Notes on statistics

Matthias Birkner[1]

Preliminary version from 28th April 2026

[1] birkner@mathematik.uni-mainz.de

Contents

Chapter 0

A prelude

An example. 48 participants in a management course were each presented with a (fictional) personnel file, and they were asked, on the basis of the file's contents, to decide whether to promote the person concerned or to hold the file and assess further candidates. The files were identical apart from the sex indication — 24 were marked as "female" and 24 as "male" — and were allocated purely at random to the participants.

The following result was obtained[1]:

	Female	Male
Promote	14	21
hold file	10	3

In particular, $14/24 \approx 58,3\%$ of the "female" files were promoted but $21/24 = 87,5\%$ of the "male" files.

Can this distribution be explained by "pure chance"?

Consider the following thought experiment: The 48 assessors, 35 "lenient" and 13 "strict", draw without replacement one file each from an urn containing 24 "female" and 24 "male" files, so that

$$X := \text{number of promoted "male" files } \sim \mathrm{Hyp}_{24,24,35}$$

and the probability of observing a deviation as large as the one actually seen in the substitution experiment would be

$$\mathbb{P}(X \geq 21) + \mathbb{P}(X \leq 14) = \mathrm{Hyp}_{24,24,35}\big(\{11,12,13,14\} \cup \{21,22,23,24\}\big) \approxeq 0.049$$

(this is the so-called p-value of the test).

Thus: The hypothesis that "the distribution arises from pure chance" appears implausible to us. (In statistical jargon: "We reject this (null) hypothesis.")

Remark. The above testing procedure is called Fisher's exact test (in the two-sided version).

[1] From Benson Rosen, Thomas H. Jerdee, Influence of sex role stereotypes on personnel decisions, J. Appl. Psych. **59**, 9–14, 1974; see Table 1 there (only the "simple job" part)

Chapter 1

Basic notions and some classical results

1.1 Statistical models

Definition 1.1.1. A *statistical model* is a triple $(\mathcal{M} =) \big(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}\big)$, where $\Omega \neq \emptyset$ is a set ("sample space"), $\mathscr{F} \subset 2^\Omega$ is a σ-algebra, Θ is a set (with $|\Theta| > 1$), and for each $\vartheta \in \Theta$, $\mathbb{P}_\vartheta$ is a probability measure on $(\Omega, \mathscr{F})$.

The model $\mathcal{M}$ is called *parametric* if $\Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$, in particular *one-parameter* if $d = 1$.

$\mathcal{M}$ is called *discrete* if Ω is countable (and $\mathbb{P}_\vartheta$ has weight function ρ_ϑ), $\mathcal{M}$ is called *continuous* if $\Omega \subset \mathbb{R}^n$ and each $\mathbb{P}_\vartheta$ has a density $\rho_\vartheta : \Omega \to [0, \infty]$.

A discrete or continuous model is called a *standard model*. The function $\Theta \times \Omega \ni (\vartheta, x) \mapsto \rho_\vartheta(x) =: \rho(\vartheta, x)$ is then called the *likelihood function*.

For $n \in \mathbb{N}$, $\big(\Omega^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta}\big)$ is called the n-fold product model of $\mathcal{M}$, this models the n-fold independent repetition of the experiment described by $\mathcal{M}$.

Definition 1.1.2. $\big(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}\big)$ a statistical model, $(S, \mathscr{A})$ a measurable space.

1. A random variable X (defined on $(\Omega, \mathscr{F})$ taking values in S, i.e. $X : \Omega \to S$ is $\mathscr{F}$-$\mathscr{A}$-measurable) is called a *statistic* (sometimes also: "sample").

2. Let $\tau : \Theta \to \mathbb{R}$ be a real parameter characteristic (or "parameter feature"), a statistic $T : \Omega \to \mathbb{R}$ is called an *estimator* (more precisely: "point estimator") for τ.

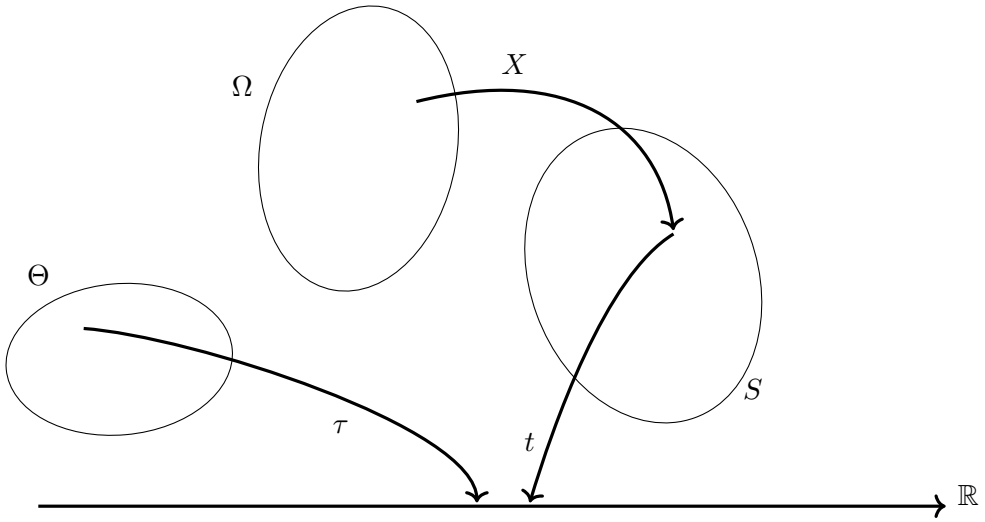
3. An estimator T for τ is called *unbiased* if

$$\forall\, \vartheta \in \Theta \ : \quad \mathbb{E}_\vartheta[T] = \vartheta.$$

   $b_\vartheta(T) \coloneqq \mathbb{E}_\vartheta[T] - \vartheta$ is called the *bias* of T.

The typical construction / situation of an estimator is $T = t(X)$ for a (measurable) function $t : S \to \mathbb{R}$.

Notation convention. An estimator for τ is often denoted $\widehat{\tau}$.



Schematic representation of an estimator $T = t(X)$ for τ

Observation 1.1.3 (Unbiased estimators for mean and variance in the product model). For $\vartheta \in \Theta$, let Q_ϑ be a probability measure on $\mathbb{R}$ with finite mean

$$m(\vartheta) := \int_{\mathbb{R}} x \, Q_\vartheta(dx)$$

and finite variance

$$v(\vartheta) := \int_{\mathbb{R}} \big(x - m(\vartheta)\big)^2 \, Q_\vartheta(dx).$$

Under $\mathbb{P}_\vartheta$, let $X_1, \dots, X_n$ be i.i.d., $X_i \sim Q_\vartheta$. (In the formalisation of Definition 1.1.1 and 1.1.2, we could choose: $\mathcal{M} = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ with $\mathbb{P}_\vartheta = Q_\vartheta^{\otimes n}$ for $\vartheta \in \Theta$, as a statistic we consider $X = (X_1, \dots, X_n)$ where $X_i : \mathbb{R}^n \to \mathbb{R}$ is the projection onto the i-th coordinate.)

Note: this may be a non-parametric model; for example, one could choose

$$\Theta := \big\{Q : Q \text{ is a probability measure on } \mathbb{R} \text{ with } \textstyle\int_{\mathbb{R}} x^2 \, Q(dx) < \infty\big\}.$$

Then

$$\overline{X} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{is an unbiased estimator for } m(\vartheta),$$

$$S^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2 \quad \text{is an unbiased estimator for } v(\vartheta).$$

In this context, $\overline{X}$ is also called the empirical mean or sample mean, S^2 the (corrected) sample variance.

For $\vartheta \in \Theta$, we have

$$\mathbb{E}_\vartheta[\overline{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta[X_i] = \frac{1}{n}\cdot n\, m(\vartheta) = m(\vartheta),$$

$$\begin{aligned}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\Big] &= n\mathbb{E}_\vartheta\big[(X_i-\overline{X})^2\big] = n\,\mathrm{Var}_\vartheta\big[X_i - \overline{X}\big] \\ &= n\,\mathrm{Var}_\vartheta\Big[\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{i=2}^n X_i\Big] = n\Big(\big(\tfrac{n-1}{n}\big)^2\mathrm{Var}_\vartheta[X_1] + \tfrac{n-1}{n^2}\mathrm{Var}_\vartheta[X_1]\Big) \\ &= (n-1)\,\mathrm{Var}_\vartheta[X_1],\end{aligned}$$

hence

$$\mathbb{E}_\vartheta\big[S^2\big] = \frac{1}{n-1}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i-\overline{X}\big)^2\Big] = v(\vartheta).$$

Definition and observation 1.1.4 (Consistency)**.** Consider in the situation of Observation 1.1.3 the sample size n as variable (formally: pass to the infinite product model $\mathcal{M} = (\mathbb{R}^\infty, \mathcal{B}(\mathbb{R})^{\otimes\infty}, (Q_\vartheta^{\otimes\infty})_{\vartheta\in\Theta})$) with $X_i : \mathbb{R}^\infty \to \mathbb{R}$ the projection onto the i-th coordinate. Then for each $\vartheta \in \Theta$

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \underset{n\to\infty}{\longrightarrow} m(\vartheta) \quad \text{in probability w.r.t. } \mathbb{P}_\vartheta := (Q_\vartheta^{\otimes\infty} \quad \text{and}$$

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 \underset{n\to\infty}{\longrightarrow} v(\vartheta) \quad \text{in probability w.r.t. } \mathbb{P}_\vartheta.$$

We say: These (sequences of) estimator(s) are *consistent*.

For $\overline{X}_n$ this follows directly from the law of large numbers, furthermore

$$\begin{aligned}\frac{n-1}{n}S_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i-\overline{X}_n\big)^2 &= \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - 2\Big(\frac{1}{n}\sum_{i=1}^n X_i\overline{X}_n\Big) + \big(\overline{X}_n\big)^2 = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - \big(\overline{X}_n\big)^2 \\ &\underset{n\to\infty}{\overset{\mathbb{P}_\vartheta}{\longrightarrow}} \int_{\mathbb{R}} x^2\, Q_\vartheta(dx) - \big(m(\vartheta)\big)^2 = v(\vartheta)\end{aligned}$$

again by the law of large numbers.

Definition 1.1.5 (UMVU estimator)**.** Let $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical standard model, $\rho(\vartheta, x)$ the likelihood function. An unbiased estimator T for a real parameter characteristic $\tau(\vartheta)$ is called *variance-minimising*, if for every other unbiased estimator $\widetilde{T}$ for τ we have

$$\mathrm{Var}_\vartheta[T] \le \mathrm{Var}_\vartheta[\widetilde{T}] \quad \text{for all } \vartheta \in \Theta.$$

Definition 1.1.6. A one-parameter standard model (i.e. $\Theta \subset \mathbb{R}$) is called *regular* if the following hold:

(i) $\Theta \subset \mathbb{R}$ is an open interval.

(ii) The likelihood function $\rho(\vartheta, x)$ is strictly positive on $\Theta \times \Omega$ and for each x, $\vartheta \mapsto \rho(\vartheta, x)$ is continuously differentiable.

(iii) $U_\vartheta(x) \coloneqq \frac{\partial}{\partial\vartheta}\log\rho(\vartheta,x)$ satisfies

$$I_\vartheta \coloneqq \mathrm{Var}_\vartheta[U_\vartheta] \in (0,\infty)$$

and it holds

$$\textstyle\int_\Omega \frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx = \frac{d}{d\vartheta}\int_\Omega \rho(\vartheta,x)\,dx \;\; (=0).$$

(U_ϑ is called the *score function* and I_ϑ is called the *Fisher information* [in the model].)

Furthermore, an estimator T is called regular if for each $\vartheta\in\Theta$ we have

$$\frac{d}{d\vartheta}\int_\Omega T(x)\rho(\vartheta,x)\,dx = \int_\Omega T(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx.$$

(If Ω is discrete, the integral $\int_\Omega \ldots dx$ should be replaced by the sum $\sum_{x\in\Omega}\ldots$.)

Theorem 1.1.7 (Cramér-Rao bound)**.** *Let $\tau:\Theta\to\mathbb{R}$ be a continuously differentiable parameter characteristic, T a regular, unbiased estimator for τ in a regular standard model. Then the Cramér-Rao bound holds:*

$$\mathrm{Var}_\vartheta[T] \geq \frac{(\tau'(\vartheta))^2}{I_\vartheta} \quad \forall\,\vartheta\in\Theta,$$

where equality holds precisely when

$$T(x)-\tau(\vartheta) = \frac{\tau'(\vartheta)}{I_\vartheta}U_\vartheta(x).$$

(For $\tau=\mathrm{Id}$ in particular: Every unbiased estimator for ϑ has variance $\geq 1/I_\vartheta$.)

Proof. Let $T:\Omega\to\mathbb{R}$ be a regular, unbiased estimator for $\tau(\vartheta)$, i.e. $\mathbb{E}_\vartheta[T]=\tau(\vartheta)$ for all $\vartheta\in\Theta$. For fixed $\vartheta\in\Theta$, we can consider both T and U_ϑ as random variables in $\mathcal{L}^2(\mathbb{P}_\vartheta)$ (if $T\notin\mathcal{L}^2(\mathbb{P}_\vartheta)$, we have $\mathrm{Var}_\vartheta[T]=\infty$ and there is nothing to prove). We consider in the following the case of a continuous model, i.e. $\mathbb{P}_\vartheta$ has a density w.r.t. Lebesgue measure. In the discrete case, the integrals over Ω in the following arguments have to be replaced by sums $\sum_{x\in\Omega}$, otherwise, the argument is the same.

Note that

$$\mathbb{E}_\vartheta[U_\vartheta] = \int_\Omega \Big(\frac{\partial}{\partial\vartheta}\log\rho(\vartheta,x)\Big)\rho(\vartheta,x)\,dx = \int_\Omega \frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)}{\rho(\vartheta,x)}\rho(\vartheta,x)\,dx = \int_\Omega \frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx = 0 \quad (1.1)$$

by assumption, i.e. the score function U_ϑ is centred under $\mathbb{P}_\vartheta$.

Then

$$\begin{aligned}\mathrm{Cov}_\vartheta[T,U_\vartheta] &= \mathbb{E}_\vartheta[TU_\vartheta]-\mathbb{E}_\vartheta[T]\mathbb{E}_\vartheta[U_\vartheta] = \mathbb{E}_\vartheta[TU_\vartheta]\\ &= \int_\Omega \Big(T(x)\frac{\partial}{\partial\vartheta}\log\rho(\vartheta,x)\Big)\rho(\vartheta,x)\,dx = \int_\Omega T(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx\\ &= \frac{\partial}{\partial\vartheta}\int_\Omega T(x)\rho(\vartheta,x)\,dx = \tau'(\vartheta)\end{aligned}$$

by assumption. The Cauchy-Schwarz inequality yields

$$\sqrt{\mathrm{Var}_\vartheta[T]\,\mathrm{Var}_\vartheta[U_\vartheta]} \geq |\,\mathrm{Cov}_\vartheta[T,U_\vartheta]| = \Big|\frac{d}{d\vartheta}\mathbb{E}_\vartheta[T]\Big|$$

and thus

$$\mathrm{Var}_\vartheta[T] \geq \frac{\Big|\frac{d}{d\vartheta}\mathbb{E}_\vartheta[T]\Big|^2}{\mathrm{Var}_\vartheta[U_\vartheta]} = \frac{(\tau'(\vartheta))^2}{I_\vartheta},$$

which is the first claim.

Furthermore, equality holds in the Cauchy-Schwarz inequality if and only if the two random variables are linearly dependent. Put $c_\vartheta := \tau'(\vartheta)/I_\vartheta$, then

$$\begin{aligned} 0 \le \operatorname{Var}_\vartheta[T - c_\vartheta U_\vartheta] &= \operatorname{Var}_\vartheta[T] - 2c_\vartheta \operatorname{Cov}_\vartheta[T, U_\vartheta] + c_\vartheta^2 \operatorname{Var}_\vartheta[U_\vartheta] \\ &= \operatorname{Var}_\vartheta[T] - 2\frac{\tau'(\vartheta)}{I_\vartheta}\tau'(\vartheta) + \frac{(\tau'(\vartheta))^2}{I_\vartheta^2} I_\vartheta = \operatorname{Var}_\vartheta[T] - \frac{(\tau'(\vartheta))^2}{I_\vartheta}. \end{aligned}$$

Thus, if T achives equality, we must have $\operatorname{Var}_\vartheta[T - c_\vartheta U_\vartheta] = 0$, which means that $T - c_\vartheta U_\vartheta$ is $\mathbb{P}_\vartheta$-a.s. equal to a constant. Since T is unbiased for $\tau(\vartheta)$ (and U_ϑ is centred under $\mathbb{P}_\vartheta$), that constant must be $\tau(\vartheta)$. This is the second claim. □

Remark 1.1.8. If the likelihood function is twice differentiable w.r.t. ϑ and $x \mapsto \frac{\partial^2}{\partial\vartheta^2} \log \rho(\vartheta, x)$ lies in $\mathcal{L}^1(\mathbb{P}_\vartheta)$ and is smooth enough so that the interchange

$$\int_\Omega \frac{\partial^2}{\partial\vartheta^2} \rho(\vartheta, x)\, dx = \frac{\partial^2}{\partial\vartheta^2} \int_\Omega \rho(\vartheta, x)\, dx \ \Big(= \frac{\partial^2}{\partial\vartheta^2} 1 = 0\Big)$$

is justified, then the Fisher information I_ϑ can be interpreted as the negative (expected) curvature of the log likelihood function (at the observed data):

$$-\mathbb{E}_\vartheta \left[\frac{\partial^2}{\partial\vartheta^2} \log \rho(\vartheta, X)\right] = \mathbb{E}_\vartheta \left[\Big(\frac{\partial}{\partial\vartheta} \log \rho(\vartheta, X)\Big)^2\right] = \mathbb{E}_\vartheta \left[U_\vartheta(X)^2\right] = \operatorname{Var}[U_\vartheta(X)] = I_\vartheta.$$

To verify this note that

$$\frac{\partial^2}{\partial\vartheta^2} \log \rho(\vartheta, x) = \frac{\frac{\partial^2}{\partial\vartheta^2}\rho(\vartheta, x)}{\rho(\vartheta, x)} - \Big(\frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta, x)}{\rho(\vartheta, x)}\Big)^2 = \frac{\frac{\partial^2}{\partial\vartheta^2}\rho(\vartheta, x)}{\rho(\vartheta, x)} - \Big(\frac{\partial}{\partial\vartheta} \log \rho(\vartheta, x)\Big)^2$$

and

$$\mathbb{E}_\vartheta \left[\frac{\frac{\partial^2}{\partial\vartheta^2}\rho(\vartheta, x)}{\rho(\vartheta, x)}\right] = \int_\Omega \frac{\frac{\partial^2}{\partial\vartheta^2}\rho(\vartheta, x)}{\rho(\vartheta, x)} \rho(\vartheta, x)\, dx = \int_\Omega \frac{\partial^2}{\partial\vartheta^2}\rho(\vartheta, x)\, dx = \frac{\partial^2}{\partial\vartheta^2} \int_\Omega \rho(\vartheta, x)\, dx = 0.$$

Definition 1.1.9 (Exponential family)**.** Let

$$\rho(\vartheta, x) = h(x) \cdot \exp\big(a(\vartheta) \cdot T(x) - b(\vartheta)\big) \tag{1.2}$$

for suitable functions $a, b : \Theta \to \mathbb{R}$ and $h : \Omega \to \mathbb{R}$. We assume that a is continuously differentiable with $a'(\vartheta) \neq 0$ for $\vartheta \in \Theta$ (in particular, a is injective and $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ for $\vartheta \neq \vartheta'$).

Then

$$U_\vartheta(x) = a'(\vartheta)T(x) - b'(\vartheta), \quad \text{hence} \quad \mathbb{E}_\vartheta[T] = \frac{b'(\vartheta)}{a'(\vartheta)} =: \tau(\vartheta)$$

(recall that $\mathbb{E}_\vartheta[U_\vartheta] = 0$ by (1.1)).

It can be shown that (see Proposition 1.4.1 in Section 1.4.1)

$$I_\vartheta = \operatorname{Var}_\vartheta[U_\vartheta] = a'(\vartheta) \cdot \tau'(\vartheta),$$

i.e.

$$T(x) = \frac{b'(\vartheta)}{a'(\vartheta)} + \frac{U_\vartheta(x)}{a'(\vartheta)} = \tau(\vartheta) + \frac{\tau'(\vartheta)U_\vartheta(x)}{I_\vartheta}$$

and T is thus (by Theorem 1.1.7) in this situation a variance-minimising unbiased estimator for τ.

Examples. 1. Binomial distributions: $\mathbb{P}_\vartheta = \mathrm{Bin}_{n,\vartheta}$, $\vartheta \in [0,1]$

$$\begin{aligned}\rho(\vartheta, x) &= \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x} \\ &= \binom{n}{x}\exp\Big(\underbrace{\frac{x}{n}}_{T(x)}\underbrace{n\log(\frac{\vartheta}{1-\vartheta})}_{=a(\vartheta)} + \underbrace{n\log(1-\vartheta)}_{=-b(\vartheta)}\Big), \quad x \in \mathbb{N}_0\end{aligned}$$

$T(x) = \dfrac{x}{n}$ is a variance-minimising unbiased estimator for $\tau(\vartheta) = \vartheta$.

2. Poisson distributions: $\mathbb{P}_\vartheta = \mathrm{Poi}_\vartheta$, $\vartheta \in (0,\infty)$

$$\rho(\vartheta, x) = e^{-\vartheta}\frac{\vartheta^x}{x!} = \underbrace{\frac{1}{x!}}_{=h(x)}\, e^{\overbrace{x}^{T(x)}\overbrace{\log\vartheta}^{=a(\vartheta)} - \overbrace{\vartheta}^{b(\vartheta)}}$$

We have $\tau(\vartheta) = \frac{1}{1/\vartheta} = \vartheta$, $T(x) = x$ is a variance-minimising unbiased estimator for ϑ, its variance is $\frac{(\tau'(\vartheta))^2}{a'(\vartheta)\tau'(\vartheta)} = \frac{1^2}{\frac{1}{\vartheta}\cdot 1} = \vartheta$.

3. Normal distributions with known variance: $\mathbb{P}_\vartheta = \mathcal{N}_{\vartheta,\sigma^2}$ with $\vartheta \in \Theta = \mathbb{R}$ and fixed $\sigma^2 > 0$,

$$\begin{aligned}\rho(\vartheta, x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}(x-\vartheta)^2\Big) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/(2\sigma^2)}}_{=h(x)}\cdot\exp\Big(-\underbrace{x}_{=T(x)}\cdot\underbrace{\frac{\vartheta}{\sigma^2}}_{=a(\vartheta)} - \underbrace{\frac{\vartheta^2}{2\sigma^2}}_{=b(\vartheta)}\Big),\end{aligned}$$

hence: $T(x) = x$ is a variance-minimising unbiased estimator for $\vartheta = \tau(\vartheta) = \frac{b'(\vartheta)}{a'(\vartheta)}$, its variance is $\sigma^2 = \frac{1}{I_\vartheta} = 1^2/(a'(\vartheta)\tau'(\vartheta))$.

Remark (Product model). The n-fold product model $\mathscr{M}^{\otimes n}$ of a regular model is again regular, its Fisher information satisfies $I_\vartheta^{(n)} = n \cdot I_\vartheta$.

If $\mathscr{M}$ is an exponential model w.r.t. the statistic T, then $\mathscr{M}^{\otimes n}$ is also an exponential model, and the underlying statistic is

$$T_n(x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^n T(x_i),$$

since

$$\rho^{\otimes n}(\vartheta,(x_1,\ldots,x_n)) = \prod_{i=1}^n \rho(\vartheta,x_i) = \Big(\prod_{i=1}^n h(x_i)\Big)\exp\left(na(\vartheta)\cdot\Big(\frac{1}{n}\cdot\sum_{i=1}^n T(x_i)\Big) - nb(\vartheta)\right).$$

1.2 Sufficiency, completeness and distribution-freeness

Notation convention. We consider $X : \Omega \to \Omega$, $X = \mathrm{Id}_\Omega$ as a random variable and interpret X as the "observed data".

Definition 1.2.1. A statistic T is called *sufficient* if the conditional distribution $\mathbb{P}_\vartheta(\,\cdot \mid T)$ of the observations does not depend on $\vartheta \in \Theta$.

Intuition: T already contains all information about ϑ that "is contained in the observed data."

Examples. 1. n-fold coin toss: $\Theta = [0,1]$, under $\mathbb{P}_\vartheta$ let $X = (X_1, \dots, X_n) \sim \mathrm{Ber}_\vartheta^{\otimes n}$, then $T := X_1 + \cdots + X_n$ is sufficient: Given $T = t \in \{0, 1, \dots, n\}$, X is uniformly distributed on

$$\big\{(x_1, \dots, x_n) \in \{0,1\}^n \,:\, x_1 + \cdots + x_n = t\big\}$$

2. Let $\Theta = \mathbb{N}$, $\Omega = \mathbb{Z}_+^2$, $\mathbb{P}_\vartheta = \mathrm{Unif}(\{0, 1, \dots, \vartheta\})^{\otimes 2}$ (considered as a probability measure on $\mathbb{Z}_+^2$) for $\vartheta \in \Theta$, i.e. under $\mathbb{P}_\vartheta$, X_1 and X_2 are independent and each uniformly distributed on $\{0, 1, \dots, \vartheta\}$. Here $T := X_1 + X_2$ is not sufficient: T takes values in $\{0, 1, \dots, 2\vartheta\}$ and

$$\mathbb{P}_\vartheta(T = t) = \frac{1 + \vartheta - |t - \vartheta|}{\vartheta^2} \quad \text{for } t = 0, 1, \dots, 2\vartheta$$

thus for $x_1, x_2 \in \{0, 1, \dots, \vartheta\}$ with $x_1 + x_2 = t$

$$\begin{aligned}\mathbb{P}_\vartheta\big((X_1, X_2) = (x_1, x_2) \,\big|\, T = t\big) &= \mathbb{1}_{\{0 \le x_1, x_2 \le \vartheta\}} \frac{1/\vartheta^2}{(1 + \vartheta - |t - \vartheta|)/\vartheta^2} \\ &= \mathbb{1}_{\{0 \le x_1, x_2 \le \vartheta\}} \frac{1}{1 + \vartheta - |t - \vartheta|}\end{aligned}$$

which depends on ϑ.

Remark. Obviously $X = \mathrm{Id}_\Omega$ is always sufficient, but this observation is generally of no use.

Theorem 1.2.2 (Fisher-Neyman factorisation theorem)**.** *Consider a standard statistical model. A statistic $T = t(X)$ with values in E is sufficient if and only if the density/weight function $\rho(\vartheta, x)$ has the form*

$$\rho(\vartheta, x) = h(x) g_\vartheta\big(t(x)\big), \quad x \in \Omega,\ \vartheta \in \Theta \tag{1.3}$$

(for a function $h : \Omega \to \mathbb{R}_+$ and functions $g_\vartheta : E \to \mathbb{R}_+$, $\vartheta \in \Theta$).

Proof of Theorem 1.2.2 in the discrete case. This is elementary:

If Ω (and thus X and w.l.o.g. also E) is discrete, then for $t \in E$ with $\mathbb{P}_\vartheta(T = t) > 0$

$$\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big) = \frac{\mathbb{P}_\vartheta(X = x, T = t)}{\mathbb{P}_\vartheta(T = t)} = \begin{cases} \dfrac{\mathbb{P}_\vartheta(X = x)}{\sum_{y\,:\,t(y) = t} \mathbb{P}_\vartheta(X = y)}, & \text{if } t(x) = t, \\ 0, & \text{otherwise.} \end{cases}$$

If $\rho(\vartheta, x)$ has the form (1.3), then

$$\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big) = \begin{cases} \dfrac{h(x) g_\vartheta(t)}{\sum_{y\,:\,t(y) = t} h(y) g_\vartheta(t)} = \dfrac{h(x)}{\sum_{y\,:\,t(y) = t} h(y)}, & \text{if } t(x) = t, \\ 0, & \text{otherwise,} \end{cases}$$

which clearly does not depend on ϑ.

Conversely, if $\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big)$ does not depend on ϑ, we can write

$$\mathbb{P}_\vartheta(X = x) = \underbrace{\mathbb{P}_\vartheta\big(T = t(x)\big)}_{=:\, g_\vartheta\left(t(x)\right)} \cdot \underbrace{\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t(x)\big)}_{=:\, h(x)}$$

□

In the continuous case we need the following lemma.

Lemma 1.2.3. *Let $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a continuous standard statistical model with $\mathbb{P}_\vartheta(dx) = \rho(\vartheta, x)dx$, $t : \Omega \to E$ measurable, $T = t \circ X$ a statistic.*

1. *(joint domination) There exist $c_1, c_2, \cdots \in [0, 1]$ with $\sum_{i=1}^\infty c_i = 1$ and $\vartheta_1, \vartheta_2, \cdots \in \Theta$, such that*

$$\mathbb{P}_\vartheta \ll \nu^* := \sum_{i=1}^\infty c_i \mathbb{P}_{\vartheta_i} \quad \textit{for all } \vartheta \in \Theta. \tag{1.4}$$

2. *If T is sufficient, then for $\vartheta \in \Theta$*

$$\nu^*(A \mid T) = \mathbb{P}_\vartheta(A \mid T) \;\; \nu^*\textit{-almost surely}$$

3. *If $\rho(\vartheta, x) = h(x)g_\vartheta(t(x))$ for $\mathbb{P}_\vartheta$-almost all x holds for all $\vartheta \in \Theta$, then*

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) = \frac{g_\vartheta(t(x))}{\sum_{i=1}^\infty c_i g_{\vartheta_i}(t(x))} \tag{1.5}$$

   is a version of the density that depends only on $t(x)$.

4. *If for each $\vartheta \in \Theta$ the density $\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x)$ depends only on $t(x)$, then T is sufficient.*

Proof. 1. Let $\Omega \subset \mathbb{R}^n$ (measurable), λ = Lebesgue measure on $\mathbb{R}^n$. Since $\mathscr{L}^1(\Omega, \lambda)$ is separable (note: $\mathcal{B}(\Omega)$ is countably generated), $\{\rho(\vartheta, \cdot) : \vartheta \in \Theta\} \subset \mathscr{L}^1(\Omega, \lambda)$ contains a countable, (with respect to the $\mathscr{L}^1$-norm) dense subset $\{\rho(\vartheta_n, \cdot) : n \in \mathbb{N}\}$, i.e. for each $\vartheta \in \Theta$ there exists a sequence $(n_j)_j \subset \mathbb{N}$ with

$$\lim_{j\to\infty} \int_\Omega \left|\rho(\vartheta, x) - \rho(\vartheta_{n_j}, x)\right| \, \lambda(dx) = 0 \tag{1.6}$$

We set

$$\nu^* := \sum_{n=1}^\infty 2^{-n} \mathbb{P}_{\vartheta_n}$$

and check that this has the desired property: Let $A \in \mathcal{B}(\Omega)$ with $\nu^*(A) = 0$ and $\vartheta \in \Theta$ be given, then with the sequence $(n_j)_j$ from (1.6) also

$$\mathbb{P}_\vartheta(A) = \int_A \rho(\vartheta, x) \, \lambda(dx) = \lim_{j\to\infty} \int_A \rho(\vartheta_{n_j}, x) \, \lambda(dx) = \lim_{j\to\infty} 0 = 0$$

i.e. $\mathbb{P}_\vartheta \ll \nu^*$.

2. By assumption there exists a kernel $\kappa : E \times \mathcal{B}(\Omega) \to [0, 1]$, such that for each $\vartheta \in \Theta$ and $A \in \mathcal{B}(\Omega)$ we have

$$\mathbb{P}_\vartheta(A \mid T = t) = \kappa(t, A) \quad \mathbb{P}_\vartheta\text{-almost surely.}$$

Thus

$$\nu^*(A \mid T = t) = \sum_{j=1}^\infty 2^{-j} \mathbb{P}_{\vartheta_j}(A \mid T = t) = \kappa(t, A) \quad \nu^*\text{-a.s. and } \mathbb{P}_\vartheta\text{-a.s..}$$

3. By assumption we have

$$\frac{d\nu^*}{d\lambda}(x) = \sum_{j=1}^{\infty} 2^{-j} \frac{d\mathbb{P}_{\vartheta_j}}{d\lambda}(x) = \sum_{j=1}^{\infty} 2^{-j} \rho(\vartheta_j, x) = \sum_{j=1}^{\infty} 2^{-j} h(x) g_{\vartheta_j}\big(t(x)\big)$$

and thus

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) = \frac{d\mathbb{P}_\vartheta}{d\lambda}\Big(\frac{d\nu^*}{d\lambda}\Big)^{-1}(x) = \frac{g_\vartheta(t(x))}{\sum_{i=1}^{\infty} c_i g_{\vartheta_i}(t(x))}$$

as required.

4. We show for $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\mathbb{1}_B\big] = \mathbb{E}_\vartheta\big[\nu^*(A \mid T)\mathbb{1}_B\big] \quad \text{for } A \in \mathcal{B}(\Omega),\ B \in \sigma(T)$$

Then $\nu^*(\cdot \mid T)$, which does not depend on ϑ, is a version of $\mathbb{P}_\vartheta(\cdot \mid T)$ and thus T is sufficient. Indeed

$$\begin{aligned}\mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\mathbb{1}_B\big] &= \mathbb{E}_\vartheta\big[\mathbb{1}_A\mathbb{1}_B\big] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbb{1}_A\mathbb{1}_B\Big] = \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbb{1}_A\mathbb{1}_B\,\Big|\,T\Big]\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbb{1}_B\mathbb{P}_{\nu^*}(A \mid T)\Big] = \mathbb{E}_\vartheta\big[\mathbb{P}_{\nu^*}(A \mid T)\mathbb{1}_B\big]\end{aligned}$$

(we use in the first equation of the second line the fact that $\frac{d\mathbb{P}_\vartheta}{d\nu^*}$ is measurable with respect to $\sigma(T)$). □

Proof of Theorem 1.2.2, continuous case. If $\rho(\vartheta, x)$ has the form (1.3), then by Lemma 1.2.3, 3. and 4., T is sufficient.

Conversely, let T be sufficient, ν^* as in Lemma 1.2.3. By Lemma 1.2.3, 2. we have for $\vartheta \in \Theta$

$$\mathbb{P}_\vartheta(\cdot \mid T) = \nu^*(\cdot \mid T)$$

For any $A \in \mathcal{B}(\Omega)$ we have

$$\begin{aligned}\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbb{1}_A\Big] &= \mathbb{P}_\vartheta(A) = \mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\big] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\nu^*(A \mid T)\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\nu^*(A \mid T)\,\Big|\,T\Big]\Big] = \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,T\Big]\nu^*(A \mid T)\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,T\Big]\mathbb{1}_A\Big]\end{aligned}$$

and therefore with $T = t(X)$

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(X) = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,t(X)\Big] =: g_\vartheta\big(t(X)\big).$$

Set $h(x) := \frac{d\nu^*}{d\lambda}(x)$, then

$$\frac{d\mathbb{P}_\vartheta}{d\lambda}(x) = \Big(\frac{d\mathbb{P}_\vartheta}{d\nu^*}\frac{d\nu^*}{d\lambda}\Big)(x) = g_\vartheta\big(t(x)\big)h(x).$$

□

Examples. 1. Normal model: $\Theta = \mathbb{R} \times (0,\infty) \ni \vartheta = (\mu, \sigma^2)$, $\Omega = \mathbb{R}^n$, $\mathbb{P}_{(\mu,\sigma^2)} = \mathcal{N}(\mu,\sigma^2)^{\otimes n}$. We have

$$\begin{aligned}\rho\big((\mu,\sigma^2),(x_1,\dots,x_n)\big) &= \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-(x_j-\mu)^2/(2\sigma^2)\big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j-\mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum_{j=1}^n x_j - \frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2\Big)\end{aligned}$$

Therefore $\big(\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2\big)$ is (minimal-)sufficient and (since it is a deterministic reparametrisation) so is

$$\big(\overline{X}, S^2\big) \coloneqq \left(\frac{1}{n}\sum_{j=1}^n X_j\,,\ \frac{1}{n-1}\sum_{j=1}^n (X_j-\overline{X})^2\right)$$

2. Uniform model: $\Theta = (0,\infty) \ni \vartheta$, $\Omega = [0,\infty)^n$, $\mathbb{P}_\vartheta = \mathrm{Unif}([0,\vartheta])^{\otimes n}$, so $T \coloneqq \max\{X_1,\dots,X_n\}$ is sufficient: We have

$$\rho\big(\vartheta,(x_1,\dots,x_n)\big) = \prod_{j=1}^n \big(\mathbb{1}_{0\le x_j\le\vartheta}/\vartheta\big) = \vartheta^{-n}\mathbb{1}_{0\le\max\{x_j:j\le n\}\le\vartheta}.$$

Definition 1.2.4. A sufficient statistic T is called *minimally sufficient* if for every sufficient statistic U there exists a function φ such that for all $\vartheta \in \Theta$

$$T = \varphi(U) \quad \mathbb{P}_\vartheta\text{-almost surely.}$$

Theorem 1.2.5 (Criterion for minimal sufficiency)**.** *Given a standard statistical model and $t : \Omega \to E$ measurable, such that*

$$t(x) = t(y) \quad \Longleftrightarrow \quad \text{there exists } 0 < \ell(x,y) < \infty \text{ with } \rho(\vartheta,y) = \ell(x,y)\rho(\vartheta,x) \text{ for } \vartheta\in\Theta. \quad (1.7)$$

Then $T = t(X)$ is minimally sufficient.

Remark. We see from Theorem 1.2.5 that in the above examples $(\overline{X}, S^2)$ in the normal product model and $\max\{X_1,\dots,X_n\}$ in the $\mathrm{Unif}([0,\vartheta])^{\otimes n}$ model are indeed minimally sufficient.

Proof of Theorem 1.2.5. Let ν^* be as in Lemma 1.2.3, $x, y \in \Omega$ with $t(x) = t(y)$, then

$$\begin{aligned}\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) &= \frac{\frac{d\mathbb{P}_\vartheta}{d\lambda}(x)}{\frac{d\nu^*}{d\lambda}(x)} = \frac{\rho(\vartheta,x)}{\sum_{i=1}^\infty c_i\rho(\vartheta_i,x)} = \frac{\rho(\vartheta,x)\ell(x,y)}{\sum_{i=1}^\infty c_i\rho(\vartheta_i,x)\ell(x,y)} \\ &= \frac{\rho(\vartheta,y)}{\sum_{i=1}^\infty c_i\rho(\vartheta_i,y)} = \frac{d\mathbb{P}_\vartheta}{d\nu^*}(y)\end{aligned}$$

i.e. $\frac{d\mathbb{P}_\vartheta}{d\nu^*}$ depends only on $t(x)$ and by Lemma 1.2.3, item 4, $T = t(X)$ is sufficient.

Now let $U = u(X)$ be another sufficient statistic. By Theorem 1.2.2, $\rho(\vartheta,x) = \widetilde{h}(x)\widetilde{g}_\vartheta(u(x))$ for certain functions $\widetilde{h}(\cdot)$ and $\widetilde{g}_\vartheta(\cdot)$ (and w.l.o.g. $\widetilde{h} > 0$, otherwise restrict Ω to $\{\widetilde{h} > 0\}$).

Let $x, y \in \Omega$ with $u(x) = u(y)$ be given, then

$$\frac{\rho(\vartheta,y)}{\rho(\vartheta,x)} = \frac{\widetilde{h}(y)\widetilde{g}_\vartheta(u(y))}{\widetilde{h}(x)\widetilde{g}_\vartheta(u(x))} = \frac{\widetilde{h}(y)}{\widetilde{h}(x)}$$

(which does not depend on ϑ), now (1.7) implies $t(x) = t(y)$.

Thus

$$u(x) = u(y) \quad \Longrightarrow \quad t(x) = t(y),$$

i.e. there exists a function f with $t(x) = f(u(x))$. Since this argument applies to every sufficient U, T is minimally sufficient. □

Definition 1.2.6. A statistic T with values in E is called *complete* if for all (measurable) functions $g : E \to \mathbb{R}$ we have

$$\forall\, \vartheta \in \Theta \,:\, \mathbb{E}_\vartheta\big[g(T)\big] = 0 \qquad \Longrightarrow \qquad \forall\, \vartheta \in \Theta \,:\, g(T) = 0 \;\; \mathbb{P}_\vartheta\text{-a.s.} \tag{1.8}$$

T is called *boundedly complete* if this is only required for bounded test functions g.

Examples. 1. $X \sim \mathrm{Poi}(\vartheta)$, $\vartheta \in \Theta = (0, \infty)$ is boundedly complete: Let $g : \mathbb{N}_0 \to \mathbb{R}$ be bounded, then the function

$$\phi : \Theta \ni \vartheta \mapsto \mathbb{E}_\vartheta\big[g(X)\big] = \sum_{x=0}^{\infty} e^{-\vartheta} \frac{\vartheta^x}{x!} g(x)$$

(as a power series with convergence radius ∞) is in particular analytic. If now $\phi(\vartheta) = 0$ for all $\vartheta > 0$ holds, then $\phi \equiv 0$ must hold, i.e. $g(x) = 0$ for all $x \in \mathbb{N}_0$.

2. In a product model (with a compact range, say) the observation $X = (X_1, \ldots, X_n)$ itself is not (boundedly) complete: Obviously $\mathbb{E}_\vartheta[X_1 - X_2] = 0$ for each ϑ, but the function $(x_1, \ldots, x_n) \mapsto x_1 - x_2$ is not constant.

3. (From [LR06, Problem 4.12 (p. 141)]) Consider $\Theta = [0, 1]$, $\Omega = \{-1, 0\} \cup \mathbb{N}$,

$$\mathbb{P}_\vartheta(x) = \begin{cases} \vartheta, & x = -1 \\ (1-\vartheta)^2 \vartheta^x, & x = 0, 1, 2, \ldots \end{cases}$$

(i.e. under $\mathbb{P}_\vartheta$, $X = -1$ with probability ϑ and with complementary probability $1 - \vartheta$, $X \sim \mathrm{Geom}(1 - \vartheta)$). Then $X \in \mathscr{L}^1(\mathbb{P}_\vartheta)$ and

$$\mathbb{E}_\vartheta[X] = -\vartheta + (1-\vartheta)\Big(\frac{1}{1-\vartheta} - 1\Big) = -\vartheta + \vartheta = 0$$

for each $\vartheta \in \Theta$.

Furthermore, for bounded $g : \Omega \to \mathbb{R}$,

$$\left|\sum_{x=0}^{\infty} \vartheta^x g(x)\right| \le \|g\|_\infty \sum_{x=0}^{\infty} \vartheta^x = \|g\|_\infty / (1-\vartheta)$$

and $\mathbb{E}_\vartheta[g(X)] = 0$ implies

$$|g(-1)| = \vartheta^{-1}(1-\vartheta)^2 \left|\sum_{x=0}^{\infty} \vartheta^x g(x)\right| \le (1-\vartheta)\|g\|_\infty / \vartheta,$$

with $\vartheta \uparrow 1$ it follows $g(-1) = 0$ and thus $\sum_{x=0}^{\infty} \vartheta^x g(x) \equiv 0$, which forces $g(x) = 0$ for $x \in \mathbb{N}_0$. (Because $\phi(\vartheta) \coloneqq \sum_{x=0}^{\infty} \vartheta^x g(x)$ is analytic in $(-1, 1)$ and $\equiv 0$ in $(0, 1)$, thus $\phi(\cdot) \equiv 0$.)

Here, X is boundedly complete, but not complete.

Theorem 1.2.7 (Bahadur's theorem)**.** *A statistic $T = t(X)$ with values in $\mathbb{R}^k$ for some k, which is boundedly complete and sufficient, is minimally sufficient.*

Proof. We write $t(x) = \big(t_1(x), \dots, t_k(x)\big)$. Let $u : \Omega \to E$ be measurable and $U = u(X)$ be (another) sufficient statistic; we want to show: $T = f(U)$ for a function $f : E \to \mathbb{R}^k$.

$\varphi : \mathbb{R} \ni z \mapsto 1/(1+e^z) \in (0,1)$ is bijective (and bi-measurable), for $t = (t_1, \dots, t_k) \in \mathbb{R}^k$ let $s = s(t) = (s_1, \dots, s_k) \in (0,1)^k$ with $s_i = \varphi(t_i)$.

Set $S \coloneqq s(T) = (\varphi(T_1), \dots, \varphi(T_k))$ and for $i \in \{1, \dots, k\}$

$$H_i(U) \coloneqq \mathbb{E}_\vartheta[S_i \mid U] \quad \Big(= \mathbb{E}_{\nu^*}[S_i \mid U] \Big)$$
$$J_i(T) \coloneqq \mathbb{E}_\vartheta[H_i(U) \mid T] \quad \Big(= \mathbb{E}_{\nu^*}[H_i(U) \mid T] \Big)$$

(with ν^* as in Lemma 1.2.3, the conditional expectations do not depend on ϑ), since S is bounded, this also holds for $H_i(U)$ and $J_i(T)$.

Furthermore for $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta\big[\varphi(T_i) - J_i(T)\big] = \mathbb{E}_\vartheta\Big[\varphi(T_i) - \mathbb{E}_\vartheta\big[\mathbb{E}_\vartheta[\varphi(T_i) \mid U] \,\big|\, T\big]\Big] = \mathbb{E}_\vartheta\big[\varphi(T_i)\big] - \mathbb{E}_\vartheta\big[\varphi(T_i)\big] = 0,$$

i.e. (since T is boundedly complete) for each $\vartheta \in \Theta$

$$J_i(T) = \varphi(T_i) \quad \mathbb{P}_\vartheta\text{-a.s.} \tag{1.9}$$

Decompose the variances (see Proposition A.1.6 in Appendix A.1):

$$\begin{aligned}
\operatorname{Var}_\vartheta\big[J_i(T)\big] &= \mathbb{E}_\vartheta\Big[\operatorname{Var}_\vartheta\big[J_i(T) \,\big|\, U\big]\Big] + \operatorname{Var}_\vartheta\big[\mathbb{E}_\vartheta[J_i(T) \mid U]\big] \\
&= \mathbb{E}_\vartheta\Big[\operatorname{Var}_\vartheta\big[J_i(T) \,\big|\, U\big]\Big] + \operatorname{Var}_\vartheta\big[H_i(U)\big] \\
\operatorname{Var}_\vartheta\big[H_i(U)\big] &= \mathbb{E}_\vartheta\Big[\operatorname{Var}_\vartheta\big[H_i(U) \,\big|\, T\big]\Big] + \operatorname{Var}_\vartheta\big[J_i(T)\big]
\end{aligned}$$

(because $\mathbb{E}_\vartheta[J_i(T) \mid U] = \mathbb{E}_\vartheta[\varphi(T_i) \mid U] = H_i(U)$ by (1.9) and definition of $H_i(U)$).

Since $\operatorname{Var}_\vartheta\big[J_i(T)\big] \le \operatorname{Var}_\vartheta\big[H_i(U)\big]$ holds (by general principles: conditioning cannot increase the variance), it follows

$$\operatorname{Var}_\vartheta\big[J_i(T) \,\big|\, U\big] = 0 \;\; \mathbb{P}_\vartheta\text{-a.s.} \quad \text{and thus also} \quad \operatorname{Var}_\vartheta\big[H_i(U) \,\big|\, T\big] = 0 \;\; \mathbb{P}_\vartheta\text{-a.s.}$$

Thus

$$\varphi(T_i) = \mathbb{E}_\vartheta[\varphi(T_i) \mid U] = H_i(U) \;\; \mathbb{P}_\vartheta\text{-a.s.}$$

and $T = (T_1, \dots, T_k)$ with $T_i = \varphi^{-1}(H_i(U))$ is (ν^*-a.s.) a function of U. □

Theorem 1.2.8 (Rao-Blackwell theorem)**.** *Let S be an estimator for a parameter characteristic $\tau(\vartheta)$ with $\mathbb{E}_\vartheta[|S|] < \infty$ for all $\vartheta \in \Theta$ and let T be a sufficient statistic. Then the "Rao-Blackwellisation"*

$$S^* := \mathbb{E}\big[S \,\big|\, T\big] \tag{1.10}$$

of S (note that the above conditional expectation does not depend on ϑ since T is sufficient) satisfies

$$\mathbb{E}_\vartheta[S^*] = \mathbb{E}_\vartheta[S] \quad \textit{for all } \vartheta \in \Theta \quad \textit{and} \tag{1.11}$$
$$\mathbb{E}_\vartheta\big[(S^* - \tau(\vartheta))^2\big] \le \mathbb{E}_\vartheta\big[(S - \tau(\vartheta))^2\big] \quad \textit{for all } \vartheta \in \Theta. \tag{1.12}$$

If $\operatorname{Var}_\vartheta[S] < \infty$ for all ϑ, then the inequality is strict whenever $S \neq S^*$.

Proof. (1.11) follows from the tower property of conditional expectation.

For (1.12): Generally,

$$\mathbb{E}_\vartheta\left[(S-\tau(\vartheta))^2\right] = \mathbb{E}_\vartheta\left[\big(S-\mathbb{E}_\vartheta[S]\big)^2\right] + \big(\mathbb{E}_\vartheta[S]-\tau(\vartheta)\big)^2$$

i.e. mean squared error = variance + (bias)^2, because

$$\begin{aligned}\mathbb{E}_\vartheta\left[(S-\tau(\vartheta))^2\right] &= \mathbb{E}_\vartheta\left[S^2\right] - 2\tau(\vartheta)\mathbb{E}_\vartheta\left[S\right] + \tau(\vartheta)^2\\ &= \mathbb{E}_\vartheta\left[S^2\right] - \big(\mathbb{E}_\vartheta\left[S\right]\big)^2 + \big(\mathbb{E}_\vartheta\left[S\right]\big)^2 - 2\tau(\vartheta)\mathbb{E}_\vartheta\left[S\right] + \tau(\vartheta)^2\\ &= \mathrm{Var}_\vartheta\left[S\right] + \big(\mathbb{E}_\vartheta[S]-\tau(\vartheta)\big)^2\end{aligned}$$

Thus (1.12) follows from (1.11) with conditional variance decomposition

$$\mathrm{Var}_\vartheta[S] = \mathbb{E}_\vartheta\big[\,\mathrm{Var}_\vartheta[S\mid T]\big] + \mathrm{Var}_\vartheta\left[\mathbb{E}_\vartheta[S\mid T]\right] \geq \mathrm{Var}_\vartheta\left[\mathbb{E}_\vartheta[S\mid T]\right] = \mathrm{Var}_\vartheta[S^*]$$

□

Corollary 1.2.9 (Lehmann-Scheffé theorem)**.** *T a sufficient and complete statistic and S an unbiased estimator for $\tau(\vartheta)$. Then S^* from Theorem 1.2.8 is variance-minimising.*

Proof. By construction $S^* = f_1(T)$ for some function f_1. Let $\widetilde{S}$ be another unbiased estimator for $\tau(\vartheta)$, $\widetilde{S}^* = \mathbb{E}_\vartheta[\widetilde{S}\mid T] =: f_2(T)$ its "Rao-Blackwellisation". Then (since both are unbiased)

$$\mathbb{E}_\vartheta[f_1(T)-f_2(T)] = \tau(\vartheta)-\tau(\vartheta) = 0$$

and completeness of T implies that $g(t) \coloneqq f_1(t)-f_2(t)$ is $\equiv 0$. □

Example. $X_1,\dots,X_n$ i.i.d. $\sim \mathrm{Unif}([0,\vartheta])$, $S \coloneqq \frac{2}{n}(X_1+\cdots+X_n)$, $T=\max\{X_1,\dots,X_n\}$; we have

$$\mathbb{E}_\vartheta\big[X_i\,\big|\,T\big] = \frac{1}{n}T + \frac{n-1}{n}\frac{1}{2}T$$

and thus

$$S^* = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\vartheta\big[X_i\,\big|\,T\big] = \frac{2}{n}T + \frac{n-1}{n}T = \frac{n+1}{n}T$$

is in this model an unbiased estimator for ϑ with minimal possible variance.

Definition 1.2.10. A statistic U is called *distribution-free* if $\mathscr{L}_\vartheta(U)$ does not depend on ϑ. U is called *maximally distribution-free* if for every other distribution-free statistic V there exists a function g with $V=g(U)$.

Examples.

1. In the normal model, $\big(X_1-\overline{X},\dots,X_n-\overline{X}\big)/\sqrt{S^2}$ is distribution-free.
2. For $(X_1,\dots,X_n)\sim \mathrm{Unif}([0,\vartheta])^{\otimes n}$ with $T:=\max\{X_1,\dots,X_n\}$, $\big(X_1/T,\dots,X_n/T\big)$ is distribution-free.

Theorem 1.2.11 (Basu's theorem)**.** *Let $(\Omega,\mathscr{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical model (we assume that $\mathbb{P}_\vartheta \ll \nu^*$ for some ν^*, cf. Theorem 1.2.2 and Lemma 1.2.3; and we again write $X=\mathrm{Id}_\Omega$ for the observations), $T=t(X)$ with $t:\Omega\to E$ measurable and $U=u(X)$ with $u:\Omega\to E'$ measurable statistics.*

1. If T is (boundedly) complete and sufficient and U is distribution-free, then T and U are independent under each $\mathbb{P}_\vartheta$.

2. If T is sufficient and T and U are independent under each $\mathbb{P}_\vartheta$, then U is distribution-free.

3. Let T and U be independent under each $\mathbb{P}_\vartheta$ and U distribution-free, suppose $\sigma(T,U)=\mathscr{F}$. Then T is sufficient.

Proof. 1. Let $A \in \mathcal{B}(E')$, $\vartheta \in \Theta$, then

$$\mathbb{P}_\vartheta(U \in A) = \mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(U \in A \mid T)\big].$$

Since U is distribution-free, $\mathbb{P}_\vartheta(U \in A)$ does not depend on ϑ, and since T is sufficient, $\mathbb{P}_\vartheta(U \in A \mid T)$ also does not depend on ϑ. Thus

$$g : E \ni t \mapsto \mathbb{P}_\vartheta(U \in A \mid T = t) - \mathbb{P}_\vartheta(U \in A) \in [-1, 1]$$

is bounded with $\mathbb{E}_\vartheta[g(T)] = 0$ for all ϑ. Completeness of T forces $g(t) \equiv 0$ ($\mathbb{P}_\vartheta$-a.s. for each ϑ), i.e.

$$\mathbb{P}_\vartheta(U \in A) = P_\vartheta(U \in A \mid T) \quad \mathbb{P}_\vartheta\text{-a.s.}$$

This means precisely that T and U are independent under $\mathbb{P}_\vartheta$.

2. For $A \in \mathscr{F}$, $\mathbb{P}_\vartheta(U \in A \mid T)$ does not depend on ϑ, i.e. there exists $\nu(A) \in [0, 1]$, such that for each $\vartheta \in \Theta$

$$\mathbb{P}_\vartheta(U \in A \mid T) = \nu(A) \quad \mathbb{P}_\vartheta\text{-a.s.}$$

Independence of U and T under $\mathbb{P}_\vartheta$ implies

$$\mathbb{P}_\vartheta(U \in A \mid T) = \mathbb{P}_\vartheta(U \in A) \quad (\mathbb{P}_\vartheta\text{-a.s.})$$

and thus $\mathbb{P}_\vartheta(U \in A) = \nu(A)$; since this does not depend on ϑ, U is distribution-free.

3. We have to show that for $A \in \mathscr{F}$

$$\mathbb{P}_\vartheta(A \mid T) \quad \text{does not depend on } \vartheta.$$

It suffices to verify this for A of the form $A = \{T \in B\} \cap \{U \in B'\}$ (because such events form a $\cap$-stable generator of $\sigma(T, U)$ and by assumption $\mathscr{F} = \sigma(T, U)$).

Now

$$\mathbb{P}_\vartheta\big(\{T \in B\} \cap \{U \in B'\} \,\big|\, T\big) = \mathbb{E}_\vartheta\big[\mathbb{1}_{\{T \in B\}} \mathbb{1}_{\{U \in B'\}} \,\big|\, T\big] = \mathbb{1}_{\{T \in B\}} \mathbb{P}_\vartheta(U \in B')$$

(where we use for the 2nd equality that T and U are independent). Since U is distribution-free, this does not depend on ϑ, i.e. T is sufficient. □

1.3 Confidence intervals (and confidence regions)

Definition 1.3.1. Let $\mathscr{M} = (\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be a statistical model, $\tau(\vartheta)$ a real-valued parameter functional, L, R statistics with $L \leq R$, $\alpha \in (0, 1)$. The (random) interval $I := [L, R]$ is called a *confidence interval* for τ at (confidence) level $1 - \alpha$ (or error level α), if

$$\forall\, \vartheta \in \Theta \;:\quad \mathbb{P}_\vartheta\big(\tau(\vartheta) \in I\big) \geq 1 - \alpha.$$

Note: I is random, but ϑ is not (at least under our present (so-called frequentist) interpretation).

More generally, a set $C(x) \subset \Theta$ constructed as a function of the observations $x \in \Omega$ is called a *confidence region* for τ at (confidence) level $1 - \alpha$, if

$$\forall\, \vartheta \in \Theta \;:\quad \mathbb{P}_\vartheta\big(\{x \in \Omega : C(x) \ni \tau(\vartheta)\}\big) \geq 1 - \alpha.$$

Obviously, one generally wishes to choose I as short as possible (as far as compatible with the required level).

Example 1.3.2 (Confidence interval for the mean in the normal model with known variance)**.** Under $\mathbb{P}_\vartheta$, let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}_{\vartheta,\sigma^2}$ with $\vartheta \in \Theta := \mathbb{R}$, $\sigma^2 > 0$ known (and fixed). Let $q := \Phi^{-1}(1-\frac{\alpha}{2})$ denote the $(1-\alpha/2)$-quantile of the standard normal distribution.

$$I := \big[\overline{X} - q \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + q \cdot \frac{\sigma}{\sqrt{n}}\big] \quad \text{with} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

is a confidence interval for ϑ at (confidence) level $1-\alpha$, since under $\mathbb{P}_\vartheta$, $\overline{X} \sim \mathcal{N}_{\vartheta,\sigma^2/n}$,

$$\begin{aligned}\mathbb{P}_\vartheta\big(\overline{X} - q \cdot \frac{\sigma}{\sqrt{n}} \le \vartheta \le \overline{X} + q \cdot \frac{\sigma}{\sqrt{n}}\big) &= \mathbb{P}_\vartheta\big(q \ge \frac{\overline{X}-\vartheta}{\sigma/\sqrt{n}} \ge -q\big) \\ &= \mathbb{P}(-q \le Z \le q) = \mathbb{P}(Z \le q) - \mathbb{P}(Z \ge -q) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha\end{aligned}$$

(with $Z \sim \mathcal{N}(0,1)$).

Example 1.3.3 (Student confidence interval for the mean in the normal model)**.** Under $\mathbb{P}_\vartheta$, $\vartheta = (\mu, \sigma^2) \in \Theta := \mathbb{R} \times (0,\infty)$, let

$$X_1, X_2, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}_{\mu,\sigma^2}.$$

Let $\alpha \in (0,1)$, $q = q_{n-1,1-\alpha/2}$ denote the $1-\frac{\alpha}{2}$-quantile of the Student distribution with $n-1$ degrees of freedom,

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Then

$$I := \Big[\overline{X} - q\sqrt{\frac{S^2}{n}}, \overline{X} + q\sqrt{\frac{S^2}{n}}\,\Big]$$

is a confidence interval for μ at error level α.

Proof. $T := \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S^2}}$ follows a Student distribution with $n-1$ degrees of freedom (for any choice of μ and σ^2, see e.g. Proposition A.2.5 in Appendix A.2), hence

$$\begin{aligned}&\mathbb{P}_{(\mu,\sigma^2)}\Big(\overline{X} - q\sqrt{\tfrac{S^2}{n}} \le \mu \le \overline{X} + q\sqrt{\tfrac{S^2}{n}}\Big) \\ &= \mathbb{P}(-q \le T \le q) = \mathbb{P}(T \le q) - \mathbb{P}(T \le -q) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

□

Example. Two sleeping pills are to be compared; 10 patients received medication A and B on consecutive nights. The data[1] (x_i = number of hours of sleep with drug A minus number of hours of sleep with drug B for patient i):

i	1	2	3	4	5	6	7	8	9	10
x_i	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

We have

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i \approxeq 1.58, \quad s = \Big(\frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2\Big)^{1/2} \approxeq 1.23.$$

[1] these are the data from Sect. IX, Illustration 1 of the original paper by Student (=W.S. Gossett), The Probable Error of a Mean, Biometrika, Vol. 6, No. 1 (Mar., 1908), pp. 1-25

Assume the data arise from a normal distribution with unknown mean μ and unknown variance σ^2 (and the results from different patients are independent). We have $q_{9,\,0.995} \approxeq 3.25$ (from a quantile table or computer program, e.g.,R), thus

$$\big[\overline{x} \pm q\frac{s}{\sqrt{n}}\big] \approxeq [0.31,\ 2.85]$$

is a confidence interval for μ (the average additional hours of sleep provided by drug A compared to drug B) at confidence level $0.99 = 1 - 0.01$.

(Note: (Meaninglessly) precise values computed to machine precision are $\overline{x} - q\frac{s}{\sqrt{n}} \approxeq 0.3159481$, $\overline{x} + q\frac{s}{\sqrt{n}} \approxeq 2.8440519$; however, one should always round the bounds of a confidence interval "conservatively", i.e., outward.)

Example 1.3.4 (Approximate confidence interval in the binomial model via normal approximation)**.** $X \sim \mathrm{Bin}_{n,\vartheta}$, $\theta \in \Theta = [0,1]$, $\widehat{\vartheta} := \frac{X}{n}$, $\widehat{\sigma} := \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, $\alpha \in (0,1)$, $q := \Phi^{-1}(1-\frac{\alpha}{2})$, then

$$I := \big[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\big]$$

is an (approximate) confidence interval for ϑ at confidence level $1-\alpha$, since under $\mathbb{P}_\vartheta$, as $n \to \infty$,

$$\widehat{\vartheta} \xrightarrow{d} \vartheta, \quad \widehat{\sigma} \xrightarrow{d} \sqrt{\vartheta(1-\vartheta)}, \quad \text{and } \frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} = \frac{X - n\vartheta}{\widehat{\sigma}\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

(by the de Moivre-Laplace theorem) and thus

$$\mathbb{P}_\vartheta\Big(\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}} \le \vartheta \le \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\Big) = \mathbb{P}_\vartheta\Big(-q \le \frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} \le q\Big) \approx \mathbb{P}(-q \le Z \le q) = 1 - \alpha \qquad (1.13)$$

Occasionally one hears the "rule of thumb" that $np(1-p) \ge 9$ should hold for the approximation in (1.13) to be useful.

1.3.1 A confidence interval for the median

Example 1.3.5 (A confidence interval for the median)**.** Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with (unknown) distribution Q, having a continuous distribution function (i.e., Q has no atoms). (In the formal terms of Definition 1.1.1: $\Theta = \{\vartheta : \vartheta \text{ non-atomic probability measure on } \mathbb{R}\}$, $\Omega = \mathbb{R}^n$, $\mathscr{F} = \mathcal{B}(\mathbb{R}^n)$, $\mathbb{P}_\vartheta = \vartheta^{\otimes n}$.)

Let $m(Q)$ denote "the" median of the probability law Q (i.e., $Q\big((-\infty, m(Q)]\big) = \frac{1}{2} = Q\big([m(Q), \infty)\big)$; if several values qualify, we take the arithmetic mean of the smallest and largest possible values).

The associated order statistics are

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

For $\alpha \in (0,1)$, choose k maximal such that $\mathrm{Bin}_{n,1/2}(\{0,\ldots,k-1\}) \le \frac{\alpha}{2}$, then

$$\big[X_{(k)}, X_{(n-k+1)}\big]$$

is a confidence interval for the median $m(\vartheta)$ at confidence level $1-\alpha$.

Proof. We have

$$\begin{aligned}\mathbb{P}_\vartheta\big(X_{(k)} > m(\vartheta)\big) &= \mathbb{P}_\vartheta\big(|\{1 \le i \le n : X_i \le m(\vartheta)\}| \le k-1\big)\\ &= \mathrm{Bin}_{n,1/2}(\{0,\ldots,k-1\}) \le \frac{\alpha}{2},\end{aligned}$$

analogously,

$$\mathbb{P}_\vartheta\big(X_{(n-k-1)} < m(\vartheta)\big) = \mathbb{P}_\vartheta\big(|\{1 \le i \le n : X_i \ge m(\vartheta)\}| \le k-1\big) \le \frac{\alpha}{2},$$

thus

$$\mathbb{P}_\vartheta\Big(\big[X_{(k)}, X_{(n-k+1)}\big] \not\ni m(\vartheta)\Big) \le \mathbb{P}_\vartheta\big(X_{(k)} > m(\vartheta)\big) + \mathbb{P}_\vartheta\big(X_{(n-k-1)} < m(\vartheta)\big) \le \alpha.$$

□

In the "sleeping pill comparison" example above with $n = 10$, for $\alpha = 0.01$: one must choose $k = 1$ (since $\mathrm{Bin}_{10,1/2}(\{0\}) \approxeq 0.001$, but $\mathrm{Bin}_{10,1/2}(\{0,1\}) \approxeq 0.012$), i.e., a confidence interval for the median (of the difference in sleep duration under drug A versus drug B) at 99% confidence level is $[X_{(1)}, X_{(10)}] = [0,\ 4.6]$.

For $\alpha = 0.05$, one could choose $k = 2$ and obtain $[X_{(2)}, X_{(9)}] = [0.8,\ 2.4]$ as a confidence interval at 95% confidence level.

1.3.2 Exact Confidence Intervals for the Success Parameter in the Binomial Distribution

Among n independent trials, x successes have been observed; we consider x as a realization of a $\mathrm{Bin}_{n,\vartheta}$-distributed random variable and wish to make inferences about ϑ based on the observation.

In Example 1.3.4 above, we considered the (approximate) confidence interval for ϑ at level $1-\alpha$ based on asymptotic normality.

Question: How can we proceed if we do not wish to rely on asymptotics? Consider $X \sim \mathbb{P}_\vartheta := \mathrm{Bin}_{n,\vartheta}$, confidence interval for $\vartheta \in \Theta = [0,1]$?

Idea: For $\vartheta \in \Theta := [0,1]$ choose $c_\vartheta \in (0,1)$ such that for

$$C_\vartheta := \big\{x \in \{0,1,\dots,n\} : \mathrm{Bin}_{n,\vartheta}(\{x\}) \ge c_\vartheta\big\}$$

we have $\mathrm{Bin}_{n,\vartheta}(C_\vartheta) \ge 1-\alpha$ (and c_ϑ as large as possible, so that C_ϑ is as small as possible).

Set $C(x) := \{\vartheta \in \Theta : x \in C_\vartheta\}$ for $x \in \Omega := \{0,1,\dots,n\}$, then we have

$$\forall\, \vartheta \in \Theta\ :\ \mathbb{P}_\vartheta(\vartheta \in C(X)) = \mathbb{P}_\vartheta(X \in C_\vartheta) \ge 1-\alpha$$

by construction. We have

1. For $\vartheta \in (0,1)$,
   the function $\{0,\dots,n\} \ni x \mapsto \mathrm{Bin}_{n,\vartheta}(\{x\})$ is strictly increasing on $\{0,1,\dots,\lceil (n+1)\vartheta - 1\rceil\}$,
   strictly decreasing on $\{\lfloor (n+1)\vartheta\rfloor,\dots,n\}$,
   thus maximal at $x = \lfloor (n+1)\vartheta\rfloor$ (and at $(n+1)\vartheta - 1$, when $(n+1)\vartheta \in \mathbb{Z}$).
2. For $x \in \{1,\dots,n\}$ the function $[0,1] \ni \vartheta \mapsto \mathrm{Bin}_{n,\vartheta}(\{x, x+1,\dots,n\})$ is continuous, strictly monotonically increasing with

   $$\mathrm{Bin}_{n,\vartheta}(\{x, x+1,\dots,n\}) = \mathrm{Beta}_{x,n-x+1}([0,\vartheta]),$$

   where $\mathrm{Beta}_{a,b}$ has density

   $$f_{\mathrm{Beta}_{a,b}}(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$

   on $(0,1)$.

Proof. 1.

$$\begin{aligned}\frac{\mathrm{Bin}_{n,\vartheta}(\{x\})}{\mathrm{Bin}_{n,\vartheta}(\{x-1\})} &= \frac{\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}}{\binom{n}{x-1}\vartheta^{x-1}(1-\vartheta)^{n-x+1}}\\ &= \frac{(n-x+1)\vartheta}{x(1-\vartheta)}\\ &> 1\\ \iff x &< (n+1)\vartheta\end{aligned}$$

2. $U_1,\dots,U_n$ independent and uniform on $[0,1]$, $S_\vartheta := \sum_{i=1}^n \mathbf{1}_{[0,\vartheta]}(U_i)$ is $\mathrm{Bin}_{n,\vartheta}$-distributed.Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the "order statistic".

$$\begin{aligned}&\mathrm{Bin}_{n,\vartheta}(\{x,\dots,n\}) = \mathbb{P}(S_\vartheta \geq x) = \mathbb{P}(U_{(x)} \leq \vartheta)\\ &= \sum_{k=1}^n \sum_{\substack{B\subseteq\{1,\dots,n\}\setminus\{k\}\\ |B|=x-1}} \underbrace{\mathbb{P}\Big(\begin{array}{l}U_k \leq \vartheta, U_m \leq U_k \text{ for } m\in B,\\ U_l > U_k \text{ for } l\in\{1,\dots,n\}\setminus(\{k\}\cup B)\end{array}\Big)}_{=\int_0^\vartheta u^{|B|}(1-u)^{n-|B|-1}\,du = \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du}\\ &= \underbrace{n\binom{n-1}{x-1}}_{} \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du\\ &= \frac{n!}{(x-1)!(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}\end{aligned}$$

□

Choose $C_\vartheta := \{x_-(\vartheta), x_-(\vartheta)+1,\dots,x_+(\vartheta)\}$ with $x_-(\vartheta) = \max\{x : \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x-1\}) \leq \frac{\alpha}{2}\}$ and $x_+(\vartheta) = \min\{x : \mathrm{Bin}_{n,\vartheta}(\{x+1,\dots,n\}) \leq \frac{\alpha}{2}\}$.

Then we have:

- $x \leq x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{x,\dots,n\}) = Beta_{x,n-x+1}([0,\vartheta]) > \frac{\alpha}{2}$
  $\iff \vartheta > p_-(x) := \frac{\alpha}{2}$-quantile of $\mathrm{Beta}_{x,n-x+1}$.
- $x \geq x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x\}) = 1 - \mathrm{Bin}(\{x+1,\dots,n\}) = \mathrm{Beta}_{x+1,n-x}([\vartheta,1]) \geq \frac{\alpha}{2}$.
  $\iff \vartheta < p_+(x) := 1 - \frac{\alpha}{2}$-quantile of $\mathrm{Beta}_{x+1,n-x}$.

We have thus proved:

Theorem 1.3.6 (exact confidence interval in the binomial model)**.** *With*

$$\begin{aligned}p_-(x) &:= \tfrac{\alpha}{2}\text{-quantile of } \mathrm{Beta}_{x,n-x+1},\\ p_+(x) &:= 1-\tfrac{\alpha}{2}\text{-quantile of } \mathrm{Beta}_{x+1,n-x}\end{aligned}$$

the mapping $x \mapsto [p_-(x), p_+(x)]$ is a confidence interval for ϑ at confidence level $1-\alpha$.

Remark. • Quantiles of the Beta distribution are tabulated; the symmetry property

$$\mathrm{Beta}_{a,b}([0,x]) = \mathrm{Beta}_{b,a}([1-x,1]) = 1 - \mathrm{Beta}_{b,a}([0,1-x])$$

can sometimes be useful when consulting tables.

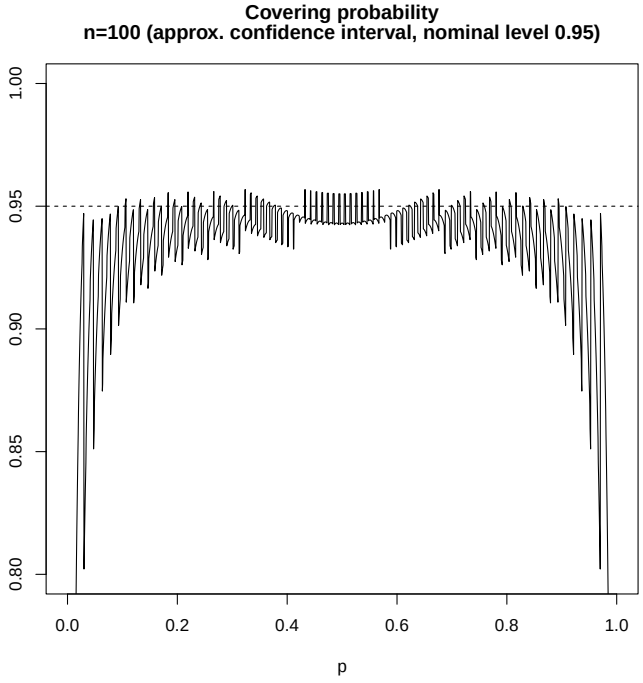
- R knows the Beta distributions, their distribution functions `pbeta(x, a, b)` and their quantiles `qbeta(p, a, b)`

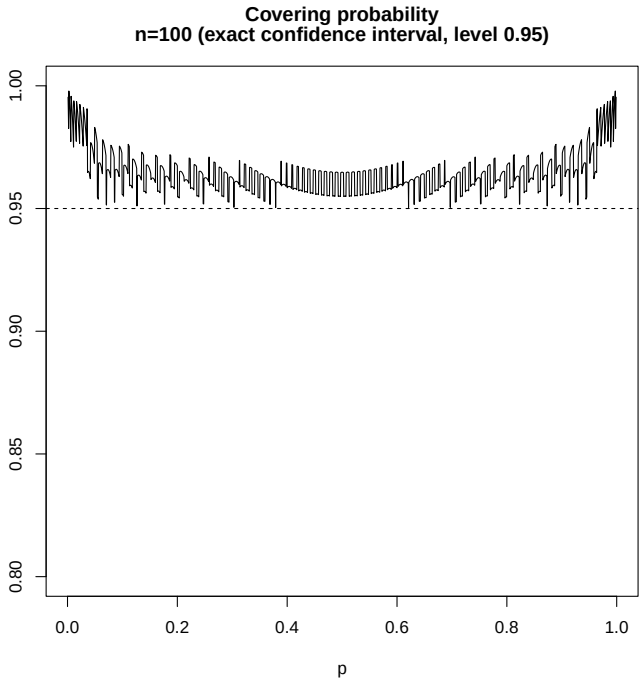
Example. $n = 53$, $x = 23$, choose $\alpha = 0.05$

$\widehat{\vartheta} = \frac{23}{53} \approx 0.434$, $\widehat{\sigma} \approx 0.496$, $q_{0.975} \approx 1.96$, $\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0.30, 0.57]$

$p_-(23) = 0.025$-quantile of $\mathrm{Beta}_{23,31} \approx 0.30$, $p_+(23) = 0.975$-quantile of $\mathrm{Beta}_{24,30} \approx 0.57$

(Absurdly precise values would be: $\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0.3005306, 0.5673939]$ $[p_-(23), p_+(23)] \approx [0.2983921, 0.5771742]$)





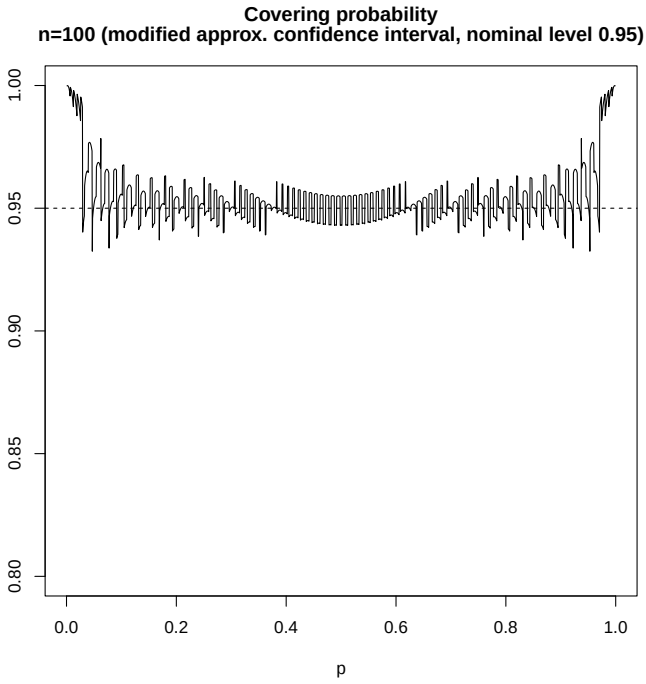
Remark 1.3.7. Sometimes one also considers the estimator

$$\widetilde{\vartheta} = \frac{x+1}{n+2}$$

for ϑ and constructs as (approximate) confidence interval at level $1-\alpha$

$$\big[\widetilde{\vartheta} - q\frac{\widetilde{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widetilde{\sigma}}{\sqrt{n}}\big]$$

with $\widetilde{\sigma} = \sqrt{\widetilde{\vartheta}(1-\widetilde{\vartheta})}$, q the $1-\frac{\alpha}{2}$-quantile of $\mathcal{N}_{0,1}$.



1.4 Further material and comments

1.4.1 On exponential families

Proposition 1.4.1. *Consider a (one-parameter) exponential family as in Definition 1.1.9 with likelihood function $\rho(\vartheta, x) = h(x) \cdot \exp\big(a(\vartheta)\cdot T(x) - b(\vartheta)\big)$ as in* (1.2)*, where a is continuously differentiable with $a'(\vartheta) \neq 0$ for $\vartheta \in \Theta$. Then we have*

1. *b is continuous differentiable on Θ with $b'(\vartheta) = a'(\vartheta)\mathbb{E}_\vartheta[T]$.*
2. *Let $S : \Omega \to \mathbb{R}$ be a statistic with $S \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ for each $\vartheta \in \Theta$. Then S is regular.*

   In particular, an exponential family describes a regular model and T is a regular statistic. $\tau(\vartheta) = \mathbb{E}_\vartheta[T]$ is continuously differentiable with $\tau'(\vartheta) = a'(\vartheta)\cdot \mathrm{Var}_\vartheta(T) > 0$.
3. *The Fisher information is given by $I_\vartheta = a'(\vartheta)\tau'(\vartheta)$.*

Proof. We formulate the in the following the proof for the case of a continuous model, in the discrete case, integrals $\int_\Omega ...dx$ are replaced by sums $\sum_{x\in\Omega} ...$, etc.

We assume w.l.o.g. that $a(\vartheta) = \vartheta$, otherwise we can re-parametrise and then use the chain rule ($a : \Theta \to a(\Theta)$ is by assumption a differentiable bijection). Note that then $a'(\vartheta) \equiv 1$.

(1) Let S be a statistic such that the expected value

$$\mathbb{E}_\vartheta[S] = \int_\Omega S(x)h(x)e^{a(\vartheta)T(x)-b(\vartheta)}\,dx$$

exists for all $\vartheta \in \Theta$ (and is finite). Put

$$u_S(\vartheta) \coloneqq e^{b(\vartheta)}\mathbb{E}_\vartheta[S] = \int_\Omega S(x)h(x)e^{a(\vartheta)T(x)}\,dx.$$

Fix $\vartheta \in \Theta$ for the moment. For $t \in \mathbb{R}$ so small that $\vartheta \pm t \in \Theta$ we have

$$\sum_{k=0}^{\infty} \frac{|t|^k}{k!} \int_\Omega h(x)|S(x)| \cdot |T(x)|^k e^{\vartheta T(x)}\,dx = \int_\Omega \sum_{k=0}^{\infty} \frac{|t|^k}{k!} h(x)|S(x)| \cdot |T(x)|^k e^{\vartheta T(x)}\,dx$$
$$= \int_\Omega |S(x)|h(x)e^{\vartheta T(x)+|tT(x)|}\,dx \le \int_\Omega |S(x)|h(x)e^{(\vartheta+t)T(x)}\,dx + \int_\Omega |S(x)|h(x)e^{(\vartheta-t)T(x)}\,dx < \infty,$$

where we used monotone convergence in the first line. In particular, $ST^k \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ for all ϑ and all $k \in \mathbb{N}_0$ and the sum $\sum_{k=0}^{\infty} \frac{t^k}{k!} \int_\Omega S(x)h(x)T(x)^k e^{\vartheta T(x)}\,dx$ converges absolutely with limit

$$\int_\Omega S(x)h(x)e^{(\vartheta+t)T(x)}\,dx = u_S(\vartheta + t).$$

Thus, $\Theta \ni \vartheta \mapsto u_S(\vartheta)$ is analytic and we can differentiate under the integral to obtain

$$u_s'(\vartheta) = \int_\Omega S(x)T(x)h(x)e^{\vartheta T(x)}\,dx = e^{b(\vartheta)}\mathbb{E}_\vartheta[ST].$$

The choice $S \equiv 1$ yields (note $u_1(\vartheta) = b(\vartheta)$)

$$\begin{aligned}
u_1'(\vartheta) &= e^{b(\vartheta)}\mathbb{E}_\vartheta[T],\ u_1''(\vartheta) = e^{b(\vartheta)}\mathbb{E}_\vartheta[T^2] \text{ with} \\
b'(\vartheta) &= \frac{d}{d\vartheta}\log u_1(\vartheta) = \frac{u_1'(\vartheta)}{u_1(\vartheta)} = \mathbb{E}_\vartheta[T] = \tau(\vartheta),\ \text{i.e. Claim 1. holds and} \\
\tau'(\vartheta) &= b'(\vartheta) = \frac{u_1''(\vartheta)}{u_1(\vartheta)} - \left(\frac{u_1'(\vartheta)}{u_1(\vartheta)}\right)^2 = \mathbb{E}_\vartheta[T^2] - (\mathbb{E}_\vartheta[T])^2 = \mathrm{Var}_\vartheta(T).
\end{aligned}$$

(2) Consider a general statistic S with $S \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ for each $\vartheta \in \Theta$.

$$\begin{aligned}
\frac{d}{d\vartheta}\mathbb{E}_\vartheta[S] &= \frac{d}{d\vartheta}(e^{-b(\vartheta)}u_S(\vartheta)) = (u_S'(\vartheta) - u_S(\vartheta)b'(\vartheta))e^{-b(\vartheta)} \\
&= \mathbb{E}_\vartheta[ST] - \mathbb{E}_\vartheta[S]\mathbb{E}_\vartheta[T] = \mathrm{Cov}_\vartheta[S,T] \\
&= \mathrm{Cov}_\vartheta[S,U_\vartheta]
\end{aligned}$$

where we used in the last step that $U_\vartheta(x) = \frac{d}{d\vartheta}\log\rho(\vartheta,x) = T(x) - b'(\vartheta)$ (recall that $a(\vartheta) = \vartheta$ by assumption) and that covariance is invariant under deterministic shifts.

The choice $S = T$ yields

$$\begin{aligned}
\tau'(\vartheta) = \frac{d}{d\vartheta}\int T(x)\rho(\vartheta,x)\,dx &= \mathbb{E}_\vartheta[TU_\vartheta] = \int T(x)\frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)}{\rho(\vartheta,x)} \cdot \rho(\vartheta,x)\,dx \\
&= \int T(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx.
\end{aligned}$$

Furthermore,

$$\frac{d}{d\vartheta}\mathbb{E}_\vartheta[S] = \frac{d}{d\vartheta}\int S(x)\rho(\vartheta,x)\,dx = \int S(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\,dx,$$

i.e. Claim 2. holds.

(3) We have $I_\vartheta = \mathrm{Var}_\vartheta[U_\vartheta] = \mathrm{Var}_\vartheta[T] = \tau'(\vartheta)$. (Recall that we parameterised via $a(\vartheta) = \vartheta$ so that $a'(\vartheta) = 1$.) □

Appendix A

Probability tools

A.1 A crash course on conditional expectations and conditional distributions

We collect here material, mostly without proofs, for ease of reference. For a more complete discussion see, e.g. [Wil91, Ch. 9], [Kle20, Kap. 8], [RW94] (in particular, Thm. (89.1) there) or [Bir24].

To set the stage, recall the basic definition of conditional probability: Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $B \in \mathscr{F}$ with $\mathbb{P}(B) > 0$.

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad A \in \mathscr{F}$$

is called the *conditional probability of A, given B*.

$\mathbb{P}(\cdot \mid B)$ defined by this formula is a probability measure on $(\Omega, \mathscr{F})$ with $\mathbb{P}(B \mid B) = 1$; for $X \in \mathcal{L}^1(\mathbb{P})$,

$$\mathbb{E}[X \mid B] = \int_\Omega X(\omega)\, \mathbb{P}(d\omega \mid B) = \frac{\mathbb{E}[\mathbb{1}_B X]}{\mathbb{P}(B)}. \tag{A.1}$$

The fact that $\mathscr{F} \ni A \mapsto \mathbb{P}(A \mid B)$ has the properties of a probability measure (σ-additivity, normalisation to total mass $= 1$) are obvious; the formula for $\mathbb{E}[X \mid B]$ is clear for indicator functions $X = \mathbb{1}_A$ and carries over in the "usual way" to all $\mathcal{L}^1$ functions.

What sense can give the conditional expectation of X, given another random variable Y (or, more generally, given the information in a certain sub-σ-algebra)? Consider first the discrete case.

Remark A.1.1 (Discrete case of conditional expectation). Let X and Y be random variables (on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$), where X is real-valued with $X \in \mathcal{L}^1(\mathbb{P})$, and Y takes values in a countable set $S = \{y_1, y_2, \dots\}$. Define

$$f(y) := \mathbb{E}[X \mid \{Y = y\}] \quad \text{for } y \in S \qquad \text{and} \qquad \mathbb{E}[X \mid Y] := f(Y). \tag{A.2}$$

Clearly, $\mathbb{E}[X \mid Y]$ is then $\sigma(Y)$-measurable (it is a function of Y) and for $A \in \sigma(Y)$, we have

$$\mathbb{E}[Y \mathbb{1}_A] = \mathbb{E}\big[\mathbb{E}[Y \mid X] \mathbb{1}_A\big]. \tag{A.3}$$

Indeed, for $A = \{Y = y_i\}$ we have (using (A.1) in the first equality)

$$\begin{aligned} \mathbb{E}[X \mathbb{1}_{\{Y=y_i\}}] &= \mathbb{P}(Y = y_i)\mathbb{E}[X \mid \{Y = y_i\}] \\ &= \mathbb{P}(Y = y_i) f(y_i) = \mathbb{E}[f(Y)\mathbb{1}_{\{Y=y_i\}}] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\mathbb{1}_{\{Y=Y_i\}}\big] \end{aligned}$$

and for $A = \{Y \in B\}$ with $B \subset S = \{y_1, y_2, \dots\}$ (note that every $A \in \sigma(Y)$ has this form) this gives

$$\begin{aligned}\mathbb{E}[X\mathbb{1}_{\{Y\in B\}}] &= \mathbb{E}\Big[Y\sum_{y\in B}\mathbb{1}_{\{Y=y\}}\Big] = \sum_{y\in B}\mathbb{E}[X\mathbb{1}_{\{Y=y\}}] \\ &= \sum_{y\in B}\mathbb{E}\big[\mathbb{E}[X\mid Y]\mathbb{1}_{\{Y=y\}}\big] = \mathbb{E}\Big[\mathbb{E}[X\mid Y]\sum_{y\in B}\mathbb{1}_{\{Y=y\}}\Big] = \mathbb{E}\big[\mathbb{E}[X\mid Y]\mathbb{1}_{\{Y=y\}}\big].\end{aligned}$$

Analogously, if $\mathscr{G} \subset \mathscr{F}$ is a sub-σ-algebra which has countably many "atoms" $A_1, A_2, \dots$ (i.e., $\emptyset \neq A_1, A_2, \dots \in \mathscr{G}$ are pairwise disjoint and every $B \in \mathscr{G}$ has the form $B = \bigcup_{i\in I} A_i$ for some – neccessarily countable – index set $I \subset \mathbb{N}$), one defines

$$\mathbb{E}[X\,|\,\mathscr{G}](\omega) = \frac{1}{\mathbb{P}(A_i)}\mathbb{E}[X1_{A_i}] \quad \text{for } \omega \in A_i \tag{A.4}$$

and analogous to (A.3) one has

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}\big[\mathbb{E}[X\,|\,\mathscr{G}]\mathbb{1}_A\big] \quad \text{for } A \in \mathscr{G}. \tag{A.5}$$

(We assume here for simplicity that $\mathbb{P}(A_i) > 0$ for all i; in general, if $\mathbb{P}(A_i) = 0$, one can set simply $\mathbb{E}[X\,|\,\mathscr{G}](\omega) = 0$ for $\omega \in A_i$).

In fact, the cases (A.2) and (A.4) are essentially equivalent by considering $A_i = \{Y = y_i\}$ and $\mathscr{G} = \sigma(Y)$.

We consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The generalisation of the discrete conditional expectation from Remark A.1.1 works as follows. (Note that a definition as in (A.1), (A.2) can not – at least without further work – be extended to the general case because one can have $\mathbb{P}(Y = y) = 0$ for all y generically. On the other hand, the property (A.3) does not use whether or not Y is discrete. This observation is the basis of Definition A.1.2.)

Definition A.1.2. Let $X \in \mathcal{L}^1(\mathbb{P})$, $\mathscr{G} \subset \mathscr{F}$ a sub-σ-algebra. A real-valued random variable Y is called (a version of the) conditional expectation of X given $\mathscr{G}$ (written $\mathbb{E}[X \mid \mathscr{G}]$), if

i) Y is $\mathscr{G}$-measurable, i.e. $\{Y \in B\} \in \mathscr{G}$ for every measurable subset $B \subset \mathbb{R}$

ii) $\mathbb{E}[Y\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A]$ for all $A \in \mathscr{G}$.

Remark. Equivalently, ii) can be replaced by ii'):

ii') $\mathbb{E}[Y \cdot H] = \mathbb{E}[X \cdot H]$ for all real-valued, bounded, $\mathscr{G}$-measurable random variables H.

If $\mathscr{G} = \sigma(Z)$ for a random variable Z, we often write $\mathbb{E}[X \mid Z] \coloneqq \mathbb{E}[X \mid \sigma(Z)]$.

Theorem A.1.3. *For $X \in \mathscr{L}^1(\mathbb{P})$, the conditional expectation $\mathbb{E}[X \mid \mathscr{G}]$ exists and is unique (up to $\mathbb{P}$-a.s. equality).*

We will not prove this theorem here (see the literature; one can use orthogonal $\mathcal{L}^2$-projection on the subspace of square integrable $\mathscr{G}$-measurable random variables or the Radon-Nikodým theorem).

Fact A.1.4. Conditional expectation has the "usual" properties of an expectation: linearity, positivity, the triangle inequality, continuity properties (under monotone convergence or under dominated convergence); these hold a.s. for every version of the conditional expectation.

In addition, conditional expectation has the following properties:

i) ("pulling out what is known": a $\mathscr{G}$-measurable random variable acts like a constant for ordinary expectation)

If $\mathbb{E}\big[|XY|\big] < \infty$ and Y is $\mathscr{G}$-measurable, then

$$\mathbb{E}[XY \mid \mathscr{G}] = Y \cdot \mathbb{E}[X \mid \mathscr{G}] \text{ a.s.},$$

in particular, $\mathbb{E}[Y|\mathscr{G}] = Y$ a.s.

ii) (behaviour under independence)

If $\sigma(X)$ and $\mathscr{G}$ are independent, then $\mathbb{E}[X \mid \mathscr{G}] = \mathbb{E}[X]$ a.s.

iii) ("tower property")

If $\mathscr{G}' \subset \mathscr{G}$ is a sub-σ-algebra, then

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathscr{G}] \,\big|\, \mathscr{G}'\big] = \mathbb{E}[X \mid \mathscr{G}'] \text{ a.s.},$$

in particular,

$$\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]] = \mathbb{E}[X]$$

iv) (Jensen's inequality for conditional expectation)

If $k : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is convex, then

$$\mathbb{E}\big[k(X) \,\big|\, \mathscr{G}\big] \geq k\big(\mathbb{E}[X \mid \mathscr{G}]\big) \quad \text{a.s.}$$

There is an analogue of the discrete case (Rem. A.1.1) for the case with density.

Remark A.1.5. Let X, Y be real random variables with joint density $f_{X,Y}$, i.e.

$$\mathbb{P}\big((X,Y) \in A\big) = \int_A f_{X,Y}(x,y)\, \lambda^{\otimes 2}\big(d(x,y)\big) \quad \text{for } A \in \mathcal{B}(\mathbb{R}^2),$$

and assume $Y \in \mathcal{L}^1(\mathbb{P})$.

For $x \in \mathbb{R}$, put

$$f_X(x) \coloneqq \int_{\mathbb{R}} f_{X,Y}(x,y)\, \lambda(dy) \quad \text{the marginal density of } X,$$
$$\varphi(x) \coloneqq \int_{\mathbb{R}} y \frac{f_{X,Y}(x,y)}{f_X(x)}\, \lambda(dy) \mathbb{1}_{f_X(x)>0}.$$

Then

$$\mathbb{E}[Y \mid X] = \varphi(X) \text{ a.s.}$$

since for $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\mathbb{E}\big[\mathbb{1}_{\{X\in B\}}\varphi(X)\big] &= \int_{\mathbb{R}} \mathbb{1}_B(x)\varphi(x) f_X(x)\, \lambda(dx) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_B(x) y f_{X,Y}(x,y) \lambda(dy)\ \lambda(dx) \\
&= \int_{\mathbb{R}^2} \mathbb{1}_B(x) y f_{X,Y}(x,y)\, \lambda^{\otimes 2}\big(d(x,y)\big) = \mathbb{E}\big[\mathbb{1}_{\{X\in B\}} Y\big].
\end{aligned}$$

For $X \in \mathcal{L}^2(\mathbb{P})$ and a σ-algebra $\mathscr{G} \subset \mathscr{F}$ the *conditional variance of X given $\mathscr{G}$* is defined (in complete analogy with "usual" variance) via

$$\mathrm{Var}[X \mid \mathscr{G}] := \mathbb{E}\big[\big(X - \mathbb{E}[X \mid \mathscr{G}]\big)^2 \,\big|\, \mathscr{G}\big] \quad (\geq 0)$$

and using the properties from Fact A.1.4 we have the alternative representation

$$\mathrm{Var}[X \mid \mathscr{G}] := \mathbb{E}\big[X^2 \,\big|\, \mathscr{G}\big] - \big(\mathbb{E}[X \mid \mathscr{G}]\big)^2.$$

Proposition A.1.6 (Variance decomposition)**.**

$$\mathrm{Var}[X] = \mathbb{E}\big[\,\mathrm{Var}[X \mid \mathscr{G}]\big] + \mathrm{Var}\,\big[\mathbb{E}[X \mid \mathscr{G}]\big] \tag{A.6}$$

Thinking of X as a "measurement" (in a random experiment) and of $\mathscr{G}$ as containing information on a "group label", one can interpret (A.6) as

$$\text{Variance = expectation of in-group variance } + \text{ variance of group means}$$

For the proof of (A.6) note

$$\begin{aligned}
&\mathbb{E}\big[\,\mathrm{Var}[X \mid \mathscr{G}]\big] + \mathrm{Var}\,\big[\mathbb{E}[X \mid \mathscr{G}]\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}[X^2 \mid \mathscr{G}] - \big(\mathbb{E}[X \mid \mathscr{G}]\big)^2\big] + \mathbb{E}\big[\big(\mathbb{E}[X \mid \mathscr{G}]\big)^2\big] - \big(\mathbb{E}\big[\mathbb{E}[X \mid \mathscr{G}]\big]\big)^2 \\
&\quad = \mathbb{E}\big[\mathbb{E}[X^2 \mid \mathscr{G}]\big] - \big(\mathbb{E}\big[\mathbb{E}[X \mid \mathscr{G}]\big]\big)^2 = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2 = \mathrm{Var}[X].
\end{aligned}$$

Fact A.1.7 (On regular versions of conditional distributions). If $Y = 1_B$ for an event $B \in \mathscr{A}$, one sometimes writes $\mathbb{P}(B \mid \mathscr{G}) = \mathbb{E}[Y \mid \mathscr{G}]$. However, one must be somewhat careful with the interpretation of $\mathbb{P}(\cdot \mid \mathscr{G})$ as a (random) measure, since in general, uncountably many B are involved and thus the compatibility of the null sets implicitly involved in the definition of the conditional expectation (cf. Def. A.1.2) is at least a priori unclear.

In "well-behaved" cases, a consistent choice is possible, the keyword here is "regular conditional distribution of a random variable". We briefly sketch the real-valued case:

Let X be a real-valued r.v. (on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$), $\mathscr{G} \subset \mathscr{F}$ a sub-σ-algebra. Then there exists a stochastic kernel $\kappa_{X|\mathscr{G}}$ from $(\Omega, \mathscr{G})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with

$$\kappa_{X|\mathscr{G}}(\omega, B) = \mathbb{E}\big[\mathbb{1}_{\{X \in B\}} \,\big|\, \mathscr{G}\big](\omega) \;\text{ a.s. for all } B \in \mathcal{B}(\mathbb{R}),$$

i.e. (cf. [Bir24, Def. 5.4] or [Kle20, Def. 8.25])

$$\begin{aligned}
&\text{for each } B \in \mathcal{B}(\mathbb{R}),\ \omega \mapsto \kappa_{X|\mathscr{G}}(\omega, B) \text{ is a version of } \mathbb{E}\big[\mathbb{1}_{\{X \in B\}} \,\big|\, \mathscr{G}\big] \text{ and} \\
&\text{for each } \omega,\ \kappa_{X|\mathscr{G}}(\omega, \cdot) \text{ is a probability measure on } (\mathbb{R}, \mathcal{B}(\mathbb{R})).
\end{aligned} \tag{A.7}$$

The main idea is to characterise the desired measure $\kappa_{X|\mathscr{G}}(\omega, \cdot)$ on $\mathbb{R}$ by its distribution function (see e.g. [Kle20, Thms. 8.29 and 8.37]); the crucial observation is that a distribution function (because of monotonicity) is already determined by its values at countably many points. So consider $B = (-\infty, r], r \in \mathbb{Q}$ and set

$$F_r := \mathbb{E}[\mathbb{1}_{(-\infty, r]}(X) \mid \mathscr{G}].$$

Then (with the properties of conditional expectation from Fact A.1.4) as desired, $\mathbb{P}$-a.s.: $F_r \leq F_{r'}$, for $r < r', (r, r' \in \mathbb{Q})$, $\lim_{n\to\infty} F_{r+\frac{1}{n}} = F_r$ for $r \in \mathbb{Q}$, $\lim_{n\to\infty} F_n = 1$, $\lim_{n\to-\infty} F_n = 0$.

Because of the countability of $\mathbb{Q}$, there exists $N \in \mathscr{F}$ with $\mathbb{P}(N) = 0$, such that the above holds for $\omega \in \Omega \setminus N$ and all $r, r' \in \mathbb{Q}$. Then

$$\widetilde{F}_s \coloneqq \begin{cases} \inf\{F_r : r \geq s, r \in \mathbb{Q}\}, & \omega \in \Omega \setminus N, \\ \overline{F}_s, & \omega \in N, \end{cases}$$

where $\overline{F}$ is any distribution function that (random) distribution function of $k_{X,\mathscr{G}}$. Details can be found, for example, in [Kle20, Ch. 8.3].

This argument can be relatively easily extended to the situation where the range $(E, \mathscr{B})$ of X is a so-called standard Borel space (also called a Borel space), i.e. if there exists an $A \in \mathcal{B}(\mathbb{R})$ and a bijection $\phi : E \to A$ such that ϕ and ϕ^{-1} are both measurable (then $(E, \mathscr{B})$ and $(A, \mathcal{B}(A))$ are isomorphic as measurable spaces). Then $X' \coloneqq \phi \circ X$ is a real-valued r.v., and the above argument applies (cf. also [Kle20, Thm 8.37]).

Finally, one can show that every separable and complete metric space E, equipped with its Borel σ-algebra, is a standard Borel space (see e.g. [RW94, Ch. II.82]; [Bre92, Appendix 7]). Such spaces are called *Polish spaces*; they play an important role in general stochastic theory (for example, $\mathbb{R}^d$ or $C([0,1])$ with the supremum norm are Polish).

A.2 The multivariate normal distribution and related distributions

We recall here some properties of the multivariate normal distribution and related distributions.

Proposition A.2.1. *Let $n \in \mathbb{N}$, $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}_{0,1}$, then*

$$X \coloneqq X_1^2 + X_2^2 + \cdots + X_n^2 \quad \textit{has density } \tfrac{1}{\Gamma(n/2)} 2^{-n/2} x^{\frac{n}{2}-1} e^{-x/2} \mathbb{1}_{[0,\infty)}(x),$$

$\chi_n^2 \coloneqq \mathscr{L}(X)$ *is called the* chi-squared distribution *with n degrees of freedom.*

Note: $\chi_n^2 = \Gamma_{1/2,n/2}$, where the gamma distribution $\Gamma_{\alpha,\nu}$ has density

$$\frac{1}{\Gamma(\nu)} \alpha^\nu x^{\nu-1} e^{-\alpha x} \mathbb{1}_{(0,\infty)}(x)$$

(α = scale, ν = shape parameter).

Proposition A.2.2. *Let $\alpha, r, s > 0$, $X \sim \Gamma_{\alpha,r}, Y \sim \Gamma_{\alpha,s}$ be independent. Then*

$$X + Y \;\; \textit{and} \;\; V \coloneqq \frac{X}{X+Y} \;\; \textit{are independent}$$

and $X + Y \sim \Gamma_{\alpha,r+s}$, $V \sim \beta_{r,s}$, where the beta distribution $\beta_{r,s}$ has density

$$\tfrac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1} \mathbb{1}_{(0,1)}(v)$$

In particular, the gamma distributions form a convolution family (with respect to the second, so-called shape parameter): $\Gamma_{\alpha,r} * \Gamma_{\alpha,s} = \Gamma_{\alpha,r+s}$.

Proof. (X, Y) has density

$$f_{(X,Y)}(x,y) = \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} x^{r-1} y^{s-1} e^{-\alpha(x+y)} \text{ on } (0,\infty)^2.$$

Let $\varphi(x,y) = \begin{pmatrix} x+y \\ \frac{x}{x+y} \end{pmatrix}$, then

$$\varphi^{-1}(z,v) = \begin{pmatrix} zv \\ z(1-v) \end{pmatrix}, \quad D\varphi(x,y) = \begin{pmatrix} 1 & \frac{y}{(x+y)^2} \\ 1 & -\frac{x}{(x+y)^2} \end{pmatrix}, \quad |\det D\varphi(x,y)| = \frac{|x+y|}{(x+y)^2} = \frac{1}{|x+y|}$$

Write $Z := X+Y$, $V := \frac{X}{X+Y}$. According to the 2-dimensional density transformation (see Report A.2.7), the density of (Z,V) is:

$$\begin{aligned} f_{(Z,V)}(z,v) &= \frac{f_{(X,Y)}(\varphi^{-1}(z,v))}{|\det D\varphi(\varphi^{-1}(z,v))|} \\ &= z \cdot \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} (zv)^{r-1} (z(1-v))^{s-1} e^{-\alpha z} \\ &= \underbrace{\frac{\alpha^{r+s}}{\Gamma(r+s)} z^{r+s-1} e^{-\alpha z}}_{\text{density of } \Gamma_{\alpha,r+s}} \underbrace{\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1}}_{\text{density of } \beta_{r,s}} \end{aligned}$$

□

Proof of Proposition A.2.1. $X \sim \mathcal{N}_{0,1}$, then $X^2 \sim \Gamma_{\frac{1}{2},\frac{1}{2}}$ $(= \chi_1^2)$: $|X|$ has density $\frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ on $(0,\infty)$; let $\varphi : (0,\infty) \to (0,\infty), x \mapsto x^2$, $\varphi^{-1}(y) = \sqrt{y}$, $\frac{d}{dy}\varphi^{-1}(y) = 2y$, thus

$$X^2 \text{ has density } \frac{1}{2\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{1}{2}y}$$

(see Observation A.2.6).

This proves the claim for $n = 1$; the general case follows inductively using Proposition A.2.2. □

Corollary and definition A.2.3. *Let $m, n \in \mathbb{N}$, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be independent, $\sim \mathcal{N}_{0,1}$.*

1. $F_{m,n} := \dfrac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}$ *has density* $f_{m,n}(x) = \dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{(m+n)}{2}}} \mathbb{1}_{(0,\infty)}(x)$. $\mathscr{L}(F_{m,n})$ *is called the* Fisher distribution[1] *with m and n degrees of freedom (more precisely: with m numerator and n denominator degrees of freedom).*

2. $T_n := \dfrac{X}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}}$ *has density* $t_n(x) = \dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} \dfrac{1}{\sqrt{n}} \left(1 + \dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$.

$\mathscr{L}(T_n)$ *is called the* Student distribution[2] *with n degrees of freedom (also called Student's t-distribution).*

Note: The Student distribution with one degree of freedom is the Cauchy distribution.

Observation A.2.4. Let T_n be Student-distributed with n degrees of freedom, then $T_n \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}$ (since $t_n(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ locally uniformly).

[1] After Ronald Aylmer Fisher, 1890–1962

[2] After William Sealy Gosset, 1876–1937, who published it in 1908 under the pseudonym "Student".

Proof of Corollary A.2.3. 1. $X \coloneqq \sum_{i=1}^{m} X_i^2 \sim \Gamma_{\frac{1}{2},\frac{m}{2}}$, $Y \coloneqq \sum_{j=1}^{n} Y_j^2 \sim \Gamma_{\frac{1}{2},\frac{n}{2}}$ are independent, thus $V \coloneqq \frac{X}{X+Y} \sim \beta_{\frac{m}{2},\frac{n}{2}}$ by Proposition A.2.2.

Then

$$F_{m,n} = \frac{nX}{mY} = \frac{n}{m}\frac{\frac{X}{X+Y}}{\frac{Y}{X+Y}} = \frac{n}{m}\frac{V}{1-V},$$

with

$$\varphi : (0,1) \to (0,\infty), v \mapsto \frac{n}{m}\frac{v}{1-v}, \quad \text{so } \varphi^{-1}(z) = \frac{mz}{n+mz}, \quad \frac{d}{dv}\varphi(v) = \frac{n}{m}\frac{1}{(1-v)^2}$$

thus $F_{m,n} = \varphi(V)$, and has density

$$\begin{aligned} f_{m,n}(z) &= \frac{mnz}{(n+mz)^2}\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{mz}{n+mz}\right)^{\frac{m}{2}-1}\left(\frac{n}{n+mz}\right)^{\frac{n}{2}-1} \\ &= \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{z^{\frac{m}{2}-1}}{(n+mz)^{\frac{(m+n)}{2}}} \end{aligned}$$

2. T_n^2 has (by 1.) density $f_{1,n}$, so $|T_n|$ has density $2t f_{1,n}(t^2)\mathbb{1}_{[0,\infty)}(t)$.

Since T_n is symmetric about 0 (obvious from the symmetry of X_1), T_n has density

$$\begin{aligned} |t| f_{1,n}(t^2) &= |t|\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} n^{n/2} \frac{(t^2)^{\frac{1}{2}-1}}{(n+t^2)^{(n+1)/2}} \\ &= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}\frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} \end{aligned}$$

□

Proposition A.2.5. *$X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}_{\mu,\sigma^2}$ with $\mu \in \mathbb{R}$, $\sigma > 0$,*

$$M \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2.$$

We have

1. M and S^2 are independent, $M \sim \mathcal{N}_{\mu,\sigma^2/n}$, $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

2. $T \coloneqq \frac{\sqrt{n}(M-\mu)}{\sqrt{S^2}}$ is Student-distributed with $n-1$ degrees of freedom.

Proof. Without loss of generality, assume $\mu = 0$, $\sigma^2 = 1$, otherwise consider $X_i' \coloneqq (X_i - \mu)/\sqrt{\sigma^2}$. 1. Let O be an orthogonal $n \times n$ matrix whose first row is $z_1 = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$, i.e. extend z_1 to an orthonormal basis $z_1, \ldots, z_n$ of $\mathbb{R}^n$, set

$$O = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

Then

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \coloneqq O \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

is n-dimensional standard normal (by Example A.2.8, invariance of the n-dimensional normal distribution under orthogonal transformations). Thus

$$\begin{aligned} Y_1 &= \sum_{i=1}^{n} \frac{1}{\sqrt{n}} X_i = \sqrt{n} M, \qquad \text{so } M \sim \mathcal{N}_{0,1/n}, \\ (n-1)S^2 &= \sum_{i=1}^{n} (X_i - M)^2 = \sum_{i=1}^{n} X_i^2 - nM^2 \\ &= ||(X_1, \dots, X_n)^T||^2 - Y_1^2 = ||(Y_1, \dots, Y_n)^T||^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2, \end{aligned}$$

hence $(n-1)S^2 \sim \chi^2_{n-1}$ and independent of M. (Geometrically: decompose $\mathbb{R}^n = D \oplus D^\perp$ into the diagonal $D = \{(x, x, \dots, x) : x \in \mathbb{R}\} \subset \mathbb{R}^n$ and its orthogonal complement $D^\perp = \{(x_1, x_2, \dots, x_n) : x_1 + \cdots + x_n = 0\}$, let $\mathcal{P}_D : \mathbb{R}^n \to D$, $\mathcal{P}_{D\perp} = \mathrm{Id}_{\mathbb{R}^n} - \mathcal{P}_D : \mathbb{R}^n \to D^\perp$ be the orthogonal projections onto D and $D^\perp$, respectively, then $\sqrt{n}M$ is the (signed) length of $\mathcal{P}_D X$ and $(n-1)S^2 = ||\mathcal{P}_{D\perp} X||^2$.)
2. This follows from 1. and the definition (cf. Corollary and Definition A.2.3, 2.) □

A.2.1 On the density transformation formula

Observation A.2.6 (Density transformation in the case $\mathbb{R}^1$). Let X be a real random variable with density f_X, i.e. $F_X(x) = \int_{-\infty}^{x} f_X(z)\, dz$, $I \subset \mathbb{R}$ a (possibly unbounded) open interval with $P(X \in I) = 1$, $J \subset I$, $\varphi : I \to J$ continuously differentiable, bijective.

Then $Y \coloneqq \varphi(X)$ has density

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

Proof. φ must clearly be strictly increasing or strictly decreasing; we consider the increasing case.

For $z < \inf J$, $P(Y \le z) = 0$; for $z > \sup J$, $P(Y \le z) = 1$.

Let $z \in J$:

$$\begin{aligned} P(Y \le z) &= P\big(\varphi(X) \le z\big) = P\big(X \le \varphi^{-1}(z)\big) \\ &= \int_{-\infty}^{\varphi^{-1}(z)} f_X(x)\, dx = \int_{-\infty}^{z} f_X\big(\varphi^{-1}(y)\big) \frac{1}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|} dy, \end{aligned}$$

where we substituted $x = \varphi^{-1}(y)$ (and thus $\frac{dx}{dy} = \frac{1}{\varphi'(\varphi^{-1}(y))}$). See also the sketch below. □

Example. $X \sim \mathcal{N}_{0,1}$, $\mu \in \mathbb{R}$, $\sigma > 0$, then $Y \coloneqq \sigma X + \mu \sim \mathcal{N}_{\mu,\sigma^2}$ (exercise).

Fact A.2.7 (Density transformation in $\mathbb{R}^d$). Let X be an $\mathbb{R}^d$-valued random variable with density f_X, $I \subset \mathbb{R}^d$ open with $P(X \in I) = 1$, $J \subset \mathbb{R}^d$ open, $\varphi : I \to J$ bijective, continuously differentiable with derivative

$$\varphi'(x) = \left(\frac{\partial \varphi_i}{\partial x_j}(x)\right)_{i,j=1}^{d} \qquad \text{("Jacobian matrix")},$$

then $Y := \varphi(X)$ has density

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\det \varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

Proofs can be found in analysis textbooks, e.g. G. Kersting and M. Brokate, *Measure and Integral*, p. 107, H. Heuser, *Analysis, Part 2*, Theorem 205.2 ("Substitution Rule"), O. Forster, *Analysis 3*, Chapter 9, Theorem 1 ("Transformation Formula"). Here we consider only the following heuristic (in the case $d = 2$): Locally,

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \end{pmatrix}$$

"looks like"

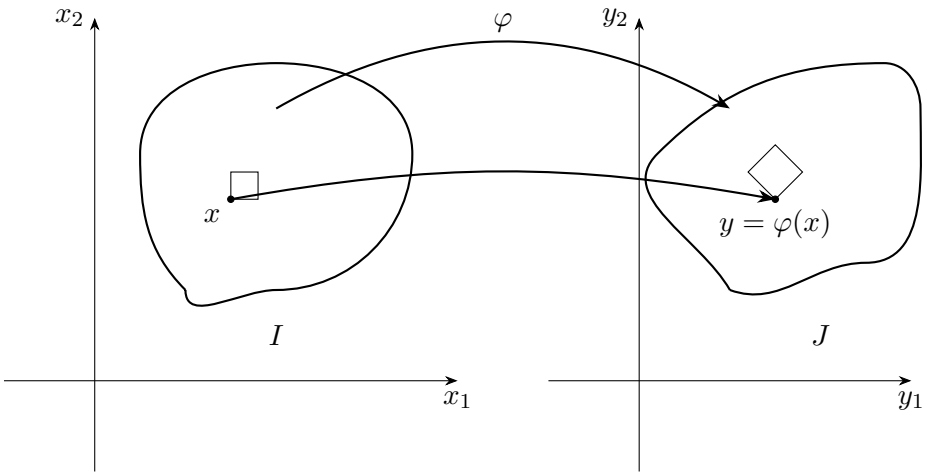
$$\begin{aligned} \varphi(x') &\approx \varphi(x) + \varphi'(x) \cdot (x' - x) \\ &= \varphi(x) + \begin{pmatrix} \frac{\partial}{\partial x_1}\varphi_1(x) & \frac{\partial}{\partial x_2}\varphi_1(x) \\ \frac{\partial}{\partial x_1}\varphi_2(x) & \frac{\partial}{\partial x_2}\varphi_2(x) \end{pmatrix} \cdot \begin{pmatrix} x_1' - x_1 \\ x_2' - x_2 \end{pmatrix} \end{aligned}$$

(plus terms of order $O(||x' - x||^2)$), thus:

the area of size $h_1 \cdot h_2$ "around x"

is mapped to

$\approx$ area $h_1 \cdot h_2 \cdot |\det \varphi'(x)|$ "around y".



Applying this to $Y = \varphi(X)$, this means intuitively: For $y = \varphi(x) \in J$ (and very small $h > 0$), we have

$$\begin{aligned} f_Y(y)h^2 &\approx \mathbb{P}\big(Y \text{ takes a value in a square of area } h^2 \text{ with "reference point" } y\big) \\ &\approx \mathbb{P}\big(X \text{ takes a value in a rectangle of area } h^2/|\det \varphi'(x)| \text{ with "reference point" } x\big) \\ &\approx f_X(x)\frac{h^2}{|\det \varphi'(x)|} = \frac{f_X(\varphi^{-1}(y))}{|\det \varphi'(\varphi^{-1}(y))|}h^2. \end{aligned}$$

Example A.2.8. 1. $X_1, \dots, X_n$ i.i.d., $X_i \sim \mathcal{N}_{0,1}$, then $X := (X_1, \dots, X_n)$ has density

$$f_X(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}||x||^2\big),\ x = (x_1, \dots, x_n) \in \mathbb{R}^n$$

(with $||x|| = \sqrt{x_1^2 + \cdots + x_n^2}$, the Euclidean norm). Let $M = (m_{ij})_{i,j=1}^n$ be an orthogonal $n \times n$ matrix (i.e. $M^T M = I$, the $n \times n$ identity matrix),

$$Y^T := MX^T \quad \text{i.e. } Y = (Y_1, \dots, Y_n) \text{ with } Y_i = \sum_{j=1}^{n} m_{ij} X_j,$$

then $Y_1, \dots, Y_n$ are i.i.d., $Y_i \sim \mathcal{N}_{0,1}$. $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, $\varphi(x) = Mx$ is bijective and differentiable with $\varphi^{-1}(y) = M^T y$, $\varphi' = M$, the density transformation formula (Fact A.2.7) shows: Y has density

$$\begin{aligned} f_Y(y) &= \frac{f_X(\varphi^{-1}(y))}{|\det \varphi'(\varphi^{-1}(y))|} = \frac{f_X(M^T y)}{|\det M|} = f_X(M^T y) \\ &= (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2} \underbrace{||M^T y||^2}_{||y||^2}\big) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2}. \end{aligned}$$

A.2.2 On general multivariate normal distributions

Observation A.2.9. i) Let $X = (X_1, \dots, X_d)^{\mathsf{t}}$ be a $\mathbb{R}^d$-valued random variable. The covariance matrix $C = (c_{ij})_{i,j=1,\dots,d}$ with $c_{ij} = \operatorname{Cov}[X_i, X_j]$ is symmetric and positive definite, since $c_{ij} = \operatorname{Cov}[X_i, X_j] = \operatorname{Cov}[X_j, X_i] = c_{ji}$ and for $a = (a_1, \dots, a_d)^{\mathsf{t}} \in \mathbb{R}^d$ we have

$$a^{\mathsf{t}} C a = \sum_{i,j=1}^{d} a_i c_{ij} a_j = \sum_{i,j=1}^{d} a_i a_j \operatorname{Cov}[X_i, X_j] = \operatorname{Cov}\left[\sum_{i=1}^{d} a_i X_i, \sum_{j=1}^{d} a_j X_j\right] = \operatorname{Var}[\langle a, X\rangle] \geq 0.$$

ii) If $Z_1, \dots, Z_d$ are independent and standard normally distributed, then $Z = (Z_1, \dots, Z_d)^{\mathsf{t}}$ has density

$$f_Z(z) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\left(z_1^2 + \cdots + z_d^2\right)\right) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|z\|^2}, \quad z \in \mathbb{R}^d.$$

$\mathcal{L}(Z)$ is called the d-dimensional standard normal distribution.

iii) Let $\mu \in \mathbb{R}^d$ and $A = (a_{ij}) \in \mathbb{R}^{d \times d}$. Then $X := \mu + AZ$ has expectation vector $\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]) = \mu$ and covariance matrix $C := AA^{\mathsf{t}}$, since

$$\begin{aligned} \operatorname{Cov}[X_k, X_l] = \operatorname{Cov}\left[\mu_k + \sum_{i=1}^{d} a_{k,i} Z_i,\ \mu_l + \sum_{j=1}^{d} a_{l,j} Z_l\right] &= \sum_{i,j=1}^{d} a_{ki} a_{lj} \operatorname{Cov}[Z_i, Z_j] \\ &= \sum_{i,j=1}^{d} a_{ki} a_{lj} \delta_{ij} = \sum_{i=1}^{d} a_{ki} a_{li} = (AA^{\mathsf{t}})_{kl}. \end{aligned}$$

If A has full rank d, then X has density

$$f_{\mu,C}(x) = \frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{1}{2}\langle x - \mu, C^{-1}(x - \mu)\rangle\right), \quad x \in \mathbb{R}^d,$$

since for $g(z) \coloneqq \mu + Az$ we have $g^{-1}(x) = A^{-1}(x-\mu)$ and $\left(\frac{\partial}{\partial x_i} g_j(z)\right)_{i,j} = \mathrm{D}g(z) = A$. Thus, with the density transformation formula and $\det C = \det(AA^{\mathsf{t}}) = (\det A)^2$:

$$f_{\mu,C}(x) = f_Z\left(g^{-1}(x)\right) \frac{1}{|\det \mathrm{D}g^{-1}(x)|}.$$

If A does not have full rank, then X has no density with respect to λ^d.

What, however, happens in the case where A (and thus also C) does not have full rank? Consider for $u \in \mathbb{R}^d$:

$$\begin{aligned}
\mathbb{E}\left[e^{i\langle u,X\rangle}\right] &= \mathbb{E}\left[e^{i\langle u,\mu\rangle} \cdot e^{i\langle u,AZ\rangle}\right] = e^{i\langle u,\mu\rangle}\mathbb{E}\left[e^{i\sum_{k,l=1}^d u_k a_{kl} Z_l}\right] = e^{i\langle u,\mu\rangle}\prod_{l=1}^d \mathbb{E}\left[e^{i\sum_{k=1}^d u_k a_{kl} Z_l}\right] \\
&= e^{i\langle u,\mu\rangle}\prod_{l=1}^d e^{-\frac{1}{2}\left(\sum_{k=1}^d u_k a_{kl}\right)^2} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\sum_{l=1}^d ((u^{\mathsf{t}}A)_l)^2} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u^{\mathsf{t}}A, u^{\mathsf{t}}A\rangle} \\
&= e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u^{\mathsf{t}}, u^{\mathsf{t}}AA^{\mathsf{t}}\rangle} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u, Cu\rangle}.
\end{aligned}$$

This suggests the following definition.

Definition A.2.10. Let $\mu \in \mathbb{R}^d$, $C \in \mathbb{R}^{d\times d}$ be symmetric and positive definite. X is called *d-dimensionally normally distributed with expectation μ and covariance matrix C* if

$$\varphi_X(u) = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u, Cu\rangle}.$$

We also write $\mathcal{L}(X) =: \mathcal{N}(\mu, C)$.

Remark A.2.11. Let $X \sim \mathcal{N}(\mu, C)$, $A \in \mathbb{R}^{d\times d}$ and $Y \coloneqq AX$. Then $Y \sim \mathcal{N}(A\mu, ACA^{\mathsf{t}})$, since

$$\mathbb{E}\left[e^{i\langle u,Y\rangle}\right] = \mathbb{E}\left[e^{i\langle u,AX\rangle}\right] = \mathbb{E}\left[e^{i\langle A^{\mathsf{t}}u,X\rangle}\right] = e^{i\langle A^{\mathsf{t}}u,\mu\rangle} e^{-\frac{1}{2}\langle A^{\mathsf{t}}u, CA^{\mathsf{t}}u\rangle} = e^{i\langle u,A\mu\rangle} e^{-\frac{1}{2}\langle u, ACA^{\mathsf{t}}u\rangle}.$$

Bibliography

[Bir24] Matthias Birkner, *Stochastik I*, Sommersemester 2024, `https://www.staff.uni-mainz.de/birkner/StochI_24/Stochastik_I_SS24.pdf`.

[Bre92] Leo Breiman, *Probability*, unabridged, corr. republ. ed., Classics in applied mathematics ; 7, SIAM, Philadelphia, Pa., 1992.

[Kle20] Achim Klenke, *Wahrscheinlichkeitstheorie*, 4th revised and supplemented edition ed., Masterclass, Berlin: Springer Spektrum, 2020 (German).

[LR06] E.L. Lehmann and J.P. Romano, *Testing statistical hypotheses*, Springer Texts in Statistics, Springer, 2006.

[RW94] L. C. G. Rogers and David Williams, *Diffusions, Markov processes, and martingales. Vol. 1: Foundations.*, 2nd ed. ed., Chichester: Wiley, 1994 (English).

[Wil91] David Williams, *Probability with martingales*, Cambridge mathematical textbooks, Cambridge Univ. Press, Cambridge, 1991.