

Introduction to Statistics

(Preliminary version ¹, December 19, 2003)

by Franz Merkl
University of Leiden

0 Preliminaries

In this section, we review some basic facts from Kansrekening en Statistiek 1 and from Analyse 1,2, and we introduce some notation.

0.1 Spaces of outcomes, events, and random variables

Consider a set Ω of possible outcomes of a random experiment. Subsets of Ω are called *events*.

(For technical reasons, sometimes not *all* subsets of Ω are called events. We ignore this fact in this lecture.)

A map $X : \Omega \rightarrow \mathbb{R}$ is called a *random variable*. More generally: a map $X : \Omega \rightarrow \mathbb{R}^n$ is called a (vector-valued) random variable or a *random vector*.

We use frequently abbreviations for events as in the following examples:

$$\{X < 7\} := \{\omega \in \Omega : X(\omega) < 7\}, \quad (1)$$

$$\{X \geq 2Y\} := \{\omega \in \Omega : X(\omega) \geq 2Y(\omega)\}, \quad (2)$$

i.e. we drop the argument ω .

0.2 Probability distributions

A probability measure (synonym: a probability distribution) on Ω is a map P which assigns a number $P[A] \in [0, 1]$ to every event $A \subseteq \Omega$, such that the following properties hold:

Kolmogorov's axioms

1. $P[A] \geq 0$,
2. $P[\Omega] = 1$,
3. If $A_1, A_2, A_3, \dots$ is a sequence of pairwise disjoint events, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$, then one has

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i]. \quad (3)$$

Here are two consequences:

$$P[\emptyset] = 0, \quad (4)$$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]. \quad (5)$$

The above mentioned abbreviation for events is frequently applied for probabilities, too. For example,

$$P[X < 7] := P[\{\omega \in \Omega : X(\omega) < 7\}]. \quad (6)$$

¹This preliminary version of a statistics script is distributed in the hope that it will be useful, but without any warranty.

0.3 Independence and conditional probabilities

Two events A and B are called *independent* if

$$P[A \cap B] = P[A]P[B]. \quad (7)$$

More generally, a family $(A_i)_{i \in I}$ of events is called *independent*, if

$$P \left[\bigcap_{i \in J} A_i \right] = \prod_{i \in J} P[A_i] \quad (8)$$

holds for all finite subsets $J \subseteq I$. A family $(X_i)_{i \in I}$ of random variables is called *independent*, if for all families $(B_i)_{i \in I}$ of intervals, the events $(\{X_i \in B_i\})_{i \in I}$ are independent. If A, B are events with $P[B] > 0$, then

$$P[A|B] := \frac{P[A \cap B]}{P[B]} \quad (9)$$

is called the *conditional probability* of A conditional on B . Then A and B are independent if and only if $P[A|B] = P[A]$. The map

$$A \mapsto P[A|B] \quad (10)$$

is a probability distribution, too. It is called the conditional distribution of P conditional on B .

0.4 Distribution of random variables and of random vectors

If $X : \Omega \rightarrow \mathbb{R}^n$ is a random variable (or a random vector), then its distribution with respect to P is following probability measure on $\mathbb{R}^n$:

$$B \mapsto P[X \in B] \quad \text{for events } B \subseteq \mathbb{R}^n. \quad (11)$$

We write $X \sim Q$ if X has the distribution Q . The probability distribution of a ($\mathbb{R}$ -valued) random variable X can be specified by its *cumulative distribution function* F_X ,

$$F_X(t) := P[X \leq t], \quad t \in \mathbb{R}. \quad (12)$$

F_X is monotonically increasing, continuous from the right, and satisfies

$$\lim_{t \rightarrow \infty} F_X[t] = 1, \quad (13)$$

$$\lim_{t \rightarrow -\infty} F_X[t] = 0. \quad (14)$$

A random variable X , or more generally a random vector $X = (X_1, \dots, X_n)$, has the density f_X if

$$P[X \in B] = \int_B f_X(t) dt \quad (15)$$

holds for all events $B \subseteq \mathbb{R}^n$. (If $n > 1$, this integral means of course a multi-dimensional integral.)

(Precisely speaking, densities are only specified up to sets of volume 0. Usually, we ignore this fact in this lecture.)

Densities are nonnegative functions, i.e. $f_X \geq 0$, which are normalized:

$$\int_{\mathbb{R}^n} f_X(t) dt = 1. \quad (16)$$

If (15) holds, we also say that the distribution of X has the density f_X .

0.5 Product of distributions and densities

If $X_1, \dots, X_n$ are *independent* random variables with distributions $P_1, \dots, P_n$, respectively, then the distribution of the random vector $(X_1, \dots, X_n)$ (i.e. the *joint* distribution) is called the product of $P_1, \dots, P_n$. It is denoted by $P_1 \otimes \dots \otimes P_n$. As a special case, if $P_1 = \dots = P_n = P$, we write $P_1 \otimes \dots \otimes P_n = P^n$, and we call $X_1, \dots, X_n$ *i.i.d.* random variables (*independent, identically distributed.*)

A random vector $X = (X_1, \dots, X_n)$ with the (joint) density f_X has independent components if and only if its density can be written as

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \cdot \dots \cdot f_{X_n}(x_n), \quad (17)$$

where f_{X_i} denotes the density of X_i .

0.6 Expectation, variance, and covariance

To random variables $X : \Omega \rightarrow \mathbb{R}$ (or $\mathbb{R}^n$), one can assign an *expected value* (synonym: expectation) $E[X] = E_P[X]$ with respect to P . If Ω is discrete, it is given by

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P[\{\omega\}] = \sum_x x P[X = x]. \quad (18)$$

If X has a density f_X , then

$$E[X] = \int_{\mathbb{R}^n} x f_X(x) dx. \quad (19)$$

More generally, if X is a function $X = g(Y)$ of a random variable (or a $\mathbb{R}^m$ -valued random vector) Y with a density f_Y , then

$$E[X] = \int_{\mathbb{R}^m} g(y) f_Y(y) dy. \quad (20)$$

The expectation might fail to exist if the integral or sum is not well defined. The expectation is linear:

$$E[X + Y] = E[X] + E[Y], \quad (21)$$

$$E[aX] = aE[X] \quad (22)$$

for $a \in \mathbb{R}$. *No independence assumption is required for this.* The *variance* of a random variable is

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \geq 0. \quad (23)$$

Its square root is called standard deviation. The variance is homogenous of degree 2:

$$\text{Var}[aX] = a^2 \text{Var}[X], \quad a \in \mathbb{R}. \quad (24)$$

One has

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \quad \text{if } X \text{ and } Y \text{ are independent.} \quad (25)$$

More generally, the *covariance* between two random variables X and Y is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]. \quad (26)$$

One has

$$\text{Var}[X + Y] = \text{Var}[X] + 2 \text{Cov}(X, Y) + \text{Var}[Y]. \quad (27)$$

More generally, the *covariance matrix* of a random vector $(X_1, \dots, X_n)$ is the matrix Σ with entries $\text{Cov}(X_i, X_j)$, $i, j = 1, \dots, n$. It is always positive semidefinite, i.e. $a \Sigma a^\top \geq 0$ for all $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Indeed, one has

$$0 \leq \text{Var}[a_1 X_1 + \dots + a_n X_n] = a \Sigma a^\top. \quad (28)$$

0.7 Some special distributions

The Bernoulli distribution $\text{Bernoulli}(p)$ is the distribution P_p on $\{0, 1\}$ with

$$P_p[\{1\}] = 1 - P_p[\{0\}] = p. \quad (29)$$

It describes e.g. random tossing of a (unfair) coin.

The binomial distribution $\text{binomial}(n, p)$ is the distribution of $X_1 + \dots + X_n$, where $(X_1, \dots, X_n) \sim \text{Bernoulli}(p)^n$. The binomial distribution has expectation np and variance $np(1-p)$.

More generally, consider a (possibly unfair) die with m sides, labeled by the numbers $1, \dots, m$. Assume it is rolled n times, giving independent random results $X_1, \dots, X_n$ with values in $\{1, \dots, m\}$. Let $p_i = P[X_j = i]$ denote the probability to get the result i in the j -th experiment; it is supposed not to depend on j . The sum

$$N_i = \sum_{j=1}^n 1_{\{X_j=i\}} \quad (30)$$

denotes the (random) number of results “ i ” in the random experiment. The *multinomial distribution* $\text{multinomial}(n, p_1, \dots, p_m)$ is defined to be the distribution of the random vector $(N_1, \dots, N_m)$.

The *Poisson distribution* $\text{Poisson}(\lambda)$ with parameter $\lambda > 0$ is the distribution on $\mathbb{N}_0$ specified by the probabilities

$$p_n = e^{-\lambda} \frac{\lambda^n}{n!}. \quad (31)$$

It appears as the limit of $\text{binomial}(n, p)$ distributions as $n \rightarrow \infty$ and $p \rightarrow 0$ with $np \rightarrow \lambda$. The Poisson distribution with parameter λ has expectation and variance λ .

The *normal distribution* $\text{normal}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (32)$$

It has the expectation μ and the variance σ^2 . In the special case $\mu = 0$, $\sigma^2 = 1$, one obtains the *standard-normal distribution*. It has the density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad (33)$$

More generally, the *multidimensional normal distribution* $\text{normal}(\mu, \Sigma)$ with expectation vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ has the following density on $\mathbb{R}^n$:

$$f(x) = (2\pi)^{-n/2} (\det \Sigma)^{-1/2} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)}. \quad (34)$$

0.8 Jacobi matrices, chain rule, and transformation of densities

Let $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(x_1, x_2, \dots, x_n) \\ g_2(x_1, x_2, \dots, x_n) \\ \vdots \\ g_m(x_1, x_2, \dots, x_n) \end{pmatrix} \quad (35)$$

be a differentiable map. Then the *Jacobi matrix* of g at x is defined to be the matrix

$$Dg(x) := \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \dots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1}(x) & \dots & \frac{\partial g_m}{\partial x_n}(x) \end{pmatrix}. \quad (36)$$

For the composition of differentiable maps $h(x) = f(g(x))$, one has the chain rule

$$Dh(x) = Df(g(x)) \cdot Dg(x), \quad (37)$$

where “ $\cdot$ ” means matrix multiplication. If $n = m$, then the determinant $\det Dg(x)$ is called *Jacobi determinant* of g at x . In particular, one has for $h(x) = f(g(x))$:

$$\det Dh(x) = \det Df(g(x)) \cdot \det Dg(x). \quad (38)$$

If a random vector $X : \Omega \rightarrow \text{range}(X) \subseteq \mathbb{R}^n$ has the density f_X , and if $g : \text{range}(X) \rightarrow \text{range}(g(X)) \subseteq \mathbb{R}^n$ is differentiable with a differentiable inverse g^{-1} , then $Y = g(X)$ has a density $f_{g(X)}$, too. It can be expressed as follows:

$$f_{g(X)}(y) = f_X(g^{-1}(y)) |\det Dg(g^{-1}(y))|^{-1} = f_X(g^{-1}(y)) |\det Dg^{-1}(y)|. \quad (39)$$

Here is a mnemonic aid to keep this formula in mind:

$$f_Y(y) dy = f_X(x) dx \quad \Rightarrow \quad f_Y(y) = f_X(x) \left| \frac{dy}{dx} \right|^{-1} = f_X(x) \left| \frac{dx}{dy} \right|. \quad (40)$$

1 Introduction

Probability theory and statistics both deal with random phenomena. However, their points of view differ:

	Probability theory	Statistics
probability distribution P	known	unknown (only general assumptions on the class of probability distributions)
outcome ω of the random experiment	unknown	known (observed data)
typical problems	calculating (or estimating) probabilities of interesting events	testing hypotheses about the unknown distribution, estimating parameters of the unknown distribution.

One could phrase it like this:

Statistics is concerned with inverse problems of probability theory:
 One wants to recover information about unknown probability distributions from observed data.
 The data is interpreted as an observed value of a random variable.

Example 1: A (possibly biased) coin is tossed randomly 100 times. We observe that it shows $\omega = 63$ times “head”.

- **Choice of a probabilistic frame model.** We make some general assumptions for this random experiment. Natural assumptions could be:

- The 100 repetitions are done independently of each other;
- For each of the 100 repetitions, the (unknown) probability p of “head” to occur is the same.
- From these assumptions we conclude:

Formalization of the frame model. (We ignore the observed value at this moment.) The random outcome ω of the number of observed “heads” is binomially(100, p)-distributed with an unknown parameter $p \in [0, 1]$. This means: If P_p describes the underlying probability measure with a fixed, but unknown parameter p , then

$$P_p[\omega = k] = \binom{100}{k} p^k (1-p)^{100-k}, \quad 0 \leq k \leq 100.$$

• **Typical problems:**

- Estimate the unknown parameter p .
- Is it plausible that the coin might be unbiased, i.e. $p = 0.5$?
- Find an interval of “plausible” values of p .

Example 2: A sample of radioactive uranium 238 is observed during one minute with a Geiger counter. One observes $\omega = 2173$ α -particles during this time interval.

- **Frame model:** Natural assumptions on this random experiment could be: The random sample consists of extremely many radioactive nuclei (order of magnitude: 10^{24} nuclei), each of which has an extremely small probability to decay and gives rise to an observed α -particle in the observation interval, independently of all other nuclei. Thus, approximating a binomial(n, p) distribution with extremely large (unknown) n and extremely low (unknown) p by a Poisson distribution, it is natural to assume that ω has a Poisson distribution P_λ with an unknown parameter $\lambda > 0$. For fixed, but unknown λ , this means

$$P_\lambda[\omega = k] = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0.$$

• **Typical problems:**

- Is it plausible that $\lambda = 2000$?
- Find an interval of plausible values for λ .
- Estimate λ .

Example 3: 30 patients suffering from the same disease were randomly divided into 2 groups. Group A (20 members) then was treated with drug A , group B (10 members) with drug B . After 1 week, $N_A = 19$ patients from group A were healthy, while only $N_B = 7$ people from group B were healthy. We would like to know: Is there a difference in the efficiency of the two drugs?

- **Frame model:** We assume: The reaction of the patients is independent of each other, with the same distribution for all people in group A , and (maybe) another distribution for all people in group B .

Thus we assume:

$$\omega = (N_A, N_B) \sim \text{binomial}(20, p_A) \otimes \text{binomial}(10, p_B),$$

with unknown parameters p_A and p_B . This means that the common distribution of N_A and N_B is a product of two different binomial distributions. $P \otimes Q$ denotes a two-dimensional joint distribution of independent random variables with marginals P and Q .

- **Typical problem:** Is $p_A = p_B$ plausible?

General abstract setup:

A *frame model* is described by

- a set Ω of possible results, and
- a family $\mathcal{P}$ of probability measures $P \in \mathcal{P}$ on Ω .

The observed result is encoded in a single $\omega \in \Omega$. Interesting properties of ω are frequently encoded in the value $X(\omega)$ of a random variable $X : \Omega \rightarrow \mathbb{R}$.

Remarks:

- Recall that a probability measure on Ω is a map P that assigns a probability $P[A] \in [0, 1]$ to every event $A \subseteq \Omega$, such that Kolmogorov's axioms hold.
- In principle, one should develop the frame model *before* one takes note of the observed data $\omega \in \Omega$.
- In Example 1 above, we may choose $\Omega = \{0, 1, \dots, 100\}$, and let $\mathcal{P}$ denote the set of all binomial(100, p) distributions, where $p \in [0, 1]$.
- In Example 2 above, we choose $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$, and $\mathcal{P}$ is chosen to be the set of all Poisson distributions.

Remember:

- The observed *known* data $\omega \in \Omega$ is interpreted as the *random* value of some random experiment.
- The underlying probability measure $P \in \mathcal{P}$ is treated as fixed, *nonrandom*, but *unknown*.

Parametric and nonparametric models

A frame model $(\Omega, \mathcal{P})$ is called *parametric* if the elements of $\mathcal{P}$ are indexed by a parameter (vector) $\theta \in \Theta$ with some parameter domain $\Theta \subseteq \mathbb{R}^n$:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

Otherwise it is called *nonparametric*.

- The frame models in the examples 1, 2, and 3 above are parametric with the parameters $p \in \Theta = [0, 1]$, $\lambda \in \Theta = [0, \infty[$, and $(p_A, p_B) \in \Theta = [0, 1]^2$, respectively.
- The set of all possible joint distributions of two independent real-valued random variables (X, Y) yields a nonparametric model over $\mathbb{R}^2$.

Definition 1.1 Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ describe a parametric model on Ω .

- **discrete case:** If Ω is discrete, then the *likelihood function* is defined to be the map

$$p : \Omega \times \Theta \rightarrow \mathbb{R}, \quad p(\omega|\theta) = P_\theta[\{\omega\}].$$

- **continuous case:** If Ω is a subset of $\mathbb{R}^n$ for some n and if all P_θ have a density, then the *likelihood function* is a function $p : \Omega \times \Theta \rightarrow \mathbb{R}$, $(\omega, \theta) \mapsto p(\omega|\theta)$ such that for all $\theta \in \Theta$, $\omega \mapsto p(\omega|\theta)$ is a density of P_θ .

Thus the likelihood function encodes all probability distributions in $\mathcal{P}$ in a single function.

2 Testing Hypotheses

Recall Example 1: A (possibly biased) coin is tossed randomly 100 times. We observe that it shows 63 times “head”. Is it reasonable to assume that the coin might be unbiased? Is the true distribution $P \in \mathcal{P}$ binomial(100,1/2)?

Definition: A hypothesis is formalized to be a subset H of the family of probability measures $\mathcal{P}$ within a frame model. If $H = \{P\}$ is a singleton, the hypothesis H is called *simple*, else it is called *composite*.

When testing hypotheses, we want to decide between two alternatives:

- a *null hypothesis* $H_0 \subset \mathcal{P}$, and
- an *alternative hypothesis* $H_A \subset \mathcal{P}$, being disjoint from H_0 .

Hypotheses are frequently described informally, like “ $p = 0.5$ ” in example 1, “ $\lambda = 2000$ ” in example 2, and “ $p_A = p_B$ ” in example 3. The last hypothesis is composite, since the values of p_A and p_B are still not determined, even if the “true” distribution $P \in \mathcal{P}$ fulfills $p_A = p_B$. The other two hypotheses are simple.

Although the roles of the null hypothesis and of the alternative hypothesis here appear to be completely exchangeable, in applications they are not:

- The null hypothesis frequently encodes some “simple explanation”, like the absence of a bias, of a physical effect, of a difference in the efficiency of two drugs. In order to gain evidence for the existence of an effect, one would like to reject such simple explanations.
- The alternative hypothesis frequently is used only to measure the sensitivity of a test.

Definition:

A statistical test for the hypotheses H_0, H_A is given by a *rejection region* $R \subset \Omega$. Its complement $\Omega \setminus R$ is called the *acceptance region*.

Depending on the observed value, we take our decision:

- If we observe $\omega \in R$, we decide to reject the null hypothesis H_0 .
- If we observe $\omega \in \Omega \setminus R$, we decide to accept the null hypothesis H_0 .

If either H_0 or H_A describes the true distribution, then there are 2 types of correct decisions and 2 types of errors:

	H_0 accepted	H_0 rejected
H_0 holds	correct decision	type I error
H_A holds	type II error	correct decision

Just like with the design of a frame model, one should develop a statistical test *before* one takes note of the observed data $\omega \in \Omega$.

In example 1, $\Omega = \{0, 1, 2, \dots, 100\}$, $H_0 = \{\text{binomial}(100, 50\%)\}$, we could choose the simple alternative $H_A = \{\text{binomial}(100, 70\%)\}$ and the rejection region $R = \{60, 61, 62, 63, \dots, 100\}$. Since we observed $\omega = 63 \in R$, we decide to reject H_0 in this test. If H_0 is true, this will be a type I error.

2.1 Goals for the design of a good test

We would like to develop tests for which both, type I errors and type II errors, do not occur too often.

- On the one hand, we would like to have for all $P_0 \in H_0$ the type I risk $P_0[R]$ as small as possible.
- On the other hand, we would like to have for all $P_A \in H_A$ the type II risk $P_A[\Omega \setminus R]$ as small as possible, too.

However, this problem is ill-posed, since the two requirements contradict each other: If we choose the rejection region R to become smaller, then type I risks decrease, but type II risks increase at the same time. In order to make the problem well-posed, we search for a compromise. Here is a measure for the type I risks and type II risks:

For a test with rejection region R , the *significance level* is defined by

$$\alpha := \sup_{P_0 \in H_0} P_0[R].$$

For a given $P_A \in H_A$, the *power* $1 - \beta$ of the test for the alternative P_A is defined by

$$1 - \beta := P_A[R].$$

Example (continued):

In the above example, we have $P_0 = \text{binomial}(100, 0.5)$, $P_A = \text{binomial}(100, 0.7)$, and $R = \{60, 61, \dots, 100\}$; thus the test has the significance level

$$\alpha = P_0[R] = \frac{\sum_{k=60}^{100} \binom{100}{k}}{2^{100}} = 2.84\%$$

and the power

$$1 - \beta = P_A[R] = \sum_{k=60}^{100} \binom{100}{k} 0.7^k 0.3^{100-k} = 98.75\%.$$

Thus, the type II risk is $\beta = 1.25\%$. (The numerical values are rounded.)

Given an upper bound α^* for the significance level that we are willing to tolerate and a fixed $P_A \in H_A$, the following problem is well-posed:

Maximize the power $1 - \beta$ among all tests with significance level $\alpha \leq \alpha^*$.

Remarks:

- Since we do not want too large type I risks, the bound α^* for the significance level should be chosen rather small: Typical values for values for α^* are 5%, 1%, or even $10^{-4}\%$.
- There is an asymmetry between rejection and acceptance in a test decision:
 - if we reject, the null hypothesis is very unplausible: If some $P_0 \in H_0$ is the true model, then the probability to reject cannot be larger than the very small value α^* .
 - On the other hand, if we accept, it resembles more an abstention: We still cannot be sure whether the null hypothesis is valid, but we did not find a significant deviation.

Recall that the null hypothesis often represents the *absence* of an effect or a simple explanation. If we decide to reject, we can rule out the *absence* of an effect or a simple explanation with a high degree of plausibility; it still does not tell us what the cause of the effect was.

- This interpretation is consistent with Popper's philosophical ideas about scientific reasoning: According to him, in science, we can never verify hypotheses. All we can do is trying to falsify them. Hypotheses that could not be falsified in a long list of trials finally are called "laws of nature".

2.2 Design of optimal tests for simple hypotheses

We take a simple null hypothesis $H_0 = \{P_0\}$ and a simple alternative hypothesis $H_A = \{P_A\}$. How can we decide whether a test has the maximal power among all tests with the same or a lower significance level α , and how can we find such an optimal test?

In order to intuitively illustrate the principle, let us introduce an analogy: We would like to pack a knapsack such that it is not heavier than what we can carry, but the value of its contents should be maximal. Here is a table for the translation in this analogy:

statistical tests	analogy: packing a knapsack
set of possible results Ω	set of goods at our disposal
rejection region R	goods that we pack into the knapsack
significance level $P_0[R]$	total mass of the goods $R \subseteq \Omega$
power $P_A[R]$	total value of the goods $R \subseteq \Omega$
bound α^* for the significance level	maximal mass that we are able to carry
likelihood ratio $\frac{P_A[\{\omega\}]}{P_0[\{\omega\}]}$	value per unit mass of the good $\omega \in \Omega$

Example: Suppose we have the following goods at our disposal:

good	value	mass	value per unit mass
Gold	1000 Euro	0.1 kg	$10000 \frac{\text{Euro}}{\text{kg}}$
Silver	500 Euro	1 kg	$500 \frac{\text{Euro}}{\text{kg}}$
Iron	100 Euro	10 kg	$10 \frac{\text{Euro}}{\text{kg}}$
Stones	200 Euro	10000 kg	$0.02 \frac{\text{Euro}}{\text{kg}}$

Every child would proceed as follows: Pack as much of the gold as possible to your knapsack. If you can

carry even more, take the silver. If you can carry even more, take the iron. If you are still not overloaded, take some of the stones.

Thus, a child knows: If you can carry 11.1 kilograms, take all the gold, all the silver, and all the iron, but no stones. Then the content of your knapsack will be worth 1600 Euros. If you pack it in a different way without exceeding the 11.1 kilograms that you can carry, it will be worth less. Better do not replace gold by stones!

Let us transfer this analogy to the design of statistical tests:

Neyman-Pearson Lemma; discrete version.

Suppose $\Omega = \{\omega_1, \dots, \omega_n\}$. Let the null hypothesis $H_0 = \{P_0\}$ and the alternative hypothesis $H_A = \{P_A\}$ be described by the probabilities

$$p_0(\omega) = P_0[\{\omega\}] > 0 \quad \text{and} \quad p_A(\omega) = P_A[\{\omega\}] \geq 0, \quad (\omega \in \Omega).$$

Assume that the elements $\omega \in \Omega$ are ordered in decreasing order according to their *likelihood ratio* $p_A(\omega)/p_0(\omega)$:

$$\frac{p_A(\omega_1)}{p_0(\omega_1)} \geq \frac{p_A(\omega_2)}{p_0(\omega_2)} \geq \dots \geq \frac{p_A(\omega_n)}{p_0(\omega_n)}.$$

For $1 \leq k \leq n$, consider the test T with rejection region $R = \{\omega_1, \dots, \omega_k\}$.

Then every other test T' with a rejection region $R' \subseteq \Omega$ and a smaller or equal significance level

$$P_0[R'] \leq P_0[R]$$

has a smaller or equal power

$$P_A[R'] \leq P_A[R].$$

Thus T is optimal (in the sense that it has maximal power) among all tests T' with the same or a smaller significance level.

Although the lemma is intuitively clear by the analogy to packing a knapsack, we will prove a more general version of the Neyman-Pearson lemma below.

Example 2.1 A (possibly biased) coin is tossed n times. It shows ω times “head”. Given fixed numbers $0 < \pi_A < \pi_0 < 1$, we want to design optimal tests for testing the null hypothesis $H_0 = \{P_0\} = \{\text{binomial}(n, \pi_0)\}$ versus the alternative hypothesis $H_A = \{P_A\} = \{\text{binomial}(n, \pi_A)\}$. We calculate the likelihood ratio for $\omega \in \Omega = \{0, 1, \dots, n\}$:

$$\frac{p_A(\omega)}{p_0(\omega)} = \frac{\binom{n}{\omega} \pi_A^\omega (1 - \pi_A)^{n-\omega}}{\binom{n}{\omega} \pi_0^\omega (1 - \pi_0)^{n-\omega}} = \left(\frac{\pi_A}{\pi_0} \right)^\omega \left(\frac{1 - \pi_0}{1 - \pi_A} \right)^{\omega - n}.$$

Since we assumed $0 < \pi_A < \pi_0 < 1$, we know

$$\frac{\pi_A}{\pi_0} < 1$$

and

$$\frac{1 - \pi_0}{1 - \pi_A} < 1;$$

thus the likelihood ratio

$$\omega \mapsto \frac{p_A(\omega)}{p_0(\omega)}$$

is monotonically decreasing. Hence the Neyman-Pearson lemma tells us that the tests with rejection region $R_k = \{0, 1, 2, \dots, k\} \subseteq \Omega$, $0 \leq k \leq n$, are all optimal in the following sense: For all other tests (rejection region R') with a lower or equal significance level

$$P_0[R'] = \sum_{\omega \in R'} \binom{n}{\omega} \pi_0^\omega (1 - \pi_0)^{n-\omega} \leq P_0[R_k] = \sum_{\omega=0}^k \binom{n}{\omega} \pi_0^\omega (1 - \pi_0)^{n-\omega},$$

the power is lower or at most equal, too:

$$P_A[R'] = \sum_{\omega \in R'} \binom{n}{\omega} \pi_A^\omega (1 - \pi_A)^{n-\omega} \leq P_A[R_k] = \sum_{\omega=0}^k \binom{n}{\omega} \pi_A^\omega (1 - \pi_A)^{n-\omega}.$$

Note that for a given π_0 , the test with rejection region R_k is optimal for all $\pi_A < \pi_0$; we say that it is *uniformly most powerful* for the composite alternative hypothesis

$$H_A = \{\text{binomial}(n, \pi_A) : 0 < \pi_A < \pi_0\}.$$

On the other hand, if we assume $0 < \pi_0 < \pi_A < 1$, we know

$$\frac{\pi_A}{\pi_0} > 1$$

and

$$\frac{1 - \pi_0}{1 - \pi_A} > 1;$$

in this case the likelihood ratio

$$\omega \mapsto \frac{p_A(\omega)}{p_0(\omega)}$$

is monotonically increasing. This time, we reverse the order of the elements of Ω , i.e. we arrange them according to their likelihood ratio dP_A/dP_0 in decreasing order. Now the Neyman-Pearson lemma tells us that the tests with rejection region $\tilde{R}_k = \{n, n-1, \dots, k\} \subseteq \Omega$, $0 \leq k \leq n$, are all optimal in the following sense: For all other tests (rejection region R') with a lower or equal significance level

$$P_0[R'] \leq P_0[\tilde{R}_k],$$

the power is lower or at most equal, too:

$$P_A[R'] \leq P_A[\tilde{R}_k].$$

General likelihood ratios. We would like to generalize the Neyman-Pearson lemma to cover continuous distributions, too. In order to do this, we generalize the notion of “likelihood ratio”:

Let P_0 and P_A be two probability measures on Ω ; their expectation operators are denoted by E_0 and E_A , respectively. The indicator function of an event $B \subseteq \Omega$ is denoted by $1_B : \Omega \rightarrow \{0, 1\}$:

$$1_B(\omega) = \begin{cases} 1 & \text{for } \omega \in B, \\ 0 & \text{for } \omega \notin B. \end{cases}$$

P_A is called *absolutely continuous* with respect to P_0 , if there is a random variable $L : \Omega \rightarrow [0, \infty[$ such that for all events $B \subseteq \Omega$ we have

$$P_A[B] = E_0[L1_B]. \quad (41)$$

L is called the *likelihood ratio* between P_A and P_0 or the *Radon-Nikodym derivative* of P_A with respect to P_0 , and we write

$$L = \frac{dP_A}{dP_0}.$$

More generally than (41),

$$E_A[X] = E_0 \left[X \frac{dP_A}{dP_0} \right] \quad (42)$$

holds for every random variable X for which at least one of the two sides is well defined.

The likelihood ratio is just a ratio of likelihood functions, as we will see now:

For discrete spaces $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$, $P_0[\{\omega\}] = p_0(\omega) > 0$, $P_A[\{\omega\}] = p_A(\omega)$ this notion of likelihood ratios just coincides with the previous one, as the following calculation shows:

$$E_A[X] = \sum_{\omega \in \Omega} X(\omega) p_A(\omega) = \sum_{\omega \in \Omega} X(\omega) \frac{p_A(\omega)}{p_0(\omega)} p_0(\omega) = E_0 \left[X \frac{dP_A}{dP_0} \right]. \quad (43)$$

This shows that (42) indeed holds for the likelihood ratio

$$\frac{dP_A}{dP_0}(\omega) = \frac{p_A(\omega)}{p_0(\omega)}. \quad (44)$$

On the other hand, if P_A and P_0 are both continuous probability distributions over $\Omega = \mathbb{R}^n$ with densities f_A and $f_0 > 0$, respectively, then the likelihood ratio is just the ratio of the densities:

$$\frac{dP_A}{dP_0}(\omega) = \frac{f_A(\omega)}{f_0(\omega)}. \quad (45)$$

Recall that a probability measure P_0 on $\Omega = \mathbb{R}^n$ is said to have the density f_0 if

$$P_0[B] = \int_B f_0(x) dx$$

holds for all events $B \subseteq \mathbb{R}^n$. In this case,

$$E_0[X] = \int_{\mathbb{R}^n} X(x) f_0(x) dx$$

holds for every random variable $X : \Omega \rightarrow \mathbb{R}$ for which the expectation exists.

To prove (45), we do a little calculation similar to (43):

$$E_A[X] = \int_{\Omega} X(x) f_A(x) dx = \int_{\Omega} X(x) \frac{f_A(x)}{f_0(x)} f_0(x) dx = E_0 \left[X \frac{f_A}{f_0} \right]. \quad (46)$$

We have prepared all tools to phrase a more general version of the Neyman-Pearson lemma:

Neyman-Pearson Lemma; general version.

Given a null hypothesis $H_0 = \{P_0\}$ and an alternative hypothesis $H_A = \{P_A\}$, assume that the likelihood ratio dP_A/dP_0 exists. Let T be a test with rejection region R such that there is $c > 0$ with

$$R \subseteq \left\{ \frac{dP_A}{dP_0} \geq c \right\} \quad \text{and} \quad \Omega \setminus R \subseteq \left\{ \frac{dP_A}{dP_0} \leq c \right\}. \quad (47)$$

Then T is optimal in the following sense: For every other test T' (rejection region $R' \subseteq \Omega$) with a lower or equal significance level

$$P_0[R'] \leq P_0[R],$$

the power is lower or at most equal, too:

$$P_A[R'] \leq P_A[R].$$

Recall that $\left\{ \frac{dP_A}{dP_0} \geq c \right\}$ is a short notation for $\left\{ \omega \in \Omega : \frac{dP_A}{dP_0}(\omega) \geq c \right\}$.

Proof of the Neyman-Pearson lemma:

Given another rejection region R' with $P_0[R'] \leq P_0[R]$, we conclude:

$$P_0[R \setminus R'] - P_0[R' \setminus R] = P_0[R] - P_0[R'] \geq 0. \quad (48)$$

Using that $dP_A/dP_0 \geq c$ holds on the event $R \setminus R'$, and that $dP_A/dP_0 \leq c$ holds on the event $R' \setminus R$, we obtain

$$E_0 \left[1_{R \setminus R'} \frac{dP_A}{dP_0} \right] \geq E_0 [1_{R \setminus R'} c] = cP_0[R \setminus R']$$

and

$$E_0 \left[1_{R' \setminus R} \frac{dP_A}{dP_0} \right] \leq E_0 [1_{R' \setminus R} c] = cP_0[R' \setminus R],$$

thus

$$\begin{aligned} P_A[R] - P_A[R'] &= P_A[R \setminus R'] - P_A[R' \setminus R] \\ &= E_0 \left[1_{R \setminus R'} \frac{dP_A}{dP_0} \right] - E_0 \left[1_{R' \setminus R} \frac{dP_A}{dP_0} \right] \\ &\geq cP_0[R \setminus R'] - cP_0[R' \setminus R] \geq 0. \end{aligned}$$

We used (48); recall that $c > 0$. Hence we get

$$P_A[R] \geq P_A[R'],$$

which proves the Neyman-Pearson lemma. □

As a consequence of the Neyman-Pearson lemma, we obtain:

Likelihood ratio tests.

The tests with rejection region

$$R := \left\{ \frac{dP_A}{dP_0} > c \right\} = \left\{ \frac{dP_0}{dP_A} < c^{-1} \right\}. \quad (49)$$

have maximal power among all tests with lower or equal significance level.

This test, described by (49), is called *likelihood ratio test*.

Of course, the right hand side in (49) only makes sense if dP_0/dP_A exists, too.

Example 2.2 Assume that X has a normal distribution with unknown parameters μ and σ^2 ; we observe a single value of X . We want to find likelihood ratio tests at a given significance level α for the simple null hypothesis

$$H_0 = \{P_0\} = \{\text{normal}(\mu_0 = 0, \sigma_0^2 = 1)\} = \{\text{standard-normal}\}$$

versus various simple alternative hypotheses

$$H_A = \{P_A\} = \{\text{normal}(\mu, \sigma^2)\}.$$

Recall that the normal distribution with parameters μ and σ^2 has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Thus we get the likelihood ratio

$$\frac{dP_0}{dP_A}(x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} = \sigma \exp \left\{ \frac{(x-\mu)^2}{2\sigma^2} - \frac{x^2}{2} \right\}.$$

Let us consider some special cases:

- If $\mu > 0$ and $\sigma^2 = 1$, then

$$\frac{dP_0}{dP_A}(x) = \exp \left\{ \frac{(x-\mu)^2}{2} - \frac{x^2}{2} \right\} = \exp \left\{ \frac{\mu^2}{2} - \mu x \right\} \quad (50)$$

is monotonically *decreasing* in x . Thus the level sets $\{dP_0/dP_A < c^{-1}\}$ are intervals of the form $]t, \infty[$. Choosing the unique intervals among these with probability α with respect to the standard normal distribution, we obtain the rejection region

$$R =]\Phi^{-1}(1 - \alpha), \infty[,$$

where

$$\Phi(t) = \frac{1}{2\pi} \int_{-\infty}^t e^{-\frac{x^2}{2}} dx$$

denotes the cumulative distribution function of the standard normal distribution.

- If $\mu < 0$ and $\sigma^2 = 1$, then (50) is monotonically *increasing* in x . This time we get the rejection region

$$R =]-\infty, \Phi^{-1}(\alpha)[.$$

- If $\mu = 0$ and $\sigma^2 > 1$, then

$$\frac{dP_0}{dP_A}(x) = \sigma \exp \left\{ (\sigma^{-2} - 1) \frac{x^2}{2} \right\}, \quad (51)$$

which is monotonically decreasing in $|x|$. This time, the level sets $\{dP_0/dP_A < c^{-1}\}$ are unions of intervals of the form $] -\infty, -t[\cup]t, \infty[$. Choosing the unique one among these sets with probability α with respect to the standard normal distribution, we obtain the rejection region

$$R =]-\infty, \Phi^{-1}\left(\frac{\alpha}{2}\right)[\cup]\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \infty[.$$

- Finally, if $\mu = 0$ and $\sigma^2 < 1$, then (51) is monotonically increasing in $|x|$. Now we get the rejection region

$$R = \left] \Phi^{-1} \left(\frac{1}{2} - \frac{\alpha}{2} \right), \Phi^{-1} \left(\frac{1}{2} + \frac{\alpha}{2} \right) \right[.$$

This shows how the rejection region depends on the choice of the alternative hypothesis.

Example 2.3 Let $X_1, X_2, \dots, X_n$ be i.i.d. (independent identically distributed) normal random variables with an unknown expectation μ and variance $\sigma^2 = 1$. Formally speaking, we use the frame model with the sample space $\Omega = \mathbb{R}^n$ and the family of probability measures

$$\mathcal{P} = \{\text{normal}(\mu, \sigma^2 = 1)^n : \mu \in \mathbb{R}\}.$$

(P^n denotes the joint distribution of n independent random variables with all marginal distributions equal to P .) We are going to develop a likelihood ratio test at significance level α for the null hypothesis

$$H_0 = \{P_0\} = \{\text{standard-normal}^n\} = \{\text{normal}(0, 1)^n\}$$

versus the alternative hypothesis

$$H_A = \{P_\mu\} = \{\text{normal}(\mu, 1)^n\}$$

for a given $\mu \neq 0$. Using the abbreviation $x = (x_1, x_2, \dots, x_n)$, we recall that P_0 and P_μ have the densities

$$f(x|0) = \prod_{j=1}^n \frac{e^{-x_j^2/2}}{\sqrt{2\pi}} = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{j=1}^n x_j^2 \right\}$$

and

$$f(x|\mu) = \prod_{j=1}^n \frac{e^{-(x_j - \mu)^2/2}}{\sqrt{2\pi}} = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{j=1}^n (x_j - \mu)^2 \right\},$$

respectively. Thus their likelihood ratio equals

$$\frac{dP_\mu}{dP_0}(x) = \frac{f(x|\mu)}{f(x|0)} = \exp \left\{ \frac{1}{2} \sum_{j=1}^n (x_j^2 - (x_j - \mu)^2) \right\} = \exp \left\{ -\frac{n\mu^2}{2} + \mu \sum_{j=1}^n x_j \right\}.$$

Note that we do not need to know the whole data $x = (x_1, \dots, x_n)$ to determine this likelihood ratio; knowing the sum $\sum_{j=1}^n x_j$ suffices.

Let us determine the level sets

$$R = \left\{ \frac{dP_\mu}{dP_0} > c \right\},$$

which are the rejection regions for likelihood ratio tests.

We distinguish two cases:

- If $\mu > 0$, then $s \mapsto \exp\{-n\mu^2/2 + \mu s\}$ is a monotonically increasing function, thus

$$R = \left\{ \frac{dP_\mu}{dP_0} > c \right\} = \left\{ \sum_{j=1}^n X_j > s \right\},$$

where we choose $c = \exp\{-n\mu^2/2 + \mu s\}$. In order to get the significance level α , we choose s such that

$$P_0 \left[\sum_{j=1}^n X_j > s \right] = \alpha.$$

- If $\mu < 0$, then $s \mapsto \exp\{-n\mu^2/2 + \mu s\}$ is a monotonically decreasing function, thus

$$R = \left\{ \frac{dP_\mu}{dP_0} > c \right\} = \left\{ \sum_{j=1}^n X_j < s \right\},$$

where we choose c as above. In order to get again the significance level α , we choose s this time such that

$$P_0 \left[\sum_{j=1}^n X_j < s \right] = \alpha.$$

By the following theorem, we know the distribution of $\sum_{j=1}^n X_j$ with respect to the null distribution P_0 :

Theorem 2.4 (Convolution of normal distributions) *If $X_1, \dots, X_n$ are independent normal random variables with expectations $\mu_1, \dots, \mu_n$ and variances $\sigma_1^2, \dots, \sigma_n^2$, then $\sum_{j=1}^n X_j$ is a normal random variable with the expectation $\sum_{j=1}^n \mu_j$ and the variance $\sum_{j=1}^n \sigma_j^2$.*

We will prove a more general statement later in this lecture.

With respect to the null distribution P_0 , the X_j are i.i.d. standard normal random variables. Thus, according to the theorem, $\sum_{j=1}^n X_j$ is a normal($\mu = 0, \sigma^2 = n$) distributed random variable with respect to P_0 ; thus

$$T = \frac{1}{\sqrt{n}} \sum_{j=1}^n X_j \tag{52}$$

is again a standard normal random variable with respect to P_0 (see section 2.3, Proposition A, in the textbook).

Thus

- **In the case $\mu > 0$:**

$$\alpha = P_0 \left[\sum_{j=1}^n X_j > s \right] = P_0 \left[T > \frac{s}{\sqrt{n}} \right] = 1 - \Phi \left(\frac{s}{\sqrt{n}} \right);$$

recall that Φ denotes the cumulative distribution function of the standard normal distribution. Thus, the likelihood ratio test at significance level α has the rejection region

$$R = \{T > \Phi^{-1}(1 - \alpha)\}.$$

- **In the case $\mu < 0$:**

$$P_0 \left[\sum_{j=1}^n X_j < s \right] = P_0 \left[T < \frac{s}{\sqrt{n}} \right] = \Phi \left(\frac{s}{\sqrt{n}} \right).$$

Thus, the likelihood ratio test at significance level α has the rejection region

$$R = \{T < \Phi^{-1}(\alpha)\}.$$

In this example, we only need to know T in order to determine the test decision at any given significance level α .

A random variable $T : \Omega \rightarrow \mathbb{R}$ which determines the rejection region R_α for every choice of the significance level α ,

$$R_\alpha = \{T \in \tilde{R}_\alpha\}$$

for some region $\tilde{R}_\alpha \subseteq \mathbb{R}$, is called *test statistic*. The region $\tilde{R}_\alpha \subseteq \mathbb{R}$ is called *rejection region* for the test statistic, too.

More generally, a random variable $X : \Omega \rightarrow \mathbb{R}$, which assigns a real number $X(\omega)$ to the observed data $\omega \in \Omega$, is also called a *statistic*.

For example, the likelihood ratio dP_A/dP_0 is a test statistic for likelihood ratio tests with null hypothesis $H_0 = \{P_0\}$ and alternative hypothesis $H_A = \{P_A\}$.

2.3 Varying the significance level: p -values.

In the last examples, we constructed not only one test, but a whole family of tests with varying significance levels. Given a fixed null hypothesis H_0 and a fixed alternative hypothesis H_A , we consider a family of tests with varying significance levels.

Speaking very roughly, the p -value is the smallest significance level at which the null hypothesis is rejected.

More precisely, we proceed as follows: We specify the family of tests by a family of rejection regions $\mathcal{R}$, every $R \in \mathcal{R}$ being an event $R \subseteq \Omega$. Assume that $\mathcal{R}$ is linearly ordered by inclusion, i.e.

$$R, R' \in \mathcal{R} \text{ implies } R \subseteq R' \text{ or } R' \subseteq R.$$

As a typical example, one could think of $\mathcal{R}$ being the set of all events $\{T < c\}$ for some given test statistic T , where $c \in \mathbb{R}$.

Definition 2.5 Given $\mathcal{R}$ as above and observed data $\omega \in \Omega$, we define the p -value for ω by

$$p = p(\omega) = \inf \left\{ \sup_{P_0 \in H_0} P_0[R] : R \in \mathcal{R}, \omega \in R \right\}. \quad (53)$$

Thus the p -value is the infimum of the significance levels of all tests in the given family for which the null hypothesis is rejected for the observed data ω .

In particular, we can phrase the test decision in terms of the p -value: Take a test with rejection region $R \in \mathcal{R}$ having the significance level $\alpha = \sup_{P_0 \in H_0} P_0[R]$.

- When $p < \alpha$, we *reject* the null hypothesis at significance level α .
- When $p > \alpha$, we *do not reject* the null hypothesis at significance level α .

In the borderline case $p = \alpha$, this criterion does not decide whether or not to reject the null hypothesis.

- Indeed, when $p < \alpha$, then there is $R' \in \mathcal{R}$ with $\omega \in R'$ and $\sup_{P_0 \in H_0} P_0[R'] < \alpha$. Since $\sup_{P_0 \in H_0} P_0[R] = \alpha$, we conclude $R \supseteq R'$, since $\mathcal{R}$ is linearly ordered. Thus $\omega \in R$ holds, too. Thus the null hypothesis is rejected for the test with rejection region R .
- On the other hand, when $p > \alpha$, we conclude $\omega \notin R$. Indeed, $\sup_{P_0 \in H_0} P_0[R] = \alpha$ and $p > \alpha$ imply that R cannot be one of the rejection regions in (53) over which the infimum in (53) is taken. Thus the null hypothesis is not rejected in this case for the test with rejection region R .

Suppose the rejection regions are given by

$$R = \{T < c\}$$

for some test statistic $T : \Omega \rightarrow \mathbb{R}$, where $c \in \mathbb{R}$. Assume we have observed the data $\omega \in \Omega$. Then the p -value equals

$$p = \inf_{c > T(\omega)} \sup_{P_0 \in H_0} P_0[T < c].$$

In almost all practical situations,

$$\inf_{c > T(\omega)} \sup_{P_0 \in H_0} P_0[T < c] = \sup_{P_0 \in H_0} P_0[T \leq T(\omega)].$$

In particular, this holds automatically whenever H_0 is simple, or, more generally, when the distribution of the test statistic T is the same with respect to all $P_0 \in H_0$. In this case we have:

For the rejection domains $R = \{T < c\}$, $c \in \mathbb{R}$, we have the following p -value at the observed data $\omega \in \Omega$:

$$p = \sup_{P_0 \in H_0} P_0[T \leq T(\omega)]. \quad (54)$$

Similarly one obtains:

For the rejection domains $R = \{T > c\}$, $c \in \mathbb{R}$, we have the following p -value at the observed data $\omega \in \Omega$:

$$p = \sup_{P_0 \in H_0} P_0[T \geq T(\omega)]. \quad (55)$$

Examples:

- In example 2.3 above, the test statistic T follows a standard normal distribution with respect to the null hypothesis. We get:

- In the case $\mu > 0$, the rejection domains are $R = \{T > c\}$, $c \in \mathbb{R}$, thus the p -value for given data $\omega \in \Omega$ equals

$$p = P_0[T \geq T(\omega)] = 1 - \Phi(T(\omega)).$$

- In the case $\mu < 0$, the rejection domains are $R = \{T < c\}$, $c \in \mathbb{R}$, thus the p -value for data $\omega \in \Omega$ equals

$$p = P_0[T \leq T(\omega)] = \Phi(T(\omega)).$$

- In example 2.1 above, the p -value in the case $\pi_A < \pi_0$ for the observed number ω of heads equals

$$p = \sum_{j=0}^{\omega} \binom{n}{j} \pi_0^j (1 - \pi_0)^{n-j}.$$

2.4 Varying the alternative distribution P_A : Uniformly most powerful tests

We have seen in example 2.2 above that the rejection region for optimal tests may strongly depend on the choice of the alternative hypothesis $H_A = \{P_A\}$. However, in some important cases it does not depend on this choice:

Lemma 2.6 (Uniformly most powerful tests for classes with monotone likelihood ratios)

Consider a one-parametric frame model

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

with some parameter domain $\Theta \subseteq \mathbb{R}$. Assume that there is a statistic $T : \Omega \rightarrow \mathbb{R}$ such that for all $\theta_0 < \theta_A$ in Θ the likelihood ratio $dP_{\theta_A}/dP_{\theta_0}$ is a *monotonically increasing* function of T , i.e.

$$\frac{dP_{\theta_A}}{dP_{\theta_0}} = f_{\theta_0, \theta_A}(T)$$

for some *monotonically increasing* function $f_{\theta_0, \theta_A} : \mathbb{R} \rightarrow \mathbb{R}$.

Given $\theta_0 \in \Theta$, consider the null hypothesis

$$H_0 = \{P_{\theta_0}\}$$

versus the alternative hypothesis

$$H_A = \{P_{\theta_A} : \theta_A > \theta_0\}$$

and a test with rejection region

$$R = \{T > c\}.$$

Then *for all* $P_{\theta_A} \in H_A$, this test has maximal power among all tests with the same or a lower significance level. This means: For all other events $R' \subseteq \Omega$:

$$\text{If } P_{\theta_0}[R'] \leq P_{\theta_0}[R], \text{ then } P_{\theta_A}[R'] \leq P_{\theta_A}[R].$$

We say that the test given by $R = \{T > c\}$ is a *uniformly most powerful test* for testing $H_0: \theta = \theta_0$ versus $H_A: \theta > \theta_0$

Of course, a similar version of the lemma for testing $H_0: \theta = \theta_0$ versus $H_A: \theta < \theta_0$ exists, too. Phrasing this version explicitly is left as an exercise.

Proof: The lemma is an immediate consequence of the Neyman-Pearson Lemma: Let $q := f_{\theta_0, \theta_A}(c)$ and $P_{\theta_A} \in H_A$. Then

$$R = \{T > c\} \subseteq \{f_{\theta_0, \theta_A}(T) \geq f_{\theta_0, \theta_A}(c)\} = \left\{ \frac{dP_{\theta_A}}{dP_{\theta_0}} \geq q \right\}$$

and

$$\Omega \setminus R = \{T \leq c\} \subseteq \{f_{\theta_0, \theta_A}(T) \leq f_{\theta_0, \theta_A}(c)\} = \left\{ \frac{dP_{\theta_A}}{dP_{\theta_0}} \leq q \right\}.$$

The Neyman-Pearson lemma yields immediately the claim of the lemma. □

Examples:

- Reconsidering example 2.1 of tossing a coin n times, (ω times “head” observed with a binomial(n, p) distribution; p unknown), we see that the lemma applies for $H_0: p = \pi_0$ versus $H_A: p > \pi_0$ with the test statistic $T = \text{id} : \omega \mapsto \omega$, and, with appropriately reversed orders, it applies to $H_A: p < \pi_0$, too.
- Reconsidering example 2.3, we see that the lemma applies for $H_0: \mu = 0, \sigma^2 = 1$ versus $H_A: \mu > 0, \sigma^2 = 1$ with the test statistic $T = n^{-1/2} \sum_{i=1}^n X_i$.

2.5 Exponential classes

An important class, to which Lemma 2.6 applies, are one-parameter exponential classes. They are defined in terms of their likelihood functions. Recall that the likelihood function just encodes the probability distribution for a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ as follows: $p(\omega|\theta) = P_\theta[\{\omega\}]$ in the discrete case, and $\omega \mapsto p(\omega|\theta)$ is the density of P_θ in the continuous case.

Definition 2.7 One-parametric *exponential classes* with natural parameter θ are classes with likelihood functions of the following form:

$$p(\omega|\theta) = \frac{g(\omega)}{Z(\theta)} \exp\{\theta T(\omega)\}$$

for some statistics $T : \Omega \rightarrow \mathbb{R}$, $g : \Omega \rightarrow \mathbb{R}$, and some normalizing constant $Z(\theta) > 0$, $\theta \in \Theta \subseteq \mathbb{R}$. T is also called “natural statistics” for the exponential class.

More generally, n -parametric exponential classes with natural parameters θ are classes with likelihood functions of the form

$$p(\omega|\theta) = \frac{g(\omega)}{Z(\theta)} \exp\left\{\sum_{j=1}^n \theta_j T_j(\omega)\right\}$$

for some statistics $T_j : \Omega \rightarrow \mathbb{R}$, $g : \Omega \rightarrow \mathbb{R}$, and some normalizing constant $Z(\theta) > 0$, $\theta = (\theta_1, \dots, \theta_n) \in \Theta \subseteq \mathbb{R}^n$.

We calculate the likelihood ratio $dP_{\theta_A}/dP_{\theta_0}$ for P_{θ_A} , P_{θ_0} in a one-parametric exponential class:

$$\frac{dP_{\theta_A}}{dP_{\theta_0}}(\omega) = \frac{\frac{g(\omega)}{Z(\theta_A)} \exp\{\theta_A T(\omega)\}}{\frac{g(\omega)}{Z(\theta_0)} \exp\{\theta_0 T(\omega)\}} = \frac{Z(\theta_0)}{Z(\theta_A)} \exp\{(\theta_A - \theta_0)T(\omega)\},$$

which is a monotonically increasing function of $T(\omega)$ for $\theta_A > \theta_0$. We conclude:

For a one-parametric exponential class ($P_\theta : \theta \in \Theta$) with natural statistics T , the following holds:

- The tests with rejection region $R = \{T > c\}$, $c \in \mathbb{R}$, are uniformly most powerful for testing $H_0: \theta = \theta_0$ versus $H_A: \theta > \theta_0$.
- The tests with rejection region $R = \{T < c\}$, $c \in \mathbb{R}$, are uniformly most powerful for testing $H_0: \theta = \theta_0$ versus $H_A: \theta < \theta_0$.

Exponential classes appear very frequently in statistics, as the following important examples illustrate:

Example 2.8

- **Binomial distributions.** Writing

$$\binom{n}{\omega} p^\omega (1-p)^{n-\omega} = \binom{n}{\omega} (1-p)^n \exp\{(\log p - \log(1-p))\omega\},$$

we see that the binomial distributions $\text{binomial}(n, p)$ for fixed n form a one-parameter exponential class with the natural parameter $\theta = \log p - \log(1-p)$.

- **Normal distributions.** Since

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

the normal(μ, σ^2) distributions form a 2-parameter exponential class with natural parameters $\theta_1 = \mu/\sigma^2$ and $\theta_2 = -1/(2\sigma^2)$ and the natural statistics $T_1(x) = x$ and $T_2(x) = x^2$.

- **Poisson distributions.** Writing

$$\frac{\lambda^\omega}{\omega!} e^{-\lambda} = \frac{e^{-\lambda}}{\omega!} e^{\omega \log \lambda},$$

we see that the Poisson distributions on $\mathbb{N}_0$ form a one-parameter exponential class with natural parameter $\log \lambda$.

- **Exponential distributions.** The life time T for the decay of a radioactive nucleus follows an exponential distribution. It is characterized by the property that for every $t \geq 0$, the remaining lifetime $T - t$ after time t of the nucleus, conditioned on not being decayed up to time t , has the same probability law as the total lifetime T . (This is why the exponential distribution is called “memory less”.) Similarly, the law for the time of the first decay in a sample of radioactive material is exponentially distributed. The exponential distribution $\text{Exponential}(\lambda)$ has the density

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad \lambda > 0,$$

thus it is obviously an exponential family.

- **Geometric distributions.** This is a discrete variant of the exponential distribution. The geometric distribution is supported on $\mathbb{N}_0$, it has the probability function

$$p(n, q) = (1 - q)q^n$$

for $n \in \mathbb{N}_0$, where $0 < q < 1$. It has a similar “memory less” property as the exponential distributions, only continuous times $t \geq 0$ are replaced by discrete times $n \in \mathbb{N}_0$. For example, assume that a (possibly biased) coin is tossed repeatedly, but independently. Let X denote the time of the first appearance of “head”. Then X has a geometric distribution, where $1 - q$ be the probability for “head” at a fixed time. Obviously, the geometric distributions are a exponential class with natural parameter $\log q$.

- **Gamma distributions.** Given a sample of extremely many radioactive nuclei, each of which has a very low probability to decay in a given time interval, a good approximation for the waiting time for the n -th decay of a particle is the convolution of n exponential distributions with some parameter λ . We call this the $\text{Gamma}(n, \lambda)$ distribution. Written formally:

$\underbrace{\text{Exponential}(\lambda) * \dots * \text{Exponential}(\lambda)}_{n \text{ factors}} = \text{Gamma}(n, \lambda)$

More generally, the $\text{Gamma}(s, \lambda)$ distribution with $s, \lambda \in \mathbb{R}^+$ is defined to have the following density on $\mathbb{R}^+$:

$$f(x|s, \lambda) = \frac{\lambda^s}{\Gamma(s)} e^{-\lambda x} x^{s-1}, \quad x > 0$$

where

$$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$$

denotes the Gamma function. From its definition, we see that the $\text{Gamma}(s, \lambda)$ distributions form a two-parameter exponential family with natural parameters $-\lambda$ and $s - 1$. Here are two important facts about Gamma distributions:

– **Convolution property:**

$$\text{Gamma}(s, \lambda) * \text{Gamma}(t, \lambda) = \text{Gamma}(s + t, \lambda)$$

If $X \sim \text{Gamma}(s, \lambda)$ and $Y \sim \text{Gamma}(t, \lambda)$ are independent, then $X + Y \sim \text{Gamma}(s + t, \lambda)$. To see this, we explicitly determine the convolution of the densities of X and Y : Recall that if X and Y are independent random variables with the probability densities f_X and f_Y , respectively, then $X + Y$ has the convolution of f_X and f_Y as its density, given by

Convolution of probability densities

$$f_{X+Y}(z) = (f_X * f_Y)(z) = \int_{\mathbb{R}} f_X(x) f_Y(z - x) dx$$

For Gamma distributions, we get for $z > 0$:

$$\begin{aligned} f_{X+Y}(z) &= \frac{\lambda^s \lambda^t}{\Gamma(s) \Gamma(t)} \int_{\mathbb{R}} 1_{\{x>0\}} e^{-\lambda x} x^{s-1} 1_{\{z-x>0\}} e^{-\lambda(z-x)} (z-x)^{t-1} dx \\ &= \frac{\lambda^s \lambda^t}{\Gamma(s) \Gamma(t)} e^{-\lambda z} \int_0^z x^{s-1} (z-x)^{t-1} dx \\ &= \frac{\lambda^s \lambda^t}{\Gamma(s) \Gamma(t)} e^{-\lambda z} z^{s+t-1} \int_0^1 u^{s-1} (1-u)^{t-1} du, \end{aligned} \quad (56)$$

thus $f_{X+Y}(z)$ is proportional to $e^{-\lambda z} z^{s+t-1}$, which is proportional to the density of a $\text{Gamma}(s+t, \lambda)$ distribution. Since we know that the convolution is normalized,

$$\int_{\mathbb{R}} f_{X+Y}(z) dz = 1,$$

we identify the normalizing constant as the same appearing in the density of $\text{Gamma}(s+t, \lambda)$

$$\frac{\lambda^s \lambda^t}{\Gamma(s) \Gamma(t)} \int_0^1 u^{s-1} (1-u)^{t-1} du = \frac{\lambda^{s+t}}{\Gamma(s+t)}.$$

As a side result, we have shown

$$\int_0^1 u^{s-1} (1-u)^{t-1} du = \frac{\Gamma(s) \Gamma(t)}{\Gamma(s+t)}.$$

– **Scaling property:** Let $a > 0$.

$$X \sim \text{Gamma}(s, \lambda) \quad \text{implies} \quad aX \sim \text{Gamma}(s, \lambda/a).$$

We note the special case

$$\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda).$$

- **Chi-Square distributions.** Let $X_1, \dots, X_n$ be i.i.d. standard normal random variables. The χ^2 -distribution with n degrees of freedom is defined to be the distribution of $\sum_{j=1}^n X_j^2$. It is often denoted by χ_n^2 . We have

$$\chi_n^2 = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

To prove this, we first look at the case $n = 1$: The χ_1^2 distribution has the cumulative distribution function

$$\begin{aligned}\chi_1^2[-\infty, t] &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{t}}^{\sqrt{t}} e^{-x^2/2} dx \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{t}} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s/2} s^{-1/2} ds,\end{aligned}\tag{57}$$

which is the cumulative distribution function of the Gamma(1/2, 1/2) distribution. This shows

$$\text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) = \chi_1^2.$$

As a side remark, note that we have no choice for the normalizing constant when comparing Gamma(1/2, 1/2) with χ_1^2 , since both are probability distributions. Thus we have proven as a side result:

$$\frac{1}{\sqrt{2\pi}} = \frac{(\frac{1}{2})^{1/2}}{\Gamma(\frac{1}{2})},$$

i.e.

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

We use the convolution property:

$$\text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)^{*n} = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right),$$

(the exponent meaning n -fold convolution). The statement

$$\chi_n^2 = (\chi_1^2)^{*n}$$

follows immediately from the definition of χ_n^2 . Thus we get the claim:

$$\chi_n^2 = (\chi_1^2)^{*n} = \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)^{*n} = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

As a special case of Gamma distributions, the χ_n^2 distributions form a one-parameter exponential class, too.

Example 2.9 Energy and temperature in statistical mechanics. In statistical mechanics, one considers systems of many atoms, e.g. 10^{23} atoms, interacting with each other: Consider n particles of mass m in a box B . In classical mechanics, the particles are described by their positions $q_1, \dots, q_{3n}$ (3 space coordinates for every particle) and their momenta (= mass $\times$ velocity) $p_1, \dots, p_{3n}$. A typical form of the total energy of the system is a sum of the kinetic and the potential energy

$$E((q_j), (p_j)) = \sum_{j=1}^{3n} \frac{p_j^2}{2m} + V(q_1, \dots, q_{3n})\tag{58}$$

With appropriate choices of the potential energy V , this is a reasonable model for matter. The positions and momenta typically are not known in detail; we choose a random distribution. A reasonable choice for the probability density for matter in thermal equilibrium is the *Boltzmann–Gibbs probability density*

$$f((q_j), (p_j)|\beta) = \frac{1}{Z(\beta)} \exp\{-\beta E((q_j), (p_j))\}.\tag{59}$$

Here

$$\beta = \frac{1}{k_B T} \quad (60)$$

is a parameter, $k_B = 1.38 \dots 10^{-23}$ Joule / Kelvin is Boltzmann's constant, T is the absolute temperature, and

$$Z(\beta) = \int_{\mathbb{R}^{6n}} \exp \{-\beta E((q_j), (p_j))\} dq_1 \dots dq_{3n} dp_1 \dots dp_{3n} \quad (61)$$

is a normalizing constant, traditionally called “partition sum”. The corresponding distribution is traditionally called “canonical ensemble” in statistical mechanics. (For a physical justification of this ansatz we refer to courses in statistical mechanics; here we just mention that the distribution is invariant when the particles move according to Newton's dynamics.)

Obviously, the canonical ensemble is an one-parameter exponential family with the energy E as a natural statistics and the inverse (scaled) temperature β as a natural parameter.

Here are a few consequences of the ansatz (58) and (59) for matter in thermal equilibrium:

- The momenta (p_j) are stochastically independent of the positions (q_j) , since the energy is the sum of a kinetic term, depending only on the momenta, and the potential term, depending only on the positions.
- The momenta p_j are i.i.d. normal($\mu = 0, \sigma^2 = mk_B T$) random variables. Thus the kinetic energy $\sum_j p_j^2 / (2m)$ is $k_B T / 2$ times a χ^2 distributed random variable with $3n$ degrees of freedom. The kinetic energy of a single particle is $k_B T / 2$ times a χ^2 distributed random variable with 3 degrees of freedom, i.e. a Gamma($\frac{3}{2}, \frac{1}{k_B T}$) distribution. This is the famous *Maxwell distribution* for the thermal energy of atoms.

To give a specific example, let us consider an ideal gas in an one sided infinite cylinder $[0, 1]^2 \times \mathbb{R}^+$ with a vertical constant gravitational field. At height x , an atom has the potential energy mgx , where g is the acceleration by the earth's gravitation. Thus we set in this example:

$$V(q_1, \dots, q_{3n}) = \sum_{j=1}^n mgq_{3j}.$$

For this example, the joint probability density for the particles' momenta and positions is

$$\frac{1}{Z} \left(\prod_{j=1}^{3n} e^{-p_j^2 / (2mk_B T)} \right) \prod_{j=1}^n \left(1_{[0,1]}(q_{3j-2}) \cdot 1_{[0,1]}(q_{3j-1}) \cdot e^{-mgq_{3j} / (k_B T)} \right).$$

with some normalizing constant Z . Thus the height coordinates q_{3j} , $j = 1, \dots, n$ are i.i.d. Exponential($mg / (k_B T)$) distributed random variables. This is the famous *barometric law* for the density of a gas in a gravitational field.

Suppose we have independent observations $\omega_1, \dots, \omega_n$ from an unknown distribution in an exponential class. Then the compound sample $\omega = (\omega_1, \dots, \omega_n)$ is described by an exponential class, too:

If

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

is a k -parameter exponential class with natural parameters $\theta = (\theta_1, \dots, \theta_k)$ and natural statistics $T = (T_1, \dots, T_k) : \Omega \rightarrow \mathbb{R}^k$, then

$$\mathcal{P}^n = \{P_\theta^n : \theta \in \Theta\}$$

is a k -parameter exponential class, too, with natural parameters $\theta = (\theta_1, \dots, \theta_k)$ and natural statistics

$$(\omega_j)_{j=1, \dots, n} \mapsto \left(\sum_{j=1}^n T_l(\omega_j) \right)_{l=1, \dots, k}, \quad (62)$$

$$\Omega^n \rightarrow \mathbb{R}^k.$$

To prove this, assume that P_θ has the likelihood function

$$p(\omega|\theta) = \frac{g(\omega)}{Z(\theta)} \exp \left\{ \sum_{l=1}^k \theta_l T_l(\omega) \right\}.$$

Then, by independence of the components $\omega_1, \dots, \omega_n$, P_θ^n has the likelihood function

$$p(\omega_1, \dots, \omega_n|\theta) = \prod_{j=1}^n \left(\frac{g(\omega_j)}{Z(\theta)} \exp \left\{ \sum_{l=1}^k \theta_l T_l(\omega_j) \right\} \right) = \frac{\prod_{j=1}^n g(\omega_j)}{Z(\theta)^n} \exp \left\{ \sum_{l=1}^k \theta_l \sum_{j=1}^n T_l(\omega_j) \right\}.$$

This proves the claim.

To give an **example**, consider n i.i.d. normal random variables $X_1, \dots, X_n$ with unknown expectation μ and unknown variance σ^2 . Thus our frame model consists of all distributions $\text{normal}(\mu, \sigma^2)^n$ on $\mathbb{R}^n$. Using our previous example of the one-dimensional normal distribution, we conclude that this is a 2-parametric exponential class with natural parameters $\theta_1 = \mu/\sigma^2$ and $\theta_2 = -1/(2\sigma^2)$ and the natural statistics

$$T_1 = \sum_{j=1}^n X_j \quad \text{and} \quad T_2 = \sum_{j=1}^n X_j^2.$$

2.6 Two-sided tests

Example 2.2 illustrates that the choice of an optimal rejection region for simple hypotheses $\{P_0\}, \{P_A\}$ may strongly depend on P_A . What to do if the null hypothesis and the alternative hypothesis are both composite, if there are no uniformly most powerful tests? There are heuristic procedures to still define good tests in this situation, even if there are no optimal ones.

Example: A coin is tossed n times. We would like to test the null hypothesis $H_0 = \{\text{binomial}(n, 50\%)\}$ versus the alternative hypothesis $H_A = \{\text{binomial}(n, p) : p \neq 50\%\}$. For each of the alternative hypotheses “ $p < 50\%$ ” and “ $p > 50\%$ ” we know uniformly most powerful tests, but there is none for H_A : “ $p \neq 50\%$ ”.

More generally, consider a one-parameter frame model $\{P_\theta : \theta \in \Theta\}$ on Ω .

Testing the null hypothesis H_0 : “ $\theta = \theta_0$ ” versus the alternative H_A : “ $\theta \neq \theta_0$ ” is called a *two-sided* testing problem.

On the other hand, testing the null hypothesis H_0 : “ $\theta = \theta_0$ ” (or H_0 : “ $\theta \leq \theta_0$ ”) versus the alternative H_A : “ $\theta > \theta_0$ ” is called a *one-sided* testing problem.

Similarly, testing the null hypothesis H_0 : “ $\theta = \theta_0$ ” (or H_0 : “ $\theta \geq \theta_0$ ”) versus the alternative H_A : “ $\theta < \theta_0$ ” is called a *one-sided* testing problem, too.

In many practically relevant cases, one has uniformly most powerful tests for one-sided testing problems, due to monotone likelihood ratios. However, uniformly most powerful tests for two-sided testing problems hardly ever exist. What to do in this case?

Heuristically, one can proceed as follows for two-sided problems:

Given a bound α^* for the significance level, construct a most powerful test for testing $H_0 : \theta = \theta_0$ versus $H'_A : \theta > \theta_0$ at the significance level $\leq \alpha^*/2$, and another most powerful test for testing $H_0 : \theta = \theta_0$ versus $H''_A : \theta < \theta_0$ at the significance level $\leq \alpha^*/2$. One gets two rejection regions, say R' and R'' . Set $R := R' \cup R''$.

Since

$$P_0[R] \leq P_0[R'] + P_0[R''] \leq \frac{\alpha^*}{2} + \frac{\alpha^*}{2} = \alpha^*,$$

the test has a significance level $\leq \alpha^*$.

Given $P_A \in H_A$, the power of the test with rejection region R is at least as large as the power for any other test at a significance level $\leq \alpha^*/2$ by the choice of R and R' . To see this, note that

$$P_A[R] \geq \max\{P_A[R'], P_A[R'']\}.$$

If the rejection regions are defined in terms of a test statistic T , say $R' = \{T > c'\}$ and $R'' = \{T < c''\}$, then we get

$$R = \{T > c' \text{ or } T < c''\}$$

as the rejection region for the two-sided test, where c' is chosen minimal and c'' is chosen maximal with the constraints

$$P_0[T > c'] \leq \frac{\alpha^*}{2} \quad \text{and} \quad P_0[T < c''] \leq \frac{\alpha^*}{2}.$$

In testing a coin by throwing it n times; H_0 : “The coin is unbiased” versus H_A : “The coin is biased”, we get the rejection domain

$$R = \{0, \dots, k\} \cup \{n - k, \dots, n\},$$

where k is chosen as large as possible with respect to the constraint

$$\frac{1}{2^n} \sum_{j=0}^k \binom{n}{j} = \frac{1}{2^n} \sum_{j=n-k}^n \binom{n}{j} \leq \frac{\alpha^*}{2}.$$

2.7 Generalized likelihood ratio tests for composite hypotheses

Consider a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with likelihood function $(\omega, \theta) \mapsto p(\omega|\theta)$. Given composite hypotheses $H_0 = \{P_\theta : \theta \in \Theta_0\}$, $H_A = \{P_\theta : \theta \in \Theta_A\}$, we know that in general no uniformly most powerful test exists. Here is a heuristic procedure what to do in this case: Suppose we have observed data $\omega \in \Omega$.

- First, determine the “most likely” parameter $\hat{\theta} \in \Theta_0 \cup \Theta_A$ (if it exists), in the sense that the likelihood function should be maximal:

$$p(\omega|\hat{\theta}) = \max_{\theta \in \Theta_0 \cup \Theta_A} p(\omega|\theta).$$

In the case $\Theta_A = \Theta \setminus \Theta_0$, i.e. $\Theta_0 \cup \Theta_A = \Theta$, the random parameter $\hat{\theta}$ is called the *maximum likelihood estimator* for the true parameter $\theta \in \Theta$; we will examine it in more detail in the section on parameter estimation, below.

- Second, determine the “most likely” parameter $\hat{\theta}_0 \in \Theta_0$ in the restricted class Θ_0 defining the null hypothesis, in the sense

$$p(\omega|\hat{\theta}_0) = \max_{\theta \in \Theta_0} p(\omega|\theta).$$

- Then use the likelihood ratio at the estimated parameters

$$T = \frac{p(\omega|\hat{\theta}_0)}{p(\omega|\hat{\theta})}$$

as a test statistic. Obviously, $T \leq 1$ holds, since the maximum in the numerator is taken over a smaller class than the maximum in the denominator.

- Choose a critical value c and define the rejection region

$$R = \{T < c\}.$$

Thus one rejects the null hypothesis if the likelihood for the estimated parameter $\hat{\theta}_0$ in the class restricted by the null hypothesis is small compared to the likelihood for the estimated parameter $\hat{\theta}$ without restriction.

Definition 2.10 Tests having the rejection region

$$R = \left\{ \omega \in \Omega : \frac{\sup_{\theta \in \Theta_0} p(\omega|\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_A} p(\omega|\theta)} < c \right\}$$

with some critical value $c \in [0, 1]$ are called *generalized likelihood ratio tests*.

Generalized likelihood ratio tests remind us to the likelihood ratio tests for simple hypotheses; the difference being that *given, nonrandom* parameters are replaced by *estimated, random* ones. However, using “estimated” parameters instead of the “true” ones, the Neyman-Pearson lemma is not anymore applicable: *Generalized likelihood ratio tests are not guaranteed to have maximal power*. However, in practice, they do reasonably well.

Example 2.11 Assume that $X_1, \dots, X_n$ are i.i.d. normal random *vectors* with values in $\mathbb{R}^k$, unknown expectation vector $\mu = (\mu_1, \dots, \mu_k) \in \mathbb{R}^k$, and covariance matrix $\sigma^2 I_k$ (σ^2 times the $k \times k$ identity matrix) with known σ^2 . In particular, every X_j , $1 \leq j \leq n$, consists of k real-valued components. Let $m \leq k$ and $V \subseteq \mathbb{R}^k$ be a m -dimensional linear subspace. We design a generalized likelihood ratio test for testing the

null hypothesis H_0 : “ $\mu \in V$ ” versus H_A : “ $\mu \notin V$ ”.

To have a specific example in mind, one could think of H_0 : “ $\mu_1 = \mu_2 = \dots = \mu_k$ ”, which describes a 1-dimensional subspace of $\mathbb{R}^k$

Formally, we take $\Omega = (\mathbb{R}^k)^n$, $\mathcal{P} = \{\text{normal}(\mu, \sigma^2 I_k)^n : \mu \in \mathbb{R}^k\}$, $H_0 = \{\text{normal}(\mu, \sigma^2 I_k)^n : \mu \in V\}$, and $H_A = \mathcal{P} \setminus H_0$.

Below, we use the following notation: For $u = (u_l)_{l=1,\dots,k}$, $v = (v_l)_{l=1,\dots,k} \in \mathbb{R}^k$ we set $\|u\|^2 = \sum_{l=1}^k u_l^2$ for the squared euclidean norm, $\langle u, v \rangle = \sum_{l=1}^k u_l v_l$ for the euclidean scalar product. Finally we write $\bar{x} = n^{-1} \sum_{j=1}^n x_j$ for the sample average, where $x_1, \dots, x_n \in \mathbb{R}^k$ are n vectors in $\mathbb{R}^k$. The likelihood function in this example at $x_1, \dots, x_n$ is

$$\begin{aligned} f(x_1, \dots, x_n | \mu) &= (2\pi\sigma^2)^{-nk/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^n \|x_j - \mu\|^2 \right\} \\ &= (2\pi\sigma^2)^{-nk/2} \exp \left\{ -\frac{1}{2\sigma^2} n \|\bar{x} - \mu\|^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n \|x_j - \bar{x}\|^2 \right\}; \end{aligned}$$

to see this, note that $x_j - \mu = (\bar{x} - \mu) + (x_j - \bar{x})$ implies

$$\sum_{j=1}^n \|x_j - \mu\|^2 = \sum_{j=1}^n \|\bar{x} - \mu\|^2 + 2 \left\langle \sum_{j=1}^n (x_j - \bar{x}), \bar{x} - \mu \right\rangle + \sum_{j=1}^n \|x_j - \bar{x}\|^2 = n \|\bar{x} - \mu\|^2 + \sum_{j=1}^n \|x_j - \bar{x}\|^2.$$

The maximum

$$\max_{\mu \in \mathbb{R}^k} f(x_1, \dots, x_n | \mu)$$

is obtained for setting μ equal to $\bar{x}$, while the maximum

$$\max_{\mu \in V} f(x_1, \dots, x_n | \mu)$$

is obtained for setting μ equal to the orthogonal projection $\Pi_V \bar{x}$ of $\bar{x}$ on the linear subspace V . We obtain the generalized likelihood ratio

$$\begin{aligned} T &= \frac{\max_{\mu \in V} f(x_1, \dots, x_n | \mu)}{\max_{\mu \in \mathbb{R}^k} f(x_1, \dots, x_n | \mu)} \\ &= \frac{(2\pi\sigma^2)^{-nk/2} \exp \left\{ -\frac{1}{2\sigma^2} n \|\bar{x} - \Pi_V \bar{x}\|^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n \|x_j - \bar{x}\|^2 \right\}}{(2\pi\sigma^2)^{-nk/2} \exp \left\{ -\frac{1}{2\sigma^2} n \|\bar{x} - \bar{x}\|^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n \|x_j - \bar{x}\|^2 \right\}} = \exp \left\{ -\frac{1}{2\sigma^2} n \|\bar{x} - \Pi_V \bar{x}\|^2 \right\}. \end{aligned}$$

Let us determine null distributions of T , i.e. the distribution of T with respect to $P_0 = \text{normal}(\mu, \sigma^2 I_k)^n \in H_0$: Since the $X_1, \dots, X_n$ are i.i.d. $\text{normal}(\mu, \sigma^2 I_k)$ random vectors, their average $\bar{X} = n^{-1} \sum_j X_j$ is a $\text{normal}(\mu, n^{-1} \sigma^2 I_k)$ distributed random vector. Introducing the normalized random vector

$$Y = \sqrt{\frac{n}{\sigma^2}} (\bar{X} - \mu),$$

we see that $Y \sim \text{normal}(0, I_k)$. However, using $P_0 \in H_0$, we know $\mu = \Pi_V \mu \in V$; thus

$$Y - \Pi_V Y = \sqrt{\frac{n}{\sigma^2}} (\bar{X} - \mu - \Pi_V (\bar{X} - \mu)) = \sqrt{\frac{n}{\sigma^2}} (\bar{X} - \Pi_V \bar{X});$$

recall that Π_V denotes the orthogonal projection onto V . We obtain

$$\|Y - \Pi_V Y\|^2 = \frac{1}{\sigma^2} n \|\bar{X} - \Pi_V \bar{X}\|^2.$$

Let $d = \dim V$ and let $Y = (Y_1, \dots, Y_k)$. In the special case

$$V = \{(u_1, \dots, u_d, 0, \dots, 0) \in \mathbb{R}^k : u_1, \dots, u_d \in \mathbb{R}\},$$

$Y - \Pi_V Y = (0, \dots, 0, Y_{d+1}, \dots, Y_k)$, and the $Y_{d+1}, \dots, Y_k$ are i.i.d. standard normal random variables. Thus $\|Y - \Pi_V Y\|^2$ is a χ_{k-d}^2 -distributed random variable with respect to P_0 for our special choice of V . However, $\|Y - \Pi_V Y\|^2$ has the same distribution χ_{k-d}^2 for every choice of a d -dimensional linear subspace V , as the following consideration shows: Choose a $k \times k$ orthogonal matrix U that transforms a given d -dimensional linear subspace $V \subseteq \mathbb{R}^n$ to the “standard” subspace

$$UV = \{Uv : v \in V\} = \{(u_1, \dots, u_d, 0, \dots, 0) \in \mathbb{R}^k : u_1, \dots, u_d \in \mathbb{R}\}.$$

We use:

Rotational invariance of the multidimensional standard normal distribution.

Let U be a orthogonal $k \times k$ matrix. If Y is a k -dimensional random vector with i.i.d. standard normal components, i.e. $Y \sim \text{normal}(0, I_k)$, then UY is a k -dimensional random vector with i.i.d. standard normal components, too.

Indeed, since the Y has the probability density

$$f(y) = (2\pi)^{-k/2} e^{-\|y\|^2/2},$$

and since $|\det U| = 1$, UY has the same probability density

$$f(z) = (2\pi)^{-k/2} e^{-\|U^{-1}z\|^2/2} |\det U|^{-1} = (2\pi)^{-k/2} e^{-\|z\|^2/2}.$$

(Recall the *transformation formula for multidimensional integrals*: If $G : B \rightarrow B'$ is a differentiable function between two open subsets B, B' of $\mathbb{R}^m$ with a differentiable inverse, then

$$\int_B f(G(x)) |\det DG(x)| dx = \int_{B'} f(y) dy, \quad (63)$$

where $\det DG$ denotes the Jacobi determinant of G . In particular, if a random vector Y has the density f , then $G^{-1}(Y)$ has the density $x \mapsto f(G(x)) |\det DG(x)|$. Furthermore, recall that $\det(AB) = (\det A)(\det B)$.)

We conclude that

$$-2 \log T = \|Y - \Pi_V Y\|^2 = \|UY - \Pi_{UV} UY\|^2$$

has a χ_{k-d}^2 distribution, too.

Summarizing, we have identified the null distribution for the generalized likelihood ratio T :

In this example, $-2 \log T$ follows a χ^2 distribution with $k - \dim V$ degrees of freedom with respect to all $P_0 \in H_0$, independently of the value of σ^2 .

Remark: The example is not as special as it might seem: Consider a k -parametric model of n i.i.d. observations

$$\mathcal{P} = \{P_\theta^n : \theta \in \Theta\}$$

with a k -dimensional open parameter set $\Theta \subseteq \mathbb{R}^k$. Consider a null hypothesis H_0 : “ $\theta \in \Theta_0$ ” parametrized by a d -dimensional submanifold Θ_0 of Θ . Typically, such a submanifold is described by $k-d$ (independent) equations for the parameter θ . In general the null distribution of the generalized likelihood ratio T depends on $P_0 \in H_0$. However, asymptotically as $n \rightarrow \infty$, this dependence typically disappears: Under rather weak smoothness and regularity assumptions for the likelihood function and with respect to *all* $\theta \in \Theta_0$ (null distribution $P_\theta^n \in H_0$), the statistic $-2\log T$ converges in distribution to a χ_{k-d}^2 distribution as $n \rightarrow \infty$, i.e. as the sample size grows to infinity. We do not rigorously describe or even prove this statement here, but the above example shows the flavor of this general statement in a simple case.

2.8 Varying the null hypothesis: Confidence regions.

Confidence intervals play a similar role in statistics as error bounds play numerical mathematics. In numerical mathematics, an interesting number is often known only up to a certain precision; one then describes intervals containing the true, unknown value. Similarly, in statistics, one would like to describe intervals containing an unknown parameter θ , based on the random observation ω . In contrast to deterministic error bounds in numerical mathematics, confidence intervals are random intervals. They do not contain the unknown parameter for sure. They only contain it with a prescribed probability (chosen close to 1), uniformly, whatever the true unknown parameter might be.

Just like the p -value encodes the test decision for many tests with varying significance level, confidence regions encode the test decision for many tests with varying null hypotheses.

Consider a parametric model

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

on the set of results Ω . Suppose that for every $\theta \in \Theta$, we have a test specified by its rejection R_θ for testing the null hypothesis $H_0(\theta) = \{P_\theta\}$. (We do not specify an alternative hypothesis for our present purpose, although the actual construction of R_θ certainly depends on it. A standard choice among many other possibilities could be $H_A(\theta) = \mathcal{P} \setminus \{P_\theta\}$.) Suppose that all tests have significance level α , i.e.

$$\text{for all } \theta \in \Theta: \quad P_\theta[R_\theta] = \alpha.$$

(In the case of discrete models, this is usually a too strong requirement, since there are only discrete possibilities for the significance level. In these cases, we just suppose

$$\text{for all } \theta \in \Theta: \quad P_\theta[R_\theta] \leq \alpha,$$

but then we try to get R_θ as large as possible in order to have large powers of the test.)

Definition 2.12 Suppose we have a family tests for testing the varying null hypotheses

$$H_0(\theta) = \{P_\theta\}, \quad \theta \in \Theta.$$

Let the tests be specified by a family of rejection domains $(R_\theta)_{\theta \in \Theta}$ with significance levels

$$P_\theta[R_\theta] \leq \alpha$$

for all $\theta \in \Theta$.

Given observed data $\omega \in \Omega$, the corresponding $(1-\alpha)$ -confidence region for the unknown parameter is defined by

$$C_{1-\alpha}(\omega) = \{\theta \in \Theta : \omega \notin R_\theta\}.$$

For one-parametric models, confidence regions usually are intervals. In this case, they are called *confidence intervals*.

Note that confidence regions are *random* regions, since they depend on the *random, but known* observation $\omega \in \Omega$, while the parameter $\theta \in \Theta$ is treated as *non-random, but unknown*.

According to the definition, every $(1 - \alpha)$ -confidence region is also a $(1 - \alpha')$ -confidence region whenever $\alpha' \geq \alpha$. However, one would like to have the confidence regions as small as possible, since confidence intervals are a probabilistic analogue of error bounds for an unknown value. The requirement to find *small* confidence intervals is usually implicitly imposed.

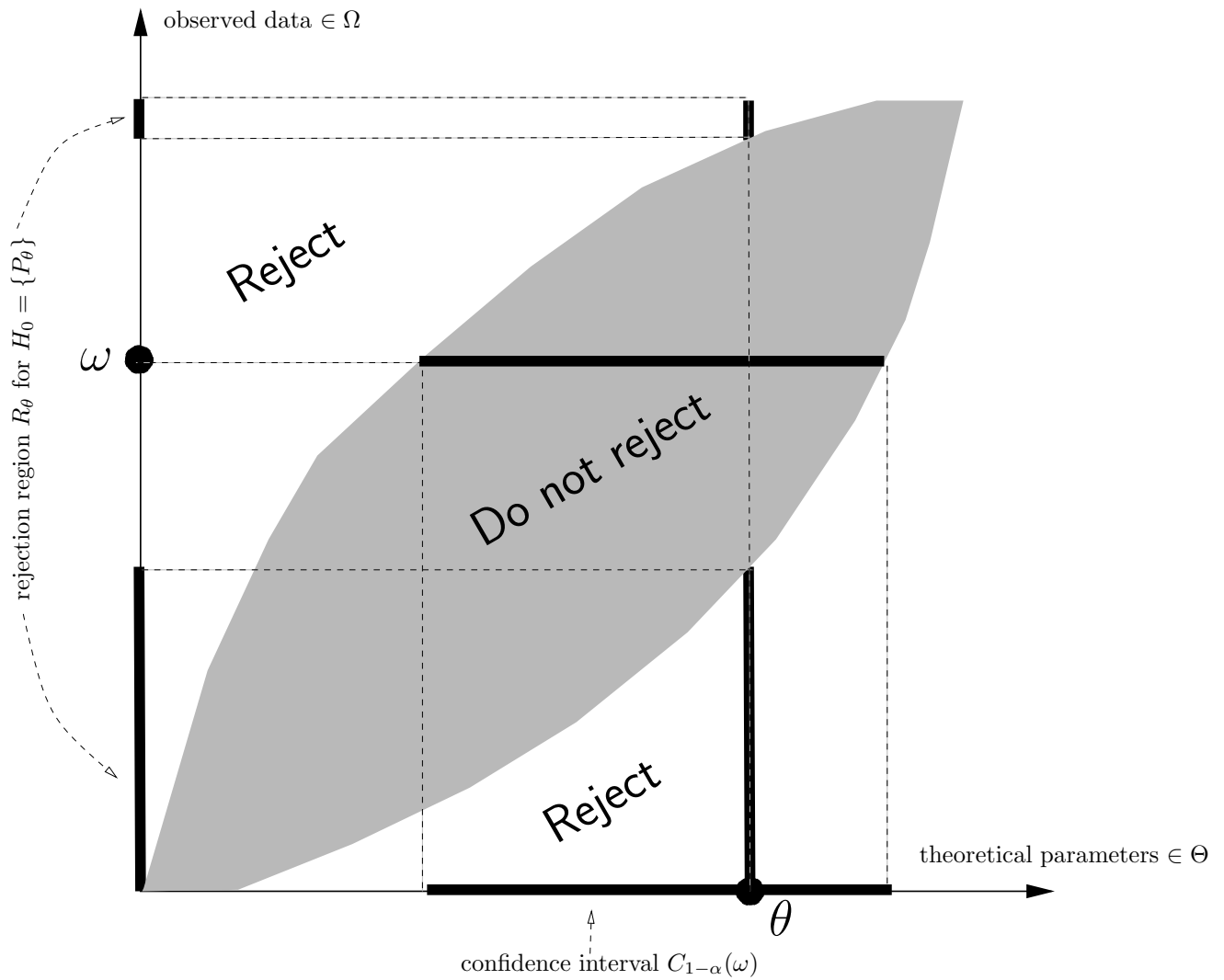
Indeed, confidence regions encode the test decision for many tests with varying null hypothesis for the given observation ω :

Duality between tests and confidence regions.

Given a parameter $\theta \in \Theta$ and data $\omega \in \Omega$, the following statements are equivalent:

- $\omega \notin R_\theta$, i.e. the null hypothesis $H_0(\theta)$: “the true parameter is θ ” is not rejected at the significance levels $\leq \alpha$;
- $\theta \in C_{1-\alpha}(\omega)$, i.e. the parameter θ is contained in the $(1 - \alpha)$ -confidence region $C_{1-\alpha}(\omega)$ for the observed data ω .

This is obvious by the definition of confidence regions.



Intuitively and roughly speaking, a confidence region describes the set of “plausible” parameters, which are “consistent” with the observed data. The closer to 1 one chooses $1 - \alpha$, the larger the confidence regions get.

Lemma 2.13 (Confidence regions contain the true parameter [at least] with the specified probability.)

For all $\theta \in \Theta$,

$$P_\theta[\theta \in C_{1-\alpha}(\omega)] \geq 1 - \alpha.$$

Thus the *random* $(1 - \alpha)$ -confidence region contains the true *non-random, but unknown* parameter θ (at least) with probability $1 - \alpha$, uniformly for all parameters $\theta \in \Theta$.

Proof: Let $\theta \in \Theta$. Then

$$P_\theta[\theta \in C_{1-\alpha}(\omega)] = P_\theta[\omega \notin R_\theta] = 1 - P_\theta[\omega \in R_\theta] \geq 1 - \alpha.$$

□

Lemma 2.14 (Recovering tests from confidence regions.)

Let random regions $C(\omega)$, $\omega \in \Omega$, be given such that for all $\theta \in \Theta$,

$$P_\theta[\theta \in C(\omega)] \geq 1 - \alpha.$$

This means that we assume that the random region $C(\omega)$ contains the parameter θ (at least) with probability $1 - \alpha$, uniformly for all parameters $\theta \in \Theta$.

Then for all $\theta \in \Theta$, the test with rejection region

$$R_\theta = \{\omega \in \Omega : \theta \notin C(\omega)\}$$

for testing the null hypothesis $H_0 = \{\theta\}$ has the significance level

$$P_\theta[R_\theta] \leq \alpha.$$

The $C(\omega)$ are $(1 - \alpha)$ -confidence intervals derived from the family of rejection domains $(R_\theta)_{\theta \in \Theta}$.

Proof:

$$P_\theta[R_\theta] = P_\theta[\theta \notin C(\omega)] = 1 - P_\theta[\theta \in C(\omega)] \leq \alpha.$$

Furthermore, $\theta \notin C(\omega)$ is equivalent to $\omega \in R_\theta$; thus the random regions $\omega \mapsto C(\omega)$ are indeed confidence intervals. □

Example 2.15 A Geiger counter detects N α -particles during one second measuring time. Assuming that $N \sim \text{Poisson}(\lambda)$ with an unknown parameter λ , we derive a (two-sided) 95% confidence interval for λ .

Using the frame model $\Omega = \mathbb{N}_0$, $\mathcal{P} = \{\text{Poisson}(\lambda) : \lambda > 0\}$, we derive two-sided tests for testing $H_0(\lambda) = \{\text{Poisson}(\lambda)\}$ versus $H_A(\lambda) = \mathcal{P} \setminus H_0(\lambda)$ on the 5% level: The rejection regions are

$$R_\lambda = \{n \in \mathbb{N}_0 : n \leq n_1(\lambda) \text{ or } n \geq n_2(\lambda)\},$$

where $n_1(\lambda)$ is taken maximal with the constraint

$$\text{Poisson}(\lambda)[\{0, \dots, n_1(\lambda)\}] \leq 2.5\%,$$

and $n_2(\lambda)$ is taken minimal with the constraint

$$\text{Poisson}(\lambda)[\{n_2(\lambda), n_2(\lambda) + 1, \dots\}] \leq 2.5\%.$$

The two-sided 95%-confidence interval is then

$$C_{95\%}(N) = \{\lambda > 0 : n_1(\lambda) < N < n_2(\lambda)\}.$$

The condition $n_1(\lambda) < N < n_2(\lambda)$ is equivalent to

$$\sum_{j=0}^N e^{-\lambda} \frac{\lambda^j}{j!} > 2.5\%$$

and

$$\sum_{j=N}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} > 2.5\%.$$

The first sum is monotonically decreasing in λ , and the second sum is monotonically increasing in λ .

(Here is one possibility to show this: Let $\lambda < \lambda'$, and take two independent random variables $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\lambda' - \lambda)$. Then $X + Y \sim \text{Poisson}(\lambda')$, and $X \leq X + Y$ implies

$$P[X \leq n] \geq P[X + Y \leq n]$$

and

$$P[X \geq n] \leq P[X + Y \geq n].$$

This is just the claim, written in a different notation.)

Thus we get the confidence interval

$$C_{95\%}(N) =]\lambda_1, \lambda_2[,$$

where λ_1 and λ_2 are determined implicitly by the equations

$$\sum_{j=0}^N e^{-\lambda_2} \frac{\lambda_2^j}{j!} = 2.5\%$$

and

$$1 - \sum_{j=0}^{N-1} e^{-\lambda_1} \frac{\lambda_1^j}{j!} = \sum_{j=N}^{\infty} e^{-\lambda_1} \frac{\lambda_1^j}{j!} = 2.5\%.$$

Given observed values for N , these transcendental equations can be solved numerically. For example, for the observation $N = 2$, one gets the following two-sided 95%-confidence interval for the unknown parameter λ (rounded numerical values):

$$C_{95\%}(N = 2) =]0.242, 7.225[.$$

Thus, for example, the null hypothesis “ $\lambda = \frac{1}{4}$ ” is not rejected on the 5% level by a two-sided test, since $\frac{1}{4} \in C_{95\%}(N = 2)$. On the other hand, the null hypothesis “ $\lambda = 8$ ” is rejected on the 5% level, since $8 \notin C_{95\%}(N = 2)$. Similarly, one finds the 99%-confidence interval (rounded numerical values):

$$C_{99\%}(N = 2) =]0.103, 9.274[;$$

of course it is larger than the 95%-confidence interval.

Confidence regions for non-parametric models. Here is a generalisation of Definition 2.12, which also applies to the non-parametric case:

Consider a (maybe non-parametric) frame model $\mathcal{P}$ over Ω and a parameter $\lambda : \mathcal{P} \rightarrow \Lambda$ onto some set Λ . In contrast to parametric models, λ need not be one-to-one.

Definition 2.16 Let $(R_\theta)_{\theta \in \Lambda}$ be the rejection regions for a family of tests with the null hypotheses

$$H_0(\theta) = \{P \in \mathcal{P} : \lambda(P) = \theta\}, \quad \theta \in \Lambda$$

with significance levels $\sup_{P: \lambda(P)=\theta} P[R_\theta] \leq \alpha$ for all $\theta \in \Lambda$. Given observed data $\omega \in \Omega$, the corresponding $(1 - \alpha)$ -confidence region is defined by

$$C_{1-\alpha}(\omega) = \{\theta \in \Lambda : \omega \notin R_\theta\}.$$

In particular,

$$\text{for all } P \in \mathcal{P}: \quad P[\lambda(P) \in C_{1-\alpha}(\omega)] \geq 1 - \alpha.$$

Example 2.17 In this example, we derive confidence intervals for the *median* of an unknown distribution.

Definition 2.18 The *median* of a random variable X (synonym: the 50% *quantile*) is a number $\text{median}_X \in \mathbb{R}$ such that

$$P[X < \text{median}_X] \leq \frac{1}{2} \leq P[X \leq \text{median}_X].$$

This condition might not determine the median uniquely. In order to have a single definite value, we take the middle of all possible choices:

$$\text{median}_X = \frac{1}{2} \left(\min\{m : P[X \leq m] \geq \frac{1}{2}\} + \max\{m : P[X \geq m] \geq \frac{1}{2}\} \right).$$

Similarly, the median of a probability distribution P is the median of X , where X has the law P .

Roughly speaking, half of the probability mass lies to the left of the median, and half of the probability mass lies to the right of the median. For example, the median of a normal(μ, σ^2) distributed random variable equals its expectation. If X has a continuous distribution, we simply get

$$P[X \leq \text{median}_X] = 50\%.$$

Let $X_1, \dots, X_n$ be i.i.d. random variables with an unknown continuous distribution P .

Recall: “ P is a continuous distribution” means that the cumulative distribution function

$$F : t \mapsto P[] - \infty, t]$$

is continuous. Equivalently, it means that $P[\{t\}] = 0$ holds for all $t \in \mathbb{R}$, i.e. the probability to hit a single given point is 0.

We want to derive a two-sided $(1 - \alpha)$ -confidence interval for the median of P . We use the nonparametric frame model $\Omega = \mathbb{R}^n$,

$$\mathcal{P} = \{P^n : P \text{ is a continuous distribution on } \mathbb{R}\}.$$

Let $m \in \mathbb{R}$. We develop a test for the null hypothesis $H_0(m)$: “ $\text{median}(P) = m$ ” versus $H_A(m)$: “ $\text{median}(P) \neq m$ ”. We use the test statistic

$$T_m = |\{j \in [1, n] : X_j \leq m\}|,$$

i.e. the number of observations less than or equal to m . Note that the distribution of T_m with respect to *all* null distributions $P_0 \in H_0(m)$ is the same. This is a main reason for choosing this test statistic. More precisely, with respect to all $P_0 \in H_0(m)$,

$$T_m \sim \text{binomial}(n, 50\%).$$

To see this, note that every X_j has the probability 50% to lie to the left of the median of its distribution. Furthermore, the X_j are independent.

We reject $H_0(m)$ for large values of $|T_m - n/2|$, taking the rejection region

$$R_m = \{|T_m - n/2| > c_n\}$$

for some critical value c_n . Thus this test has the significance level

$$\alpha = 1 - \text{binomial}(n, 50\%)[[\frac{n}{2} - c_n, \frac{n}{2} + c_n]].$$

We get the following $1 - \alpha$ confidence interval for the unknown median:

$$C_{1-\alpha}(X_1, \dots, X_n) = \{m \in \mathbb{R} : |T_m - n/2| \leq c_n\} = [X_{(n/2-c_n)}, X_{(n/2+c_n+1)}]$$

where

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

are the data $X_1, \dots, X_n$ rearranged in increasing order. $(X_{(1)}, \dots, X_{(n)})$ is called the *order statistic* of $X_1, \dots, X_n$. Let us examine which confidence level $1 - \alpha$ we get for large n : By the de Moivre – Laplace theorem (= central limit theorem for binomial random variables) and with respect to the null distribution,

$$\frac{T_m - n/2}{\sqrt{n}/2}$$

converges in distribution to a standard normal distribution as $n \rightarrow \infty$. Thus, choosing a sequence (c_n) with

$$\frac{c_n}{\sqrt{n}/2} \xrightarrow{n \rightarrow \infty} c$$

for some $c \geq 0$, the confidence levels $1 - \alpha$ converge to

$$\text{standard-normal}([-c, c]) = 2\Phi(c) - 1$$

as $n \rightarrow \infty$, where Φ denotes again the distribution function of the standard normal distribution. Roughly speaking, this means:

If one wants to have a fixed confidence level $1 - \alpha$ for large n , one should roughly choose

$$c_n \approx \frac{\sqrt{n}}{2} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

2.9 Linear transformation of normal distributions

Here is an important transformation property of multidimensional normal distributions:

Lemma 2.19 (Linear transformation of multidimensional normal distributions) Let

$$X \sim \text{normal}(\mu, \Sigma)$$

be a random vector with values in $\mathbb{R}^m$, expectation vector $\mu \in \mathbb{R}^m$, and covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$. Let $n \leq m$ and $A \in \mathbb{R}^{n \times m}$ be a $n \times m$ -matrix of rank n and $b \in \mathbb{R}^n$. Then

$$AX + b \sim \text{normal}(A\mu + b, A\Sigma A^\top).$$

(Here $A^\top$ denotes the transposed matrix of A :

$$A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} \quad \text{implies} \quad A^\top = \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1m} & \dots & a_{nm} \end{pmatrix}.$$

Recall that the covariance of two random variables Y and Z is defined by $\text{Cov}(Y, Z) = E[(Y - E[Y])(Z - E[Z])]$, and that the covariance matrix of a random vector $(X_1, \dots, X_n)$ is a $n \times n$ -matrix with entries $\text{Cov}(X_i, X_j)$, $1 \leq i, j \leq n$.)

Proof. X has the (joint) probability density

$$f(x) = (2\pi)^{-m/2} (\det \Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}.$$

(Recall: a random vector $X = (X_1, \dots, X_n)$ is said to have the density f , if for all events $B \subseteq \mathbb{R}^n$ one has $P[X \in B] = \int_B f(x) dx$; here the integral is n -dimensional.)

We prove the lemma first for the case $m = n$. In this case, we use the transformation rule for probability densities to find the following density g of $AX + b$:

$$\begin{aligned} g(y) &= (2\pi)^{-m/2} (\det \Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} (A^{-1}(y - b) - \mu)^\top \Sigma^{-1} (A^{-1}(y - b) - \mu) \right\} |\det A|^{-1} \\ &= (2\pi)^{-m/2} (\det(A\Sigma A^\top))^{-1/2} \exp \left\{ -\frac{1}{2} (y - A\mu - b)^\top (A\Sigma A^\top)^{-1} (y - A\mu - b) \right\}, \end{aligned}$$

which is the density of the normal($A\mu + b, A\Sigma A^\top$) distribution. We used that the Jacobi determinant of the map $x \mapsto Ax + b$ is $\det A$ and that $\det(A^\top \Sigma A) = (\det A)^2 \det \Sigma$.

(Recall the transformation formula (63) for multidimensional integrals, and recall that $\det(AB) = (\det A)(\det B)$.)

In the case $n < m$ we examine first the most simple example: Assume that X has i.i.d. standard normal components, i.e. $X \sim \text{normal}(0, I_m)$. Furthermore, let $x \mapsto \Pi x$, $\mathbb{R}^m \rightarrow \mathbb{R}^n$ denote the projection on the first m components, i.e.

$$\Pi = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix} = (I_n \mid 0_{n \times (m-n)})$$

with the $n \times n$ identity matrix I_n and the $n \times (m - n)$ null matrix $0_{n \times (m-n)}$, then ΠX has i.i.d. standard normal components, too, i.e. $X \sim \text{normal}(0, I_n)$. Since $\Pi I_m \Pi^\top = I_n$, this proves the lemma for this particularly simple example.

We reduce the general case to the two cases we treated so far:

We decompose A in a composition of several matrices. First, since the covariance matrix Σ is positive definite, we obtain $M\Sigma M^\top = I_m$ for some invertible $m \times m$ matrix M .

(Recall that a symmetric $m \times m$ -matrix Σ is called *positive definite*, if $x^\top \Sigma x > 0$ holds for all vectors $x \in \mathbb{R}^m$ with $x \neq 0$. Furthermore, recall that symmetric matrices can be diagonalized: $\Sigma = TDT^{-1} = TDT^\top$ with some orthogonal matrix T , $T^{-1} = T^\top$, and some diagonal matrix D . For positive definite matrices Σ , the diagonal entries of D will be positive. Let $\sqrt{D}$ denote a diagonal matrix with the square root of the entries of D as diagonal entries. We set $M = (T\sqrt{D})^{-1}$; then $M\Sigma M^\top = (T\sqrt{D})^{-1} (TDT^{-1}) [(T\sqrt{D})^{-1}]^\top = \sqrt{D}^{-1} T^{-1} T D T^{-1} T \sqrt{D}^{-1} = I_m$.)

For $n \times m$ -matrices S , let

$$\text{Null}(S) = \{x \in \mathbb{R}^m : Sx = 0\}$$

denote the null space of S . Let U be a orthogonal matrix that transforms the $m - n$ -dimensional space $\text{Null}(AM^{-1})$ onto the $m - n$ -dimensional space $\text{Null}(\Pi)$. We get $\text{Null}(\Pi U) = \text{Null}(AM^{-1})$; hence

$$\text{Null}(\Pi U M) = \text{Null}(A).$$

As a consequence, A can be written as

$$A = C \Pi U M$$

for some invertible $n \times n$ -matrix C . We conclude:

- $M(X - \mu)$ is $\text{normal}(0, M\Sigma M^\top) = \text{normal}(0, I_m)$ distributed;
- thus $UM(X - \mu)$ is $\text{normal}(0, I_m)$ distributed, too;
- thus $\Pi UM(X - \mu)$ is $\text{normal}(0, I_n)$ distributed;
- thus $A(X - \mu) = C\Pi UM(X - \mu)$ is $\text{normal}(0, CC^\top)$ distributed;
- thus $AX + b$ is $\text{normal}(A\mu + b, CC^\top)$ distributed.

We calculate:

$$CC^\top = C\Pi\Pi^\top C^\top = C\Pi UU^\top \Pi^\top C^\top = C\Pi UM\Sigma M^\top U^\top \Pi^\top C^\top = A\Sigma A^\top;$$

hence we get our claim:

$$AX + b \sim \text{normal}(A\mu + b, A\Sigma A^\top).$$

□

As a consequence, we obtain a proof of Theorem 2.4:

If $X_1, \dots, X_n$ are independent normal random variables with expectations $\mu_1, \dots, \mu_n$ and variances $\sigma_1^2, \dots, \sigma_n^2$, then $\sum_{j=1}^n X_j$ is a normal random variable with the expectation $\sum_{j=1}^n \mu_j$ and the variance $\sum_{j=1}^n \sigma_j^2$.

Indeed, we just apply the lemma to the special case $A = (1, 1, \dots, 1) \in \mathbb{R}^n$ and $b = 0$:

$$X = (X_1, \dots, X_n)^\top \sim \text{normal}((\mu_1, \dots, \mu_n)^\top, \text{diag}(\sigma_1^2, \dots, \sigma_n^2))$$

implies

$$\sum_j X_j = AX \sim \text{normal}(A(\mu_1, \dots, \mu_n)^\top, A \text{diag}(\sigma_1^2, \dots, \sigma_n^2) A^\top) = \text{normal}(\sum_j \mu_j, \sum_j \sigma_j^2).$$

Example 2.20 (Linear approximation for error propagation) As an application, consider imprecisely measured physical quantities $X_1, \dots, X_m$. A frequently used assumption is that the data are normally distributed, with a small covariance matrix.

Assume that $(X_1, \dots, X_m) \sim \text{normal}(\mu, \epsilon^2 \Sigma)$ with a small value $\epsilon > 0$. Consider a smooth, but possibly nonlinear transformation

$$f : (X_1, \dots, X_m) \mapsto Y = (Y_1, \dots, Y_m).$$

As $\epsilon \rightarrow 0$,

$$\epsilon^{-1}(Y - f(\mu))$$

converges in distribution to a $\text{normal}(0, J\Sigma J^\top)$ distribution. Here $J = Df(\mu)$ denotes the Jacobi matrix of f at μ .

Indeed, by Taylor's formula,

$$Y = f(\mu) + J(X - \mu) + O(\|X - \mu\|^2),$$

thus

$$\epsilon^{-1}(Y - f(\mu)) = J(\epsilon^{-1}(X - \mu)) + \epsilon^{-1}O(\|X - \mu\|^2).$$

The error term in the last expression is of the order ϵ ; thus it converges in distribution to 0 as $\epsilon \rightarrow 0$. It only remains the linear transformation $X \mapsto \epsilon^{-1}J(X - \mu)$; thus the lemma yields the claim.

Roughly speaking:

If we neglect nonlinear terms, the covariance matrix of the transformed quantities Y is roughly

$$\text{Cov}_Y \approx J \text{Cov}_X J^\top,$$

where J denotes the Jacobi matrix of the transformation.

This formula for error propagation is extensively used in applied sciences.

As a specific example, assume that the polar coordinates (r, ϕ) of a point in the plane are measured only with a certain (very small) standard deviation $\Delta r, \Delta \phi$, independently of each other. We derive a linear approximation for the covariance matrix of the cartesian coordinates

$$x = r \cos \phi,$$

$$y = r \sin \phi.$$

We get the Jacobi matrix

$$J = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \cos \phi & -r \sin \phi \\ \sin \phi & r \cos \phi \end{pmatrix}$$

and thus

$$\begin{aligned} \text{Cov}_{x,y} &\approx \begin{pmatrix} \cos \phi & -r \sin \phi \\ \sin \phi & r \cos \phi \end{pmatrix} \begin{pmatrix} \Delta r^2 & 0 \\ 0 & \Delta \phi^2 \end{pmatrix} \begin{pmatrix} \cos \phi & \sin \phi \\ -r \sin \phi & r \cos \phi \end{pmatrix} \\ &= \begin{pmatrix} \Delta r^2 \cos^2 \phi + \Delta \phi^2 r^2 \sin^2 \phi & (\Delta r^2 - r^2 \Delta \phi^2) \cos \phi \sin \phi \\ (\Delta r^2 - r^2 \Delta \phi^2) \cos \phi \sin \phi & \Delta r^2 \sin^2 \phi + \Delta \phi^2 r^2 \cos^2 \phi \end{pmatrix}. \end{aligned}$$

This formula gets asymptotically precise in the limit as $\Delta r, \Delta \phi \rightarrow 0$. Note that x and y may be correlated, even if r and ϕ are independent.

What happens when we linearly transform a m -dimensional normal(μ, Σ) random vector X by a $n \times m$ matrix A of rank less than n ? Then the image is supported in a linear subspace of $\mathbb{R}^n$; thus the components of AX are linearly dependent: Some of the components of AX then are redundant. The covariance matrix of AX still equals $A\Sigma A^\top$, but this matrix is not positive definite, but only positive semidefinite. We call the distribution of $AX + b$ again the normal($A\mu + b, A\Sigma A^\top$) distribution, even if $A\Sigma A^\top$ is only semidefinite.

2.10 The χ^2 test for simple hypotheses.

Example 2.21 The elections in some country in the last year gave the following result:

party	A	B	C	D
votes	39.6%	31.5%	21.7%	7.2%

Now, one year later, 1000 people were asked again which party they would vote for at present. Here is the result:

party	A	B	C	D
number of votes	375	326	209	90

Does this result reflect a change in the opinions, or could it be just explained as a random effect, caused by using only a small sample? In this section, we develop a statistical test for this problem and more general “goodness of fit” problems.

Let us first develop a frame model: The reply of the j -th voter in the sample ($j = 1, \dots, n$) is modeled by a random variable $Y_j \in \{A, B, C, D\}$. More generally, if there are d possible answers, we take random

variables $Y_i \in \Omega = \{A_1, \dots, A_d\}$. Let us assume that the Y_i are independent identically distributed with an unknown distribution P . Thus we use the frame model with result space Ω^n and

$$\mathcal{P} = \{P^n : P \text{ is a probability distribution on } \Omega\}. \quad (64)$$

The unknown distribution P can be described in terms of the d probabilities

$$p_i = P[\{A_i\}], \quad i = 1, \dots, d.$$

Of course, these probabilities sum up to 1: One has $p_i \geq 0$ with the constraint

$$\sum_{i=1}^d p_i = 1.$$

The number of votes for party A_i is

$$N_i = \sum_{j=1}^n 1_{\{Y_j = A_i\}}.$$

(Recall that $1_{\{Y_j = A_i\}} = 1$ for $Y_j = A_i$, but $1_{\{Y_j = A_i\}} = 0$ for $Y_j \neq A_i$.)

Definition 2.22 (Multinomial distribution) *The multinomial($n, p_1, \dots, p_d$) distribution is defined to be the joint distribution of*

$$N_i = \sum_{j=1}^n 1_{\{Y_j = A_i\}}, \quad i = 1, \dots, d,$$

where $Y_1, \dots, Y_n$ are i.i.d. random variables with distribution P , where $p_i = P[\{A_i\}]$. One has

$$n = \sum_{i=1}^d N_i.$$

Explicitly, we have for $n_1, \dots, n_d \in \mathbb{N}_0$ with $n_1 + \dots + n_d = n$:

$$P^n[N_i = n_i \text{ for } i = 1 \dots d] = \frac{n!}{\prod_{i=1}^d n_i!} \prod_{i=1}^d p_i^{n_i}. \quad (65)$$

Note that N_i just follows a binomial(n, p_i) distribution.

We take

$$X_j = (1_{\{Y_j = A_1\}}, 1_{\{Y_j = A_2\}}, \dots, 1_{\{Y_j = A_d\}})^\top, \quad j = 1, \dots, n;$$

note that

$$(N_1, \dots, N_d)^\top = \sum_{j=1}^n X_j \sim \text{multinomial}(n, p_1, \dots, p_d). \quad (66)$$

Let us calculate the expectation vector μ and the covariance matrix Σ of X_j . We write E^n for the expectation with respect to P^n .

$$\mu = E^n[X_j] = (E^n[1_{\{Y_j = A_i\}}])_{i=1 \dots d} = (P^n[Y_j = A_i])_{i=1 \dots d} = (p_1, \dots, p_d)^\top \quad (67)$$

and

$$\text{Cov}(1_{\{Y_j = A_i\}}, 1_{\{Y_j = A_{i'}\}}) = E^n[1_{\{Y_j = A_i\}} 1_{\{Y_j = A_{i'}\}}] - p_i p_{i'} \quad (68)$$

$$= P^n[Y_j = A_i = A_{i'}] - p_i p_{i'} = p_i \delta_{ii'} - p_i p_{i'}. \quad (69)$$

(Recall the formula

$$\text{Cov}(U, V) = E[UV] - E[U]E[V].$$

We use the Kronecker's delta:

$$\delta_{ii'} = \begin{cases} 1 & \text{for } i = i' \\ 0 & \text{for } i \neq i' \end{cases}$$

which describes the entries of the identity matrix.)

Thus we get the following covariance matrix of the random vector X_j :

$$\Sigma = \text{diag}(p_1, \dots, p_d) - (p_1, \dots, p_d)^\top (p_1, \dots, p_d).$$

Recall that $X_1, \dots, X_n$ are i.i.d. random vectors. Using (66), we get

Lemma 2.23 (Expectation and covariance matrix of the multinomial distribution)

The multinomial($n, p_1, \dots, p_d$) distribution has the expectation

$$E[(N_1, \dots, N_d)] = (np_1, \dots, np_d)$$

and the covariances

$$\text{Cov}(N_i, N_{i'}) = \begin{cases} n(p_i - p_i^2) & \text{for } i = i', \\ -np_i p_{i'} & \text{for } i \neq i'. \end{cases}$$

In matrix notation, this means:

$$(\text{Cov}(N_i, N_{i'}))_{i, i'=1, \dots, d} = n\Sigma = n \text{diag}(p_1, \dots, p_d) - n(p_1, \dots, p_d)^\top (p_1, \dots, p_d).$$

In order to understand the asymptotic behaviour of N_i for large n , we use the multidimensional central limit theorem:

Central limit theorem – multidimensional version. Let $X_1, \dots, X_n$ be i.i.d. random vectors with expectation vector $E[X_1] = \mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then as $n \rightarrow \infty$,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n (X_j - \mu)$$

converges in distribution to a normal($0, \Sigma$) distribution.

This theorem is proved in the same way as the one-dimensional version.

(In our case, where we have a continuous distribution as limit, “convergence in distribution to a normal distribution” just means:

$$\lim_{n \rightarrow \infty} P \left[\frac{1}{\sqrt{n}} \sum_{j=1}^n (X_j - \mu) \in S \right] = \text{normal}(0, \Sigma)[S]$$

for all open (or closed) sets $S \subseteq \mathbb{R}^d$.)

Roughly speaking, the central limit theorem states that the normal($n\mu, n\Sigma$) distribution is a good approximation for the distribution of $\sum_{j=1}^n X_j$ as n gets larger and larger.

We apply the central limit theorem to the multinomial distribution. We get

$$\frac{1}{\sqrt{n}} (N_1 - np_1, \dots, N_d - np_d)^\top = \frac{1}{\sqrt{n}} \sum_{j=1}^n (X_j - \mu) \tag{70}$$

converges in distribution to a normal($0, \Sigma$) distribution as $n \rightarrow \infty$.

The matrix Σ is not positive definite; it is only positive semidefinite. This is not astonishing, since $(N_1 - np_1, \dots, N_d - np_d)^\top$ takes values in the $d - 1$ dimensional subspace $\{(x_1, \dots, x_d)^\top \in \mathbb{R}^d : x_1 + \dots + x_d = 0\}$ of $\mathbb{R}^d$, as a consequence of the constraint

$$\sum_{i=1}^d (N_i - np_i) = \sum_{i=1}^d N_i - n \sum_{i=1}^d p_i = n - n = 0.$$

In order to get a simpler geometrical structure, it is convenient to scale (70) by $1/\sqrt{p_i}$:

Definition 2.24 (X^2 test statistic) We define the scaled difference vector X of the observed values $N_1, \dots, N_d$ from their expectation:

$$X = \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)_{i=1 \dots d}. \quad (71)$$

Its squared euclidean norm is called the X^2 test statistic:

$$X^2 = \|X\|^2 = \sum_{i=1}^d \frac{(N_i - np_i)^2}{np_i}. \quad (72)$$

Roughly speaking, the X^2 test statistic measures, how far the observed counts $N_1, \dots, N_d$ are from their expectation $np_1, \dots, np_d$. If X^2 is large, then one has observed considerable deviations from the expected values.

Since X is just a linear transformation of $n^{-1/2}(N_1 - np_1, \dots, N_d - np_d)^\top$, it converges in distribution to some multidimensional (degenerate) normal distribution, too.

Let us calculate the covariance matrix $\text{Cov}(X)$ of X :

$$\begin{aligned} \text{Cov}(X) &= \text{diag} \left(\frac{1}{\sqrt{p_1}}, \dots, \frac{1}{\sqrt{p_d}} \right) \Sigma \text{diag} \left(\frac{1}{\sqrt{p_1}}, \dots, \frac{1}{\sqrt{p_d}} \right) \\ &= I_d - (\sqrt{p_1}, \dots, \sqrt{p_d})^\top (\sqrt{p_1}, \dots, \sqrt{p_d}). \end{aligned} \quad (73)$$

We interpret $\text{Cov}(X)$ geometrically: $v = (\sqrt{p_1}, \dots, \sqrt{p_d})^\top$ is a unit vector, since

$$\|v\|^2 = \sum_{i=1}^d p_i = 1.$$

Thus $vv^\top$ is the orthogonal projection onto the one-dimensional space spanned by v .

(Indeed: If $a \in \mathbb{R}^d$ is proportional to v , say $a = \lambda v$, then $vv^\top a = \lambda vv^\top v = \lambda v\|v\|^2 = a$. On the other hand, if $a \in \mathbb{R}^d$ is orthogonal to v , i.e. $v^\top a = 0$, then $vv^\top a = 0$.)

As a consequence, $\text{Cov}(X) = I_d - vv^\top$ is the orthogonal projection to the $(d - 1)$ -dimensional orthogonal complement

$$v^\perp = \{a \in \mathbb{R}^d : v^\top a = 0\}$$

of the one-dimensional space spanned by v . Thus we get:

Lemma 2.25 As $n \rightarrow \infty$, the scaled difference vector X converges in distribution to a $\text{normal}(0, I_d - vv^\top)$ distribution. The covariance matrix $I_d - vv^\top$ is the orthogonal projection to the orthogonal complement $v^\perp$ of v , which has $d - 1$ dimensions.

The X^2 statistic globally measures the deviations of $(N_1, \dots, N_d)$ from the expectation vector $(np_1, \dots, np_d)$. This motivates the following test:

χ^2 test for simple hypotheses.

Suppose we have n i.i.d. observations $Y_1, \dots, Y_n$ with values in a finite set $\Omega = \{A_1, \dots, A_d\}$ and an unknown distribution $Y_j \sim P$. For $i = 1 \dots d$, let N_i denote the number of occurrences of the value A_i among $Y_1, \dots, Y_n$; thus $(N_1, \dots, N_d)$ follows a multinomial distribution with unknown parameters.

Given a probability distribution P_0 over Ω , $P_0[\{A_i\}] = p_i > 0$, consider the null hypothesis

$$H_0 : \quad "P = P_0"$$

versus the alternative hypothesis

$$H_A : \quad "P \neq P_0".$$

With respect to H_0 ,

$$(N_1, \dots, N_d) \sim \text{multinomial}(n, p_1, \dots, p_d).$$

The χ^2 test with critical value c has the rejection region

$$R = \{X^2 > c\},$$

where the test statistic X^2 is defined by (72).

Given a significance level α , the following theorem helps to choose critical values c for the χ^2 -test. Also, it helps us to approximately determine the p-value for the χ^2 test.

Theorem 2.26 (Pearson's theorem for simple hypotheses)
Asymptotics of the X^2 test statistics.

Assume that $(N_1, \dots, N_d) \sim \text{multinomial}(n, p_1, \dots, p_d)$. As the number n of observations tends to ∞ , the X^2 test statistic

$$X^2 = \sum_{i=1}^d \frac{(N_i - np_i)^2}{np_i}$$

converges in distribution to a χ^2 distribution with $d - 1$ degrees of freedom.

Roughly speaking, this means that the critical value c for a χ^2 test at significance level α is approximately the $(1 - \alpha)$ -quantile of the χ_{d-1}^2 distribution ($d - 1$ degrees of freedom), at least for large values of n . But what is a "large value"? For reasonable approximations, one may apply the following rule of thumb (due to F. Hampel, ETH Zurich):

For about 4/5 of all categories A_i , the expected number np_i has to be ≥ 4 , and for all other categories it has to be ≥ 1 .

Proof of Theorem 2.26. We know already:

$$X \xrightarrow{d} \text{normal}(0, \Pi)$$

as $n \rightarrow \infty$, where " $\xrightarrow{d}$ " means "convergence in distribution", and $\Pi = I_d - vv^T$ denotes the orthogonal projection onto the $(d - 1)$ -dimensional space $v^\perp$. On the other hand, if $Z \sim \text{normal}(0, I_d)$, i.e. if Z is

a d -dimensional standard normal vector, then $\Pi Z \sim \text{normal}(0, \Pi I_d \Pi^\top) = \text{normal}(0, \Pi)$, i.e. ΠZ has the same distribution as X in the limit as $n \rightarrow \infty$. Thus X^2 converges in distribution to the distribution of $\|\Pi Z\|^2$, which is the χ_{d-1}^2 distribution.

(Recall that χ_{d-1}^2 was defined to be the distribution of $\|Y\|^2$, where $Y \sim \text{normal}(0, I_{d-1})$. By rotational invariance of the d -dimensional standard normal distribution, it is the same as the distribution of $\|\Pi Z\|^2$, where $Z \sim \text{normal}(0, I_d)$, for every orthogonal projection Π to some $(d-1)$ -dimensional space. This is most obvious for the standard projection $(x_1, \dots, x_d) \mapsto (x_1, \dots, x_{d-1}, 0)$, and then, by applying an orthogonal transformation, for any orthogonal projection Π of rank $d-1$.)

□

Here is another rule of thumb for the 95% quantile of the χ^2 -distribution:

The 95% quantile of the χ_ν^2 distribution is approximately $\nu + 2\sqrt{2\nu}$. Thus the rejection region for the χ^2 test with ν degrees of freedom on the 5% significance level is approximately

$$R_{5\%} \approx \{X^2 > \nu + 2\sqrt{2\nu}\}. \quad (\text{rough rule of thumb})$$

(To justify this rule of thumb, note that the χ_ν^2 distribution has the expectation ν , which is n times the expectation of χ_1^2 , and the variance 2ν , which is ν times the variance of χ_1^2 . Let X_ν^2 denote a χ_ν^2 distributed random variable. By the central limit theorem, $(X_\nu^2 - \nu)/\sqrt{2\nu}$ converges to a standard normal distribution as $\nu \rightarrow \infty$. The 95% quantile of the standard normal distribution is $\Phi^{-1}(0.95) = 1.64485$ (rounded). Thus for large degrees of freedom ν , $R_{5\%} \approx \{X^2 > \nu + 1.64485\sqrt{2\nu}\}$. Replacing $\Phi^{-1}(0.95)$ by 2 makes the rule worse for large values of ν , but better and easier to use for moderate values of ν .)

Example 2.21 (continued)

Recall the data from the introductory example: $n = 1000$,

party	A	B	C	D
elections (p_i)	39.6%	31.5%	21.7%	7.2%
survey (N_i)	375	326	209	90

We calculate the X^2 test statistic (rounded result):

$$X^2 = \frac{(375 - 396)^2}{396} + \frac{(326 - 315)^2}{315} + \frac{(209 - 217)^2}{217} + \frac{(90 - 72)^2}{72} = 6.293.$$

With respect to H_0 : $(N_1, N_2, N_3, N_4) \sim \text{multinomial}(1000, 39.6\%, 31.5\%, 21.7\%, 7.2\%)$; the X^2 statistic is approximately distributed according to a χ^2 distribution with $4 - 1 = 3$ degrees of freedom. Using this approximation, we get the approximate p -value

$$p = \chi_3^2[[6.293, \infty]] = \frac{2^{-3/2}}{\Gamma(3/2)} \int_{6.293}^{\infty} e^{-x/2} x^{3/2-1} dx = 9.82\%.$$

(Recall that the χ_d^2 distribution has the probability density

$$f_d(x) = \frac{2^{-d/2}}{\Gamma(d/2)} e^{-x/2} x^{d/2-1},$$

since $\chi_d^2 = \text{Gamma}(d/2, 1/2)$.)

On significance levels smaller than 9.82%, we cannot reject H_0 . This number is not extremely small; we better do not reject H_0 unless we are willing to tolerate a type I error in roughly 10% of all cases when H_0 is true. We interpret this as follows: Either there is really no change in the opinion of the people, or

our sample size is too small to significantly (e.g. on the 5% level) give evidence for the change of opinions.

(The rule of thumb leads to the same conclusion: $R_{5\%} \approx \{X^2 > 3 + 2\sqrt{6}\} \approx \{X^2 > 7.9\}$; thus we do not reject the null hypothesis on the 5% level, according to the rule of thumb. One may compare this with the true 95% quantile of the χ^2_3 distribution: $\chi^2_3[[7.81472 \dots, \infty]] = 5\%$.)

The χ^2 test (and its more general version, treated below) can also be applied for i.i.d. random variables with values in any result space Ω , possibly infinite: one just splits Ω in “cells”:

Choosing cells. Let $Y_1, \dots, Y_n$ be i.i.d. random variables with values in a (possibly large or infinite) result space Ω and an unknown distribution $Y_j \sim P$. We partition the result space in d pairwise disjoint events, called “cells”:

$$\Omega = A_1 \cup \dots \cup A_d.$$

Let

$$N_i = \sum_{j=1}^n 1_{\{Y_j \in A_i\}}, \quad i = 1 \dots d,$$

denote the number of Y_j with value in the i -th cell A_i . Then

$$(N_1, \dots, N_d) \sim \text{multinomial}(n, p_1, \dots, p_d),$$

where $p_i = P[A_i]$.

In particular, one can apply the χ^2 test to $(N_1, \dots, N_d)$ for testing a null hypothesis H_0 : “ $Y_j \sim P$ ” versus the alternative H_0 : “ $Y_j \not\sim P$ ”, if the numbers np_i are not too small.

How to choose the cells A_i ? On the one hand, choosing them too large and too few of them, one loses a lot of information, thus reducing the power of the test. On the other hand, choosing too many of them, the approximation by the central limit theorem is weakened. There is no general rule; however, the first rule of thumb mentioned above helps.

Example 2.27 (Goodness of fit to a standard normal distribution) Suppose we are given $n = 1000$ i.i.d. random numbers $Y_1, \dots, Y_n$. We want to test the null hypothesis that $Y_j \sim \text{standard-normal}$ versus the alternative $Y_j \not\sim \text{standard-normal}$. For example, let us choose the cells $A_1 =]-\infty, -2]$, $A_2 =]-2, -1]$, $A_3 =]-1, -0.5]$, $A_4 =]-0.5, 0]$, $A_5 =]0, 0.5]$, $A_6 =]0.5, 1]$, $A_7 =]1, 2]$, $A_8 =]2, \infty[$. Let N_i denote the number of Y_j in cell A_i . To have a concrete example, we take the following data:

i	1	2	3	4	5	6	7	8
N_i	29	128	153	209	191	134	130	26
p_i	2.275%	13.591%	14.988%	19.146%	19.146%	14.988%	13.591%	2.275%

The third line in this table contains the (rounded) values

$$p_i = \text{standard-normal}[A_i] = \frac{1}{\sqrt{2\pi}} \int_{A_i} e^{-x^2/2} dx.$$

Let us calculate the X^2 test statistic (rounded):

$$X^2 = \sum_{i=1}^8 \frac{(N_i - np_i)^2}{np_i} = 6.25.$$

We use a χ^2 distribution with $7 = 8 - 1$ degrees of freedom. Since the p-value

$$\chi_7^2[[6.25, \infty]] = \frac{2^{-7/2}}{\Gamma(7/2)} \int_{6.25}^{\infty} e^{-x/2} x^{7/2-1} dx = 51.1\% \quad (\text{rounded})$$

is certainly not a small number, we do not reject the null hypothesis.

(The rule of thumb yields the approximate 95% quantile $7 + 2\sqrt{2 \cdot 7} \approx 14.5$ of the χ_7^2 distribution. Since $X^2 = 6.25$ is much smaller than 14.5, the rough rule of thumb yields the same conclusion as the precise calculation. In fact, the above data was really generated by using standard normal random numbers.)

2.11 The χ^2 test for composite hypotheses.

Example 2.28 $n = 100$ candidates participated in a multiple choice test, consisting of two questions a) and b). For the question a), there were 2 possible answers (“T, F”), and for the second question b), there were 3 possible answers (“A, B, C”). Here is a table, how many candidates decided for a possible combination of answers:

	A	B	C
T	8	15	7
F	25	21	24

We would like to test the hypothesis that the answers to the two questions were given independently of each other.

Let us put this example in a more general framework. We first develop a frame model: Let $Y_j^a \in \Omega_a = \{T, F\}$ be the random answer of candidate no. j to question a), and $Y_j^b \in \Omega_b = \{A, B, C\}$ be the answer of the same candidate to question b). More generally, let us assume that there are a possible answers for question a) and b possible answers for question b):

$$Y_j^a \in \Omega_a = \{A_1, \dots, A_a\}$$

and

$$Y_j^b \in \Omega_b = \{B_1, \dots, B_b\}$$

The compound random variable

$$Y_j = (Y_j^a, Y_j^b) \in \Omega = \Omega_a \times \Omega_b$$

encodes both answers of the j -th candidate. The number of candidates giving the compound answer (A_i, B_k) is

$$N_{i,k} = \sum_{j=1}^n 1_{\{Y_j = (A_i, B_k)\}}.$$

Let us assume that the Y_j , $j = 1, \dots, n$, are i.i.d. random variables; i.e. we use the frame model with the result space Ω^n and the joint probability distributions

$$\mathcal{P} = \{P^n : P \text{ is a non-degenerate probability distribution on } \Omega\}. \quad (74)$$

(“non-degenerate” here means that $P[\{\omega\}] > 0$ should hold for all $\omega \in \Omega$. We include this condition for technical reasons, to avoid a division by 0 later. Intuitively, it means that every possible combination of answers should have a positive probability to occur.)

Our null hypothesis “the answers to different questions are independent” can be formalized as

$$H_0 = \left\{ (P_a \otimes P_b)^n : \begin{array}{l} P_a \text{ and } P_b \text{ are non-degenerate probability} \\ \text{distributions on } \Omega_a \text{ and } \Omega_b, \text{ respectively} \end{array} \right\} \subseteq \mathcal{P}. \quad (75)$$

$\mathcal{P}$ is a parametric model; it can be parametrized by the parameter vector

$$(p_{ik})_{i=1\dots a, k=1\dots b} = (P^n[Y_j = (A_i, B_k)])_{i=1\dots a, k=1\dots b} \in \Theta = \{(p_{ik}) \in]0, 1[^{a \times b} : \sum_{i=1}^a \sum_{k=1}^b p_{ik} = 1\}.$$

Note that the parameters are not independent, due to the constraint $\sum_i \sum_k p_{ik} = 1$. Thus Θ is not ab -dimensional; it is only $ab - 1$ dimensional. Writing P_θ^n for the distribution corresponding to the parameter vector $\theta \in \Theta$, we have

$$H_0 = \{P_\theta^n : \theta \in \Theta_0\}$$

with

$$\Theta_0 = \{(p_i q_k)_{i=1\dots a, k=1\dots b} : \sum_{i=1}^a p_i = 1, \sum_{k=1}^b q_k = 1, \text{ all } p_i, q_k > 0\}.$$

Θ_0 is a $a + b - 2$ -dimensional *submanifold* of Θ .

(A subset M of an open set $\Theta \subseteq \mathbb{R}^m$ is called a ν -dimensional *submanifold*, if for every point $\theta \in M$ there is a neighborhood U of θ , an open set V in $\mathbb{R}^\nu$, and a bijective smooth function $f : V \rightarrow M \cap U$ with Jacobi matrix Df of full rank ν at every point, such that f^{-1} is continuous. Roughly speaking, this means:

M can be locally parametrized by ν independent smooth parameters.

Equivalently, M is a ν -dimensional submanifold of Θ when for every point $\theta \in M$ there is an open neighbourhood U of θ and a smooth map $g : U \rightarrow \mathbb{R}^{m-\nu}$ with Jacobi matrix Dg of full rank $m - \nu$ at every point, such that $M \cap U = \{x \in U : g(x) = 0\}$. Roughly speaking, this means:

M can be locally described by $m - \nu$ independent smooth equations $g_1(x_1, \dots, x_m) = 0, \dots, g_{m-\nu}(x_1, \dots, x_m) = 0$.

Thus a 1-dimensional submanifold is just a smooth curve, a 2-dimensional submanifold is a smooth surface, and so on.)

Indeed, Θ_0 is parametrized by

$$(p_1, \dots, p_a, q_1, \dots, q_b) \mapsto (p_i q_k)_{i=1\dots a, k=1\dots b}.$$

Two of the parameters (e.g. p_a and q_b) are redundant, due to the two “sum = 1” constraints. Thus there are only $(a - 1) + (b - 1)$ *independent* parameters. The parametrization has full rank $(a - 1) + (b - 1)$, since it has a smooth inverse, given by

$$p_i = \sum_{k=1}^b p_{ik}, \quad q_k = \sum_{i=1}^a p_{ik}.$$

Thus Θ_0 is indeed a $a + b - 2 = (a - 1) + (b - 1)$ -dimensional submanifold of Θ .

We generalize our frame model even further, taking an arbitrary submanifold Θ_0 of Θ as parameter set for the null hypothesis:

χ^2 test for composite hypotheses. Assume we have observed numbers of counts

$$(N_1, \dots, N_d) \sim P_\theta = \text{multinomial}(n, p_1, \dots, p_d)$$

with unknown parameters $\theta = (p_1, \dots, p_d) \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is described by the constraints $\sum_{i=1}^d p_i = 1$, all $p_i > 0$. Let $\Theta_0 \subseteq \Theta$ be a ν -dimensional submanifold of Θ . Consider the null hypothesis

$$H_0: "(p_1, \dots, p_d) \in \Theta_0"$$

versus the alternative hypothesis

$$H_A: "(p_1, \dots, p_d) \notin \Theta_0".$$

Let $\hat{\theta} = (\hat{p}_1, \dots, \hat{p}_d) \in \Theta_0$ be the *maximum-likelihood estimator* for the unknown parameters, i.e. the parameter vector $\hat{\theta} \in \Theta_0$ for which the likelihood function $\Theta_0 \ni \theta \mapsto P_\theta[\{(N_1, \dots, N_d)\}]$ takes its maximal value, assuming that there is a unique maximum.

Consider the X^2 statistic with *estimated* parameters

$$X^2 = \sum_{i=1}^d \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}.$$

The χ^2 test for testing H_0 with critical value c has the rejection domain

$$R = \{X^2 > c\}.$$

Roughly speaking, when using composite hypotheses for the χ^2 test, we just replace the parameters p_i in the X^2 statistic by their estimated values.

(Recall that the likelihood $P_\theta[\{(N_1, \dots, N_d)\}]$ is the probability of getting the observation $(N_1, \dots, N_d)$, provided the true parameter is θ . One can show that for all $\theta \in \Theta$, the probability (w.r.t. P_θ) of the estimator $\hat{\theta}$ to be well defined converges to 1 as $n \rightarrow \infty$. We will examine maximum likelihood estimators in more detail in the next chapter.)

Theorem 2.29 (Fisher's generalisation of Pearson's theorem) Let Θ_0 be a ν -dimensional submanifold of the $d - 1$ dimensional parameter space $\Theta \subset \mathbb{R}^d$. Then, as $n \rightarrow \infty$, with respect to all null distributions P_θ , $\theta \in \Theta_0$, the X^2 test statistic *with ν estimated parameters* converges in distribution to a χ^2 distribution *with $d - 1 - \nu$ degrees of freedom*.

Roughly speaking, starting with a d dimensional space, 1 degree of freedom is lost by the constraint $\sum_i p_i = 1$, and ν degrees of freedom are lost by estimating ν free parameters.

(Unfortunately, the proof of this theorem is rather complicated; we skip it. Roughly speaking, the spirit of the proof is the same as in example 2.11. Here is only some heuristics for the proof: Locally, it is a good approximation to replace the submanifold by its tangential space T . With respect to any given null distribution P_θ , $\theta \in \Theta_0$, as n gets large, the maximum likelihood estimator $\hat{\theta}$ converges in probability to the "true" value θ ; thus the problem gets "local". Locally, after scaling with $(1/\sqrt{\hat{p}_i})_{i=1, \dots, d}$, the maximum likelihood estimator $\hat{\theta}$ is well approximated by the orthogonal projection of the observed data to the tangential space of the submanifold. Thus the vector $X = ((N_i - n\hat{p}_i)/\sqrt{n\hat{p}_i})_i$, up to small error terms, varies only in the $d - 1 - \nu$ dimensional orthogonal space $T^\perp$ within the parameter space, and by the central limit theorem, it converges to a multidimensional standard normal distribution in $T^\perp$. Thus $\|X\|^2$ converges to a $\chi_{d-1-\nu}^2$ distribution.)

Let us apply this theory to get a χ^2 test on independence. Reconsider the setup described in (74) and (75). Given observed counts $(N_{i,k})_{i=1\dots a, k=1\dots b}$, we determine the maximum likelihood estimator for the parameters $p_1, \dots, p_a, q_1, \dots, q_b$. Thus we maximize the likelihood function

$$f((p_i), (q_k), (N_{i,k})) = \frac{n!}{\prod_{i=1}^a \prod_{k=1}^b N_{i,k}!} \prod_{i=1}^a \prod_{k=1}^b (p_i q_k)^{N_{i,k}} \quad (76)$$

with respect to the constraints

$$\sum_{i=1}^a p_i = 1, \quad \sum_{k=1}^b q_k = 1. \quad (77)$$

We use Lagrange's method to take care of the constraints.

(Recall Lagrange's method: Suppose you want to maximize a smooth function $f(x_1, \dots, x_n)$ with respect to the k smooth (non-degenerate) constraints $g_i(x_1, \dots, x_n) = 0$, $i = 1, \dots, k$. If there is an extremum at x , then there are numbers $\mu_1, \dots, \mu_k$, called Lagrange multipliers, such that

$$\nabla f(x) = \sum_{i=1}^k \mu_i \nabla g_i(x).$$

Here

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^\top$$

denotes the gradient of f .)

The first $N_{i,k}$ dependent factors in (76) are irrelevant for the maximization; and we simplify the maximization problem by taking logarithms. Thus we want to maximize the concave function

$$\sum_{i=1}^a \sum_{k=1}^b N_{i,k} (\log p_i + \log q_k) \stackrel{!}{=} \text{maximal}$$

with respect to the two constraints (77), which define an affine-linear subspace. In particular, every critical point for this maximization problem is really a maximum. Introducing the Lagrange multipliers μ_1, μ_2 , we get the equations

$$\sum_{k=1}^b \frac{N_{i,k}}{\hat{p}_i} = \mu_1, \quad i = 1, \dots, a$$

and

$$\sum_{i=1}^a \frac{N_{i,k}}{\hat{q}_k} = \mu_2, \quad k = 1, \dots, b.$$

Thus $\hat{p}_i$ is proportional to $\sum_k N_{i,k}$ and $\hat{q}_k$ is proportional to $\sum_i N_{i,k}$. Using the constraints (77), we conclude

$$\hat{p}_i = \frac{\sum_{k=1}^b N_{i,k}}{\sum_{i'=1}^a \sum_{k=1}^b N_{i',k}} = \frac{1}{n} \sum_{k=1}^b N_{i,k}$$

and

$$\hat{q}_k = \frac{\sum_{i=1}^a N_{i,k}}{\sum_{i=1}^a \sum_{k'=1}^b N_{i,k'}} = \frac{1}{n} \sum_{i=1}^a N_{i,k}.$$

The result is just what one would intuitively guess: For estimating the probability that a candidate answers "A_i" in the first question, just ignore the second question, and take the frequency $\sum_i N_{i,k}/n$ of the answer "A_i" in the observations as your estimate.

Since we have estimated $a + b - 2$ out of $ab - 1$ free parameters, the remaining number of degrees of freedom is

$$(ab - 1) - (a + b - 2) = (a - 1)(b - 1).$$

Let us summarize:

χ^2 test on independence. In the setup (74) and (75), assume we are given a table with numbers of counts of compound observations $Y_j = (A_i, B_k)$:

	B ₁	B ₂	...	B _b	sum
A ₁	$N_{1,1}$	$N_{1,2}$	...	$N_{1,b}$	$N_{1,*}$
A ₂	$N_{2,1}$	$N_{2,2}$	...	$N_{2,b}$	$N_{2,*}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A _a	$N_{a,1}$	$N_{a,2}$	...	$N_{a,b}$	$N_{a,*}$
sum	$N_{*,1}$	$N_{*,2}$	...	$N_{*,b}$	$N_{*,*} = n$

Using the estimators $\hat{p}_i = N_{i,*}/n$ and $\hat{q}_k = N_{*,k}/n$, evaluate the X^2 statistic

$$X^2 = \sum_{i=1}^a \sum_{k=1}^b \frac{(N_{i,k} - n\hat{p}_i\hat{q}_k)^2}{n\hat{p}_i\hat{q}_k}.$$

The rejection domain for the null hypothesis H_0 : “ Y_j^a, Y_j^b are independent” on the significance level α is approximately

$$R = \{X^2 > c\},$$

where c denotes the $(1 - \alpha)$ -quantile of the χ^2 distribution with $(a - 1)(b - 1)$ degrees of freedom: $\chi_{(a-1)(b-1)}^2 \llbracket c, \infty \rrbracket = \alpha$. The approximation gets asymptotically exact in the limit $n \rightarrow \infty$.

Example 2.28 (continued) Recall the data from the introductory example:

	A	B	C	sum
T	8	15	7	30
F	25	21	24	70
sum	33	36	31	100

We estimate the unknown parameters within the null hypothesis: $\hat{p}_1 = 30\%$, $\hat{p}_2 = 70\%$, $\hat{q}_1 = 33\%$, $\hat{q}_2 = 36\%$, $\hat{q}_3 = 31\%$. We calculate the X^2 statistic:

$$\begin{aligned} X^2 &= \frac{(8 - 100 \cdot 0.3 \cdot 0.33)^2}{100 \cdot 0.3 \cdot 0.33} + \frac{(15 - 100 \cdot 0.3 \cdot 0.36)^2}{100 \cdot 0.3 \cdot 0.36} + \frac{(7 - 100 \cdot 0.3 \cdot 0.31)^2}{100 \cdot 0.3 \cdot 0.31} \\ &\quad + \frac{(25 - 100 \cdot 0.7 \cdot 0.33)^2}{100 \cdot 0.7 \cdot 0.33} + \frac{(21 - 100 \cdot 0.7 \cdot 0.36)^2}{100 \cdot 0.7 \cdot 0.36} + \frac{(24 - 100 \cdot 0.7 \cdot 0.31)^2}{100 \cdot 0.7 \cdot 0.31} \\ &= 3.66685 \quad (\text{rounded}). \end{aligned} \tag{78}$$

We use the χ^2 distribution with $(2 - 1) \cdot (3 - 1) = 2$ degrees of freedom. We get the (approximate) p-value

$$\chi_2^2 \llbracket 3.66685, \infty \rrbracket = 15.98\% \quad (\text{rounded}).$$

This is not a small number; thus we do not reject the null hypothesis that the questions were independently answered.

(The thumb rule yields the approximate 95% quantile $2 + 2\sqrt{2 \cdot 2} = 6$ of the χ_2^2 distribution, which is much larger than X^2 . It thus leads to the same conclusion: We do not reject the null hypothesis on the 5% significance level.)

2.12 The one-sample t test

Example 2.30 In the year 1879, the famous American physicist A. Michelson measured repeatedly the speed of light. In one series of $n = 20$ trials, he obtained the following data (in km/sec):

299850, 299740, 299900, 300070, 299930, 299850, 299950, 299980, 299980, 299880,
300000, 299980, 299930, 299650, 299760, 299810, 300000, 300000, 299960, 299960.

Assuming that the measurement yields i.i.d. data with a normal distribution with unknown parameters μ and σ , we ask for a 99% confidence interval for μ , and another one for σ . (We ignore rounding effects in the data.)

More generally, let us assume that the joint distribution of random values $X_1, \dots, X_n$ follows the frame model

$$\mathcal{P} = \{\text{normal}(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

over $\Omega = \mathbb{R}^n$. Given $\mu_0 \in \mathbb{R}$, let us develop a test for testing $H_0: \mu = \mu_0$ versus $H_A: \mu \neq \mu_0$. If we knew σ^2 , we could take the normalized test statistic:

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma},$$

where

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$$

is the sample average. Note that with respect to H_0 , $\bar{X} \sim \text{normal}(\mu_0, \sigma^2/n)$ holds; thus $Z \sim \text{standard-normal}$.

However, we do not know σ . Thus we replace the (unknown) σ by an estimate, derived from the data.

Definition 2.31 The empirical variance s_X^2 of data $X_1, \dots, X_n$ is defined by

$$s_X^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2.$$

Its square root s_X is called the empirical standard deviation.

(Note that in the definition of s_X^2 we divide by $n-1$, not by n . The reason for this choice is that $E[s_X^2] = \sigma^2$ only holds for this definition, but not when dividing by n ; see below. The smaller denominator “ $n-1$ ” compensates that we take $\sum_j (X_j - \bar{X})^2$, not $\sum_j (X_j - \mu)^2$, since we cannot observe the latter.)

Thus we define the test statistic

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s_X},$$

which can be determined using only the observed data. Astonishingly, with respect to $H_0: \mu = \mu_0$, the numerator $\sqrt{n}(\bar{X} - \mu)$ and the denominator s_X in the definition of T are *independent* random variables, although both are derived from the same sample:

Lemma 2.32 Let $X_1, \dots, X_n$ be i.i.d. $\text{normal}(\mu, \sigma^2)$ random variables. Then

$$\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}, \frac{(n-1)s_X^2}{\sigma^2} \right) \sim \text{standard-normal} \otimes \chi_{n-1}^2.$$

As a consequence,

$$E[s_X^2] = \sigma^2.$$

As a consequence, with respect to H_0 , the test statistic T has the same distribution as

$$T_{n-1} = \frac{X}{\sqrt{\frac{1}{n-1}Y_{n-1}}},$$

where $(X, Y_{n-1}) \sim \text{standard-normal} \otimes \chi_{n-1}^2$.

Proof of Lemma 2.32. Consider the unit vector

$$v = \frac{1}{\sqrt{n}}(1, \dots, 1)^\top \in \mathbb{R}^n.$$

Let

$$\begin{aligned} Z_j &= \frac{X_j - \mu}{\sigma} \sim \text{standard-normal}, \\ Z &= (Z_1, \dots, Z_n)^\top \sim \text{normal}(0, I_n). \end{aligned}$$

We get

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \sqrt{n}\bar{Z} = \frac{1}{\sqrt{n}} \sum_{j=1}^n Z_j = v^\top Z \quad (79)$$

and

$$(n-1)\frac{s_X^2}{\sigma^2} = (n-1)s_Z^2 = \sum_{j=1}^n (Z_j - \bar{Z})^2 = \|Z - vv^\top Z\|^2 = \|(I_n - vv^\top)Z\|^2. \quad (80)$$

(Recall that $I_n - vv^\top$ is the orthogonal projection to the $n-1$ dimensional subspace $v^\perp$ orthogonal to v .)

By rotational invariance of the n dimensional standard normal distribution, the joint distribution of $(v^\top Z, \|(I_n - vv^\top)Z\|^2)$ is the same for all unit vectors v . In particular, we may replace v by the standard unit vector $e = (0, \dots, 0, 1)^\top \in \mathbb{R}^n$:

$$\begin{aligned} &(v^\top Z, \|(I_n - vv^\top)Z\|^2) \\ &\sim (e^\top Z, \|(I_n - ee^\top)Z\|^2) = (Z_n, \|(Z_1, \dots, Z_{n-1})\|^2) \\ &\sim \text{standard-normal} \otimes \chi_{n-1}^2. \end{aligned}$$

Combining this with (79) and (80) proves the first part of the lemma. Furthermore,

$$E[s_x^2] = \frac{\sigma^2}{n-1} E[\chi_{n-1}^2] = \sigma^2 E[\chi_1^2] = \sigma^2 \text{Var}[\text{standard-normal}] = \sigma^2.$$

(We use the same symbol for the expectation of a random variable and for the expectation of the corresponding distribution.)

□

The distribution of the test statistic T is known as the *Student t distribution with $n-1$ degrees of freedom*:

Definition 2.33 (The Student t distribution) *The t distribution with n degrees of freedom, denoted by t_n , is defined to be the distribution of*

$$T_n = \frac{X}{\sqrt{\frac{1}{n}Y_n}},$$

where $(X, Y_n) \sim \text{standard-normal} \otimes \chi_n^2$.

Lemma 2.34 Let $T_n \sim t_n$. Then as $n \rightarrow \infty$, T_n converges in distribution to a standard normal distribution.

Proof. Recall: The χ_n^2 distribution is the distribution of $\|Z\|^2 = \sum_{i=1}^n Z_i^2$, where $Z = (Z_1, \dots, Z_n) \sim \text{normal}(0, I_n)$. By the weak law of large numbers, using $E[Z_i^2] = \text{Var}[Z_i] = 1$, we know

$$\frac{Y_n}{n} = \frac{1}{n} \sum_{i=1}^n Z_i^2 \xrightarrow{n \rightarrow \infty} 1 \quad \text{in probability.}$$

(Recall the *weak law of large numbers*: For i.i.d. random variables $U_1, U_2, U_3, \dots$ with finite expected value $\mu = E[U_i]$, the following holds for all $\delta > 0$:

$$P \left[\left| \frac{1}{n} \sum_{i=1}^n U_i - \mu \right| < \delta \right] \xrightarrow{n \rightarrow \infty} 1,$$

i.e. $\sum_{i=1}^n U_i/n \rightarrow \mu$ in probability.)

Taking $X \sim \text{standard-normal}$, independent of Y_n , we get that $T_n = X/\sqrt{Y_n/n}$ converges in distribution to the distribution of $X/1$, i.e. to the standard normal distribution. □

Lemma 2.35 (Density of the t distribution) The t_n distribution has the density

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(\frac{t^2}{n} + 1\right)^{-(n+1)/2}. \quad (81)$$

(As a special case, the t_1 distribution is the Cauchy distribution $\text{Cauchy}(1)$. By definition, the Cauchy distribution $\text{Cauchy}(a)$ with parameter $a > 0$ has the probability density

$$g(t) = \frac{1}{\pi} \frac{a}{a^2 + t^2}.$$

The Cauchy distribution has neither an expectation nor a variance, since the density $g(t)$ decreases only slowly as $t \rightarrow \infty$. Note that the density of t_n decreases only like $O(|t|^{-n-1})$ as $|t| \rightarrow \infty$, i.e. only like a power law. Thus when $T_n \sim t_n$, one has only $P[|T_n| > t] = O(t^{-n})$ as $t \rightarrow \infty$. This should be opposed to normal distributions, which decay superexponentially fast at infinity. Intuitively, writing $T_n = X/\sqrt{Y_n/n}$ according to the definition of t_n , T_n takes large values with a much higher probability than X does. Typically, this does not occur because X takes large values, but because of small values of Y_n . This fact is important for t tests, explained below, at very small significance levels.)

Proof of Lemma 2.35. Let $(X, Y_n) \sim \text{standard-normal} \otimes \chi_n^2$. These random variables have the joint density

$$f_{(X, Y_n)}(x, y) = f_X(x) f_{Y_n}(y),$$

where

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and

$$f_{Y_n}(y) = \frac{2^{-n/2}}{\Gamma(n/2)} y^{n/2-1} e^{-y/2}.$$

(Recall that $\chi_n^2 = \text{Gamma}(n/2, 1/2)$.)

We set $Q = X/\sqrt{Y_n}$. We derive first the joint probability density $f_{(Q, Y_n)}$ of (Q, Y_n) : Consider the

transformation $G : \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R} \times \mathbb{R}^+$, $G(x, y) = (x/\sqrt{y}, y)$. It has the inverse $G^{-1}(q, y) = (q\sqrt{y}, y)$ with the Jacobi determinant

$$\det DG^{-1}(q, y) = \begin{vmatrix} \sqrt{y} & \frac{q}{2\sqrt{y}} \\ 0 & 1 \end{vmatrix} = \sqrt{y}.$$

By the transformation rule for probability densities, we get

$$f_{(Q, Y_n)}(q, y) = f_{(X, Y_n)}(G^{-1}(q, y)) |\det DG^{-1}(q, y)| = f_X(q\sqrt{y}) f_Y(y) \sqrt{y}.$$

To get the marginal density f_Q for Q , we integrate over y :

$$\begin{aligned} f_Q(q) &= \int_0^\infty f_X(q\sqrt{y}) f_Y(y) \sqrt{y} dy \\ &= \frac{1}{\sqrt{2\pi}} \frac{2^{-n/2}}{\Gamma(\frac{n}{2})} \int_0^\infty e^{-(q\sqrt{y})^2/2} y^{(n-1)/2} e^{-y/2} dy \\ &= \frac{1}{\sqrt{2\pi}} \frac{2^{-n/2}}{\Gamma(\frac{n}{2})} \int_0^\infty e^{-(q^2+1)y/2} y^{(n-1)/2} dy \\ &= \frac{1}{\sqrt{2\pi}} \frac{2^{-n/2}}{\Gamma(\frac{n}{2})} \Gamma\left(\frac{n+1}{2}\right) \left(\frac{q^2+1}{2}\right)^{-(n+1)/2} \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi} \Gamma(\frac{n}{2})} (q^2+1)^{-(n+1)/2}. \end{aligned}$$

(We used

$$\int_0^\infty y^{s-1} e^{-\lambda y} dy = \Gamma(s) \lambda^{-s}$$

in the last but one step.)

By scaling with $\sqrt{n}$, we finally get the formula (81) for the density f_{T_n} of $T_n = \sqrt{n}Q$:

$$f_{T_n}(t) = \frac{1}{\sqrt{n}} f_Q\left(\frac{t}{\sqrt{n}}\right) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(\frac{t^2}{n} + 1\right)^{-(n+1)/2}.$$

□

Let us summarize:

One-sample t-test for the expectation of i.i.d. normal variables. Assume that $X_1, \dots, X_n$ are i.i.d. $\text{normal}(\mu, \sigma^2)$ random variables with unknown μ and unknown σ^2 . Given $\mu_0 \in \mathbb{R}$, consider the null hypothesis H_0 : “ $\mu = \mu_0$ ” versus one of the following alternative hypotheses:

1. “ $\mu > \mu_0$ ”,
2. “ $\mu < \mu_0$ ”,
3. “ $\mu \neq \mu_0$ ”.

With respect to H_0 , the test statistic

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s},$$

follows a Student t distribution with $n - 1$ degrees of freedom. For the three alternatives above, one gets the following rejection regions at significance level α :

1. $R_1 = \{T > q_{n-1, 1-\alpha}\}$,
2. $R_2 = \{T < q_{n-1, \alpha}\}$,
3. $R_3 = \{|T| > q_{n-1, 1-\alpha/2}\}$

respectively, where $q_{f, \alpha}$ denotes the α -quantile of the t_f distribution, i.e. $t_f[] - \infty, q_{f, \alpha}] = \alpha$. For large n and moderate significance levels, the t_{n-1} distribution is well approximated by the standard normal distribution.

Example 2.30 (continued) For the data $X_1, \dots, X_n$, $n = 20$, in the introductory example, we get the mean

$$\bar{X} = 299909 \frac{\text{km}}{\text{sec}}$$

and the empirical standard deviation

$$s_X = \sqrt{\frac{1}{19} \sum_{j=1}^{20} (X_j - \bar{X})^2} = 105 \frac{\text{km}}{\text{sec}} \quad (\text{rounded}).$$

Given μ_0 , the rejection region for the 2-sided t-test of the null hypothesis “ $\mu = \mu_0$ ” on the significance level α equals

$$R_\alpha(\mu_0) = \{|T_{\mu_0}| > q_{n-1, 1-\alpha/2}\},$$

where

$$T_{\mu_0} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s_X}$$

denotes the Student t test statistic, and $q_{n-1, 1-\alpha/2}$ denotes the $1 - \alpha/2 = 99.5\%$ quantile of the Student t distribution with $n - 1 = 19$ degrees of freedom. Thus we get the $1 - \alpha$ confidence interval

$$C_{1-\alpha} = \{\mu_0 : |T_{\mu_0}| \leq q_{n-1, 1-\alpha/2}\} = \left[\bar{X} - \frac{s_X q_{n-1, 1-\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{s_X q_{n-1, 1-\alpha/2}}{\sqrt{n}} \right]$$

Using the numerical value

$$q_{19, 99.5\%} = 2.861 \quad (\text{rounded}),$$

which solves the equation

$$\frac{\Gamma(10)}{\sqrt{19\pi}\Gamma(9.5)} \int_{q_{19, 99.5\%}}^{\infty} \left(\frac{t^2}{19} + 1 \right)^{-10} dt = 0.5\%,$$

we get the two-sided confidence interval

$$C_{99\%} = \left[299909 - \frac{105 \cdot 2.861}{\sqrt{20}}, 299909 + \frac{105 \cdot 2.861}{\sqrt{20}} \right] = [299842, 299976] \quad (\text{rounded; unit: km/sec}).$$

(Let us compare the 99.5% quantile $q_{19,99.5\%} = 2.861$ of the t distribution with 19 degrees of freedom with the corresponding quantile $q_{\infty,99.5\%} = 2.576$ of the standard normal distribution: We still see a considerable difference, due to the rather slow decay of the t distributions.)

The present-day value of the speed of light is 299792.5 km/sec, which is *not* contained in our confidence interval. Two interpretations are possible:

- We could be in one of the unlucky 1% cases, where the true parameter is not contained in the 99% confidence interval.
- More reasonable, the assumption that μ equals the true speed of light might fail to hold, due to systematic errors, or due to failure of the normality assumption.

A confidence interval for the variance. Next, we derive a one-sided confidence interval of the form $C_{1-\alpha} =]0, \sigma_{\max}^2]$ for the unknown variance σ^2 , given i.i.d. $\text{normal}(\mu, \sigma^2)$ data $X_1, \dots, X_n$ with unknown μ and σ . Given $\sigma_0^2 > 0$, consider the null hypothesis H_0 : “ $\sigma^2 = \sigma_0^2$ ” versus the alternative H_A : “ $\sigma^2 < \sigma_0^2$ ”. We take the empirical variance s_X^2 as a test statistic. By Lemma 2.32, with respect to H_0 , $(n-1)s_X^2/\sigma_0^2 \sim \chi_{n-1}^2$ holds irrespectively of the value of μ . Thus we get the rejection region at significance level α :

$$R_\alpha = \left\{ \frac{(n-1)s_X^2}{\sigma_0^2} < c_{n-1,\alpha} \right\} = \left\{ \sigma_0^2 > \frac{(n-1)s_X^2}{c_{n-1,\alpha}} \right\},$$

where $c_{n-1,\alpha}$ denotes the α quantile of the χ_n^2 distribution. By duality between tests and confidence regions, this yields the $1 - \alpha$ confidence region

$$C_{1-\alpha} = \left] 0, \frac{(n-1)s_X^2}{c_{n-1,\alpha}} \right]$$

for the unknown value of σ^2 .

2.13 Comparing two coupled samples.

Example 2.36 Do trouts need more oxygen in fast flowing water than in slowly flowing water? Here are experimental data, oxygen consumption in mm^3 per hour and per gramm weight of the fish:

fast (X_i)	108	122	96	129	98
slow (Y_i)	85	83	69	95	122

Is the oxygen consumption in the first row significantly higher than in the second row? For every column, let us assume that the data were obtained with the same fish individuals, once experimenting with fast water, and then with slowly flowing water. In this case, it is not reasonable to assume that the X_i are independent of the Y_i .

2.13.1 Comparing two coupled normal samples.

In this subsection, we treat this problem in a *parametric* model: We assume that the data follow a joint normal distribution. A non-parametric model for the same setup is treated in the next subsection, below. More precisely, we assume:

Frame model for coupled normal samples: (X_i, Y_i) , $i = 1, \dots, n$ are i.i.d. random vectors with an unknown normal distribution $(X_i, Y_i) \sim \text{normal}((\mu_X, \mu_Y), \Sigma)$ over $\mathbb{R}^2$, the expectation vector $(\mu_X, \mu_Y) \in \mathbb{R}^2$ and the covariance matrix $\Sigma \in \mathbb{R}^{2 \times 2}$ being unknown.

If there is no effect of the speed of the water on the oxygen consumption, then we could just exchange the X_i and Y_i without changing their joint distribution. Thus we are lead to the

Null hypothesis H_0 : “ X_i, Y_i are exchangeable random variables, i.e. (X_i, Y_i) has the same distribution as (Y_i, X_i) .”

In this frame model, $X_i - Y_i, i = 1, \dots, n$, are i.i.d. normal random variables.

With respect to H_0 , $\mu_X = \mu_Y$ holds, thus $E_0[X_i - Y_i] = \mu_X - \mu_Y = 0$ holds with respect to all null distributions $P_0 \in H_0$. Thus, with respect to H_0 , $X_i - Y_i, i = 1, \dots, n$, are i.i.d. $\text{normal}(0, \sigma^2)$ random variables with unknown variance $\sigma^2 = (1, -1)\Sigma(1, -1)^\top$.

t test for two coupled normal samples. We may just apply the one-sample t test to the difference data $X_i - Y_i, i = 1, \dots, n$.

We may use, for example, the alternative hypothesis “ $\mu_X > \mu_Y$ ”, i.e. “the expected oxygen consumption in fast water is higher than in slow water”.

2.13.2 Nonparametric methods for comparing two coupled samples.

Reconsider example 2.36: We have no good theoretical evidence that the data is described by a joint normal distribution.

Nonparametric frame model $\mathcal{P}$ for coupled samples: $(X_i, Y_i), i = 1, \dots, n$ are i.i.d. random vectors with an unknown continuous distribution $(X_i, Y_i) \sim P$ over $\mathbb{R}^2$.

The null hypothesis, informally described, just looks the same as in the parametric case, although it is certainly a much larger (nonparametric) class of probability distributions:

Null hypothesis H_0 : “ X_i, Y_i are exchangeable random variables, i.e. (X_i, Y_i) has the same distribution as (Y_i, X_i) .”

Let us not specify an alternative hypothesis at this moment; a standard choice could be $H_A: (X_i, Y_i) \not\sim (Y_i, X_i)$.

The sign test. Here is a simple nonparametric method for this setup:

With respect to all P in the frame model, $1_{\{X_i < Y_i\}}, i = 1, \dots, n$ are i.i.d. Bernoulli(p) random variables with unknown $p = P[X_i < Y_i]$.

(Recall: a $\{0, 1\}$ valued random variable X with $P[X = 1] = p = 1 - P[X = 0]$ is called a Bernoulli(p) random variable.)

As a consequence:

The sign test statistic

$$N = \sum_{i=1}^n 1_{\{X_i < Y_i\}}$$

is a binomial(n, p) random variable with unknown p with respect to the frame model $\mathcal{P}$.

Let $P_0 \in H_0$. Using that P_0 is a continuous distribution, we get

$$P_0[X_i < Y_i] = P_0[X_i \leq Y_i] = 1 - P_0[Y_i < X_i] = 1 - P_0[X_i < Y_i];$$

recall that X_i, Y_i are exchangeable with respect to P_0 . Thus

$$P_0[X_i < Y_i] = \frac{1}{2}.$$

Null distribution of the sign test statistic. For all $P_0 \in H_0$,

$$N \stackrel{P_0}{\sim} \text{binomial}(n, 1/2).$$

For testing H_0 , we may use just the rejection domains, depending on the alternative hypothesis, $\{N < c_1\}$, $\{N > c_2\}$, $\{|N - n/2| > c_3\}$, for some critical values c_1, c_2, c_3 . In the above example, the oxygen consumption increased in one out of 5 cases. For a one-sided test (rejection region $R = \{N < c_1\}$), we get the p value

$$p = \text{binomial}(5, 1/2)[\{0, 1\}] = 18.75\%.$$

For example on the significance level 5%, we cannot reject the hypothesis that the oxygen consumption does not decrease.

The main advantages of the sign test are its simplicity and its robustness. However, the power of the sign test is not very high.

A rank test for coupled samples. We will develop another nonparametric method for this setup in this section, without paying the price of additional assumptions, like restricting the frame model to a smaller class. When one does not have a good indication for using normal distributions in the frame model, the rank test described in this section is a good choice: It is robust, and its power is not much lower than the power of the t test.

Setup for the rank test of two coupled samples. The frame model and the null hypothesis are the same as for the sign test.

With respect to H_0 , $Z_j = X_j - Y_j$ has the same (continuous) distribution as $-Z_j = Y_j - X_j$, due to exchangeability. Thus $Z_1, \dots, Z_n$ are *symmetric* i.i.d. random variables. Let $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ be the (unique) permutation that reorders $|Z_1|, \dots, |Z_n|$ in increasing order:

$$|Z_{\sigma(1)}| < |Z_{\sigma(2)}| < |Z_{\sigma(3)}| < \dots < |Z_{\sigma(n)}|$$

(Since Z_j are i.i.d. variables with a *continuous* distribution, all numbers $|Z_1|, \dots, |Z_n|$ will be different with probability 1. Thus σ is well-defined with probability 1.)

Definition 2.37 (Rank sum test statistic - coupled samples) *The rank sum test statistic for coupled samples is defined by*

$$S = \sum_{j=1}^n j 1_{\{Z_{\sigma(j)} > 0\}}.$$

This means: we reorder the data $Z_1, \dots, Z_n$ in increasing order of their absolute values, and then take just the sum of the ranks of positive values among $Z_1, \dots, Z_n$.

Example 2.36 (continued) Let us consider the difference data:

X_i	108	122	96	129	98
Y_i	85	83	69	95	122
$Z_i = X_i - Y_i$	23	39	27	34	-24

We reorder the data according to $|Z_i|$:

j	1	2	3	4	5
$\sigma(j)$	1	5	3	4	2
$Z_{\sigma(j)}$	23	-24	27	34	39
$j1_{\{Z_{\sigma(j)} > 0\}}$	1	0	3	4	5

Thus the rank sum test statistic equals

$$S = 1 + 3 + 4 + 5 = 13.$$

(Intuitively, it is plausible that the rank sum S encodes more relevant information than the test statistic N for the sign test: Not only the sign of the Z_j is used, but we also compare their relative sizes. Also, it is intuitively clear that S is a sensitive measure for deviations of the distribution of Z_j from a symmetric distribution.)

The following lemma shows why S is useful as a test statistic in a nonparametric setup: Its null distribution is universal; it is the same distribution with respect to all $P_0 \in H_0$:

Lemma 2.38 *Let $Z_1, \dots, Z_n$ be any symmetric, i.i.d. random variables with a continuous distribution P , and let the permutation σ be defined as above. Then $B_j := 1_{\{Z_{\sigma(j)} > 0\}}$, $j = 1, \dots, n$, are i.i.d. Bernoulli(1/2) random variables. In particular, the null distribution of the rank sum test statistic*

$$S = \sum_{j=1}^n jB_j$$

does not depend on $P_0 \in H_0$.

Proof. Since Z_i has a continuous distribution, we know $Z_i \neq 0$ with probability 1. So let us assume $Z_i \neq 0$ for sure. Then

$$\text{sign}(Z_i) = \begin{cases} 1 & \text{for } Z_i > 0 \\ -1 & \text{for } Z_i < 0 \end{cases}$$

is well defined. Since Z_i is symmetric, the conditional distributions of $|Z_i|$, conditioned on $\text{sign}(Z_i) = +1$ and on $\text{sign}(Z_i) = -1$, are the same. In other words: $|Z_i|$ and $\text{sign}(Z_i)$ are independent.

(Recall the definition of *conditional probabilities*: $P[A|B] = P[A \cap B]/P[B]$, where $P[B] > 0$. Note that A and B are independent if and only if $P[A|B] = P[A]$. Furthermore, recall the definition of a conditional distribution of a random variable X , conditioned on an event B with $P[B] > 0$: It is just the distribution of X with respect to the conditional probability measure $A \mapsto P[A|B]$.)

Furthermore,

$$P[\text{sign}(Z_i) = +1] = P[\text{sign}(Z_i) = -1] = \frac{1}{2}.$$

Thus $1_{\{Z_i > 0\}}$, $i = 1, \dots, n$, are i.i.d. Bernoulli(1/2) random variables. The random permutation σ , which rearranges $|Z_1|, \dots, |Z_n|$ in increasing order, depends only on the $|Z_1|, \dots, |Z_n|$; hence it is independent of $(\text{sign}(Z_1), \dots, \text{sign}(Z_n))$. Reordering the i.i.d. Bernoulli(1/2) variables $1_{\{Z_i > 0\}}$, $i = 1, \dots, n$, by the random permutation σ , which is *independent* of $(1_{\{Z_1 > 0\}}, \dots, 1_{\{Z_n > 0\}})$, yields again Bernoulli(1/2) random variables. Thus

$$(B_1, \dots, B_n) = (1_{\{Z_{\sigma(1)} > 0\}}, \dots, 1_{\{Z_{\sigma(n)} > 0\}}) \sim \text{Bernoulli}(1/2)^n.$$

This proves the lemma. □

Note that the random variables jB_j in the definition of the test statistic S are independent with respect to H_0 , but they do not have the same distribution. In order to find the asymptotic distribution of S as $n \rightarrow \infty$, we use the following version of the central limit theorem (without proof), which applies to non-identically distributed independent random variables:

Central limit theorem – version of Lindeberg-Feller. Let $U_j, j \in \mathbb{N}$, be independent random variables with expectation $\mu_j = E[U_j]$ and variance $\sigma_j^2 = \text{Var}[U_j] < \infty$. Set

$$\Sigma_n^2 = \text{Var}[U_1 + \dots + U_n] = \sigma_1^2 + \dots + \sigma_n^2.$$

Assume that the following *Lindeberg condition* holds for all $\epsilon > 0$:

$$\frac{1}{\Sigma_n^2} \sum_{j=1}^n E[(U_j - \mu_j)^2 1_{\{|U_j - \mu_j| \geq \epsilon \Sigma_n\}}] \xrightarrow{n \rightarrow \infty} 0.$$

Then

$$\frac{1}{\Sigma_n} \sum_{j=1}^n (U_j - \mu_j) \xrightarrow{n \rightarrow \infty} \text{standard-normal},$$

where the convergence is in distribution.

(Intuitively speaking, the Lindeberg condition states that the large deviations of U_j from their expected value μ_j should only contribute a negligible fraction of the total variance Σ_n^2 .)

Let us check that the Lindeberg condition is fulfilled for the random variables $U_j = jB_j$, where $B_j, j \in \mathbb{N}$, are i.i.d. Bernoulli(1/2) random variables: We calculate:

$$\mu_j = E[jB_j] = \frac{j}{2}, \quad \text{Var}[jB_j] = \frac{j^2}{4},$$

hence

$$\Sigma_n^2 = \sum_{j=1}^n \frac{j^2}{4} = \frac{1}{24} n(n+1)(2n+1).$$

Thus Σ_n grows roughly like $n^{3/2}/\sqrt{12}$, while the largest possible value of $|jB_j - E[jB_j]|$ for $j \leq n$ is $n/2$, which grows slower than $n^{3/2}/\sqrt{12}$ for large n . Thus for all $\epsilon > 0$ there is $n_0 > 0$ such that for all $n \geq n_0$ and all $j = 1 \dots n$ one has surely $|jB_j - E[jB_j]| < \epsilon \Sigma_n$. Thus the Lindeberg condition is certainly satisfied. We conclude, using $\sum_{j=1}^n \mu_j = n(n+1)/4$:

The null distribution of the rank sum test statistic S for coupled samples has the following asymptotic behaviour for large sample sizes:

$$\text{with respect to } H_0: \quad \frac{S - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \xrightarrow{n \rightarrow \infty} \text{normal}(0, 1)$$

(convergence in distribution).

Roughly speaking, this means that the null distribution of S is well approximated by a normal($\mu = n(n+1)/4, \sigma^2 = n(n+1)(2n+1)/24$) distribution for large n . In particular, this approximation allows us to calculate approximately the rejection domain for the rank sum test for coupled samples for given significance levels.

Let us summarize:

One-sample rank sum test.

Frame model $\mathcal{P}$: $Z_1, \dots, Z_n$ are i.i.d. random variables with a continuous distribution.

Null hypothesis H_0 : “ $Z_i \sim -Z_i$ for all i .”

Test statistic:

$$S = \sum_{j=1}^n j 1_{\{Z_{\sigma(j)} > 0\}},$$

where the permutation σ rearranges the data in increasing order of their absolute value.

Rejection domain: For two-sided tests:

$$R = \{|S - n(n+1)/4| > c\}.$$

There are similar one-sided versions.

Asymptotic significance level: For large n , the significance level is approximately

$$\alpha \approx 2\Phi\left(\frac{-c}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. As $n \rightarrow \infty$, this formula gets asymptotically exact. (More precisely, the differences of both sides converges to 0, uniformly in c .)

2.14 Comparing two independent samples

Two groups of patients (maybe of different sample sizes n and m) are treated with two different substances A and B. One day later, their systolic blood pressure is measured. One gets data $X_1, \dots, X_n$ for group A and $Y_1, \dots, Y_m$ for group B. We would like to test: Are the two substances equally active, at least concerning the systolic blood pressure? Let us consider a parametric and a non-parametric model for the joint distribution of $(X_1, \dots, X_n, Y_1, \dots, Y_m)$:

	parametric (normal) model	nonparametric model
Frame model $\mathcal{P}$	$\text{normal}(\mu_X, \sigma^2)^n \otimes \text{normal}(\mu_Y, \sigma^2)^m$ (unknown parameters)	$P^n \otimes Q^m$ (unknown continuous distr. P, Q)
Null hypothesis H_0	$\mu_X = \mu_Y$	$P = Q$

The tests to be developed in this subsection are *not sensitive to differences in the variance*; they are only sensitive to differences in the expectation. This is why we even assume the variances of X_i and Y_j to be equal in the parametric frame model.

2.14.1 Comparing two independent samples (normal frame model)

Let us examine first the parametric model: From Lemma 2.32, we know

$$\left(\bar{X}, \bar{Y}, \frac{(n-1)s_X^2}{\sigma^2}, \frac{(m-1)s_Y^2}{\sigma^2}\right) \sim \text{normal}\left(\mu_X, \frac{\sigma^2}{n}\right) \otimes \text{normal}\left(\mu_Y, \frac{\sigma^2}{m}\right) \otimes \chi_{n-1}^2 \otimes \chi_{m-1}^2.$$

As a consequence,

$$\left(\bar{X} - \bar{Y}, \frac{1}{\sigma^2}[(n-1)s_X^2 + (m-1)s_Y^2] \right) \sim \text{normal}(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}) \otimes \chi_{(n-1)+(m-1)}^2.$$

Recall the following convolution properties:

$$\begin{aligned} \text{normal}(\mu_1, \sigma_1^2) * \text{normal}(\mu_2, \sigma_2^2) &= \text{normal}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \\ \chi_a^2 * \chi_b^2 &= \chi_{a+b}^2. \end{aligned}$$

We get

$$(Z, S^2) := \left(\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}, \frac{1}{\sigma^2}[(n-1)s_X^2 + (m-1)s_Y^2] \right) \sim \text{standard-normal} \otimes \chi_{n+m-2}^2.$$

and thus

$$\sqrt{\frac{n+m-2}{\frac{1}{n} + \frac{1}{m}}} \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}} = \frac{Z}{\sqrt{\frac{1}{n+m-2} S^2}} \sim t_{n+m-2}$$

by the definition of the Student t distribution. Let us summarize:

Two-sample t test for independent samples.

Frame model:

$$(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim \text{normal}(\mu_X, \sigma^2)^n \otimes \text{normal}(\mu_Y, \sigma^2)^m$$

with unknown parameters μ_X, μ_Y, σ^2 .

Null hypothesis: “ $\mu_X = \mu_Y$ ”.

Test statistic:

$$T = \sqrt{\frac{n+m-2}{\frac{1}{n} + \frac{1}{m}}} \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}}.$$

With respect to the null hypothesis, $T \sim t_{n+m-2}$.

Rejection region (two-sided version): For significance level α ,

$$R_\alpha = \{|T| > q_{n+m-2, 1-\alpha/2}\},$$

where $q_{n+m-2, 1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the Student t distribution with $n + m - 2$ degrees of freedom. There are similar one-sided versions.

2.14.2 Comparing two independent samples (nonparametric model)

With respect to our *nonparametric* null hypothesis, we have $(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim P^{n+m}$ with an unknown continuous distribution P ; i.e. we have i.i.d. random variables. What is a reasonable alternative hypothesis? We define a partial order between probability distributions on $\mathbb{R}$:

Definition 2.39 We say that P is stochastically smaller than Q if there are random variables $X \sim P$ and $Y \sim Q$ with $X \leq Y$. If additionally $X < Y$ holds with positive probability, we say P is strictly stochastically smaller than Q .

Here is a test, based on counting how often an X_i is bigger than a Y_j .

Wilcoxon two-sample rank test; synonym: Mann Whitney U test.

Frame model: $(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim P^n \otimes Q^m$ with unknown continuous distributions P and Q .

Null hypothesis H_0 : “ $P = Q$ ”.

Alternative hypothesis H_A : “ P is stochastically strictly smaller than Q .”

Test statistic:

$$U = \sum_{i=1}^n \sum_{j=1}^m 1_{\{X_i > Y_j\}}.$$

Rejection region:

$$R = \{U < c\}$$

for some critical value c .

Null distribution of the test statistic: The null distribution of U does not depend on P . With respect to all $P_0 \in H_0$,

$$E_0[U] = \frac{mn}{2}$$

and

$$\text{Var}_0[U] = \frac{1}{12}mn(m+n+1).$$

As $m, n \rightarrow \infty$, the normalized test statistic $(U - E_0[U])/\sqrt{\text{Var}_0[U]}$ converges in distribution to a standard-normal distribution. As a consequence, choosing the critical value

$$c = E_0[U] + \Phi^{-1}(\alpha)\sqrt{\text{Var}_0[U]},$$

one gets approximately the significance level α . The approximation gets asymptotically exact in the limit $m, n \rightarrow \infty$.

The U test statistic can also be calculated with the help of a rank sum. For details, see exercise 2 on homework sheet 11.

Proof of the statements on the null distribution of U : Let F denote the cumulative distribution function of P . Then, with respect to H_0 , $F[X_i] \sim F[Y_j] = \text{uniform}[0, 1]$. Furthermore, $\{X_i > Y_j\} = \{F[X_i] > F[Y_j]\}$ (at least up to a set of probability 0) for all i and j , by monotonicity of F ; recall that P is a continuous distribution. Thus the null distribution of U does not change when we replace the X_i, Y_j by i.i.d. $\text{uniform}[0, 1]$ random variables; hence the null distribution of U does not depend on P . We calculate for $P_0 \in H_0$:

$$E_0[U] = \sum_{i=1}^n \sum_{j=1}^m P_0[X_i > Y_j] = \sum_{i=1}^n \sum_{j=1}^m \frac{1}{2} = \frac{mn}{2}$$

and

$$\begin{aligned}\text{Var}_0[U] &= \sum_{i=1}^n \sum_{j=1}^m \sum_{i'=1}^n \sum_{j'=1}^m \text{Cov}_0(1_{\{X_i > Y_j\}}, 1_{\{X_{i'} > Y_{j'}\}}) \\ &= \sum_{i=1}^n \sum_{j=1}^m \sum_{i'=1}^n \sum_{j'=1}^m (P_0[X_i > Y_j, X_{i'} > Y_{j'}] - P_0[X_i > Y_j]P_0[X_{i'} > Y_{j'}])\end{aligned}$$

(Recall $\text{Cov}(U, V) = E[UV] - E[U]E[V]$ and $E[1_A] = P[A]$.)

We calculate: $P_0[X_i > Y_j] = P_0[X_{i'} > Y_{j'}] = 1/2$ and

$$P_0[X_i > Y_j, X_{i'} > Y_{j'}] = \begin{cases} \frac{1}{2} & \text{for } i = i', j = j' \\ \frac{1}{3} & \text{for } i = i', j \neq j' \text{ or } i \neq i', j = j' \\ \frac{1}{4} & \text{for } i \neq i', j \neq j' \end{cases}$$

(Note that $\{X_i > Y_j\}, \{X_{i'} > Y_{j'}\}$ are independent for $i \neq i', j \neq j'$, and that the probability for X_i to be the smallest one among $X_i, Y_j, Y_{j'}$ ($j \neq j'$) equals $1/3$ by symmetry.)

Consequently,

$$\text{Cov}_0(1_{\{X_i > Y_j\}}, 1_{\{X_{i'} > Y_{j'}\}}) = \begin{cases} \frac{1}{4} & \text{for } i = i', j = j' & (mn \text{ summands}) \\ \frac{1}{12} & \text{for } i = i', j \neq j' \text{ or } i \neq i', j = j' & (mn(n-1) + m(m-1)n \text{ summands}) \\ 0 & \text{for } i \neq i', j \neq j' & (m(m-1)n(n-1) \text{ summands}) \end{cases} \quad (82)$$

We get:

$$\text{Var}_0[U] = mn \frac{1}{4} + (mn(n-1) + m(m-1)n) \frac{1}{12} = \frac{1}{12} mn(m+n+1).$$

Note that the nondiagonal covariance terms yield the main part of the total variance.

The rest of this proof is optional material.

Unfortunately, due to lack of independence of the summands, we cannot apply the central limit theorem directly to prove the statement about asymptotic normality.

Since the null distribution of U does not depend on the distribution P of the X_i, Y_j , we may choose a particular P , which is most convenient for us. Let us again choose

$$X_i \sim Y_j \sim P = \text{uniform}[0, 1].$$

The key idea is to approximate $1_{\{X_i - Y_j\}} - 1/2$ by the differences $X_i - Y_j$. Indeed, we get the cross covariances

$$\text{Cov}_0(1_{\{X_i - Y_j\}}, X_{i'} - Y_{j'}) = \begin{cases} \frac{1}{6} & \text{for } i = i', j = j' \\ \frac{1}{12} & \text{for } i = i', j \neq j' \text{ or } i \neq i', j = j' \\ 0 & \text{for } i \neq i', j \neq j' \end{cases},$$

which coincides with the covariances in (82) *outside* the diagonal (the main contribution!), but not on the diagonal.

(To check this formula, note that

$$\text{Cov}_0(1_{\{X_i > Y_j\}}, X_i) = \text{Cov}_0(1_{\{X_i > Y_j\}}, -Y_j) = \int_{[0,1]^2} 1_{\{x > y\}} x \, dx \, dy - \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{12},$$

and

$$\text{Cov}_0(1_{\{X_i > Y_j\}}, X_{i'}) = 0 = \text{Cov}_0(1_{\{X_i > Y_j\}}, Y_{j'})$$

for $i \neq i'$ and $j \neq j'$ by independence.)

Furthermore, $\text{Var}_0[X_i] = \text{Var}_0[Y_j] = 1/12$ together with the independence structure implies

$$\text{Cov}_0(X_i - Y_j, X_{i'} - Y_{j'}) = \begin{cases} \frac{1}{6} & \text{for } i = i', j = j' \\ \frac{1}{12} & \text{for } i = i', j \neq j' \text{ or } i \neq i', j = j' \\ 0 & \text{for } i \neq i', j \neq j' \end{cases},$$

i.e. we get again the same covariance structure. Taking differences, all the nondiagonal terms cancel, and only some diagonal (lower order) terms remain:

$$\text{Cov}_0(1_{\{X_i > Y_j\}} - (X_i - Y_j), 1_{\{X_{i'} > Y_{j'}\}} - (X_{i'} - Y_{j'})) = \begin{cases} \frac{1}{12} & \text{for } i = i', j = j' \\ 0 & \text{for } i \neq i' \text{ or } j \neq j' \end{cases} \quad (83)$$

(We used $\text{Cov}(A - B, A' - B') = \text{Cov}(A, A') - \text{Cov}(B, A') - \text{Cov}(A, B') + \text{Cov}(B, B')$.)

We set

$$V = \sum_{i=1}^n \sum_{j=1}^m (X_i - Y_j) = m \sum_{i=1}^n X_i - n \sum_{j=1}^m Y_j \quad (84)$$

in analogy to

$$U = \sum_{i=1}^n \sum_{j=1}^m 1_{\{X_i > Y_j\}}.$$

As a consequence of (83), no nondiagonal terms remain in the following calculation:

$$\text{Var}_0[U - V] = \sum_{i=1}^n \sum_{j=1}^m \text{Var}_0[1_{\{X_i > Y_j\}} - (X_i - Y_j)] = \frac{nm}{12}.$$

Using $\text{Var}_0[U] = mn(m + n + 1)/12$ and Chebyshev's inequality, we get for all $\epsilon > 0$:

$$\epsilon^2 P_0 \left[\left| \frac{U - V - E_0[U - V]}{\sqrt{\text{Var}_0[U]}} \right| \geq \epsilon \right] \leq \frac{\text{Var}_0[U - V]}{\text{Var}_0[U]} \rightarrow 0$$

as $m, n \rightarrow \infty$, i.e.

$$\frac{U - V - E_0[U - V]}{\sqrt{\text{Var}_0[U]}} \rightarrow 0 \quad (85)$$

in probability.

(Recall Chebyshev's inequality: $\epsilon^2 P[|X - E[X]| \geq \epsilon] \leq \text{Var}[X]$.)

Furthermore,

$$E_0[V] = 0$$

and, by independence,

$$\text{Var}_0[V] = m^2 \sum_{i=1}^n \text{Var}_0[X_i] + n^2 \sum_{j=1}^m \text{Var}_0[Y_j] = \frac{1}{12}(m^2 n + n^2 m).$$

By the Lindeberg-Feller version of the central limit theorem,

$$\frac{V}{\sqrt{\text{Var}_0[V]}} \rightarrow \text{standard-normal}$$

in distribution as $m, n \rightarrow \infty$.

(Indeed, we use the right hand side in (84): For all $\epsilon > 0$,

$$|nX_i|, |mY_j| \leq n + m \leq \frac{\epsilon}{12}(m^2 n + n^2 m) = \epsilon \text{Var}_0[V]$$

holds for all large n, m ; thus the Lindeberg condition is fulfilled.)

Since

$$\frac{\text{Var}_0[U]}{\text{Var}_0[V]} = \frac{\frac{1}{12}mn(m + n + 1)}{\frac{1}{12}(m^2 n + n^2 m)} \rightarrow 1$$

as $m, n \rightarrow \infty$, we also get

$$\frac{V}{\sqrt{\text{Var}_0[U]}} \rightarrow \text{standard-normal}$$

in distribution as $m, n \rightarrow \infty$; thus, using $E_0[V] = 0$ and (85), we have proven the claim

$$\frac{U - E_0[U]}{\sqrt{\text{Var}_0[U]}} = \frac{U - V - E_0[U - V]}{\sqrt{\text{Var}_0[U]}} + \frac{V}{\sqrt{\text{Var}_0[U]}} \rightarrow \text{standard-normal}$$

in distribution as $m, n \rightarrow \infty$. □

Let us summarize the tests treated in the last two subsections, which are all sensitive to differences in the expectation of two samples:

Comparing two samples	normal frame model	nonparametric frame model
coupled samples	1-sample t test on $X_i - Y_i$	sign test, 1-sample rank test on $X_i - Y_i$
independent samples	2-sample t test	Wilcoxon 2-sample rank test

2.15 Comparing variances in a parametric model

Let $X_1, \dots, X_n$ be an i.i.d. $\text{normal}(\mu_X, \sigma_X^2)$ sample, and, independently, $Y_1, \dots, Y_m$ be an i.i.d. $\text{normal}(\mu_Y, \sigma_Y^2)$ sample. In this section, we are not interested in tests on μ_X, μ_Y ; rather we develop a test for comparing the variances.

Fisher's f test.

Frame model:

$$(X_1, \dots, X_n, Y_1, \dots, Y_m) \sim \text{normal}(\mu_X, \sigma_X^2)^n \otimes \text{normal}(\mu_Y, \sigma_Y^2)^m$$

with unknown parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$.

Null hypothesis H_0 : “ $\sigma_X = \sigma_Y$ ”

Test statistic:

$$F = \frac{s_X^2}{s_Y^2}.$$

Rejection regions (one-sided version): $R_+ = \{F > c_+\}$ or $R_- = \{F < c_-\}$, depending on the alternative hypothesis H_A^+ : “ $\sigma_X > \sigma_Y$ ” or H_A^- : “ $\sigma_X < \sigma_Y$ ”.

(Recall that the empirical variance s_X^2 was defined by $s_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$.)

Let us determine the null distribution of the test statistic: By Lemma 2.32, we know

$$\left(\frac{(n-1)s_X^2}{\sigma_X^2}, \frac{(m-1)s_Y^2}{\sigma_Y^2} \right) \sim \chi_{n-1}^2 \otimes \chi_{m-1}^2.$$

Thus, with respect to H_0 , the Fisher's f test statistic is distributed according to the quotient of $S_1/(n-1)$ and $S_2/(m-1)$, where $(S_1, S_2) \sim \chi_{n-1}^2 \otimes \chi_{m-1}^2$.

Definition 2.40 (Fisher's f distributions) *The Fisher f distribution with degrees of freedom n, m is defined to be the distribution of $F = \frac{S_1/n}{S_2/m}$, where $(S_1, S_2) \sim \chi_n^2 \otimes \chi_m^2$. It is denoted by $f_{n,m}$.*

In particular:

The null distribution of the Fisher f test statistic is $f_{n-1, m-1}$.

The Fisher f distributions are closely related to Beta distributions. Recall their definition:

Definition 2.41 (Beta distributions) Let $a, b > 0$. The $\text{Beta}(a, b)$ distribution is defined to have the following probability density on the unit interval:

$$\beta_{a,b}(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1.$$

Here

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

denotes the Beta function.

Lemma 2.42 If

$$(S_1, S_2) \sim \text{Gamma}(r, \lambda) \otimes \text{Gamma}(s, \lambda),$$

then

$$(S_1 + S_2, S_1/(S_1 + S_2)) \sim \text{Gamma}(r + s, \lambda) \otimes \text{Beta}(r, s).$$

Proof: Homework sheet 10, exercise 2. □

We apply this to the case

$$(S_1, S_2) \sim \text{Gamma}(n/2, 1/2) \otimes \text{Gamma}(m/2, 1/2) = \chi_n^2 \otimes \chi_m^2.$$

Then we know on the one hand

$$B := \frac{S_1}{S_1 + S_2} \sim \text{Beta}(n/2, m/2),$$

and on the other hand

$$F = \frac{S_1/n}{S_2/m} \sim f_{n,m}.$$

Using

$$B = \frac{nF}{nF + m}, \quad F = \frac{m}{n} \frac{B}{1 - B}$$

and the fact

$$\frac{d}{dx} \frac{nx}{nx + m} = \frac{nm}{(nx + m)^2},$$

we get the following relation between the density f_B of B and f_F of F :

$$f_F(x) = f_B \left(\frac{nx}{nx + m} \right) \frac{nm}{(nx + m)^2}, \quad (x > 0)$$

and thus

$$\begin{aligned} f_F(x) &= \frac{1}{B(n/2, m/2)} \left(\frac{nx}{nx + m} \right)^{n/2-1} \left(1 - \frac{nx}{nx + m} \right)^{m/2-1} \frac{nm}{(nx + m)^2} \\ &= \frac{n^{n/2} m^{m/2}}{B(n/2, m/2)} \frac{x^{n/2-1}}{(nx + m)^{n/2+m/2}}. \end{aligned}$$

Let us summarize:

Density of Fisher's f distribution. The Fisher f distribution $f_{n,m}$ has the density

$$f_F(x) = \frac{n^{n/2} m^{m/2}}{B(n/2, m/2)} \frac{x^{n/2-1}}{(nx+m)^{n/2+m/2}}, \quad x > 0.$$

The Fisher f test is not robust: Its significance level may change drastically as soon as one deviates a little bit from the normality assumption, especially in the tails of the distributions. Thus one should not take its result too serious unless one has good (theoretical) evidence that the data really are normally distributed. In practise, this is rarely the case. In these cases, one should prefer a non-parametric alternative, like a rank test on the deviations from the median.

2.16 Asymptotic power of likelihood ratio tests

Consider n i.i.d. random variables $X_1, \dots, X_n$ with an unknown distribution. Given two distributions P_0 and P_A (absolutely continuous with respect to each other), let us consider the hypotheses H_0 : “ $X_j \sim P_0$ ” versus H_A : “ $X_j \sim P_A$ ”. How much do optimal tests distinguish between P_0 and P_A ? Given a significance level α , how fast does the power converge to 1 as $n \rightarrow \infty$?

Formally speaking, we use a result space Ω^n , $H_0 = \{P_0^n\}$, $H_A = \{P_A^n\}$. E_0^n and E_A^n denote the expectations with respect to P_0^n and P_A^n , respectively; in the case $n = 1$ we drop the superscript n . Here is a tool to measure how much P_0 and P_A can be distinguished:

Definition 2.43 The *relative entropy* (synonym: *Kullback-Leibler distance*) of P_0 with respect to P_A is defined by

$$H(P_0, P_A) = E_0[\log f] = E_A[f \log f]$$

with the abbreviation

$$f = \frac{dP_0}{dP_A}.$$

The function $x \mapsto x \log x$ is convex with value 0 and derivative 1 at $x = 1$. Thus $f \log f \geq f - 1$ (with equality only for $f = 1$). Consequently,

$$H(P_0, P_A) \geq E_A[f - 1] = E_0[1] - E_A[1] = 0,$$

with equality only for $P_A = P_0$.

(Recall the following basic property of likelihood ratios:

$$E_A \left[\frac{dP_0}{dP_A} X \right] = E_0[X] \tag{86}$$

holds for all random variables X for which at least one of the two sides is well defined.)

Indeed, the relative entropy measures the asymptotic power of tests. More precisely:

Lemma 2.44 (Stein's Lemma) Consider a sequence of likelihood ratio tests with rejection domains

$$R_n = \left\{ \frac{dP_0^n}{dP_A^n} < c_n \right\} \subseteq \Omega^n$$

and significance levels

$$\alpha_n = P_0^n[R_n].$$

Assume that the α_n are bounded away from 0 and 1, i.e. there is $\epsilon > 0$ such that for all $n \in \mathbb{N}$ holds $\epsilon \leq \alpha_n \leq 1 - \epsilon$. Furthermore, assume that the relative entropy $H(P_0, P_A)$ is finite. Then the power of the tests converges exponentially fast to 1 as $n \rightarrow \infty$. More precisely,

$$P_A^n[R_n] = 1 - e^{-nH(P_0, P_A) + o(n)}, \quad (87)$$

where $o(n)/n \rightarrow 0$ as $n \rightarrow \infty$. Furthermore, the critical values c_n have the following asymptotic behaviour:

$$\frac{1}{n} \log c_n \xrightarrow{n \rightarrow \infty} H(P_0, P_A) \quad (88)$$

(Recall that " $a_n \xrightarrow{n \rightarrow \infty} a$ " means: For all $\epsilon > 0$, there is $N \in \mathbb{N}$ such that for all $n \geq N$ one has $|a_n - a| < \epsilon$.)

Proof. We estimate:

$$1 - P_A^n[R_n] = P_A^n[\Omega^n \setminus R_n] = P_A^n \left[\frac{1}{c_n} \frac{dP_0^n}{dP_A^n} \geq 1 \right] \leq E_A^n \left[\frac{1}{c_n} \frac{dP_0^n}{dP_A^n} \right] = E_0^n \left[\frac{1}{c_n} \right] = \frac{1}{c_n}. \quad (89)$$

(To understand this estimate, recall (86), and recall the general Chebyshev inequality: For a random variable $X \geq 0$, one has $1_{\{X \geq 1\}} \leq X$; thus $P[X \geq 1] = E[1_{\{X \geq 1\}}] \leq E[X]$ holds.)

We observe

$$\frac{dP_0^n}{dP_A^n} = \prod_{j=1}^n f(X_j), \quad \log \frac{dP_0^n}{dP_A^n} = \sum_{j=1}^n \log f(X_j). \quad (90)$$

Recall that $X_j : \Omega^n \rightarrow \Omega$ denotes the j -th observation. Taking logarithms, we rewrite the rejection regions as

$$R_n = \left\{ \frac{1}{n} \sum_{j=1}^n \log f(X_j) < \frac{1}{n} \log c_n \right\}, \quad (91)$$

thus

$$P_0^n \left[\frac{1}{n} \sum_{j=1}^n \log f(X_j) < \frac{1}{n} \log c_n \right] = P_0^n[R_n] \in [\epsilon, 1 - \epsilon]. \quad (92)$$

By the weak law of large numbers, we know for all $\delta > 0$:

$$P_0^n \left[\left| \frac{1}{n} \sum_{j=1}^n \log f(X_j) - E_0[\log f] \right| < \delta \right] \xrightarrow{n \rightarrow \infty} 1. \quad (93)$$

(Recall the *weak law of large numbers*: For i.i.d. random variables $Z_1, Z_2, Z_3, \dots$ with finite expected value $\mu = E[Z_j]$, the following holds for all $\delta > 0$:

$$P \left[\left| \frac{1}{n} \sum_{j=1}^n Z_j - \mu \right| < \delta \right] \xrightarrow{n \rightarrow \infty} 1.$$

Note that we use here a version of the weak law of large numbers which does not assume that the variance of $\log f$ with respect to P_0 exists.)

As a consequence, using $E_0[\log f] = H(P_0, P_A)$:

$$P_0^n \left[\frac{1}{n} \sum_{j=1}^n \log f(X_j) \leq H(P_0, P_A) + \delta \right] \xrightarrow{n \rightarrow \infty} 1 \quad (94)$$

and

$$P_0^n \left[\frac{1}{n} \sum_{j=1}^n \log f(X_j) \leq H(P_0, P_A) - \delta \right] \xrightarrow{n \rightarrow \infty} 0. \quad (95)$$

Comparing (92) with (94) and (95), we conclude that

$$H(P_0, P_A) - \delta \leq \frac{1}{n} \log c_n \leq H(P_0, P_A) + \delta$$

holds for all n large enough. This holds for all $\delta > 0$; thus we have proven (88). Combining this with (89), we obtain one direction of (87):

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log(1 - P_A^n[R_n]) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{c_n} = -H(P_0, P_A).$$

(Recall that $\limsup_{n \rightarrow \infty} a_n$ is the infimum of all $c \in \mathbb{R}$ for which $a_n > c$ holds for at most finitely many $n \in \mathbb{N}$. Similarly, $\liminf_{n \rightarrow \infty} a_n$ is the supremum of all $c \in \mathbb{R}$ for which $a_n < c$ holds for at most finitely many $n \in \mathbb{N}$. Furthermore, if $\limsup_{n \rightarrow \infty} a_n = \liminf_{n \rightarrow \infty} a_n$, then $a_n \xrightarrow{n \rightarrow \infty} a$.)

We derive the other direction next. Let $\delta > 0$. We use the notation $E[X, B] = E[X1_B]$ for random variables X and events B in the following estimate; recall (86) for the third step:

$$\begin{aligned} 1 - P_A^n[R_n] &= P_A^n \left[\frac{dP_0^n}{dP_A^n} \geq c_n \right] \\ &\geq P_A^n \left[e^{-nH(P_0, P_A) - n\delta} \leq \frac{dP_A^n}{dP_0^n} \leq \frac{1}{c_n} \right] \\ &= E_0^n \left[\frac{dP_A^n}{dP_0^n}, e^{-nH(P_0, P_A) - n\delta} \leq \frac{dP_A^n}{dP_0^n} \leq \frac{1}{c_n} \right] \\ &\geq e^{-nH(P_0, P_A) - n\delta} P_0^n \left[e^{-nH(P_0, P_A) - n\delta} \leq \frac{dP_A^n}{dP_0^n} \leq \frac{1}{c_n} \right] \\ &\geq e^{-nH(P_0, P_A) - n\delta} \left(P_0^n \left[e^{-nH(P_0, P_A) - n\delta} \leq \frac{dP_A^n}{dP_0^n} \right] - P_0^n \left[\frac{dP_A^n}{dP_0^n} > \frac{1}{c_n} \right] \right). \end{aligned} \quad (96)$$

(For the last estimate, recall that for all events A and B one has $P[A \cap B] \geq P[A] - P[B^c]$, where $B^c = \Omega \setminus B$ denotes the complement of B .)

We estimate the terms in the last expression:

$$P_0^n \left[\frac{dP_A^n}{dP_0^n} > \frac{1}{c_n} \right] = P_0^n[R_n] \leq 1 - \epsilon. \quad (97)$$

Furthermore, by (90) and (94),

$$P_0^n \left[e^{-nH(P_0, P_A) - n\delta} \leq \frac{dP_A^n}{dP_0^n} \right] = P_0^n \left[\frac{1}{n} \sum_{j=1}^n \log f(X_j) \leq H(P_0, P_A) + \delta \right] \geq 1 - \frac{\epsilon}{2} \quad (98)$$

holds for all sufficiently large n . Combining (96) with (97) and (98), we see that

$$1 - P_A^n[R_n] \geq e^{-nH(P_0, P_A) - n\delta} \left(\left(1 - \frac{\epsilon}{2}\right) - (1 - \epsilon) \right) = e^{-nH(P_0, P_A) - n\delta + \log(\epsilon/2)}$$

holds for all large n . Since this holds for all $\delta > 0$, we obtain the other direction of (87):

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log(1 - P_A^n[R_n]) \geq -H(P_0, P_A).$$

(Note that “ $\lim_{n \rightarrow \infty} \frac{1}{n} \log(1 - P_A^n[R_n]) = -H(P_0, P_A)$ ” is just a different notation for “ $P_A^n[R_n] = 1 - e^{-nH(P_0, P_A) + o(n)}$, where $o(n)/n \rightarrow 0$ as $n \rightarrow \infty$.”)

This finishes the proof of Stein’s lemma. □

Example 2.45 Consider again a possibly biased coin, being tossed n times. Consider the frame model $\Omega^n = \{0, 1\}^n$, $\mathcal{P} = \{P_p^n : 0 < p < 1\}$, where $P_p[\omega = 1] = 1 - P_p[\omega = 0] = p$. Consider most powerful tests (rejection region R_n) for testing H_0 : “ $p = p_0$ ” versus H_A : “ $p = p_A$ ” at a significance level $\alpha_n \in [1\%, 10\%]$. What is the asymptotic behavior of the power $1 - \beta_n = P_{p_A}^n[R_n]$ as $n \rightarrow \infty$?

We calculate the relative entropy of P_{p_0} with respect to P_{p_A} :

$$H(P_{p_0}, P_{p_A}) = E_{p_0} \left[\log \frac{dP_{p_0}}{dP_{p_A}} \right] = (1 - p_0) \log \frac{1 - p_0}{1 - p_A} + p_0 \log \frac{p_0}{p_A}.$$

By Stein’s Lemma, we get

$$\beta_n = e^{-nH(P_{p_0}, P_{p_A}) + o(n)} = e^{o(n)} \left(\frac{1 - p_A}{1 - p_0} \right)^{n(1 - p_0)} \left(\frac{p_A}{p_0} \right)^{np_0},$$

where $o(n)/n \rightarrow 0$ as $n \rightarrow \infty$. In particular, the type II risk β_n converges exponentially fast to 0.

Here is a numerical example to illustrate the fast speed of convergence: For $p_0 = 50\%$, $p_A = 20\%$ we get $H(P_{50\%}, P_{20\%}) = 0.223144$ (rounded); thus one has the rounded values

$$\frac{n}{\exp(-nH(P_{50\%}, P_{20\%}))} \parallel \begin{array}{c|c|c|c|c} 10 & 100 & 200 & 1000 \\ \hline 0.107 & 2.037 \cdot 10^{-10} & 4.15 \cdot 10^{-20} & 1.23 \cdot 10^{-97} \end{array}$$

This shows: already with moderate sample sizes n , one can get extremely small numbers for $e^{-nH(P_0, P_A)}$.

3 Estimating parameters

3.1 Basic definitions

In the previous chapter, we encountered already several times the problem to *estimate* unknown parameters:

- In the χ^2 test for composite hypotheses, we replaced unknown parameters by their maximum likelihood estimation.
- In the derivation of the t test, we replaced the unknown true standard deviation σ by an estimation, the empirical standard deviation s .
- Confidence intervals also have some flavor of an estimation problem: Instead of giving only *one* estimated value, we give a whole region of “plausible” values for an unknown parameter.

Let us first formally define the notion of estimators:

Definition 3.1 (Estimators) Consider a frame model $(\Omega, \mathcal{P})$ and a parameter $\theta : \mathcal{P} \rightarrow \mathbb{R}^n$. An estimator $\hat{\theta}$ for θ is a random variable $\hat{\theta} : \Omega \rightarrow \mathbb{R}^n$.

Given data $\omega \in \Omega$, the estimated value for the unknown true parameter $\theta(P)$ is $\hat{\theta}(\omega)$; it can be determined from the known data ω . Here $P \in \mathcal{P}$ denotes the “true”, but unknown underlying probability distribution.

(In the definition of estimators, we did not impose any condition asking that the estimator should be a “good” one. For example, 0 is an estimator for every parameter, but in many cases, it is a rather bad one.)

Let us discuss a few criteria to evaluate the quality of an estimator:

Definition 3.2 (Unbiased estimators) An estimator $\hat{\theta}$ for a parameter θ is called *unbiased*, if for all $P \in \mathcal{P}$ one has

$$E_P[\hat{\theta}] = \theta(P).$$

Here E_P means the expectation with respect to $P \in \mathcal{P}$.

Thus an estimator is unbiased, if its expectation $E_P[\hat{\theta}]$ is the true parameter $\theta(P)$, whatever the “true” $P \in \mathcal{P}$ is.

Example 3.3 Let $X_1, \dots, X_n$ be i.i.d. random variables with an unknown distribution P and an unknown finite expectation $\mu(P) = E_P[X_i]$.

(Formally speaking, we use the nonparametric model

$$\mathcal{P} = \{P^n : P \text{ is a probability distribution on } \mathbb{R} \text{ with finite expectation } \mu(P)\}$$

on $\Omega = \mathbb{R}^n$, and $X_i(\omega)$ is the i -th component of $\omega \in \Omega$.)

The sample mean

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

is an unbiased estimator for the unknown expectation $\mu(P)$, since

$$E_P[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E_P[X_i] = \mu(P).$$

Every X_i is an unbiased estimator for $\mu(P)$, too.

Example 3.4 In addition to the previous example, let us assume that the $X_1, \dots, X_n \sim P$ have a finite variance $\sigma^2(P)$. Then the empirical variance

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

is an unbiased estimator for the unknown variance $\sigma^2(P)$. We have proven this in the last chapter if P is restricted to be a normal distribution. Here is a general proof:

We observe the following orthogonality relation between $(X_i - \bar{X})$ and $(\bar{X} - \mu(P))$:

$$\sum_{i=1}^n E_{P^n}[(X_i - \bar{X})(\bar{X} - \mu(P))] = E_{P^n}[(n\bar{X} - n\bar{X})(\bar{X} - \mu(P))] = 0.$$

Using $X_i - \mu(P) = (X_i - \bar{X}) + (\bar{X} - \mu(P))$, we get the following Pythagoras-like conclusion:

$$\begin{aligned}
n\sigma^2(P) &= \sum_{i=1}^n E_{P^n}[(X_i - \mu(P))^2] \\
&= \sum_{i=1}^n E_{P^n}[(X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu(P)) + (\bar{X} - \mu(P))^2] \\
&= \sum_{i=1}^n E_{P^n}[(X_i - \bar{X})^2] + 2 \sum_{i=1}^n E_{P^n}[(X_i - \bar{X})(\bar{X} - \mu(P))] + \sum_{i=1}^n E_{P^n}[(\bar{X} - \mu(P))^2] \\
&= (n-1)E_{P^n}[s_X^2] + 0 + n \operatorname{Var}_{P^n}[\bar{X}] \\
&= (n-1)E_{P^n}[s_X^2] + \sigma^2(P),
\end{aligned}$$

where we have used

$$\operatorname{Var}_{P^n}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^n \operatorname{Var}_{P^n}[X_i] = \frac{\sigma^2(P)}{n},$$

which follows by the independence of the X_i . This implies the claim for all P :

$$E_{P^n}[s_X^2] = \sigma^2(P).$$

On the other hand, the empirical standard deviation s_X is *not* an unbiased estimator for the standard deviation $\sigma(P)$.

(To see this, one may use Jensen's inequality: If f is a concave function and X is a random variable with values in the domain of f , then

$$E[f(X)] \leq f(E[X]).$$

Furthermore, if f is strictly concave and X is not constant with probability 1, then

$$E[f(X)] < f(E[X]).$$

In particular, using that the square root function is concave, we conclude

$$E_{P^n}[s_X] = E_{P^n}[\sqrt{s_X^2}] < \sqrt{E_{P^n}[s_X^2]} = \sqrt{\sigma^2(P)} = \sigma(P)$$

whenever the X_i are not constant with probability 1.)

Although it is nice to have unbiased estimators, there is no unbiased estimator in many cases, and in some cases, one even prefers to have a bias.

Here is a real world example, communicated by R. Gill: The length needed for a transatlantic sea cable was estimated. When the cable finally was positioned, it turned out to be a few hundred meter too short. The end of the cable fell into the sea, causing a big financial loss. The engineers should have better used a biased estimator for the length of the cable!

Consider an i.i.d. random sample $X_1, \dots, X_n \sim P$, i.e. we consider the joint frame model

$$\mathcal{P}^n = \{P^n : P \in \mathcal{P}\}$$

over Ω^n for some single-observation frame model $(\Omega, \mathcal{P})$.

Definition 3.5 (Consistent estimators) A sequence of estimators $\hat{\theta}_n : \Omega^n \rightarrow \mathbb{R}^m$, $n \in \mathbb{N}$, for a parameter $\theta : \mathcal{P} \rightarrow \mathbb{R}^m$ is called *consistent*, if for all $P \in \mathcal{P}$ and for all $\epsilon > 0$ one has

$$\lim_{n \rightarrow \infty} P^n \left[|\hat{\theta}_n - \theta(P)| > \epsilon \right] = 0,$$

i.e. $\hat{\theta}_n \rightarrow \theta(P)$ in probability with respect to P^n .

Speaking intuitively, the probability to have a large error $|\hat{\theta}_n - \theta(P)|$ for the estimation should get arbitrarily small as the sample size n grows to infinity.

Example 3.6 As a consequence of the weak law of large numbers, the means $\bar{X}_n = \sum_{i=1}^n X_i/n$, $n \in \mathbb{N}$, yield a consistent sequence estimators for the expectation $\mu(P) = E_{P^n}[X_i]$.

3.2 Maximum likelihood estimators and the method of least squares

We considered one method to find estimators already several times in this lecture. Consider a parametric frame model $\mathcal{P}$ with likelihood function $p : \Omega \times \Theta \rightarrow [0, \infty[$, $(\omega, \theta) \mapsto p(\omega|\theta)$. Recall:

Definition 3.7 (Maximum likelihood estimators) *The maximum likelihood estimator for the parameter $\theta \in \Theta$ for the observation $\omega \in \Omega$ is defined by*

$$\hat{\theta}_{\text{ML}}(\omega) = \arg \max_{\theta \in \Theta} p(\omega|\theta),$$

provided there is a unique maximum of $p(\omega|\cdot)$.

Example 3.8 A (possibly biased) coin is tossed n times. Using the frame model

$$\mathcal{P} = \{\text{binomial}(n, p) : 0 \leq p \leq 1\}$$

for the observed number of “heads” $\omega \in \Omega = \{0, 1, \dots, n\}$, we find the maximum likelihood estimator for the unknown parameter p : The likelihood function is

$$f(\omega, p) = \binom{n}{\omega} p^\omega (1-p)^{n-\omega}.$$

Given ω , we maximize $f(\omega, p)$: We calculate for $0 < p < 1$, $0 < \omega < n$:

$$\frac{\partial}{\partial p} \log f(\omega, p) = \frac{\partial}{\partial p} (\omega \log p + (n - \omega) \log(1 - p)) = \frac{\omega}{p} - \frac{n - \omega}{1 - p},$$

which equals 0 at the point

$$\hat{p}_{\text{ML}}(\omega) = \frac{\omega}{n}. \tag{99}$$

Note that the log-likelihood function $p \mapsto \log f(\omega, p)$ is concave; thus the zero of the derivative is indeed a maximum. (The case $\omega \in \{0, n\}$ needs a special consideration: there, the maximum is obtained at the boundary $\hat{p}_{\text{ML}} \in \{0, 1\}$. However, (99) holds in this case, too.) The result yields what one would intuitively guess: The maximum likelihood estimator of the unknown probability to see “head” is just the observed relative frequency of “heads”.

Example 3.9 The oxygen contents y_i (in mg/l) of the water in a lake at different depth x_i was measured. Here are the data:

x_i	15	20	30	40	50	60	70
y_i	6.5	5.6	5.4	6.0	4.6	1.4	0.1

Let us use the model that there is an unknown linear relation between the oxygen contents and the depth, up to unknown random errors:

Frame model:

$$y_i = ax_i + b + \epsilon_i, \quad i = 1, \dots, n$$

where $(\epsilon_1, \dots, \epsilon_n) \sim \text{normal}(0, \sigma^2)^n$ with unknown parameters a, b , and σ^2 . Let us treat the x_i as fixed numbers, defined by the experimental design, and y_i as random numbers.

Let us find the maximum likelihood estimator for a and b : The likelihood function equals

$$\begin{aligned} f(y_1, \dots, y_n | a, b, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^n e^{-\epsilon_i^2/(2\sigma^2)} \\ &= \exp \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - ax_i - b)^2 \right\} \end{aligned}$$

Given σ^2 , let us examine when this expression gets maximal:

Method of least squares for the regression line. The maximum likelihood estimator $\hat{a}, \hat{b}$ for the unknown line parameters is specified by the optimization problem

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - ax_i - b)^2 \stackrel{!}{=} \text{minimal}.$$

This means: One tries to minimize the sum of the squared error terms.

Let us interpret the problem geometrically: Consider the vectors $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{x} = (x_1, \dots, x_n)$, and $\mathbf{1} = (1, \dots, 1)$ in $\mathbb{R}^n$. Then

$$\sum_{i=1}^n (y_i - ax_i - b)^2 = \|\mathbf{y} - a\mathbf{x} - b\mathbf{1}\|^2,$$

i.e. we are searching for the point in the 2-dimensional linear subspace $\text{span}(\mathbf{x}, \mathbf{1})$ of $\mathbb{R}^n$ which is closest to $\mathbf{y}$. Taking partial derivatives with respect to a and b , we get

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial}{\partial a} \|\mathbf{y} - a\mathbf{x} - b\mathbf{1}\|^2 = 2 \langle \mathbf{x}, a\mathbf{x} + b\mathbf{1} - \mathbf{y} \rangle \\ 0 &\stackrel{!}{=} \frac{\partial}{\partial b} \|\mathbf{y} - a\mathbf{x} - b\mathbf{1}\|^2 = 2 \langle \mathbf{1}, a\mathbf{x} + b\mathbf{1} - \mathbf{y} \rangle \end{aligned}$$

Geometrically, this is obvious: the error vector $\hat{a}\mathbf{x} + \hat{b}\mathbf{1} - \mathbf{y}$ has to be orthogonal to the vectors $\mathbf{x}$ and $\mathbf{1}$ that span the linear subspace. We get the linear equation

$$\begin{pmatrix} \langle \mathbf{x}, \mathbf{x} \rangle & \langle \mathbf{x}, \mathbf{1} \rangle \\ \langle \mathbf{1}, \mathbf{x} \rangle & \langle \mathbf{1}, \mathbf{1} \rangle \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}, \mathbf{y} \rangle \\ \langle \mathbf{1}, \mathbf{y} \rangle \end{pmatrix}, \quad (100)$$

or, equivalently

$$\begin{aligned} \hat{a} \sum_{i=1}^n x_i^2 + \hat{b} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i \\ \hat{a} \sum_{i=1}^n x_i + \hat{b} n &= \sum_{i=1}^n y_i, \end{aligned}$$

which has the following solution:

Coefficients of the regression line.

$$\hat{a} = \frac{\begin{vmatrix} \langle \mathbf{x}, \mathbf{y} \rangle & \langle \mathbf{x}, \mathbf{1} \rangle \\ \langle \mathbf{1}, \mathbf{y} \rangle & \langle \mathbf{1}, \mathbf{1} \rangle \end{vmatrix}}{\begin{vmatrix} \langle \mathbf{x}, \mathbf{x} \rangle & \langle \mathbf{x}, \mathbf{1} \rangle \\ \langle \mathbf{1}, \mathbf{x} \rangle & \langle \mathbf{1}, \mathbf{1} \rangle \end{vmatrix}} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$\hat{b} = \frac{\begin{vmatrix} \langle \mathbf{x}, \mathbf{x} \rangle & \langle \mathbf{x}, \mathbf{y} \rangle \\ \langle \mathbf{1}, \mathbf{x} \rangle & \langle \mathbf{1}, \mathbf{y} \rangle \end{vmatrix}}{\begin{vmatrix} \langle \mathbf{x}, \mathbf{x} \rangle & \langle \mathbf{x}, \mathbf{1} \rangle \\ \langle \mathbf{1}, \mathbf{x} \rangle & \langle \mathbf{1}, \mathbf{1} \rangle \end{vmatrix}} = \frac{(\sum_{i=1}^n x_i^2) (\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i) (\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\overline{x^2}\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - (\bar{x})^2}$$

We abbreviated $\overline{f(x, y)} := \sum_i f(x_i, y_i)/n$ to denote the average of the $f(x_i, y_i)$. Note that the denominator is closely related to the empirical variance of the x_i , since

$$\frac{n-1}{n} s_x^2 = \overline{x^2} - (\bar{x})^2.$$

We get:

Regression line. The regression line can be written as

$$y = \hat{a}x + \hat{b},$$

or, equivalently, as

$$y - \bar{y} = \hat{a}(x - \bar{x}).$$

In particular, the regression line goes through the center of mass $(\bar{y}, \bar{x})$ of the data points.

To prove this, we just note that $\bar{y} = \hat{a}\bar{x} + \hat{b}$.

Next, consider a more general model: Consider a (possibly overdetermined) linear system of equations

$$A\mathbf{x} = \mathbf{b},$$

where

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n},$$

$\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$, $\mathbf{b} = (b_1, \dots, b_m)^\top \in \mathbb{R}^m$. We assume that $m > n$ and $\text{rank}(A) = n$, i.e. the system is indeed overdetermined; thus there is at most one solution $\mathbf{x}$. Now consider a random perturbation $\tilde{\mathbf{b}} = (\tilde{b}_1, \dots, \tilde{b}_m)^\top$ of the right hand side $\mathbf{b}$, where

$$\epsilon := \tilde{\mathbf{b}} - \mathbf{b} \sim \text{normal}(0, \sigma^2 I_m)$$

is a normally distributed error vector with expectation 0. Suppose we know the matrix A , and we can observe the perturbed (random) right hand side $\tilde{\mathbf{b}}$, while $\mathbf{b}$ and $\mathbf{x}$ are unknown; $\mathbf{x}$ is treated as a (unknown) parameter. Since the system is overdetermined, the perturbed version $A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ in general will not have a solution $\tilde{\mathbf{x}}$ anymore. Let us find the maximum likelihood estimator for the unknown $\mathbf{x}$: Abbreviating $c = -\frac{m}{2} \log(2\pi\sigma^2)$, the log-likelihood function equals

$$\log f(\tilde{\mathbf{b}}|\mathbf{x}) = c - \frac{\|\epsilon\|^2}{2\sigma^2} = c - \frac{\|A\mathbf{x} - \tilde{\mathbf{b}}\|^2}{2\sigma^2}.$$

(Recall that the density of the $\text{normal}(0, \sigma^2 I_m)$ distribution equals $g(\epsilon) = e^c \exp\{-\|\epsilon\|^2/(2\sigma^2)\}$.)

Geometrically, we phrase our optimization problem as follows:

The method of least squares. The maximum likelihood estimator $\hat{\mathbf{x}}$ for the unknown parameter $\mathbf{x}$ is characterized by

$$\|A\hat{\mathbf{x}} - \tilde{\mathbf{b}}\|^2 \stackrel{!}{=} \text{minimal},$$

i.e. it is determined by the point $A\hat{\mathbf{x}}$ in the range $\text{Im } A$ of the linear map $\mathbf{x} \mapsto A\mathbf{x}$ which is closest to the perturbed right hand side $\tilde{\mathbf{b}}$.

The error vector $A\hat{\mathbf{x}} - \tilde{\mathbf{b}}$ must be orthogonal to $\text{Im } A$. This is geometrically obvious, but let us do a calculation to check it algebraically:

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial}{\partial \hat{x}_j} \|A\hat{\mathbf{x}} - \tilde{\mathbf{b}}\|^2 \\ &= \frac{\partial}{\partial \hat{x}_j} \sum_{i=1}^m \left(\sum_{k=1}^n a_{ik} \hat{x}_k - \tilde{b}_i \right)^2 \\ &= 2 \sum_{i=1}^m a_{ij} \left(\sum_{k=1}^n a_{ik} \hat{x}_k - \tilde{b}_i \right); \end{aligned}$$

thus we have for all $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$

$$0 \stackrel{!}{=} \sum_{j=1}^n \sum_{i=1}^m a_{ij} y_j \left(\sum_{k=1}^n a_{ik} \hat{x}_k - \tilde{b}_i \right) = \langle A\mathbf{y}, A\hat{\mathbf{x}} - \tilde{\mathbf{b}} \rangle = \langle \mathbf{y}, A^\top A\hat{\mathbf{x}} - A^\top \tilde{\mathbf{b}} \rangle.$$

Hence $A\mathbf{y}$ is indeed orthogonal to the error vector $A\hat{\mathbf{x}} - \tilde{\mathbf{b}}$. Let us summarize:

Least square estimator for a perturbed overdetermined linear system. Consider the overdetermined linear system

$$A\mathbf{x} = \mathbf{b}$$

with a given matrix $A \in \mathbb{R}^{m \times n}$, $\text{rank } A = n < m$. Assume that $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are unknown parameters, but a random perturbation

$$\tilde{\mathbf{b}} \sim \text{normal}(\mathbf{b}, \sigma^2 I_m)$$

is known. Then the maximum likelihood estimator $\hat{\mathbf{x}}$ for $\mathbf{x}$ is determined by the $n \times n$ linear system

$$A^\top A\hat{\mathbf{x}} = A^\top \tilde{\mathbf{b}},$$

i.e.

$$\hat{\mathbf{x}} = (A^\top A)^{-1} A^\top \tilde{\mathbf{b}}.$$

One has

$$\Pi \tilde{\mathbf{b}} = A\hat{\mathbf{x}},$$

where $\Pi := A(A^\top A)^{-1} A^\top$ denotes the orthogonal projection to the n -dimensional linear subspace $\text{Im } A \subseteq \mathbb{R}^m$.

($\Pi = A(A^\top A)^{-1} A^\top$ is indeed the orthogonal projection to $\text{Im } A$, since for all $\mathbf{y} \in \mathbb{R}^m$ one has

$$\Pi A\mathbf{y} = A(A^\top A)^{-1} A^\top A\mathbf{y} = A\mathbf{y},$$

and since for all $\mathbf{u} \in (\text{Im } A)^\perp$ and $\mathbf{z} \in \mathbb{R}^m$ one has

$$\langle \mathbf{z}, \Pi \mathbf{u} \rangle = \langle \Pi^\top \mathbf{z}, \mathbf{u} \rangle = \langle \Pi \mathbf{z}, \mathbf{u} \rangle = \langle A(A^\top A)^{-1} A^\top \mathbf{z}, \mathbf{u} \rangle = 0,$$

because of $A(A^\top A)^{-1} A^\top \mathbf{z} \in \text{Im } A \perp \mathbf{u}$, and hence $\Pi \mathbf{u} = 0$.)

Let us now consider the residuals $A\hat{\mathbf{x}} - \tilde{\mathbf{b}}$, i.e. the error terms, caused by the failure of the maximum likelihood estimator $\hat{\mathbf{x}}$ to solve the overdetermined linear system “ $A\hat{\mathbf{x}} = \tilde{\mathbf{b}}$ ” precisely. Its squared norm $r^2 = \|A\hat{\mathbf{x}} - \tilde{\mathbf{b}}\|^2$ is called the “residual sum of squares”.

Lemma 3.10 (Joint distribution of the least square estimator and of the residuals)

For the normal frame model, the least square estimator $\hat{\mathbf{x}}$ and the residual vector $A\hat{\mathbf{x}} - \tilde{\mathbf{b}}$ are independent of each other. Furthermore,

$$(\hat{\mathbf{x}}, \frac{1}{\sigma^2} \|A\hat{\mathbf{x}} - \tilde{\mathbf{b}}\|^2) \sim \text{normal}(\mathbf{x}, \sigma^2(A^\top A)^{-1}) \otimes \chi_{m-n}^2.$$

Proof. We abbreviate

$$\epsilon := \tilde{\mathbf{b}} - \mathbf{b} \sim \text{normal}(0, \sigma^2 I_m).$$

Then

$$A\hat{\mathbf{x}} - \mathbf{b} = \Pi \epsilon$$

and

$$\tilde{\mathbf{b}} - A\hat{\mathbf{x}} = \tilde{\mathbf{b}} - \Pi \tilde{\mathbf{b}} = (I_m - \Pi) \tilde{\mathbf{b}} = (I_m - \Pi) \epsilon,$$

since $\mathbf{b} = \Pi \tilde{\mathbf{b}} \in \text{Im } A$. Thus $A\hat{\mathbf{x}} - \mathbf{b}$ and $\tilde{\mathbf{b}} - A\hat{\mathbf{x}}$ are the components of ϵ in $\text{Im } A$ and in $(\text{Im } A)^\perp$, respectively. We get

$$\begin{aligned} \mathbb{R}^{2m} \ni \begin{pmatrix} A\hat{\mathbf{x}} - \mathbf{b} \\ \tilde{\mathbf{b}} - A\hat{\mathbf{x}} \end{pmatrix} &= \begin{pmatrix} \Pi \\ I_m - \Pi \end{pmatrix} \epsilon \\ &\sim \text{normal}\left(0, \sigma^2 \begin{pmatrix} \Pi \\ I_m - \Pi \end{pmatrix} I_m (\Pi^\top, I_m - \Pi^\top)\right) = \text{normal}\left(0, \sigma^2 \begin{pmatrix} \Pi & 0 \\ 0 & I_m - \Pi \end{pmatrix}\right) \end{aligned}$$

(Recall that $\Pi = \Pi^\top = \Pi^2$, and recall the behaviour of the normal distribution under linear transformations: $X \sim \text{normal}(\mu, \Sigma)$ implies $BX + c \sim \text{normal}(B\mu + c, B\Sigma B^\top)$.)

Thus $A\hat{\mathbf{x}} - \mathbf{b}$ and $\tilde{\mathbf{b}} - A\hat{\mathbf{x}}$ are independent.

(Geometrically, this is clear because of the rotational invariance of the $\text{normal}(0, \sigma^2 I_m)$ distribution.)

As a consequence,

$$\hat{\mathbf{x}} = (A^\top A)^{-1} A^\top [(A\hat{\mathbf{x}} - \mathbf{b}) + \mathbf{b}] \tag{101}$$

and $A\hat{\mathbf{x}} - \tilde{\mathbf{b}}$ are independent, too. Furthermore,

$$A\hat{\mathbf{x}} - \mathbf{b} \sim \text{normal}(0, \sigma^2 \Pi)$$

and (101) imply

$$\hat{\mathbf{x}} \sim \text{normal}((A^\top A)^{-1} A^\top \mathbf{b}, \sigma^2 (A^\top A)^{-1} A^\top \Pi A (A^\top A)^{-1}) = \text{normal}(\mathbf{x}, \sigma^2 (A^\top A)^{-1}).$$

Finally,

$$\frac{1}{\sigma} (\tilde{\mathbf{b}} - A\hat{\mathbf{x}}) \sim \text{normal}(0, I_m - \Pi)$$

implies

$$\frac{1}{\sigma^2} \|A\hat{\mathbf{x}} - \tilde{\mathbf{b}}\|^2 \sim \chi_{m-n}^2$$

by the rotational invariance of the standard normal distribution, since $I_m - \Pi$ is an orthogonal projection to a $(m - n)$ -dimensional subspace of $\mathbb{R}^m$. □

Example 3.11 (Regression line – reconsidered) Given $x_1, \dots, x_n$ and $y_1, \dots, y_n$, consider the overdetermined system of equations

$$\begin{aligned} ax_1 + b &= y_1, \\ ax_2 + b &= y_2, \\ &\dots \\ ax_n + b &= y_n \end{aligned}$$

for two unknown parameters a and b . In matrix notation, we get

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

To find an approximate solution of this overdetermined system, we use the method of least squares:

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

i.e.

$$\begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & n \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix},$$

which is again the 2×2 linear system for the coefficients of the regression line.

3.3 Sufficient statistics

Example 3.12 (Comparing two frame models for coin tossing) Consider the following two frame models for tossing a (possibly biased) coin n times:

- $\Omega = \{0, 1\}^n$, $\mathcal{P} = \{\text{Bernoulli}(p)^n : 0 \leq p \leq 1\}$. In this model, the whole list of “heads” and “tails” is recorded.
- $\Omega' = \{0, 1, \dots, n\}$, $\mathcal{P}' = \{\text{binomial}(n, p) : 0 \leq p \leq 1\}$. Here, only the number of “heads” is recorded.

The first model can be mapped to the second model by using the sum statistic

$$X : \Omega \rightarrow \Omega', \quad X(\omega_1, \dots, \omega_n) = \sum_{i=1}^n \omega_i.$$

Intuitively, one has the impression that no “relevant” information gets lost in this “compression procedure”: The precise order of the “heads” and “tails” seems to be irrelevant. Assume that we observe $k = X(\omega)$, but not $\omega = (\omega_1, \dots, \omega_n)$. What do we know about ω , when the “true” distribution for the

random experiment is $P_p = \text{Bernoulli}(p)^n$? The conditional distribution of ω with respect to P_p is the uniform distribution on

$$\{X = k\} = \{(\omega_1, \dots, \omega_n) \in \Omega : \sum_{i=1}^n \omega_i = k\},$$

since every $\omega \in \{X = k\}$ has the same probability

$$P_p[\{\omega\}] = p^k(1-p)^{n-k}.$$

In particular, *this conditional distribution does not depend on the choice of $P_p \in \mathcal{P}$* . The extra information “precise order of heads and tails”, which gets lost by keeping only $X(\omega)$, has *the same conditional distribution given $X(\omega)$ for all $P_p \in \mathcal{P}$* . Consequently, it cannot provide any information about the unknown P_p . We say that the statistic X is *sufficient*; it contains already all relevant information.

Recall the notion of conditional distributions: Let $X : \Omega \rightarrow \mathbb{R}^m$ be a random variable on Ω , endowed with the probability distribution P . In the discrete case, the conditional distribution $P[\cdot | X = k]$ of P given $X = k$ is given by the conditional probabilities

$$P[A|X = k] = \frac{P[A \cap \{X = k\}]}{P[X = k]}$$

for all events $A \subseteq \Omega$.

Here is the property of conditional distributions which is most important for us:

General conditional distributions. The conditional distribution $P[\cdot | X]$ with respect to P given X is a probability measure $A \mapsto P[A|X]$, depending only on the value of X , such that for all events $A \subset \Omega$ and for all random variables $f(X)$, depending only on X , one has

$$E_P[P[A|X]f(X)] = E_P[1_A f(X)], \quad (102)$$

whenever the right hand side is well-defined.

This characteristic property of conditional distributions holds in general, for the discrete case and for the continuous case. To illustrate the characteristic property (102), let us check it in the discrete case:

Let $f(X)$ be a random variable depending only on X . Then

$$\begin{aligned} E_P[P[A|X]f(X)] &= \sum_k P[A|X = k]f(k)P[X = k] \\ &= \sum_k \frac{P[A \cap \{X = k\}]}{P[X = k]}f(k)P[X = k] \\ &= \sum_k P[A \cap \{X = k\}]f(k) \\ &= \sum_k \sum_{\omega: X(\omega)=k} 1_A(\omega)f(k)P[\{\omega\}] \\ &= \sum_{\omega \in \Omega} 1_A(\omega)f(X(\omega))P[\{\omega\}] \\ &= E_P[1_A f(X)]. \end{aligned}$$

We have prepared all tools to define sufficient statistics in general:

Definition 3.13 (Sufficient statistics) Consider a statistic $X : \Omega \rightarrow \mathbb{R}^m$ on a frame model $(\Omega, \mathcal{P})$. The statistic X is called *sufficient*, if the conditional distribution $P[\cdot | X]$ is the same for all $P \in \mathcal{P}$.

(Precisely speaking, one has the freedom to modify “the” conditional distribution on events of probability 0. We ignore this fact in this lecture.)

Intuitively speaking, the extra information, which is not encoded in X , is described by $P[\cdot | X]$. If this extra information does not depend on P , it does not help us at all to recover relevant information about $P \in \mathcal{P}$.

Theorem 3.14 (A criterion for sufficiency) Let $(\Omega, \mathcal{P})$ be a frame model and $X : \Omega \rightarrow \mathbb{R}^m$ be a statistic. For all $P, P' \in \mathcal{P}$, assume that the likelihood ratio dP/dP' exists and depends only on the value of X :

$$\frac{dP}{dP'}(\omega) = f_{P,P'}(X(\omega)), \quad \omega \in \Omega.$$

Then X is a sufficient statistic.

Proof. Let $A \subseteq \Omega$ be an event. We check that $P[A|X] = P'[A|X]$, i.e. the conditional probability of A given X does not depend on the choice of $P \in \mathcal{P}$. Let $f(X) \geq 0$ be a random variable that depends only on X . Let E, E' denote the expectation corresponding to P, P' , respectively. We get, using that $dP/dP'(\omega)$ is a function of $X(\omega)$:

$$\begin{aligned} E[P'[A|X]f(X)] &= E' \left[P'[A|X]f(X) \frac{dP}{dP'} \right] \\ &= E' \left[1_A f(X) \frac{dP}{dP'} \right] \\ &= E[1_A f(X)] \\ &= E[P[A|X]f(X)], \end{aligned}$$

and thus $P'[A|X] = P[A|X]$ (with P -probability 1), since the calculation holds for all choices of $f(X)$.

(Recall that for $|E[h(X)]|, |E[h'(X)]| < \infty$, $E[h(X)f(X)] = E[h'(X)f(X)]$ for all f implies $h(X) = h'(X)$ with probability 1.)

Hence X is sufficient. □

The following corollary yields a large class of examples for sufficient statistics:

Corollary 3.15 (Sufficient statistics for exponential classes) Let $(\Omega, \mathcal{P})$ be an exponential class with natural statistic X . Then X is sufficient.

Proof. Let $\theta \in \Theta$ denote the natural parameter corresponding to the natural statistic X (recall Definition 2.7). According to the definition of exponential classes, one has a likelihood function

$$f(\omega|\theta) = \frac{g(\omega)}{Z(\theta)} e^{\langle \theta, X(\omega) \rangle}.$$

Let $\theta, \theta' \in \Theta$ be two parameters, and $P_\theta, P_{\theta'} \in \mathcal{P}$ be the corresponding probability distributions. Then

$$\frac{dP_\theta}{dP_{\theta'}}(\omega) = \frac{f(\omega|\theta)}{f(\omega|\theta')} = \frac{Z(\theta')}{Z(\theta)} e^{\langle \theta - \theta', X(\omega) \rangle}$$

is a function $f_{\theta, \theta'}(X(\omega))$ of $X(\omega)$; hence Theorem 3.14 is applicable. □

Example 3.16 A Geiger counter is read off every second out of n seconds. In second i , it detects N_i α -particles. We use the frame model

$$(N_1, \dots, N_n) \sim \text{Poisson}(\lambda)^n, \quad \lambda > 0.$$

This is a n -th power of an exponential class; thus it is an exponential class. The natural statistic is the total number of detected particles

$$N = \sum_{i=1}^n N_i.$$

By corollary 3.15, the total number N of counts is a sufficient statistic.

Example 3.17 Let $X_1, \dots, X_n$ be i.i.d. $\text{normal}(\mu, \sigma^2)$ random variables with unknown μ and unknown σ^2 . By Example 2.8, the one-dimensional normal distributions $X \sim \text{normal}(\mu, \sigma^2)$ form an exponential class with natural statistic X, X^2 ; thus $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a natural statistic for the power class $\{\text{normal}(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 > 0\}$. Hence $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient. The average $\bar{X}$ and the empirical variance s_X^2 encode the same information as this statistic, since one can determine $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ from $(\bar{X}, s_X^2)$ and vice versa:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \tag{103}$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{1}{n(n-1)} \left(\sum_{i=1}^n X_i \right)^2 \tag{104}$$

and

$$\sum_{i=1}^n X_i = n\bar{X} \tag{105}$$

$$\sum_{i=1}^n X_i^2 = (n-1)s_X^2 + n\bar{X}^2 \tag{106}$$

Thus $(\bar{X}, s_X^2)$ is sufficient, too. Intuitively speaking, this means:

The sample mean $\bar{X}$ and the empirical standard deviation s_X^2 contain already the whole relevant information about an i.i.d. normal sample $X_1, \dots, X_n$.

3.4 Uniform reduction of the expected quadratic error

Here is another measure to evaluate the quality of an estimator:

Definition 3.18 (Expected quadratic error) Let $\hat{\theta}$ be an estimator for the parameter $\theta : \mathcal{P} \rightarrow \mathbb{R}^m$. The expected quadratic error is defined to be

$$E_P[|\hat{\theta} - \theta(P)|^2], \quad P \in \mathcal{P}.$$

If $\hat{\theta}$ is unbiased, then

$$E_P[|\hat{\theta} - \theta(P)|^2] = \text{Var}_P[\hat{\theta}].$$

An important tool in this section (and in many other contexts) is the notion of *conditional expectation*

$E[Y|X]$ of a random variable Y , given another random variable X .

For discrete X , the conditional expectation equals

$$E[Y|X = k] = \frac{E[Y 1_{\{X=k\}}]}{P[X = k]}.$$

Let us recall some basic properties of conditional expectations, which hold in general, for discrete and for continuous random variables:

Conditional expectations. Let X and Y be random variables on $(\Omega, \mathcal{P})$. Assume that $E[Y]$ is finite. The conditional expectation $E[Y|X]$ of Y given X is the expectation of Y with respect to the conditional distribution $P[\cdot|X]$. It is a random variable depending on the value of X , such that

$$E[E[Y|X]f(X)] = E[Yf(X)] \quad (107)$$

holds for all random variables $f(X)$ depending only on X for which $E[Yf(X)]$ is well defined. In particular,

$$E[Yg(X)|X] = E[Y|X]g(X) \quad (108)$$

holds for every random variable $g(X)$ depending only on X for which the right hand side is well defined.

(Strictly speaking, many statements about conditional probabilities hold only with probability 1. We ignore this fact here.)

Indeed (108) follows from (107), since $E[Y|X]g(X)$ depends only on X , and since

$$E[E[Y|X]g(X)f(X)] = E[Yg(X)f(X)]$$

holds for all f for which the right hand side is well defined.

(Recall again that for $|E[h(X)]|, |E[h'(X)]| < \infty$, $E[h(X)f(X)] = E[h'(X)f(X)]$ for all f implies $h(X) = h'(X)$ with probability 1.)

To illustrate (107), let us check it in the discrete case:

$$\begin{aligned} E[E[Y|X]f(X)] &= \sum_k E[Y|X = k]f(k)P[X = k] \\ &= \sum_k \frac{E[Y 1_{\{X=k\}}]}{P[X = k]} f(k)P[X = k] \\ &= \sum_k E[Y 1_{\{X=k\}}] f(k) \\ &= E \left[Y \sum_k 1_{\{X=k\}} f(k) \right] \\ &= E[Yf(X)]. \end{aligned}$$

Let us now use conditional expectations to lower the expected quadratic error of an estimator in the presence of a sufficient statistic:

Given a frame model $(\Omega, \mathcal{P})$, let $\hat{\theta}$ be an estimator of a parameter $\theta : \mathcal{P} \rightarrow \Theta$. Assume that $\hat{\theta}$ has a finite variance $\text{Var}_P[\hat{\theta}] < \infty$ for all $P \in \mathcal{P}$. Let X be a sufficient statistic. By the definition of sufficiency, the conditional distribution $P[\cdot|X]$ is the same for all $P \in \mathcal{P}$; thus the conditional expectation $E_P[\hat{\theta}|X]$ with respect to $P \in \mathcal{P}$ is the same for all $P \in \mathcal{P}$, too; we write $E_P[\hat{\theta}|X] = E_{\mathcal{P}}[\hat{\theta}|X]$.

Theorem 3.19 (Rao–Blackwell) *If X is a sufficient statistic, then for all $P \in \mathcal{P}$,*

$$E_P[|E_P[\hat{\theta}|X] - \theta(P)|^2] \leq E_P[|\hat{\theta} - \theta(P)|^2]. \quad (109)$$

If $\hat{\theta}$ is an unbiased estimator for θ , then $E_P[\hat{\theta}|X]$ is an unbiased estimator, too, and

$$\text{Var}_P[E_P[\hat{\theta}|X]] \leq \text{Var}_P[\hat{\theta}]. \quad (110)$$

Equality in (109) and (110) holds if and only if $E_P[\hat{\theta}|X] = \hat{\theta}$, i.e. if $\hat{\theta}$ is a function of X (with probability 1).

This means that one can lower the expected quadratic error of an estimator uniformly by taking the expectation with respect to a sufficient statistic.

Proof. Let $P \in \mathcal{P}$. We abbreviate $\hat{\theta}' := E_P[\hat{\theta}|X]$ and $\theta = \theta(P)$. By (108),

$$E_P[\hat{\theta}'^2] = E_P[E_P[\hat{\theta}\hat{\theta}'|X]] = E_P[\hat{\theta}\hat{\theta}']$$

and

$$E_P[(\hat{\theta} - \hat{\theta}')(\hat{\theta}' - \theta)] = E_P[\hat{\theta}\hat{\theta}'] - E_P[\hat{\theta}'^2] + E_P[\hat{\theta}']\theta - E_P[\hat{\theta}]\theta = 0$$

holds; note that $E_P[\hat{\theta}'] = E_P[\hat{\theta}]$. We use the decomposition

$$\hat{\theta} - \theta = (\hat{\theta} - \hat{\theta}') + (\hat{\theta}' - \theta)$$

in a “Pythagoras-like” argument:

$$E_P[(\hat{\theta} - \theta)^2] = E_P[(\hat{\theta} - \hat{\theta}')^2] + 2E_P[(\hat{\theta} - \hat{\theta}')(\hat{\theta}' - \theta)] + E_P[(\hat{\theta}' - \theta)^2] \quad (111)$$

$$= E_P[(\hat{\theta} - \hat{\theta}')^2] + E_P[(\hat{\theta}' - \theta)^2] \quad (112)$$

$$\geq E_P[(\hat{\theta}' - \theta)^2], \quad (113)$$

where we have equality in the last step if and only if $\hat{\theta}' = \hat{\theta}$ with probability 1. This proves (109). (110) is an immediate consequence of (109), since for unbiased estimators we have

$$E_P[\hat{\theta}] = E_P[\hat{\theta}'] = \theta(P)$$

and thus

$$\text{Var}_P[\hat{\theta}'] = E_P[(\hat{\theta}' - \theta)^2] \leq E_P[(\hat{\theta} - \theta)^2] = \text{Var}_P[\hat{\theta}].$$

This finishes the proof of Rao–Blackwell’s theorem. □

Example 3.20 Let $X_1, \dots, X_n$ be i.i.d. normal(μ, σ^2) random variables with unknown parameters μ, σ^2 . Since $E_{\mu, \sigma^2}[X_1] = \mu$ holds, X_1 is an unbiased estimator for μ with variance σ^2 . By example 3.17, $T = (\bar{X}, s_X^2)$ is a sufficient statistic. Let us calculate the conditional expectation of X_i given T : $E_{\mu, \sigma^2}[X_i|T]$ is a function of T only, and by permutation symmetry of the $X_1, \dots, X_n$, it cannot depend on i . Thus

$$E_{\mu, \sigma^2}[X_1|T] = \frac{1}{n} \sum_{i=1}^n E_{\mu, \sigma^2}[X_i|T] = E_{\mu, \sigma^2} \left[\frac{1}{n} \sum_{i=1}^n X_i \middle| T \right] = E_{\mu, \sigma^2}[\bar{X}|T] = \bar{X}.$$

Since $\bar{X}$ is not equal to X_1 with probability 1 (at least for $n > 1$), Rao–Blackwell’s theorem implies that $\bar{X}$ is an unbiased estimator, too, with a lower variance than X_1 . Indeed, we know

$$\text{Var}_{\mu, \sigma^2}[\bar{X}] = \frac{\sigma^2}{n} < \sigma^2 = \text{Var}_{\mu, \sigma^2}[X_1].$$

3.5 A lower bound for the expected quadratic estimation error

In the last subsection, we described a method to reduce the variance of an unbiased estimator. However, the variance cannot be reduced arbitrarily far. There is a principal lower bound:

Theorem 3.21 (Cramér–Rao inequality, synonym: information inequality) *Consider a one-parametric frame model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with likelihood function $f : \Omega \times \Theta \rightarrow]0, \infty[$, $\Theta \subseteq \mathbb{R}$ being an open set of parameters. Let $\hat{\theta} : \Omega \rightarrow \mathbb{R}$ be an unbiased estimator for θ . Assume that for all $\theta_0 \in \Theta$, there is a neighborhood U of θ_0 such that*

$$E_{\theta_0} \left[\sup_{\theta \in U} \left| \frac{\partial}{\partial \theta} \frac{dP_\theta}{dP_{\theta_0}} \right| \right] < \infty \quad \text{and} \quad E_{\theta_0} \left[\sup_{\theta \in U} \left| \hat{\theta} \frac{\partial}{\partial \theta} \frac{dP_\theta}{dP_{\theta_0}} \right| \right] < \infty. \quad (114)$$

Then the following lower bound for the variance of the estimator holds for all $\theta \in \Theta$:

$$\text{Var}_\theta[\hat{\theta}] \geq \frac{1}{\text{Var}_\theta \left[\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right]}. \quad (115)$$

In many applications, the technical condition (114) is fulfilled. Then (115) yields a lower bound for the variance of an unbiased estimator which does *not* depend on the estimator.

Definition 3.22 (Fisher information) *The quantity*

$$I_{\mathcal{P}}(\theta) = \text{Var}_\theta \left[\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right]$$

is called Fisher information.

The main tools in the proof of Cramér–Rao’s theorem are an exchange of differentiation and expectation and the Cauchy-Schwarz inequality.

The Cauchy-Schwarz inequality. If U and V are random variables with finite variance, then the following inequality holds:

$$E[UV]^2 \leq E[U^2]E[V^2].$$

(Here is a proof of the Cauchy-Schwarz inequality: If $E[U^2] = 0$ or $E[V^2] = 0$, then $UV = 0$ with probability one. Thus the Cauchy-Schwarz inequality holds in this case. In the case $E[U^2] > 0$ and $E[V^2] > 0$, we set

$$U' = \frac{U}{\sqrt{E[U^2]}}, \quad V' = \frac{V}{\sqrt{E[V^2]}}.$$

We get

$$0 \leq E[(U' \pm V')^2] = E[U'^2] \pm 2E[U'V'] + E[V'^2] = 1 \pm 2 \frac{E[UV]}{\sqrt{E[U^2]E[V^2]}} + 1.$$

The Cauchy-Schwarz inequality is an immediate consequence.)

Proof of the Cramér–Rao inequality. We check the information inequality at a point $\theta_0 \in \Theta$. Without loss of generality, we may assume $\theta_0 = 0$; otherwise we replace the parameter θ by the shifted parameter $\theta - \theta_0$, and replace the unbiased estimator $\hat{\theta}$ by $\hat{\theta} - \theta_0$, too. This assumption just serves to

simplify the notation. Since $\hat{\theta}$ is an unbiased estimator for θ , we get for all $\theta \in U$:

$$\theta = E_{\theta}[\hat{\theta}] = E_0 \left[\hat{\theta} \frac{dP_{\theta}}{dP_0} \right].$$

We differentiate with this equation respect to θ and use the Cauchy-Schwarz inequality. The technical condition (114) guarantees that we may exchange differentiation and expectation:

$$1 = \frac{d}{d\theta} E_0 \left[\hat{\theta} \frac{dP_{\theta}}{dP_0} \right] = E_0 \left[\hat{\theta} \frac{\partial}{\partial \theta} \frac{dP_{\theta}}{dP_0} \right] \leq \sqrt{E_0[\hat{\theta}^2] E_0 \left[\left(\frac{\partial}{\partial \theta} \frac{dP_{\theta}}{dP_0} \right)^2 \right]}. \quad (116)$$

(We use the following fact without proof: Let X be a random variable and U be an open set, and g be a function of two variables which is differentiable in the second variable. If

$$\sup_{y \in U} \left| \frac{\partial}{\partial y} g(X, y) \right|$$

has a finite expectation, then

$$E \left[\frac{\partial}{\partial y} g(X, y) \right] = \frac{d}{dy} E[g(X, y)]$$

holds for all $y \in U$.)

We evaluate (116) at $\theta = \theta_0 = 0$ and observe $\text{Var}_0[\hat{\theta}] = E_0[\hat{\theta}^2]$. Furthermore,

$$\left. \frac{\partial}{\partial \theta} \frac{dP_{\theta}}{dP_0}(\omega) \right|_{\theta=0} = \left. \frac{\partial}{\partial \theta} \frac{f(\omega|\theta)}{f(\omega|0)} \right|_{\theta=0} = \left. \frac{\partial}{\partial \theta} \log f(\omega|\theta) \right|_{\theta=0}$$

holds.

We get the following version of the Cramér-Rao inequality:

$$\text{Var}_0[\hat{\theta}] E_0 \left[\left(\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right)^2 \right] \Big|_{\theta=0} \geq 1. \quad (117)$$

To derive the information inequality (115) from this version, we calculate:

$$\begin{aligned} E_0 \left[\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right] \Big|_{\theta=0} &= E_0 \left[\frac{\partial}{\partial \theta} \frac{f(\omega|\theta)}{f(\omega|0)} \right] \Big|_{\theta=0} \\ &= E_0 \left[\frac{\partial}{\partial \theta} \frac{dP_{\theta}}{dP_0} \right] \Big|_{\theta=0} \\ &= \frac{d}{d\theta} E_0 \left[\frac{dP_{\theta}}{dP_0} \right] \Big|_{\theta=0} \\ &= \frac{d}{d\theta} E_{\theta}[1] \Big|_{\theta=0} = 0, \end{aligned}$$

where exchanging the derivative and expectation is again allowed because of the technical condition (114). We get

$$E_0 \left[\left(\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right)^2 \right] \Big|_{\theta=0} = \text{Var}_0 \left[\frac{\partial}{\partial \theta} \log f(\omega|\theta) \right] \Big|_{\theta=0}.$$

Combining this with (117) finishes the proof of the Cramér Rao theorem. □

Lemma 3.23 (Additivity of the Fisher information) *The Fisher information is additive with respect to i.i.d. repetitions:*

$$I_{\mathcal{P}^n}(\theta) = nI_{\mathcal{P}}(\theta).$$

Proof: If $f(\omega|\theta)$ is the likelihood function for $\mathcal{P}$, then $\prod_{i=1}^n f(\omega_i|\theta)$ is the likelihood function for an n -fold i.i.d. repetition. Let $(\omega_1, \dots, \omega_n) \sim P_{\theta}^n$. Then

$$\frac{\partial}{\partial \theta} \log f(\omega_i|\theta), \quad i = 1, \dots, n$$

are i.i.d. random variables. We get

$$I_{\mathcal{P}^n}(\theta) = \text{Var}_{P_{\theta}^n} \left[\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(\omega_i|\theta) \right] = \sum_{i=1}^n \text{Var}_{P_{\theta}^n} \left[\frac{\partial}{\partial \theta} \log f(\omega_i|\theta) \right] = nI_{\mathcal{P}}(\theta).$$

□

Example 3.24 (Optimal unbiased estimators for the expectation in a normal model) Let us calculate the Fisher information for the parameter μ in the parametric model

$$\mathcal{P} = \{\text{normal}(\mu, \sigma^2) : \mu \in \mathbb{R}\},$$

where $\sigma^2 > 0$ is fixed. Let $X \sim \text{normal}(\mu, \sigma^2)$. Then

$$I_{\mathcal{P}}(\mu) = \text{Var}_{\mu, \sigma^2} \left[\frac{\partial}{\partial \mu} \left(\frac{(X - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} \right) \right] = \text{Var}_{\mu, \sigma^2} \left[\frac{\mu - X}{\sigma^2} \right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Let $X_1, \dots, X_n$ be i.i.d. $\text{normal}(\mu, \sigma^2)$ random variables and $\hat{\mu}$ be an unbiased estimator for μ such that the technical condition (114) holds. Then the Cramér-Rao theorem yields

$$\text{Var}_{\mu, \sigma^2}[\hat{\mu}] \geq \frac{1}{I_{\mathcal{P}^n}(\mu)} = \frac{1}{nI_{\mathcal{P}}(\mu)} = \frac{\sigma^2}{n}.$$

This lower bound is reached for the unbiased estimator $\hat{\mu} = \bar{X}$. In this sense, the mean of the data is an optimal *unbiased* estimator.

(A famous result of Stein shows that one cannot drop the assumption that $\hat{\mu}$ is unbiased in this example: At least for normal distributions in higher dimensions, there are (biased) estimators for μ that have a uniformly smaller expected quadratic error than the expected quadratic error of the mean $\bar{X}$ as an estimator of the expectation μ .)