) - nq_i)^2}{S_n(i)} \tag{8d2}$$

can be used to test the problem (8d1) and the critical value of the level α test can be approximated, using the limit $\chi^2(S_n) \xrightarrow{d} \chi^2_{k-1}$ under H_0.

Here is the first classical application of the Pearson's χ^2 test:

EXAMPLE 8d1 (Goodness of fit test). Consider the problem of testing whether a sample comes from a particular distribution or not. More precisely, given a sample $X_1, ..., X_n$ from a c.d.f. F, test

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Formally, the parametric space Θ here is the space of c.d.f's, which is infinite dimensional, unlike all the examples we have seen so far. Such statistical models are called nonparametric and their study require a different theory and tools, than those covered in this course.

However, a reasonable test can be constructed using the following idea. Partition the range of X_1 into k disjoint sets $I_1, ..., I_k$ (usually intervals if the range of X_1 is $\mathbb{R}$) and consider $S_n(i) = \sum_{m=1}^{n} \mathbf{1}_{\{X_m \in I_i\}}$, $i = 1, ..., k$. The vector S_n has multinomial distribution with probabilities $p_i = \mathbb{P}_F(X_1 \in I_i) = \int_{I_i} dF$. This reduces the problem to (8d1) with $q_i = \mathbb{P}_{F_0}(X_1 \in I_i)$, to which the χ^2-test is applicable. Of course, strictly speaking, the obtained test is not suitable for the original problem, since two different distributions may assign the same probabilities to all I_i's. However, under appropriate restrictions on the set of alternative distributions it is known to give very reasonable powers and is widely used in practice. The choice of the intervals I_i is another practical issue to be addressed.

■

For the next application, we shall need the following variation on the Pearson's χ^2 theme. Suppose we want to test

$$H_0 : p = \psi(p)$$
$$H_1 : p \neq \psi(p),$$

given the mult(n, p) counts S_n as before, where ψ is a fixed known function and the relation $p = \psi(p)$ defines a d-dimensional submanifold on the $k-1$ dimensional simplex $\mathcal{S}^{k-1}$ of probability vectors.

The GLRT statistic is

$$\lambda_n(S_n) = \prod_{i=1}^{k} \frac{(S_n(i)/n)^{S_n(i)}}{\left(\hat{p}_i\right)^{S_n(i)}}$$

where $\hat{p}_i$ is the MLE of p_i under H_0. Hence

$$2\log\lambda_n(S_n) = 2\sum_{i=1}^{k} S_n(i)\big(\log S_n(i)/n - \log\hat{p}_i\big) = \\ -2\sum_{i=1}^{k}\big(S_n(i)/n - \hat{p}_i\big) + \sum_{i=1}^{k}\frac{n^2}{S_n(i)}\big(S_n(i)/n - \hat{p}_i\big)^2 + r_n.$$

As before, the first term vanishes and the residual term converges to zero in probability under H_0. Moreover, since under H_0 $S_n(i)/n - \hat{p}_i = (S_n(i)/n - p_i) + (p_i - \hat{p}_i) \to 0$ in probability,

$$\sum_{i=1}^{k}\frac{n^2}{S_n(i)}\big(S_n(i)/n - \hat{p}_i\big)^2 = \sum_{i=1}^{k}\frac{\Big(\sqrt{n}\big(S_n(i)/n - \hat{p}_i\big)\Big)^2}{S_n(i)/n} = \sum_{i=1}^{k}\frac{\big(S_n(i) - n\hat{p}_i\big)^2}{n\hat{p}_i} + r'_n,$$

with a residual term r'_n converging to zero. Hence by Wilks' theorem the dominant term

$$\chi^2(S_n) = \sum_{i=1}^{k}\frac{\big(S_n(i) - n\hat{p}_i\big)^2}{n\hat{p}_i} \tag{8d3}$$

converges to $\chi^2_{(k-1)-d}$ distribution under H_0. The main application of this version of the Pearson's χ^2-test is testing independence in contingency tables.

EXAMPLE 8d2. Suppose we want to test whether smoking and lung cancer are related. For this purpose, we may survey n individuals from the population for their smoking habits and the disease history. If the individuals are assumed i.i.d., the sufficient statistic is the vector of counters of each one of the four cases, which can be conveniently summarized in the following *contingency table*:

smoking/cancer	no	yes
no	S_{00}	S_{01}
yes	S_{10}	S_{11}

Denote by p_{ij}, $i,j \in \{0,1\}$ the corresponding probabilities (e.g. p_{00} is the probability that an individual doesn't smoke and didn't develop cancer, etc.) and by p_{i*} and p_{*j} the marginal probabilities (e.g. $p_{1*} = p_{10} + p_{11}$ is the probability that a sampled individual smokes). We want to test

$$H_0 : p_{ij} = p_{i*}p_{*j}, \quad \text{for all } i,j \in \{0,1\}$$
$$H_1 : p_{ij} \neq p_{i*}p_{*j}, \quad \text{for some } i,j \in \{0,1\}.$$

More generally, we may consider testing the above hypotheses for the contingency table of the form

	x_1	x_2	...	x_c	total
y_1	S_{11}	S_{12}	...	S_{1c}	S_{1*}
y_2	S_{21}	S_{22}	...	S_{2c}	S_{2*}
...	..	...	...	...	...
y_r	S_{r1}	S_{r2}	...	S_{rc}	S_{r*}
total	S_{*1}	S_{*2}	...	S_{*c}	n

where c and r are the numbers of columns and rows respectively, and $S_{i*} = \sum_{j=1}^{c} S_{ij}$ and $S_{*j} = \sum_{i=1}^{r} S_{ij}$. A calculation reveals that the MLE of p_{ij} under H_0 equals $\bar{X}_{i*}\bar{X}_{*j} := (S_{i*}/n)(S_{*j}/n)$ and the statistic (8d3) reads

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(S_{ij} - n\bar{X}_{i*}\bar{X}_{*j}\right)^2}{n\bar{X}_{i*}\bar{X}_{*j}}.$$

The dimension of the constrained subspace under H_0 is $(r-1)+(c-1)$ and the dimension of the whole parameter space is $rc-1$, hence the statistic converges in law to χ-square distribution with $rc-1-r-c+2 = (r-1)(c-1)$ degrees of freedom. ■

Goodness of fit tests: Kolmogorov–Smirnov and Cramer–von Mises. Suppose we observe the sample X_1, ...,X_n and would like to decide whether it comes from a particular c.d.f. F_0 or not, i.e. to test

$$\begin{aligned} H_0 &: F = F_0 \\ H_1 &: F \neq F_0. \end{aligned} \tag{8d4}$$

This is the goodness of fit problem, briefly introduced in Example 8d1 as an application of Pearson's χ^2-test, which has an appealing simple structure, but also has both practical and theoretical limitations. In particular, it is not immediately clear how to choose the intervals to construct the appropriate multinomial distribution. Moreover, since the vector of corresponding counts do not in general form a sufficient statistic, it is clear that Pearson's χ^2-test inevitably discards some statistical information, contained in the sample.

Alternative solutions to the problem are based on the empirical distribution of the sample. Let $\hat{F}_n$ be the empirical distribution of the sample, i.e.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}, \quad x \in \mathbb{R}.$$

Clearly, $\hat{F}_n$ is a legitimate discrete c.d.f., which is random due to its dependence on the sample. By the LLN,

$$\hat{F}_n(x) \xrightarrow{\mathbb{P}_F} F(x), \quad \forall x \in \mathbb{R},$$

where $\mathbb{P}_F$ denotes the probability law of the data, corresponding to the c.d.f. F. Remarkably, a stronger, uniform over x, result holds

THEOREM 8d3 (Glivenko Cantelli). *Assume $X_1, ..., X_n$ are i.i.d., then*

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\mathbb{P}_F} 0.$$

Even more remarkably, the convergence above can be quantified as follows:

THEOREM 8d4 (Dvoretzky-Kiefer-Wolfowitz inequality). *For i.i.d. r.v.'s $X_1, ..., X_n$ and an $\varepsilon > 0$,*

$$\mathbb{P}_F\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

These results, whose proof is beyond our scope, hint to uniform over $x \in \mathbb{R}$ version of CLT:

THEOREM 8d5 (Kolmogorov Smirnov). *Assume $X_1, ..., X_n$ are i.i.d. with continuous distribution, then*

$$\sqrt{n}\sup_{x\in\mathbb{R}}|\hat{F}_n(x) - F(x)| \xrightarrow{d_F} \kappa$$

where κ is a random variable with the Kolmogorov distribution:

$$\mathbb{P}(\xi \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2}.$$

The latter theorem can be applied to test (8d4): given the sample, calculate[12] the statistic

$$D_n(X) = \sqrt{n}\sup_{x\in\mathbb{R}}|\hat{F}_n(x) - F_0(x)|$$

and reject H_0 if $D_n \ge c(\alpha)$, where $c(\alpha)$ is chosen as the α-quantile of the Kolmogorov distribution.

REMARK 8d6. It is not hard to find the distribution of D_n for any fixed n, however, for large n the computations become quite involved and the Kolmogorov-Smirnov asymptotic is often preferred.

A computationally appealing feature of the statistic D_n is that its limit distribution does not depend on F_0, i.e. it is asymptotically *distribution-free*. This property should be expected and, in fact, the statistic D_n is distribution-free for each $n \ge 1$. To see this, recall that if F is a continuous distribution on $\mathbb{R}$ and $X_1 \sim F$, then $F(X_1) \sim U([0,1])$. Hence

$$\sup_{x\in\mathbb{R}}|\hat{F}_n(x) - F_0(x)| = \sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{X_i\le x\}} - F_0(x)\right| = \sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{F_0(X_i)\le F_0(x)\}} - F_0(x)\right| =$$
$$\sup_{u\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{F_0(X_i)\le u\}} - u\right| = \sup_{u\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{U_i\le u\}} - u\right|,$$

which shows that the distribution of D_n is independent of F_0.

Let us see how the particular form of the limit distribution emerges. Note that we can restrict our consideration to $F_0(x) = x$, $x \in [0,1]$ corresponding to $U([0,1])$. To this end, for any fixed $x \in [0,1]$, $\sum_{i=1}^{n}\mathbf{1}_{\{U_i\le x\}} \sim \mathrm{Bin}(n,x)$. Hence by the classical CLT,

$$\sqrt{n}\big(\hat{F}_n(x) - F_0(x)\big) \xrightarrow{d} N(0, x(1-x)).$$

The statistic D_n is a functional of $\sqrt{n}\big(\hat{F}_n(x) - F_0(x)\big)$, $x \in \mathbb{R}$, i.e. it depends on this expression at all points x simultaneously. Hence we shall need the multivariate version of the CLT:

THEOREM 8d7. *Let (X_n) be an i.i.d. sequence of random vectors in $\mathbb{R}^d$ with $\mathbb{E}X_1 = \mu$ and the covariance matrix $S = \mathrm{cov}(X_1, X_1)$. Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, S).$$

[12] calculating supremum is easy in this case since $x \mapsto F_0(x)$ increases and $\hat{F}_n(x)$ is piecewise constant

The proof of this theorem is similar to the proof of the scalar CLT.

Consider now the random vector $\big(\hat{F}_n(x_1), \hat{F}_n(x_2)\big)$ for $0 \le x_1 < x_2 \le 1$. As before we have $\mathbb{E}_{F_0}\hat{F}_n(x_i) = x_i$ and $\mathrm{var}_{F_0}(\hat{F}_n(x_1)) = \frac{1}{n}x_i(1-x_i)$ for $i = 1, 2$. Also

$$\mathrm{cov}\big(\hat{F}_n(x_1), \hat{F}_n(x_2)\big) = \frac{1}{n^2}\mathbb{E}_{F_0}\sum_{i,j}\mathbf{1}_{\{U_i \le x_1\}}\mathbf{1}_{\{U_j \le x_2\}} - x_1x_2 =$$

$$\frac{1}{n^2}\big((n^2-n)x_1x_2 + n(x_1 \wedge x_2)\big) - x_1x_2 = 1/n\Big((x_1 \wedge x_2) - x_1x_2\Big).$$

Hence by the multivariate CLT,

$$\sqrt{n}\big(\hat{F}_n(x_1) - F_0(x_1), \hat{F}_n(x_2) - F_0(x_2)\big) \xrightarrow{d} N\Big(0, C(x_1, x_2)\Big),$$

where

$$C(x_1, x_2) = \begin{pmatrix} x_1(1-x_1) & x_1 \wedge x_2 - x_1x_2 \\ x_1 \wedge x_2 - x_1x_2 & x_2(1-x_2) \end{pmatrix}.$$

Similarly, one can check that the vector with entries $\sqrt{n}\big(\hat{F}_n(x) - F_0(x)\big)$, $x \in \{x_1, ..., x_n\}$ converges weakly to a zero mean Gaussian vector with the covariance matrix with the entries $x_i \wedge x_j - x_ix_j$, $i, j \in \{1, ..., n\}$.

More sophisticated mathematics, namely functional CLT, shows that the whole function $\sqrt{n}\big(\hat{F}_n(x) - F_0(x)\big)$, $x \in \mathbb{R}$ converges weakly[13] to a zero mean Gaussian process $V(x)$ with the correlation function $\mathbb{E}V(x)V(y) = x \wedge y - xy$, $x, y \in [0, 1]$. The process with this correlation function is called Brownian bridge and the Kolmogorov distribution is the one of $\sup_{x\in[0,1]} V(x)$, which can be found using appropriate techniques.

Similarly, it can be shown that the Cramer–von Mises statistic

$$H_n := n\int_{\mathbb{R}}(\hat{F}_n(x) - F_0(x))^2 dF_0(x)$$

converges weakly to the random variable

$$H = \int_0^1 V^2(x)dx,$$

whose distribution is also independent of F_0 and can be found explicitly.

e. Testing multiple hypotheses

Familywise Error Rate. When a new medicine is being considered for use, the decision is based not only on its direct efficiency, but also on the presence of side-effects. Suppose we know what the possible side-effects are and we want to detect them, using appropriate statistical tests. Even if the false positive probability of each test is small, the probability of getting at least one false positive result can be large by 'pure chance', if the number of tests is large.

More precisely, consider m hypothesis testing problems in which the null hypotheses H_{0i} are tested against the corresponding alternatives H_{1i}, $i = 1, ..., m$ and assume that in each problem a size α test δ_i is used. The *familywise error rate* (FWER) is defined as the probability of rejecting at least one of H_{0i}'s erroneously. In the medicine example, FWER is the probability of erroneous detection of at least one side-effect.

[13] weak convergence should and can be defined for random processes

Note that it is the type I error of the test $\delta = \max_{i=1,...,m} \delta_i$ in the problem

$$H_0 : (\theta_1, ..., \theta_m) \in \Theta_0 = \Theta_{01} \times ... \times \Theta_{0m}$$
$$H_1 : (\theta_1, ..., \theta_m) \notin \Theta_0 = \Theta_{01} \times ... \times \Theta_{0m}$$

where Θ_{0i} and Θ_{1i} are the parameter subspaces, corresponding to the null hypothesis and the alternative in the i-th problem. If the samples in the problems are independent, then for $\theta \in \Theta_0$,

$$\text{FWER} = \mathbb{E}_\theta \delta = \mathbb{E}_\theta \max_i \delta_i = 1 - \mathbb{E}_\theta \prod_i (1 - \delta_i) = 1 - (1 - \alpha)^m.$$

If FWER is required to be less or equal than $\bar{\alpha} \in (0, 1)$, then the size in each test must satisfy

$$1 - (1 - \alpha)^m = \bar{\alpha} \quad \Longrightarrow \quad \alpha \leq 1 - (1 - \bar{\alpha})^{1/m},$$

and hence the sizes of the individual tests are to be *corrected* to achieve the desired *control* over the FWER. The latter simple formula is called Sidak's correction. For example, if $m = 20$ and $\bar{\alpha} = 0.05$ is needed, then $\alpha \leq 0.0026$ must be chosen. Requiring such small sizes in the tests may significantly decrease their powers and consequently in the familywise power.

If the samples are dependent, the *Bonferroni correction* suggests to replace the individual sizes with $\bar{\alpha}/m$, in which case

$$\text{FWER} = \mathbb{P}_\theta \Big(\bigcup_{i=1}^m \{\delta_i = 1\} \Big) \leq m \mathbb{E}_\theta \delta_i = m \bar{\alpha}/m = \bar{\alpha}.$$

Note that if the samples are in fact independent, the Bonferroni correction yields smaller individual sizes and hence worse overall power. More sophisticated approaches to controlling the FWER exist.

False Discovery Rate. Suppose we want to identify the genes, which are related to certain disease. For a single given gene such relation can be tested, e.g., by Pearson's χ^2 test as in Example 8d2. In practice, a very large number of genes is screened and it makes sense to seek for a screening procedure, which makes as few erroneous detections as possible. To this end, consider the quantities, summarized in the following table

	# of true H_{0i}'s	# of false H_{0i}'s	total
# of rejected H_{0i}'s	V	S	R
# of accepted H_{0i}'s	U	T	m-R
total	m_0	$m - m_0$	m

TABLE 1. The standard terminology of FDR

Note that the random variables V, S, U and T are not observable, while R is. The *false discovery rate* (FDR) is defined as the expected ratio of false discoveries (erroneous rejections of the null hypotheses) over the total number of discoveries:

$$\text{FDR} = \mathbb{E} \frac{V}{R} \mathbf{1}_{\{R>0\}} = \mathbb{E} \frac{V}{R \vee 1}.$$

Here $\mathbb{E}$ is the expectation with respect to the probability $\mathbb{P}$, under which m_0 out of m null hypotheses are true[14].

REMARK 8e1. Note that if $m_0 = m$, i.e. all the null hypotheses are true, $V = R$ and $\text{FWER} = \mathbb{P}(V > 0) = \mathbb{P}(R > 0)$. Hence in this case

$$\text{FDR} = \mathbb{E}\frac{R}{R \vee 1} = \mathbb{P}(R > 0) = \text{FWER}.$$

The following algorithm, proposed by Y.Benjamini and Y.Hochberg, controls the FDR at any desired level $\alpha \in (0, 1)$.

(1) Order the p-values p_i of the individual tests in the increasing order and denote the ordered sequence by $p_{(i)}$, $i = 1, ..., m$

(2) For a fixed number find the integer

$$k = \max\Big\{i : p_{(i)} \leq \frac{i}{m}\alpha\Big\},$$

where $\max\{\emptyset\} = 0$.

(3) Reject $H_{0(i)}$ if and only if $i \leq k$ (if $k = 0$, then none is rejected).

PROPOSITION 8e2. *If p_i's are independent and uniformly distributed on the interval $[0, 1]$, then the FDR of BH procedure is less or equal than α.*

REMARK 8e3. The assumptions of the proposition are satisfied if the tests are based on independent samples and the test statistic in each test has continuous distribution.

To prove this assertion we shall need some additional tools from probability theory. Let us start with some basic definitions. An increasing sequence of σ-algebras $\mathcal{F}_j \subseteq \mathcal{F}_{j+1}$, $j \in \mathbb{Z}_+$ is called *filtration* (of σ-algebras). For example, if (X_j) is a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the σ-algebras $\mathcal{F}_j^X = \sigma\{X_1, ..., X_j\}$ form a filtration, called the *natural* filtration of the process (X_j). The filtration should be thought of as an increasing collection of events, which become certain as time evolves. For the natural filtration of the sequence (X_j), this means the collection of events revealed by observing the sequence up to current time.

A process (X_j) is *adapted* to the filtration $(\mathcal{F}_j)$ if X_j is measurable with respect to (i.e. determined by) $\mathcal{F}_j$ for all j's. A process (M_j) is a *martingale* with respect to filtration $(\mathcal{F}_j)$ if it is adapted to it and for $j \geq i$

$$\mathbb{E}(M_j|\mathcal{F}_i) = M_i.$$

A random variable τ with nonnegative integer values is called *stopping time* if $\{\tau \leq j\} \in \mathcal{F}_j$ for all j's, i.e. the event $\{\tau \leq j\}$ becomes certain if $\mathcal{F}_j$ is observed. For example, if (X_j) is an adapted process, then the *first passage time* of $a \in \mathbb{R}$

$$\tau_a = \min\big\{j : X_j \geq a\big\}$$

[14]there are $\binom{m}{m_0}$ such probabilities, corresponding to the combinations of indices of the true null hypotheses, but this will not play a role in the calculations to be presented below due to independence

is a stopping time. Indeed, $\{\tau_a > j\} = \cap_{i=1}^{j}\{X_i < a\}$ and since (X_j) is adapted to $(\mathcal{F}_j)$, we have $\{X_i \le a\} \in \mathcal{F}_j$ for all $i \le j$ and hence $\{\tau_a > j\} \in \mathcal{F}_j$ and hence $\{\tau_a \le j\} \in \mathcal{F}_j$ for all j's. Convince yourself, that, for example, the last passage time $s_a = \max\{j : X_j \ge a\}$ is not a stopping time.

EXAMPLE 8e4. Let (ξ_j) be a sequence of i.i.d. equiprobable random signs, i.e. $\mathbb{P}(\xi_1 = \pm 1) = 1/2$ and define the *simple random walk*

$$S_j = \sum_{i=1}^{j} \xi_i, \quad j \ge 1,$$

and $S_0 = 0$. Since for each j, all ξ_i's with $i \le j$ can be recovered from $S_1, ..., S_j$ and vise versa, the natural filtrations of (ξ_j) and (S_j) coincide. Denote this filtration by $\mathcal{F}_j$. The sequence (ξ_j) is not a martingale w.r.t. $(\mathcal{F}_j)$ since for $j > i$, $\mathbb{E}(\xi_j|\mathcal{F}_i) = \mathbb{E}\xi_j = 0 \ne \xi_i$. However (S_j) is a martingale w.r.t. $(\mathcal{F}_j)$

$$\mathbb{E}(S_j|\mathcal{F}_i) = S_i + \mathbb{E}\left(\sum_{\ell=i+1}^{j} \Big|\mathcal{F}_i\right) = S_i, \quad j \ge i.$$

■

The martingale property implies that $\mathbb{E}M_j = \mathbb{E}M_0$ and one may expect that $\mathbb{E}M_\tau = \mathbb{E}M_0$ holds for Markov times as well. This is true, at least for martingales on finite intervals:

THEOREM 8e5 (a rudiment of the Optional Sampling Theorem). *Let (M_j), $j = 0, 1, ..., n$ be a martingale with respect to a filtration $(\mathcal{F}_j)$ and let τ be a stopping time with values in $\{0, ..., n\}$. Then*

$$\mathbb{E}M_\tau = \mathbb{E}M_0.$$

PROOF. Since $\sum_{i=0}^{n} \mathbf{1}_{\{\tau=i\}} = 1$,

$$\mathbb{E}M_\tau = \mathbb{E}M_0\mathbf{1}_{\{\tau=0\}} + \mathbb{E}\sum_{i=1}^{n} M_i\mathbf{1}_{\{\tau=i\}} = \mathbb{E}M_0\mathbf{1}_{\{\tau=0\}} + \mathbb{E}\sum_{i=1}^{n} M_i\big(\mathbf{1}_{\{\tau\le i\}} - \mathbf{1}_{\{\tau<i\}}\big) =$$

$$\mathbb{E}M_0\mathbf{1}_{\{\tau=0\}} + \sum_{i=1}^{n}\big(\mathbb{E}M_i\mathbf{1}_{\{\tau\le i\}} - \mathbb{E}\mathbb{E}\big(M_i\mathbf{1}_{\{\tau\le i-1\}}|\mathcal{F}_{i-1}\big)\big) =$$

$$\mathbb{E}M_0\mathbf{1}_{\{\tau=0\}} + \sum_{i=1}^{n}\big(\mathbb{E}M_i\mathbf{1}_{\{\tau\le i\}} - \mathbb{E}\mathbf{1}_{\{\tau\le i-1\}}\mathbb{E}\big(M_i|\mathcal{F}_{i-1}\big)\big) =$$

$$\mathbb{E}M_0\mathbf{1}_{\{\tau=0\}} + \sum_{i=1}^{n}\big(\mathbb{E}M_i\mathbf{1}_{\{\tau\le i\}} - \mathbb{E}M_{i-1}\mathbf{1}_{\{\tau\le i-1\}}\big) = \mathbb{E}M_n\mathbf{1}_{\{\tau\le n\}} = \mathbb{E}M_n = \mathbb{E}M_0.$$

□

REMARK 8e6. The Optional Sampling Theorem remains valid under more general conditions. For example, if τ is a stopping time with $\mathbb{E}\tau < \infty$ and (M_j) is a martingale with

$$\mathbb{E}\big(\big|M_{j+1} - M_j\big|\big|\mathcal{F}_j\big) \le C, \quad \mathbb{P} - a.s. \forall j,$$

then $\mathbb{E}M_\tau = \mathbb{E}M_0$. Do not think, however, that the latter property holds without appropriate assumptions. For example, if (S_j) is the simple random walk as above and

$$\tau_1 = \inf\{j \geq 0 : S_j = 1\},$$

then obviously $S_{\tau_1} = 1$ and thus $\mathbb{E}S_{\tau_1} \neq \mathbb{E}S_0 = 0$. It can be seen that $\mathbb{E}\tau_1 = \infty$ in this case.

In the proof to follow, we shall need similar results and notions for *continuous time* processes, i.e. families of random variables indexed by real numbers, rather than sequences of random variables indexed by integers. While the analogous theory is similar in spirit, it is much more technically involved.

Proof of Proposition 8e2. Let I be the set of true null hypotheses, $|I| = m_0$, and for any $t \in [0,1]$, define the total number of hypotheses, rejected at level t

$$r(t) = \#\{i : p_i \leq t\}$$

and the number of *erroneously* rejected hypotheses at level t

$$v(t) = \#\{i \in I : p_i \leq t\}.$$

Let $\mathrm{fdr}(t)$ be the (random) proportion of false discoveries:

$$\mathrm{fdr}(t) = \frac{v(t)}{r(t)}\mathbf{1}_{\{r(t)>0\}} = \frac{v(t)}{r(t) \vee 1}, \quad t \in [0,1]$$

Note that $r\big(p_{(i)}\big) = i$ and hence

$$\begin{aligned}\Big\{i : p_{(i)} \leq \frac{i}{m}\alpha\Big\} = \Big\{i : \frac{m}{i}p_{(i)} \leq \alpha\Big\} = \Big\{i : \frac{m}{r(p_{(i)})}p_{(i)} \leq \alpha\Big\} = \\ \Big\{i : \frac{m}{r(p_{(i)}) \vee 1}p_{(i)} \leq \alpha\Big\} := \Big\{i : Q(p_{(i)}) \leq \alpha\Big\},\end{aligned} \tag{8e1}$$

where

$$Q(t) = \frac{mt}{r(t) \vee 1}, \quad t \in [0,1].$$

Define $\tau_\alpha = \sup\big\{t \in [0,1] : Q(t) \leq \alpha\big\}$, and note that

$$k = \max\Big\{i : p_{(i)} \leq \frac{i}{m}\alpha\Big\} = \max\Big\{i : Q(p_{(i)}) \leq \alpha\Big\} = \max\big\{i : p_{(i)} \leq \tau_\alpha\big\},$$

where the latter holds since $Q(t)$ has negative jumps at $p_{(i)}$'s and increases otherwise (sketch a plot). Since both $r(t)$ and $v(t)$ are constant between $p_{(i)}$'s, it follows that $v(p_{(k)}) = v(\tau_\alpha)$ and $r(p_{(k)}) = r(\tau_\alpha)$. For the BH procedure (with $p_{(0)} := 0$)

$$\frac{V}{R \wedge 1} = \frac{v(p_{(k)})}{r(p_{(k)}) \vee 1} = \frac{v(\tau_\alpha)}{r(\tau_\alpha) \vee 1} = \frac{v(\tau_\alpha)}{\tau_\alpha}\frac{Q(\tau_\alpha)}{m} \leq \frac{v(\tau_\alpha)}{\tau_\alpha}\frac{\alpha}{m}$$

and hence

$$\mathrm{FDR} = \mathbb{E}\frac{V}{R \wedge 1} \leq \frac{\alpha}{m}\mathbb{E}M(\tau_\alpha), \tag{8e2}$$

where we defined $M(t) = v(t)/t$. Let $(\mathcal{F}_t)$, $t \in [0,1]$ be the filtration of σ-algebras generated by the events $\{p_i \leq t\}$. Since p_i's are independent and $p_i \sim U([0,1])$ for $i \in I$, for $s \leq t$ and $i \in I$

$$\mathbb{E}\Big(\mathbf{1}_{\{p_i \leq s\}}\big|\mathcal{F}_t\Big) = \mathbb{E}\Big(\mathbf{1}_{\{p_i \leq s\}}\big|\mathbf{1}_{\{p_i \leq t\}}\Big) = \mathbb{P}(p_i \leq s | p_i \leq t)\mathbf{1}_{\{p_i \leq t\}} = \frac{s}{t}\mathbf{1}_{\{p_i \leq t\}}$$

and hence

$$\mathbb{E}\big(M(s)|\mathcal{F}_t\big) = \mathbb{E}\Big(\frac{1}{s}\sum_{i\in I}\mathbf{1}_{\{p_i\le s\}}\big|\mathcal{F}_t\Big) = \frac{1}{s}\sum_{i\in I}\frac{s}{t}\mathbf{1}_{\{p_i\le t\}} = M(t).$$

Clearly $M(t)$ is measurable with respect to $\mathcal{F}_t$ for all $t \in [0,1]$, and hence $M(t)$ is a martingale in *reversed* time. Moreover, in reversed time, τ_α is the first hitting time of the level α by the process $Q(t)$, which is adapted to the time reversed filtration and hence is a stopping time with respect to the time reversed filtration. Then by the appropriate Optional Sampling Theorem

$$\mathbb{E}M(\tau_\alpha) = \mathbb{E}M(1) = \mathbb{E}v(1)/1 = m_0.$$

Plugging this equality into (8e2), we obtain the required claim. □

The recent text [**4**] is a good starting point for further exploration of the large scale inference problems.

Exercises

PROBLEM 8.1. Consider the problem of testing the hypothesis $H_0 : p = 1/2$ against the alternative $H_1 : p > 1/2$, given a sample $X_1, ..., X_n$ from $\text{Ber}(p)$ (n independent tosses of a coin).

(1) Suggest a reasonable level α test, using the CLT approximation
(2) Calculate the power of the test at $p = 3/4$
(3) For $n = 30$, $\alpha = 0.05$ find the critical region C of the test and evaluate your answer in the previous questions
(4) Find the approximate value of n so that the power in (2) is greater than 0.99
(5) Argue why the approximation in (1) is well justified by CLT, while (2) is not (consult the footnotes in Example 8a4)

PROBLEM 8.2. Let $x \sim \exp(\theta)$. It is required to test $H_0 : \theta = 1$ against $H_1 : \theta = 2$ and the following critical regions are suggested:

$$C_1 = [0, 1), \quad C_2 = [1, \infty), \quad C_3 = [0, 1/2) \cup (2, \infty), \quad C_4 = \mathbb{R} \setminus \{1\}.$$

Calculate the level and find the power functions of the corresponding tests.

PROBLEM 8.3. In each one of the following problems, find the N-P test based on the i.i.d. sample $X = (X_1, ..., X_n)$ and simplify it as much as possible:

(1) $H_0 : X \sim N(0, 1)$ against $H_1 : X \sim N(2, 1)$
(2) $H_0 : X \sim \chi^2_{(2)}$ against $H_1 : X \sim \chi^2_{(6)}$
(3) $H_0 : X \sim \exp(2)$ against $H_1 : X \sim \exp(3)$

Find the power of corresponding level $\alpha = 0.05$ tests.

PROBLEM 8.4. Let $X_i \sim N(\theta_i, 1)$ $i = 1, ..., 9$ independent r.v.'s.

(1) Find the NP level α test for the following problem:

$$H_0 : \theta_i = 0, \forall i$$
$$H_1 : \theta_i = \frac{1}{2}, i = 1, ..., 5 \text{ and } \theta_i = -1/2, i = 6, ..., 9$$

(2) Specify the test for $\alpha = 0.05$ and calculate its power

PROBLEM 8.5 (Slope detection problem). Let $X_i \sim N(i\theta, 1)$, $i = 1, ..., 10$ be independent r.v.'s. Find the level α N-P test for testing $H_0 : \theta = 0$ against $H_1 : \theta = 1$.

PROBLEM 8.6.

(1) Find the N-P level α test, based on a single sample X for

$$H_0 : X \sim \exp(1)$$
$$H_1 : X \sim f(x) = \frac{2}{2\pi} \exp(-x^2/2) I(x \in (0, \infty)).$$

(2) Does the test with the critical region $C = (0.8, 1.2)$ yields maximal power in the above problem. Find its power and level.

PROBLEM 8.7. Let X be a single sample from the c.d.f.

$$F(x) = \begin{cases} 0 & x < \theta \\ \frac{1}{4}(x-\theta)^2 & x \in [\theta, \theta+2) \\ 1 & x \geq \theta + 2 \end{cases}, \quad \theta \in \mathbb{R}.$$

(1) Find the most powerful level α test for the problem:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

with $\theta_1 > \theta_0$.

(2) Specify your test for $\theta_0 = 0, \theta_1 = 1$ and $\alpha = 0.36$. Calculate its power.

PROBLEM 8.8. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$ with known μ.

(1) Find the most powerful α level test for the following problem:

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma = \sigma_1,$$

with $\sigma_1 > \sigma_0$.

(2) How would your test change if $H_1 : \sigma = \sigma_2$ with $\sigma_2 > \sigma_1$?

(3) Do your answers above allow to conclude that your test is UMP for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma > \sigma_0$?

(4) Find the p-value for your test for $n = 15$, $\sigma_1^2 = 2$, $\sigma_0^2 = 1$, $\mu = 0$ and the realization x, which yields $\frac{1}{n-1} \sum_{i=1}^n x_i^2 = 1.25$

PROBLEM 8.9. Let X_1 be a single sample from the p.d.f.

$$f(x; \theta) = (2\theta x + 1 - \theta) I(x \in [0, 1]), \quad \theta \in [-1, 1].$$

(1) Find the most powerful level α test for testing $H_0 : \theta = 0$ against $H_1 : \theta = 1$.

(2) To test the hypotheses $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$, consider the critical region $C = \{x \in \mathbb{R} : x \geq 1/2\}$. Calculate the corresponding power function and significance level.

(3) Find the UMP level α test in the previous question (or argue that it doesn't exist)

PROBLEM 8.10. Let $Y \sim N(0,1)$ and define $X = \sqrt{\theta}|Y| + a$, where $a \in \mathbb{R}$ and $\theta \in \mathbb{R}_+$ are unknown parameters.

(1) Find the p.d.f. of X
(2) Assuming that $\theta = 1$, find the MLE of a
(3) Assuming that $\theta = 1$, find the GLRT for the problem:

$$H_0 : a = a_0$$
$$H_1 : a \neq a_0$$

PROBLEM 8.11. Let $X \sim Bin(n, \theta)$ and consider the hypotheses testing problem:

$$H_0 : \theta = 1/2$$
$$H_1 : \theta \neq 1/2$$

(1) Find that the GLRT has the form $\{|2X - n| > c\}$.
(2) Use the normal approximation to find the test with level $\alpha = 0.05$, when $n = 36$.
(3) Show that under H_0, approximately $2 \log \lambda(X) \sim \chi^2_1$.

PROBLEM 8.12. Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ and consider testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma \neq \sigma_0$

(1) Find the GLRT statistic and show that it is a function of

$$U = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2}.$$

(2) Show that the acceptance region of the GLRT has the form $\{c_1 < U < c_2\}$.
(3) Show that level α test is obtained if $F(c_2) - F(c_1) = 1 - \alpha$, where F is the c.d.f. of χ^2_{n-1} distribution.
(4) Show that $c_1 - c_2 = n \log(c_1/c_2)$ (**Hint:** use the result in (2)).

PROBLEM 8.13. The horse races fans claim that on the circle hippodrome, the outer track is disadvantageous against the inner track. Suppose that the hippodrome has 8 circle tracks, which are numbered from 1 to 8, where the 8-th track has the largest radius. Test whether the position of the track affects the win rate, summarized in the following table:

track	1	2	3	4	5	6	7	8
win rate	29	19	18	25	17	10	15	11

TABLE 2. Track position versus win rate

Hint: test the null hypothesis, under which the win rate has uniform distribution, against the alternative, under which the win rate has multinomial distribution.

number of girls	0	1	2	3	4	5	6
frequency	13	68	154	185	136	68	16

TABLE 3. Frequency table

PROBLEM 8.14. A random sample from families with 6 children is summarized in the following table:

(1) Test the hypothesis[15] that the number of girls has Bin$(6, 1/2)$ distribution. Calculate the p-value of the test.
(2) Solve the previous question for the hypothesis Bin$(6, p)$

PROBLEM 8.15. Let $X_1, ..., X_n \sim N(\mu_x, \sigma_x^2)$ and $Y_1, ..., Y_m \sim N(\mu_y, \sigma_y^2)$ be independent samples and consider testing $H_0 : \sigma_x = \sigma_y$ against $H_1 : \sigma_x \neq \sigma_y$ at the significance level α. None of the parameters are known.

(1) Find the MLEs of all the parameters under both H_0 and H_1
(2) Show that the GLRT statistic is equivalent to
$$\lambda(X) = \frac{\left(\hat{\sigma}^2\right)^{(m+n)/2}}{\left(\hat{\sigma}_x^2\right)^{n/2}\left(\hat{\sigma}_y^2\right)^{m/2}},$$
where $\hat{\sigma}^2$ is the MLE under H_0 and $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are MLEs under H_1.
(3) Show that
$$\lambda(X) = C\frac{\left(1 + \frac{n-1}{m-1}T\right)^{(n+m)/2}}{\left(\frac{n-1}{m-1}T\right)^{n/2}} =: f(T)$$
where C is a constant and
$$T(X, Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2}.$$
(4) Plot f and show that[16] the critical regions $C_k = \{\lambda(x, y) > k\}$ are equivalent to the critical regions
$$C_{k_1,k_2} = \{T(x, y) < k_1 \cup T(x, y) > k_2\}.$$

PROBLEM 8.16 (Bayes hypotheses testing). Assuming the prior π on Θ, derive the Bayes test for the problem of testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, which minimizes the sum of erros of the two types. Compare your result to GLRT.

[15] i.e. come up with a reasonable alternative to be tested versus H_0

[16] recall that T has F-distribution with $(m-1, n-1)$ degrees of freedom under H_0

Exams/solutions

a. 2009/2010 (A) 52303

Problem 1. (The sunrise problem of Laplace)

"What is the probability that the sun will rise tomorrow?" In this problem we shall explore how the question was answered by P-S. Laplace in 18-th century. Suppose that n days ago[17] a random variable R was sampled from uniform distribution $U([0,1])$ and since then the sun raises each day with probability R. More precisely, let X_i be the indicator of the sunrise on the i-th morning and assume that $X_1, X_2, ...,$ are *conditionally* independent r.v.'s given R, with $\mathrm{Ber}(R)$ distribution: for any $n \geq 1$

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n | R) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i | R) = \prod_{i=1}^{n} R^{x_i}(1-R)^{1-x_i}.$$

(1) Find the distribution of X_1.

Solution

X_1 takes values in $\{0,1\}$ and $\mathbb{P}(X_1 = 1) = \mathbb{E}\mathbb{P}(X_1 = 1|R) = \mathbb{E}R = 1/2$. Hence $X_1 \sim \mathrm{Ber}(1/2)$.

(2) Are $X_1, ..., X_n$ identically distributed ? independent ?

Solution

All X_i's are $\mathrm{Ber}(1/2)$, as was shown in the previous question. They are not independent:

$$\mathbb{P}(X_1 = 1, X_2 = 1) = \mathbb{E}\mathbb{P}(X_1 = 1, X_2 = 1|R) =$$

$$\mathbb{E}\mathbb{P}(X_1 = 1|R)\mathbb{P}(X_2 = 1|R) = \mathbb{E}R^2 = \int_0^1 u^2 du = 1/3,$$

while

$$\mathbb{P}(X_1 = 1) = \mathbb{E}\mathbb{P}(X_1 = 1|R) = \mathbb{E}R = 1/2,$$

which implies

$$1/3 = \mathbb{P}(X_1 = 1, X_2 = 1) \neq \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = 1/4.$$

[17] P-S.Laplace has literally taken n to be the number of days from the origin according to the Bible

(3) Explain why the conditional distribution of $S_n(X) = \sum_{i=1}^n X_i$, given R is $\mathrm{Bin}(R, n)$.

Solution

Since X_i's are conditionally independent $\mathrm{Ber}(R)$ r.v.'s, given R, their sum has conditionally Binomial distribution, given R (just as if [18] R were a number).

(4) Find[19] the p.m.f. of $S_n(X)$:

$$\mathbb{P}(S_n(X) = k) = ...?$$

Hint: *use the "tower property" of conditional expectations*

Solution

$S_n(X)$ has uniform distribution on $\{0, ..., n\}$:

$$\mathbb{P}(S_n(X) = k) = \mathbb{E}\mathbb{P}(S_n(X) = k|R) = \binom{n}{k} \mathbb{E}R^k(1-R)^{n-k} =$$
$$\frac{n!}{k!(n-k)!} \int_0^1 u^k(1-u)^{n-k} du = \frac{n!}{k!(n-k)!} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}$$

(5) Prove the following Bayes formula: for a random variable ξ with $\mathbb{E}|\xi| < \infty$ and discrete random variable η (taking integer values), show that

$$\mathbb{E}(\xi|\eta = k) = \frac{\mathbb{E}\xi I(\eta = k)}{\mathbb{P}(\eta = k)}.$$

Solution

[18] As X_i's are Bernoulli r.v., $S_n(X)$ takes values in $\{0, ..., n\}$. The conditional j.p.m.f. of the vector X, given R is

$$p_{X|R}(x; R) = \prod_{i=1}^n R^{x_i}(1-R)^{1-x_i} = R^{S_n(x)}(1-R)^{n-S_n(x)}, \quad x \in \{0,1\}^n.$$

Note that for all binary vectors x, such that $S_n(x) = k$, $p_{X|R}(x; R) = R^k(1-R)^{n-k}$ and there are $\binom{n}{k}$ such strings. Hence

$$\mathbb{P}(S_n = k|R) = \mathbb{P}\Big(\sum_{i=1}^n X_i = k|R\Big) = \binom{n}{k} R^k(1-R)^{n-k},$$

which verifies the claim.

[19] you will find the following integral useful: for nonnegative integers $k \le n$

$$\int_0^1 u^k(1-u)^{n-k} du = \frac{k!(n-k)!}{(n+1)!}$$

Denote

$$\phi(k) := \frac{\mathbb{E}\xi I(\eta = k)}{\mathbb{P}(\eta = k)}.$$

To verify the formula we have to check the orthogonality property:

$$\mathbb{E}\big(\xi - \phi(\eta)\big)h(\eta) = 0, \quad \forall h.$$

To this end, we have

$$\mathbb{E}\xi h(\eta) = \sum_k h(k)\mathbb{E}\xi I(\eta = k)$$

and on the other hand:

$$\mathbb{E}\phi(\eta)h(\eta) = \sum_k \phi(k)h(k)\mathbb{E}I(\eta = k) = \sum_k h(k)\frac{\mathbb{E}\xi I(\eta = k)}{\mathbb{P}(\eta = k)}\mathbb{P}(\eta = k) = \sum_k h(k)\mathbb{E}\xi I(\eta = k).$$

These two equalities verify the orthogonality property and thus prove the Bayes formula.

(6) Find $\mathbb{E}\big(R|S_n(X)\big)$.

Solution

Following the hint, we obtain:

$$\mathbb{E}\big(R|S_n(X) = k\big) = \frac{\mathbb{E}RI(S_n(X) = k)}{\mathbb{P}(S_n(X) = k)}.$$

Further,

$$\mathbb{E}RI(S_n(X) = k) = \mathbb{E}R\mathbb{E}\big(I(S_n = k)|R\big) = \mathbb{E}R\mathbb{P}(S_n = k|R) =$$

$$\mathbb{E}R\binom{n}{k}R^k(1-R)^{n-k} = \mathbb{E}\binom{n}{k}R^{k+1}(1-R)^{n-k} =$$

$$\binom{n}{k}\int_0^1 u^{k+1}u^{n+1-(k+1)}du = \frac{n!}{k!(n-k)!}\frac{(k+1)!(n-k)!}{(n+2)!} = \frac{k+1}{(n+2)(n+1)}$$

and as $\mathbb{P}(S_n(X) = k) = \frac{1}{n+1}$,

$$\mathbb{E}\big(R|S_n(X) = k\big) = \frac{k+1}{n+2}$$

(7) Calculate the probability "that the sun will rise tomorrow":

$$\mathbb{P}(X_{n+1} = 1|X_n = 1, ..., X_1 = 1).$$

Solution

Using the conditional independence of X_i's

$$\mathbb{P}(X_{n+1}=1|X_n=1,...,X_1=1)=\frac{\mathbb{P}(X_{n+1}=1,X_n=1,...,X_1=1)}{\mathbb{P}(X_n=1,...,X_1=1)}=$$
$$\frac{\mathbb{E}\mathbb{P}(X_{n+1}=1,X_n=1,...,X_1=1|R)}{\mathbb{E}\mathbb{P}(X_n=1,...,X_1=1|R)}=\frac{\mathbb{E}R^{n+1}}{\mathbb{E}R^n}=\frac{1/n+2}{1/(n+1)}=\frac{n+1}{n+2}.$$

Problem 2.

A circle C is drawn on a rectangular sheet of paper. Alice and Bob want to estimate the area A of the circle and suggest two different approaches. Alice exploits the fact that the sheet of paper is rectangular and the length of its side is known precisely (denote it by b). She suggests to sample i.i.d. random vectors from the uniform distribution over the sheet and to estimate A by the proportion of the vectors falling inside the circle.

(1) Show that if X and Y are i.i.d. r.v. with $U([0,b])$ distribution, then the random vector (X,Y) has uniform distribution (i.e. constant j.p.d.f.) on the planar rectangular $[0,b]\times[0,b]$

Solution

The j.c.d.f. of (X,Y) is constant outside the rectangular and otherwise is given by:

$$F_{X,Y}(x,y)=\mathbb{P}(X\le x,Y\le y)=\mathbb{P}(X\le x)\mathbb{P}(Y\le y)=\frac{1}{b^2}xy$$

whenever $x,y\in[0,b]\times[0,b]$ and thus the j.p.d.f. vanishes outside the rectangular and otherwise is given by

$$f_{XY}(x,y)=\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)=\frac{1}{b^2}.$$

(2) Alice generates i.i.d. random vectors $(X_1,Y_1),...,(X_n,Y_n)$ and writes down $Z_i=I\big((X_i,Y_i)\in C\big)$, $i=1,...,n$. Specify a statistical model, parameterized by A (the area of the circle), which supports this experiment. Find the minimal sufficient statistic.

Solution

Z_i's are i.i.d. Bernoulli r.v.'s with $\mathbb{P}_A(Z_i=1)=A/b^2$. Since the sheet is rectangular, the largest circle which is possible in this problem has area $A_{\max}:=\frac{\pi}{4}b^2$, i.e. the parametric space is $(0,A_{\max})$. As we already know, $\bar{Z}_n$ is the minimal sufficient statistic for this model.

(3) Alice's estimator of A is $\tilde{A}_n(Z)=b^2\bar{Z}_n$. Is it an unbiased estimator ? If yes, is it UMVUE ? Calculate the corresponding MSE risk.

Solution

$\mathbb{E}_A\tilde{A}_n(Z) = b^2A/b^2 = A$ and hence $\tilde{A}_n$ is unbiased. It is also UMVUE by the L-S theorem ($\bar{Z}_n$ is a complete sufficient statistic for the Bernoulli model). The MSE risk is the variance of $\bar{Z}_n$, i.e. $b^4\frac{1}{n}A/b^2(1-A/b^2) = \frac{1}{n}A(b^2-A)$.

Bob suggests the more traditional method: to measure the diameter of the circle with a ruler and calculate its area by means of the formula $A(D) = \frac{\pi}{4}D^2$. Since his measurement is noisy, he repeats it n times and assumes that his data $X_1, ..., X_n$ is a sample from $N(D, \sigma^2)$, where σ^2 is known. He suggests to estimate A by the method of moments: $\hat{A}_n(X) = \frac{\pi}{4}\bar{X}_n^2$.

(4) Is $\hat{A}_n(X)$ unbiased ? If not, calculate the bias and suggest an unbiased estimator $\hat{A}_n^u(X)$ on the basis of $\hat{A}_n(X)$.

Solution

The estimator $\hat{A}_n$ is biased: recall that $\bar{X}_n \sim N(D, \sigma^2/n)$ and thus

$$\mathbb{E}_D\frac{\pi}{4}(\bar{X}_n)^2 = \frac{\pi}{4}\big(D^2 + \sigma^2/n\big) = A + \frac{\pi}{4}\frac{\sigma^2}{n},$$

i.e. $b(A, \hat{A}_n) = \frac{\pi}{4}\frac{\sigma^2}{n}$. The corresponding unbiased estimator is

$$\hat{A}_n^u(X) = \frac{\pi}{4}(\bar{X}_n)^2 - \frac{\pi}{4}\frac{\sigma^2}{n}$$

(5) Are the MSE risks of the estimators $\hat{A}_n$ and $\hat{A}_n^u$

$$R(A(D), \hat{A}_n) = \mathbb{E}_D(\hat{A}_n(X) - A(D))^2$$

and

$$R(A_n(D), \hat{A}_n^u) = \mathbb{E}_D(\hat{A}_n^u(X) - A(D))^2$$

comparable ? If yes, which of the two estimators is inadmissible ?

Hint: *the answer can be given without calculations.*

Solution

Note that the two estimators differ by a constant and hence have the same variance. Thus the estimator with smaller squared bias has smaller risk: $\hat{A}_n^u$ has zero bias and hence $\hat{A}_n$ is inadmissible.

(6) Find the UMVUE estimator of the area $A(D)$.

Solution

Recall that $\bar{X}_n$ is a complete sufficient statistic. Hence by L-S theorem the R-B procedure yields the UMVUE. However, $\hat{A}^u_n(X)$ is already a function of the sufficient statistic $\bar{X}_n$, and hence it is the UMVUE.

(7) Calculate the C-R bound for the MSE risk of unbiased estimators of the area $A(D)$. Is the C-R bound attained by the risk of UMVUE from the previous question ? Discuss the same question for large n ?

Solution

As we have seen the Fisher information for the model is $I(D) = n/\sigma^2$. Moreover, $\psi(D) = \frac{\pi}{4}D^2$ and hence for any unbiased estimator $T(X)$ of $A(D)$

$$\mathrm{var}_D(T) \geq \frac{\big(\psi'(D)\big)^2}{I(D)} = \frac{\pi^2 D^2 \sigma^2}{4n}.$$

Let's calculate the risk of the UMVUE from the previous question:

$$\begin{aligned}
&\mathrm{var}_D\big(\hat{A}^u_n(X)\big) = \mathrm{var}_D\Big(\frac{\pi}{4}(\bar{X}_n)^2\Big) = \frac{\pi^2}{16}\mathrm{var}_D\Big((\bar{X}_n - D + D)^2\Big) \\
&= \frac{\pi^2}{16}\mathrm{var}_D\Big((\bar{X}_n - D)^2 + 2D(\bar{X}_n - D) + D^2\Big) \\
&= \frac{\pi^2}{16}\mathrm{var}_D\Big((\bar{X}_n - D)^2 + 2D(\bar{X}_n - D)\Big) \\
&= \frac{\pi^2}{16}\mathrm{var}_D\Big((\bar{X}_n - D)^2\Big) + \frac{\pi^2}{16}2\mathrm{cov}_D\Big((\bar{X}_n - D)^2, 2D(\bar{X}_n - D)\Big) + \\
&\qquad\qquad \frac{\pi^2}{16}\mathrm{var}_D\Big(2D(\bar{X}_n - D)\Big).
\end{aligned}$$

Recall that $\bar{X}_n - D \sim N(0, \sigma^2/n)$, hence

$$\mathrm{var}_D\Big((\bar{X}_n - D)^2\Big) = \mathbb{E}_D(\bar{X}_n - D)^4 - \Big(\mathbb{E}_D(\bar{X}_n - D)^2\Big)^2 = 3(\sigma/\sqrt{n})^4 - (\sigma/\sqrt{n})^4 = 2\sigma^4/n^2$$

$$\mathrm{var}_D\Big(2D(\bar{X}_n - D)\Big) = 4D^2\sigma^2/n$$

$$\mathrm{cov}_D\Big((\bar{X}_n - D)^2, 2D(\bar{X}_n - D)\Big) = 2D\mathbb{E}_D\big((\bar{X}_n - D)^2 - \sigma^2\big)(\bar{X}_n - D) = 0.$$

Assembling all parts together, we get

$$\mathrm{var}_D\big(\hat{A}^u_n(X)\big) = \frac{\pi^2\sigma^4}{8n^2} + \frac{\pi^2 D^2 \sigma^2}{4n}.$$

Hence the variance of the UMVUE in this case is strictly greater than the C-R bound. However, as n increases, the risk is dominated[20] by the second term, which is precisely the C-R lower bound.

[20] More precisely, $\mathrm{var}_D\big(\hat{A}^u_n(X)\big) = \mathrm{CR}(D) + o(1/n)$

(8) Are the estimators sequences $\hat{A}_n$ and $\hat{A}^u_n$, $n \geq 1$ consistent ?

Solution

By the LLN $\bar{X}_n \to D$ as $n \to \infty$. $\hat{A}_n$ is a continuous function of $\bar{X}_n$ and hence converges to A in probability, i.e. $\bar{A}_n$ is a consistent estimator of A. $\bar{A}^u_n$ differs from $\bar{A}_n$ be a deterministic sequence, which converges to zero. Hence $\bar{A}^u_n$ is also consistent.

Problem 3. (Change point detection)

A factory produces bottles of mineral water. By the health standards, the acidity of the water should not deviate from zero too much. The quality control department revealed that the acidity of the water raised to the level a, considered unhealthy. The last measurements of acidity has been taken n days ago, when it was still normal. The factory has to call out the bottles from the shops, and it has to find out how many days ago the acidity went wrong.

The home statistician suggests the following way to discover the change: he measures acidity of the water from the bottles, produced at each one of the last n days and assumes that the obtained measurements $X = (X_0, ..., X_n)$ are sampled from the following statistical model: all X_i's are independent, $X_0, ..., X_{\theta-1} \sim N(0,1)$ and $X_\theta, ..., X_n \sim N(a,1)$, where the unknown parameter $\theta \in \Theta = \{1, ..., n\}$ is the day, at which the change occurred[21].

(1) Show that this model has the likelihood:

$$L_n(x;\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left(-\frac{1}{2}\sum_{i=0}^{\theta-1} x_i^2 - \frac{1}{2}\sum_{i=\theta}^{n}(x_i - a)^2 \right), \quad x \in \mathbb{R}^{n+1}.$$

Solution

This likelihood is the j.p.d.f. of X

(2) Is this model identifiable ?

Solution

Clearly for each θ, the vector X has different j.p.d.f: e.g.

$$\mathbb{E}_\theta \sum_{i=0}^{n} X_i = a(n - \theta + 1)$$

which is a one to one function of θ. Hence the model is identifiable.

(3) Show that $T(X) = X_1, ..., X_{n-1}$ is a sufficient statistic.

Solution

[21] pay attention that $X_0 \sim N(0,1)$ and $X_n \sim N(a,1)$, which are the days the acidity is known to be normal and excessive respectively

Note that $X_0 \sim N(0,1)$ and $X_n \sim (a,1)$ independently of θ (these are the days at which the acidity is known precisely). Since the sample is i.i.d., the conditional distribution of X given $T(X) = (X_1, ..., X_{n-1})$ trivially does not depend on θ:

$$\mathbb{P}(X_0 \le x_0, ..., X_n \le x_n | X_1, ..., X_{n-1}) = I(X_1 \le x_1)...I(X_{n-1} \le x_{n-1})\Phi(x_0)\Phi(x_n - a),$$

where Φ is $N(0,1)$ c.d.f.

(4) Is $T(X) = (X_1, ..., X_{n-1})$ minimal sufficient ?

Solution

We have

$$L_n(x;\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left(-\frac{1}{2}\sum_{i=0}^{\theta-1} x_i^2 - \frac{1}{2}\sum_{i=\theta}^{n} x_i^2 + a\sum_{i=\theta}^{n} x_i - \frac{1}{2}(n-\theta+1)a^2 \right) =$$
$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left(-\frac{1}{2}\sum_{i=0}^{n} x_i^2 + a\sum_{i=\theta}^{n} x_i - \frac{1}{2}(n-\theta+1)a^2 \right).$$

Let x and y be vectors in $\mathbb{R}^{n+1}$, then

$$\frac{L_n(x;\theta)}{L_n(y;\theta)} = \exp\left(-\frac{1}{2}\sum_{i=0}^{n}(x_i^2 - y_i^2) + a\sum_{i=\theta}^{n}(x_i - y_i) \right)$$

is not a function of θ if and only if $x_i = y_i$ for all $i = 1, ..., n-1$. Hence $T(X)$ is the minimal statistic.

(5) For $n = 2$, find the MLE of θ.

Solution

Note that $L_n(x;\theta)$ is maximal if and only if

$$\phi(x;\theta) := \sum_{i=\theta}^{n} x_i + \frac{1}{2}a^2\theta$$

is maximal (why?). For $n = 2$, $\theta \in \Theta = \{1, 2\}$ and

$$\phi(x;1) = x_1 + x_2 + \frac{1}{2}a^2$$
$$\phi(x;2) = x_2 + a^2$$

and thus the MLE of θ is given by:

$$\hat{\theta}(X) = \begin{cases} 1 & \phi(X;1) \ge \phi(X;2) \\ 2 & \text{otherwise} \end{cases} = \begin{cases} 1 & X_1 \ge \frac{1}{2}a^2 \\ 2 & \text{otherwise} \end{cases}$$

(6) Suggest an unbiased estimator of θ for any $n \geq 2$.

Solution

Since

$$\mathbb{E}\sum_{i=0}^{n} X_i = \sum_{i=\theta}^{n} a = (n-\theta+1)a,$$

the estimator

$$\tilde{\theta}(X) = n + 1 - \frac{1}{a}\sum_{i=0}^{n} X_i$$

is unbiased.

b. 2009/2010 (B) 52303

Problem 1.

Let $Z_1, Z_2, ...$ be i.i.d. Ber(θ), $\theta \in (0,1)$ r.v.'s. We have seen that there is no unbiased estimator of the odds $\theta/(1-\theta)$, based on $Z_1, ..., Z_n$ for any *fixed* $n \geq 1$. Let

$$X_1 = \min\{n : Z_n = 0\},$$

be the number of tosses till the first 0 occurs. As is well known, X_1 has Geo($1-\theta$) distribution[22]:

$$\mathbb{P}_\theta(X_1 = k) = \theta^{k-1}(1-\theta), \quad k \in \{1, 2, ...\}.$$

(1) Find the MLE $\hat{\theta}_1$ of θ on the basis of X_1.

Solution

The log-likelihood

$$\log L(X_1; \theta) = (X_1 - 1)\log\theta + \log(1-\theta).$$

On the event $\{X_1 = 1\}$, the maximum is attained at $\hat{\theta}(1) = 0$ and otherwise $\hat{\theta}(X_1)$ solves

$$\frac{\partial}{\partial\theta}\Big((X_1-1)\log\theta + \log(1-\theta)\Big) = (X_1-1)\frac{1}{\theta} - \frac{1}{1-\theta} = 0,$$

which yields $\hat{\theta}_1 = 1 - 1/X_1$.

(2) Show that the "plug-in" estimator[23] $\hat{\eta}_1 := \hat{\theta}_1/(1-\hat{\theta}_1)$ is unbiased for $\eta(\theta) = \theta/(1-\theta)$. Explain the "contradiction" with the non-existence claim above.

Solution

[22] $\mathbb{E}_\theta = \frac{1}{1-\theta}$ and $\mathrm{var}_\theta(X_1) = \frac{\theta}{(1-\theta)^2}$

[23] since $\eta(\theta) = \theta/(1-\theta)$ is a one-to-one function, $\hat{\eta} := \hat{\theta}/(1-\hat{\theta})$ is in fact the MLE of $\eta(\theta)$.

$\hat\theta(X_1) = X_1 - 1$ is an unbiased estimator of $\theta/(1-\theta)$:

$$\mathbb{E}_\theta \hat\theta(X_1) = \mathbb{E}_\theta X_1 - 1 = \frac{1}{1-\theta} - 1 = \frac{\theta}{1-\theta}.$$

This does not contradict the above non-existence result, since the statistic $\hat\theta(X_1)$ is a function of the whole sequence $(Z_i)_{i\geq 1}$, rather than a fixed number of Z_i's.

(3) Is the obtained estimator above UMVUE ? Calculate its MSE risk.

Solution

Geo$(1-\theta)$ distribution belongs to the one parameter exponential family with $c(\theta) = \log\theta$, whose range clearly has a non-empty interior. Hence the sufficient statistic X_1 is complete and by the L-S theorem, $\hat\theta(X_1)$ is the UMVUE. Its MSE risk is given by

$$\mathrm{var}_\theta\big(\hat\theta(X_1)\big) = \mathrm{var}_\theta(X_1) = \frac{\theta}{(1-\theta)^2}.$$

(4) Calculate the C-R lower bound for the MSE risk of unbiased estimators of $\theta/(1-\theta)$. Is C-R bound attained ?

Solution

The Fisher information is

$$I_{X_1}(\theta) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2}\Big(\log\theta^{X_1-1}(1-\theta)\Big) = -\mathbb{E}_\theta \frac{\partial}{\partial\theta}\Big((X_1-1)\frac{1}{\theta} - \frac{1}{1-\theta}\Big) =$$
$$-\mathbb{E}_\theta\Big(-(X_1-1)\frac{1}{\theta^2} - \frac{1}{(1-\theta)^2}\Big) = \Big(\frac{1}{1-\theta} - 1\Big)\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)^2}.$$

The C-R bound for unbiased estimators of $\psi(\theta) = \theta/(1-\theta)$ is given by:

$$\mathrm{CR}(\theta) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)} = \left(\frac{1}{(1-\theta)^2}\right)^2 \theta(1-\theta)^2 = \frac{\theta}{(1-\theta)^2},$$

which coincides with the risk of UMVUE, i.e. the C-R bound is attained in this case.

(5) Encouraged by his progress, the statistician suggests to count the tosses till $m \geq 1$ zeros occur:

$$X_m = \min\Big\{j : \sum_{i=1}^{j}(1-Z_i) = m\Big\}.$$

Argue that the p.m.f. of X_m is given by:

$$p_{X_m}(k;\theta) = \mathbb{P}_\theta(X_m = k) = \binom{k-1}{m-1}\theta^{k-m}(1-\theta)^m, \quad k \geq m$$

Solution

The event $\{X_m = k\}$ occurs if and only if the last toss yields zero and there are $m-1$ zero tosses among $k-1$ tosses. The binomial coefficient is the number of strings with $m-1$ zeros among $k-1$ bits, and the claimed formula follows by independence.

(6) Find the MLE $\hat{\theta}_m$ of θ on the basis of X_m and the corresponding "plug-in" estimator $\hat{\eta}_m$ of $\eta(\theta)$

Solution

We have

$$\log L(X_m;\theta) = \log \binom{X_m - 1}{m-1} + (X_m - m)\log\theta + m\log(1-\theta).$$

Again, if $X_m = m$, then $\hat{\theta}(X_m) = 0$, otherwise it solves

$$(X_m - m)\frac{1}{\theta} = \frac{m}{1-\theta},$$

which gives:

$$\hat{\theta}_m = 1 - \frac{m}{X_m}.$$

Hence

$$\hat{\eta}_m = \frac{X_m}{m} - 1.$$

(7) (bonus[24] +5) Are the plug-in estimators $(\hat{\eta}_m)_{m\geq 1}$ consistent for $\theta/(1-\theta)$? If yes, find the asymptotic rate and the asymptotic error distribution.

Hint: Show that $X_m = \sum_{j=1}^{m} \xi_i$, where ξ_i are i.i.d. Geo$(1-\theta)$ r.v.'s and apply the LLN.

Solution

Define

$$\xi_1 = \min\{k : Z_k = 0\}$$
$$\xi_i = \min\{k > \xi_{i-1} : Z_k = 0\} - \xi_{i-1}, \quad j = 2, ..., m$$

[24] this question was excluded from the final version of the exam

the times between consecutive 0's. Clearly $X_m = \sum_{i=1}^m \xi_i$ and ξ_i's are independent Geo$(1-\theta)$ r.v.'s:

$$\mathbb{P}_\theta(\xi_1 = k_1, ..., \xi_m = k_m) =$$
$$\mathbb{P}_\theta(Z_1...Z_{k_1-1} = 1, Z_{k_1} = 0, ..., Z_{k_{m-1}+1}Z_{k_m-1} = 1, Z_{k_m} = 0) =$$
$$\underbrace{\mathbb{P}_\theta\big(Z_1...Z_{k_1-1} = 1, Z_{k_1} = 0\big)}_{\text{Geo}(1-\theta)} ... \underbrace{\mathbb{P}_\theta\big(Z_{k_{m-1}+1}Z_{k_m-1} = 1, Z_{k_m} = 0\big)}_{\text{Geo}(1-\theta)}.$$

Now by the LLN $\lim_{m\to\infty} X_m/m = 1/(1-\theta)$ in $\mathbb{P}_\theta$-probability and hence $X_m/m - 1 \to \theta/(1-\theta)$, which verifies consistency.

Note that $\mathbb{E}_\theta X_m/m = \mathbb{E}_\theta \bar{\xi}_m = 1/(1-\theta)$. Then by the CLT:

$$\sqrt{m}\Big(X_m/m - 1 - \frac{\theta}{1-\theta}\Big) = \sqrt{m}\Big(\bar{\xi}_m - \frac{1}{1-\theta}\Big) \xrightarrow[m\to\infty]{d} N(0, \text{var}_\theta(\xi_1)),$$

where $\text{var}_\theta(\xi_1) = \theta/(1-\theta)^2$.

Problem 2.

It is required to estimate the expected weight of the fish in a pond, using the fishes sample caught by the net. The weight of a fish in a pond is assumed to be a r.v. $X \sim U([0,\theta])$, $\theta > 0$.

Let Z be a random variable taking value 1 if the fish is captured by the net and 0 otherwise. It is known that smaller fishes are less likely to be caught by the net and a statistician assumes:

$$\mathbb{P}_\theta(Z = 1|X) = \frac{X}{\theta}.$$

(1) Calculate the probability that a fish is caught by the net $\mathbb{P}_\theta(Z = 1)$.

Solution

$$\mathbb{P}_\theta(Z = 1) = \mathbb{E}_\theta \mathbb{P}_\theta(Z = 1|X) = \mathbb{E}_\theta \frac{X}{\theta} = \frac{\theta/2}{\theta} = 1/2.$$

(2) Prove that the conditional p.d.f of X, given the event $\{Z = 1\}$, is

$$f(x;\theta) := \frac{d}{dx}\mathbb{P}_\theta(X \le x|Z = 1) = \frac{2x}{\theta^2} I(x \in [0,\theta])$$

Solution

We have

$$\mathbb{P}_\theta(X \le x|Z = 1) = \frac{\mathbb{P}_\theta(X \le x, Z = 1)}{\mathbb{P}_\theta(Z = 1)},$$

where

$$\mathbb{P}_\theta(X \le x, Z = 1) = \mathbb{E}_\theta I(X \le x) I(Z = 1) =$$
$$\mathbb{E}_\theta I(X \le x) \mathbb{E}_\theta\big(I(Z = 1)|X\big) = \mathbb{E}_\theta I(X \le x) \mathbb{P}_\theta\big(Z = 1|X\big) =$$
$$\mathbb{E}_\theta I(X \le x) \frac{X}{\theta} = \frac{1}{\theta} \int_0^\theta I(s \le x) s \frac{1}{\theta} ds.$$

Hence for $x \le \theta$,

$$\mathbb{P}_\theta(X \le x, Z = 1) = \frac{1}{\theta^2} \int_0^x s ds = \frac{1}{2} \Big(\frac{x}{\theta}\Big)^2,$$

and for $x > \theta$,

$$\mathbb{P}_\theta(X \le x, Z = 1) = 1/2.$$

Hence,

$$\mathbb{P}_\theta(X \le x | Z = 1) = \Big(\frac{x}{\theta}\Big)^2, \quad x \in [0, \theta].$$

(3) The statistician suggests that the weights $Y_1, ..., Y_n$ of the n caught fishes are i.i.d. r.v. sampled from $f(x; \theta)$ from the previous question. Explain how this statistical model fits the inference problem at hand and elaborate the assumptions it is based upon.

Solution

This model takes into account that smaller fishes are less likely to be included in the sample: the fact that a fish is caught tilts the distribution of its weight towards higher values, or more precisely, we observe random variables with the c.d.f. $\mathbb{P}_\theta(X \le x | Z = 1)$, rather than the original uniform c.d.f Moreover, the i.i.d. assumption means that the weights of the fishes are i.i.d. $U([0, \theta])$ and that the probability of catching the fish depends only on its own weight and not on the weights of the others:

$$\mathbb{P}_\theta(Z_i = 1 | X_1, X_2, ...) = \mathbb{P}(Z_i = 1 | X_i).$$

(4) Is the model identifiable ?

Solution

Obviously, identifiable, since e.g. the support of $f(x; \theta)$ is at one-to-one correspondence with $\theta > 0$: e.g. its support is proportional to θ.

(5) Find the minimal sufficient statistic[25]

Solution

The likelihood function is

$$L(Y;\theta) = \Big(\frac{2}{\theta^2}\Big)^n \Big(\prod_{i=1}^n Y_i\Big) I(\max_j Y_j \le \theta).$$

By the F-N factorization theorem, $\max_i Y_i$ is a sufficient statistic. Further,

$$L(x;\theta)/L(y;\theta) = \frac{\prod_{i=1}^n x_i I(\max_j x_j \le \theta)}{\prod_{i=1}^n y_i I(\max_j y_j \le \theta)}.$$

Since $\prod_{i=1}^n (x_i/y_i)$ does not depend on θ, the latter is independent of θ if and only if the ratio of indicators is independent of θ. Now minimality follows exactly as for the $U([0,\theta])$ case (see the lecture notes).

(6) Find the MLE for θ (twice the expected weight). Is it a function of the sufficient statistic from the previous question ?

Solution

Since $1/\theta^{2n}$ is a decreasing function of θ, the maximum of $L(Y;\theta)$ is attained at

$$\hat{\theta}_n(Y) = \max_i Y_i.$$

(7) Suggest an estimator of θ, using the method of moments.

Solution

The first moment of Y_1 is

$$\mathbb{E}_\theta Y_1 = \int_0^\theta s \frac{2}{\theta^2} s ds = \frac{2}{\theta^2}\frac{\theta^3}{3} = \frac{2}{3}\theta,$$

and the corresponding method of moments estimator

$$\tilde{\theta}_n(Y) = \frac{3}{2}\bar{Y}_n.$$

[25] This question was replaced by "Find two non-equivalent sufficient statistics" in the exam. One obvious sufficient statistic is all the data $Y_1, ..., Y_n$ and the other one is $\max_i Y_i$, as discussed in the solution

(8) Are your estimators consistent for θ as $n \to \infty$?

Solution

$\tilde{\theta}_n(Y)$, $n \geq 1$ is consistent by the LLN. To show that the MLE is consistent, we have to show that

$$\mathbb{P}_\theta(\theta - \max_i Y_i \geq \varepsilon) \to 0, \quad \forall \varepsilon.$$

To this end,

$$\mathbb{P}_\theta(\theta - \max_i Y_i \geq \varepsilon) = \mathbb{P}_\theta(\max_i Y_i \leq \theta - \varepsilon) = \prod_{i=1}^{n} \mathbb{P}_\theta(Y_i \leq \theta - \varepsilon) =$$

$$\left(\frac{\theta - \varepsilon}{\theta}\right)^{2n} = \left(1 - \frac{\varepsilon}{\theta}\right)^{2n} \xrightarrow{n\to\infty} 0, \quad \forall \varepsilon > 0.$$

Problem 3.

A cat is following a mouse on a real line. The current position of the mouse is a random variable $X \sim N(\mu, \sigma^2)$ (with known μ and σ^2). The cat is old and hence it doesn't see the mouse clearly: it observes $Y = X + Z$, where $Z \sim N(0, 1)$ is the "noise", independent of X.

(1) The cat tries to predict the position of the mouse by the simple predictor $g_c(Y) = Y$. Calculate the corresponding MSE.

Solution

The MSE of $g_c(Y)$ is $\mathbb{E}(Y - X)^2 = \mathbb{E}Z^2 = 1$.

(2) Find cat's optimal (in the MSE sense) predictor $g_c^*(Y)$ of the mouse's position X

Solution

The optimal predictor is the conditional expectation

$$g_c^*(Y) = \mathbb{E}(X|Y) = \mu + \frac{\sigma^2}{\sigma^2 + 1}(Y - \mu).$$

(3) Calculate the MSE of the cat's optimal prediction:

$$\mathbb{E}\Big(X - g_c^*(Y)\Big)^2$$

and compare with the MSE obtained in (1).

Solution

The MSE of the cat's predictor is:

$$\mathbb{E}(X - g_c^*(Y))^2 = \mathrm{var}(X|Y) = \sigma^2 - \frac{\sigma^4}{\sigma^2+1} = \frac{\sigma^2}{\sigma^2+1}.$$

(4) The cat crawls to the predicted position of the mouse $g_c^*(Y)$, so that the mouse cannot see him. Hence the mouse has to predict the position of the cat $g_c^*(Y)$, when it knows only its own position X. What is the mouse's optimal predictor $g_m^*(X)$ of the cat's position ?

Solution

The cat's optimal predictor is

$$g_m^*(X) = \mathbb{E}\Big(\mu + \frac{\sigma^2}{\sigma^2+1}(Y-\mu)\big|X\Big) = \mu + \frac{\sigma^2}{\sigma^2+1}\big(\mathbb{E}(Y|X) - \mu\big).$$

Since X and Z are independent, $\mathbb{E}(Y|X) = \mathbb{E}(X+Z|X) = \mathbb{E}(X|X) + \mathbb{E}(Z|X) = X$ and thus

$$g_m^*(X) = \mu + \frac{\sigma^2}{\sigma^2+1}\big(X - \mu\big).$$

(5) Calculate the MSE of the mouse's predictor:

$$\mathbb{E}\big(g_c^*(Y) - g_m^*(X)\big)^2.$$

Compare it to the cat's MSE from (3).

Solution

The MSE of the mouse's predictor is

$$\mathbb{E}\big(g_c^*(Y) - g_m^*(X)\big)^2 = \mathbb{E}\Big(\mu + \frac{\sigma^2}{\sigma^2+1}(Y-\mu) - \mu - \frac{\sigma^2}{\sigma^2+1}\big(X-\mu\big)\Big)^2 =$$
$$\left(\frac{\sigma^2}{\sigma^2+1}\right)^2 \mathbb{E}\big(Y-X\big)^2 = \left(\frac{\sigma^2}{\sigma^2+1}\right)^2 \mathbb{E}\big(Z\big)^2 = \left(\frac{\sigma^2}{\sigma^2+1}\right)^2.$$

The precision of the mouse is better, since $\sigma^2/(\sigma^2+1) < 1$.

(6) Suppose now that X and Z are no longer Gaussian, but still independent with $\mathbb{E}X = \mu$, $\mathrm{var}(X) = \sigma^2$ and $\mathbb{E}Z = 0$, $\mathrm{var}(Z) = 1$. Show by a counterexample, that the predictor from (2) is no longer optimal in general.

Solution

E.g. if $X \sim \text{Ber}(1/2)$ and Z has p.d.f. $f(x)$ with zero mean and unit variance, then, as saw,

$$\mathbb{E}(X|Y) = \mathbb{P}(X = 1|Y) = \frac{f(Y-1)}{f(Y-1)+f(Y)},$$

which certainly doesn't have the form of the predictor in (2).

(7) Argue that cat's *optimal* predictor of the mouse's position in the previous question is *at least* as precise as before (i.e. that its MSE is not greater than the MSE obtained in (3)), regardless of the distributions of X and Z.

Hint: what MSE the cat gets if it uses the predictor from (2).

Solution

If the cat uses the predictor from (2), it gets the same precision as before, since all the calculations involve only the mean and variance. But it is not necessarily its *optimal* predictor and hence it's optimal MSE may be even better.

(8) (bonus[26]+5) Show that mouse's predictor has smaller MSE than cat's predictor, if X and Z are independent, but otherwise have arbitrary distributions (with finite variances).

Solution

The MSE of the cat's predictor is

$$\begin{aligned}
&\mathbb{E}\big(X - \mathbb{E}(X|Y)\big)^2 = \\
&\mathbb{E}\Big(X - \mathbb{E}\big(\mathbb{E}(X|Y)|X\big) + \mathbb{E}\big(\mathbb{E}(X|Y)|X\big) - \mathbb{E}(X|Y)\Big)^2 \overset{\dagger}{=} \\
&\mathbb{E}\Big(X - \mathbb{E}\big(\mathbb{E}(X|Y)|X\big)\Big)^2 + \mathbb{E}\Big(\mathbb{E}(X|Y)|X\big) - \mathbb{E}(X|Y)\Big)^2 \geq \\
&\mathbb{E}\Big(\mathbb{E}\big(\mathbb{E}(X|Y)|X\big) - \mathbb{E}(X|Y)\Big)^2,
\end{aligned}$$

which is the MSE of the mouse's predictor. The equality † holds since the prediction error $\mathbb{E}\big(\mathbb{E}(X|Y)|X\big) - \mathbb{E}(X|Y)$ is orthogonal to any function of X and in particular to $X - \mathbb{E}\big(\mathbb{E}(X|Y)|X\big)$. Hence the mouse always sees the cat with smaller MSE.

[26] this question was excluded from the final version of the exam

c. 2009/2010 (A) 52314

Problem 1. Let $Z_1, Z_2, ...$ be i.i.d. Ber(θ), $\theta \in (0,1)$ r.v.'s. We have seen that there is no unbiased estimator of the odds $\theta/(1-\theta)$, based on $Z_1, ..., Z_n$ for any *fixed* $n \geq 1$. Let

$$X_1 = \min\{n : Z_n = 0\},$$

be the number of tosses till the first 0 occurs. As is well known, X_1 has Geo($1-\theta$) distribution[27]:

$$\mathbb{P}_\theta(X_1 = k) = \theta^{k-1}(1-\theta), \quad k \in \{1, 2, ...\}.$$

(1) Find the MLE of θ on the basis of X_1.

Solution

The log-likelihood

$$\log L(X_1; \theta) = (X_1 - 1)\log\theta + \log(1-\theta).$$

On the event $\{X_1 = 1\}$, the maximum is attained at $\hat{\theta}(1) = 0$ and otherwise $\hat{\theta}(X_1)$ solves

$$\frac{\partial}{\partial\theta}\Big((X_1 - 1)\log\theta + \log(1-\theta)\Big) = (X_1 - 1)\frac{1}{\theta} - \frac{1}{1-\theta} = 0,$$

which yields $\hat{\theta} = 1 - 1/X_1$.

(2) Find the MLE of $\eta(\theta) := \theta/(1-\theta)$.

Hint: $\eta(\theta)$ is a one-to-one function of $\theta \in (0,1)$.

Solution

The MLE $\hat{\eta}$ is $\hat{\eta} = \frac{\hat{\theta}}{1-\hat{\theta}} = \frac{1-1/X_1}{1/X_1} = X_1 - 1$.

(3) Show that MLE $\hat{\eta}$ is unbiased. Explain the "contradiction" with the non-existence claim above.

Solution

$\hat{\theta}(X_1) = X_1 - 1$ is an unbiased estimator of $\theta/(1-\theta)$:

$$\mathbb{E}_\theta\hat{\theta}(X_1) = \mathbb{E}_\theta X_1 - 1 = \frac{1}{1-\theta} - 1 = \frac{\theta}{1-\theta}.$$

This does not contradict the above non-existence result, since the statistic $\hat{\theta}(X_1)$ is a function of the whole sequence $(Z_i)_{i\geq 1}$, rather than a fixed number of Z_i's.

[27] $\mathbb{E}_\theta = \frac{1}{1-\theta}$ and $\mathrm{var}_\theta(X_1) = \frac{\theta}{(1-\theta)^2}$

(4) Is the obtained estimator above UMVUE ? Calculate its MSE risk.

Solution

Geo$(1-\theta)$ distribution belongs to the one parameter exponential family with $c(\theta) = \log\theta$, whose range clearly has a non-empty interior. Hence the sufficient statistic X_1 is complete and by the L-S theorem, $\hat\theta(X_1)$ is the UMVUE. Its MSE risk is given by

$$\mathrm{var}_\theta\big(\hat\theta(X_1)\big) = \mathrm{var}_\theta(X_1) = \frac{\theta}{(1-\theta)^2}.$$

(5) Calculate the C-R lower bound for the MSE risk of unbiased estimators of $\theta/(1-\theta)$. Is C-R bound attained ?

Solution

The Fisher information is

$$I_{X_1}(\theta) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2}\Big(\log\theta^{X_1-1}(1-\theta)\Big) = -\mathbb{E}_\theta\frac{\partial}{\partial\theta}\Big((X_1-1)\frac{1}{\theta} - \frac{1}{1-\theta}\Big) =$$
$$-\mathbb{E}_\theta\Big(-(X_1-1)\frac{1}{\theta^2} - \frac{1}{(1-\theta)^2}\Big) = \Big(\frac{1}{1-\theta}-1\Big)\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)^2}.$$

The C-R bound for unbiased estimators of $\psi(\theta) = \theta/(1-\theta)$ is given by:

$$\mathrm{CR}(\theta) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)} = \left(\frac{1}{(1-\theta)^2}\right)^2\theta(1-\theta)^2 = \frac{\theta}{(1-\theta)^2},$$

which coincides with the risk of UMVUE, i.e. the C-R bound is attained in this case.

(6) Encouraged by his progress, the statistician suggests to count the tosses till $m \geq 1$ zeros occur:

$$X_m = \min\Big\{n : \sum_{i=1}^{n}(1-Z_i) = m\Big\}.$$

Argue that the p.m.f. of X_m is given by:

$$p_{X_m}(k;\theta) = \mathbb{P}_\theta(X_m = k) = \binom{k-1}{m-1}\theta^{k-m}(1-\theta)^m, \quad k \geq m$$

Solution

The event $\{X_m = k\}$ occurs if and only if the last toss yields zero and there are $m-1$ zero tosses among $k-1$ tosses. The binomial coefficient is the number of strings with $m-1$ zeros among $k-1$ bits, and the claimed formula follows by independence.

(7) Find the MLE of $\eta(\theta) = \theta/(1-\theta)$ on the basis of X_m.

Solution

We have

$$\log L(X_m;\theta) = \log \binom{X_m - 1}{m-1} + (X_m - m)\log\theta + m\log(1-\theta).$$

Again, if $X_m = m$, then $\hat\theta(X_m) = 0$, otherwise it solves

$$(X_m - m)\frac{1}{\theta} = \frac{m}{1-\theta},$$

which gives:

$$\hat\theta(X_m) = 1 - \frac{m}{X_m}.$$

Hence

$$\hat\eta = \frac{X_m}{m} - 1.$$

(8) Do MLEs from the previous question form a consistent sequence of estimators of $\theta/(1-\theta)$?

Hint: Show that $X_m = \sum_{j=1}^m \xi_i$, where ξ_i are i.i.d. Geo$(1-\theta)$ r.v.'s and apply the LLN.

Solution

Define

$$\xi_1 = \min\{k : Z_k = 0\}$$
$$\xi_i = \min\{k > \xi_{i-1} : Z_k = 0\} - \xi_{i-1}, \quad j = 2, ..., m$$

the times between consecutive 0's. Clearly $X_m = \sum_{i=1}^m \xi_i$ and ξ_i's are independent:

$$\mathbb{P}_\theta(\xi_1 = k_1, ..., \xi_m = k_m) =$$
$$\mathbb{P}_\theta(Z_1 ... Z_{k_1 - 1} = 1, Z_{k_1} = 0, ..., Z_{k_{m-1}+1} Z_{k_m - 1} = 1, Z_{k_m} = 0) =$$
$$\underbrace{\mathbb{P}_\theta\big(Z_1 ... Z_{k_1-1} = 1, Z_{k_1} = 0\big)}_{\text{Geo}(1-\theta)} ... \underbrace{\mathbb{P}_\theta\big(Z_{k_{m-1}+1} Z_{k_m - 1} = 1, Z_{k_m} = 0\big)}_{\text{Geo}(1-\theta)}.$$

Now by the LLN $\lim_{m\to\infty} X_m/m = 1/(1-\theta)$ in $\mathbb{P}_\theta$-probability and hence $X_m/m - 1 \to \theta/(1-\theta)$, which verifies consistency.

(9) Find the asymptotic rate and the asymptotic error distribution of the MLE's found above

Hint: The hint from the previous question applies

Solution

Note that $\mathbb{E}_\theta X_m/m = \mathbb{E}_\theta \bar{\xi}_m = 1/(1-\theta)$. Then by the CLT:

$$\sqrt{m}\Big(X_m/m - 1 - \frac{\theta}{1-\theta}\Big) = \sqrt{m}\Big(\bar{\xi}_m - \frac{1}{1-\theta}\Big) \xrightarrow[m\to\infty]{d} N(0, \mathrm{var}_\theta(\xi_1)),$$

where $\mathrm{var}_\theta(\xi_1) = \theta/(1-\theta)^2$.

Problem 2. A factory started production of a new series of electric lamps and it is required to estimate their mean lifetime as soon as possible. For this purpose, the statistician suggests the following experiment: N lamps are powered on for a *known* period of a hours, during which the lifetimes of the burnt out lamps are recorded for the purpose of estimation.

Let n be the number of the burnt out lamps and assume that the lamp lifetime has p.d.f. $f(u;\theta)$, $u \in \mathbb{R}_+$, where θ is the unknown parameter.

(1) Assuming that $n \geq 1$, the statistician claims that the obtained data $X_1, ..., X_n$ is a sample from the p.d.f.

$$f_t(x;\theta) = \frac{f(x;\theta)}{F(a;\theta)} I(x \in [0,a]),$$

where $F(x;\theta) = \int_0^x f(u;\theta)du$. Explain his claim.

Solution

The p.d.f. $f_t(x;\theta)$ is the density of the *truncated* c.d.f. $\mathbb{P}(\xi \leq x | \xi \leq a)$, where $\xi \sim f(x;\theta)$. This corresponds to the fact that an observed lifetime ξ is reported only if it is less than a, i.e. is conditioned on the event $\{\xi \leq a\}$.

(2) Is the model identifiable if the lifetime has $U([0,\theta])$ distribution with $\theta > 0$?

Solution

For $U([0,\theta])$ with $\theta > a$,

$$f_t(x;\theta) = \frac{\frac{1}{\theta} I(x \in [0,\theta])}{\frac{a}{\theta}} I(x \in [0,a]) = \frac{1}{a} I(x \in [0, \min(a,\theta)]).$$

Thus if the parameter space $\Theta = \mathbb{R}_+$, $f_t(x;\theta)$ ceases to depend on θ for all θ large enough, which means that the model is not identifiable.

(3) Assuming hereafter Exp(θ) lifetime with the p.d.f.[28]

$$f(x;\theta) = \theta e^{-\theta x} I(x \in \mathbb{R}_+)$$

find the minimal sufficient statistic. Is it complete ?

Solution

The likelihood for this model is

$$L_n(x;\theta) = \frac{\theta^n \exp\Big(-\theta \sum_{i=1}^n x_i\Big)}{\prod_{i=1}^n \big(1 - e^{-\theta a}\big)}, \quad x \in \mathbb{R}_+^n,$$

which by the F-N factorization theorem implies that $S_n(x) = \sum_{i=1}^n x_i$ is a sufficient statistic. This likelihood clearly belongs to one parameter exponential family with $c(\theta) = -\theta$, whose range has a nonempty interior. Hence $S_n(X)$ is complete and thus minimal sufficient.

(4) Suggest a consistent sequence of estimators for $1/\theta$ (the mean lifetime).
Hint: consider the method of moments estimator based on the first two moments:

$$\mathbb{E}_\theta X_1 = \frac{1}{\theta} - \frac{a}{e^{a\theta} - 1}$$
$$\mathbb{E}_\theta X_1^2 = \frac{2}{\theta^2} - \frac{a^2 + 2a\frac{1}{\theta}}{e^{a\theta} - 1}$$

Solution

Let $\eta = 1/\theta$ for brevity, then combining the two equalities and eliminating the term $e^{a\theta} - 1$, we get

$$\frac{\eta - \mathbb{E}_\theta X_1}{2\eta^2 - \mathbb{E}_\theta X_1^2} = \frac{1}{a + 2\eta},$$

which, solved for η, gives:

$$\eta = \frac{a\mathbb{E}_\theta X_1 - \mathbb{E}_\theta X_1^2}{a - 2\mathbb{E}_\theta X_1}.$$

Since by the law of large numbers $\hat{m}_n^1(X) := \frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}_\theta X_1$ and $\hat{m}_n^2(X) := \frac{1}{n}\sum_{i=1}^n X_i^2 \to \mathbb{E}_\theta X_1^2$ in $\mathbb{P}_\theta$-probability, the estimators

$$\hat{\eta}_n(X) = \frac{a\hat{m}_n^1 - \hat{m}_n^2}{a - 2\hat{m}_n^1}$$

converge to $1/\theta$ in $\mathbb{P}_\theta$-probability for θ's with $\mathbb{E}_\theta X_1 \neq a/2$.

[28] the corresponding c.d.f. is $F(x;\theta) = (1 - e^{-\theta x})I(x \geq 0)$ and $\mathbb{E}_\theta = 1/\theta$, etc.

Unfortunately, none of the lamps in the experiment burnt out, i.e. $n = 0$ has been realized. Thus the model above is not applicable. However the statistician doesn't give up and suggests another approach as follows.

(5) Let Z_i be a random variable, taking value 1 if the i-th lamp among N ones, used in the experiment, does not burn out by a hours and 0 otherwise. Specify the statistical model, which assumes that all N lamps used in the experiment are i.i.d. Exp(θ) r.v.'s

Solution

The model $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}_+}$ is given by the j.p.m.f. of i.i.d. Ber($e^{-a\theta}$) r.v.'s.

(6) Suggest a consistent sequence of estimators of $1/\theta$, based on $Z_1, ..., Z_N$, $N \geq 1$.

Hint: Pay attention that $(\bar{Z}_N)_{N\geq 1}$ is a consistent sequence of estimators of θ.

Solution

Note that $\bar{Z}_N \to e^{-a\theta}$ in $\mathbb{P}_\theta$-probability. Since $x \mapsto 1/\log(1/x)$ is a continuous function on $x > 0$,

$$\hat{\theta}_N(Z) := \frac{a}{\log \frac{1}{\bar{Z}_N}} \xrightarrow[N\to\infty]{\mathbb{P}_\theta} \frac{1}{\theta},$$

which means that $\hat{\theta}_N$, $N \geq 1$ is a consistent estimator of $1/\theta$.

(7) Does there exist an unbiased estimator of $1/\theta$, based on $Z = (Z_1, ..., Z_N)$?

Solution

Let $T(Z)$ be an arbitrary statistic, then

$$\mathbb{E}_\theta T(Z) = \sum_{u\in\{0,1\}^N} T(u)\big(e^{-a\theta}\big)^{S_n(u)}\big(1 - e^{-a\theta}\big)^{n-S_n(u)}$$

is a bounded function of θ, e.g.:

$$\big|\mathbb{E}_\theta T(Z)\big| \leq 2^N \max_{u\in\{0,1\}^N} |T(u)|,$$

which means that it cannot equal $1/\theta$, which is an unbounded function.

Problem 3.

Let $X_i = \mu_i + \varepsilon_i$, $i = 1, ..., n$ where ε_i's are i.i.d. $N(0, \sigma^2)$ with known $\sigma^2 > 0$. The vector $\mu_1, ..., \mu_n$ is a signal to be detected in the noisy observations $(X_1, ..., X_n)$.

(1) Let $(\beta_i)_{i\geq 1}$ be a deterministic sequence of nonzero real numbers and assume that $\mu_i = \beta_i$, i.e. the received signal has a precisely known shape. Construct the most powerful level α test for

$$\begin{array}{ll} H_0 : \mu_i = 0, & \text{for all } i \in \{1, ..., n\} \\ H_1 : \mu_i = \beta_i, & \text{for all } i \in \{1, ..., n\} \end{array},$$

assuming that σ^2 is known.

Solution

We are faced with the problem of testing a simple hypothesis against a simple alternative, for which the N-P test is the most powerful. Denote by $\mu_{1:n} = (\mu_1, ..., \mu_n)$ and let $\mu^0_{1:n} := 0_{1:n}$ and $\mu^1_{1:n} := \beta_{1:n}$. Since $\sigma^2 > 1$ is known, the likelihood ratio is

$$\frac{L(x; \mu^0_{1:n})}{L(x; \mu^1_{1:n})} = \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \beta_i)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 \Big) = \exp\Big(\frac{1}{2\sigma^2}\sum_{i=1}^n x_i\beta_i - \frac{1}{2\sigma^2}\sum_{i=1}^n \beta_i^2 \Big).$$

Hence the N-P test rejects H_0 if and only if

$$\Big\{ \sum_{i=1}^n X_i\beta_i \geq c \Big\}.$$

Note that under H_0, the test statistic is Gaussian with zero mean variance $B_n := \sigma^2 \sum_{i=1}^n \beta_i^2$, hence the α level test is obtained with the critical value solving the equation:

$$\alpha = \mathbb{P}_0\Big(\sum_{i=1}^n X_i\beta_i / \sqrt{B_n} \geq c/\sqrt{B_n} \Big) = 1 - \Phi\Big(c/\sqrt{B_n} \Big),$$

which yields

$$c(\alpha) = \sqrt{B_n}\Phi^{-1}(1-\alpha) = \sigma\Phi^{-1}(1-\alpha)\sqrt{\sum_{i=1}^n \beta_i^2}.$$

(2) Assume now that $\mu_i = \mu\beta_i$, where $\mu > 0$ is the unknown amplitude. Is your test from the previous question applicable for testing:

$$\begin{array}{ll} H_0 : \mu_i = 0, & \text{for all } i \in \{1, ..., n\} \\ H_1 : \mu_i = \mu\beta_i, \text{with } \mu > 0 & \text{for all } i \in \{1, ..., n\} \end{array}.$$

If yes, is it UMP ?

Solution

The likelihood in this case is

$$L(x;\mu_{1:n}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu\beta_i)^2\right) =$$
$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2+\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i\beta_i-\frac{\mu^2}{2\sigma^2}\sum_{i=1}^{n}\beta_i^2\right),$$

which is one parametric exponential family with the canonical sufficient statistic $T(x) = \sum_{i=1}^{n} x_i\beta_i$ and monotonically increasing $c(\mu) = \mu$. Hence by the K-R theorem, the N-P level α test is UMP in this case.

(3) Formulate the sufficient and necessary conditions on the sequence $(\beta_i)_{i\geq 1}$, so that the power of your test at any $\mu > 0$ in the previous question converges to 1 as $n \to \infty$?

Solution

The power function of the N-P test is given by:

$$\pi(\mu,\delta^*) = \mathbb{P}_\mu\left(\sum_{i=1}^{n} X_i\beta_i \geq \sqrt{B_n}\Phi^{-1}(1-\alpha)\right) =$$
$$\mathbb{P}_\mu\Bigg(\underbrace{\Big(\sum_{i=1}^{n} X_i\beta_i - \mu B_n\Big)/\sqrt{B_n}}_{\sim N(0,1)} \geq \Phi^{-1}(1-\alpha) - \mu B_n/\sqrt{B_n}\Bigg) =$$
$$1-\Phi\Big(\Phi^{-1}(1-\alpha)-\mu\sqrt{B_n}\Big).$$

The power of the test converges to one for any $\mu > 0$ if and only if [29] $\sum_{i=1}^{n}\beta_i^2 \to \infty$ as $n \to \infty$. Note that the latter condition means that the signal waveform must not converge to zero too fast: if it does, the test is not *consistent*, i.e. its power cannot be made arbitrarily close to one by taking n large enough.

(4) Suppose now that the shape of the signal is unknown. Construct the level α GLRT test for

$$\begin{array}{ll} H_0 : \mu_i = 0, & \text{for all } i \in \{1,...,n\} \\ H_1 : \mu_i \neq 0, & \text{for some } i \in \{1,...,n\} \end{array}.$$

Solution

[29] tests with this property are called consistent (think why)

Under H_1, the MLE of $\mu_{1:n}$ is $\hat{\mu}^1_{1:n} = X_{1:n}$ and hence the GLRT statistic is

$$\log \lambda(X) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \hat{\mu}^1_i)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 = \frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2,$$

i.e. the test rejects H_0 if and only if

$$\Big\{\sum_{i=1}^n X_i^2 \geq c\Big\}.$$

Under H_0, the statistic $\sum_{i=1}^n X_i^2/\sigma^2$ has χ^2_n distribution (denote its c.d.f. by F_n) and hence α level test is obtained with the critical level, solving

$$\alpha = \mathbb{P}_0\Big(\sum_{i=1}^n X_i^2/\sigma^2 \geq c/\sigma^2\Big) = 1 - F_n\big(c/\sigma^2\big),$$

which gives

$$c(\alpha) = \sigma^2 F_n^{-1}(1-\alpha).$$

(5) Can the GLRT be constructed for the detection problem in (4), if σ^2 is unknown ? Explain.

Solution

The GLRT is not well defined, since under H_1, σ^2 and $\mu_{1:n}$ can be chosen to yield arbitrary large values of the likelihood, namely if $\hat{\mu}^1_{1:n} = X_{1:n}$ is taken,

$$L(x;\sigma^2,\hat{\mu}^1_{1:n}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \to \infty, \quad \sigma^2 \to 0.$$

d. 2009/2010 (B) 52314

Problem 1.

Let $Z_1, ..., Z_n$ be i.i.d. r.v.'s with $U([0,1])$ distribution and define

$$X_i = \theta + \theta Z_i, \quad i = 1, ..., n,$$

where $\theta > 0$ is the unknown parameter. It is required to estimate θ, given the sample $X = (X_1, ..., X_n)$.

(1) Show that $X \sim U([\theta, 2\theta])$, i.e. its p.d.f. is given by

$$f(x;\theta) = \frac{1}{\theta} I(x \in [\theta, 2\theta])$$

Solution

Let g be the p.d.f. of Z_1, then since $\theta > 0$,

$$\mathbb{P}(X_1 \leq u) = \mathbb{P}(\theta + \theta Z_1 \leq u) = \mathbb{P}(Z_1 \leq u/\theta - 1)$$

and thus

$$f(u;\theta) = \frac{d}{du}\mathbb{P}(Z_1 \leq u/\theta - 1) = \frac{1}{\theta}g(u/\theta - 1) = \frac{1}{\theta}I(u/\theta - 1 \in [0,1]) = \frac{1}{\theta}I(u \in [\theta, 2\theta]).$$

(2) Show that (X^*, X_*), where

$$X_* := \min_{i\in\{1,\ldots,n\}} X_i, \quad X^* := \max_{i\in\{1,\ldots,n\}} X_i,$$

is a sufficient statistic for θ

Solution

The sufficiency follows from the F-N theorem, since the likelihood has the form

$$L_n(x;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} I(x_i \in [\theta, 2\theta]) = \frac{1}{\theta^n}I(x^* \leq 2\theta)I(x_* \geq \theta),\ x \in \mathbb{R}^n,\ \theta \in \mathbb{R}_+$$

(3) Is (X^*, X_*) minimal sufficient ?

Solution

Let $x, y \in \mathbb{R}^n$, such that either $x^* \neq y^*$ or $x_* \neq y_*$, then

$$\frac{L_n(x;\theta)}{L_n(x;\theta)} = \frac{I(x^* \leq 2\theta)I(x_* \geq \theta)}{I(y^* \leq 2\theta)I(y_* \geq \theta)}$$

is clearly a function[30] of θ: e.g. if, say, $x^* = y^*$ and $x_* < y_*$, then it equals 1, if $\theta < x^*$ and takes value 0, if $\theta \in (x^*, y^*)$. Other cases are examined similarly.

(4) Sketch the likelihood as a function of θ and find the MLE $\hat{\theta}$. Calculate[31]its MSE risk.

Solution

$L_n(x;\theta)$ vanishes strictly decreases on the interval $[X^*/2, X_*]$ and vanishes elsewhere. Hence

$$\hat{\theta} = X^*/2.$$

[30]with e.g. the conventions $\frac{0}{0} := 0$ and $\frac{1}{0} := \infty$

The risk is given by

$$R(\theta,\hat{\theta}) = \mathrm{var}(\hat{\theta}) + (\mathbb{E}_\theta\hat{\theta} - \theta)^2 = \frac{1}{4}\mathrm{var}(X^*) + \Big(\frac{1}{2}\mathbb{E}_\theta X^* - \theta\Big)^2 =$$
$$\frac{\theta^2}{4}\mathrm{var}(Z^*) + \Big(\frac{1}{2}(\theta + \theta\mathbb{E}_\theta Z^*) - \theta\Big)^2 = \frac{\theta^2}{4}\frac{n}{(n+1)^2(n+2)} + \frac{\theta^2}{4}\Big(\frac{n}{n+1} - 1\Big)^2 =$$
$$\frac{\theta^2}{4}\frac{2n+1}{(n+1)^2(n+2)}$$

(5) Is the estimator $\tilde{\theta} = \frac{1}{3}(X_* + X^*)$ unbiased ?

Solution

Yes:

$$\mathbb{E}_\theta\frac{1}{3}(X_* + X^*) = \mathbb{E}_\theta\frac{1}{3}(2\theta + \theta Z_* + \theta Z^*) = \frac{1}{3}\Big(2\theta + \theta\frac{1}{n+1} + \theta\frac{n}{n+1}\Big) = \theta.$$

(6) Find the j.p.d.f. of (Z_*, Z^*)

Hint: Note that $\mathbb{P}(Z_* > u, Z^* \le v) = \Big(\mathbb{P}(Z_1 \in [u,v])\Big)^n$, for $u \le v$.

Solution

Since

$$F_{Z_*Z^*}(u,v) = \mathbb{P}(Z_* \le u, Z^* \le v) = \mathbb{P}(Z^* \le v) - \mathbb{P}(Z_* > u, Z^* \le v).$$

and,

$$\mathbb{P}(Z_* > u, Z^* \le v) = \Big(\mathbb{P}(Z_1 \in [u,v])\Big)^n = \begin{cases}(v-u)^n, & v > u\\ 0 & v \le u\end{cases}$$

we have

$$f_{Z_*Z^*}(u,v) = \frac{\partial^2}{\partial u\partial v}F_{Z_*Z^*}(u,v) = -\frac{\partial^2}{\partial u\partial v}\mathbb{P}(Z_* > u, Z^* \le v) =$$
$$n(n-1)(v-u)^{n-2}I(0 \le u \le v \le 1).$$

[31] You might find the following formulae useful:

$$\mathbb{E}Z_* = \frac{1}{n+1}, \qquad \mathbb{E}Z^* = \frac{n}{n+1}$$
$$\mathbb{E}(Z_*)^2 = \frac{2}{(n+1)(n+2)}, \qquad \mathbb{E}(Z^*)^2 = \frac{n}{n+2}$$
$$\mathrm{var}_\theta(Z_*) = \frac{n}{(n+1)^2(n+2)} \qquad \mathrm{var}(Z^*) = \frac{n}{(n+1)^2(n+2)}$$

(7) Write down the integral, required to compute[32]

$$\mathrm{cov}(Z_*, Z^*) = \frac{1}{(n+1)^2(n+2)}.$$

Solution

$$\mathbb{E}Z_*Z^* = n(n-1)\int_{\{(u,v):v\geq u\}} uv(v-u)^{n-2}dudv =$$

$$n(n-1)\int_0^1 u\left(\int_u^1 v(v-u)^{n-2}dv\right)du = \ldots = \frac{1}{n+2},$$

and

$$\mathrm{cov}(Z_*, Z^*) = \mathbb{E}Z_*Z^* - \mathbb{E}Z_*\mathbb{E}Z^* = \frac{1}{n+2} - \frac{1}{n+1}\frac{n}{n+1} = \frac{1}{(n+1)^2(n+2)}$$

(8) Calculate the MSE risk of $\tilde{\theta}$ and compare it to the risk of $\hat{\theta}$. On the basis of your calculation, if either of the estimators is inadmissible ?

Solution

Since the estimator is unbiased,

$$R(\theta,\tilde{\theta}) = \mathrm{var}(\tilde{\theta}) = \mathrm{var}\Big(\frac{1}{3}X_* + \frac{1}{3}X^*\Big) = \frac{\theta^2}{9}\mathrm{var}\big(Z_* + Z^*\big) =$$

$$\frac{\theta^2}{9}\Big(\mathrm{var}(Z_*) + 2\mathrm{cov}(Z_*, Z^*) + \mathrm{var}(Z^*)\Big) =$$

$$\frac{2\theta^2}{9}\Big(\frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2(n+2)}\Big) = \frac{2\theta^2}{9}\frac{1}{(n+1)(n+2)}.$$

To compare the risks, consider the ratio:

$$\frac{R(\theta,\hat{\theta})}{R(\theta,\tilde{\theta})} = \frac{9}{8}\frac{2n+1}{n+1}, \quad n \geq 1.$$

Note that the function $n \mapsto \frac{2n+1}{n+1}$ is increasing and attains its minimum $3/2$ at $n = 1$. Hence

$$\frac{R(\theta,\hat{\theta})}{R(\theta,\tilde{\theta})} \geq \frac{27}{16} > 1.68... \quad \forall n \geq 1$$

This shows that the MLE $\hat{\theta}$ is inadmissible and, moreover, is strictly inferior to $\tilde{\theta}$ asymptotically as $n \to \infty$.

[32] no need to proceed with the calculations

Problem 2.

Bob and Alice toss two different coins with heads probabilities θ_1 and θ_2, which Claude wants to estimate. In order to trick Claude, they reveal the outcomes of their tosses, without telling whose coin was tossed first. Claude knows that Bos's coin has greater heads probabilities than Alice's, i.e. $\theta_2 > \theta_1$.

To estimate $\theta := (\theta_1, \theta_2)$ Claude assumes that he observes n i.i.d. pairs $(X_1, Y_1), ..., (X_n, Y_n)$ with

$$X_i = Z_i A_i + (1 - Z_i) B_i$$
$$Y_i = (1 - Z_i) A_i + Z_i B_i,$$

where $A_i \sim \text{Ber}(\theta_1)$ (the tosses of Alice), $B_i \sim \text{Ber}(\theta_2)$ (the tosses of Bob) and $Z_i \sim \text{Ber}(1/2)$. All A_i's, B_i's and Z_i's are independent.

(1) Explain how this model fits the experiment and what is the role of Z_i's in it ?

Solution

If the event $\{Z_i = 1\}$ occurs, then $(X_i, Y_i) = (A_i, B_i)$, i.e. the first coin comes from Alice and the second from Bob, and if $\{Z_i = 1\}$ occurs, then $(X_i, Y_i) = (B_i, A_i)$. Hence Claude's model assumes that in fact Alice and Bob toss a third fair coin and swap their outcomes accordingly. Such a model supports any sequence of swaps.

(2) Find the distribution of X_1

Solution

X_1 is Bernoulli r.v. with

$$\mathbb{P}_\theta(X_1 = 1) = \mathbb{P}_\theta(Z_i = 1, A_i = 1) + \mathbb{P}_\theta(Z_i = 0, B_i = 1) =$$
$$\mathbb{P}_\theta(Z_i = 1)\mathbb{P}_\theta(A_i = 1) + \mathbb{P}_\theta(Z_i = 0)\mathbb{P}_\theta(B_i = 1) = \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2.$$

(3) Show that if Claude uses only $X_1, ..., X_n$, he gets non-identifiable model with respect to θ.

Solution

The distribution of X_1 and, by the i.i.d. property, of the whole vector $(X_1, ..., X_n)$ depends on θ, only through $\theta_1 + \theta_2$.

(4) Show that the j.p.m.f of (X_1, Y_1) is

$$p_{X_1,Y_1}(u,v;\theta) = \Big(\theta_1\theta_2\Big)^{uv}\Big((1-\theta_1)(1-\theta_2)\Big)^{(1-u)(1-v)} * \left(\frac{\theta_1(1-\theta_2)}{2} + \frac{(1-\theta_1)\theta_2}{2}\right)^{1-uv-(1-u)(1-v)}, \quad u,v \in \{1,0\}$$

Solution

We have

$$\mathbb{P}_\theta(X_1 = 1, Y_1 = 1) = \mathbb{P}(A_1 = 1, B_1 = 1) = \theta_1\theta_2,$$

and similarly

$$\mathbb{P}_\theta(X_1 = 0, Y_1 = 0) = \mathbb{P}(A_1 = 0, B_1 = 0) = (1-\theta_1)(1-\theta_2).$$

Further,

$$\mathbb{P}_\theta(X_1 = 0, Y_1 = 1) = \mathbb{P}_\theta(A_1 = 0, B_1 = 1, Z_1 = 1) + \mathbb{P}_\theta(A_1 = 1, B_1 = 0, Z_1 = 0) = (1-\theta_1)\theta_2\frac{1}{2} + \theta_1(1-\theta_2)\frac{1}{2},$$

and similarly,

$$\mathbb{P}_\theta(X_1 = 0, Y_1 = 1) = \theta_1(1-\theta_2)\frac{1}{2} + (1-\theta_1)\theta_2\frac{1}{2}.$$

(5) Find a two-dimensional sufficient statistic for estimating θ from $(X_1, Y_1), ..., (X_n, Y_n)$.

Solution

The j.p.m.f. of the r.v.'s $(X_1, Y_1), ..., (X_n, Y_n)$ is

$$p(x,y) = \Big(\theta_1\theta_2\Big)^{T_0(x,y)}\Big((1-\theta_1)(1-\theta_2)\Big)^{T_1(x,y)} * \left(\frac{\theta_1(1-\theta_2)}{2} + \frac{(1-\theta_1)\theta_2}{2}\right)^{n-T_0(x,y)-T_1(x,y)}, \quad x,y \in \{1,0\}^n,$$

where $T_0(x,y) = \sum_{i=1}^n x_i y_i$ and $T_1(x,y) = \sum_{i=1}^n (1-x_i)(1-y_i)$. Note that these two statistics is at one-to-one correspondence to the statistic $\Big(\sum_{i=1}^n X_i Y_i, \sum_{i=1}^n (X_i + Y_i)\Big)$, which by the F-N factorization theorem is sufficient. The usual calculation also shows that it is minimal sufficient.

(6) Is the model identifiable, if all the data $(X_1, Y_1), ..., (X_n, Y_n)$ is used?

Solution

Note that

$$s(\theta_1,\theta_2) := \mathbb{E}_\theta(X_1+Y_1) = \theta_1+\theta_2$$

and

$$r(\theta_1,\theta_2) := \mathbb{E}_\theta X_1Y_1 = \mathbb{P}_\theta(X_1=1,Y_1=1) = \mathbb{P}_\theta(A_1=1,B_1=1) = \theta_1\theta_2.$$

The function $(\theta_1,\theta_2)\mapsto(s,r)$ is invertible[33] on $\theta_2>\theta_1$: the inverse is obtained by solving[34] the quadratic equation $\theta^2-s\theta+r=0$:

$$\theta_1 = \frac{s-\sqrt{s^2-4r}}{2},\quad \theta_2 = \frac{s+\sqrt{s^2-4r}}{2}.$$

Hence the model is identifiable.

(7) How your answer to the previous question would change, if Claude doesn't know whose coin has greater heads probabilities ?

Solution

Notice that the distribution of the data is invariant with respect to permuting θ_1 and θ_2. Hence e.g. for $\theta=(1/2,1/3)$ and $\theta'=(1/3,1/2)$ one gets exactly the same distribution, i.e. the model is not identifiable. The condition $\theta_2>\theta_1$ reduces the parameter space and eliminates this ambiguity.

(8) Suggest consistent estimators for θ_1 and θ_2, based on all the data $(X_1,Y_1),...,(X_n,Y_n)$.

Solution

The formulas, obtained in the previous question, can be used to construct the method of moments estimators for θ_1 and θ_2:

$$\tilde\theta_1 = \frac{\overline{X+Y}-\sqrt{\overline{X+Y}^2-4\overline{XY}}}{2},$$
$$\tilde\theta_2 = \frac{\overline{X+Y}+\sqrt{\overline{X+Y}^2-4\overline{XY}}}{2},$$

where $\overline{X+Y} = \frac{1}{n}\sum_{i=1}^n(X_i+Y_i)$ and $\overline{XY} = \frac{1}{n}\sum_{i=1}^n X_iY_i$. However, the expression under the root may be negative with positive probability (e.g. on the event $\{X_1=Y_1=1,X_i=Y_i=0,i>1\}$). Hence the estimators must be modified to be well defined

[33] note that it is not invertible on $(0,1)\times(0,1)$!

[34] Note that only real roots are possible.

for any possible outcome of the experiment. For example,

$$\tilde{\theta}_1 = \frac{\overline{X+Y} - \sqrt{|\overline{X+Y}^2 - 4\overline{XY}|}}{2},$$
$$\tilde{\theta}_2 = \frac{\overline{X+Y} + \sqrt{|\overline{X+Y}^2 - 4\overline{XY}|}}{2},$$

The consistency follows by the LLN and continuity of the corresponding functions in the empirical means.

Problem 3.(Change Detection)

A plant produces bottles of mineral water. Production is monitored by a statistician, who samples the water acidity daily. She suspects that the production line stopped to work properly within the last n days and suggests to perform a test.

Let $(X_1, ..., X_n)$ denote the acidity of the water measured on each one of the days and let $\theta \in \{1, ..., n\}$ denote the day index at which the change have occurred. The statistician wants to test the hypothesis:

$$H_0 : X_1, ..., X_n \text{ are i.i.d. } N(0,1) \text{ r.v.'s}$$

against the alternative

$$H_1 : X_1, ..., X_n \text{ are independent, } \begin{cases} X_1, ..., X_{\theta-1} \sim N(0,1) \\ X_\theta, ..., X_n \sim N(a,1) \end{cases} \quad \theta \in \{1, ..., n\}$$

where $a > 0$ is a known constant.

(1) Find the likelihood function of the data under the alternative

Solution

$$L(x;\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{\theta-1} x_i^2 - \frac{1}{2}\sum_{i=\theta}^{n}(x_i - a)^2 \right),$$

where $\sum_{i=1}^{0}(...) = 0$ is understood.

(2) For $n = 1$, find the most powerful level α test

Solution

When $n = 1$, we are faced with the simple hypothesis testing problem:

$$H_0 : X_1 \sim N(0,1)$$
$$H_1 : X_1 \sim N(a,1),$$

for which the most powerful test is given by the N-P likelihood ratio test, rejecting H_0 iff

$$X_1 > c.$$

The α-level test is obtained by choosing:

$$\mathbb{P}_0\big(X_1 > c\big) = \alpha \quad \Longrightarrow \quad c := \Phi^{-1}(1-\alpha).$$

(3) Find the α level test, using the test statistic $\bar{X}_n$ and calculate its power function.

Solution

The test rejects H_0 if and only if

$$\bar{X}_n > c.$$

Note that $\bar{X}_n$ is an $N(0,1/n)$ r.v. under H_0 and hence the α-level test is obtained by choosing c, which solves

$$\mathbb{P}_0(\bar{X}_n > c) = \mathbb{P}_0(\underbrace{\sqrt{n}\bar{X}_n}_{\sim N(0,1)} > \sqrt{n}c) = 1 - \Phi(c\sqrt{n}) = \alpha,$$

that is

$$c(\alpha) = \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha).$$

Under H_1, $\bar{X}_n$ is a Gaussian r.v. with unit variance and mean

$$\mathbb{E}_\theta \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta X_i = \frac{1}{n}\sum_{i=\theta}^{n} a = a\frac{n-\theta+1}{n}.$$

The power function of the test (for $\theta \in \{1,...,n\}$) is given by

$$\pi(\theta) = \mathbb{P}_\theta\big(\bar{X}_n > c(\alpha)\big) = \mathbb{P}_\theta\Big(\underbrace{\sqrt{n}\Big(\bar{X}_n - a\frac{n-\theta+1}{n}\Big)}_{\sim N(0,1)} > \sqrt{n}\Big(c(\alpha) - a\frac{n-\theta+1}{n}\Big)\Big) =$$

$$1 - \Phi\Big(\sqrt{n}\Big(c(\alpha) - a\frac{n-\theta+1}{n}\Big)\Big) = 1 - \Phi\Big(\Phi^{-1}(1-\alpha) - a\frac{n-\theta+1}{\sqrt{n}}\Big).$$

(4) Prove that the GLRT test with a critical value c rejects H_0 iff

$$\max_{\theta\in\{1,...,n\}} \sum_{i=\theta}^{n}(X_i - a/2) > c.$$

Solution

The GLRT statistic in this case is:

$$\frac{\sup_{\theta\in\Theta_1} L(x;\theta)}{\sup_{\theta\in\Theta_0} L(x)} = \frac{\max_{\theta\in\{1,\ldots,n\}} \exp\left(-\frac{1}{2}\sum_{i=1}^{\theta-1} x_i^2 - \frac{1}{2}\sum_{i=\theta}^{n}(x_i-a)^2\right)}{\exp(-\frac{1}{2}\sum_{i=1}^{n} x_i^2)} =$$

$$\max_{\theta\in\{1,\ldots,n\}} \exp\left(a\sum_{i=\theta}^{n}(x_i - a/2)\right)$$

and the GLRT test rejects H_0 if and only if

$$\max_{\theta\in\{1,\ldots,n\}} \sum_{i=\theta}^{n}(X_i - a/2) > c,$$

where c must be chosen to meet the significance error requirement.

Note that this test has a nice implementation advantage, being *sequential*: H_0 is rejected once the cumulative sum

$$\sum_{i=1}^{m}(X_i - a/2), \quad m = 1, 2, \ldots$$

exceeds threshold c.

(5) Find the GLRT statistic, if a is unknown and is to be considered as a parameter as well.

Solution

The test statistic in this case is:

$$\frac{\max_{\theta\in\{1,\ldots,n\}} \max_{a\in\mathbb{R}} \exp\left(-\frac{1}{2}\sum_{i=1}^{\theta-1} x_i^2 - \frac{1}{2}\sum_{i=\theta}^{n}(x_i-a)^2\right)}{\exp(-\frac{1}{2}\sum_{i=1}^{n} x_i^2)} =$$

$$\max_{\theta\in\{1,\ldots,n\}} \max_{a\in\mathbb{R}} \exp\left(a\sum_{i=\theta}^{n}(x_i - a/2)\right)$$

For each θ, the maximum w.r.t. a is attained at

$$a^* = \frac{1}{n-\theta+1}\sum_{i=\theta}^{n} x_i$$

and thus the GLRT rejects H_0 if and only if

$$\max_{\theta\in\{1,\ldots,n\}} \frac{1}{n-\theta+1}\left(\sum_{i=\theta}^{n} X_i\right)^2 > c.$$

e. 2010/2011 (A) 52303

Problem 1.

Let $X_1, ..., X_n$ be a sample from the Gaussian distribution $N(\theta, \theta^2)$, where $\theta > 0$ is the unknown parameter.

(1) Find a sufficient statistic, coarser than the sample X itself

Solution

The likelihood function is

$$L(x;\theta) = \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n \exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^n (x_i-\theta)^2\right) = \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n \exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 + \frac{1}{\theta}\sum_{i=1}^n x_i - n\right)$$

By F-N theorem, the statistic $T(X) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ *is sufficient (in fact, minimal).*

(2) Calculate the Fisher information for the sample and derive the lower bound for the unbiased estimators of θ.

Solution

By the i.i.d. property, $I_X(\theta) = nI_{X_1}(\theta)$. *Further,*

$$\log f_{X_1}(X_1;\theta) = -\log\sqrt{2\pi} - \log\theta - \frac{1}{2\theta^2}X_1^2 + \frac{1}{\theta}X_1 - 1/2$$

Hence

$$\frac{\partial}{\partial\theta}\log f_{X_1}(X_1;\theta) = -\frac{1}{\theta} + \frac{1}{\theta^3}X_1^2 - \frac{1}{\theta^2}X_1$$

and

$$\frac{\partial^2}{\partial\theta^2}\log f_{X_1}(X_1;\theta) = \frac{1}{\theta^2} - \frac{3}{\theta^4}X_1^2 + \frac{2}{\theta^3}X_1.$$

Note that $\mathbb{E}_\theta X_1^2 = \mathrm{var}_\theta(X_1) + (\mathbb{E}_\theta X_1)^2 = 2\theta^2$ *and so*

$$I_{X_1}(\theta) = -\mathbb{E}_\theta\frac{\partial^2}{\partial\theta^2}\log f_{X_1}(X_1;\theta) = -\left(\frac{1}{\theta^2} - \frac{3}{\theta^4}2\theta^2 + \frac{2}{\theta^3}\theta\right) = \frac{3}{\theta^2}.$$

For any unbiased estimator $\hat\theta_n(X)$ *of* θ, $\mathrm{var}_\theta(\hat\theta_n) \geq \frac{1}{3}\frac{\theta^2}{n}$.

(3) Show that the estimator $\bar{X}_n$ of θ is not efficient, i.e. its risk does not attain the C-R bound.

Solution

$\bar{X}_n$ is unbiased and hence its risk is the variance $\mathrm{var}_\theta(\bar{X}_n) = \theta^2/n$, which is strictly greater than C-R bound for all $\theta > 0$. Hence it is not efficient.

(4) Is $\bar{X}_n$ efficient asymptotically as $n \to \infty$, i.e.

$$\lim_n \frac{n/I_{X_1}(\theta)}{\mathrm{var}_\theta(\bar{X}_n)} = 1$$

Solution

$\bar{X}_n$ is not efficient asymptotically either:

$$\frac{n/I_{X_1}(\theta)}{\mathrm{var}_\theta(\bar{X}_n)} = 1/3, \quad \forall n \geq 1$$

(5) Explain how the estimator

$$T_n(X) = \sqrt{\frac{1}{2}\frac{1}{n}\sum_{i=1}^n X_i^2},$$

is obtained by the method of moments and show that the sequence (T_n) is consistent for θ.

Solution

The estimator can be obtained by the method of moments using $\mathbb{E}_\theta X_1^2 = 2\theta^2$. By the law of large numbers, $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{\mathbb{P}_\theta} 2\theta^2$ and since $u \mapsto \sqrt{u/2}$ is a continuous function, $T_n(X) \xrightarrow{\mathbb{P}_\theta} \sqrt{\frac{1}{2}2\theta^2} = \theta$, i.e. T_n is consistent for θ.

(6) Show that

$$\mathrm{var}_\theta(X_1^2) = 6\theta^4$$

Hint: note that $X_1 = \theta(1+\xi)$ with $\xi \sim N(0,1)$ and recall that $\mathbb{E}\xi^4 = 3$ and $\mathbb{E}\xi^3 = 0$.

Solution

Note that $X_1 = \theta(1+\xi)$ *and*

$$\mathrm{var}_\theta(X_1^2) = \theta^4 \mathrm{var}\big((1+\xi)^2\big).$$

$\mathbb{E}(1+\xi)^2 = 2$ *and*

$$\mathrm{var}\big((1+\xi)^2\big) = \mathbb{E}(1+\xi)^4 - \big(\mathbb{E}(1+\xi)^2\big)^2 = \mathbb{E}(1+\xi)^4 - 4.$$

Moreover,

$$\mathbb{E}(1+\xi)^4 = \mathbb{E}(1+4\xi+6\xi^2+4\xi^3+\xi^4) = 1+6+3 = 10,$$

and the claim follows.

(7) Using the Delta method, find the asymptotic error distribution for (T_n) and the corresponding rate. Compare your result with the estimator in (3) and with the C-R bound from (2). Explain.

Solution

By the LLN, $S_n(X) := \frac{1}{n}\sum_{i=1}^n \frac{1}{2}X_i^2 \xrightarrow{\mathbb{P}} \theta^2$ *and by the CLT,*

$$\sqrt{n}(S_n(X) - \theta^2) \xrightarrow{d} N\left(0, \mathrm{var}_\theta\Big(\frac{1}{2}X_1^2\Big)\right) = N\left(0, \frac{3}{2}\theta^4\right)$$

Hence by the Delta method, applied to $g(u) = \sqrt{u}$,

$$\sqrt{n}\big(T_n - \theta\big) = \sqrt{n}\Big(g(S_n) - g(\theta^2)\Big) \xrightarrow{d} N\left(0, \big(g'(\theta^2)\big)^2 \frac{3}{2}\theta^4\right) =$$
$$N\left(0, \Big(\frac{1}{2}\frac{1}{\sqrt{\theta^2}}\Big)^2 \frac{3}{2}\theta^4\right) = N\left(0, \frac{3}{8}\theta^2\right).$$

The asymptotic variance is better than that of $\bar{X}_n$. *Comparing to the C-R bound, found above, is strictly speaking*[35] *irrelevant, since* T_n*'s are not unbiased (revealed by an additional calculation).*

Problem 2.

A coin is tossed an unknown number of times $\nu \in \{1, 2, ...\}$ and the number of heads X is revealed.

(1) A statistician assumes that $X \sim \mathrm{Bin}(\nu, p)$, where both ν and $p \in (0,1)$ are unknown parameters. What are the assumptions of the model. Specify the parametric space.

Solution

[35] in fact, the sequence (T_n) is *asymptotically* unbiased and hence the C-R bound is still very much relevant (alas, not in the scope of our course)

The tosses are assumed to be i.i.d. with probability of heads p. The parameter (ν, p) takes values in $(0,1) \times \mathbb{N}$.

(2) If p is known and ν is unknown, is the model identifiable ? If both p and ν are unknown, is the model identifiable ?

Solution

If ν is unknown and p is known, the model is identifiable: the number of integers, on which the p.m.f. of X is positive, equals $\nu + 1$ and hence for two different values of ν different p.m.f.'s emerge. Alternatively, consider e.g. the statistic $T(X) = \mathbf{1}_{\{X=0\}}$. The function $\nu \mapsto \mathbb{E}_\nu T(X) = \mathbb{P}_\nu(X = 0) = (1-p)^\nu$ is a one-to-one function of ν.

The model is identifiable also if both parameters are unknown. Let $(p, \nu) \neq (p', \nu')$. If $\nu \neq \nu'$, the corresponding p.m.f's are different by the same argument as in the previous question: the number of nonzero probabilities is different, regardless of p and p'. If $\nu = \nu'$ and $p \neq p'$, the p.m.f.'s are different, since e.g. their means are different: $\nu p \neq \nu p'$.

(3) Assume that $p \in (0,1)$ is known and denote by $\hat{\nu}(X)$ the MLE of ν. Find $\hat{\nu}(0)$.

Solution

The likelihood function is

$$L(x; \nu) = \frac{\nu!}{x!(\nu - x)!} p^x (1-p)^{\nu - x}, \quad x \in \{0, ..., \nu\},$$

and at $x = 0$, $L(0; \nu) = (1-p)^\nu$. Hence

$$\hat{\nu}(0) = \operatorname{argmax}_{\nu \in \mathbb{N}} (1-p)^\nu = 1,$$

where the latter equality holds, since $\nu \mapsto (1-p)^\nu$ is a decreasing function.

(4) Calculate $\hat{\nu}(1)$, if p is known and $p = 1 - e^{-1} = 0.6321...$
How your answer generalizes to the case of $p = 1 - e^{-\ell}$, $\ell \in \{1, 2, ...\}$?
To $p = 1 - e^{-1/\ell}$, $\ell \in \{1, 2, ...\}$? To any $p \in (0,1)$?

Solution

$$\hat{\nu}(1) = \operatorname{argmax}_{\nu \in \mathbb{N}} \log L(1; \nu) = \operatorname{argmax}_{\nu \in \mathbb{N}} \log \left(\nu p (1-p)^{\nu - 1} \right) =$$
$$\operatorname{argmax}_{\nu \in \mathbb{N}} \left(\log \nu + \log p + (\nu - 1) \log(1-p) \right) =: \operatorname{argmax}_{\nu \in \mathbb{N}} \phi(\nu).$$

The function $\phi(u)$ is well defined for all $u \geq 1$ (and not just for integer arguments) and

$$\phi'(u) = 1/u + \log(1-p),$$

which vanishes at $u^* = -1/\log(1-p)$. *Moreover, since* $\phi''(u) = -1/u^2 < 0$, $\phi(u)$ *attains its local maximum at* u^*. *As* $\lim_{u\to\infty}\phi(u) = -\infty$, *the global maximum over* $u \in [1,\infty)$ *is attained at* $\max\left(1, u^*\right)$.

Note that

$$\max_{\nu\in\mathbb{N}}\phi(\nu) \le \max_{u\in[1,\infty)}\phi(u),$$

since the maximum in the right hand side is taken over a larger set. For $p = 1 - e^{-1}$, $u^* = 1$ *is an integer and the latter inequality is attained, i.e.*

$$\hat{\nu}(1) = \operatorname{argmax}_{\nu\in\mathbb{N}}\phi(\nu) = \operatorname{argmax}_{u\in[1,\infty)}\phi(u) = 1.$$

Similarly, for $p = 1 - e^{-\ell}$ *with* $\ell \in \mathbb{N}$, $u^* = 1/\ell$ *and hence* $\hat{\nu}(1) = \max(1, 1/\ell) = 1$. *For* $p = 1 - e^{-1/\ell}$, $\ell \ge 1$ *we get* $u^* = \ell$.

In the general case, $p \in (0,1)$

$$\hat{\nu}(1) = \begin{cases} \max(1, \lfloor u^* \rfloor), & \text{if } \log L(1; \lfloor u^* \rfloor) \ge L(1; \lceil u^* \rceil) \\ \max(1, \lceil u^* \rceil), & \text{if } \log L(1; \lfloor u^* \rfloor) < L(1; \lceil u^* \rceil) \end{cases},$$

where $\lfloor u^* \rfloor$ *denotes the greatest integer less than or equal to* u^* *and* $\lceil u^* \rceil$ *denotes the smallest integer greater or equal to* u^*.

(5) Explain why the model does not belong to the exponential family, if ν is unknown

Solution

If ν *is unknown, the support of the Binomial p.m.f. depends on the parameter and hence doesn't fit the exponential family form.*

(6) Show that if both p and ν are unknown, X is a complete statistic.

Solution

Suppose $\mathbb{E}_{p,\nu}g(X) = 0$ *for all* $p \in (0,1)$ *and* $\nu \in \mathbb{N}$, *i.e.*

$$\mathbb{E}_{p,\nu}g(X) = \sum_{i=0}^{\nu} g(i)\binom{\nu}{i}p^i(1-p)^{\nu-i} = (1-p)^{\nu}\sum_{i=0}^{\nu} g(i)\binom{\nu}{i}\left(\frac{p}{1-p}\right)^i = 0, \quad p \in (0,1), \quad \nu \in \mathbb{N}.$$

Since for $p \in (0,1)$, *the function* $p/(1-p)$ *takes values in* $(0,\infty)$, *the latter implies*[36] *that* $g(i) = 0$ *for all* $i \in \mathbb{N}$ *and hence* X *is complete.*

[36] recall that a polynomial equals to zero on an open interval if and only if all its coefficients are zeros

(7) Find the UMVUE for $p\nu$, if both p and ν are unknown.

Solution

X is an unbiased estimator of $p\nu$ and X is a sufficient and complete statistic, hence by the L-S theorem X is UMVUE for $p\nu$.

(8) Is X a complete statistic, if ν is unknown and p is known and equals $1/2$?

Hint: recall $(1-1)^\nu = \sum_{i=0}^{\nu} \binom{\nu}{i} (-1)^i = 0$ for any $\nu \in \mathbb{N}$.

Solution

Consider the statistic $g(X) = (-1)^X$.

$$\mathbb{E}_\nu g(X) = (1/2)^\nu \sum_{i=0}^{\nu} g(i) \binom{\nu}{i} = (1/2)^\nu \sum_{i=0}^{\nu} (-1)^i \binom{\nu}{i} =$$

$$(1/2)^\nu \sum_{i=0}^{\nu} (-1)^i 1^{\nu-i} \binom{\nu}{i} = (1/2)^\nu (1-1)^\nu = 0, \quad \forall \nu \in \mathbb{N}.$$

However $g(X) \neq 0$, hence X is not complete.

f. 2010/2011 (B) 52303

Problem 1.[Neyman-Scott] (see in 2010/2011 (A) 52314 below)

Problem 2.

Let $X = (X_1, ..., X_n)$ be a sample from the probability density

$$f(x;\theta) = p\psi(x) + (1-p)\psi(x-\mu), \quad x \in \mathbb{R},$$

where $\psi(x) = \frac{1}{2}e^{-|x|}$ is the density of Laplace (two-sided exponential) r.v., $\mu > 0$ is a known constant and $p \in [0,1]$ is the unknown parameter.

(1) Is the model identifiable ? Would the model be identifiable if both p and μ were unknown ?

Solution

If μ is known, the model is identifiable: e.g. $\mathbb{E}_p X_1 = (1-p)\mu$ is a one-to-one function of p (recall that $\mu \neq 0$). The model is not identifiable if both p and μ are unknown: if $p = 1$, the density doesn't depend on μ.

(2) Answer the previous question, when $p \in (0,1)$.

Solution

If μ is known, then the model is still identifiable as before. However, if both $p \in (0,1)$ and $\mu > 0$ are unknown, the model remains identifiable. To see this, consider for example the first and the second moments of X_1:

$$m_1(p,\mu) := \mathbb{E}X_1 = (1-p)\mu$$

$$m_2(p,\mu) := \mathbb{E}X_1^2 \stackrel{\dagger}{=} 2p + (1-p)\big(\mu^2+2\big) = 2 + (1-p)\mu^2,$$

where in $\dagger$, we used the expressions

$$\int_{-\infty}^{\infty} x^2 \frac{1}{2} e^{-|x|} dx = \int_0^{\infty} x^2 e^{-x} dx = 2.$$

If $p \in (0,1)$, then the function $(p,\mu) \mapsto \big(m_1(p,\mu), m_2(p,\mu)\big)$ is one-to-one with the inverse:

$$\begin{pmatrix} p \\ \mu \end{pmatrix} = \begin{pmatrix} 1 - \dfrac{m_1^2}{m_2 - 2} \\ \dfrac{m_2 - 2}{m_1} \end{pmatrix}.$$

(3) Find the MLE of p on the basis of one sample X_1

Solution

The likelihood $L_1(X;p) = p\psi(X_1) + (1-p)\psi(X_1-\mu)$ is a linear function of p and is maximized at an endpoint of the interval $[0,1]$. Hence

$$\hat{p}(X_1) = \mathbf{1}_{\{\psi(X_1) \geq \psi(X_1-\mu)\}} = \mathbf{1}_{\{|X_1| \leq |X_1-\mu|\}} = \mathbf{1}_{\{X_1 \leq \mu/2\}}.$$

(4) Using the first moment of X_1, suggest a sequence of unbiased consistent estimators of p

Solution

Note $\mathbb{E}_p X_1 = (1-p)\mu$, and the method of moments gives $\hat{p}_n(X) = 1 - \bar{X}_n/\mu$. Since by the LLN, $\bar{X}_n \stackrel{\mathbb{P}_p}{\to} (1-p)\mu$, the sequence of estimators $(\hat{p}_n)$ is consistent.

(5) Find the asymptotic error distribution and the corresponding rate for the sequence of estimators from the previous question. What happens to the asymptotic variance when μ is close to zero ? When μ is very large ? Explain.

Solution

We have

$$\mathrm{var}_p(X_1) = \mathbb{E}_p X_1^2 - (\mathbb{E}_p X_1)^2 = 2p + (1-p)(\mu^2+2) - (1-p)^2\mu^2 = 2 + p(1-p)\mu^2$$

and by the CLT

$$\sqrt{n}\big(1 - \bar{X}_n/\mu - p\big) = \frac{1}{\mu}\sqrt{n}\Big((1-p)\mu - \bar{X}_n\Big) \xrightarrow{d} N\left(0, \frac{1}{\mu^2}\Big(2 + p(1-p)\mu^2\Big)\right) = N\left(0, \frac{2}{\mu^2} + p(1-p)\right)$$

Note that $X_1 = \xi_1 Z_1 + (1-\xi_1)(Z_1+\mu)$, where $\xi_1 \sim \mathrm{Ber}(p)$ and $Z_1 \sim \mathrm{Lap}(1)$ are independent. If μ is large, it is easy to guess ξ_i's and hence the variance of the error is as if we observe the ξ_i's directly. When μ is large, the observations are not very informative and hence the observation variance is large.

(6) Prove that the statistic

$$T(X) = \Big(|X_1| - |X_1 - \mu|, ..., |X_n| - |X_n - \mu|\Big)$$

is sufficient and that it is strictly coarser than the whole sample $(X_1, ..., X_n)$.

Solution

The likelihood is

$$L(X;p) = \prod_{i=1}^n \Big(p\psi(X_i) + (1-p)\psi(X_i - \mu)\Big) = \prod_{i=1}^n \Big(p\frac{1}{2}e^{-|X_i|} + (1-p)\frac{1}{2}e^{-|X_i-\mu|}\Big) =$$

$$\prod_{i=1}^n e^{-|X_i|}\Big(p\frac{1}{2} + (1-p)\frac{1}{2}e^{|X_i|-|X_i-\mu|}\Big) = e^{-\sum_i |X_i|}\prod_{i=1}^n \Big(p\frac{1}{2} + (1-p)\frac{1}{2}e^{|X_i|-|X_i-\mu|}\Big).$$

By the F-N factorization theorem $T(X)$ is sufficient. $T(X)$ is strictly coarser than X, since the function:

$$|x| - |x-\mu| = \begin{cases} \mu & x \geq \mu \\ 2x - \mu & 0 \leq x < \mu \\ -\mu & x < 0 \end{cases}$$

is not one-to-one on the range of X_1 (why ?)

(7) Are your estimators in (4) UMVUE ? Are they admissible ?

Solution

Note that

$$\mathbb{E}_p\big(\bar{X}_n|T(X)\big)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_p\big(X_i|T(X)\big)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_p\big(X_i\big||X_i|-|X_i-\mu|\big).$$

Since $|X_i|-|X_i-\mu|$ is strictly coarser than X_i,

$$\mathbb{E}_p\big(X_i\big||X_i|-|X_i-\mu|\big)\neq X_i$$

with positive probability[37] *and hence by R-B theorem, conditioning $\hat{p}_n(X)$ on the sufficient statistic $T(X)$ from (6) yields an estimator with strictly better risk. Consequently $\hat{p}_n$ is not UMVUE and is inadmissible.*

Appendix

The conditional expectation $\mathbb{E}_p(X_1|T_1)$, where $T_1=|X_1|-|X_1-\mu|$, is not hard to calculate:

$$\mathbb{E}_p(X_1|T_1)=\begin{cases}\dfrac{\mathbb{E}_p(X_1\mathbf{1}_{\{X_1<0\}})}{\mathbb{P}_p(X_1<0)}, & T_1<-\mu\\ \dfrac{1}{2}(T_1+\mu), & -\mu\le T_1\le\mu\\ \dfrac{\mathbb{E}_p(X_1\mathbf{1}_{\{X_1>\mu\}})}{\mathbb{P}_p(X_1>\mu)}, & T_1>\mu\end{cases}$$

and

$$\frac{\mathbb{E}_p(X_1\mathbf{1}_{\{X_1\le 0\}})}{\mathbb{P}_p(X_1<0)}=\frac{\frac{1}{2}\int_{-\infty}^0 x\Big(pe^x+(1-p)e^{x+\mu}\Big)dx}{\frac{1}{2}\int_{-\infty}^0\Big(pe^x+(1-p)e^{x+\mu}\Big)dx}=\frac{\int_{-\infty}^0 xe^x dx}{\int_{-\infty}^0 e^x dx}=-1,$$

and, similarly,

$$\frac{\mathbb{E}_p(X_1\mathbf{1}_{\{X_1>\mu\}})}{\mathbb{P}_p(X_1>\mu)}=\frac{\frac{1}{2}\int_{\mu}^\infty x\Big(pe^{-x}+(1-p)e^{-x+\mu}\Big)dx}{\frac{1}{2}\int_{\mu}^\infty\Big(pe^{-x}+(1-p)e^{-x+\mu}\Big)dx}=\frac{\int_{\mu}^\infty xe^{-x}dx}{\int_{\mu}^\infty e^{-x}dx}=\frac{e^{-\mu}+\mu e^{-\mu}}{e^{-\mu}}=1+\mu.$$

By the R-B theorem, the estimator

$$\tilde{p}_n(X)=1-\frac{1}{\mu}\frac{1}{n}\sum_{i=1}^n\Big((1+\mu)\mathbf{1}_{\{T_i(X_i)>\mu\}}-\mathbf{1}_{\{T_i(X_i)<-\mu\}}+\frac{1}{2}\big(T_i(X_i)+\mu\big)\mathbf{1}_{\{T_i(X_i)\in[-\mu,\mu]\}}\Big)$$

is an unbiased estimator of p with the better MSE risk than $1-\bar{X}_n/\mu$.

g. 2010/2011 (A) 52314

Problem 1.[Neyman-Scott[38]]

[37] see the Appendix, if you are curious what the improved estimator look like

[38] this classical example in asymptotic statistics shows that in the presence of high-dimensional nuisance parameter μ, the MLE may be inconsistent

Let $\xi_1, ..., \xi_n, \eta_1, ..., \eta_n$ be i.i.d. $N(0,1)$ random variables and set

$$X_i = \mu_i + \sigma\xi_i, \quad i = 1, ..., n$$
$$Y_i = \mu_i + \sigma\eta_i$$

where $\mu = (\mu_1, ..., \mu_n) \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}_+$ are unknown parameters. It is required to estimate σ^2, given the data $(X_1, Y_1), ..., (X_n, Y_n)$. The vector μ is regarded as nuisance parameter.

(1) Find the likelihood function for this model.

Solution

By independence,

$$L(x, y; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2}\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{\sigma^2}\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{\sigma^2}\right\}$$

with $x, y \in \mathbb{R}^n$ *and* $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}^n \times \mathbb{R}_+$.

(2) Find the MLE of (μ, σ^2)

Solution

Using the the elementary formula $a^2 + b^2 = \frac{1}{2}(a-b)^2 + \frac{1}{2}(a+b)^2$, *we get*

$$L(X, Y; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i - \mu_i)^2 + (Y_i - \mu_i)^2}{\sigma^2}\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\frac{1}{2}(X_i - Y_i)^2 + \frac{1}{2}(Y_i + X_i - 2\mu_i)^2}{\sigma^2}\right\}.$$

The latter is maximized by the choice $\hat{\mu}_i = (X_i + Y_i)/2$ *and* $\hat{\sigma}_n^2 = \frac{1}{4}\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2$.

(3) Are MLE's of μ_i's biased ? Is MLE of σ^2 biased ?

Solution

$\mathbb{E}_{\mu,\sigma^2}\hat{\mu}_i = \mu_i$, *i.e. the MLE's of* μ_i*'s are unbiased. However,* $\mathbb{E}_{\mu,\sigma^2}(X_i - Y_i)^2 = 2\sigma^2$ *and hence*

$$\mathbb{E}_{\mu,\sigma^2}\hat{\sigma}_n^2 = \frac{1}{2}\sigma^2,$$

i.e. the MLE of σ^2 is biased.

(4) Are MLE's of μ_i's consistent ? Is MLE of σ^2 consistent ?

Solution

The MLE's of μ_i's depend only on two random variables and hence are trivially not consistent. Note that $X_i - Y_i \sim N(0, 2\sigma^2)$ are i.i.d. and hence by the LLN

$$\hat{\sigma}_n^2(X,Y) \xrightarrow{n\to\infty} \frac{1}{2}\sigma^2$$

in probability, i.e MLE of σ^2 is not consistent either.

(5) Find the minimal sufficient statistic.

Solution

$$L(x,y;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^2+y_i^2) + \frac{1}{\sigma^2}\sum_{i=1}^n \mu_i(x_i+y_i) - \frac{1}{\sigma^2}\sum_{i=1}^n \mu_i^2\right\}$$

and by the F-N factorization theorem, the statistic

$$T(X,Y) = (T_1,...,T_n,T_{n+1}) = \left(X_1+Y_1,...,X_n+Y_n,\sum_{i=1}^n (X_i^2+Y_i^2)\right)$$

is sufficient. Further, for (x,y) and $(\tilde{x},\tilde{y})$ in $\mathbb{R}^n \times \mathbb{R}^n$,

$$\frac{L(x,y;\mu,\sigma^2)}{L(\tilde{x},\tilde{y};\mu,\sigma^2)} = \exp\left\{-\frac{1}{2\sigma^2}\Big(T_{n+1}(x,y) - T_{n+1}(\tilde{x},\tilde{y})\Big) + \frac{1}{\sigma^2}\sum_{i=1}^n \mu_i\Big(T_i(x,y) - T_i(\tilde{x},\tilde{y})\Big)\right\}.$$

The latter is not a function of (μ,σ^2) only if $T(x,y) = T(\tilde{x},\tilde{y})$. Hence $T(X,Y)$ is minimal sufficient.

(6) Is the minimal sufficient statistic complete ?

Solution

The likelihood belongs to the $(n+1)$-parameter exponential family with

$$c(\mu,\sigma^2) = \left(\frac{\mu_1}{\sigma^2},...,\frac{\mu_n}{\sigma^2},-\frac{1}{2\sigma^2}\right).$$

The range of $c(\mu,\sigma^2)$ is $\mathbb{R}^n \times \mathbb{R}_-$, which is obviously not empty. Hence $T(X,Y)$ is a complete statistic.

(7) Find the UMVUE of σ^2 and calculate its MSE risk

Solution

The statistic (a modification of MLE)

$$\tilde{\sigma}^2(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(X_i - Y_i)^2$$

is an unbiased estimator of σ^2. *Since*

$$\frac{1}{2}\sum_i (X_i - Y_i)^2 = \sum_i (X_i^2 + Y_i^2) - \frac{1}{2}\sum_i (X_i + Y_i)^2,$$

$\tilde{\sigma}^2(X,Y)$ *is a function of the complete sufficient statistic and hence by the L-S theorem it is UMVUE. Note that* $\xi_i := (X_i - Y_i)/\sqrt{2} \sim N(0,\sigma^2)$ *and the MSE risk is*

$$R(\sigma^2,\tilde{\sigma}^2) = \mathrm{var}(\tilde{\sigma}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(X_i - Y_i)^2 - \sigma^2\right)^2 =$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i^2 - \sigma^2)\right)^2 = \frac{1}{n}\mathbb{E}(\xi_1^2 - \sigma^2)^2 = \frac{1}{n}\Big(3(\sigma^2)^2 - 2(\sigma^2)^2 + (\sigma^2)^2\Big) = \frac{2}{n}(\sigma^2)^2$$

(8) Find the C-R lower bound for MSE risk of unbiased estimators of σ^2, assuming that μ_i's are known. Is it attained by the estimator from the previous question?

Solution

$$\log L(x,y;\sigma^2) = -n\log 2\pi - n\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2 + (y_i-\mu_i)^2}{\sigma^2}$$

and hence

$$\frac{\partial}{\partial\sigma^2}\log L(X,Y;\sigma^2) = -n\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\Big((X_i-\mu_i)^2 + (Y_i-\mu_i)^2\Big)$$

and

$$\frac{\partial^2}{\big(\partial\sigma^2\big)^2}\log L(X,Y;\sigma^2) = n\frac{1}{(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}\Big((X_i-\mu_i)^2 + (Y_i-\mu_i)^2\Big).$$

The Fisher information is then

$$I_n(\sigma^2) = -\mathbb{E}_{\sigma^2}\frac{\partial^2}{\big(\partial\sigma^2\big)^2}\log L(X,Y;\sigma^2) = -n\frac{1}{(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}n2\sigma^2 = \frac{n}{(\sigma^2)^2}.$$

The bound is not attained, which shall be expected, since if μ_i's are unknown, the problem of estimating σ^2 is harder (and, of course, the model is different).

Problem 2. An electronic device monitors the radiation activity, by counting the number of particles, emitted by a source. The outputs of the device at consecutive time units $i = 1, ..., n$ are independent random variables $X_1, ..., X_n$ with Poisson distribution $X_i \sim \text{Poi}(1 + \lambda_i)$, where $\lambda_i \geq 0$ is the unknown intensity of the source at time i (and 1 models the known radiation of the background).

(1) Find the UMP test statistic for the problem

$$H_0 : \lambda_1 = ... = \lambda_n = 0$$
$$H_1 : \lambda_1 = ... = \lambda_n > 0$$

Solution

The problem can be rephrased as testing $H_0 : X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Poi}(1)$ against $H_1 : X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Poi}(1 + r)$, where $r > 0$. The likelihood of the model is:

$$L(x; r) = \prod_{i=1}^{n} e^{-(1+r)} \frac{(1+r)^{X_i}}{X_i!} = e^{-n(1+r)}(1+r)^{n\bar{X}_n} \Big/ \prod_i X_i!$$

and the likelihood ratio

$$R(X; r_1, r_0) = \frac{L(X; r_1)}{L(X; r_0)} = e^{-n(r_1 - r_0)} \left(\frac{1 + r_1}{1 + r_0} \right)^{n\bar{X}_n}$$

is a strictly increasing function of the statistic $\bar{X}_n$ for $r_1 > r_0$. Hence by K-R theorem the likelihood ratio test is UMP. The test statistic is given by

$$\frac{L(X; r)}{L(X; 0)} = e^{-nr} (1 + r)^{n\bar{X}_n}$$

and hence the UMP test rejects H_0 if and only if $\{\bar{X}_n \geq c\}$, where c is the critical value to be chosen to meet the size requirement.

(2) For which values of the size α, an exact critical value can be found for the UMP test ?

Solution

Recall that under H_0, $S(X) = n\bar{X}_n$ has $\text{Poi}(n)$ distribution, hence c is the smallest integer satisfying $\mathbb{P}_0(\bar{X}_n \geq c) = \alpha$, or

$$\sum_{k \geq nc} e^{-n} \frac{n^k}{k!} = \alpha,$$

which can be evaluated numerically.

(3) Use the CLT to approximate the critical value of the UMP test.

Solution

Write c_n to emphasize the dependence of the critical value on n, then

$$\mathbb{P}_0\big(\bar{X}_n \geq c_n\big) = \mathbb{P}_0\left(\sqrt{n}\big(\bar{X}_n - 1\big) \geq \sqrt{n}\big(c_n - 1\big)\right).$$

Now if we choose $c_n := 1 + z/\sqrt{n}$ with $z \in \mathbb{R}$, then

$$\lim_{n\to\infty} \mathbb{P}_0\big(\bar{X}_n \geq c_n\big) = 1 - \Phi(z),$$

and hence $z := \Phi^{-1}(1-\alpha)$ yields the test with critical value $c_n = 1 + \Phi^{-1}(1-\alpha)/\sqrt{n}$, whose size is approximately α.

(4) Is the test with approximate critical value found in (3) consistent[39]?

Hint: you may find the LLN useful.

Solution

The power function of the approximate UMP test is

$$\pi_n(r) = \mathbb{P}_r\Big(\bar{X}_n \geq 1 + \Phi^{-1}(1-\alpha)/\sqrt{n}\Big) =$$

$$\mathbb{P}_r\Big(\bar{X}_n - 1 - r \geq -r + \Phi^{-1}(1-\alpha)/\sqrt{n}\Big) \overset{\dagger}{\geq} \mathbb{P}_r\Big(\bar{X}_n - 1 - r \geq -r/2\Big) \geq$$

$$\mathbb{P}_r\Big(|\bar{X}_n - 1 - r| \leq r/2\Big) = 1 - \mathbb{P}_r\Big(|\bar{X}_n - 1 - r| > r/2\Big) \xrightarrow{n\to\infty} 1,$$

where the inequality † holds for all sufficiently large n (think why) and the convergence holds, since $\bar{X}_n \to 1 + r$ in $\mathbb{P}_r$-probability.

(5) Find the GLRT statistic for the problem in (1). Does the GLRT coincide with UMP test ?

Solution

[39] a sequence of tests (δ_n) is consistent, if the corresponding power functions $\pi_n(\theta) := \mathbb{E}_\theta \delta_n$ satisfy:

$$\lim_n \pi_n(\theta) = 1, \quad \forall \theta \in \Theta_1.$$

Under H_1, the log*-likelihood*

$$\log L(X;r) = -n(1+r) + n\bar{X}_n \log(1+r) - \log \prod_i X_i!$$

is a continuous function of r on $[0,\infty)$. Taking the derivative and equating the result to zero we get the extremum:

$$-n + n\bar{X}_n \frac{1}{1+r} = 0 \quad \Longrightarrow \quad r^* = \bar{X}_n - 1,$$

which is a local maximum (the second derivative is negative). Since $\lim_{r\to\infty} \log L(X;r) = -\infty$ and $\lim_{r\to -1} \log L(X;r) = \infty$, r^* is a global maximum on $(-1,\infty)$ and hence the MLE of r is[40]

$$\hat{r}_n(X) = (\bar{X}_n - 1)^+.$$

Hence the GLRT statistic is

$$\Lambda_n(X) := \frac{\sup_{r>0} L(X;r)}{L(X;0)} = e^{-n(1+\hat{r}_n)+n}(1+\hat{r}_n)^{n\bar{X}_n} = \begin{cases} 1 & \bar{X}_n \le 1 \\ e^{-n(\bar{X}_n - 1)}(\bar{X}_n)^{n\bar{X}_n} & \bar{X}_n > 1 \end{cases}$$

Let $\varphi(x) = e^{-n(x-1)}(x)^{nx}$ and note that $(\log \varphi(x))' = (-nx + n + nx\log x)' = -n + n(\log x + 1) = n\log x > 0$ for $x > 1$. Since $(\log\varphi(x))' = \varphi'/\varphi(x)$ and $\varphi(x) > 0$ for $x > 1$, it follows that $x \mapsto \varphi(x)$ is strictly increasing on $(1,\infty)$. It can be shown that the median of the Poisson distribution with mean $n \ge 1$ is greater than n and hence for $\alpha < 1/2$, the critical value of the UMP test is greater than 1, i.e it accepts H_0 if $\bar{X}_n \le 1$. Since the GLRT also accepts H_0 for $\bar{X}_n \le 1$ and the GLRT statistic is strictly increasing in $\bar{X}_n$ for $\bar{X}_n > 1$, the GLRT coincides with the UMP test.

(6) Find the GLRT statistic for the problem

$$H_0 : \lambda_1 = ... = \lambda_n = 0,$$
$$H_1 : \lambda_1 + ... + \lambda_n > 0$$

Solution

The log*-likelihood under H_1 is*

$$\log L(X;\lambda_1,...,\lambda_n) = \sum_{i=1}^n \Big(-(1+\lambda_i) + X_i \log(1+\lambda_i) - \log X_i! \Big),$$

[40] we use the notation $x^+ = \max(x,0)$

which is maximized over $\mathbb{R}^n_+$ *by* $\hat{\lambda}_i := (X_i - 1)^+$. *Hence the GLRT statistic is*

$$\frac{\sup_{\lambda_1+...+\lambda_n>0} L(X;\lambda_1,...,\lambda_n)}{L(X;0)} = \frac{\prod_i e^{-1-(X_i-1)^+}(1+(X_i-1)^+)^{X_i}}{e^{-n}} =$$

$$\prod_i e^{-(X_i-1)^+}(1+(X_i-1)^+)^{X_i} = \prod_{i:X_i\geq 1} e^{1-X_i}(X_i)^{X_i}$$

(7) It is known that the radiation jumped from 0 to the known level $r > 0$ at the unknown time $\nu \in \{1, ..., n\}$:

$$\lambda_1 = ... = \lambda_{\nu-1} = 0, \quad \lambda_\nu = ... = \lambda_n = r.$$

Assuming uniform prior on ν, find the Bayes estimator $\hat{\nu}(X)$ with respect to the loss function $\ell(k, m) = \mathbf{1}_{\{k\neq m\}}$.

Solution

$\nu \in \{1, ..., n\}$ is the only unknown parameter in this problem and the corresponding likelihood is

$$L(X;\nu) := \prod_{i=1}^{\nu-1} e^{-1}\frac{1}{X_i!}\prod_{i=\nu}^{n} e^{-(1+r)}\frac{(1+r)^{X_i}}{X_i!} = \prod_{i=1}^{n} e^{-1}\frac{1}{X_i!}\prod_{i=\nu}^{n} e^{-r}(1+r)^{X_i}.$$

where $\prod_{i=1}^{0}(...) = 1$ *is understood. Recall that the Bayes estimator for this loss function is the posterior mode:*

$$\hat{\nu}(X) = \text{argmax}_{\ell\in\{1,...,n\}}\frac{L(X;\ell)\pi(\ell)}{\sum_{j=1} L(X;j)\pi(j)} \overset{\dagger}{=} \text{argmax}_{\ell\in\{1,...,n\}}\frac{L(X;\ell)}{\sum_{j=1} L(X;j)} =$$

$$\text{argmax}_{\ell\in\{1,...,n\}} L(X;\ell) = \text{argmax}_{\ell\in\{1,...,n\}} \log L(X;\ell)$$

where the equality † *holds, since the posterior* π *is uniform over* $\{1, ..., n\}$, *i.e.* $\pi(i) = 1/n$. *Hence*

$$\hat{\nu}(X) = \text{argmax}_{\ell\in\{1,...,n\}} \log L(X;\ell) = \text{argmax}_{\ell\in\{1,...,n\}} \log \prod_{i=\ell}^{n} e^{-r}(1+r)^{X_i} =$$

$$\text{argmax}_{\ell\in\{1,...,n\}} \sum_{i=\ell}^{n} \Big(X_i \log(1+r) - r\Big).$$

h. 2011/2012 (A) 52314

Problem 1. (variations on the regression theme)

Consider the problem of estimating the unknown parameter $\theta \in \mathbb{R}$ from the sample (X_1, Y_1), ..., (X_n, Y_n), $n \geq 3$ where

$$Y_i = X_i\theta + \varepsilon_i, \quad i = 1, ..., n$$

where *covariates* X_i's and the *noise* ε_i's are i.i.d. random variables with $N(0, 1)$ distribution.

(1) Show that the likelihood function is given by

$$L(X,Y;\theta)=\Big(\frac{1}{2\pi}\Big)^n\exp\Big(-\frac{1}{2}\sum_i X_i^2-\frac{1}{2}\sum_i(Y_i-\theta X_i)^2\Big).$$

Solution

For any $y\in\mathbb{R}$,

$$\mathbb{P}_\theta(Y_1\le y|X_1)=\mathbb{P}_\theta(\varepsilon_1\le y-\theta X_1|X_1)=\Phi(y-\theta X_1),$$

where Φ is the standard Gaussian c.d.f. This means that Y_1 is conditionally Gaussian given X_1 with mean θX_1 and unit variance. Hence

$$f_{X_1Y_1}(x,y)=f_{Y_1|X_1}(y;x)f_{X_1}(x)=\frac{1}{2\pi}\exp\Big(-\frac{1}{2}(y-\theta x)^2-\frac{1}{2}x^2\Big)$$

and the claimed expression follows by the i.i.d. assumption.

(2) Find the minimal sufficient statistic for this model.

Solution

We have

$$L(X,Y;\theta)=\Big(\frac{1}{2\pi}\Big)^n\exp\Big(-\frac{1}{2}\sum_i(X_i^2+Y_i^2)+\theta\sum_i X_iY_i-\frac{1}{2}\theta^2\sum_i X_i^2\Big)$$

and by the factorization theorem the statistic

$$T(X,Y)=\Big(\sum_i X_iY_i,\sum_i X_i^2\Big)$$

is sufficient. Further, for $x,y,x',y'\in\mathbb{R}^n$ and $\theta\in\mathbb{R}$,

$$\log\frac{L(x,y;\theta)}{L(x',y';\theta)}=-\frac{1}{2}\sum_i(x_i^2+y_i^2)+\frac{1}{2}\sum_i(x_i'^2+y_i'^2)+\theta\Big(\sum_i x_iy_i-\sum_i x_i'y_i'\Big)-\frac{1}{2}\theta^2\Big(\sum_i x_i^2-\sum_i x_i'^2\Big).$$

The latter does not depend on θ, only if $T(x,y)=T(x',y')$, which implies minimality of T.

(3) Find the MLE of θ.

Solution

the log*-likelihood function is a parabola in* θ *with the unique maximum at*

$$\hat{\theta}(X,Y) := \frac{\sum_i X_i Y_i}{\sum_i X_i^2}.$$

(4) Is the MLE unbiased ?

Solution

The MLE is unbiased

$$\mathbb{E}_\theta \hat{\theta}(X,Y) = \mathbb{E}_\theta \frac{\sum_i X_i(\theta X_i + \varepsilon_i)}{\sum_i X_i^2} = \theta + \mathbb{E}_\theta \frac{\sum_i X_i \mathbb{E}_\theta(\varepsilon_i | X_1, ..., X_n)}{\sum_i X_i^2} = \theta.$$

(5) Is the statistic found in (2) complete ?

Solution

For the function $g(t) = t_2 - n$, $t = (t_1, t_2) \in \mathbb{R}^2$ *we have*

$$\mathbb{E}_\theta g(T(X,Y)) = \mathbb{E}_\theta \sum_i X_i^2 - n = 0, \quad \forall \theta \in \mathbb{R},$$

while $\mathbb{P}_\theta\Big(g(T(X,Y)) = 0\Big) = 0$. *Hence the statistic is not complete.*

(6) Find the Cramer-Rao bound for the MSE risk of the unbiased estimators of θ.

Solution

The Fisher information for this problem is

$$I_n(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L(X,Y;\theta) = \mathbb{E}_\theta \sum_{i=1}^n X_i^2 = n,$$

and hence

$$\mathbb{E}_\theta(\tilde{\theta}(X,Y) - \theta)^2 \geq 1/n, \quad \forall \theta \in \mathbb{R},$$

for any unbiased estimator $\tilde{\theta}$.

(7) Is the MLE from (2) efficient for n ?

Hint: for $\xi \sim \chi^2_n$ and $n \geq 3$, $\mathbb{E}\frac{1}{\xi} = \frac{1}{n-2}$

Solution

Since $\xi = \sum_{i=1}^n X_i^2 \sim \chi_n^2$,

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta \left(\frac{\sum_i X_i \varepsilon_i}{\sum_i X_i^2} \right)^2 = \mathbb{E}_\theta \left(\frac{1}{\sum_i X_i^2} \right)^2 \sum_i \sum_j X_i X_j \mathbb{E}_\theta(\varepsilon_i \varepsilon_j | X_1, ..., X_n) =$$

$$\mathbb{E}_\theta \left(\frac{1}{\sum_i X_i^2} \right)^2 \sum_i X_i^2 = \mathbb{E}_\theta \frac{1}{\sum_i X_i^2} = \frac{1}{n-2}.$$

Hence the MLE is not efficient for any n, but the sequence of MLEs is asymptotically efficient as $n \to \infty$ *(in agreement with the general asymptotic results on MLEs).*

(8) Is the estimator $\tilde{\theta}(X, Y) = \frac{1}{n} \sum_{i=1}^n X_i Y_i$ unbiased ? efficient ? Do the risk of $\tilde{\theta}$ and the risk of MLE compare ? Is it improvable through the R-B procedure ?

Solution

We have

$$\mathbb{E}_\theta \tilde{\theta} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta X_i (X_i \theta + \varepsilon_i) = \theta,$$

i.e. $\tilde{\theta}$ *is unbiased. It's risk is given by*

$$R(\theta, \tilde{\theta}) = \mathbb{E}_\theta \left(\frac{1}{n} \sum_{i=1}^n (X_i^2 - 1)\theta + \frac{1}{n} \sum_{i=1}^n X_i \varepsilon_i \right)^2 = \frac{1}{n} \Big(\theta^2 \mathbb{E}_\theta (X_i^2 - 1)^2 + 1 \Big) = \frac{1}{n} \Big(2\theta^2 + 1 \Big).$$

Hence $\tilde{\theta}$ *is not efficient, either for a fixed n or asymptotically. It's risk is not comparable with the risk of the MLE. The R-B procedure does not change the estimator and hence doesn't yield an improvement.*

(9) Argue that if the UMVUE exists for this problem it has to be efficient (i.e. to attain the C-R bound).

Hint: Consider the estimators of the form

$$\frac{1}{n} \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n \theta_0 (X_i^2 - 1)$$

for various values of θ_0.

Solution

Note that for any θ_0*, the estimator*

$$\check{\theta} = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n \theta_0 (X_i^2 - 1)$$

is unbiased. Its risk is given by

$$R(\theta,\check{\theta}) = \mathbb{E}_\theta\Big(\frac{1}{n}\sum_{i=1}^n (X_i^2-1)(\theta-\theta_0) + \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i\Big)^2 = \frac{1}{n}\Big(2(\theta-\theta_0)^2+1\Big).$$

Suppose the UMVUE $\hat{\theta}^*$ exists, then

$$R(\theta_0,\hat{\theta}^*) \le R(\theta_0,\check{\theta}) = 1/n.$$

Since θ_0 was arbitrary, the claim follows.

Problem 2.

Consider the problem of detection of a *sparse* signal in noise. More precisely, we sample independent random variables $X = (X_1,...,X_n)$ with $X_i \sim N(\theta_i, 1)$ and would like to test the hypotheses

$$H_0 : \theta = 0$$
$$H_1 : \theta \in \Theta_1$$

where Θ_1 is a subset of $\mathbb{R}^n \setminus \{0\}$ with a particular structure, specified below.

(1) Let e_i be a vector with 1 at the i-th entry and zeros at all others and assume that

$$\Theta_1 = \{e_1,...,e_n\}.$$

Find the level α test which rejects H_0 if and only if $\{\bar{X}_n \ge c\}$, and find its power function. Does the power function converge as $n \to \infty$ and if yes, find the limit? Explain the result.

Solution

This is the standard calculation:

$$\alpha = \mathbb{P}_0(\bar{X}_n \ge c) = \mathbb{P}_0(\sqrt{n}\bar{X}_n \ge \sqrt{n}c) = 1-\Phi(\sqrt{n}c),$$

which yields $c_\alpha = \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha)$. The power function is

$$\pi(e_i;\delta) = \mathbb{P}_{e_i}(\bar{X} \ge c_\alpha) = \mathbb{P}_{e_i}\big(\sqrt{n}(\bar{X}_n - 1/n) \ge \sqrt{n}(c_\alpha - 1/n)\big) =$$
$$= 1-\Phi\Big(\sqrt{n}(c_\alpha-1/n)\Big) = 1-\Phi\Big(\Phi^{-1}(1-\alpha)-1/\sqrt{n}\Big)$$

which does not depend on the alternative. Note that $\lim_n \pi(e_1,\delta) = \alpha$, which should be expected, since for large n the zero signal and the signal with just one unit entry are hard to distinguish.

(2) Find the level α GLRT for the problem from (1) and calculate its power function. Does the power function converge as $n \to \infty$? If yes, find the limit. Explain the result.

Solution

The GLRT statistic is

$$\lambda(X) = \frac{\sup_{\theta\in\Theta_1} L(X;\theta)}{L(X;0)} = \frac{\max_{i\le n} \exp\Big(-\frac{1}{2}\sum_{m\neq i}^{n} X_m^2 - \frac{1}{2}(X_i-1)^2\Big)}{\exp\Big(-\frac{1}{2}\sum_{m=1}^{n} X_m^2\Big)} =$$

$$\max_{i\le n} \exp\Big(-\frac{1}{2}(X_i-1)^2 + \frac{1}{2}X_i^2\Big) = \max_{i\le n} \exp\Big(X_i - \frac{1}{2}\Big) = \exp\Big(\max_i X_i - \frac{1}{2}\Big)$$

and hence GLRT rejects H_0 on the event $\{\max_i X_i \ge c\}$. Further,

$$\mathbb{P}_0\big(\max_i X_i \ge c\big) = 1 - \mathbb{P}_0\big(\max_i X_i < c\big) = 1 - \Phi^n(c) = \alpha,$$

which gives $c_\alpha = \Phi^{-1}(\sqrt[n]{1-\alpha})$. Using the symmetry of the test statistic,

$$\pi(e_i;\delta) = \mathbb{P}_{e_i}(\max_j X_j \ge c_\alpha) = 1 - \mathbb{P}_{e_1}\Big(\max\big(X_1, \max_{j>1} X_j\big) < c_\alpha\Big) =$$

$$1 - \Phi(c_\alpha - 1)\Phi^{n-1}(c_\alpha) = 1 - \Phi\Big(\Phi^{-1}(\sqrt[n]{1-\alpha}) - 1\Big)(1-\alpha)^{1-1/n}.$$

Once again $\lim_n \pi(e_i;\delta^*) = \alpha$.

(3) Is either of the tests in (1) and (2) UMP ?

Solution

Suppose $\delta^*(X) = \{\max_i X_i \ge c_\alpha\}$ is UMP and let $\delta(X)$ be the N-P test for the problem

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta = e_1 \end{aligned} \qquad (\dagger)$$

Note that δ is still a legitimate level α test in the composite problem of (2). Since δ^* is UMP

$$\mathbb{E}_{e_1}\delta^*(X) \ge \mathbb{E}_{e_1}\delta(X).$$

But then by the converse of N-P lemma, $\delta^*(X)$ and $\delta(X)$ coincide off the event $\{\delta(X) = c_\alpha\}$. The N-P test for $(\dagger)$ rejects H_0 on the event $\{X_1 \ge c'_\alpha\}$ and hence differs from δ^* with positive probability (why?). The obtained contradiction shows that δ^* is not UMP. Similarly, the test from (1) is not UMP.

(4) Consider the alternative subspace

$$\Theta_1 = \Big\{\theta \in \mathbb{R}^n : \sum_{i=1}^{n} \mathbf{1}_{\{\theta_i \neq 0\}} = 1\Big\},$$

i.e. all vectors with only one nonzero entry. Find the level α GLRT and its power function (leave your answer in terms of appropriate c.d.f)

Solution

The GLRT statistic is

$$\log\lambda(X)=\log\frac{\sup_{\theta\in\Theta_1}L(X;\theta)}{L(X;0)}=\max_{a\in\mathbb{R}}\max_i\Big(-\frac{1}{2}(X_i-a)^2+X_i^2\Big)=$$

$$\max_{a\in\mathbb{R}}\Big(a\max_i X_i-\frac{1}{2}a^2\Big)=\frac{1}{2}\max_i X_i^2$$

and GLRT rejects H_0 one the event $\{\max_i X_i^2\geq c\}$. The critical value c_α and the power function of the level α test are those found in (2), with Φ replaced by the c.d.f. Ψ of χ_1^2 distribution.

(5) Consider the alternative subspace

$$\Theta_1=\Big\{\theta\in\mathbb{R}^n:\theta_m\in\{0,1\},\sum_{m=1}^n\theta_m\leq p\Big\},$$

where p is a fixed integer. In words, Θ is the set of binary vectors with no more than p nonzero entries. Find the GLRT test statistic in this case.

Solution

The GLRT statistic is

$$\log\lambda(X)=\log\frac{\sup_{\theta\in\Theta_1}L(X;\theta)}{L(X;0)}=\max_{|I|\leq p}\Big(-\frac{1}{2}\sum_{m\notin I}X_m^2-\frac{1}{2}\sum_{m\in I}(X_m-1)^2+\frac{1}{2}\sum_{m=1}^n X_m^2\Big)=$$

$$\max_{|I|\leq p}\sum_{m\in I}(X_m-1/2)=\max_{|I|\leq p}\Big(\sum_{m\in I}X_m-|I|/2\Big)$$

i. 2012/2013 (A) 52314

Problem 1. (German tank problem)

The enemy has produced an unknown amount of tanks $N\in\{1,2,...\}$. According to the intelligence, all the tanks are numbered from 1 to N. Our goal is to estimate N, given the serial numbers of k tanks, spotted in the battlefield. Let $X=(X_1,...,X_k)$ be the list of the observed serial numbers in the increasing order.

(1) Construct the statistical model for the obtained data and find the corresponding likelihood function, assuming that any k serial numbers out of N could have been observed with equal probability.

Solution

There are $\binom{N}{k}$ ways to choose k serial numbers out of N. Since the list of the k numbers is ordered, the likelihood function is

$$L(X;N) = \mathbb{P}_n(X_1 = x_1, ..., X_k = x_k) = \frac{1}{\binom{N}{k}} \mathbf{1}_{\{x_1 < ... < x_k \leq N\}}, \quad x_i \in \{1, ..., N\}$$

where the unknown parameter N takes its values in the natural numbers $\mathbb{N}$.

(2) Find the MLE of N

Solution

For a fixed value of $M(x)$, the function $L(x;N)$ vanishes for $N < M(x)$ and decreases in N for $N \geq M(x)$. Hence the MLE of N is $\widehat{N} = M(X) = X_k$, i.e. the maximal serial number among those observed.

(3) Find the minimal sufficient statistic.

Solution

Note that

$$L(x;N) = \frac{1}{\binom{N}{k}} \mathbf{1}_{\{x_1 < ... < x_k\}} \mathbf{1}_{\{x_k \leq N\}}$$

and by the F-N factorization theorem, the statistic $M(x) = \max_i x_i = x_k$ is sufficient. Further, let x and y be two k-tuples of ordered distinct integers, such that $M(x) \neq M(y)$. Then the ratio

$$\frac{L(x;N)}{L(y;N)} = \frac{\mathbf{1}_{\{M(x) \leq N\}}}{\mathbf{1}_{\{M(y) \leq N\}}}$$

is a nonconstant function of N and hence $M(x)$ is minimal sufficient.

(4) Prove that the p.m.f. of $M(X) = \max_i X_i = X_k$ is

$$\mathbb{P}_N\big(M(X) = j\big) = \frac{\binom{j-1}{k-1}}{\binom{N}{k}}, \quad j = k, ..., N.$$

Solution

Clearly, if k serial numbers are observed, the maximal one cannot be less than k and hence $\mathbb{P}_N(M(X) = j) = 0$ for $j < k$. For $j \geq k$, the event $\{M(X) = j\}$ is comprised of k-tuples of serial numbers, containing the integer j and any $k-1$ of the $j-1$ integers, smaller than j. There are $\binom{j-1}{k-1}$ such numbers and the claimed formula follows.

(5) Prove that $M(X)$ is a complete statistic

Solution

Let g be a real valued function on $\{k, ..., N\}$. Then

$$\mathbb{E}_N g(M) = \sum_{j=k}^{N} g(j) \frac{\binom{j-1}{k-1}}{\binom{N}{k}}, \quad N \geq k.$$

We shall argue that $\mathbb{E}_N g(M) = 0$ for all $N \geq k$ implies $g(i) = 0$ for all $i \geq k$. To this end, suppose that $g(i) = 0$ for all $i = k, ..., n$, then

$$\mathbb{E}_{n+1} g(M) = g(n+1) \frac{\binom{n}{k-1}}{\binom{n+1}{k}}$$

and $\mathbb{E}_{n+1} g(M) = 0$ implies $g(n+1) = 0$, i.e. $g(i) = 0$ for all $i = 1, ..., n+1$. Since $\mathbb{E}_k g(M) = g(k)$, it follows that $\mathbb{E}_k g(M) = 0$ implies $g(k) = 0$ and the claim holds by induction.

(6) Find the UMVUE of N

Hint: you may find useful the combinatorial identity

$$\sum_{j=k}^{N} \binom{j}{k} = \binom{N+1}{k+1}, \quad N \geq k.$$

Solution

Let's first see where the hinted combinatorial identity comes from. Summing up over the p.m.f. of M we get

$$\sum_{j=k}^{N} \frac{\binom{j-1}{k-1}}{\binom{N}{k}} = 1.$$

Replacing k with $k+1$ and N with $N+1$ we obtain

$$1 = \sum_{j=k+1}^{N+1} \frac{\binom{j-1}{k}}{\binom{N+1}{k+1}} = \sum_{j=k}^{N} \frac{\binom{j}{k}}{\binom{N+1}{k+1}},$$

which is the claimed identity. Hence

$$\mathbb{E}_N M(X) = \frac{1}{\binom{N}{k}} \sum_{j=k}^{N} j \binom{j-1}{k-1} = \frac{1}{\binom{N}{k}} \sum_{j=k}^{N} j \frac{(j-1)!}{(k-1)!(j-k)!} =$$

$$\frac{k}{\binom{N}{k}} \sum_{j=k}^{N} \frac{j!}{k!(j-k)!} = \frac{k}{\binom{N}{k}} \binom{N+1}{k+1} = \frac{k(N-k)!k!(N+1)!}{N!(k+1)!(N-k)!} = (N+1)\frac{k}{k+1}.$$

Hence the estimator

$$\widetilde{N} = M(X)\big(1 + 1/k\big) - 1$$

is unbiased and, since M is complete, it is the UMVUE.

(7) A calculation shows that

$$\mathrm{var}_N\big(M(X)\big) = \frac{k}{(k+1)^2(k+2)}(N-k)(N+1), \quad 1 \le k \le N.$$

Are the estimator $M(X)$ and the UMVUE comparable for $k \ge 2$? Is any of these estimators inadmissible ?

Solution

The formula for $\mathrm{var}_N(M)$ is obtained by calculations, similar to those in the previous question. The risk of the UMVUE is

$$R(\widetilde{N}, N) = \mathrm{var}_N(\widetilde{N}) = \mathrm{var}_N\big(M\big)\big(1 + 1/k\big)^2,$$

while the risk of $M(X)$ is

$$R(M, N) = \mathrm{var}_N(M) + \big(\mathbb{E}_N M\big)^2.$$

Hence for all N and k

$$\frac{R(M,N)}{R(\widetilde{N},N)} = \left(\frac{k}{1+k}\right)^2 \left(1 + \frac{\big(\mathbb{E}_N M\big)^2}{\mathrm{var}_N\big(M\big)}\right) =$$

$$\left(\frac{k}{1+k}\right)^2 \left(1 + \frac{(N+1)k(k+2)}{(N-k)}\right) \ge \left(\frac{k}{1+k}\right)^2 \left(1 + k^2\right).$$

For $k \ge 2$, the right hand side is greater than 1 for all N and hence the MLE $M(X)$ is uniformly inferior to UMVUE. In particular, MLE is inadmissible.

Problem 2.

It is required to test whether the blood pressure of the patients depends on the deviation of their weight from the nominal value. To this end, n patients are examined and their blood pressures and weight deviations are recorded. The blood pressure of the i-th patient is assumed to satisfy the linear regression model

$$Y_i = ax_i + b + \varepsilon_i,$$

where ε_i's are i.i.d. $N(0,1)$ random variables, x_i is the (non-random) measured weight deviation of the i-th patient and a and b are parameters. Under these assumptions, the lack of dependence corresponds to the value $a = 0$.

(1) Assuming that b is known and $x \neq 0$, find the most powerful test of size α for the problem:

$$H_0 : a = 0$$
$$H_1 : a = a_1,$$

where $a_1 > 0$ is a known number. Calculate the corresponding power.

Solution

By N-P lemma the likelihood ratio test is the most powerful. The likelihood function in this problem is

$$L(Y; a, b) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i - ax_i - b)^2\right)$$

and hence the LRT statistic for the problem at hand is

$$\frac{L(Y; a_1, b)}{L(Y; 0, b)} = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i - a_1x_i - b)^2 + \frac{1}{2}\sum_{i=1}^{n}(Y_i - b)^2\right) =$$
$$\exp\left(\frac{1}{2}\sum_{i=1}^{n} a_1x_i(2Y_i - 2b - a_1x_i)\right) = \exp\left(a_1\sum_{i=1}^{n} x_i(Y_i - b) - \frac{1}{2}\sum_{i=1}^{n} a_1^2x_i^2\right).$$

Hence the LRT test rejects H_0 if and only if $\langle x, Y - b\rangle/\|x\| \geq c$. Note that under $\mathbb{P}_0$,

$$\langle x, Y - b\rangle/\|x\| = \sum_{i=1}^{n} x_i\varepsilon_i/\|x\| \sim N(0,1)$$

and hence

$$\mathbb{P}_0\big(\langle x, Y - b\rangle/\|x\| \geq c\big) = 1 - \Phi(c).$$

Equating this expression to α and solving for c gives

$$c_\alpha = \Phi^{-1}(1 - \alpha).$$

To recap the most powerful α-level test rejects H_0 if and only if

$$\frac{\langle x, Y - b\rangle}{\|x\|} \geq \Phi^{-1}(1 - \alpha).$$

The power of this test is given by

$$\mathbb{P}_1\big(\langle x, Y - b\rangle/\|x\| \geq c_\alpha\big) = \mathbb{P}_1\big(\langle x, a_1x + \varepsilon\rangle/\|x\| \geq c_\alpha\big) =$$
$$\mathbb{P}_1\big(\langle x, \varepsilon\rangle/\|x\| \geq c_\alpha - a_1\|x\|\big) = 1 - \Phi\Big(\Phi^{-1}(1 - \alpha) - a_1\|x\|\Big)$$

(2) Again assuming that b is known, find the UMP test for the problem of deciding whether the dependence is negative or positive, i.e.

$$H_0 : a < 0$$
$$H_1 : a \geq 0.$$

Solution

By the K-R theorem, the LRT test from the previous question is UMP for this problem, since the likelihood functions corresponds to the 1-exponential family.

(3) Assuming that both a and b are unknown and that *not all* x_i's are equal, find the GLRT statistic for the problem of testing the dependence

$$H_0 : a = 0$$
$$H_1 : a \neq 0$$

Solution

The MLE of b under H_0 is $\hat{b}_0 = \bar{Y}$ and

$$\sup_{\theta \in \Theta_0} \log L(Y;\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\frac{1}{n}\sum_i (Y_i - \bar{Y})^2 =: -\frac{n}{2}\log(2\pi) - \frac{n}{2}\widehat{\mathrm{var}}(Y)$$

Under H_1, the MLE's of a and b are the familiar regression coefficients

$$\hat{b}_1 = \bar{Y} - \hat{a}_1 \bar{x},$$

and, assuming $\widehat{\mathrm{var}}(x) > 0$,

$$\hat{a}_1 = \frac{\frac{1}{n}\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} =: \frac{\widehat{\mathrm{cov}}(x, Y)}{\widehat{\mathrm{var}}(x)}.$$

Hence

$$\sup_{\theta \in \Theta_1} \log L(Y;\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\left(\widehat{\mathrm{var}}(Y) - \frac{\widehat{\mathrm{cov}}^2(x, Y)}{\widehat{\mathrm{var}}(x)}\right).$$

Consequently, the GLRT statistic is given by

$$\log \lambda(Y) = -\frac{n}{2}\left(\widehat{\mathrm{var}}(Y) - \frac{\widehat{\mathrm{cov}}^2(x, Y)}{\widehat{\mathrm{var}}(x)}\right) + \frac{n}{2}\widehat{\mathrm{var}}(Y) = \frac{n}{2}\frac{\widehat{\mathrm{cov}}^2(x, Y)}{\widehat{\mathrm{var}}(x)},$$

and H_0 is rejected if and only if

$$\frac{|\widehat{\mathrm{cov}}(x, Y)|}{\sqrt{\widehat{\mathrm{var}}(x)}} \geq c.$$

(4) Find the critical value of level α GLRT test.

Solution

Under H_0,

$$\widehat{\mathrm{cov}}(x,Y) = \frac{1}{n}\sum_i (Y_i - \bar{Y})(x_i - \bar{x}) = \frac{1}{n}\sum_i (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x}).$$

Let $\xi_i := \varepsilon_i - \bar{\varepsilon}$ and note that

$$\mathbb{E}\xi_i\xi_j = \begin{cases} 1 - 1/n, & i = j \\ -1/n, & i \neq j \end{cases} = \delta_{ij} - 1/n$$

Hence

$$\mathbb{E}\left(\sum_i (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})\right)^2 = \sum_i \sum_j \mathbb{E}\xi_i\xi_j (x_i - \bar{x})(x_j - \bar{x}) =$$
$$\sum_i \sum_j (\delta_{ij} - 1/n)(x_i - \bar{x})(x_j - \bar{x}) =$$
$$\sum_i (x_i - \bar{x})^2 - \frac{1}{n}\sum_i (x_i - \bar{x}) \sum_j (x_j - \bar{x}) = n\widehat{\mathrm{var}}(x)$$

and

$$\frac{\widehat{\mathrm{cov}}(x,Y)}{\sqrt{\widehat{\mathrm{var}}(x)}} \sim N\left(0, \frac{1}{n}\right). \tag{0i1}$$

Consequently,

$$\mathbb{P}_0\left(\left|\frac{\widehat{\mathrm{cov}}(x,Y)}{\sqrt{\widehat{\mathrm{var}}(x)}}\right| \geq c\right) = \mathbb{P}_0\left(\sqrt{n}\left|\frac{\widehat{\mathrm{cov}}(x,Y)}{\sqrt{\widehat{\mathrm{var}}(x)}}\right| \geq \sqrt{n}c\right) = 2\Phi(-\sqrt{n}c),$$

and $c_\alpha = -\frac{1}{\sqrt{n}}\Phi^{-1}(\alpha/2)$.

(5) A sequence of tests is said to be *consistent* if their powers converge to 1 as $n \to \infty$ at all alternatives. Derive the sufficient and necessary condition on the sequence (x_i) under which the corresponding sequence of tests from the previous question is consistent.

Solution

Note that for $a \neq 0$ and $b \in \mathbb{R}$,

$$\widehat{\mathrm{cov}}(x,Y) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) =$$
$$\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\big(a(x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}\big) = a\widehat{\mathrm{var}}(x) + \sqrt{\frac{\widehat{\mathrm{var}}(x)}{n}}Z,$$

where $Z \sim N(0,1)$. Hence the power of GLRT is given by

$$\mathbb{P}_{a,b}\left(\left|\frac{\widehat{\mathrm{cov}}(x,Y)}{\sqrt{\widehat{\mathrm{var}}(x)}}\right| \geq c_\alpha\right) = \mathbb{P}_{a,b}\left(\left|\sqrt{n}a\sqrt{\widehat{\mathrm{var}}(x)} + Z\right| \geq -\Phi^{-1}(\alpha/2)\right) =$$
$$1 - \mathbb{P}_{a,b}\left(\left|\sqrt{n}a\sqrt{\widehat{\mathrm{var}}(x)} + Z\right| \leq -\Phi^{-1}(\alpha/2)\right) \xrightarrow{n\to\infty} 1,$$

where the convergence holds if and only if $n\widehat{\mathrm{var}}(x) \to \infty$ as $n \to \infty$.

(6) Show that if all x_i's are equal to a nonzero constant, no consistent sequence of tests exists for the problem from (3).

Solution

Suppose δ_n is a consistent sequence of level α tests, i.e.

$$\mathbb{E}_0\delta_n \leq \alpha, \quad \forall n \geq 1$$

and $x_i \equiv x$ for all i's. Let (a^*, b^*) be such that $b^* = -ax^*$, e.g., $a^* = 1$ and $b^* = -x$, then $\mathbb{P}_{a^*,b^*} = \mathbb{P}_0$ and hence the power satisfies $\mathbb{E}_{a^*,b^*}\delta_n \leq \alpha$ for all $n \geq 1$, which contradicts consistency.

(7) Answer the questions (3) and (4), assuming that the covariates X_i's are i.i.d. r.v. with density f, independent of ε_i's.

Solution

The log-likelihood in this case is

$$\log L(Y, X; \theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_i (Y_i - a - bX_i)^2 + \sum_{i=1}^n \log f(X_i)$$

Since X_i's have density, $\widehat{\mathrm{var}}(X) > 0$ with probability one and hence the GLRT statistic doesn't change. Further, under H_0,

$$\frac{\widehat{\mathrm{cov}}(X,Y)}{\sqrt{\widehat{\mathrm{var}}(X)}} = \frac{\widehat{\mathrm{cov}}(X,\varepsilon)}{\sqrt{\widehat{\mathrm{var}}(X)}}.$$

Since ε_i's and X_i's are independent,

$$\mathbb{E}\exp\left(it\frac{\widehat{\mathrm{cov}}(X,Y)}{\sqrt{\widehat{\mathrm{var}}(X)}}\right) = \mathbb{E}\mathbb{E}\left(\exp\left(it\frac{\widehat{\mathrm{cov}}(X,\varepsilon)}{\sqrt{\widehat{\mathrm{var}}(X)}}\right)\middle| X_1, ..., X_n\right) = \exp\left(-\frac{t^2}{2}\frac{1}{n}\right),$$

where the last equality follows from (0i1). Hence

$$\frac{\widehat{\mathrm{cov}}(X,Y)}{\sqrt{\widehat{\mathrm{var}}(X)}} \sim N\left(0, \frac{1}{n}\right),$$

and the critical value remains the same as in the previous question.

j. 2012/2013 (B) 52314

Problem 1. (Fisher's Nile problem[41])

Let $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$ be independent samples from $\mathrm{Exp}(\theta)$ and $\mathrm{Exp}(1/\theta)$ respectively, where $\theta \in \mathbb{R}_+$ is the unknown parameter to be estimated from (X, Y).

(1) Find the minimal sufficient statistic.

Solution

The likelihood function is

$$L(X, Y; \theta) = \exp\left(-\theta \sum_i X_i - \frac{1}{\theta} \sum_i Y_i\right),$$

and $(\bar{X}_n, \bar{Y}_n)$ is a sufficient statistic. To check minimality, let (x, y) and (x', y') be such that $(\bar{x}, \bar{y}) \neq (\bar{x}', \bar{y}')$, then

$$\frac{L(x, y; \theta)}{L(x', y'; \theta)} = \exp\left(-\theta n(\bar{x} - \bar{x}') - \frac{n}{\theta}(\bar{y} - \bar{y}')\right),$$

is a non-constant function of θ. Hence the statistic is minimal sufficient.

(2) Is the minimal sufficient statistic complete?

Solution

Note that

$$\mathbb{E}_\theta \bar{X}_n \bar{Y}_n - 1 = \mathbb{E}_\theta \bar{X}_n \mathbb{E}_\theta \bar{Y}_n - 1 = 0, \quad \forall \theta > 0,$$

while $\bar{X}_n \bar{Y}_n - 1$ is not a constant random variable. Hence the statistic is incomplete.

(3) Apply the R-B lemma to Y_1, using the minimal sufficient statistic, to obtain an improved unbiased estimator of θ

Hint: avoid complicated calculations

Solution

By independence of X and Y,

$$\mathbb{E}_\theta(Y_1 | \bar{X}_n, \bar{Y}_n) = \mathbb{E}_\theta(Y_1 | \bar{Y}_n) = \bar{Y}_n.$$

(4) Find the C-R bound for the MSE risk of unbiased estimators of θ. Is the estimator, found in the previous question, efficient?

Solution

[41] The question of existence of the UMVUE for the setting of this question is known as Fisher's Nile problem and remains open already for many years (see, however, the recent progress in `http://arxiv.org/abs/1302.0924`)

The Fisher information of (X_1, Y_1) is

$$I(\theta) := -\mathbb{E}_\theta \partial_\theta^2 \log \exp\Big(-\theta X_1 - \frac{1}{\theta} Y_1 \Big) = \frac{2}{\theta^3} \mathbb{E}_\theta Y_1 = \frac{2}{\theta^2}$$

and hence the variance of any unbiased estimator is greater or equal to $\frac{\theta^2}{2}\frac{1}{n}$. Since $\mathrm{var}_\theta(\bar{Y}_n) = \theta^2 \frac{1}{n}$, the estimator $\bar{Y}_n$ is not efficient.

(5) Find the MLE of θ.

Solution

Differentiating the log-likelihood w.r.t. θ and equating to zero, we get

$$\hat{\theta} = \sqrt{\bar{Y}_n / \bar{X}_n},$$

which is readily checked to be the only maximum, by inspecting the second derivative.

(6) Is the MLE consistent?

Solution

By the LLN, $\bar{X}_n \xrightarrow{\mathbb{P}_\theta} 1/\theta$ and $\bar{Y}_n \xrightarrow{\mathbb{P}_\theta} \theta$ and hence by Slutsky's lemma and continuity of the square root, $\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta$, for all $\theta > 0$, i.e., MLE is consistent.

(7) Consider the sequence of estimators

$$\hat{\theta}_n = \frac{1}{2}\left(\bar{Y}_n + \frac{1}{\bar{X}_n} \right).$$

Is it consistent? Asymptotically normal? If yes, find the asymptotic variance and compare it to the C-R bound from (4).

Solution

The sequence of estimators is consistent, similarly to (6). Since $\sqrt{n}(\bar{X}_n - 1/\theta) \xrightarrow{d} N(0, 1/\theta^2)$, the Δ-method applies to $g(s) = 1/s$ and $s = 1/\theta$:

$$\sqrt{n}\Big(\frac{1}{\bar{X}_n} - \frac{1}{1/\theta} \Big) = \sqrt{n}\Big(g(\bar{X}_n) - g(s) \Big) \xrightarrow{d} N\Big(0, V(s)\Big),$$

with

$$V(s) = \big(g'(s)\big)^2 s^2 = \Big(\frac{1}{s^2}\Big)^2 s^2 = \frac{1}{s^2} = \theta^2.$$

Since $\bar{Y}_n$ and $\bar{X}_n$ are independent, and $\sqrt{n}(\bar{Y}_n - \theta) \xrightarrow{d} N(0, \theta^2)$, it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\Big(0, \frac{\theta^2}{2}\Big).$$

The asymptotic variance coincides with the Fisher information rate, i.e. the estimator is asymptotically efficient in Fisher's sense.

Problem 2.

An integer valued quantity (e.g., the number of emitted particles per second) is measured by a device, which is suspected to introduce censoring at some level K. More precisely, let $Z_1, ..., Z_n$ be a sample from Geo(p) distribution with known $p \in (0,1)$ and define $X_i = \min(Z_i, K+1)$, where K is the unknown censoring level.

Below we shall explore hypothesis testing problems for K, given the censored data $X = (X_1, ..., X_n)$. For notational convenience, we shall regard $K = \infty$ as a point in the parametric space $\mathbb{N} \cup \{\infty\}$ and interpret it as no censoring, i.e. $X_1, ..., X_n$ are i.i.d Geo(p) r.v. under $\mathbb{P}_\infty$.

(1) Show that for $K \in \mathbb{N}$ the likelihood function is given by the formula

$$L_n(X;K) = (1-p)^{S_n(X)-n} p^{n-C_n(X;K)} \mathbf{1}_{\{M_n(X) \le K+1\}},$$

where

$$S_n(X) = \sum_{i=1}^n X_i$$
$$M_n(X) = \max_i X_i$$
$$C_n(X;K) = \sum_{i=1}^n \mathbf{1}_{\{X_i = K+1\}}$$

Solution

For $K < \infty$

$$\mathbb{P}_K(X_1 = m) = \begin{cases} p(1-p)^{m-1} & m = 1, ..., K \\ (1-p)^K & m = K+1 \\ 0 & m > K+1 \end{cases}$$

and hence the likelihood function is given by

$$\begin{aligned} L_n(X^n;K) &= \prod_{i=1}^n \left(p(1-p)^{X_i-1} \mathbf{1}_{\{X_i \le K\}} + (1-p)^K \mathbf{1}_{\{X_i = K+1\}} \right) \mathbf{1}_{\{X_i \le K+1\}} \\ &= \mathbf{1}_{\{\max_i X_i \le K+1\}} \prod_{i=1}^n \left(p(1-p)^{X_i-1} \mathbf{1}_{\{X_i \le K\}} + (1-p)^{X_i-1} \mathbf{1}_{\{X_i = K+1\}} \right) \\ &= \mathbf{1}_{\{\max_i X_i \le K+1\}} \prod_{i=1}^n (1-p)^{X_i-1} \left(p \mathbf{1}_{\{X_i \le K\}} + \mathbf{1}_{\{X_i = K+1\}} \right) \\ &= \mathbf{1}_{\{\max_i X_i \le K+1\}} (1-p)^{\sum_i (X_i-1)} p^{\sum_i \mathbf{1}_{\{X_i \le K\}}} \\ &= \mathbf{1}_{\{M_n(X) \le K+1\}} (1-p)^{S_n(X)-n} p^{n-C_n(X;K)} \end{aligned}$$

(2) Find the most powerful test statistic for the problem of testing for presence of the known censoring level K_0:

$$H_0 : K = K_0$$
$$H_1 : K = \infty$$

Solution

The most powerful test rejects H_0 if and only if

$$\Big\{L_n(X^n;\infty) \geq cL_n(X^n;K_0)\Big\} =$$
$$\Big\{\mathbf{1}_{\{M_n(X)\leq K_0+1\}}p^{-C_n(X;K_0)} \leq \frac{1}{c}\Big\} =$$
$$\Big\{M_n(X) > K_0+1\Big\}\bigcup\Big\{p^{-C_n(X;K_0)} \leq \frac{1}{c}\Big\} =$$
$$\Big\{M_n(X) > K_0+1\Big\}\bigcup\Big\{C_n(X;K_0) \leq c'\Big\},$$

where $c' = \log c/\log p$, i.e., either when the maximum exceeds K_0+1 or when the number of censorings is low.

(3) Using the CLT approximation, find the asymptotic MP test of size $\alpha \in (0,1)$.

Solution

Under H_0, $\Big\{M_n(X) > K_0+1\Big\} = \emptyset$ and $C_n(X;K_0) \sim \text{Bin}(p^{K_0},n)$ and hence

$$\mathbb{P}_{K_0}\Big(\{M_n(X) > K_0+1\}\cup\{C_n(X;K_0)\leq c'\}\Big) = \mathbb{P}_{K_0}\Big(C_n(X;K_0)\leq c'\Big) =$$
$$\mathbb{P}_{K_0}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbf{1}_{\{X_i=K_0+1\}}-(1-p)^{K_0}}{\sqrt{(1-p)^{K_0}(1-(1-p)^{K_0})}} \leq \frac{\frac{c'}{\sqrt{n}}-\sqrt{n}(1-p)^{K_0}}{\sqrt{(1-p)^{K_0}(1-(1-p)^{K_0})}}\right).$$

If we choose

$$c' := \Phi^{-1}(\alpha)\sqrt{n(1-p)^{K_0}(1-(1-p)^{K_0})}+n(1-p)^{K_0},$$

the level of the test will converge to α as $n\to\infty$ by the CLT.

(4) Is the obtained test consistent, i.e. does its power converge to 1 as $n\to\infty$?

Solution

The test is consistent:

$$\mathbb{P}_\infty\Big(\{M_n(X) > K_0+1\}\cup\{C_n(X;K_0)\leq c'\}\Big) =$$
$$1-\mathbb{P}_\infty\Big(\{M_n(X) \leq K_0+1\}\cap\{C_n(X;K_0) > c'\}\Big) \geq$$
$$1-\mathbb{P}_\infty\Big(M_n(X)\leq K_0+1\Big) = 1-\Big(\mathbb{P}_\infty\big(X_1\leq K_0+1\big)\Big)^n \xrightarrow{n\to\infty} 1.$$

(5) Prove that the GLRT for the problem of testing for presence of an unknown censoring level:

$$H_0 : K < \infty$$
$$H_1 : K = \infty$$

rejects H_0 if and only if

$$\left\{ \sum_{i=1}^{n} \mathbf{1}_{\{X_i = M_n(X)\}} \leq c \right\}$$

for some critical value $c > 0$.

Solution

Note that the likelihood $L_n(X; K)$ is maximized by the maximizer $\hat{K}$ of

$$p^{n - C_n(X;K)} \mathbf{1}_{\{M_n(X) \leq K+1\}},$$

over $K \in \mathbb{N}$. Clearly, $\hat{K} \geq M_n(X) - 1$ and, since $C_n(X; K) = 0$ for $K > M_n(X) - 1$ and $C_n(X; K) > 0$ for $K = M_n(X) - 1$, we get $\hat{K} = M_n(X) - 1$. Consequently, the GLRT statistic is given by

$$\lambda_n(X) = \frac{L_n(X; \infty)}{\max_{K \in \mathbb{N}} L_n(X; K)} = \frac{1}{\max_{K \in \mathbb{N}} p^{-C_n(X;K)} \mathbf{1}_{\{M_n(X) \leq K+1\}}} =$$
$$p^{C_n(X; M_n(X) - 1)} = p^{\sum_i \mathbf{1}_{\{X_i = M_n(X)\}}}$$

and hence the GLRT rejects H_0 if and only if the number of maxima in the sample is small

$$\sum_{i=1}^{n} \mathbf{1}_{\{X_i = M_n(X)\}} \leq c,$$

where c is the critical value.

(6) Show that for any $K \in \mathbb{N}$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i = M_n(X)\}} \xrightarrow[n \to \infty]{\mathbb{P}_K} (1 - p)^K.$$

Hint: Find the limit of $M_n(X)$ under $\mathbb{P}_K$ and figure out how to apply LLN

Solution

Under $\mathbb{P}_K$, $M_n(K)$ converges to $K+1$ and hence $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}}$ essentially behaves as $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=K+1\}}$ for large n. More precisely, since $M_n(X) \le K+1$, $\mathbb{P}_K$-a.s.

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} =$$
$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=K+1\}}\mathbf{1}_{\{M_n(X)=K+1\}} + \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}}\mathbf{1}_{\{M_n(X)<K+1\}} =$$
$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=K+1\}} + \frac{1}{n}\sum_{i=1}^n \Big(\mathbf{1}_{\{X_i=M_n(X)\}} - \mathbf{1}_{\{X_i=K+1\}}\Big)\mathbf{1}_{\{M_n(X)<K+1\}}.$$

The first term converges in $\mathbb{P}_K$-probability to $(1-p)^K$ by the LLN, while the second term converges to zero:

$$\mathbb{E}_K\left|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{1}_{\{X_i=M_n(X)\}} - \mathbf{1}_{\{X_i=K+1\}}\Big)\mathbf{1}_{\{M_n(X)<K+1\}}\right| \le$$
$$2\mathbb{P}_K(M_n(X) < K+1) \le 2\big(\mathbb{P}_K(X_1 < K+1)\big)^n = 2\big(1-(1-p)^K\big)^n \xrightarrow{n\to\infty} 0,$$

and the claim follows by Slutsky's lemma.

(7) It can be shown that for any $\gamma > 0$,

$$\frac{1}{n^\gamma}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} \xrightarrow[n\to\infty]{\mathbb{P}_\infty} 0.$$

Using this limit and the result of the previous question, suggest a sequence of critical values c_n, so that the corresponding sequence of GLRT's significance levels converge to 0 for any fixed null hypothesis and the power converge to 1 as $n \to \infty$.

Solution

Take any $\gamma \in (0,1)$ and let $c_n = n^\gamma$, then for any $K \in \mathbb{N}$, the significance level converges to zero:

$$\mathbb{P}_K\left(\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} \le c_n\right) = \mathbb{P}_K\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} \le n^{\gamma-1}\right) =$$
$$\mathbb{P}_K\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} - (1-p)^K \le n^{\gamma-1} - (1-p)^K\right) \xrightarrow{n\to\infty} 0.$$

where the convergence holds by (6). On the other hand, the power converges to 1:

$$\mathbb{P}_\infty\left(\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} \le c_n\right) = \mathbb{P}_\infty\left(\frac{1}{n^\gamma}\sum_{i=1}^n \mathbf{1}_{\{X_i=M_n(X)\}} \le 1\right) \xrightarrow{n\to\infty} 1.$$

Where does the claimed limit come from? Under $\mathbb{P}_\infty$, $M_n(X) \to \infty$ and hence the number of times the maximum is attained till time n might grow slower than linearly.

In fact, according to the claim, it grows slower than polynomially. Let's see why this is true. Since X_i's are i.i.d.

$$\mathbb{P}_\infty\big(X_i = M_n(X)\big) = \mathbb{P}_\infty\big(X_1 = M_n(X)\big) = \mathbb{P}_\infty\Big(\bigcap_{i=2}^{n}\{X_i \le X_1\}\Big) =$$

$$\mathbb{E}_\infty\mathbb{P}_\infty\Big(\bigcap_{i=2}^{n}\{X_i \le X_1\}|X_1\Big) = \mathbb{E}_\infty\big(1-(1-p)^{X_1}\big)^{n-1}.$$

For any $z > 0$,

$$\mathbb{E}_\infty\big(1-(1-p)^{X_1}\big)^{n-1} =$$
$$\mathbb{E}_\infty\big(1-(1-p)^{X_1}\big)^{n-1}\mathbf{1}_{\{X_1\le z\}} + \mathbb{E}_\infty\big(1-(1-p)^{X_1}\big)^{n-1}\mathbf{1}_{\{X_1> z\}} \le$$
$$\big(1-(1-p)^{z}\big)^{n-1} + \mathbb{P}_\infty(X_1 > z) = \big(1-(1-p)^{z}\big)^{n-1} + (1-p)^z =$$
$$\big(1-e^{-z|\log(1-p)|}\big)^{n-1} + e^{-z|\log(1-p)|}$$

Let $\delta \in (0,1)$ and $z_n := \frac{\log n^\delta}{|\log(1-p)|}$. Then $e^{-z_n|\log(1-p)|} = n^{-\delta}$ and

$$\mathbb{P}_\infty\big(X_1 = M_n(X)\big) \le 2(1-e^{-z_n|\log(1-p)|})^n + e^{-z_n|\log(1-p)|} =$$
$$2(1-n^{-\delta})^{n\delta+n(1-\delta)} + n^{-\delta} \le 2\exp(-n(1-\delta)) + n^{-\delta} \le 2n^{-\delta},$$

for all n large enough, where we used the bound $(1-x^{-1})^x \le e^{-1}$ for $x \ge 1$. Consequently, with $\gamma := 1-\delta(1-\delta)$ and some constant C,

$$\mathbb{E}_\infty n^{-\gamma}\sum_{i=1}^{n}\mathbf{1}_{\{X_i=M_n(X)\}} = \mathbb{E}_\infty n^{-1+\delta(1-\delta)}\sum_{i=1}^{n}\mathbf{1}_{\{X_i=M_n(X)\}} \le$$
$$Cn^{-1+\delta(1-\delta)}nn^{-\delta} = Cn^{-\delta^2} \xrightarrow{n\to\infty} 0,$$

as claimed.

Bibliography

[1] P.Bickel, K.Doksum, Mathematical Statistics: basic ideas and selected topics, 1977

[2] Borovkov, A. A. Mathematical statistics. Gordon and Breach Science Publishers, Amsterdam, 1998.

[3] Casella, George; Berger, Roger L. Statistical inference, 1990.

[4] Efron, Bradley. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Vol. 1. Cambridge University Press, 2010.

[5] Ghosh, Malay, Basu's theorem with applications: a personalistic review. Special issue in memory of D. Basu. Sankhya- Ser. A 64 (2002), no. 3, part 1, 509531.

[6] Ibragimov, I. A.; Khasminskii, R. Z. Statistical estimation. Asymptotic theory. Applications of Mathematics, 16. Springer-Verlag, New York-Berlin, 1981

[7] Kagan, Abram; Yu, Tinghui, Some inequalities related to the Stam inequality. Appl. Math. 53 (2008), no. 3, 195–205.

[8] Lehmann, E. L. Casella, George Theory of point estimation, 2nd ed., Springer Texts in Statistics. Springer-Verlag, New York, 1998.

[9] Lehmann, E. L.; Romano, Joseph P. Testing statistical hypotheses. 3d edition. Springer Texts in Statistics. Springer, New York, 2005

[10] Reeds, James A. "Asymptotic number of roots of Cauchy location likelihood equations." The Annals of Statistics (1985): 775-784.

[11] Samworth, R. J., Stein's Paradox. Eureka, 62, 38-41, 2012

[12] Shao, Jun, Mathematical statistics. Springer Texts in Statistics. Springer-Verlag, New York, 2003

[13] Shao, Jun, Mathematical statistics: exercises and solutions. Springer, New York, 2005

[14] A.Shiryaev, Probability, Spinger

[15] A.Tsybakov, Introduction to Nonparametric Estimation, Springer New York 2009

[16] R. A. Wijsman, On the Attainment of the Cramer-Rao Lower Bound, The Annals of Statistics, Vol. 1, No. 3 (May, 1973), pp. 538-542