

Ramp-Up Mathematics: Stochastics

Basics of probability theory and statistics

Benedikt Jahnel

Technische Universität Braunschweig

Lecture notes, 23 October 2026

Abstract

These notes accompany a short course of four lectures over two weeks. Part 1 (Lectures 1 and 2) introduces the basic model of probability theory, random variables and their characteristics. Part 2 (Lectures 3 and 4) presents the statistical perspective: statistical models, estimators, tests and linear regression.

Contents

1	Basics of probability theory	1
1.1	The basic model	2
1.2	Independence and conditional probabilities	4
1.3	Product spaces	5
1.4	Random variables and their distributions	6
1.5	Expectation and variance	7
1.6	Several random variables	9
1.7	Conditional expectations	11
2	Basics of statistics	11
2.1	A motivating example	12
2.2	Statistical models, statistics and estimators	13
2.3	Mean, variance and maximum-likelihood estimators	15
2.4	Statistical tests	17
2.5	Linear regression	18

Further reading

The presentation follows [Geo13], which covers both parts of the course at a comparable level; a German-language alternative is [Kre05]. For probability theory, [Fel68] and [GS20] offer many examples, while [Kle20] and [Dur19] develop the measure-theoretic foundations. For statistics, we recommend [CB02] and [Was04], and, for more advanced topics, [LC98] on estimation, [LR22] on testing and [SL03] on linear regression.

1 Basics of probability theory

The goal of probability theory is the modelling and analysis of systems that include (assumed) randomness.

1.1 The basic model

Definition 1.1 (Probability space). A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ with the following components.

- (i) The *sample space* $\Omega \neq \emptyset$. Its elements $\omega \in \Omega$ are called *outcomes*, *realisations* or *simple events*.
- (ii) The σ -algebra $\mathcal{F} \subset \mathcal{P}(\Omega) := \{A : A \subset \Omega\}$, a subset of the power set of Ω . Its elements $A \in \mathcal{F}$ are called *events*, and $\mathcal{F}$ has the following properties:
 - 1) $\Omega \in \mathcal{F}$;
 - 2) if $A \in \mathcal{F}$, then $A^c := \Omega \setminus A \in \mathcal{F}$;
 - 3) if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
- (iii) The *probability measure* $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$, which assigns the probability $\mathbb{P}(A)$ to the event $A \in \mathcal{F}$. It has the following properties:
 - 1) $\mathbb{P}(\Omega) = 1$;
 - 2) σ -additivity: for all $A_1, A_2, \dots \in \mathcal{F}$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Remark 1.2 (Properties of σ -algebras and probability measures). Let $A, B \in \mathcal{F}$.

- (i) The following sets are again events:

- 1) $A \cup B \in \mathcal{F}$, “ A or B has occurred”;
- 2) $A \cap B \in \mathcal{F}$, “ A and B have occurred”;
- 3) $B \setminus A \in \mathcal{F}$, “ B has occurred, but not A ”;
- 4) $A^c \in \mathcal{F}$, “ A has not occurred”.

- (ii) Probability measures have the following properties:

- 1) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
- 2) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
- 3) *monotonicity*: if $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Sketch of proof. For (i), finite unions are covered by choosing $A_1 = A$, $A_2 = B$ and $A_i = \emptyset = \Omega^c$ for $i \geq 3$. De Morgan's law gives $A \cap B = (A^c \cup B^c)^c$, and $B \setminus A = B \cap A^c$. For (ii), σ -additivity applied to $\Omega = A \cup A^c$ yields 1). The set $A \cup B$ is the disjoint union of A and $B \setminus A$, and B is the disjoint union of $B \setminus A$ and $A \cap B$; combining both decompositions gives 2). If $A \subset B$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$, which is 3). $\square$

Example 1.3 (Discrete models). If Ω is at most countably infinite (for example $\Omega = \mathbb{N}_0$), then we can take $\mathcal{F} = \mathcal{P}(\Omega)$. Every probability measure is then determined by its (*discrete*) *density* $\rho : \Omega \rightarrow [0, 1]$ with $\sum_{k \in \Omega} \rho(k) = 1$ via

$$\mathbb{P}(A) = \sum_{k \in A} \rho(k) \quad \text{for all } A \in \mathcal{F}.$$

- (i) Let $\Omega = \{0, 1, \dots, n\}$, so that $|\Omega| = n + 1$.

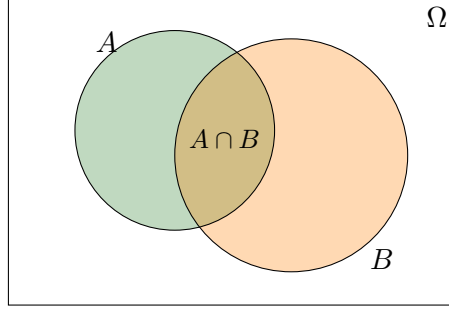


Figure 1: Two events $A, B \in \mathcal{F}$ in the sample space Ω . In $\mathbb{P}(A) + \mathbb{P}(B)$ the overlap $A \cap B$ is counted twice, which explains the correction term in [Remark 1.2](#) (ii) 2).

- 1) The *Laplace experiment* or *discrete uniform distribution* has density $\rho(k) = 1/|\Omega|$ for all $k \in \Omega$, and hence

$$\mathbb{P}(A) = \sum_{k \in A} \rho(k) = \frac{|A|}{|\Omega|} \quad \text{for all } A \in \mathcal{F}.$$

- 2) The *binomial distribution* with success parameter $p \in [0, 1]$ has density

$$\text{Bin}_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for all } k \in \Omega.$$

- (ii) Let $\Omega = \mathbb{N}_0$.

- 1) The *Poisson distribution* with parameter $\lambda \in (0, \infty)$ has density

$$\text{Poi}_\lambda(k) := e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for all } k \in \mathbb{N}_0.$$

- 2) The *geometric distribution* with parameter $p \in (0, 1]$ has density

$$\text{Geo}_p(k) := (1-p)^k p \quad \text{for all } k \in \mathbb{N}_0.$$

It describes the number of failures before the first success in independent trials with success probability p .

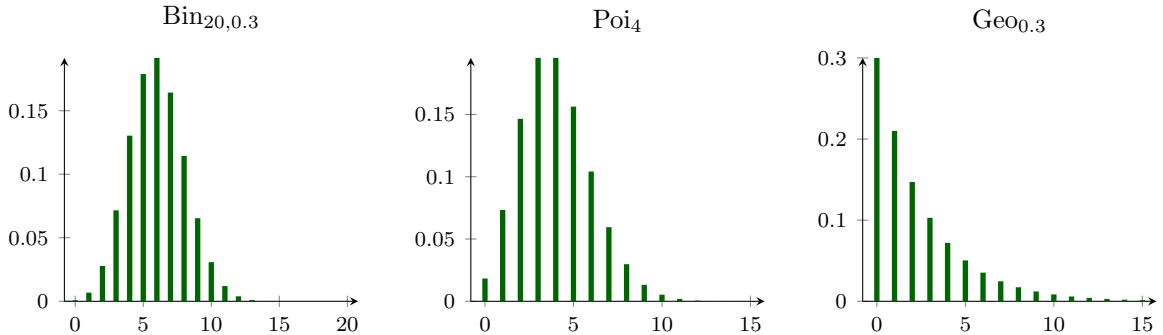


Figure 2: Discrete densities from [Example 1.3](#) for illustrative parameter values. The binomial and the Poisson density are concentrated around $np = 6$ and $\lambda = 4$, respectively, while the geometric density decreases geometrically from $k = 0$.

Example 1.4 (Models with densities). If $\Omega = \mathbb{R}$, we usually work with $\mathcal{F} = \mathcal{B}(\mathbb{R})$, the *Borel σ -algebra* on $\mathbb{R}$. It is the smallest σ -algebra containing all intervals $(a, b]$ with $a < b \in \mathbb{R}$; see [Kle20] for its construction and properties. A large class of models can be defined as

$$\mathbb{P}(A) = \int_A f(x) dx \quad \text{for all } A \in \mathcal{B}(\mathbb{R}),$$

where the (*Lebesgue*) *density* $f: \mathbb{R} \rightarrow [0, \infty)$ satisfies $\int_{\mathbb{R}} f(x) dx = 1$.

- 1) The *uniform distribution* on $[a, b]$, with $a < b \in \mathbb{R}$, has density

$$f(x) = \frac{1}{b-a} \mathbf{1}_{\{a \leq x \leq b\}}.$$

- 2) The *exponential distribution* with parameter $\lambda > 0$ has density

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

- 3) The *normal distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has density

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

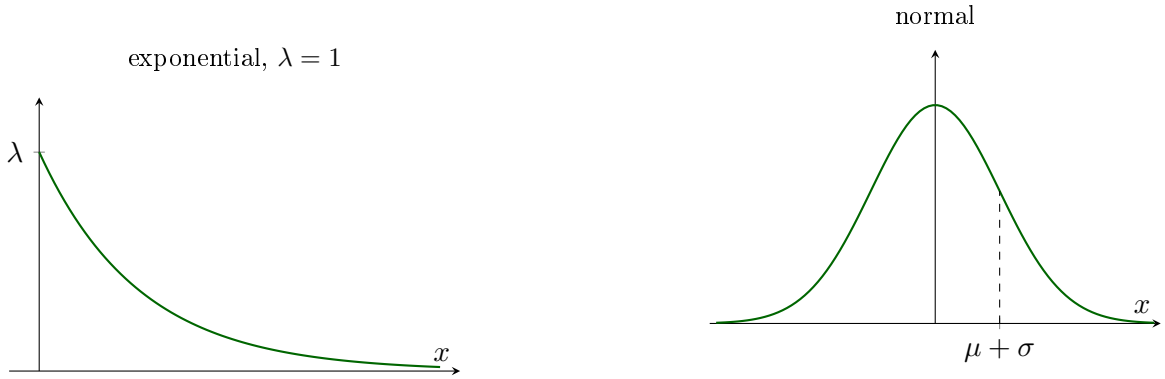


Figure 3: Densities of the exponential and of the normal distribution from Example 1.4.

1.2 Independence and conditional probabilities

Definition 1.5 (Independence). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

- (i) Two events $A, B \in \mathcal{F}$ are called *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- (ii) A family $(A_i)_{i \in I}$ of events is called *independent* if, for all $J \subset I$ with $|J| < \infty$,

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

It is called *pairwise independent* if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j \in I$.

Example 1.6 (Two dice). Consider two dice rolls, modelled by $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$ and $\rho((\ell, k)) = 1/36$ for all $(\ell, k) \in \Omega$. The events

$$A = \{(1, k) : k = 1, \dots, 6\} \quad \text{and} \quad B = \{(\ell, 2) : \ell = 1, \dots, 6\}$$

(“the first die shows 1” and “the second die shows 2”) are independent, since

$$\mathbb{P}(A \cap B) = \rho((1, 2)) = \frac{1}{36} \quad \text{and} \quad \mathbb{P}(A)\mathbb{P}(B) = \sum_{k=1}^6 \rho((1, k)) \sum_{\ell=1}^6 \rho((\ell, 2)) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

		second die k						
		1	2	3	4	5	6	
first die ℓ	1							A
	2							
	3							
	4							
	5							
	6							
		B						

Figure 4: The 36 equally likely outcomes of [Example 1.6](#). The row A and the column B each contain 6 outcomes and meet in exactly one, so $\mathbb{P}(A \cap B) = 1/36 = (6/36) \cdot (6/36)$.

Definition 1.7 (Conditional probability). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then

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

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

Remark 1.8 (Properties of conditional probabilities). Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$.

- 1) The map $\mathbb{P}(\cdot|B)$ is again a probability measure on $(\Omega, \mathcal{F})$.
- 2) If A and B are independent, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.
- 3) *Law of total probability.* Let $B_1, B_2, \dots \in \mathcal{F}$ with $B_i \cap B_j = \emptyset$ for all $i \neq j$, $\bigcup_{i \geq 1} B_i = \Omega$ and $\mathbb{P}(B_i) > 0$ for all $i \geq 1$. Then, for all $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i) \mathbb{P}(B_i). \quad (1.1)$$

- 4) *Bayes' formula.* If, in the setting of 3), also $\mathbb{P}(A) > 0$, then, for all $j \geq 1$,

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j) \mathbb{P}(B_j)}{\sum_{i=1}^{\infty} \mathbb{P}(A|B_i) \mathbb{P}(B_i)}.$$

Proof. Property 1) follows from $\mathbb{P}(\Omega \cap B) = \mathbb{P}(B)$ and the σ -additivity of $\mathbb{P}$. Property 2) is immediate from [Definition 1.5](#). For 3), the events $A \cap B_i$ are pairwise disjoint with union A , so σ -additivity gives the first identity, and [Definition 1.7](#) the second. For 4), we write $\mathbb{P}(B_j|A) = \mathbb{P}(A \cap B_j)/\mathbb{P}(A)$ and use [Definition 1.7](#) in the numerator and (1.1) in the denominator. $\square$

1.3 Product spaces

Definition 1.9 (Product spaces). Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two probability spaces. We define the *product probability space* as $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$, where:

- 1) the *product σ -algebra* $\mathcal{F}_1 \otimes \mathcal{F}_2$ is the smallest σ -algebra on $\Omega_1 \times \Omega_2$ containing all sets $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$;

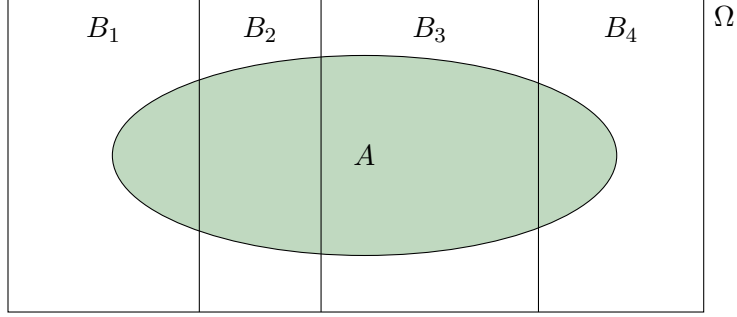


Figure 5: A partition $B_1, \dots, B_4$ of Ω cuts the event A into the disjoint pieces $A \cap B_i$. Summing their probabilities gives the law of total probability (1.1); Bayes' formula asks for the share of the piece $A \cap B_j$ in A .

- 2) the *product measure* $\mathbb{P}_1 \otimes \mathbb{P}_2$ is the unique probability measure on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ such that

$$(\mathbb{P}_1 \otimes \mathbb{P}_2)(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \quad \text{for all } A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2.$$

The model of [Example 1.6](#) is the product of two copies of the Laplace experiment on $\{1, \dots, 6\}$. Product spaces are therefore the natural model for independent repetitions of an experiment. The existence and uniqueness of the product measure are proved, for instance, in [\[Kle20, Dur19\]](#).

1.4 Random variables and their distributions

Often we are interested only in certain aspects of a random model. These are represented by a mapping $X: \Omega \rightarrow \mathbb{R}$.

Definition 1.10 (Real-valued random variables). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A mapping $X: \Omega \rightarrow \mathbb{R}$ is called a *random variable* if it is measurable, that is,

$$X^{-1}(B) = \{\omega \in \Omega: X(\omega) \in B\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

The random variable X induces a probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ via

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)) \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

The measure $\mathbb{P}_X$ is called the *distribution* of X . We write $\mathbb{P}(X \in B) := \mathbb{P}(X^{-1}(B))$, and $\mathbb{P}(X = x) := \mathbb{P}(X^{-1}(\{x\}))$.

Example 1.11 (Fair die). Consider a fair die roll, that is, $\Omega = \{1, \dots, 6\}$ and $\rho(k) = 1/6$ for all $k \in \Omega$. Let $X: \Omega \rightarrow \{0, 1\}$, $X(k) = \mathbf{1}\{k \text{ even}\}$. Then

$$\mathbb{P}_X(\{1\}) = \mathbb{P}(X^{-1}(\{1\})) = \mathbb{P}(\{k: k \text{ even}\}) = \sum_{k \text{ even}} \rho(k) = \frac{1}{6} |\{2, 4, 6\}| = \frac{1}{2}.$$

The distribution $\mathbb{P}_X$ is called the *Bernoulli distribution* with parameter $p = 1/2$.

We call X *discrete* if it takes at most countably many values $x_1, x_2, \dots$, and *continuous* with density f_X if $\mathbb{P}_X(B) = \int_B f_X(x) dx$ for all $B \in \mathcal{B}(\mathbb{R})$.

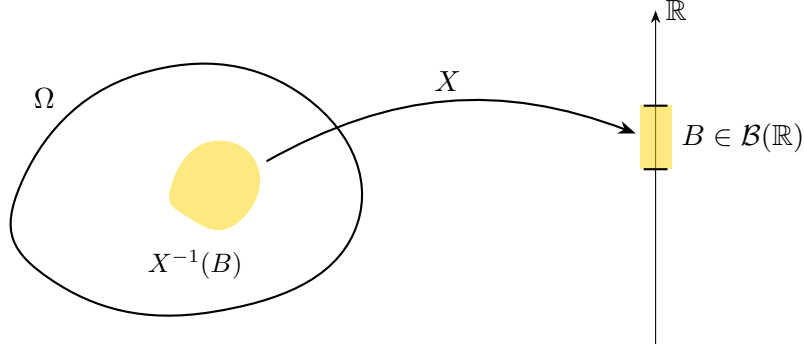


Figure 6: A random variable X maps outcomes to real numbers; the preimage of every Borel set B must be an event.

1.5 Expectation and variance

We are often interested in certain characteristics of probability measures, such as mean values.

Definition 1.12 (Expectations of discrete and continuous models).

- (i) Let X be a discrete random variable with values $x_i \in \mathbb{R}, i \in \mathbb{N}$. If $\mathbb{E}[|X|] := \sum_i |x_i| \mathbb{P}(X = x_i) < \infty$, then

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

is called the *expected value* of X .

- (ii) Let X be a continuous random variable with density f_X . If $\mathbb{E}[|X|] := \int_{\mathbb{R}} |x| f_X(x) dx < \infty$, then

$$\mathbb{E}[X] := \int_{\mathbb{R}} x f_X(x) dx$$

is called the *expected value* of X .

Remark 1.13 (Properties of expected values). Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation.

- 1) *Linearity*: for all $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$.
- 2) *Monotonicity*: if $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- 3) *Jensen's inequality*: if $g: \mathbb{R} \rightarrow \mathbb{R}$ is convex and $\mathbb{E}[|g(X)|] < \infty$, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.
- 4) *Markov's inequality*: for all $\varepsilon > 0$ and all non-decreasing functions $f: \mathbb{R} \rightarrow [0, \infty)$ with $f(\varepsilon) > 0$,

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{1}{f(\varepsilon)} \mathbb{E}[f(X)].$$

Proofs of 1)–3) for general random variables require the Lebesgue integral; see [Kle20, Dur19].

Proof of 4). Since f is non-decreasing and non-negative, $f(\varepsilon) \mathbf{1}\{X \geq \varepsilon\} \leq f(X)$ pointwise on Ω . Monotonicity and $\mathbb{E}[\mathbf{1}\{X \geq \varepsilon\}] = \mathbb{P}(X \geq \varepsilon)$ give the claim. $\square$

Example 1.14.

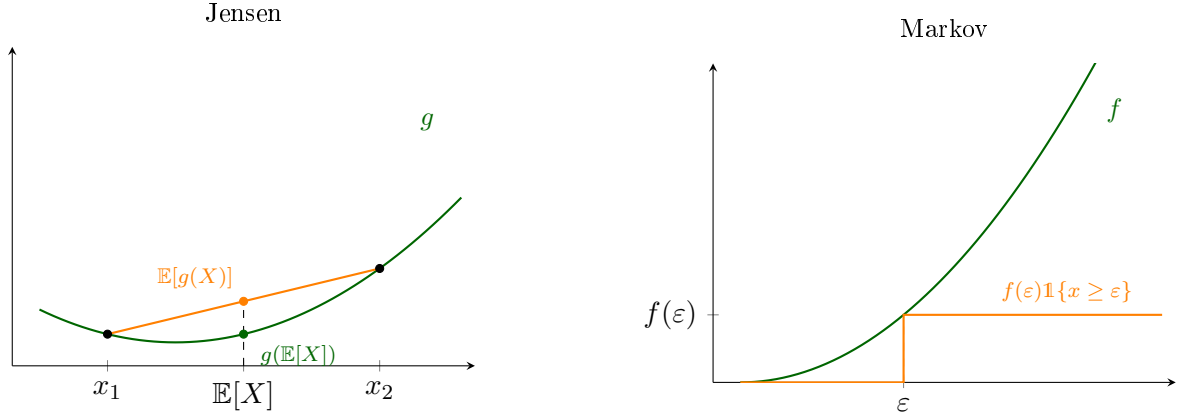


Figure 7: Left: for a convex function g and X taking the values x_1, x_2 with probability $1/2$ each, the chord lies above the graph, so $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. Right: the proof of Markov's inequality rests on the pointwise bound $f(\varepsilon)\mathbb{1}\{x \geq \varepsilon\} \leq f(x)$ for non-decreasing $f \geq 0$; taking expectations at $x = X$ gives the claim.

- 1) *Binomial distribution.* Let $\Omega = \{0, \dots, n\}$ with $\mathbb{P} = \text{Bin}_{n,p}$ and $X(k) = k$. Using $k \binom{n}{k} = n \binom{n-1}{k-1}$ for $k \geq 1$ and the binomial theorem,

$$\mathbb{E}[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np(p + (1-p))^{n-1} = np.$$

- 2) *Exponential distribution.* Let X be exponentially distributed with parameter $\lambda > 0$. Integration by parts gives

$$\mathbb{E}[X] = \int_0^\infty x \lambda e^{-\lambda x} dx = [-x e^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

The second important characteristic measures the expected deviation from the mean value.

Definition 1.15 (Variance). Let X be a random variable with $\mathbb{E}[X^2] < \infty$ and $\mu := \mathbb{E}[X] \in \mathbb{R}$. Then

$$\text{Var}[X] := \mathbb{E}[(X - \mu)^2]$$

is called the *variance* of X , and $\sqrt{\text{Var}[X]}$ is called the *standard deviation* of X . More generally, for $k \in \mathbb{N}$ with $\mathbb{E}[|X|^k] < \infty$, the number $\mathbb{E}[X^k]$ is called the *k-th moment* of X .

Remark 1.16 (Properties of variances). Let X be a random variable with $\mathbb{E}[X^2] < \infty$.

- 1) $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$.
- 2) $\text{Var}[aX + b] = a^2 \text{Var}[X]$ for all $a, b \in \mathbb{R}$.
- 3) $\text{Var}[X] = \min_{a \in \mathbb{R}} \mathbb{E}[(X - a)^2]$.
- 4) *Chebyshev's inequality:* for all $\varepsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \text{Var}[X].$$

Proof. Write $\mu = \mathbb{E}[X]$. Property 1) follows from expanding $(X - \mu)^2$ and linearity; the inequality is Jensen's inequality for $g(x) = x^2$. Property 2) follows from $aX + b - \mathbb{E}[aX + b] = a(X - \mu)$. For 3), $\mathbb{E}[(X - a)^2] = \text{Var}[X] + (\mu - a)^2$, which is minimal for $a = \mu$. Property 4) is Markov's inequality applied to the random variable $|X - \mu|$ with $f(x) = x^2 \mathbf{1}\{x \geq 0\}$. $\square$

Example 1.17.

- 1) *Binomial distribution.* In the setting of Example 1.14 1), the same trick applied twice gives $\mathbb{E}[X(X - 1)] = n(n - 1)p^2$. Hence

$$\text{Var}[X] = \sum_{k=0}^n (k - np)^2 \text{Bin}_{n,p}(k) = n(n - 1)p^2 + np - (np)^2 = np(1 - p).$$

- 2) *Exponential distribution.* Integrating by parts twice yields $\mathbb{E}[X^2] = 2/\lambda^2$, and therefore

$$\text{Var}[X] = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

1.6 Several random variables

Often we consider multiple aspects of a random experiment, that is, two (or more) random variables X, Y on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define independence of random variables as follows.

Definition 1.18 (Independent random variables). A family $X_1, \dots, X_n$ of real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is called *independent* if

$$\mathbb{P}\left(\bigcap_{i=1}^n X_i^{-1}(A_i)\right) = \prod_{i=1}^n \mathbb{P}(X_i^{-1}(A_i)) \quad \text{for all } A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}).$$

Remark 1.19.

- 1) Discrete random variables $X_1, \dots, X_n$ are independent if and only if their joint discrete density factorises,

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

for all $x_1, \dots, x_n \in \mathbb{R}$, where $\rho_{X_i}(x) := \mathbb{P}(X_i = x)$.

- 2) Continuous random variables $X_1, \dots, X_n$ with joint density $f_{X_1, \dots, X_n}$ are independent if and only if

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \quad \text{for (almost) all } x_1, \dots, x_n \in \mathbb{R}.$$

- 3) If $X_1, \dots, X_n$ are independent with finite expectation, then

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

Definition 1.20 (Covariance and correlation). Let X, Y be real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. We define the *covariance* of X and Y as

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The random variables X and Y are called *uncorrelated* if $\text{Cov}(X, Y) = 0$. If moreover $\text{Var}[X] \neq 0 \neq \text{Var}[Y]$, then

$$\text{Corr}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}} \in [-1, 1]$$

is called the *correlation* between X and Y .

The bound $|\text{Corr}(X, Y)| \leq 1$ follows from the Cauchy–Schwarz inequality

$$|\mathbb{E}[UV]| \leq \mathbb{E}[U^2]^{1/2} \mathbb{E}[V^2]^{1/2},$$

applied to $U = X - \mathbb{E}[X]$ and $V = Y - \mathbb{E}[Y]$.

Remark 1.21.

- 1) Independent random variables are uncorrelated by [Remark 1.19 3\)](#), but the reverse implication is wrong in general. For example, let X be uniformly distributed on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\text{Cov}(X, Y) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0$, but $\mathbb{P}(X = 0, Y = 1) = 0 \neq (1/3) \cdot (2/3) = \mathbb{P}(X = 0)\mathbb{P}(Y = 1)$.
- 2) *Bienaymé's formula*: for pairwise uncorrelated random variables $X_1, \dots, X_n$ with finite variance,

$$\text{Var} \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n \text{Var}[X_i].$$

Indeed, the left-hand side equals $\sum_{i,j=1}^n \text{Cov}(X_i, X_j)$, and all terms with $i \neq j$ vanish.

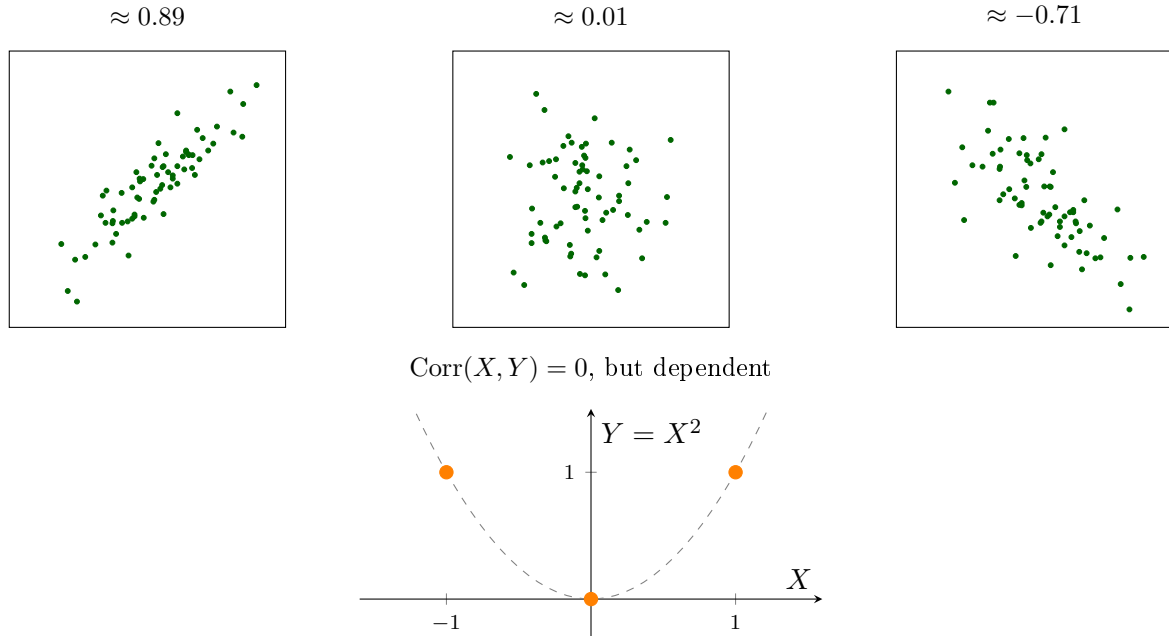


Figure 8: Top: 70 simulated samples of pairs (X, Y) with the indicated empirical correlation (numerical illustration). Correlation measures how well the points are described by a line. Bottom: the example of [Remark 1.21 1\)](#). The three equally likely points lie on a parabola, so Y is a function of X , yet the best linear description is flat.

1.7 Conditional expectations

Based on our understanding of conditional probabilities $\mathbb{P}(A|B)$, we would like to extend this concept to random variables.

Definition 1.22 (Discrete conditional expectations). Let X, Y be discrete real-valued random variables with joint density $\rho(x, y) = \mathbb{P}(X = x, Y = y)$. For all x, y with $\mathbb{P}(Y = y) > 0$, we write

$$\rho(x|y) := \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

for the *conditional density* of X given $Y = y$. If $\mathbb{E}[|X|] < \infty$, we define the *conditional expectation* of X given Y as the random variable

$$\mathbb{E}[X|Y](\omega) := \sum_i x_i \rho(x_i|Y(\omega)), \quad \omega \in \Omega,$$

where the sum runs over all values x_i of X .

Remark 1.23.

- 1) The conditional expectation $\mathbb{E}[X|Y]$ is a random variable, and its expectation is that of X :

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X|Y]] &= \sum_{y_j: \mathbb{P}(Y=y_j)>0} \sum_i x_i \rho(x_i|y_j) \mathbb{P}(Y = y_j) \\ &= \sum_{i,j} x_i \rho(x_i, y_j) = \sum_i x_i \rho_X(x_i) = \mathbb{E}[X]. \end{aligned}$$

- 2) The interpretation is that $x \mapsto \rho(x|y)$ is the probability to see the outcome x in the experiment X when we already know that the experiment Y has the outcome y .
- 3) If X and Y are independent, then $\rho(x|y) = \rho_X(x)$ by [Remark 1.19 1\)](#), and hence $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
- 4) Similarly, for continuous random variables X, Y with joint density $f(x, y)$, one defines the conditional density “morally” as

$$f(x|y) := \frac{f(x, y)}{f_Y(y)}, \quad \text{where } f_Y(y) = \int_{\mathbb{R}} f(\tilde{x}, y) d\tilde{x}.$$

More precisely, $f(\cdot|\cdot)$ is a function such that, for all $A, B \in \mathcal{B}(\mathbb{R})$,

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

Accordingly, $\mathbb{E}[X|Y](\omega) := \int_{\mathbb{R}} x f(x|Y(\omega)) dx$. For the general theory of conditional expectations, see [\[Kle20, Dur19\]](#).

2 Basics of statistics

We now take a new perspective. *Probability theory* is concerned with the modelling and analysis of probabilistic systems. *Statistics* aims to extract information from observations about the underlying (assumed) random generation mechanism, for example a probabilistic model. Our terminology follows [\[Geo13\]](#); see also [\[CB02, Kre05\]](#).

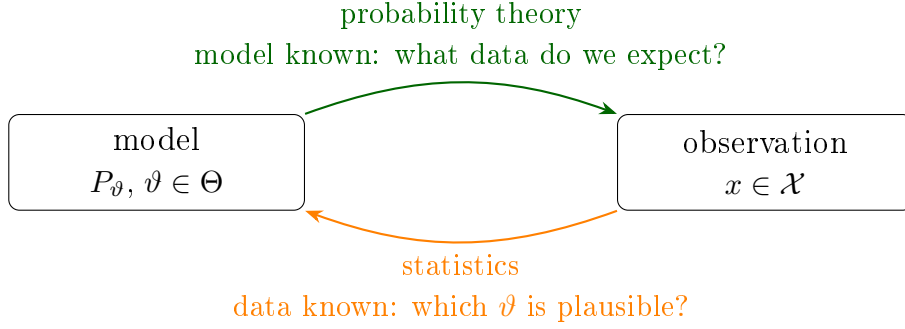


Figure 9: Probability theory and statistics look at the same objects in opposite directions.

2.1 A motivating example

Example 2.1 (Quality control). An apple importer receives $N = 10\,000$ apples. How many apples are bad? She takes a random sample of $n = 50$ apples and sees x bad apples. What can be said about the number w of bad apples among all N ?

Approach 1 (naive estimate). We assume that $w/N \approx x/n$, that is, the importer assumes that there are

$$W(x) = \frac{Nx}{n}$$

bad apples. Such a map W is called an *estimator*.

Approach 2 (estimate with error). The importer does not try to predict w exactly, but rather exhibits an interval $C(x)$ of values which should contain the true value with high probability, that is,

$$P_w(\{x: w \in C(x)\}) \geq 1 - \alpha$$

for all true (unknown) w and the “right” probability measure P_w . Such an interval $C(x)$ is called a *confidence interval* with *confidence level* $1 - \alpha$.

In our example it seems reasonable to assume that the sample is governed by a *hypergeometric distribution*. This is an urn model with w white and $N - w$ black balls, from which n balls are drawn without replacement:

$$P_w(\{x\}) = \text{Hyp}_{n;w,N-w}(\{x\}) := \frac{\binom{w}{x} \binom{N-w}{n-x}}{\binom{N}{n}}, \quad x = 0, \dots, n.$$

Here $\binom{w}{x}$ counts the choices of x bad apples among the w bad ones, $\binom{N-w}{n-x}$ the choices of $n - x$ good apples among the $N - w$ good ones, and $\binom{N}{n}$ the total number of ways to pick n apples out of N . The expected value of this distribution is nw/N , and we look for a map $x \mapsto C(x)$ such that

$$\text{Hyp}_{n;w,N-w}(\{x: C(x) \ni w\}) \geq 1 - \alpha \quad \text{for all } w = 0, \dots, N.$$

Approach 3 (decision making). Maybe the importer is not interested in w itself, but wants to know whether a certain threshold for the percentage of bad apples is exceeded. For example, she can refuse to pay if more than 5% of the apples are bad. This leads to the

null hypothesis $H_0: w \in \{0, \dots, 500\}$ and the alternative $H_1: w \in \{501, \dots, 10\,000\}$.

We need a decision-making criterion of the form

$$x \leq c \implies \text{decide for } H_0, \quad x > c \implies \text{decide for } H_1.$$

Hence, we must find c such that

- 1) $\text{Hyp}_{n;w,N-w}(\{x > c\})$ is small for $w \leq 500$, and
- 2) $\text{Hyp}_{n;w,N-w}(\{x > c\})$ is large for $w > 500$.

Such a procedure is called a *test*.

2.2 Statistical models, statistics and estimators

Definition 2.2 (Statistical model). A *statistical model* is a triple $(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta)$ consisting of

- a *sample space* $\mathcal{X} \neq \emptyset$,
- a σ -algebra $\mathcal{F}$ on $\mathcal{X}$, and
- a family $(P_\vartheta)_{\vartheta \in \Theta}$ of probability measures on $(\mathcal{X}, \mathcal{F})$ with at least two elements.

The index set Θ is called the *parameter space* (often $\Theta \subset \mathbb{R}^d$), and its elements ϑ are called *parameters*.

Remark 2.3.

- 1) If $\Theta \subset \mathbb{R}$, the statistical model is called a *one-parameter model*.
- 2) If $\mathcal{X}$ is discrete (and $\mathcal{F} = \mathcal{P}(\mathcal{X})$), the model is called *discrete*. It is called *continuous* if $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{F} = \mathcal{B}(\mathcal{X})$ and every P_ϑ has a density ρ_ϑ . The class of discrete and continuous models is called the class of *standard models*.
- 3) For a statistical model $(E, \mathcal{E}, Q_\vartheta: \vartheta \in \Theta)$ and $n \geq 2$, we call

$$(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta) = (E^n, \mathcal{E}^{\otimes n}, Q_\vartheta^{\otimes n}: \vartheta \in \Theta)$$

the *n-fold product model*. Here $E^n = E \times \cdots \times E$, the product σ -algebra $\mathcal{E}^{\otimes n}$ is generated by the sets $B_1 \times \cdots \times B_n$ with $B_i \in \mathcal{E}$, and $Q_\vartheta^{\otimes n}$ is the unique probability measure with

$$Q_\vartheta^{\otimes n}(B_1 \times \cdots \times B_n) = \prod_{i=1}^n Q_\vartheta(B_i) \quad \text{for all } B_1, \dots, B_n \in \mathcal{E},$$

as in [Definition 1.9](#). The projection $X_i: \mathcal{X} \rightarrow E$ onto the i -th coordinate is a random variable, and $X_1, \dots, X_n$ are independent with distribution Q_ϑ under every P_ϑ .

Definition 2.4 (Statistics and estimators). Let $(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta)$ be a statistical model and $(\Sigma, \mathcal{S})$ another measurable space.

- (a) Any random variable $S: (\mathcal{X}, \mathcal{F}) \rightarrow (\Sigma, \mathcal{S})$ is called a *statistic*.
- (b) Let $\tau: \Theta \rightarrow \Sigma$ be a mapping that associates to any $\vartheta \in \Theta$ a characteristic $\tau(\vartheta) \in \Sigma$. A statistic $T: \mathcal{X} \rightarrow \Sigma$ is then called an *estimator* for τ .

Remark 2.5.

- 1) *Interpretation.* In probability theory, a random variable $\omega \mapsto X(\omega)$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that we understand, and its outcome is unknown. In statistics, the probability measure is unknown, and a statistic $x \mapsto S(x)$ is a map that we construct in order to extract information from the observation x .

2) *Point estimators.* An important example is $\Sigma = \Theta$ and $\tau(\vartheta) = \vartheta$, so that an estimator $T: \mathcal{X} \rightarrow \Theta$ aims to estimate the parameter ϑ itself. Such an estimator is called a *point estimator* for ϑ and is often denoted by $\hat{\vartheta}$.

Example 2.6 (Point estimators). In a TV show, the host presents random numbers in $[0, \vartheta]$, where ϑ is only known to her. Two players estimate ϑ based on n samples, and the better estimate wins.

Modelling. We take the n -fold product model with $\mathcal{X} = [0, \infty)^n$, $\Theta = (0, \infty)$ and $P_\vartheta = U_{[0, \vartheta]}^{\otimes n}$, the product measure of the uniform distribution on $[0, \vartheta]$.

Player 1. By the law of large numbers (see, for instance, [Geo13, Kle20]), the empirical mean should be close to the expected value:

$$\frac{1}{n} \sum_{i=1}^n X_i \approx \mathbb{E}_\vartheta[X_1] = \frac{1}{\vartheta} \int_0^\vartheta x \, dx = \frac{1}{\vartheta} \left[\frac{x^2}{2} \right]_0^\vartheta = \frac{\vartheta}{2}.$$

Hence, player 1 uses

$$T_n := \frac{2}{n} \sum_{i=1}^n X_i.$$

Player 2. The maximum should be close to ϑ , so player 2 uses

$$\tilde{T}_n := \max\{X_1, \dots, X_n\}.$$

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

$$P_\vartheta(|T_n - \vartheta| \geq \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \text{and} \quad P_\vartheta(|\tilde{T}_n - \vartheta| \geq \varepsilon) \xrightarrow{n \rightarrow \infty} 0.$$

For T_n this follows from Chebyshev's inequality and the variance computed below. For $\tilde{T}_n \leq \vartheta$ and $\varepsilon < \vartheta$, independence gives $P_\vartheta(|\tilde{T}_n - \vartheta| \geq \varepsilon) = P_\vartheta(X_1 \leq \vartheta - \varepsilon)^n = (1 - \varepsilon/\vartheta)^n$.

Unbiasedness. The estimator 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 $\tilde{T}_n$, independence gives, for $c \in [0, \vartheta]$,

$$P_\vartheta(\tilde{T}_n \leq c) = P_\vartheta(X_1 \leq c)^n = \left(\frac{c}{\vartheta}\right)^n =: F(c),$$

so $\tilde{T}_n$ has density $F'(x) = nx^{n-1}\vartheta^{-n}$ on $[0, \vartheta]$, and

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

Hence, $\tilde{T}_n$ is not unbiased, but

$$\tilde{T}_n^* := \frac{n+1}{n} \tilde{T}_n$$

is unbiased and consistent.

Which estimator fluctuates more? By Bienaymé's formula and $\text{Var}_\vartheta[X_1] = \vartheta^2/12$,

$$\text{Var}_\vartheta[T_n] = \left(\frac{2}{n}\right)^2 \text{Var}_\vartheta\left[\sum_{i=1}^n X_i\right] = \frac{4}{n} \text{Var}_\vartheta[X_1] = \frac{4}{n\vartheta} \int_0^\vartheta \left(x - \frac{\vartheta}{2}\right)^2 \, dx = \frac{\vartheta^2}{3n}.$$

On the other hand,

$$\begin{aligned}\mathrm{Var}_\vartheta[\tilde{T}_n] &= \mathbb{E}_\vartheta[\tilde{T}_n^2] - \mathbb{E}_\vartheta[\tilde{T}_n]^2 = \int_0^\vartheta x^2 n x^{n-1} \vartheta^{-n} dx - \left(\frac{n\vartheta}{n+1}\right)^2 \\ &= \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) \vartheta^2 = \frac{n\vartheta^2}{(n+1)^2(n+2)},\end{aligned}$$

and therefore

$$\mathrm{Var}_\vartheta[\tilde{T}_n^*] = \frac{(n+1)^2}{n^2} \mathrm{Var}_\vartheta[\tilde{T}_n] = \frac{\vartheta^2}{n(n+2)}.$$

Both T_n and $\tilde{T}_n^*$ are unbiased, but the variance of $\tilde{T}_n^*$ decays like $1/n^2$ instead of $1/n$, so $\tilde{T}_n^*$ is the winner.

Lesson. We can apply different criteria to estimators and evaluate them accordingly.

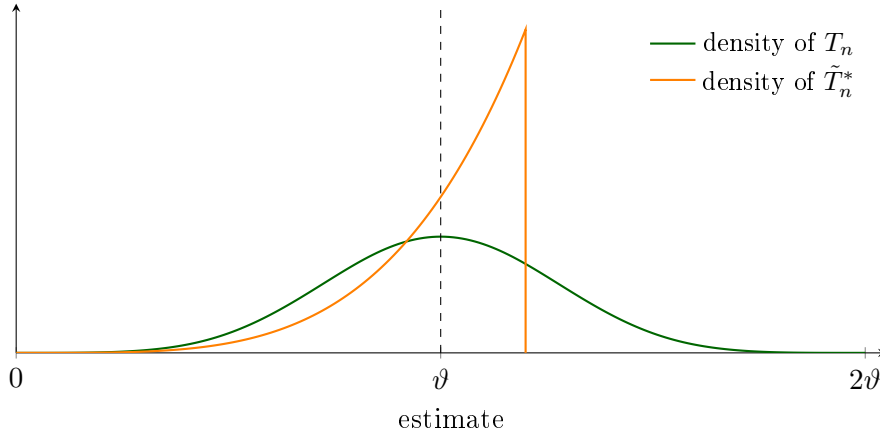


Figure 10: Exact densities of the two unbiased estimators of [Example 2.6](#) for $n = 5$ (illustrative choice). Both have mean ϑ , but $\tilde{T}_n^*$ is concentrated much more closely around ϑ ; its variance is $\vartheta^2/35$ compared with $\vartheta^2/15$ for T_n . The density of $\tilde{T}_n^*$ vanishes above $(n+1)\vartheta/n$.

Definition 2.7 (Bias). Let $(\mathcal{X}, \mathcal{F}, P_\vartheta : \vartheta \in \Theta)$ be a statistical model and $\tau : \Theta \rightarrow \mathbb{R}$ a characteristic. An estimator $T : \mathcal{X} \rightarrow \mathbb{R}$ for τ is called *unbiased* if

$$\mathbb{E}_\vartheta[T] = \tau(\vartheta) \quad \text{for all } \vartheta \in \Theta.$$

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

2.3 Mean, variance and maximum-likelihood estimators

Theorem 2.8 (Mean and variance estimators). *Let $n \geq 2$ and let $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), Q_\vartheta^{\otimes n} : \vartheta \in \Theta)$ be the real n -fold product model. Assume that, for every $\vartheta \in \Theta$, the distribution Q_ϑ has finite second moment, with expected value and variance*

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

Then

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

are unbiased estimators for m and v , respectively.

Proof. Fix $\vartheta \in \Theta$. Linearity gives $\mathbb{E}_\vartheta[M] = m(\vartheta)$. Expanding the square,

$$\sum_{i=1}^n (X_i - M)^2 = \sum_{i=1}^n (X_i - m(\vartheta))^2 - n(M - m(\vartheta))^2.$$

The first sum has expectation $nv(\vartheta)$. Since the X_i are independent, Bienaymé's formula gives $\mathbb{E}_\vartheta[(M - m(\vartheta))^2] = \text{Var}_\vartheta[M] = v(\vartheta)/n$. Hence the expectation of the left-hand side is $(n-1)v(\vartheta)$, and $\mathbb{E}_\vartheta[V^*] = v(\vartheta)$. $\square$

The normalisation $1/(n-1)$ instead of $1/n$ compensates for the fact that M is itself estimated from the data.

We now introduce an important class of estimators; for a detailed treatment, see [CB02, LC98].

Definition 2.9 (Maximum-likelihood estimators). Let $(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta)$ be a standard model, so that every P_ϑ has a density $x \mapsto \rho_\vartheta(x)$. The function

$$\rho: \mathcal{X} \times \Theta \rightarrow [0, \infty), \quad \rho(x, \vartheta) := \rho_\vartheta(x),$$

is called the *likelihood function*, and $\rho_x: \Theta \rightarrow [0, \infty)$, $\rho_x(\vartheta) := \rho(x, \vartheta)$, is called the *likelihood function for the observation* $x \in \mathcal{X}$. An estimator $T: \mathcal{X} \rightarrow \Theta$ for ϑ is called a *maximum-likelihood estimator* if

$$\rho(x, T(x)) = \max_{\vartheta \in \Theta} \rho(x, \vartheta) \quad \text{for all } x \in \mathcal{X},$$

that is, $T(x) \in \arg \max \rho_x$.

How do we find such a T ? Since the logarithm is strictly increasing, ρ_x and the *log-likelihood function* $\vartheta \mapsto \log \rho_x(\vartheta)$ have the same maximisers, and products of densities become sums.

Example 2.10 (Binomial distribution). Consider the model

$$(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), \text{Bin}_{n, \vartheta}: \vartheta \in [0, 1]).$$

For $x \in \{1, \dots, n-1\}$ and $\vartheta \in (0, 1)$,

$$\frac{d}{d\vartheta} \log \rho_x(\vartheta) = \frac{d}{d\vartheta} \left(\log \binom{n}{x} + x \log \vartheta + (n-x) \log(1-\vartheta) \right) = \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta}.$$

This derivative vanishes if and only if $\vartheta = x/n$, and it is positive to the left and negative to the right of this point. For $x = 0$ and $x = n$, the likelihood $\rho_x(\vartheta) = (1-\vartheta)^n$, respectively ϑ^n , is maximal at $\vartheta = 0$, respectively $\vartheta = 1$. Hence, the maximum-likelihood estimator is

$$T(x) = \frac{x}{n},$$

the relative frequency of successes.

Definition 2.11 (Best estimator). Let $(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta)$ be a statistical model. An unbiased estimator T for the real-valued characteristic $\tau(\vartheta)$ is called *variance-minimising* or a *best estimator* if, for all other unbiased estimators S for τ ,

$$\text{Var}_\vartheta[T] \leq \text{Var}_\vartheta[S] \quad \text{for all } \vartheta \in \Theta.$$

Remark 2.12. A large class of statistical models, the so-called *exponential families*, possesses best estimators; see, for instance, [Geo13, LC98].

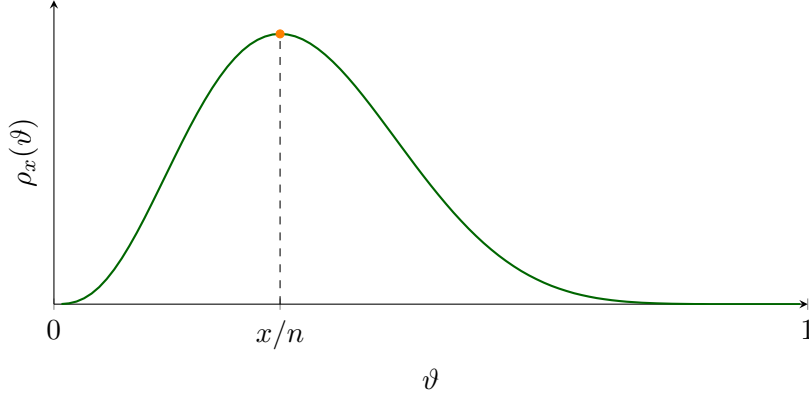


Figure 11: Likelihood function $\vartheta \mapsto \rho_x(\vartheta) = \text{Bin}_{n,\vartheta}(x)$ from [Example 2.10](#) for $n = 10$ and the observation $x = 3$. It is maximal at $T(x) = x/n = 0.3$.

2.4 Statistical tests

Recall the apple importer from [Example 2.1](#). We now give a formal definition and set-up for a statistical test. The design of a statistical test follows five steps.

Step 1: define a statistical model. In our example, $\mathcal{X} = \{0, \dots, n\}$, $\Theta = \{0, \dots, N\}$ and $P_\vartheta = \text{Hyp}_{n;\vartheta, N-\vartheta}$ for $\vartheta \in \Theta$.

Step 2: formulate the (null) hypothesis and the alternative. We split the parameter set Θ into disjoint sets Θ_0 and Θ_1 such that

$$\vartheta \in \Theta_0 \iff \vartheta \text{ is acceptable}, \quad \vartheta \in \Theta_1 \iff \vartheta \text{ is problematic.}$$

We say that the *hypothesis* $H_0: \vartheta \in \Theta_0$ is tested against the *alternative* $H_1: \vartheta \in \Theta_1$. In our example, $\Theta_0 = \{0, \dots, 500\}$ (good quality) and $\Theta_1 = \{501, \dots, 10000\}$ (bad quality).

Step 3: choose an acceptance level. Choose $0 < \alpha < 1$, for example $\alpha = 0.05$. The probability to choose the alternative although the hypothesis holds should be at most α .

Step 4: choose a testing rule. We choose a statistic $\varphi: \mathcal{X} \rightarrow [0, 1]$ as follows. For the observation $x \in \mathcal{X}$, the value $\varphi(x)$ represents the degree to which we tend to the alternative:

$$\begin{aligned} \varphi(x) = 0 &\iff \text{we accept the hypothesis,} \\ \varphi(x) = 1 &\iff \text{we accept the alternative,} \\ 0 < \varphi(x) < 1 &\iff \text{we are not sure; we perform an auxiliary random experiment} \\ &\quad \text{in which we accept the alternative with probability } \varphi(x). \end{aligned}$$

For example,

$$\varphi(x) = \begin{cases} 1, & x > c, \\ \gamma, & x = c, \\ 0, & x < c, \end{cases}$$

where c is the smallest number for which the requirement of Step 3 is maintained by the non-randomised rule $\mathbb{1}\{x > c\}$, and $\gamma \in [0, 1]$ is then chosen as large as possible. In words, we choose the alternative as often as possible, but not so often that cases in which the hypothesis actually holds lead to the alternative too often.

Step 5: perform the experiment.

Definition 2.13 (Statistical test). Let $(\mathcal{X}, \mathcal{F}, P_\vartheta: \vartheta \in \Theta)$ be a statistical model with $\Theta = \Theta_0 \dot{\cup} \Theta_1$, where Θ_0 is the null hypothesis and Θ_1 the alternative.

- (a) Every statistic $\varphi: \mathcal{X} \rightarrow [0, 1]$ is called a *test* of Θ_0 against Θ_1 . A test is called *non-randomised* if $\varphi(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$; otherwise it is called *randomised*. For a non-randomised test, the set $\{x \in \mathcal{X}: \varphi(x) = 1\}$ is called the *critical region* of φ .
- (b) The number $\sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}[\varphi]$ is called the *effective level* of φ . If $\sup_{\vartheta \in \Theta_0} \mathbb{E}_{\vartheta}[\varphi] \leq \alpha$ for some $\alpha \in (0, 1)$, then φ is called a test of *level* α .
- (c) The function $G_{\varphi}: \Theta \rightarrow [0, 1]$, $G_{\varphi}(\vartheta) := \mathbb{E}_{\vartheta}[\varphi]$, is called the *power function* of φ . For $\vartheta \in \Theta_1$, it is the probability with which the alternative is detected if it is present.

Remark 2.14 (Two kinds of errors). For $\vartheta \in \Theta_0$, the number $\mathbb{E}_{\vartheta}[\varphi]$ is called the *type I error*; for a test of level α it is at most α . For $\vartheta \in \Theta_1$, the number $1 - \mathbb{E}_{\vartheta}[\varphi]$ is called the *type II error*. The theory of optimal tests, which minimise the type II error among all tests of level α , is developed in [LR22].

Example 2.15 (Apple importer, continued). For $N = 10\,000$, $n = 50$ and $\alpha = 0.05$, a numerical evaluation of the hypergeometric probabilities gives $\text{Hyp}_{50;500,9500}(\{x > 4\}) \approx 0.103$ and $\text{Hyp}_{50;500,9500}(\{x > 5\}) \approx 0.037$. Since $\text{Hyp}_{n;\vartheta,N-\vartheta}(\{x > c\})$ is non-decreasing in ϑ , the supremum over Θ_0 is attained at $\vartheta = 500$, and Step 4 yields $c = 5$ and $\gamma \approx 0.19$. For $\vartheta = 1000$, that is, 10% bad apples, the power is only $G_{\varphi}(1000) \approx 0.42$. A sample of 50 apples is thus too small to detect this alternative reliably: the type II error is about 0.58.

2.5 Linear regression

Linear regression is treated in [Geo13], and in depth in [SL03].

Example 2.16 (Metal bar). Consider a metal bar. We measure its length deformation at temperatures $t_1, \dots, t_n \in \mathbb{R}$ and assume that

$$X_k = \gamma_0 + \gamma_1 t_k + \sqrt{v} \xi_k, \quad 1 \leq k \leq n. \quad (2.1)$$

Here, X_k is the measurement, $\gamma_0, \gamma_1 \in \mathbb{R}$ are the unknown coefficients (the target variables), t_k is the *regression variable*, $v > 0$ is an unknown *fluctuation parameter*, and ξ_k is a random measurement error with $\mathbb{E}[\xi_k] = 0$ and $\text{Var}[\xi_k] = 1$. In vector notation, with $\mathbf{X} = (X_1, \dots, X_n)^\top$, $\mathbf{1} = (1, \dots, 1)^\top$, $\mathbf{t} = (t_1, \dots, t_n)^\top$ and $\xi = (\xi_1, \dots, \xi_n)^\top$, the model reads

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

Let $P_{\gamma_0, \gamma_1, v}$ denote the distribution of this random vector, and consider the statistical model

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\gamma_0, \gamma_1, v}: \gamma_0, \gamma_1 \in \mathbb{R}, v \in (0, \infty)).$$

Principle of least squares. We choose $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1)$ such that the mean-square error

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

is minimal at $\gamma = \hat{\gamma}$. To find the minimum, we set the partial derivatives to zero:

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

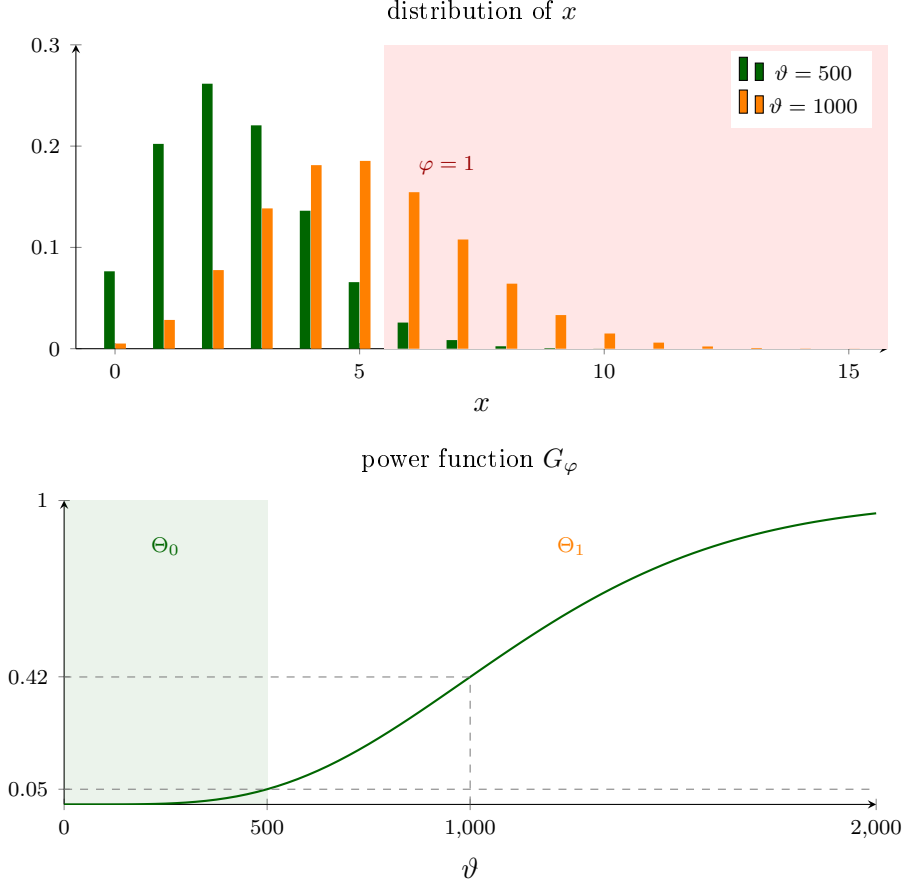


Figure 12: The test of [Example 2.15](#) with $c = 5$ (numerical values). Top: hypergeometric densities of the number x of bad apples in the sample for $\vartheta = 500$ and $\vartheta = 1000$; the test decides for the alternative in the shaded region $x > 5$ and with probability $\gamma \approx 0.19$ at $x = 5$. Bottom: the power function stays below $\alpha = 0.05$ on Θ_0 and increases only slowly on Θ_1 ; the type II error is $1 - G_\varphi(\vartheta)$.

With the empirical mean $M(\mathbf{x}) := (1/n) \sum_{k=1}^n x_k$ for $\mathbf{x} \in \mathbb{R}^n$, this yields the *normal equations*

$$M(\mathbf{X}) = \gamma_0 + \gamma_1 M(\mathbf{t}), \quad (2.2)$$

$$\frac{1}{n} \sum_{k=1}^n t_k X_k = \gamma_0 M(\mathbf{t}) + \gamma_1 \frac{1}{n} \sum_{k=1}^n t_k^2. \quad (2.3)$$

We introduce the empirical variance and covariance

$$V(\mathbf{t}) := \frac{1}{n} \sum_{k=1}^n t_k^2 - M(\mathbf{t})^2 \quad \text{and} \quad c(\mathbf{t}, \mathbf{x}) := \frac{1}{n} \sum_{k=1}^n t_k x_k - M(\mathbf{t})M(\mathbf{x}).$$

Then (2.3) reads $\gamma_0 M(\mathbf{t}) + \gamma_1 (V(\mathbf{t}) + M(\mathbf{t})^2) = c(\mathbf{t}, \mathbf{X}) + M(\mathbf{t})M(\mathbf{X})$. Subtracting $M(\mathbf{t})$ times (2.2), we obtain

$$\gamma_1 V(\mathbf{t}) = c(\mathbf{t}, \mathbf{X}).$$

Theorem 2.17 (Least-squares estimators). *In the setting of [Example 2.16](#), assume that not all t_k are equal, so that $V(\mathbf{t}) > 0$. Then*

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

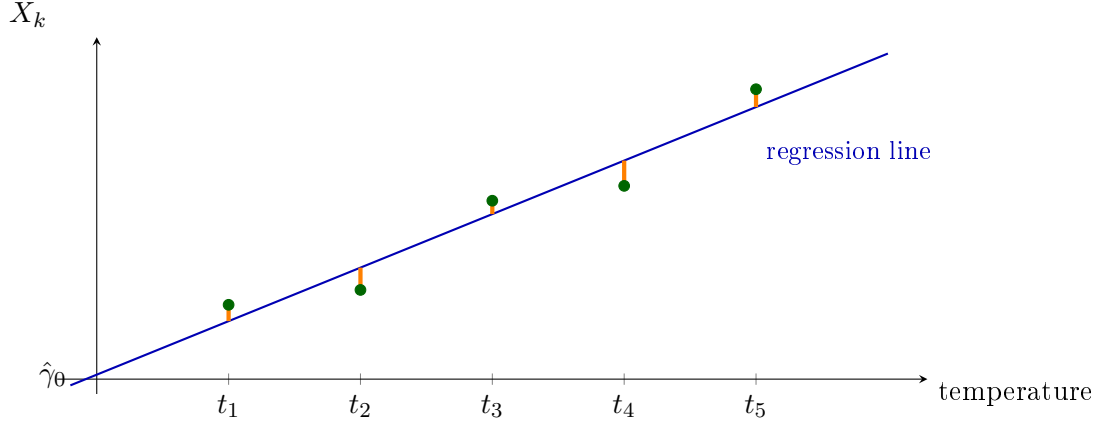


Figure 13: Measurements X_k at temperatures t_k (illustrative data) and the least-squares regression line $t \mapsto \hat{\gamma}_0 + \hat{\gamma}_1 t$. The line minimises the mean of the squared lengths of the vertical residuals $X_k - (\hat{\gamma}_0 + \hat{\gamma}_1 t_k)$, drawn in orange.

are the unique least-squares estimators for γ_0 and γ_1 . They are unbiased.

Proof. By the computation above, $(\hat{\gamma}_0, \hat{\gamma}_1)$ is the unique solution of the normal equations (2.2)–(2.3). The function $\gamma \mapsto F_\gamma$ is a quadratic polynomial whose Hessian

$$2 \begin{pmatrix} 1 & M(\mathbf{t}) \\ M(\mathbf{t}) & \frac{1}{n} \sum_k t_k^2 \end{pmatrix}$$

has positive trace and determinant $4V(\mathbf{t}) > 0$. Hence F_γ is strictly convex, and its unique critical point is its unique minimiser.

For unbiasedness, $\mathbf{x} \mapsto c(\mathbf{t}, \mathbf{x})$ is linear with $c(\mathbf{t}, \mathbf{1}) = 0$ and $c(\mathbf{t}, \mathbf{t}) = V(\mathbf{t})$. By (2.1) and $\mathbb{E}[\xi_k] = 0$, we have $\mathbb{E}_{\gamma_0, \gamma_1, v}[\mathbf{X}] = \gamma_0 \mathbf{1} + \gamma_1 \mathbf{t}$, and thus

$$\mathbb{E}_{\gamma_0, \gamma_1, v}[\hat{\gamma}_1] = \frac{c(\mathbf{t}, \gamma_0 \mathbf{1} + \gamma_1 \mathbf{t})}{V(\mathbf{t})} = \gamma_1, \quad \mathbb{E}_{\gamma_0, \gamma_1, v}[\hat{\gamma}_0] = \gamma_0 + \gamma_1 M(\mathbf{t}) - \gamma_1 M(\mathbf{t}) = \gamma_0. \quad \square$$

Remark 2.18 (Geometric interpretation). Equip $\mathbb{R}^n$ with the Euclidean inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n x_k y_k$. The observation satisfies $\mathbf{X} \in \text{span}(\mathbf{1}, \mathbf{t}, \xi)$, but not necessarily $\mathbf{X} \in \text{span}(\mathbf{1}, \mathbf{t})$. The least-squares principle says that $\hat{\mathbf{X}} := \hat{\gamma}_0 \mathbf{1} + \hat{\gamma}_1 \mathbf{t} \in \text{span}(\mathbf{1}, \mathbf{t})$ has the property

$$\frac{1}{n} \sum_{k=1}^n (X_k - \hat{X}_k)^2 = \min_{\mathbf{y} \in \text{span}(\mathbf{1}, \mathbf{t})} \frac{1}{n} \sum_{k=1}^n (X_k - y_k)^2.$$

In fact, $\hat{\mathbf{X}}$ is the orthogonal projection of $\mathbf{X}$ onto the linear subspace

$$L = L(\mathbf{1}, \mathbf{t}) := \{\gamma_0 \mathbf{1} + \gamma_1 \mathbf{t} : \gamma_0, \gamma_1 \in \mathbb{R}\}.$$

Indeed, by the normal equations (2.2) and (2.3),

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

References

- [CB02] G. Casella and R. L. Berger, *Statistical inference*, 2nd ed., Duxbury, Pacific Grove, CA, 2002.

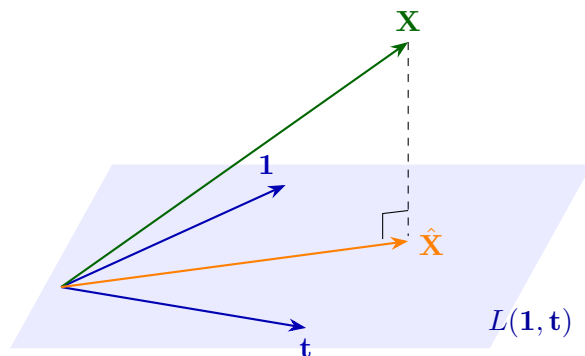


Figure 14: The least-squares fit $\hat{\mathbf{X}}$ is the orthogonal projection of $\mathbf{X}$ onto the plane spanned by $\mathbf{1}$ and $\mathbf{t}$.

- [Dur19] R. Durrett, *Probability: Theory and examples*, 5th ed., Cambridge Series in Statistical and Probabilistic Mathematics, vol. 49, Cambridge University Press, Cambridge, 2019.
- [Fel68] W. Feller, *An introduction to probability theory and its applications. Vol. I*, 3rd ed., John Wiley & Sons, New York, 1968.
- [Geo13] H.-O. Georgii, *Stochastics: Introduction to probability and statistics*, 2nd ed., De Gruyter Textbook, De Gruyter, Berlin, 2013.
- [GS20] G. R. Grimmett and D. R. Stirzaker, *Probability and random processes*, 4th ed., Oxford University Press, Oxford, 2020.
- [Kle20] A. Klenke, *Probability theory: A comprehensive course*, 3rd ed., Universitext, Springer, Cham, 2020.
- [Kre05] U. Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 8th ed., Vieweg, Wiesbaden, 2005.
- [LC98] E. L. Lehmann and G. Casella, *Theory of point estimation*, 2nd ed., Springer Texts in Statistics, Springer, New York, 1998.
- [LR22] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, 4th ed., Springer Texts in Statistics, Springer, Cham, 2022.
- [SL03] G. A. F. Seber and A. J. Lee, *Linear regression analysis*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, 2003.
- [Was04] L. Wasserman, *All of statistics: A concise course in statistical inference*, Springer Texts in Statistics, Springer, New York, 2004.