Ramp Up Mathematics
Stochastics

Summer Semester 2024

Institute for Mathematical Stochastics
TU Braunschweig

Preface

These notes contain basics in probability theory and statistics as a crash course to bring us up to speed in the modelling and analysis of phenomena involving some form of randomness or at least complexity or unpredictability that justifies the probabilistic perspective. This is a shortened version of an earlier manuscript and limited, with a few exceptions, to non-asymptotic probability. This has the advantage of greater accessibility but necessarily falls short in presenting the immensely important concepts of convergence of random variables and their distributions. In essence, these notes cover the definitions and main results (without proofs of course) of the first few weeks of an introductory course in probability and an introductory course in statistics. In order to be prepared to attend a masters-level course in probability theory or statistics, we highly recommend to consult further textbooks on the subject like the following:

- Georgii, H.-O. (2008). Stochastics: Introduction to Probability and Statistics. De Gruyter.
- Bickel, P.J. and Doksum, K.A. (2001). Mathematical Statistics: Basic Ideas and Selected Topics. Prentice Hall.
- Wassermann, L. (2004). All of Statistics. A Concise Course in Statistical Inference. Springer Texts in Statistics.

Your lecturers of the Institute of Mathematical Stochastics.

May 01, 2024

Contents

1 Basics of probability theory 1

1.1 Introduction . . . 1

1.2 Discrete distributions and distributions on $\mathbb{R}$ . . . 2

1.3 Independent events and conditional probabilities . . . 5

1.4 Random variables and their properties . . . 6

1.5 Conditional expectations . . . 10

1.6 Normal distributions . . . 11

2 Basics of statistics 13

2.1 Introduction . . . 13

2.2 Point estimation . . . 14

2.3 Confidence intervals . . . 17

2.4 Statistical tests . . . 19

2.5 Linear regression . . . 25

1 Basics of probability theory

1.1 Introduction

In probability, an ***experiment*** refers to any action or activity whose outcome is subject to uncertainty. Such experiments are mathematically described by means of a **probability space** which is a triplet $(\Omega, \mathcal{A}, \mathbb{P})$ such that:

- Ω is a nonempty set. It is called the **sample space** of an experiment. It contains all possible outcomes of that experiment. Elements $\omega \in \Omega$ are called **outcomes**, ***realizations*** or ***simple events***.
- $\mathcal{A}$ is a system of subsets of Ω which describes all possible ***events*** occurring in the experiment. The system $\mathcal{A}$ has the following properties:

  i) $\Omega \in \mathcal{A}$;
  ii) $A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A} \quad (A^c := \Omega \setminus A$ is the complement of $A)$;
  iii) $A_1, \dots, A_n, \dots \in \mathcal{A} \quad \Rightarrow \quad \bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$.

  A system of subsets of Ω with (i)-(iii) is called a ***σ-algebra***. From the properties (i)-(iii) we can derive $\emptyset \in \mathcal{A}$, and that for any two events $A, B \in \mathcal{A}$, also their union $A \cup B$, their intersection $A \cap B$ and their set difference $A \setminus B$ are events (i.e., elements of $\mathcal{A}$). Interpretation:

$A \cup B$	A or B has occured
$A \cap B$	A and B have occured
$B \setminus A$	B has occured, but not A
$A^c := \Omega \setminus A$	A has not occured
$A \subset B$	if A occurs, so does B

  Events A and B are called ***disjoint*** or ***mutually exclusive***, if $A \cap B = \emptyset$.

- $\mathbb{P}$ is a ***probability (measure)***, i.e., a mapping from $\mathcal{A}$ to $[0, 1]$ which assigns to each event A a number $\mathbb{P}(A) \in [0, 1]$ called ***probability of the event*** A. A probability measure $\mathbb{P}$ has the following properties:

  i) $\mathbb{P}(\emptyset) = 0$;
  ii) $\mathbb{P}(\Omega) = 1$;
  iii) ***σ-additivity:*** For any countable collection of disjoint events $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ holds
  $$\mathbb{P}\left(\dot{\bigcup_{k=1}^{\infty}} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

Proposition 1.1

i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

ii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any events A, B. In particular, $\mathbb{P}(A \dot{\cup} B) = \mathbb{P}(A) + \mathbb{P}(B)$ for any two disjoint events A, B.

We call an experiment ***discrete***, if the set Ω is finite or countably infinite. In this case, consider a mapping $p : \Omega \to [0, \infty)$ such that $\sum_{\omega \in \Omega} p(\omega) = 1$. Then there exists a unique probability measure $\mathbb{P}$ on $\mathfrak{P}(\Omega)$ such that $\mathbb{P}(\{\omega\}) = p(\omega)$ for all $\omega \in \Omega$. The mapping p is called ***probability mass function (PMF)*** of $\mathbb{P}$. The probability of each event A can then be calculated as

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega).$$

1.2 Discrete distributions and distributions on $\mathbb{R}$

Definition 1.2: Discrete uniform distributions

Let Ω be a non-empty and finite set. We denote by $|A|$ the number of elements in $A \subset \Omega$. The mapping $p : \Omega \to \mathbb{R}$, $\omega \mapsto \frac{1}{|\Omega|}$ defines a PMF on Ω. The corresponding discrete experiment is called ***Laplace experiment*** on Ω. The probability measure $\mathbb{P}$ induced by p,

$$\mathbb{P} : \mathfrak{P}(\Omega) \to \mathbb{R}, \qquad A \mapsto \frac{|A|}{|\Omega|},$$

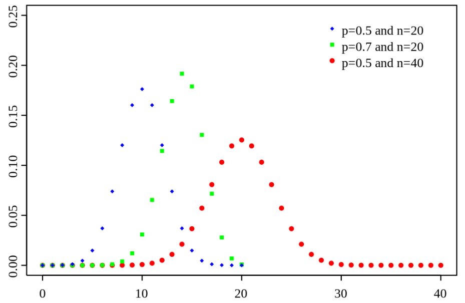
is called ***discrete uniform distribution*** on $(\Omega, \mathfrak{P}(\Omega))$.

Definition 1.3: Binomial and Bernoulli distributions

Let $\Omega = \{0, 1, \dots, n\}$. For $k \in \Omega$ and some fixed $p \in [0, 1]$, define

$$B_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}.$$

Then $B_{n,p}$ is a PMF on Ω. The corresponding experiment is called ***binomial experiment***. The probability measure on $(\Omega, \mathfrak{P}(\Omega))$ induced by $B_{n,p}$ is called ***binomial(n, p) distribution***. In case $n = 1$, we speak of a ***Bernoulli experiment*** and the ***Bernoulli(p)-distribution***.



https://commons.wikimedia.org/w/index.php?curid=3646951

PMF of the binomial distribution. Source:

Theorem 1.4: Poisson approximation of the binomial distribution

If, in a binomial experiment, the success probability p_n depends on the number n of trials such that

$$\lim_{n\to\infty} np_n = \lambda \in (0,\infty),$$

then, for every $k \in \mathbb{N}_0$,

$$B_{n,p_n}(k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \longrightarrow e^{-\lambda}\frac{\lambda^k}{k!}, \qquad n \to \infty.$$

The probability measure on $(\mathbb{N}_0, \mathfrak{P}(\mathbb{N}_0))$ induced by the PMF $\mathrm{Poiss}_\lambda(k) := e^{-\lambda}\lambda^k/k!$ is called the ***Poisson(λ)-distribution***.

The sample space Ω is sometimes given by $\mathbb{R}$ (or by an interval $I \subset \mathbb{R}$). In this case, we take as $\mathcal{A}$ the smallest σ-algebra which contains all intervals (or all subintervals of I). This σ-algebra is called ***Borel σ-algebra*** and is denoted by $\mathcal{B}(\mathbb{R})$ (or $\mathcal{B}(I)$ respectively).

Definition 1.5: Distribution functions

A function $F : \mathbb{R} \to \mathbb{R}$ is called a ***cumulative distribution function*** (CDF) if F is a monotone increasing and right-continuous function such that

$$\lim_{x\to-\infty} F(x) = 0 \qquad \text{and} \qquad \lim_{x\to+\infty} F(x) = 1.$$

Proposition 1.6: Characterization via distribution functions

For each CDF F there exists a unique probability measure $\mathcal{P}_F$ on $\mathcal{B}(\mathbb{R})$ such that

$$\mathcal{P}_F((a,b]) = F(b) - F(a) \qquad \text{for all} \quad a \le b \in \mathbb{R}. \tag{1.2.1}$$

And, vice versa, for each probability measure $\mathcal{P}$ on $\mathcal{B}(\mathbb{R})$ there exists a unique CDF F with $\mathcal{P} = \mathcal{P}_F$.

Definition 1.5 and Formula (1.2.1) yield $F(x) = \mathcal{P}_F((-\infty, x])$ for each $x \in \mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ be (Riemann-)integrable over any bounded interval and fulfill

- $f(x) \ge 0$ for all $x \in \mathbb{R}$;
- $\int_{-\infty}^{+\infty} f(x)dx$ exists and $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Then f is called a ***probability density function (PDF)***. The function

$$F : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \int_{-\infty}^{x} f(y)dy$$

is then a CDF and hence defines a unique probability measure $\mathcal{P}_F$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\mathcal{P}_F((a,b]) = \int_a^b f(y)dy. \tag{1.2.2}$$

For probability measures constructed via PDFs holds $\mathcal{P}_F(\{x\}) = 0$ for any $x \in \mathbb{R}$.

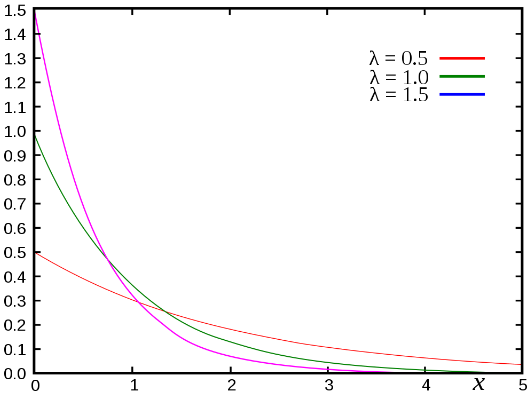
Example 1.7: Continuous distributions

- ***Uniform distribution on** $[a,b]$* (with parameters $a < b \in \mathbb{R}$):

$$f(x) := \frac{1}{b-a}\mathbb{1}_{[a,b]}(x).$$

- ***Exponential distribution*** (with parameter $\lambda > 0$):

$$f(x) := \lambda e^{-\lambda x}\mathbb{1}_{\{x\geq 0\}}(x).$$

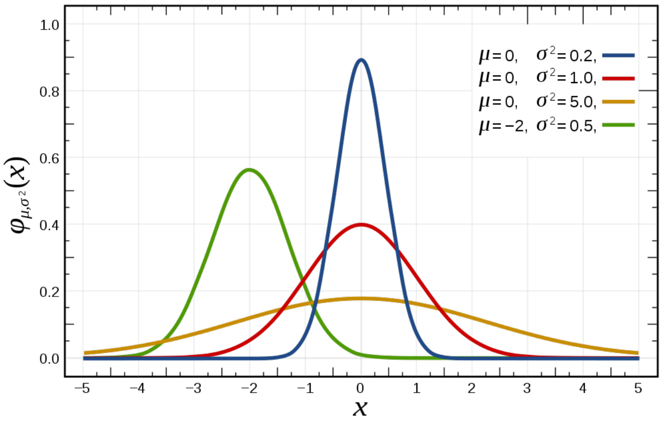


Density of the exponential distribution. Source: https://commons.wikimedia.org/w/index.php?curid=90524175

- ***Normal distribution*** (with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$):

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

With $\mu = 0$ and $\sigma = 1$ we obtain the ***standard normal distribution***.



Density of the normal distribution. Source: https://commons.wikimedia.org/w/index.php?curid=3817954

1.3 Independent events and conditional probabilities

One of the key concepts in probability theory is stochastic independence which should be understood as a precise description of the intuitive idea that events do not influence each other.

Definition 1.8: Independence

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

i) Events A, $B \in \mathcal{A}$ are called ***independent*** (notation: $A \perp\!\!\!\perp B$), if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

ii) A family $(A_i)_{i \in \mathcal{I}}$ of events, $\mathcal{I} \neq \emptyset$ arbitrary, is called ***independent***, if for all finite $\mathcal{J} \subset \mathcal{I}$,

$$\mathbb{P}\left(\cap_{j \in \mathcal{J}} A_j\right) = \prod_{j \in \mathcal{J}} \mathbb{P}(A_j).$$

The family is called ***pairwise independent***, if for all $i, j \in \mathcal{I}$ with $i \neq j$,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Note that independent events are pairwise independent. However, the reverse does not hold in general.

Definition 1.9: Conditional probabilities

For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and A, $B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$, we call

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

the ***conditional probability of*** A ***given*** B. This is the probability for A to occur under the precondition that B has occurred.

Note that if $A \perp\!\!\!\perp B$ and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Theorem 1.10: Conditional probability measures are probability measures

For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$, we call $\mathcal{A}_B := \{A \cap B \,|\, A \in \mathcal{A}\}$ the ***trace-σ-algebra of*** $\mathcal{A}$ ***on*** B.

$$\mathbb{P}_B : \mathcal{A}_B \to [0, \infty), \quad A \mapsto \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

is a probability measure on $\mathcal{A}_B$ and thus, $(B, \mathcal{A}_B, \mathbb{P}_B)$ is a probability space.

Theorem 1.11: Formulas for conditional probabilities

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(B_i)_{i\in\mathcal{I}}$ a countable family of pairwise disjoint events with $\uplus_{i\in\mathcal{I}} B_i = \Omega$, $\mathbb{P}(B_i) > 0$ for all $i \in \mathcal{I}$ and $A \in \mathcal{A}$. Then,

i) ***Law of total probability***:

$$\mathbb{P}(A) = \sum_{i\in\mathcal{I}} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

ii) ***Bayes' formula***: If $\mathbb{P}(A) > 0$, then for all $j \in \mathcal{I}$ holds

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i\in\mathcal{I}} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Let $(\Omega_1, \mathcal{A}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{A}_2, \mathbb{P}_2)$ be two probability spaces. We construct the ***product space*** as follows.

- As the sample space we take $\Omega := \Omega_1 \times \Omega_2$.
- As the set $\mathcal{A}$ of all events we take the smallest σ-algebra containing all sets of the form $A_1 \times A_2$, $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$. $\mathcal{A}$ is called the ***product σ-algebra*** and is denoted by $\mathcal{A}_1 \otimes \mathcal{A}_2$.
- There exists a unique probability measure $\mathbb{P}$ on $\mathcal{A}_1 \otimes \mathcal{A}_2$ such that

  $$\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2), \quad \forall\ A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2.$$

  This $\mathbb{P}$ is called ***product probability measure*** and is denoted by $\mathbb{P}_1 \otimes \mathbb{P}_2$.
- We call the triplet $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ a ***product*** of $(\Omega_1, \mathcal{A}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{A}_2, \mathbb{P}_2)$.

The product of finitely many probability spaces is defined in a similar way. Product spaces are used to model the independent coupling of experiments.

1.4 Random variables and their properties

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A mapping $X : \Omega \to \mathbb{R}$ is called ***measurable*** if for each set $B \in \mathcal{B}(\mathbb{R})$ holds $X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$. A mapping $X : \Omega \to \mathbb{R}$ is measurable if and only if for each $x \in \mathbb{R}$ holds $X^{-1}((-\infty, x]) \in \mathcal{A}$. A measurable mapping is called a ***random variable (RV)***. Each RV X induces a probability measure $\mathcal{P}_X$ on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ by

$$\mathcal{P}_X(B) := \mathbb{P}(X^{-1}(B)), \qquad \forall\ B \in \mathcal{B}(\mathbb{R}).$$

The probability measure $\mathcal{P}_X$ is called the ***distribution*** of the RV X, the corresponding CDF $F_X : F_X(x) = \mathcal{P}_X((-\infty, x]) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$, is called ***CDF of*** X. If a RV X takes only countably many different values, we call it a ***discrete RV***. If the CDF F_X has a density (PDF) f_X, X is called a ***continuous RV***.

Proposition 1.12: Combinations of random variables

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of RVs on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then

i) $\sum\limits_{n=1}^{N} X_n$ and $\prod\limits_{n=1}^{N} X_n$ are RVs for each $N \in \mathbb{N}$;

ii) $\sup_{n\in\mathbb{N}} X_n$, $\inf_{n\in\mathbb{N}} X_n$, $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are RVs;

iii) $g(X_1, \dots, X_n)$ is a RV for each piecewise continuous function $g : \mathbb{R}^n \to \mathbb{R}$.

Definition 1.13: Expected value

i) Let X be a discrete RV with values x_i, $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty} |x_i| \cdot \mathbb{P}(X = x_i) < \infty$. Then

$$\mathbb{E}[X] := \sum_{i=1}^{\infty} x_i \cdot \mathbb{P}(X = x_i)$$

is called the ***expectation*** or ***mean (value)*** of X.

ii) Let X be a continuous RV with PDF f_X and $\int_{-\infty}^{+\infty} |x| \cdot f_X(x)dx < \infty$. Then

$$\mathbb{E}[X] := \int_{-\infty}^{+\infty} x \cdot f_X(x)dx$$

is called the ***expectation*** or ***mean (value)*** of X.

The expectation of a RV X is a kind of weighted average of the values of X, where the probability $\mathbb{P}$ prescribes the weights. For a general RV X, the expectation $\mathbb{E}[X]$ is defined as the ***Lebesgue integral*** $\mathbb{E}[X] := \int_\Omega X(\omega)d\mathbb{P}(\omega)$. This implies the ***linearity*** of the expectation: For any two RVs X and Y and $a, b \in \mathbb{R}$ holds

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Proposition 1.14: Properties of expectations

i) If X is a discrete RV taking only non-negative integer values, then $\mathbb{E}[X] = \sum\limits_{n=1}^{\infty} \mathbb{P}(X \geq n)$.

ii) Let X be a RV and $g : \mathbb{R} \to \mathbb{R}$ a piecewise continuous function. Then $Y := g(X)$ is again a RV with $\mathbb{E}[Y] = \sum\limits_{i=1}^{\infty} g(x_i) \cdot \mathbb{P}(X = x_i)$ in the discrete case and $\mathbb{E}[Y] = \int_{-\infty}^{+\infty} g(x) \cdot f_X(x)dx$ in the continuous case.

Definition 1.15: Variance

Let X be a RV with finite expectation. Then the ***variance*** of X is given by

$$\operatorname{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

The value $\sqrt{\operatorname{Var}(X)}$ is called ***standard deviation (SD)*** of X. The values $\mathbb{E}[X^k]$ and $\mathbb{E}[|X|^k]$ are called ***k-th moment*** and ***k-th absolute moment*** of X respectively.

Proposition 1.16: Properties of variances

For any RV X holds

i) $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$;

ii) $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X) \quad \forall a, b \in \mathbb{R}$;

iii) $\operatorname{Var}(X) = \min_{a \in \mathbb{R}} \mathbb{E}\left[(X - a)^2\right]$.

Theorem 1.17: Chebyshev-Markov-Inequality

Let X be a RV with finite expectation. For each $\varepsilon > 0$ and each $r > 0$ holds

$$\mathbb{P}\Big(\{|X| \geq \varepsilon\}\Big) \leq \frac{\mathbb{E}[|X|^r]}{\varepsilon^r}.$$

Especially,

$$\mathbb{P}\Big(\{|X - \mathbb{E}[X]| \geq \varepsilon\}\Big) \leq \frac{\operatorname{Var}(X)}{\varepsilon^2}.$$

A family of real-valued RVs $X_1, \ldots, X_n$ can be considered as a single RV $X := (X_1, \ldots, X_n)$ taking values in $\mathbb{R}^n$. Then X gives rise to a probability measure $\mathcal{P}_X$ on the Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ (which is the smallest σ-algebra over $\mathbb{R}^n$ containing all cuboids, i.e., n-dimensional intervals) and to a CDF F_X. This F_X is called the ***joint probability distribution function*** of the RVs $X_1, \ldots, X_n$. We have

$$F_X(x_1, \ldots, x_n) := \mathcal{P}_X\Big((-\infty, x_1] \times \ldots \times (-\infty, x_n]\Big) := \mathbb{P}\Big(X_1 \leq x_1, \ldots, X_n \leq x_n\Big).$$

If all RVs $X_1, \ldots, X_n$ are discrete, their ***joint probability mass function*** p_X is defined such that $p_X(x_1, \ldots, x_n) := \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ for all values x_i of the RV X_i. Then for each set $B \in \mathcal{B}(\mathbb{R}^n)$ holds

$$\mathcal{P}_X(B) = \mathbb{P}\Big((X_1, \ldots, X_n) \in B\Big) = \sum_{x_1, \ldots, x_n \in B} p_X(x_1, \ldots, x_n).$$

To find the ***marginal probability mass function*** p_{X_i} of each particular RV X_i from p_X, sum up over all possible values of the other RVs:

$$p_{X_i}(x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} p_X(x_1, \ldots, x_n).$$

In the continuous case, if there exists a ***joint probability density function*** f_X, the joint probability distribution function F_X is given by

$$F_X(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(t_1, \dots, t_n) dt_1 \cdots dt_n.$$

Then we have for any cuboid $B := \prod_{i=1}^{n} [a_i, b_i] \in \mathcal{B}(\mathbb{R}^n)$

$$\mathcal{P}_X(B) = \mathbb{P}\Big(a_1 \leq X_1 \leq b_1, \dots, a_n \leq X_n \leq b_n\Big) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_X(t_1, \dots, t_n) dt_1 \cdots dt_n.$$

To find the ***marginal probability density function*** f_{X_i} of each particular RV X_i from f_X, integrate f_X w.r.t. all variables which correspond to the other RVs:

$$f_{X_i}(x_i) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(t_1, \dots, t_{i-1}, x_i, t_{i+1}, \dots, t_n) dt_1 \cdots dt_{i-1} dt_{i+1} \cdots dt_n.$$

Definition 1.18: Independence of random variables

Real valued RVs $X_1, \dots, X_n$ are called ***independent***, if for all $x_1, \dots, x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n),$$

i.e., if $F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$ or, equivalently, if $\mathcal{P}_X = \mathcal{P}_{X_1} \otimes \cdots \otimes \mathcal{P}_{X_n}$ with $X = (X_1, \dots, X_n)$.

If all RVs $X_1, \dots, X_n$ are discrete, they are independent if and only if $p_X(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$. If all RVs $X_1, \dots, X_n$ are continuous with marginal densities $f_{X_1}, \dots, f_{X_n}$ and joint density f_X, they are independent if and only if $f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$. For an $\mathbb{R}^n$-valued RV $X = (X_1, \dots, X_n)$ and a piecewise continuous function $g : \mathbb{R}^n \to \mathbb{R}$, the expectation of the real-valued RV $g(X_1, \dots, X_n)$ is defined as follows.

$$\mathbb{E}[g(X_1, \dots, X_n)] := \begin{cases} \sum\limits_{x_1, \dots, x_n} g(x_1, \dots, x_n) p_X(x_1, \dots, x_n), & \text{in the disc. case;} \\ \int\limits_{-\infty}^{+\infty} \cdots \int\limits_{-\infty}^{+\infty} g(t_1, \dots, t_n) f_X(t_1, \dots, t_n) dt_1 \cdots dt_n, & \text{in the cont. case.} \end{cases}$$

Definition 1.19: Covariance and correlation

Let X and Y be two real valued RVs with finite expectation. Then

$$\operatorname{Cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right]$$

is called the ***covariance*** of X and Y. RVs X and Y with $\operatorname{Cov}(X, Y) = 0$ are called ***uncorrelated***. If X and Y possess finite second moments and $\operatorname{Var}(X) \neq 0 \neq \operatorname{Var}(Y)$, then

$$\operatorname{Corr}(X, Y) := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}}$$

is called the ***correlation*** between X and Y.

From the definition follows that $\mathrm{Var}(X) = 0 \Leftrightarrow X = \mathbb{E}[X]$ $\mathbb{P}$-almost surely, i.e., $\mathbb{P}(X = \mathbb{E}[X]) = 1$. Further, $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$ and $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Moreover, $|\mathrm{Corr}(X, Y)| \leq 1$ for all RVs X and Y as well as $|\mathrm{Corr}(X, Y)| = 1 \;\Leftrightarrow\; \exists\, a, b \in \mathbb{R}$, $a \neq 0$, such that $Y = aX + b$ $\mathbb{P}$-a.s. Further, the following ***Cauchy–Schwartz inequality*** holds for any two RVs X and Y with finite variance:

$$\big(\mathrm{Cov}(X, Y)\big)^2 \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$$

Proposition 1.20: Expectation and variance under independence

i) Let $X_1, \ldots, X_n$ be independent RVs with finite expectation. Then

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]. \tag{1.4.1}$$

In particular, $X_1, \ldots, X_n$ are pairwise uncorrelated.

ii) Let $X_1, \ldots, X_n$ be pairwise uncorrelated RVs with finite variancce. Then

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Note that if two RVs X and Y are uncorrelated, they need not be independent.

Proposition 1.21

Let X, Y be independent continuous RVs with PDFs f_X and f_Y, respectively. Then $Z := X + Y$ is a continuous RV whose PDF f_Z is given as the ***convolution*** of f_X and f_Y:

$$f_Z(z) := \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y) dy = \int_{-\infty}^{+\infty} f_Y(z - x) f_X(x) dx.$$

1.5 Conditional expectations

Definition 1.22: Discrete conditional densities

Let X, Y be two discrete RVs on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with joint density $p(x, y) = \mathbb{P}(X = x, Y = y)$. Then, for all x, y with $p_2(y) = \mathbb{P}(Y = y) > 0$ we define

$$p(x|y) := \frac{p(x, y)}{p_2(y)}$$

the ***conditional density of*** x ***given*** y. This is the probability for $X = x$ to occur under the precondition that $Y = y$ has occurred. Then, $\mathbb{E}[X|Y]$, defined as

$$\mathbb{E}[X|Y](\omega) = \sum_{i \geq 0} x_i p(x_i | Y(\omega)),$$

is called the ***conditional expectation of*** X ***given*** Y.

Note that in case X and Y are independent, we have $p(x|y) = p_1(x)$. Further, the conditional expectation $\mathbb{E}[X|Y]$ is a RV with

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{i\geq 0}\sum_{j\geq 0} x_j p(x_j|y_i)p_2(y_i) = \sum_{i,j\geq 0} x_j p(x_j, y_i) = \mathbb{E}[X].$$

Similarly, for continuous RVs X, Y with joint density $f(x, y)$ any function $f(x|y)$ that satisfies

$$\int_A \left(\int_B f(x|y)dx\right) f(y)dy = \int_A \int_B f(x,y)dxdy$$

for all $A, B \in \mathcal{B}(\mathbb{R}^d)$, where $f(y)$ is the density corresponding to Y, is called a conditional density.

1.6 Normal distributions

The normal distribution plays a prominent role in the modelling of complex behaviors. The main reason for this is the following central limit theorem.

Theorem 1.23: Lindeberg-Lévy central limit theorem (CLT)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of iidRVs with $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathrm{Var}(X_1) \in (0, \infty)$. Let

$$\Phi(y) := \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

be the distribution function of the $\mathcal{N}\big(0, 1\big)$-distribution. Then,

$$\sup_{y\in\mathbb{R}} \left| \mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) \leq y\right) - \Phi(y)\right| \xrightarrow{n\to\infty} 0.$$

The CLT presents an example of the so-called ***convergence in distribution***, which is defined as the pointwise convergence of distribution functions F_n, in the CLT case

$$F_n(y) = \mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu) \leq y\right),$$

to some limiting distribution function F, in the CLT case Φ, for all continuity points of F.

Definition 1.24: Multivariate normal distributions

Let $X = (X_1, \ldots, X_d)$ be a RV with values in $\mathbb{R}^d$, $\mu \in \mathbb{R}^d$ and A a positive definite $d \times d$ matrix. If for all $x \in \mathbb{R}^d$

$$\mathbb{P}(X \leq x) = \int_{(-\infty, x]} \frac{1}{\sqrt{\det(2\pi A)}} e^{-\frac{1}{2}\langle y-\mu, A^{-1}(y-\mu)\rangle} dy,$$

we say that X is ***normal distributed*** with expected value μ and covariance matrix A and write $X \sim \mathcal{N}\big(\mu, A\big)$.

Note that since A is positive definite, i.e., $\langle x, Ax \rangle > 0$ for all $x \in \mathbb{R}^d \backslash \{0\}$ where $\langle \cdot, \cdot \rangle$ the d-dimensional scalar product, the inverse A^{-1} exists. Further note that indeed $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X_i, X_j) = A_{ij}$ for all $i, j \in \{1, \dots, d\}$. Degenerate multivariate normal distributions can be defined also for A only positive semidefinite, however densities do not exists if $\langle x, Ax \rangle = 0$ for some $x \neq 0$.

Theorem 1.25: Properties of normal distributions

We have that $X \sim \mathcal{N}\big(\mu, A\big)$ if and only if $\langle \lambda, X \rangle \sim \mathcal{N}\big(\langle \lambda, \mu \rangle, \langle \lambda, A\lambda \rangle\big)$ for all $\lambda \in \mathbb{R}^d$. Further, for $X \sim \mathcal{N}\big(\mu, A\big)$ and $Y \sim \mathcal{N}\big(\nu, B\big)$ independent and with values in $\mathbb{R}^d$ we have that $X + Y \sim \mathcal{N}\big(\mu + \nu, A + B\big)$.

We can construct a Gaussian vector Y with prescribed expected value and covariance matrix from an iidstandard Gaussian vector as follows. Consider $X = (X_1, \dots, X_d)^T$ with $X_i \sim \mathcal{N}\big(0, 1\big)$, $i \in \{1, \dots, d\}$, independent standard-normal RVs in $\mathbb{R}$. Then, for $\mu \in \mathbb{R}^d$ and any $d \times d$ matrix B such that BB^T is positive definite, we have that $Y := BX + \mu \sim \mathcal{N}\big(\mu, BB^T\big)$. Note that BB^T is always positive semidefinite.

Gaussian vectors have the special property that independence and uncorrelatedness are equivalent, i.e., for $X \sim \mathcal{N}\big(\mu, A\big)$ it holds that X_i, X_j are independent if and only if $A_{ij} = 0$.

Let us finally mention that the class of multivariate normal distributions is closed under conditioning. That is, for $(X, Y) \sim \mathcal{N}\big(\mu, A\big)$ a two-dimensional Gaussian vector, then the conditional density $f(x|y)$ is the density of a normal distribution in $\mathbb{R}$ with $\mu_{x|y} = \mu_x + A_{xy}(A^{-1})_{yy}(y - \mu_y)$ and $A_{x|y} = A_{xx} - A_{xy}(A^{-1})_{yy}A_{yx}$.

2 Basics of statistics

2.1 Introduction

The main goal in statistics is to extract information about an underlying random generating mechanism from observations. Let us consider first an example.

Example 2.1

An apple importer receives $N = 10000$ apples and wants to estimate how many apples are bad. The importer takes a random sample of $n = 50$ apples and sees $0 \leq x \leq n$ bad apples. What can be said about $0 \leq w \leq N$, the total number of bad apples?

Ansatz 1: We assume that $w/N \approx x/n$, i.e., the importer anticipates $W(x) = Nx/n$ bad apples. Such a mapping is called a (point) ***estimator***.

Ansatz 2: The importer does not try to predict w but rather an interval $C(x)$ of values which should contain the true value w with high probability, i.e.,

$$\mathbb{P}_w(w \in C(x)) \geq 1 - \alpha$$

for all w, where $\mathbb{P}_w$ is a well chosen probability distribution depending on w. Such an interval is called a ***confidence interval with confidence level*** $1 - \alpha$.

In our example it seems reasonable to assume that the sample n is governed by a hypergeometric distribution, i.e.,

$$\mathbb{P}_w(k) = \binom{w}{k}\binom{N-w}{n-k}\Big/\binom{N}{n}.$$

Ansatz 3: The importer may not be interested in the value w but rather needs to make a decision wether to reject the entire apple shipment due to a too large percentage of bad apples. For example, by agreement, the importer can refuse to pay in case more than 5% of the apples are bad.

$$\begin{aligned} \textbf{null} - \textbf{hypothesis}: &\quad H_o\colon 0 \leq w \leq 500 \\ \textbf{alternative}: &\quad H_1\colon 500 < w \leq 10000 \end{aligned}$$

We need a decision-making criterion of the form

$$x \leq c \Rightarrow \text{ decide for } H_o \qquad \text{and} \qquad x > c \Rightarrow \text{ decide for } H_1,$$

and thus need to find c such that

$$\mathbb{P}_w(x > c) \text{ small for } w \leq 500 \qquad \text{and} \qquad \mathbb{P}_w(x > c) \text{ big for } w > 500.$$

This is called a ***test***.

2.2 Point estimation

One of the basic tasks of statistics is to estimate parameters of a statistical model based on observations.

Definition 2.2: Statistical Model

A ***statistical model*** is a triple $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$, where $\mathfrak{X}$ is the ***sample space***, $\mathcal{L}$ is a σ-algebra on $\mathfrak{X}$, Θ is the ***parameter space*** and $\mathbb{P}_\vartheta$ is a family of probability measures on $(\mathfrak{X}, \mathcal{L})$ with at least two elements.

If $\Theta \subset \mathbb{R}$ the statistical model is called a ***one-parameter model***. In case $\mathfrak{X}$ is discrete, the model is called ***discrete***. It is called ***continuous*** if $\mathfrak{X} \subset \mathbb{R}^n$, $\mathcal{L} = \mathcal{B}(\mathfrak{X})$ and $\mathbb{P}_\vartheta$ has a density ρ_ϑ with respect to the Lebesgue measure. A discrete and continuous model is called a ***standard model***.

For a statistical model $(E, \mathcal{E}, Q_\vartheta \colon \vartheta \in \Theta)$ and $n \geq 2$ we call

$$(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta) = (E^n, \mathcal{E} \otimes \cdots \otimes \mathcal{E}, Q_\vartheta \otimes \cdots \otimes Q_\vartheta \colon \vartheta \in \Theta)$$

the n***-fold product model***. We will often denote by $X_i \colon \mathfrak{X} \to E$ the projection to the i-th coordinate and note that under $\mathbb{P}_\vartheta$ the RVs $X_1, \ldots, X_n$ are independent.

Definition 2.3: Statistics

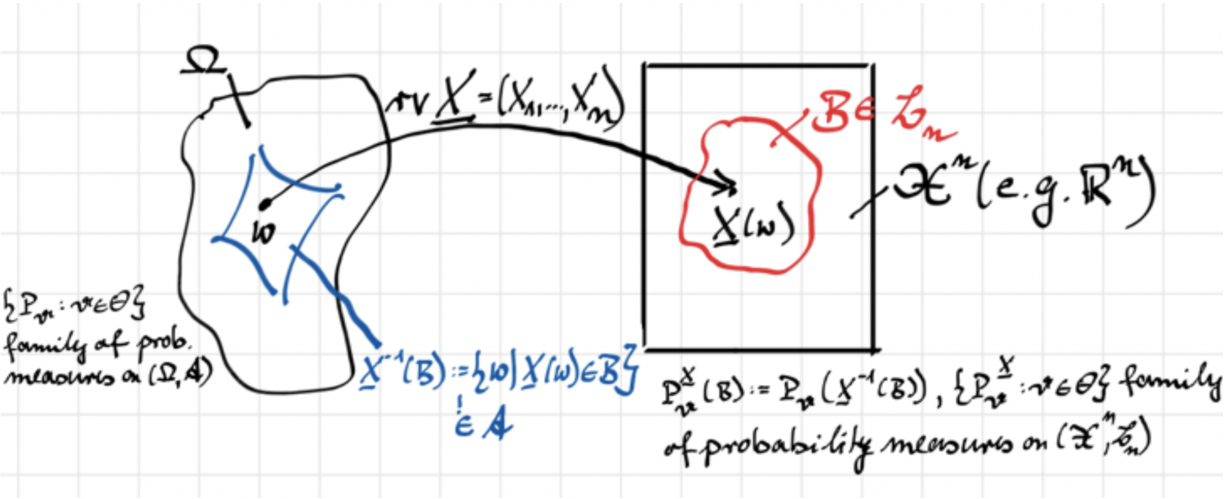
Let $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model and $(\mathcal{S}, \mathfrak{S})$ another measurable space. Then,

i) any RV $S \colon (\mathfrak{X}, \mathcal{L}) \to (\mathcal{S}, \mathfrak{S})$ is called a ***statistic*** and

ii) for $\tau \colon \Theta \to \mathcal{S}$ a mapping that associates any $\vartheta \in \Theta$ to a characteristic $\tau(\vartheta) \in \mathcal{S}$, a statistic $T \colon \mathfrak{X} \to \mathcal{S}$ is an ***estimator*** for τ.

There is a fundamental difference in the interpretation of RVs versus statistics. In probability theory typically the probability space is a known model and a RV is interpreted as a random outcome of an experiment. However, in statistics, the underlying probability space is unknown and a statistics is a well-constructed mapping in order to extract information about the underlying randomness.

An important class of examples is given when $\mathcal{S} = \Theta$ and $T \colon \mathfrak{X} \to \Theta$ aims to estimate ϑ. Such an estimator is often called a ***point estimator*** for ϑ and denotes as $\hat{\vartheta}$. Here is an illustration for the general strategy in case of the n-fold product model.


Ω
rv $\underline{X}=(X_1,\dots,X_n)$
$B \in \mathfrak{B}_n$
$\mathfrak{X}^n$ (e.g. $\mathbb{R}^n$)
$\underline{X}(\omega)$
ω
$\{P_\vartheta : \vartheta \in \Theta\}$ family of prob. measures on $(\Omega, \mathcal{A})$
$\underline{X}^{-1}(B) := \{\omega \mid \underline{X}(\omega) \in B\} \overset{!}{\in} \mathcal{A}$
$P_\vartheta^{\underline{X}}(B) := P_\vartheta(\underline{X}^{-1}(B))$, $\{P_\vartheta^{\underline{X}} : \vartheta \in \Theta\}$ family of probability measures on $(\mathfrak{X}^n, \mathfrak{B}_n)$

Let us give an example.

Example 2.4

Consider a TV show where the host presents random numbers in $[0, \vartheta]$ where $\vartheta > 0$ is only known to the host. Two players can estimate ϑ based on n-many independent samples. Hence, $\mathfrak{X} = [0, \infty)^n$, $\Theta = (0, \infty)$ and $\mathbb{P}_\vartheta$ is the n-fold product of the uniform distribution on $[0, \vartheta]$.

Ansatz 1: Use the ***law of large numbers***, that is

$$\frac{1}{n}\sum_{i=1}^n X_i \approx \mathbb{E}_\vartheta[X_1] = \vartheta^{-1}\int_0^\vartheta x dx = \frac{\vartheta}{2}.$$

In words, the empirical mean converges to the expected value as n tends to infinity. Player 1 sets $T_n = \frac{2}{n}\sum_{i=1}^n X_i$.

Ansatz 2: Use idea that the maximum of the sample should be close to ϑ, i.e., Player 2 sets $\tilde{T}_n = \max\{X_1, \dots, X_n\}$.

Both estimators are ***consistent*** in the sense that for all $\varepsilon > 0$,

$$\mathbb{P}_\vartheta(|T_n - \vartheta| > \varepsilon) \to 0 \qquad \text{and} \qquad \mathbb{P}_\vartheta(|\tilde{T}_n - \vartheta| > \varepsilon) \to 0$$

as n tends to infinity. This is an example for the so-called ***convergence in probability***. Moreover, T_n is ***unbiased*** in the sense that

$$\mathbb{E}_\vartheta[T_n] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\vartheta[X_i] = 2\mathbb{E}_\vartheta[X_1] = \vartheta.$$

For $\tilde{T}_n$ note that by independence $\mathbb{P}_\vartheta(\tilde{T}_n \le c) = \mathbb{P}_\vartheta(X_1 \le c)^n = (c/\vartheta)^n = F_\vartheta(c)$ and hence

$$\mathbb{E}_\vartheta[\tilde{T}_n] = \int_0^\vartheta x F_\vartheta'(x) dx = \frac{n}{\vartheta^n}\int_0^\vartheta x^n dx = \frac{n}{n+1}\vartheta,$$

which tends to ϑ as n tends to infinity. Thus, $\tilde{T}_n$ is not unbiased, but $\tilde{T}_n^* := \frac{n+1}{n}\tilde{T}_n$ is both, unbiased and consistent.

Which estimator fluctuates more? Let us calculate

$$\mathbb{V}_\vartheta(T_n) = \left(\frac{2}{n}\right)^2 \mathbb{V}_\vartheta(\sum_{i=1}^n X_i) = \frac{4}{n}\mathbb{V}_\vartheta(X_1) = \frac{\vartheta^2}{3n} \qquad \text{and}$$

$$\mathbb{V}_\vartheta(\tilde{T}_n) = \mathbb{E}_\vartheta[\tilde{T}_n^2] - \mathbb{E}_\vartheta[\tilde{T}_n]^2 = \frac{n\vartheta^2}{(n+1)^2(n+2)}.$$

However, $\mathbb{V}_\vartheta(\tilde{T}_n^*) = \frac{\vartheta^2}{(n+1)(n+2)}$, which makes $\tilde{T}_n^*$ the least fluctuating estimator.

Definition 2.5: Bias

Let $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model and $\tau \colon \Theta \to \mathbb{R}$ a characteristic. The estimator $T \colon \mathfrak{X} \to \mathbb{R}$ for τ is called ***unbiased*** if

$$\mathbb{E}_\vartheta[T] = \tau(\vartheta).$$

Otherwise, $\mathbb{B}_T(\vartheta) := \mathbb{E}_\vartheta[T] - \tau(\vartheta)$ is called the ***bias*** of T.

Theorem 2.6: Mean and variance estimators in product models

Let $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), Q_\vartheta^{\otimes n} \colon \vartheta \in \Theta)$, $n \geq 2$, the real-valued n-fold product model with existing expected value $m(\vartheta) = \mathbb{E}_\vartheta[X_1]$ and variance $v(\vartheta) = \mathbb{V}_\vartheta(X_1)$. Then,

$$M := \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad V = \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2$$

are unbiased estimators for m and v, respectively.

Let us introduce an important class of estimators.

Definition 2.7: Maximum Likelihood (ML) Estimation

Consider a standard model $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ where $\mathbb{P}_\vartheta$ has a density $x \mapsto \varrho_\vartheta(x)$. Then, $\varrho \colon \mathfrak{X} \times \Theta \to (0, \infty)$ with $\varrho(x, \vartheta) = \varrho_\vartheta(x)$ is called ***likelihood function*** and $\varrho_x := \varrho(x, \cdot)$ the ***likelihood function for*** $x \in \mathfrak{X}$. A estimator $T \colon \mathfrak{X} \to \Theta$ for ϑ is called ***maximum-likelihood estimator*** if $\varrho(x, T(x)) = \max\{\varrho(x, \vartheta) \colon \vartheta \in \Theta\}$. In short $T(x) = \operatorname{argmax} \varrho_x$.

How do we find such T? A useful trick is to consider the ***log-likelihood function*** $\log \varrho(x, \vartheta)$.

Example 2.8: Maximum likelihood in binomial models

Consider the statistical model $(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), \mathcal{B}_{n,\vartheta} \colon \vartheta \in (0,1))$. Then,

$$\frac{d}{d\vartheta}\log \varrho_x(\vartheta) = \frac{d}{d\vartheta}\Big(x \log \vartheta + (n-x)\log(1-\vartheta)\Big) = \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0$$

implies that $\vartheta = x/n$ and hence $T(x) = x/n$.

And here some further examples.

Example 2.9: Maximum likelihood in further models

In the real-valued n-fold product model with

i) $Q_\vartheta = \mathcal{N}\big(\mu, \sigma^2\big)$, i.e., $\vartheta = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$, the ML estimators are given through
$$\widehat{\mu}_n^{ML} = \frac{1}{n}\sum_{i=1}^n X_i, \; \widehat{\sigma}_n^{2,ML} = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n^{ML})^2. \tag{2.2.1}$$

ii) Q_ϑ the exponential distribution with parameter $\lambda \in (0, \infty)$, i.e., X_i possesses the density $x \mapsto \lambda e^{-\lambda x}\mathbb{1}\{x \geq 0\}$, the ML estimator for λ is
$$\widehat{\lambda}_n^{ML} = \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^{-1} \tag{2.2.2}$$

iii) Q_ϑ the uniform distribution on $[0, \vartheta]$, $\vartheta \in (0, \infty)$ (see Example 2.4), the ML estimator for ϑ is
$$\widehat{\vartheta}_n^{ML} = \max\{X_1, ..., X_n\}. \tag{2.2.3}$$

Under the great variety of possible estimators, which is the best one? It depends on the requirement we impose on the estimator. Here is a characterization of a best estimator via minimal variance.

Definition 2.10: Best estimator

Let $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model. An unbiased estimator T for the real-valued characteristic $\tau(\vartheta)$ is called ***variance minimizing*** or ***best estimator***, if for all other unbiased estimators S we have $\mathbb{V}_\vartheta(T) \leq \mathbb{V}_\vartheta(S)$ for all $\vartheta \in \Theta$.

We note that a large class of statistical models (***exponential models***) possesses a best estimator.

2.3 Confidence intervals

As introduced in the initial example, another way to derive meaningful statistical statements is to estimate sets rather than points.

Definition 2.11: Confidence sets

Let $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model, $\mathcal{S}$ a set and $\tau \colon \Theta \to \mathcal{S}$ a characteristic for the parameter, and $0 < \alpha < 1$. A mapping $C \colon \mathfrak{X} \to \mathcal{P}(\mathcal{S})$ is called a ***confidence***

region for τ ***of level*** $1-\alpha$ if

$$\mathbb{P}_\vartheta\Big\{C \ni \tau(\vartheta)\Big\} \geq 1-\alpha \quad \forall \vartheta \in \Theta. \tag{2.3.1}$$

Of course it is wanted to achieve (2.3.1) for confidence sets C as small as possible. $C = \mathcal{S}$ obviously fulfills (2.3.1) for every $\alpha \in (0,1)$, but is of no information to the investigator.

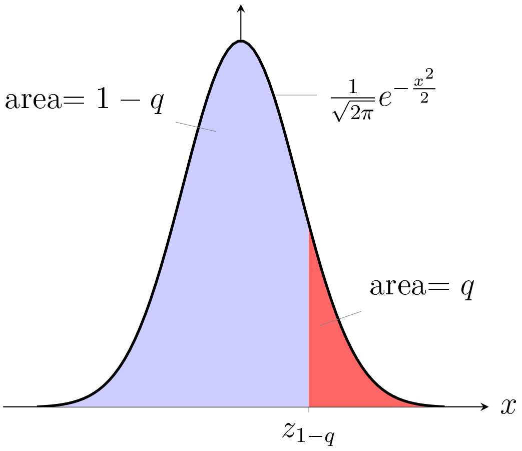
Example 2.12

Let $X_1, \ldots, X_n$ be iid random variables distributed according to $\mathcal{N}\big(\mu, \sigma^2\big)$. The mean μ is unknown.

i) Assume that we know the value of σ^2. Then,

$$\left[\overline{X}_n - \frac{\sigma \cdot z_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma \cdot z_{1-\alpha/2}}{\sqrt{n}}\right], \text{ for short } \overline{X}_n \pm \frac{\sigma \cdot z_{1-\alpha/2}}{\sqrt{n}}, \tag{2.3.2}$$

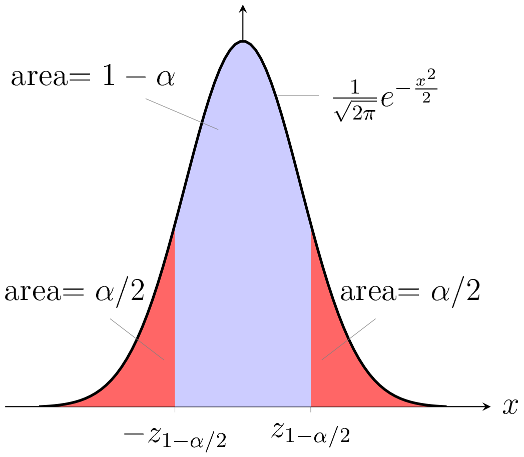
is a $(1-\alpha)$-confidence interval for μ, where $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. The value z_{1-q}, $q \in (0,1)$, is defined as is described in the following figure:



z_{1-q} is denoted the $(1-q)$***-quantile of the standard normal distribution***. To see that the interval (2.3.2) has confidence level $(1-\alpha)$ observe:

$$\mathbb{P}_\mu\left\{\mu \in \overline{X}_n \pm \frac{\sigma \cdot z_{1-\alpha/2}}{\sqrt{n}}\right\} \overset{!}{=} \mathbb{P}_\mu\left\{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \in \Big[-z_{1-\alpha/2}, z_{1-\alpha/2}\Big]\right\} = 1-\alpha,$$

since $\sqrt{n}\dfrac{\overline{X}_n - \mu}{\sigma}$ (in case that μ is the true expectation of the X_i) possesses a standard normal distribution.



Some values of z_{1-q} are:

$z_{0.95}$	$z_{0.975}$	$z_{0.995}$
1.645	1.960	2.576

ii) In case that σ^2 is also unknown

$$\overline{X}_n \pm \frac{S \cdot t_{n-1;1-\alpha/2}}{\sqrt{n}}, \tag{2.3.3}$$

with $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, is an estimator of σ^2, and $t_{n-1;1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the so-called ***Student t-distribution*** with $(n-1)$ degrees of freedom, defines a $(1-\alpha)$-confidence interval for μ.

iii) If we drop the assumption of a normal distribution and only assume that $X_1, \ldots$ are iid random variables with unknown mean μ and unknown variance σ^2, then

$$\overline{X}_n \pm \frac{S \cdot z_{1-\alpha/2}}{\sqrt{n}} \tag{2.3.4}$$

still is a confidence interval for μ but now with asymptotical level $(1-\alpha)$, only. More precisely we obtain from the Lindeberg-Lévy CLT:

$$\mathbb{P}_{\mu,\sigma}\left\{\mu \in \overline{X}_n \pm \frac{S \cdot z_{1-\alpha/2}}{\sqrt{n}}\right\} \xrightarrow{n\to\infty} 1-\alpha \quad \forall \mu \in \mathbb{R}\ \forall \sigma \in (0,\infty). \tag{2.3.5}$$

2.4 Statistical tests

Consider a statistical model $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$. For a subset $\Theta_o \subset \Theta$ we are interested in the question whether or not the true underlying parameter belongs to this set Θ_o.

Example 2.13: Statistical tests

Assume $X_1, \ldots, X_n$ iid according to Bernoulli(p), $p \in (0,1)$. The X_i for example describe whether or not a dice shows the figure 6. One may be interested to check whether the dice is fair, more precisely to see whether $\mathbb{P}\{X_i = 1\} = 1/6$ is true or not. In this case we choose $\Theta_o = \{1/6\} \subset \Theta = (0,1)$. A possible check procedure or statistical test would likely look as follows:

$$\text{Reject } \Theta_o \text{ if } \overline{X}_n \text{ (the relative number of successes)}$$
$$\text{is too far away from } 1/6 \text{, e.g. if } \left|\overline{X}_n - 1/6\right| > c \text{ for some critical value } c.$$

An important question is how to choose c.

A further example is to develop a statistical test (a check procedure) for the problem whether a specific type of vehicle complies with the prescribed limit value of a maximum NO_x-emission of 80 milligram per kilometer driven. For this purpose $n = 9$ test drives are carried out with the corresponding vehicle type and the respective NO_x-emissions are measured. A possible precise stochastic modeling is as follows: The emissions are assumed to vary randomly and therefore they are described as random variables $X_1, \dots, X_n$. The design of the test drives should be in a way that an iid assumption is justifiable. The legal limit of 80 mg/km translates to $\mu = \mathbb{E}[X_i] \leq 80$. We probably want to reject the hypothesis $H_o : \mu \leq 80$ if the observed NO_x-emission average $\overline{X}_n$ is greater than $80 + c$ for some positive tolerance value. The key question is how to choose c. We could further restrict our model by assuming that $X_1, \dots, X_n$ are iid according to $\mathcal{N}\big(\mu, \sigma^2\big)$ with unknown parameter $\vartheta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty) =: \Theta$. Then the hypothesis $H_o : \mu \leq 80$ means that $\vartheta = (\mu, \sigma) \in \Theta_o := (-\infty, 80] \times (0, \infty) \subset \Theta$.

We now given a formal definition of a statistical test.

Definition 2.14: Statistical Test

We assume a statistical model $(\mathfrak{X}, \mathcal{L}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$. Θ separates in the two disjoint sets Θ_o, the ***null-hypothesis*** and $\Theta_1 = \Theta \setminus \Theta_o$, the ***alternative***. Every statistics $\varphi \colon \mathfrak{X} \to [0, 1]$ is called a ***test*** of Θ_o against Θ_1. A test is called ***non-random*** if $\varphi(x) \in \{0, 1\}$ for all $x \in \mathfrak{X}$, otherwise it is called ***randomized***. In the first case, $\{x \in \mathfrak{X} \colon \varphi(x) = 1\}$ is called ***critical region*** of the test φ. Further, $\sup_{\vartheta \in \Theta_o} \mathbb{E}_\vartheta[\varphi]$ is called ***effective level*** of φ and if $\sup_{\vartheta \in \Theta_o} \mathbb{E}_\vartheta[\varphi] \leq \alpha$, then φ is a ***test of level*** α. Finally, the function $\beta_\varphi \colon \Theta \to [0, 1]$, $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi]$ is called ***power function*** of φ at ϑ.

The value $\varphi \in [0, 1]$ is interpreted as the probability of deciding for the alternative in the presence of the observation x. $\beta_\varphi(\vartheta)$ is the expected rejection rate of the null-hypothesis if the observations were generated under the probability law $\mathbb{P}_\vartheta$. Of course it is wanted to have $\beta_\varphi(\vartheta)$ close to zero for $\vartheta \in \Theta_o$ and to have $\beta_\varphi(\vartheta)$ close to one for $\vartheta \in \Theta_1$.

The error behavior of a statistical test φ is quite clear and can be illustrated as follows

decision for actually correct	H_o	H_1 (rejection of H_o)
H_o	correct decision	type I error
H_1	type II error	correct decision

(2.4.1)

The quantity $\beta_\varphi(\vartheta)$ for $\vartheta \in \Theta_o$ is called ***error probability of first kind*** or ***type I error probability***. For a ***level*** α ***test*** the type I error probability is bounded by the predefined value α. The quantity $1 - \beta_\varphi(\vartheta)$ for $\vartheta \in \Theta_1$ is called ***error probability of second kind*** or ***type II error probability***. The rationale behind $\sup_{\vartheta\in\Theta} \beta_\varphi(\vartheta) \leq \alpha$ is that the type I error is considered as more severe compared to the type II error.

Remark 2.15

It is important to keep in mind that for level α tests the type I error is under control because we know that the type I error probability is below a value α which can be prescribed by the investigator. This is by far not true for type II errors. Typically type II error probabilities can increase up to the value of $1-\alpha$. Recalling the error table this means that only a ***rejection of the hypothesis*** H_o is a reliable decision. Either the rejection of H_o is right or, if it is wrong, this happens only with a probability less than or equal to α. This matter of fact is very important to have in mind when designing a statistical test for a situation of interest. One typically aims at constructing a test for which the power $\beta_\varphi(\vartheta)$ increases to 1 as quickly as possible $\forall \theta \in \Theta_1$ while complying with the desired size condition.

In the following we go through a small number of basic but important statistical tests.

Theorem 2.16: One-Sided Gauss-Tests

Assume the n-fold product model with $X_1, \dots, X_n$ are iid according to $\mathcal{N}\big(\mu, \sigma_0^2\big)$, where $\mu \in \mathbb{R} = \Theta$ denotes the parameter of the model and $\sigma_0^2 \in (0, \infty)$ is known. We consider the testing problem

$$H_o : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0 \quad (\mu_0 \text{ given value}). \tag{2.4.2}$$

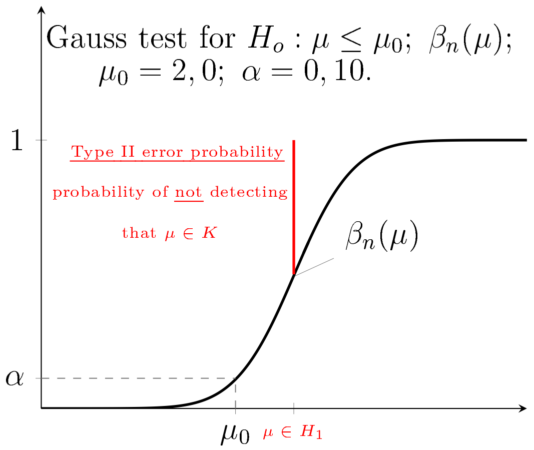
The following ***one-sided Gauss test*** (for H_o versus H_1) has level α:

$$\varphi(X_1, \dots, X_n) = \begin{cases} 1 \text{ (rejection of } H_o), \\ \\ 0 \text{ (no rejection of } H_o), \end{cases} \quad \overline{X}_n \begin{matrix} > \\ \\ \leq \end{matrix} \quad \mu_0 + \frac{z_{1-\alpha} \cdot \sigma_0}{\sqrt{n}} \quad . \tag{2.4.3}$$

The power function of φ reads as follows

$$\beta_\varphi(\mu) = \mathbb{P}_\mu\Big\{\overline{X}_n > \mu_0 + \frac{z_{1-\alpha}\cdot\sigma_0}{\sqrt{n}}\Big\} = \mathbb{P}_\mu\Bigg\{\underbrace{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma_0}}_{\sim\mathcal{N}(0,1)} > \sqrt{n}\frac{\mu_0-\mu}{\sigma_0} + z_{1-\alpha}\Bigg\}$$

$$= \int\limits_{z_{1-\alpha}+\sqrt{n}\frac{\mu_0-\mu}{\sigma_0}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{2.4.4}$$



From the illustration the following can be seen: φ has level $\alpha : \beta_\varphi(\mu) \le \alpha \ \forall \mu \le \mu_0$. The type II error probability can be as large as $1-\alpha$. The type II error probability is smaller the larger μ is.

Remark 2.17

In the same underlying situation as in Theorem 2.16 but now for the testing problem

$$H_o : \mu \ge \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0 \quad (\mu_0 \text{ given value}) \tag{2.4.5}$$

the following corresponding one-sided Gauss test also has level α:

$$\varphi(X_1,\dots,X_n) = \begin{cases} 1 \text{ (rejection of } H_o), & \overline{X}_n < \mu_0 - \frac{z_{1-\alpha}\cdot\sigma_0}{\sqrt{n}} \\ 0 \text{ (no rejection of } H_o), & \overline{X}_n \ge \mu_0 - \frac{z_{1-\alpha}\cdot\sigma_0}{\sqrt{n}} \end{cases}. \tag{2.4.6}$$

The power function of φ is

$$\beta_\varphi(\mu) = \mathbb{E}_\mu[\varphi] = \mathbb{P}_\mu\Big\{\overline{X}_n < \mu_0 - \frac{z_{1-\alpha}\cdot\sigma_0}{\sqrt{n}}\Big\} = \int\limits_{-\infty}^{\sqrt{n}\frac{\mu_0-\mu}{\sigma_0}-z_{1-\alpha}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

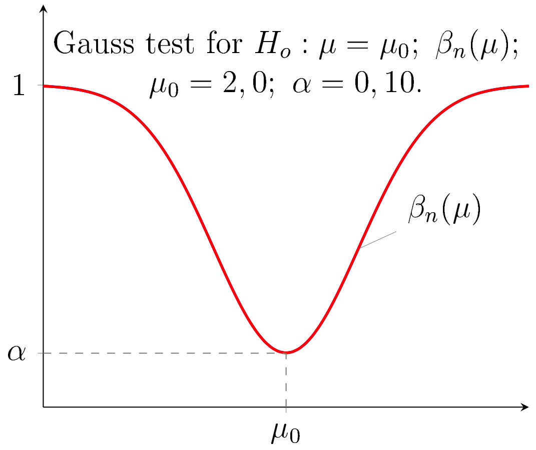
$$= \Phi\Big(\sqrt{n}\frac{\mu_0-\mu}{\sigma_0} - z_{1-\alpha}\Big) \tag{2.4.7}$$

with $\Phi(u) := \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ the (cumulative) distribution function of $\mathcal{N}\big(0,1\big)$.

Theorem 2.18: Two-Sided Gauss Test

We consider the same statistical model as in Theorem 2.16. But the testing problem now reads for a fixed value $\mu_0 \in \mathbb{R}$ as follows

$$H_o : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \tag{2.4.8}$$



This, in contrast to $H_o : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, is a two-sided testing problem. The level α corresponding ***two-sided Gauss test*** is

$$\varphi(X_1, \dots, X_n) = \begin{cases} 1, & > \\ & \sqrt{n}\dfrac{|\overline{X}_n - \mu_0|}{\sigma_0} \qquad z_{1-\alpha/2}. \\ 0, & \leq \end{cases} \tag{2.4.9}$$

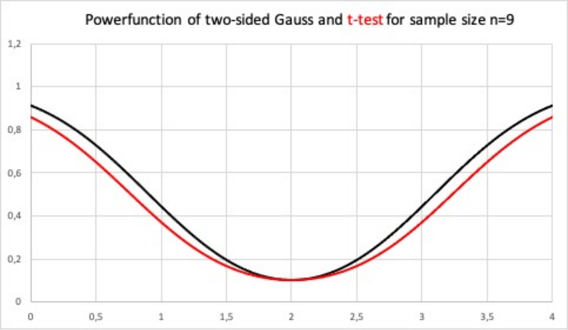
The power function of the two-sided Gauss test reads as follows

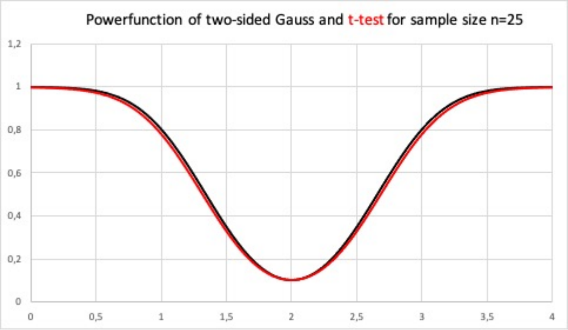
$$\beta_\varphi(\mu) = \mathbb{P}_\mu \left\{ \sqrt{n}\frac{|\overline{X}_n - \mu_0|}{\sigma_0} > z_{1-\alpha/2} \right\}. \tag{2.4.10}$$

Theorem 2.19: Student's t-Tests

Consider iid random variables $X_1, \dots, X_n$ distributed according to a normal (Gaussian) distribution $\mathcal{N}\big(\mu, \sigma^2\big)$, where both μ and σ^2 are unknown. The parameter of the model then is $\vartheta = (\mu, \sigma)$. For the testing problems $H_o : \mu \leq \mu_0$, $H_o : \mu \geq \mu_0$ and $H_o : \mu = \mu_0$ the corresponding Gauss tests (2.4.8), (2.4.11) and (2.4.14) can be modified by replacing σ^2 (which in the model considered herein is unknown) by $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$ and $z_{1-\alpha}$ respectively $z_{1-\alpha/2}$ through $t_{n-1;1-\alpha}$ respectively $t_{n-1;1-\alpha/2}$ (which are quantiles of Student t-distribution with $(n-1)$ degrees of freedom), such that the resulting so-called ***t-tests*** possess level α and are consistent.

In the final illustration we compare the power function of the two-sided Gauss test with that of the two-sided t-test for two different sample sizes $n = 9$ and $n = 25$.





It can be seen that the t-test has less power compared to the Gauss test, which is clear because the t-test faces the situation that the variance is not known and therefore has to be estimated. This fact adds some additional uncertainty to the test decision of the t-test which manifests itself in a lower power. However, for increasing sample size the lost is not really substantial.

Example 2.20

We return to Example 2.2.1 where are discussed testing problems for NO$_x$-emissions of vehicles. Assume that the following NO$_x$-emissions have been observed during $n = 9$ carried out test drives (in mg per km driven)

$$78, 81, 79, 85, 68, 72, 75, 70, 80.$$

We obtain from the observations: $\overline{X}_n = 76.44$ and $S = 5.59$. We assume that the measurements are (at least approximately) normally distributed. The mean $\mu = \mathbb{E}[X_i]$ and the variance $\sigma^2 = \text{Var}(X_i)$ both are unknown. This implies that a t-test is appropriate. But how to choose the hypothesis H_o?

Assume that we work for the vehicle manufacturer and we intend to confirm that the specific make and model complies with the legal limit of 80 mg/km. That is, we would like to have a rejection of the hypothesis $H_o : \mu \geq \mu_0 = 80$. We choose the level $\alpha = 5\%$ Since $t_{8;0.95} = 1.860$ we get

$$76.44 = \overline{X}_n < \mu_0 - \frac{t_{n-1;1-\alpha} \cdot S}{\sqrt{n}} = 80 - \frac{1.86 \cdot 5.59}{3} = 76.53.$$

We actually get the desired rejection of the hypothesis $H_o : \mu \geq 80$, which confirms that the vehicles comply with the legal limit. Having the error table in mind it is seen that in case of a rejection of the hypothesis either a correct decision of a type I error with an error probability less than or equal to $\alpha = 5\%$ is present. So we are quite safe with our decision!

Now imagine that we work for the state supervisory authority. In this case we might be interested to find evidence that the specific vehicle does **not** comply with the legal limit of 80 mg/km NO$_x$-emission. In such a situation we want to reject the hypothesis $H_o : \mu \leq 80$. For the same data and the same level $\alpha = 0.05$ we could not reject the hypothesis $H_o : \mu \leq 80$ because

$$76.44 = \overline{X}_n \not> \mu_0 + \frac{t_{n-1;1-\alpha} \cdot S}{\sqrt{n}} = 83.47.$$

This decision (no rejection of H_o) is either correct or a type II error occurs. Since type II error probabilities are not under control the decision of not rejecting $H_o : \mu \leq 80$ is subject to a high degree of uncertainty.

2.5 Linear regression

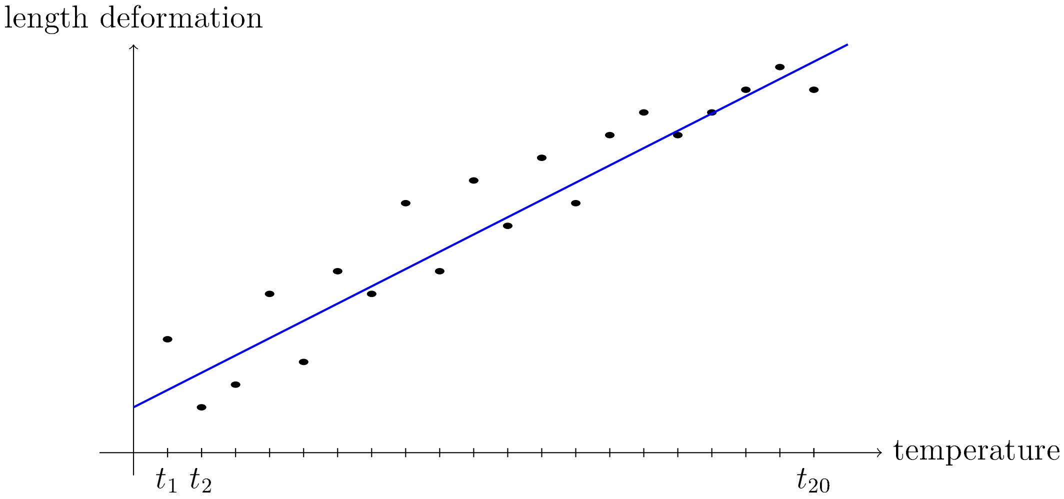
Let us start with an example.

Example 2.21: Linear regression

Consider a metal bar for which we want to measure the length deformation at temperatures $t = (t_1, \dots, t_n)$. We assume that the length is given by

$$X_k = \gamma_0 + \gamma_1 t_k + \sqrt{v}\xi_k \qquad \forall 1 \leq k \leq n,$$

where $\gamma = (\gamma_0, \gamma_1)$ and v are unknown real-valued coefficients, $X = (X_1, \dots, X_n)$ is a vector of measurements and $\xi = (\xi_1, \dots, \xi_n)$ a vector of iid random measurement errors with $\mathbb{E}[\xi_1] = 0$ and $\mathbb{V}(\xi_1) = 1$.



Let $P_{\gamma,v}$ denote the distribution of the random vector

$$\gamma_0 \mathbf{1} + \gamma_1 t + \sqrt{v}\xi$$

and consider the statistical model $\Big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\gamma,v} : \gamma \in \mathbb{R}^2, v \in (0, \infty)\Big)$.

We want to estimate γ using the ***principle of least squares***. That is, we set $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1)$ such that the mean quadratic error

$$F_\gamma := \frac{1}{n} \sum_{k=1}^{n} \Big(X_k - (\gamma_0 + \gamma_1 t_k)\Big)^2$$

is minimal at $\hat{\gamma} = \gamma$.

In order to find the minimum we calculate

$$\frac{d}{d\gamma_0}F_\gamma = -\frac{2}{n}\sum_{k=1}^{n}(X_k - \gamma_0 - \gamma_1 t_k) = 0 = \frac{d}{d\gamma_1}F_\gamma = -\frac{2}{n}\sum_{k=1}^{n} t_k(X_k - \gamma_0 - \gamma_1 t_k)$$

which implies the so-called ***normal equations***

$$M(X) = \gamma_0 + \gamma_1 M(t) \qquad \text{and} \qquad \frac{1}{n}\sum_{k=1}^{n} t_k X_k = \gamma_0 M(t) + \gamma_1 \frac{1}{n}\sum_{k=1}^{n} t_k^2,$$

where $M(X) = \frac{1}{n}\sum_{k=1}^{n} X_k$. Combining these equations, we see that

$$\hat{\gamma}_0 = M(X) - \frac{M(t)}{V(t)}c(t,X) \qquad \text{and} \qquad \hat{\gamma}_1 = \frac{c(t,X)}{V(t)},$$

where $V(t) = \frac{1}{n}\sum_{k=1}^{n} t_k^2 - M(t)^2$ the variance of t and $c(t,X) = \frac{1}{n}\sum_{k=1}^{n} t_k X_k - M(t)M(X)$ the covariance of t and X. By convexity, $\hat{\gamma}$ is the unique ***least-square estimator*** for γ. One can check that $\hat{\gamma}$ is unbiased.

The form of the estimator also has a geometric interpretation. Note that we assume that $X \in \operatorname{span}(\mathbf{1}, t, \xi)$ however usually $X \notin \operatorname{span}(\mathbf{1}, t)$. The least-square principle says that $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 t \in \operatorname{span}(\mathbf{1}, t)$ with

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \hat{X}_k)^2 = \min\left\{\frac{1}{n}\sum_{k=1}^{n}(X_k - y_k)^2 \colon y \in \operatorname{span}(\mathbf{1}, t)\right\}.$$

So $\hat{X}$ is the minimizer. But $\hat{X}$ is also the projection of X to the linear subspace $L(\mathbf{1}, t) := \{\alpha\mathbf{1} + \beta t \colon \alpha, \beta \in \mathbb{R})$ since, by the normal equations,

$$\langle X - \hat{X}, \mathbf{1}\rangle = nM(X) - n\hat{\gamma}_0 - n\hat{\gamma}_1 M(t) = 0 = \langle X - \hat{X}, t\rangle.$$

Definition 2.22: Linear model

Let $s, n \in \mathbb{N}$ with $s < n$. A ***linear model*** with n real-valued measurements, (unknown) s-dimensional parameters $\gamma = (\gamma_0, \dots, \gamma_s)^T \in \mathbb{R}^s$ and (unknown) scalar parameter $v > 0$, consists of a real-valued $n \times s$ matrix A with full rank s, the ***design matrix*** and a real-valued random vector $\xi = (\xi_1, \dots, \xi_n)^T$ of n standardized RVs ξ_k, the ***error***.

The n-.dimensional observation $X = (X_1, \dots, X_n)^T$ is given via the linear equation

$$X = A\gamma + \sqrt{v}\xi.$$

The suitable statistical model is $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\gamma,v} \colon \gamma \in \mathbb{R}^s, v \in (0,\infty)\big)$, where $P_{\gamma,v}$ is the distribution of $A\gamma + \sqrt{v}\xi$.

Note that, since the error is centered, $\mathbb{E}[X] = A\gamma$.

Example 2.23: Gaussian product model

We set $s = 1$, $A = \mathbf{1}$, $\gamma = m \in \mathbb{R}$ and $\xi \sim \mathcal{N}\big(\mathbf{0}, E\big)$, where E is the n-dimensional unit matrix. Then,

$$X = \mathbf{m} + \sqrt{v}\xi \sim \mathcal{N}\Big(\mathbf{m}, vE\Big),$$

the n-dimensional product normal distribution with mean m and variance v. In case of the metal-bar example, $s = 2$, $A = \mathbf{1}t$ and $X = A\gamma + \sqrt{v}\xi$.

Let us finally remark that on a general level, least-square estimators are of the form

$$\hat{\gamma} = (A^TA)^{-1}A^TX.$$

Indeed, as mentioned above, $\hat{X} = A\hat{\gamma}$ is the projection of X to $\mathrm{span}(A)$ and thus,

$$A^T(X - \hat{X}) = A^TX - A^TA(A^TA)^{-1}A^TX = 0,$$

as desired. Note that $\hat{\gamma}$ is unbiased since $\mathbb{E}[\hat{\gamma}] = (A^TA)^{-1}A^T\mathbb{E}[X] = (A^TA)^{-1}A^TA\gamma = \gamma$.