Ramp-Up Mathematics: Stochastics

Lecture notes and extended notes

Benedikt Jahnel
Institute for Mathematical Stochastics, Technische Universität Braunschweig

Winter semester 2026/27

Abstract

These notes accompany a short course on the basics of probability theory and statistics. Part I contains the material of the four lectures: the basic model of probability theory, random variables and their characteristics, statistical models, estimators, tests and linear regression. Part II contains the extended notes. They revisit the lecture material in more depth and add the results that students are expected to know when they enter a master's-level course in probability theory or statistics: distribution functions and quantiles, random vectors, limit theorems, normal distributions and their relatives, confidence intervals, tests in normal models and general linear models.

Contents

I Lecture notes **2**

1 Basics of probability theory **3**
1.1 The basic model . . . 3
1.2 Independence and conditional probabilities . . . 5
1.3 Product spaces . . . 7
1.4 Random variables and their distributions . . . 7
1.5 Expectation and variance . . . 8
1.6 Several random variables . . . 10
1.7 Conditional expectations . . . 12

2 Basics of statistics **13**
2.1 A motivating example . . . 13
2.2 Statistical models, statistics and estimators . . . 14
2.3 Mean, variance and maximum-likelihood estimators . . . 17
2.4 Statistical tests . . . 19
2.5 Linear regression . . . 20

II Extended notes **22**

3 Probability spaces and distributions on the real line **23**
3.1 Further properties of probability measures . . . 23
3.2 Distribution functions and quantiles . . . 24
3.3 Independence and conditioning . . . 25
3.4 Products of finitely many probability spaces . . . 26

4 Random variables, moments and random vectors **27**
4.1 Measurability . . . 27
4.2 Expectation and moments . . . 27
4.3 Random vectors . . . 28
4.4 Conditional expectations . . . 31

5 Limit theorems **32**
5.1 Modes of convergence . . . 32
5.2 Laws of large numbers . . . 34
5.3 The central limit theorem . . . 35

6 Normal distributions and their relatives **37**
6.1 One-dimensional normal distributions . . . 37
6.2 Multivariate normal distributions . . . 38
6.3 Chi-square and Student distributions . . . 39

7 Point estimation **41**
7.1 Statistical models and observations . . . 41
7.2 Consistency and mean squared error . . . 41
7.3 Maximum-likelihood estimators in product models . . . 42
7.4 Best estimators . . . 43

8 Confidence intervals **43**

9 Statistical tests **45**
9.1 Errors and their asymmetry . . . 45
9.2 Gauss tests . . . 46
9.3 t-tests and p-values . . . 47

10 Linear models **48**

How to read these notes

Part I is self-contained and is what we cover in the lectures. Part II follows the same order, uses the same notation and refers back to Part I instead of repeating definitions. Most results in Part II are stated without proof or with a short proof; references for the omitted proofs are given in the text.

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 [BD01], [CB02] and [Was04], and, for more advanced topics, [LC98] on estimation, [LR22] on testing and [SL03] on linear regression.

Part I
Lecture notes

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* $\mathcal{F} \subset \mathcal{P}(\Omega) := \{A \colon A \subset \Omega\}$, 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 $A^{\mathrm{c}} := \Omega \setminus A \in \mathcal{F}$;
3) if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

(iii) The *probability measure* $\mathbb{P} \colon \mathcal{F} \to [0,1]$, 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 $i \neq j$,

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

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) $A^{\mathrm{c}} \in \mathcal{F}$, "$A$ has not occurred".

(ii) Probability measures have the following properties:

1) $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$;
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 $A_i = \emptyset = \Omega^{\mathrm{c}}$ for $i \geq 3$. De Morgan's law gives $A \cap B = (A^{\mathrm{c}} \cup B^{\mathrm{c}})^{\mathrm{c}}$, and $B \setminus A = B \cap A^{\mathrm{c}}$. For (ii), we first note that $\mathbb{P}(\emptyset) = 0$, since σ-additivity applied to $\emptyset = \bigcup_{i=1}^{\infty} \emptyset$ gives $\mathbb{P}(\emptyset) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$. Choosing $A_i = \emptyset$ for all but finitely many i, we see that $\mathbb{P}$ is also finitely additive. Applied to $\Omega = A \cup A^{\mathrm{c}}$, finite additivity 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). □

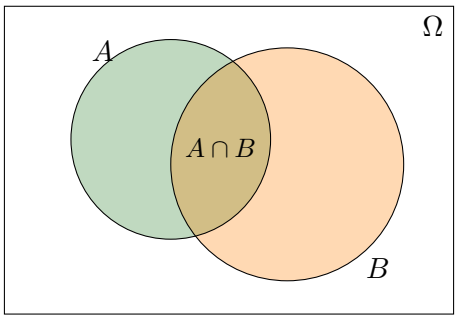

Figure 1: Two events $A, B \in \mathcal{F}$ in the sample space Ω. In $\mathbb{P}(A) + \mathbb{P}(B)$ the overlap $A \cap B$ is counted twice, which explains the correction term in Remark 1.2 (ii) 2).

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* $\rho\colon \Omega \to [0,1]$ with $\sum_{k\in\Omega} \rho(k) = 1$ via

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where the *(Lebesgue) density* $f\colon \mathbb{R} \to [0, \infty)$ satisfies $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.

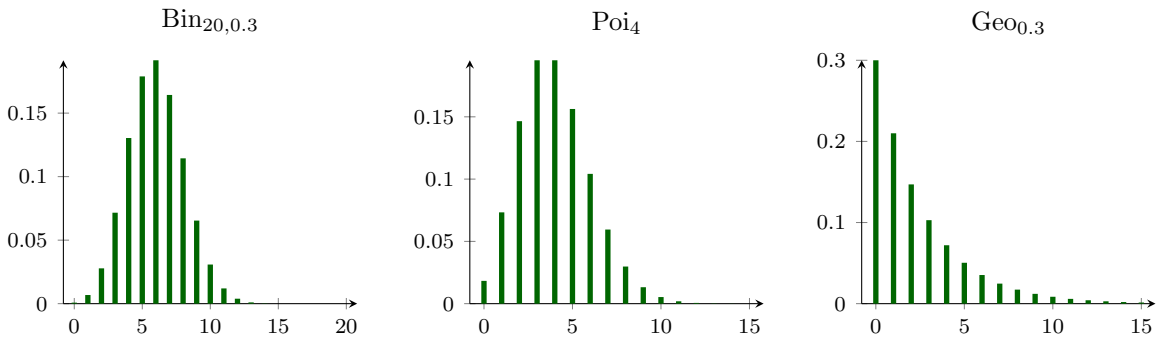

Figure 2: Discrete densities from Example 1.3 for illustrative parameter values. The binomial and the Poisson density are concentrated around $np = 6$ and $\lambda = 4$, respectively, while the geometric density decreases geometrically from $k = 0$.

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

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

(ii) The *exponential distribution* with parameter $\lambda > 0$ has density

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

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

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

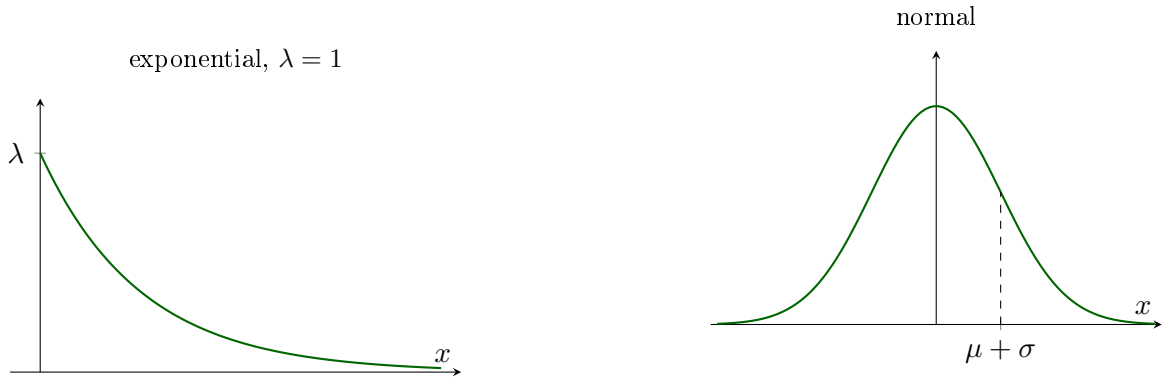

Figure 3: Densities of the exponential and of the normal distribution from Example 1.4.

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 $|J| < \infty$,

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

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

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

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

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

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

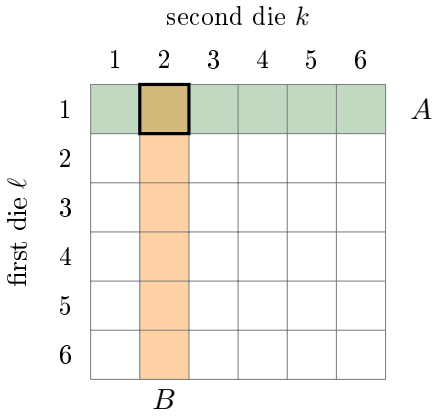

Figure 4: The 36 equally likely outcomes of Example 1.6. The row A and the column B each contain 6 outcomes and meet in exactly one, so $\mathbb{P}(A \cap B) = 1/36 = (6/36)\cdot(6/36)$.

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 $\mathbb{P}(B) > 0$.

(i) The map $\mathbb{P}(\cdot|B)$ is again a probability measure on $(\Omega, \mathcal{F})$.

(ii) If A and B are independent, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

(iii) *Law of total probability.* Let $I = \{1,\dots,m\}$ for some $m \in \mathbb{N}$ or $I = \mathbb{N}$, and let $(B_i)_{i\in I}$ be a finite or countable *partition* of Ω, that is, $B_i \in \mathcal{F}$, $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i\in I} B_i = \Omega$. Assume that $\mathbb{P}(B_i) > 0$ for all $i \in I$. Then, for all $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \cap B_i) = \sum_{i\in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i). \tag{1.1}$$

(iv) *Bayes' formula.* If, in the setting of (iii), also $\mathbb{P}(A) > 0$, then, for all $j \in I$,

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

Proof. Property (i) follows from $\mathbb{P}(\Omega \cap B) = \mathbb{P}(B)$ and the σ-additivity of $\mathbb{P}$. Property (ii) is immediate from Definition 1.5. For (iii), the events $A \cap B_i$ are pairwise disjoint with union A, so σ-additivity (or finite additivity, see Remark 1.2, if I is finite) gives the first identity, and Definition 1.7 the second. For (iv), 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. □

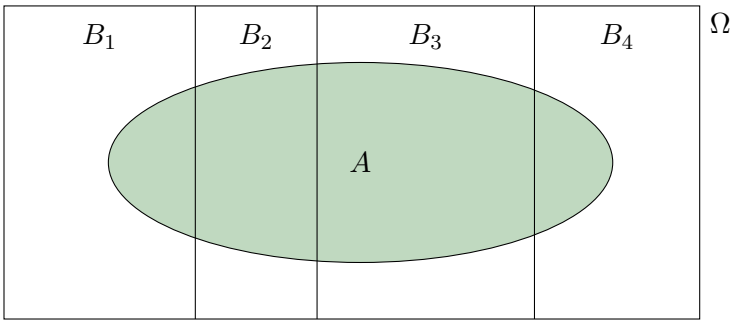

Figure 5: A partition $B_1, \dots, B_4$ of Ω cuts the event A into the disjoint pieces $A \cap B_i$. Summing their probabilities gives the law of total probability (1.1); Bayes' formula asks for the share of the piece $A \cap B_j$ in A.

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:

(i) 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 $A_2 \in \mathcal{F}_2$;

(ii) 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

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

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 $X \colon \Omega \to \mathbb{R}$.

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

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

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

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

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

More generally, let $(E, \mathcal{E})$ be a *measurable space*, that is, a set $E \neq \emptyset$ together with a σ-algebra $\mathcal{E}$ on E. A map $X \colon \Omega \to E$ with $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{E}$ is called an *E-valued random variable*, or a *measurable map* from $(\Omega, \mathcal{F})$ to $(E, \mathcal{E})$. Its distribution $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$ is a probability measure on $(E, \mathcal{E})$. Real-valued random variables correspond to $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and random vectors to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, where $\mathcal{B}(\mathbb{R}^n)$ is generated by the products of intervals.

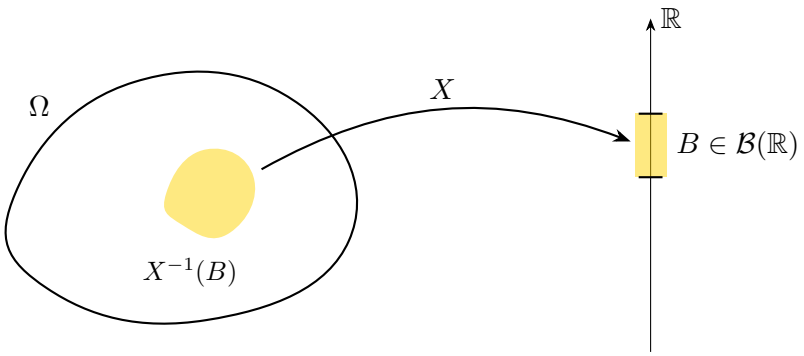

Figure 6: A random variable X maps outcomes to real numbers; the preimage of every Borel set B must be an event.

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 $X \colon \Omega \to \{0, 1\}$, $X(k) = \mathbf{1}\{k \text{ even}\}$. Then

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

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 $\mathbb{P}_X(B) = \int_B f_X(x)\,\mathrm{d}x$ for all $B \in \mathcal{B}(\mathbb{R})$.

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 taking the values $x_i \in \mathbb{R}$ for $i \in \mathbb{N}$. If $\mathbb{E}[|X|] \coloneqq \sum_i |x_i|\,\mathbb{P}(X = x_i)$ is finite, then

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

is called the *expected value* of X.

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

$$\mathbb{E}[X] \coloneqq \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$$

is called the *expected value* of X.

Remark 1.13 (Transformation formula and general expectations).

(i) For non-negative X, the series and the integral in Definition 1.12 are always defined in $[0, \infty]$, and we use them as the definition of $\mathbb{E}[X] \in [0, \infty]$ also when they are infinite.

(ii) Let $g \colon \mathbb{R} \to \mathbb{R}$ be measurable, that is, $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for all $B \in \mathcal{B}(\mathbb{R})$; this holds, for example, for continuous or monotone g. Then $g(X)$ is again a random variable, and the *transformation formula*

$$\mathbb{E}[g(X)] = \sum_i g(x_i)\,\mathbb{P}(X = x_i) \qquad \text{respectively} \qquad \mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x$$

holds for discrete, respectively continuous, X, whenever $g \geq 0$ or the right-hand side is finite with $|g|$ in place of g. In particular, moments and variances below are computed in this way.

(iii) For general random variables, which are neither discrete nor continuous, $\mathbb{E}[X]$ is defined as the Lebesgue integral $\int_\Omega X \, \mathrm{d}\mathbb{P}$; see [Kle20, Dur19]. The properties below hold in this generality.

Remark 1.14 (Properties of expected values). Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation.

(i) *Linearity*: for all $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$.

(ii) *Monotonicity*: if $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

(iii) *Jensen's inequality*: if $g \colon \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[|g(X)|] < \infty$, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.

(iv) *Markov's inequality*: for all $\varepsilon > 0$ and all non-decreasing functions $f \colon \mathbb{R} \to [0, \infty)$ with $f(\varepsilon) > 0$,

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

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

Proof of (iv). Since f is non-decreasing and non-negative, $f(\varepsilon)\mathbf{1}\{X \geq \varepsilon\} \leq f(X)$ pointwise on Ω. Monotonicity, which also holds for expectations in $[0, \infty]$ of non-negative random variables, and $\mathbb{E}[\mathbf{1}\{X \geq \varepsilon\}] = \mathbb{P}(X \geq \varepsilon)$ give the claim. □

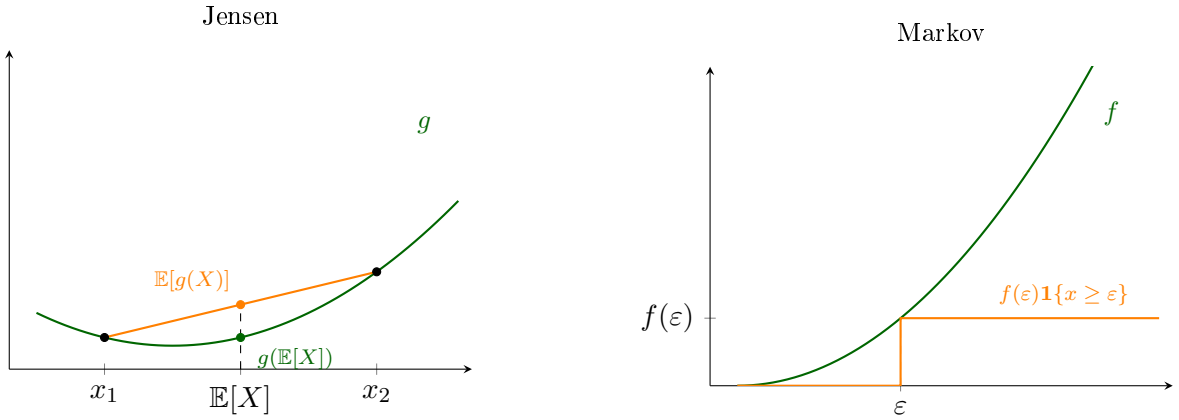

Figure 7: Left: for a convex function g and X taking the values x_1, x_2 with probability $1/2$ each, the chord lies above the graph, so $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. Right: the proof of Markov's inequality rests on the pointwise bound $f(\varepsilon)\mathbf{1}\{x \geq \varepsilon\} \leq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

Example 1.15.

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

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

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

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

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

Definition 1.16 (Variance)**.** Let X be a random variable with $\mathbb{E}[X^2] < \infty$ and $\mu := \mathbb{E}[X] \in \mathbb{R}$. Then

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

is called the *variance* of X, and $\sqrt{\mathrm{Var}[X]}$ 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 1.17 (Properties of variances). Let X be a random variable with $\mathbb{E}[X^2] < \infty$.

(i) $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$.

(ii) $\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$ for all $a, b \in \mathbb{R}$.

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

(iv) *Chebyshev's inequality*: for all $\varepsilon > 0$,

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

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

Example 1.18.

(i) *Binomial distribution.* In the setting of Example 1.15 (i), the same trick applied twice gives $\mathbb{E}[X(X-1)] = n(n-1)p^2$. Hence

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

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

$$\mathrm{Var}[X] = \int_0^\infty \Big(x - \frac{1}{\lambda}\Big)^2 \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

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

$$\mathbb{P}\Big(\bigcap_{i=1}^n X_i^{-1}(A_i)\Big) = \prod_{i=1}^n \mathbb{P}\big(X_i^{-1}(A_i)\big) \qquad \text{for all } A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R}).$$

Remark 1.20.

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

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

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

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

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

(iii) If $X_1, \dots, X_n$ are independent with finite expectation, then

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

Definition 1.21 (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

$$\operatorname{Cov}[X, Y] \coloneqq \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big].$$

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

$$\operatorname{Corr}[X, Y] \coloneqq \frac{\operatorname{Cov}[X, Y]}{\sqrt{\operatorname{Var}[X] \operatorname{Var}[Y]}} \in [-1, 1]$$

is called the *correlation* between X and Y.

The bound $|\operatorname{Corr}[X, Y]| \leq 1$ 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 1.22.

(i) Independent random variables are uncorrelated by Remark 1.20 (iii), but the reverse implication is wrong in general. For example, let X be uniformly distributed on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\operatorname{Cov}[X, Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0$, but $\mathbb{P}(X = 0, Y = 1) = 0 \neq (1/3) \cdot (2/3) = \mathbb{P}(X = 0)\mathbb{P}(Y = 1)$.

(ii) *Bienaymé's formula*: for pairwise uncorrelated random variables $X_1, \dots, X_n$ with finite variance,

$$\operatorname{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

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

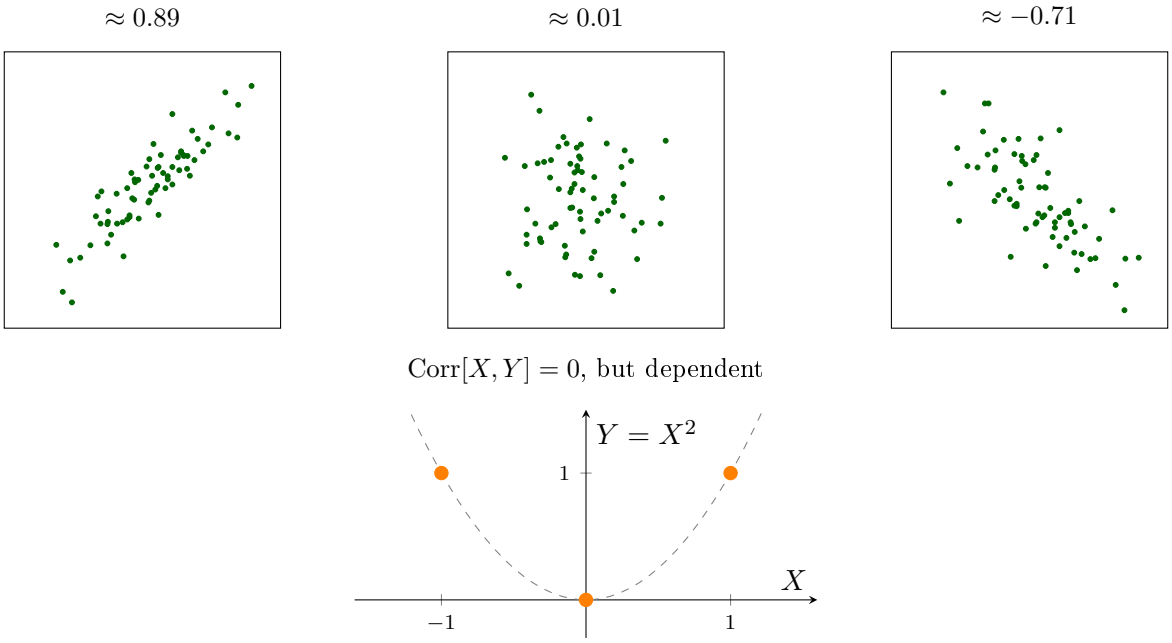

Figure 8: Top: 70 simulated samples of pairs (X, Y) with the indicated empirical correlation (numerical illustration). Correlation measures how well the points are described by a line. Bottom: the example of Remark 1.22 (i). The three equally likely points lie on a parabola, so Y is a function of X, 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 1.23 (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

$$\mathbb{E}[X|Y](\omega) := \sum_i x_i \, \rho\big(x_i|Y(\omega)\big) \qquad \text{for all } \omega \in \Omega \text{ with } \mathbb{P}\big(Y = Y(\omega)\big) > 0,$$

where the sum runs over all values x_i of X. The remaining ω form the event $\bigcup_y \{Y = y\}$, where the union runs over the at most countably many values y of Y with $\mathbb{P}(Y = y) = 0$; it is therefore a null set, and on it we set $\mathbb{E}[X|Y](\omega) := 0$.

Remark 1.24.

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

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

(ii) 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.

(iii) If X and Y are independent, then $\rho(x|y) = \rho_X(x)$ by Remark 1.20 (i), and hence $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

(iv) Similarly, for continuous random variables X, Y with joint density $f(x,y)$, one defines the conditional density "morally" as

$$f(x|y) := \frac{f(x,y)}{f_Y(y)} \qquad \text{for all } x \in \mathbb{R} \text{ and all } y \text{ with } f_Y(y) > 0,$$

where $f_Y(y) = \int_{\mathbb{R}} f(\tilde{x}, y)\,\mathrm{d}\tilde{x}$ is the density of Y. More precisely, $f(\cdot|\cdot)$ is a function such that, for all $A, B \in \mathcal{B}(\mathbb{R})$,

$$\int_A \Big(\int_B f(x|y)\,\mathrm{d}x \Big) f_Y(y)\,\mathrm{d}y = \int_A \int_B f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Accordingly, $\mathbb{E}[X|Y](\omega) := \int_{\mathbb{R}} x f\big(x|Y(\omega)\big)\,\mathrm{d}x$. 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].

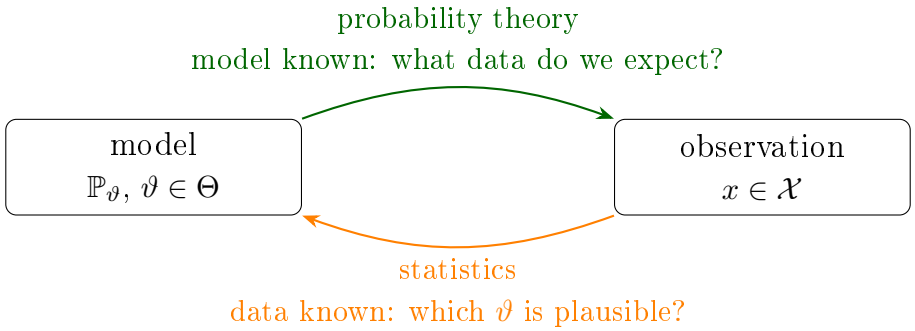

Figure 9: Probability theory and statistics look at the same objects in opposite directions.

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,

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

for all true (unknown) w and the "right" probability measure $\mathbb{P}_w$. 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:

$$\mathbb{P}_w(\{x\}) = \mathrm{Hyp}_{n;w,N-w}(\{x\}) \coloneqq \frac{\binom{w}{x}\binom{N-w}{n-x}}{\binom{N}{n}}, \qquad x = 0,\dots,n.$$

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

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

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

We need a decision-making criterion of the form

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

Hence, we must find c such that

(i) $\mathrm{Hyp}_{n;w,N-w}(\{x > c\})$ is small for $w \leq 500$, and

(ii) $\mathrm{Hyp}_{n;w,N-w}(\{x > c\})$ 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 $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta\colon \vartheta \in \Theta)$ consisting of a *sample space* $\mathcal{X} \neq \emptyset$, a σ-algebra $\mathcal{F}$ on $\mathcal{X}$, and a family $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ 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.

(i) If $\Theta \subset \mathbb{R}$, the statistical model is called a *one-parameter model*.

(ii) 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 $\mathbb{P}_\vartheta$ has a density ρ_ϑ. The class of discrete and continuous models is called the class of *standard models*.

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

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

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

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

as in Definition 1.9. The projection $X_i \colon \mathcal{X} \to E$ onto the i-th coordinate is a random variable, and $X_1, \ldots, X_n$ are independent with distribution Q_ϑ under every $\mathbb{P}_\vartheta$.

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

(i) Any random variable $S \colon (\mathcal{X}, \mathcal{F}) \to (\Sigma, \mathcal{S})$ is called a *statistic*.

(ii) Let $\tau \colon \Theta \to \Sigma$ be a mapping that associates to any $\vartheta \in \Theta$ a characteristic $\tau(\vartheta) \in \Sigma$. A statistic $T \colon \mathcal{X} \to \Sigma$ is then called an *estimator* for τ.

Remark 2.5.

(i) *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.

(ii) *Point estimators.* An important example is $\Sigma = \Theta$ and $\tau(\vartheta) = \vartheta$, so that an estimator $T \colon \mathcal{X} \to \Theta$ 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 $\mathbb{P}_\vartheta = \mathrm{U}_{[0,\vartheta]}^{\otimes n}$, 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:

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

Hence, player 1 uses

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

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

$$\tilde{T}_n \coloneqq \max\{X_1, \ldots, X_n\}.$$

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

$$\mathbb{P}_\vartheta\big(|T_n - \vartheta| \geq \varepsilon\big) \xrightarrow[n \to \infty]{} 0 \qquad \text{and} \qquad \mathbb{P}_\vartheta\big(|\tilde{T}_n - \vartheta| \geq \varepsilon\big) \xrightarrow[n \to \infty]{} 0.$$

For T_n this follows from Chebyshev's inequality and the variance computed below. Since $\tilde{T}_n \le \vartheta$, independence gives $\mathbb{P}_\vartheta(|\tilde{T}_n - \vartheta| \ge \varepsilon) = \mathbb{P}_\vartheta(X_1 \le \vartheta - \varepsilon)^n = (1 - \varepsilon/\vartheta)^n$ for $\varepsilon < \vartheta$, and this probability vanishes for $\varepsilon \ge \vartheta$.

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 $s \in [0, \vartheta]$,

$$\mathbb{P}_\vartheta(\tilde{T}_n \le s) = \mathbb{P}_\vartheta(X_1 \le s)^n = \left(\frac{s}{\vartheta}\right)^n =: F(s),$$

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

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

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 $\mathrm{Var}_\vartheta[X_1] = \vartheta^2/12$,

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

On the other hand,

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

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.

Definition 2.7 (Bias)**.** Let $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model and $\tau \colon \Theta \to \mathbb{R}$ a characteristic. An estimator $T \colon \mathcal{X} \to \mathbb{R}$ for τ with $\mathbb{E}_\vartheta[|T|] < \infty$ for all $\vartheta \in \Theta$ is called *unbiased* if

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

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

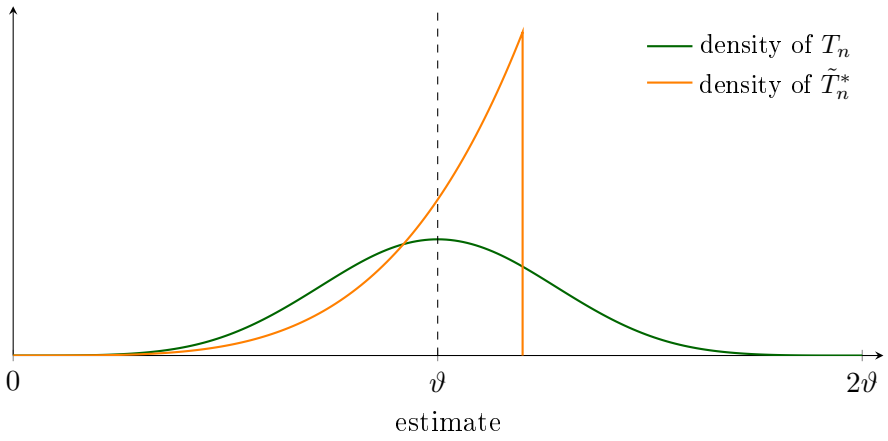

Figure 10: Exact densities of the two unbiased estimators of Example 2.6 for $n = 5$ (illustrative choice). Both have mean ϑ, but $\tilde{T}_n^*$ is concentrated much more closely around ϑ; its variance is $\vartheta^2/35$ compared with $\vartheta^2/15$ for T_n. The density of $\tilde{T}_n^*$ vanishes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

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*

$$m(\vartheta) := \int_{\mathbb{R}} x\, Q_\vartheta(\mathrm{d}x) \qquad \text{and} \qquad v(\vartheta) := \int_{\mathbb{R}} \big(x - m(\vartheta)\big)^2\, Q_\vartheta(\mathrm{d}x).$$

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.

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

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

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

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 $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$ be a standard model, so that every $\mathbb{P}_\vartheta$ has a density $x \mapsto \rho_\vartheta(x)$. The function

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

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

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

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

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

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

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

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.

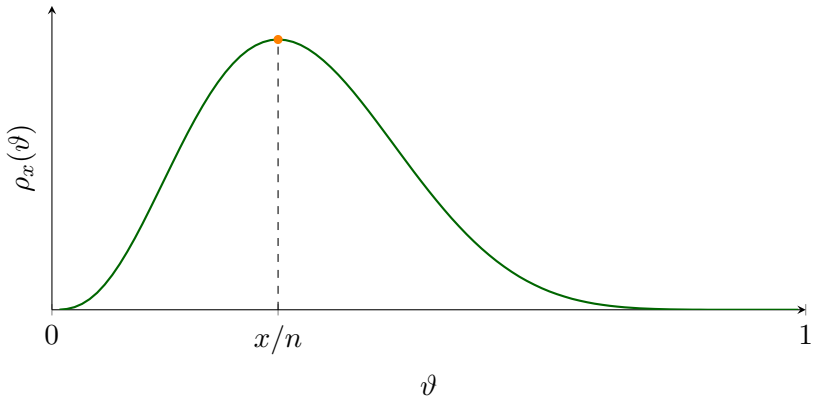

Figure 11: Likelihood function $\vartheta \mapsto \rho_x(\vartheta) = \mathrm{Bin}_{n,\vartheta}(x)$ from Example 2.10 for $n = 10$ and the observation $x = 3$. It is maximal at $T(x) = x/n = 0.3$.

Definition 2.11 (Best estimator)**.** Let $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model. An unbiased estimator T for the real-valued characteristic $\tau(\vartheta)$ with $\mathrm{Var}_\vartheta[T] < \infty$ for all $\vartheta \in \Theta$ is called *variance-minimising* or a *best estimator* if, for all other unbiased estimators S for τ,

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

Remark 2.12. Best estimators need not exist for a given characteristic τ. An important class of models in which they do exist are the one-parameter *exponential families*: under suitable regularity conditions, their canonical statistic T is a best estimator for the characteristic $\tau(\vartheta) \coloneqq \mathbb{E}_\vartheta[T]$; see, for instance, [Geo13, LC98].

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 $\mathbb{P}_\vartheta = \mathrm{Hyp}_{n;\vartheta,N-\vartheta}$ 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

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

We say that the *hypothesis* $H_0 \colon \vartheta \in \Theta_0$ is tested against the *alternative* $H_1 \colon \vartheta \in \Theta_1$. In our example, $\Theta_0 = \{0, \dots, 500\}$ (good quality) and $\Theta_1 = \{501, \dots, 10\,000\}$ (bad quality).

Step 3: choose a significance 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 $\varphi \colon \mathcal{X} \to [0,1]$ as follows. For the observation $x \in \mathcal{X}$, the value $\varphi(x)$ represents the degree to which we tend to the alternative:

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

For example,

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

where c is the smallest number for which the requirement of Step 3 is maintained by the non-randomised rule $\mathbf{1}\{x > c\}$, and $\eta \in [0,1]$ 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 $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model with $\Theta = \Theta_0 \dot\cup \Theta_1$, where Θ_0 is the null hypothesis and Θ_1 the alternative.

(i) Every statistic $\varphi \colon \mathcal{X} \to [0,1]$ 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 $\{x \in \mathcal{X} \colon \varphi(x) = 1\}$ is called the *critical region* of φ.

(ii) The number $\sup_{\vartheta \in \Theta_0} \mathbb{E}_\vartheta[\varphi]$ is called the *effective level* of φ. Let $\alpha \in (0,1)$. If the effective level of φ is at most α, then φ is called a test of *level* α.

(iii) The function $G_\varphi \colon \Theta \to [0,1]$, $G_\varphi(\vartheta) := \mathbb{E}_\vartheta[\varphi]$, 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 $\mathrm{Hyp}_{50;500,9500}(\{x > 4\}) \approx 0.103$ and $\mathrm{Hyp}_{50;500,9500}(\{x > 5\}) \approx 0.037$. Since $\mathrm{Hyp}_{n;\vartheta,N-\vartheta}(\{x > c\})$ is non-decreasing in ϑ, the supremum over Θ_0 is attained at $\vartheta = 500$, and Step 4 yields $c = 5$ and $\eta \approx 0.19$. 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.

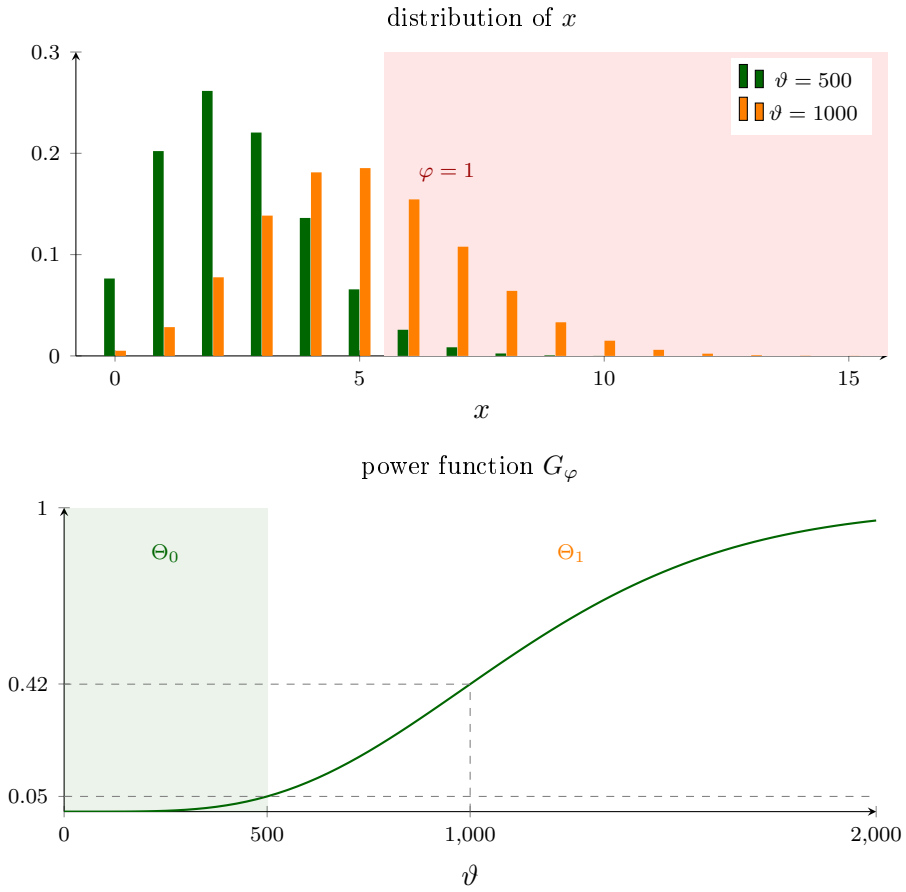

Figure 12: The test of Example 2.15 with $c = 5$ (numerical values). Top: hypergeometric densities of the number x of bad apples in the sample for $\vartheta = 500$ and $\vartheta = 1000$; the test decides for the alternative in the shaded region $x > 5$ and with probability $\eta \approx 0.19$ at $x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

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

$$X_k = \gamma_0 + \gamma_1 t_k + \sigma \xi_k, \qquad 1 \le k \le n. \tag{2.1}$$

Here, X_k is the measurement, $\gamma_0, \gamma_1 \in \mathbb{R}$ are the unknown *regression coefficients*, t_k is the *regression variable*, $\sigma > 0$ is an unknown *fluctuation parameter*, and ξ_k is a random measurement error. We assume that $\xi_1, \dots, \xi_n$ are uncorrelated with $\mathbb{E}[\xi_k] = 0$ and $\mathrm{Var}[\xi_k] = 1$,

and that the joint distribution of $(\xi_1, \dots, \xi_n)$ is fixed, that is, it does not depend on γ_0, γ_1 and σ. In vector notation, with $\boldsymbol{X} = (X_1, \dots, X_n)^\top$, $\mathbf{1} = (1, \dots, 1)^\top$, $\boldsymbol{t} = (t_1, \dots, t_n)^\top$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$, the model reads

$$\boldsymbol{X} = \gamma_0 \mathbf{1} + \gamma_1 \boldsymbol{t} + \sigma \boldsymbol{\xi}.$$

Then the distribution $\mathbb{P}_{\gamma_0,\gamma_1,\sigma}$ of $\boldsymbol{X}$ is determined by the parameters, and we consider the statistical model

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_{\gamma_0,\gamma_1,\sigma} \colon \gamma_0, \gamma_1 \in \mathbb{R},\ \sigma \in (0,\infty)\big).$$

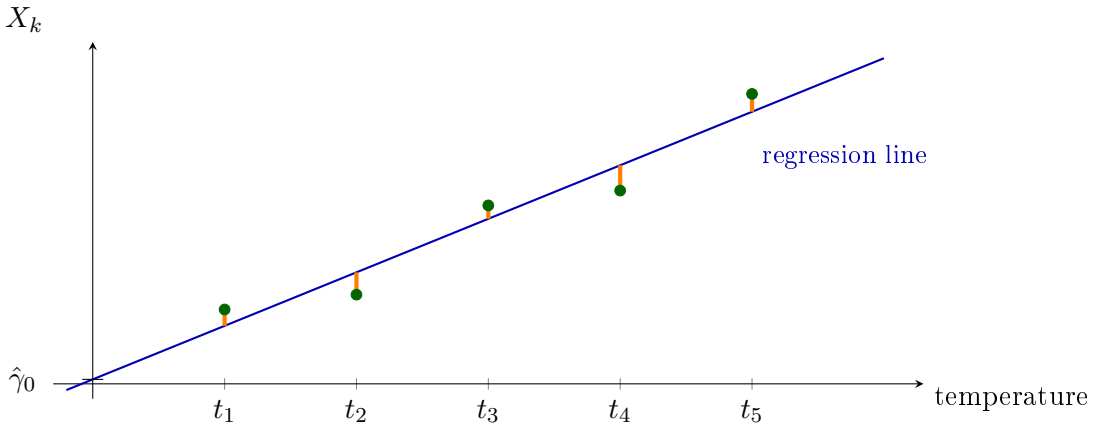

Figure 13: Measurements X_k at temperatures t_k (illustrative data) and the least-squares regression line $t \mapsto \hat{\gamma}_0 + \hat{\gamma}_1 t$. The line minimises the mean of the squared lengths of the vertical residuals $X_k - (\hat{\gamma}_0 + \hat{\gamma}_1 t_k)$, drawn in orange.

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

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

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

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

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

$$M(\boldsymbol{X}) = \gamma_0 + \gamma_1 M(\boldsymbol{t}), \tag{2.2}$$

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

We introduce the empirical variance and covariance

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

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

$$\gamma_1 V(\boldsymbol{t}) = K(\boldsymbol{t}, \boldsymbol{X}).$$

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

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

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 $\gamma \mapsto D_\gamma$ is a quadratic polynomial whose Hessian

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

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

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

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

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

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

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

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

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

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

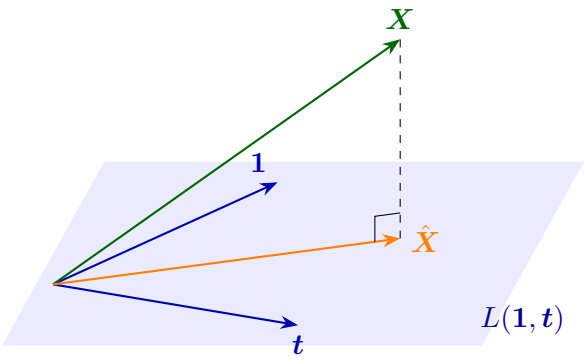

Figure 14: The least-squares fit $\hat{\boldsymbol{X}}$ is the orthogonal projection of $\boldsymbol{X}$ onto the plane spanned by $\mathbf{1}$ and $\boldsymbol{t}$.

Part II
Extended notes

This part complements Part I. We use the notation introduced there and refer to it instead of repeating definitions. Vectors and matrices are printed in bold, as in Section 2.5, and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_k x_k y_k$ denotes the Euclidean inner product with norm $|\boldsymbol{x}| := \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2}$.

3 Probability spaces and distributions on the real line

3.1 Further properties of probability measures

We collect consequences of Definition 1.1 that are used throughout.

Proposition 3.1 (Further properties of probability measures)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, A_2, \dots \in \mathcal{F}$.*

(i) *σ-subadditivity: $\mathbb{P}\big(\bigcup_{n=1}^{\infty} A_n\big) \le \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.*

(ii) Continuity from below*: if $A_1 \subset A_2 \subset \cdots$ and $A := \bigcup_{n=1}^{\infty} A_n$, then $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.*

(iii) Continuity from above*: if $A_1 \supset A_2 \supset \cdots$ and $A := \bigcap_{n=1}^{\infty} A_n$, then $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.*

(iv) Inclusion–exclusion formula*: for all $n \in \mathbb{N}$,*

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{|J|+1}\, \mathbb{P}\Big(\bigcap_{j \in J} A_j\Big).$$

Proof. Let $B_1 := A_1$ and $B_n := A_n \setminus \bigcup_{k<n} A_k$ for $n \ge 2$. The sets B_n are pairwise disjoint, $B_n \subset A_n$, and $\bigcup_{k \le n} B_k = \bigcup_{k \le n} A_k$ for all $n \in \mathbb{N} \cup \{\infty\}$. Hence (i) follows from σ-additivity and monotonicity (Remark 1.2). In the situation of (ii), $\mathbb{P}(A_n) = \sum_{k \le n} \mathbb{P}(B_k)$ converges to $\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}(A)$. For (iii), we apply (ii) to the increasing sequence A_n^{c} with union A^{c} and use $\mathbb{P}(A_n^{\mathrm{c}}) = 1 - \mathbb{P}(A_n)$. Finally, (iv) follows from Remark 1.2 (ii) 2) by induction over n. □

The binomial distribution counts successes in n independent trials, see Example 3.12 below. If the trials are many and successes are rare, it is well approximated by the Poisson distribution from Example 1.3.

Theorem 3.2 (Poisson approximation of the binomial distribution)**.** *Let $\lambda \in (0, \infty)$ and let $(p_n)_{n \in \mathbb{N}}$ be a sequence in $[0,1]$ with $n p_n \to \lambda$ as $n \to \infty$. Then, for every $k \in \mathbb{N}_0$,*

$$\mathrm{Bin}_{n,p_n}(k) \xrightarrow[n \to \infty]{} \mathrm{Poi}_{\lambda}(k).$$

Proof. Let $k \in \mathbb{N}_0$. Since $p_n \to 0$, we may assume $n \ge k$ and $p_n < 1$, and we write

$$\mathrm{Bin}_{n,p_n}(k) = \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{(n p_n)^k}{k!} \cdot (1-p_n)^n \cdot (1-p_n)^{-k}.$$

As $n \to \infty$, the first factor tends to 1, the second to $\lambda^k/k!$ and the last to 1. For the third factor, $\log(1-u) = -u + O(u^2)$ as $u \to 0$ gives $n \log(1-p_n) = -n p_n + O(n p_n^2) \to -\lambda$, and hence $(1-p_n)^n \to \mathrm{e}^{-\lambda}$. □

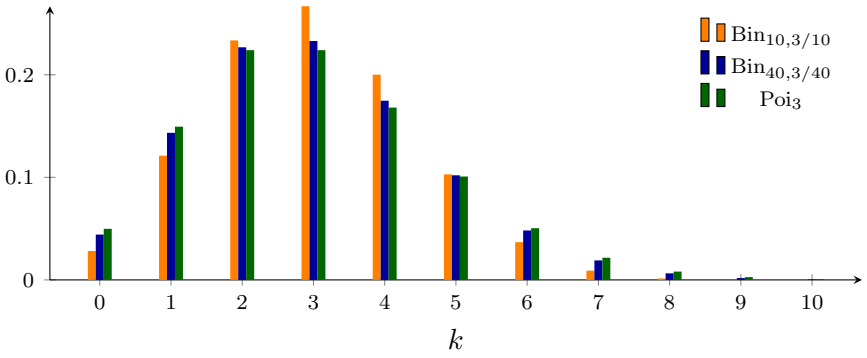

Figure 15: Poisson approximation (Theorem 3.2) for $\lambda = 3$: the binomial densities $\mathrm{Bin}_{n,3/n}$ for $n = 10$ and $n = 40$ approach the Poisson density Poi_3.

3.2 Distribution functions and quantiles

Probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are described by a single real function.

Definition 3.3 (Distribution function)**.** A function $F\colon \mathbb{R} \to [0,1]$ is called a *(cumulative) distribution function* if it is non-decreasing and right-continuous with

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

Proposition 3.4 (Distribution functions and probability measures)**.** *For every distribution function F, there is a unique probability measure ν_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that*

$$\nu_F\big((a,b]\big) = F(b) - F(a) \qquad \text{for all } a < b.$$

Conversely, for every probability measure ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the function $F(x) := \nu((-\infty, x])$ is a distribution function, and $\nu = \nu_F$.

Sketch of proof. For the converse statement, F is non-decreasing by monotonicity, right-continuous by continuity from above applied to $(-\infty, x + 1/n]$, and has the stated limits by continuity from above and from below (Proposition 3.1). Two probability measures that agree on all intervals $(a,b]$ agree, by continuity from below, on all half-lines $(-\infty, x]$. This family is closed under intersections and generates $\mathcal{B}(\mathbb{R})$, so the two measures agree on $\mathcal{B}(\mathbb{R})$, which proves uniqueness. The existence of ν_F is the construction of Lebesgue–Stieltjes measures; see [Kle20]. □

For a real random variable X, we call $F_X(x) := \mathbb{P}(X \le x)$ the *distribution function* of X. By Proposition 3.4, $\mathbb{P}_X = \nu_{F_X}$, so the distribution of X is determined by F_X.

Remark 3.5 (Jumps and atoms). Let F be a distribution function and $F(x-) := \lim_{y \uparrow x} F(y)$. Continuity from above applied to $(x - 1/n, x] \downarrow \{x\}$ gives $\nu_F(\{x\}) = F(x) - F(x-)$. Hence F is continuous at x if and only if $\nu_F(\{x\}) = 0$. The distribution function of a discrete random variable is a step function with jumps of height $\mathbb{P}(X = x)$ at the values x of X.

If f is a Lebesgue density as in Example 1.4, then $F(x) := \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ is a continuous distribution function with $\nu_F(B) = \int_B f(y)\,\mathrm{d}y$ for all $B \in \mathcal{B}(\mathbb{R})$. By the fundamental theorem of calculus, $F'(x) = f(x)$ at every point x where f is continuous. Conversely, if F is a continuous distribution function that is continuously differentiable outside a finite set, then F', extended arbitrarily to this finite set, is a density of ν_F. For example, the exponential distribution with parameter $\lambda > 0$ has the distribution function $F(x) = (1 - \mathrm{e}^{-\lambda x})\mathbf{1}\{x \ge 0\}$.

Definition 3.6 (Quantiles)**.** Let F be a distribution function and $q \in (0,1)$. The number

$$F^{-1}(q) := \inf\{x \in \mathbb{R} \colon F(x) \geq q\}$$

is called the *q-quantile* of F, and also of ν_F and of every random variable with distribution function F. The map $F^{-1} \colon (0,1) \to \mathbb{R}$ is called the *generalised inverse* of F, and $F^{-1}(1/2)$ is called the *median*.

The infimum is finite, because $F(x) \to 1 > q$ as $x \to \infty$ and $F(x) \to 0 < q$ as $x \to -\infty$. If F is continuous and strictly increasing, then F^{-1} is the usual inverse function.

Lemma 3.7 (Generalised inverse)**.** *Let F be a distribution function. For all $q \in (0,1)$ and $x \in \mathbb{R}$, we have $F^{-1}(q) \leq x$ if and only if $q \leq F(x)$.*

Proof. If $q \leq F(x)$, then x belongs to the set in Definition 3.6, so $F^{-1}(q) \leq x$. Conversely, choose $x_n \downarrow F^{-1}(q)$ with $F(x_n) \geq q$. Right-continuity gives $F(F^{-1}(q)) \geq q$, and if $F^{-1}(q) \leq x$, then monotonicity gives $F(x) \geq q$. □

Proposition 3.8 (Quantile transformation)**.** *Let F be a distribution function and let U be uniformly distributed on $[0,1]$. Then $X := F^{-1}(U)$, defined on the event $\{0 < U < 1\}$ of probability one, has distribution function F.*

Proof. By Lemma 3.7, $\mathbb{P}(X \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$ for all $x \in \mathbb{R}$. □

This is how random numbers with a prescribed distribution are generated from uniform random numbers. For instance, $-\lambda^{-1}\log(1-U)$ is exponentially distributed with parameter λ.

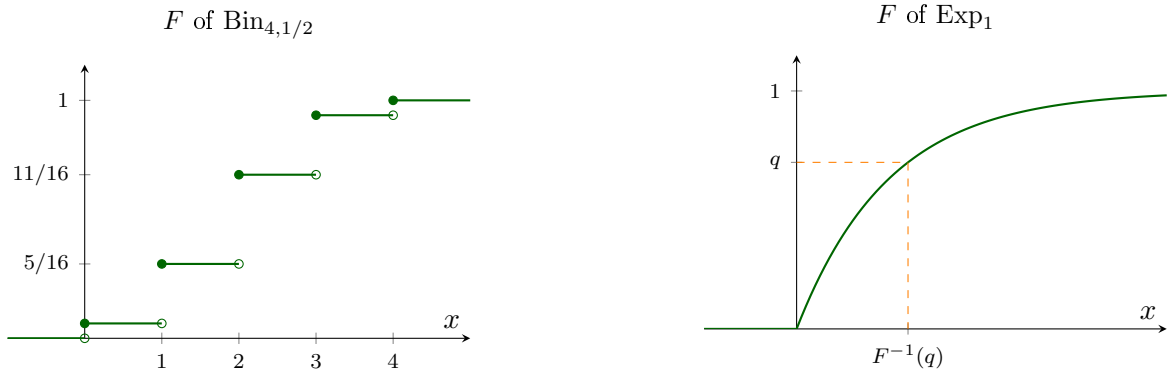

Figure 16: Distribution functions. Left: the step function of the binomial distribution $\mathrm{Bin}_{4,1/2}$; it is right-continuous, and the jump at x equals the probability of $\{x\}$ (Remark 3.5). Right: the continuous distribution function $F(x) = 1 - \mathrm{e}^{-x}$ of the exponential distribution and its q-quantile $F^{-1}(q) = -\log(1-q)$ for $q = 0.7$ (Definition 3.6).

3.3 Independence and conditioning

Independence of a family of events (Definition 1.5) is a stronger requirement than pairwise independence.

Example 3.9 (Pairwise independent, but not independent)**.** Consider two fair coin tosses, modelled by the Laplace experiment on $\Omega = \{0,1\}^2$, and the events

$$A = \{\text{first toss shows 1}\}, \qquad B = \{\text{second toss shows 1}\}, \qquad C = \{\text{both tosses agree}\}.$$

Each event has probability $1/2$, and each pairwise intersection equals $\{(1,1)\}$, which has probability $1/4$. Hence A, B, C are pairwise independent. They are not independent, since $\mathbb{P}(A \cap B \cap C) = 1/4 \neq 1/8 = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

If A and B are independent, then so are A and B^{c}, since $\mathbb{P}(A \cap B^{\mathrm{c}}) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B^{\mathrm{c}})$. More generally, in an independent family of events, any of the events may be replaced by their complements.

Remark 3.10 (Conditional probability on a subset). Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The probability measure $\mathbb{P}(\cdot|B)$ from Remark 1.8 is concentrated on B. Equivalently, it defines a probability measure on B with the *trace σ-algebra* $\mathcal{F}_B := \{A \cap B \colon A \in \mathcal{F}\}$. Iterating Definition 1.7 yields the *multiplication rule*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}),$$

valid whenever $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$.

Example 3.11 (Medical tests)**.** A disease affects 1% of a population. A test detects the disease with probability 0.99, and it gives a positive result for a healthy person with probability 0.05. Let B_1 be the event that a randomly chosen person is ill, $B_2 = B_1^{\mathrm{c}}$, and let A be the event of a positive test result. Bayes' formula (Remark 1.8) gives

$$\mathbb{P}(B_1|A) = \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.05 \cdot 0.99} = \frac{1}{6}.$$

Although the test is accurate, only one in six persons with a positive result is ill, because healthy persons are far more frequent than ill ones.

3.4 Products of finitely many probability spaces

Definition 1.9 extends to finitely many factors. For probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, \ldots, n$, the *product space* is $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n, \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)$, where the product σ-algebra is generated by the sets $A_1 \times \cdots \times A_n$ with $A_i \in \mathcal{F}_i$, and the product measure is the unique probability measure with

$$(\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mathbb{P}_i(A_i) \qquad \text{for all } A_i \in \mathcal{F}_i.$$

On the product space, the coordinate projections are independent and the i-th projection has distribution $\mathbb{P}_i$.

Example 3.12 (Bernoulli experiments and the binomial distribution)**.** Let $p \in [0,1]$ and consider the n-fold product of the Bernoulli distribution $\mathrm{Bin}_{1,p}$, that is, $\Omega = \{0,1\}^n$ with discrete density $\rho(\omega) = p^{S(\omega)}(1-p)^{n-S(\omega)}$, where $S(\omega) := \sum_{i=1}^{n} \omega_i$ is the number of successes. Since exactly $\binom{n}{k}$ outcomes satisfy $S(\omega) = k$,

$$\mathbb{P}(S = k) = \binom{n}{k} p^k (1-p)^{n-k} = \mathrm{Bin}_{n,p}(k), \qquad k = 0, \ldots, n.$$

Thus S is binomially distributed and is the sum of n independent Bernoulli variables $X_i(\omega) = \omega_i$ with $\mathbb{E}[X_i] = p$ and $\mathrm{Var}[X_i] = p(1-p)$. Linearity and Bienaymé's formula (Remark 1.22) give $\mathbb{E}[S] = np$ and $\mathrm{Var}[S] = np(1-p)$ without the computations of Examples 1.15 and 1.18.

4 Random variables, moments and random vectors

4.1 Measurability

Since the half-lines $(-\infty, x]$ generate $\mathcal{B}(\mathbb{R})$, a map $X\colon \Omega \to \mathbb{R}$ is a random variable if and only if $\{X \le x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$; see [Kle20]. A map $g\colon \mathbb{R}^n \to \mathbb{R}$ is called *measurable* if $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$ for all $B \in \mathcal{B}(\mathbb{R})$; continuous, monotone and piecewise continuous functions are measurable.

Proposition 4.1 (Operations on random variables)**.** *Let $X_1, X_2, \dots$ be real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) *For every measurable $g\colon \mathbb{R}^n \to \mathbb{R}$, the map $g(X_1, \dots, X_n)$ is a random variable. In particular, sums, products and linear combinations of random variables are random variables.*

(ii) *If $\sup_n X_n(\omega) < \infty$ for all $\omega \in \Omega$, then $\sup_n X_n$ is a random variable; the same holds for $\inf_n X_n$, $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ when they are finite. Moreover, $\{\omega\colon \lim_{n\to\infty} X_n(\omega) \text{ exists}\} \in \mathcal{F}$.*

Sketch of proof. For (i), $(X_1, \dots, X_n)$ is an $\mathbb{R}^n$-valued random variable, and compositions of measurable maps are measurable. For (ii), $\{\sup_n X_n \le x\} = \bigcap_n \{X_n \le x\} \in \mathcal{F}$; the other cases follow similarly; see [Kle20]. □

4.2 Expectation and moments

For a general random variable, $\mathbb{E}[X]$ is the Lebesgue integral $\int_\Omega X \,\mathrm{d}\mathbb{P}$ (Remark 1.13). The following results from integration theory are used repeatedly; see [Kle20].

Remark 4.2 (Convergence theorems and Fubini's theorem). Let $X, X_1, X_2, \dots$ be random variables with $X_n \to X$ pointwise on Ω.

(i) *Monotone convergence*: if $0 \le X_1 \le X_2 \le \cdots$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$ in $[0, \infty]$.

(ii) *Dominated convergence*: if $|X_n| \le Y$ for all n and a random variable Y with $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

(iii) *Fubini's theorem*: let $g\colon \mathbb{R}^2 \to \mathbb{R}$ be measurable. If $g \ge 0$ or $\int_{\mathbb{R}^2} |g(x,y)| \,\mathrm{d}(x,y) < \infty$, then
$$\int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g(x,y) \,\mathrm{d}x \Big) \,\mathrm{d}y = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} g(x,y) \,\mathrm{d}y \Big) \,\mathrm{d}x;$$
the same holds for sums and for products of probability measures.

Proposition 4.3 (Tail formula)**.** *Let X be a random variable with values in $\mathbb{N}_0$. Then*
$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}(X \ge n) \in [0, \infty].$$

Proof. Since all terms are non-negative, we may interchange the order of summation, which gives
$$\sum_{n=1}^{\infty} \mathbb{P}(X \ge n) = \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} k\, \mathbb{P}(X = k) = \mathbb{E}[X]. \qquad \square$$

For example, a geometrically distributed X (Example 1.3) satisfies $\mathbb{P}(X \ge n) = (1-p)^n$, and hence $\mathbb{E}[X] = (1-p)/p$. Similarly, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x) \,\mathrm{d}x$ for every random variable $X \ge 0$.

Proposition 4.4 (Moments). *Let X be a random variable.*

(i) *If $0 < p \le q$ and $\mathbb{E}[|X|^q] < \infty$, then $\mathbb{E}[|X|^p] < \infty$. In particular, every random variable with $\mathbb{E}[X^2] < \infty$ has a finite expectation and a finite variance.*

(ii) Markov's inequality for moments*: for all $\varepsilon > 0$ and $r > 0$,*

$$\mathbb{P}(|X| \ge \varepsilon) \le \varepsilon^{-r}\,\mathbb{E}[|X|^r].$$

Proof. For (i), we use $|x|^p \le 1 + |x|^q$ for all $x \in \mathbb{R}$ and the monotonicity of the expectation. Part (ii) follows as in the proof of Markov's inequality (Remark 1.14 (iv)): the pointwise bound $\varepsilon^r \mathbf{1}\{|X| \ge \varepsilon\} \le |X|^r$ and monotonicity of the expectation give the claim. □

Table 1 collects the distributions used in these notes. The moments follow from Remark 1.13; for instance, for the Poisson distribution, $\mathbb{E}[X(X-1)] = \sum_{k\ge 2} k(k-1)\mathrm{e}^{-\lambda}\lambda^k/k! = \lambda^2$, and hence $\mathrm{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$.

distribution	density	expectation	variance
$\mathrm{Bin}_{n,p}$, $p \in [0,1]$	$\binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \dots, n$	np	$np(1-p)$
$\mathrm{Hyp}_{n;w,N-w}$	$\binom{w}{k}\binom{N-w}{n-k}\big/\binom{N}{n}$, $k = 0, \dots, n$	nw/N	$n\frac{w}{N}\big(1-\frac{w}{N}\big)\frac{N-n}{N-1}$
Poi_λ, $\lambda > 0$	$\mathrm{e}^{-\lambda}\lambda^k/k!$, $k \in \mathbb{N}_0$	λ	λ
Geo_p, $p \in (0,1]$	$(1-p)^k p$, $k \in \mathbb{N}_0$	$(1-p)/p$	$(1-p)/p^2$
$\mathrm{U}_{[a,b]}$, $a < b$	$(b-a)^{-1}\mathbf{1}\{a \le x \le b\}$	$(a+b)/2$	$(b-a)^2/12$
Exp_λ, $\lambda > 0$	$\lambda \mathrm{e}^{-\lambda x}\mathbf{1}\{x \ge 0\}$	$1/\lambda$	$1/\lambda^2$
$\mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$	$(2\pi\sigma^2)^{-1/2}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2

Table 1: Common distributions with their discrete or Lebesgue densities, expectations and variances.

4.3 Random vectors

A *random vector* is an $\mathbb{R}^n$-valued random variable $\boldsymbol{X} = (X_1, \dots, X_n)^\top$, see Section 1.4; equivalently, $X_1, \dots, X_n$ are real random variables on the same probability space. Its distribution $\mathbb{P}_{\boldsymbol{X}}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called the *joint distribution* of $X_1, \dots, X_n$, and

$$F_{\boldsymbol{X}}(\boldsymbol{x}) := \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n), \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

is its *joint distribution function*. As in one dimension, $F_{\boldsymbol{X}}$ determines $\mathbb{P}_{\boldsymbol{X}}$. If $\boldsymbol{X}$ takes at most countably many values, its *joint discrete density* is $\rho_{\boldsymbol{X}}(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$, and the *marginal densities* are obtained by summing over the other coordinates,

$$\rho_{X_i}(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \rho_{\boldsymbol{X}}(x_1, \dots, x_n).$$

A measurable $f_{\boldsymbol{X}} \colon \mathbb{R}^n \to [0, \infty)$ is a *joint density* of $\boldsymbol{X}$ if $\mathbb{P}(\boldsymbol{X} \in B) = \int_B f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for all $B \in \mathcal{B}(\mathbb{R}^n)$. Then X_i has the *marginal density* obtained by integrating out the other coordinates,

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_{\boldsymbol{X}}(x_1, \dots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n.$$

Proposition 4.5 (Characterisations of independence)**.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ be a random vector. The following are equivalent.*

(i) *$X_1, \dots, X_n$ are independent in the sense of Definition 1.19.*

(ii) *$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

(iii) *$\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.*

If $\boldsymbol{X}$ is discrete, they are also equivalent to $\rho_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n \rho_{X_i}(x_i)$ for all $\boldsymbol{x}$. If every X_i has a density f_{X_i}, they are also equivalent to $\boldsymbol{x} \mapsto \prod_{i=1}^n f_{X_i}(x_i)$ being a joint density of $\boldsymbol{X}$. Moreover, if $X_1, \dots, X_n$ are independent and $g_1, \dots, g_n \colon \mathbb{R} \to \mathbb{R}$ are measurable, then $g_1(X_1), \dots, g_n(X_n)$ are independent.

Sketch of proof. Choosing $A_i = (-\infty, x_i]$ in Definition 1.19 shows that (i) implies (ii). Under (ii), both measures in (iii) agree on the products of half-lines, a family that is closed under intersections and generates $\mathcal{B}(\mathbb{R}^n)$, so they agree. Evaluating (iii) on $A_1 \times \cdots \times A_n$ gives (i). The last claim follows from $g_i(X_i)^{-1}(B) = X_i^{-1}(g_i^{-1}(B))$; see [Kle20] for details. □

For a measurable $g \colon \mathbb{R}^n \to \mathbb{R}$, the transformation formula of Remark 1.13 becomes

$$\mathbb{E}[g(\boldsymbol{X})] = \sum_{\boldsymbol{x}} g(\boldsymbol{x}) \rho_{\boldsymbol{X}}(\boldsymbol{x}) \qquad \text{respectively} \qquad \mathbb{E}[g(\boldsymbol{X})] = \int_{\mathbb{R}^n} g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

whenever $g \geq 0$ or the right-hand side is finite with $|g|$ in place of g.

Proposition 4.6 (Properties of the covariance)**.** *Let X, Y, Z be random variables with finite second moments and $a, b \in \mathbb{R}$.*

(i) $\operatorname{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, $\operatorname{Cov}[X, X] = \operatorname{Var}[X]$, *and* Cov *is symmetric and bilinear, for instance* $\operatorname{Cov}[aX + bY, Z] = a \operatorname{Cov}[X, Z] + b \operatorname{Cov}[Y, Z]$.

(ii) $\operatorname{Var}[X] = 0$ *if and only if* $X = \mathbb{E}[X]$ *almost surely, that is,* $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

(iii) Cauchy–Schwarz inequality*:* $\operatorname{Cov}[X, Y]^2 \leq \operatorname{Var}[X] \operatorname{Var}[Y]$.

(iv) *If* $\operatorname{Var}[X] \neq 0 \neq \operatorname{Var}[Y]$, *then* $|\operatorname{Corr}[X, Y]| = 1$ *if and only if there are* $a \neq 0$ *and* $b \in \mathbb{R}$ *with* $Y = aX + b$ *almost surely; in this case,* a *has the sign of* $\operatorname{Corr}[X, Y]$.

Proof. Part (i) follows from linearity of the expectation. For (ii), Chebyshev's inequality (Remark 1.17) gives $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = 0$ for all $\varepsilon > 0$ if $\operatorname{Var}[X] = 0$, and continuity from below yields $\mathbb{P}(X \neq \mathbb{E}[X]) = 0$; the converse is immediate. For (iii) and (iv), we may assume $\operatorname{Var}[X] > 0$ and consider

$$h(t) \coloneqq \operatorname{Var}[Y - tX] = \operatorname{Var}[Y] - 2t \operatorname{Cov}[X, Y] + t^2 \operatorname{Var}[X] \geq 0, \qquad t \in \mathbb{R}.$$

A quadratic polynomial that is non-negative has a non-positive discriminant, which is (iii). Equality in (iii) holds if and only if h has a zero t_0, that is, if and only if $Y - t_0 X$ is almost surely constant by (ii). In this case, $t_0 = \operatorname{Cov}[X, Y] / \operatorname{Var}[X]$, which is non-zero and has the sign of the correlation. □

For a random vector $\boldsymbol{X}$ whose coordinates have finite second moments, we write $\mathbb{E}[\boldsymbol{X}] \coloneqq (\mathbb{E}[X_i])_i$ and call

$$\operatorname{Cov}[\boldsymbol{X}] \coloneqq \big(\operatorname{Cov}[X_i, X_j]\big)_{i,j=1}^n$$

the *covariance matrix* of $\boldsymbol{X}$.

Proposition 4.7 (Affine transformations of random vectors)**.** *Let $\boldsymbol{X}$ be a random vector with values in $\mathbb{R}^n$ whose coordinates have finite second moments, and let $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.*

(i) $\operatorname{Cov}[\boldsymbol{X}]$ *is symmetric and positive semidefinite, since* $\langle \boldsymbol{\lambda}, \operatorname{Cov}[\boldsymbol{X}]\boldsymbol{\lambda}\rangle = \operatorname{Var}[\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle] \geq 0$ *for all* $\boldsymbol{\lambda} \in \mathbb{R}^n$.

(ii) $\mathbb{E}[\boldsymbol{B}\boldsymbol{X} + \boldsymbol{b}] = \boldsymbol{B}\,\mathbb{E}[\boldsymbol{X}] + \boldsymbol{b}$ *and* $\operatorname{Cov}[\boldsymbol{B}\boldsymbol{X} + \boldsymbol{b}] = \boldsymbol{B}\operatorname{Cov}[\boldsymbol{X}]\boldsymbol{B}^\top$.

Proof. Both parts follow from the linearity of the expectation and the bilinearity of the covariance (Proposition 4.6). □

The distribution of a sum of independent random variables is the *convolution* of their distributions.

Proposition 4.8 (Convolution)**.** *Let X and Y be independent random variables.*

(i) *If X and Y have densities f_X and f_Y, then $X + Y$ has the density*

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\,\mathrm{d}y = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\,\mathrm{d}x, \qquad z \in \mathbb{R}.$$

(ii) *If X and Y are discrete, then $\mathbb{P}(X+Y=z) = \sum_x \mathbb{P}(X=x)\,\mathbb{P}(Y=z-x)$ for all $z \in \mathbb{R}$.*

Proof. For (i), Proposition 4.5 gives the joint density $f_X(x)f_Y(y)$, and Fubini's theorem and the substitution $u = x + y$ yield

$$\mathbb{P}(X+Y \leq z) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}\{x+y \leq z\} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{z} \Big(\int_{\mathbb{R}} f_X(u-y) f_Y(y)\,\mathrm{d}y \Big)\,\mathrm{d}u.$$

Part (ii) follows from $\{X+Y=z\} = \bigcup_x \{X=x, Y=z-x\}$ and independence. □

Example 4.9 (Convolutions)**.**

(i) If U_1, U_2 are independent and uniformly distributed on $[0,1]$, then $U_1 + U_2$ has the triangular density $f(z) = z$ for $z \in [0,1]$, $f(z) = 2 - z$ for $z \in [1,2]$, and $f(z) = 0$ otherwise.

(ii) If X, Y are independent and exponentially distributed with parameter λ, then $X + Y$ has the density $f(z) = \int_0^z \lambda \mathrm{e}^{-\lambda x} \lambda \mathrm{e}^{-\lambda(z-x)}\,\mathrm{d}x = \lambda^2 z \mathrm{e}^{-\lambda z}$ for $z \geq 0$.

(iii) If $X \sim \mathrm{Poi}_\lambda$ and $Y \sim \mathrm{Poi}_\mu$ are independent, then the binomial theorem gives

$$\mathbb{P}(X+Y=n) = \sum_{k=0}^{n} \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,\mathrm{e}^{-\mu}\frac{\mu^{n-k}}{(n-k)!} = \mathrm{e}^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!},$$

so $X + Y \sim \mathrm{Poi}_{\lambda+\mu}$.

(iv) Sums of independent normally distributed random variables are normally distributed; see Proposition 6.1.

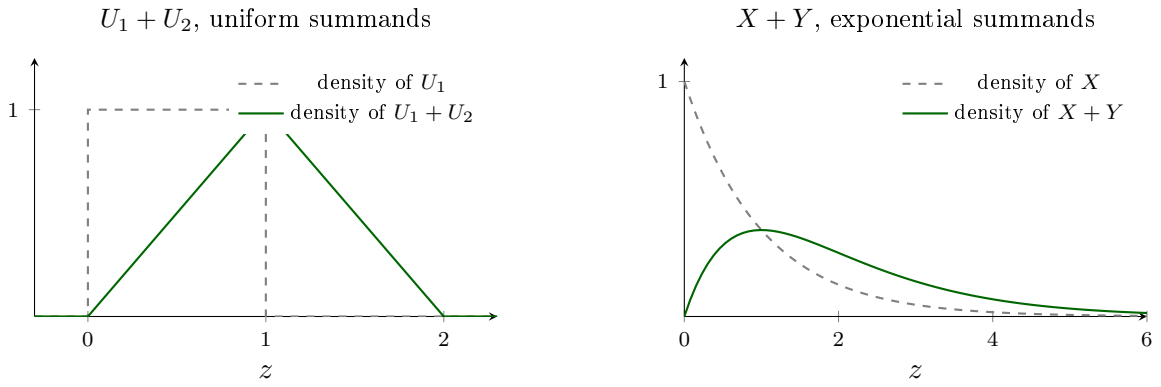

Figure 17: Convolutions from Example 4.9. Left: the sum of two independent uniform random variables on $[0, 1]$ has a triangular density. Right: the sum of two independent exponential random variables with parameter $\lambda = 1$ has the density $z\mathrm{e}^{-z}$, $z \geq 0$.

4.4 Conditional expectations

We extend Section 1.7 for discrete random variables X, Y with $\mathbb{E}[|X|] < \infty$. For a value y of Y with $\mathbb{P}(Y = y) > 0$, we write $\mathbb{E}[X|Y = y] := \sum_i x_i \, \rho(x_i|y)$, so that $\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = Y(\omega)]$.

Proposition 4.10 (Properties of conditional expectations)**.** *Let X, X', Y be discrete random variables with $\mathbb{E}[|X|], \mathbb{E}[|X'|] < \infty$, and let $a, b \in \mathbb{R}$. The following hold almost surely.*

(i) Linearity*:* $\mathbb{E}[aX + bX'|Y] = a\mathbb{E}[X|Y] + b\mathbb{E}[X'|Y]$.

(ii) Taking out what is known*:* $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$ *for every* $g\colon \mathbb{R} \to \mathbb{R}$ *with* $\mathbb{E}[|g(Y)X|] < \infty$*; in particular,* $\mathbb{E}[g(Y)|Y] = g(Y)$.

(iii) Defining property*:* $\mathbb{E}\big[\mathbb{E}[X|Y]\, h(Y)\big] = \mathbb{E}[X\, h(Y)]$ *for every bounded* $h\colon \mathbb{R} \to \mathbb{R}$*; for* $h = 1$*, this is Remark 1.24 (i).*

(iv) Independence*: if X and Y are independent, then* $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

(v) Best prediction*: if* $\mathbb{E}[X^2] < \infty$*, then* $\mathbb{E}\big[(X - \mathbb{E}[X|Y])^2\big] \leq \mathbb{E}\big[(X - h(Y))^2\big]$ *for every* $h\colon \mathbb{R} \to \mathbb{R}$.

Proof. Parts (i), (ii) and (iv) follow by direct computation from Definition 1.23. For (iii), we sum over the values y of Y with $\mathbb{P}(Y = y) > 0$ and obtain

$$\mathbb{E}\big[\mathbb{E}[X|Y]h(Y)\big] = \sum_y h(y) \sum_i x_i \, \rho(x_i|y)\mathbb{P}(Y = y) = \sum_y \sum_i h(y)x_i \, \mathbb{P}(X = x_i, Y = y),$$

which equals $\mathbb{E}[Xh(Y)]$. For (v), we write

$$\mathbb{E}\big[(X - h(Y))^2\big] = \sum_y \mathbb{P}(Y = y) \sum_i \big(x_i - h(y)\big)^2 \rho(x_i|y).$$

For fixed y, the inner sum is the mean squared deviation of the probability distribution $\rho(\cdot|y)$ from $h(y)$. This distribution has a finite second moment, since $\sum_i x_i^2 \rho(x_i|y) \leq \mathbb{E}[X^2]/\mathbb{P}(Y = y)$. By Remark 1.17 (iii), the inner sum is minimal for $h(y) = \mathbb{E}[X|Y = y]$. □

Part (v) generalises Remark 1.17 (iii): the best constant prediction of X is $\mathbb{E}[X]$, and the best prediction of X by a function of Y is $\mathbb{E}[X|Y]$.

Example 4.11 (Two dice)**.** Let X_1, X_2 be the results of two fair dice, as in Example 1.6, and $S := X_1 + X_2$. The pairs (X_1, S) and (X_2, S) have the same distribution, so $\mathbb{E}[X_1|S] = \mathbb{E}[X_2|S]$. By (i) and (ii), $\mathbb{E}[X_1|S] + \mathbb{E}[X_2|S] = \mathbb{E}[S|S] = S$, and hence $\mathbb{E}[X_1|S] = S/2$.

For continuous random variables, the conditional density of Remark 1.24 (iv) leads to the same properties; the general theory is developed in [Kle20, Dur19]. An explicit continuous example is the normal distribution in Proposition 6.6.

5 Limit theorems

The notes of Part I are restricted to finitely many random variables. Statistics, however, relies on the behaviour of estimators as the sample size grows, which requires notions of convergence for random variables.

5.1 Modes of convergence

Definition 5.1 (Convergence of random variables)**.** Let $X, X_1, X_2, \dots$ be real random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

(i) X_n converges to X *in probability*, written $X_n \xrightarrow{\mathbb{P}} X$, if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$.

(ii) X_n converges to X *almost surely*, written $X_n \xrightarrow{\text{a.s.}} X$, if $\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1$.

(iii) For $p \geq 1$, X_n converges to X *in* L^p, written $X_n \xrightarrow{L^p} X$, if $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$.

(iv) X_n converges to X *in distribution*, written $X_n \xrightarrow{\text{d}} X$, if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for every $x \in \mathbb{R}$ at which F_X is continuous.

Convergence in distribution only involves the distributions of the random variables, which may then be defined on different probability spaces. We also write $X_n \xrightarrow{\text{d}} \nu$ for a probability measure ν on $\mathbb{R}$, if the distribution functions converge to that of ν. Throughout, *almost surely* (a.s.) means outside an event of probability zero.

Theorem 5.2 (Relations between the modes of convergence)**.** *Let $X, X_1, X_2, \dots$ be real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$.*

(ii) *If $X_n \xrightarrow{L^p} X$ for some $p \geq 1$, then $X_n \xrightarrow{\mathbb{P}} X$.*

(iii) *If $1 \leq p \leq q$ and $X_n \xrightarrow{L^q} X$, then $X_n \xrightarrow{L^p} X$.*

(iv) *If $X_n \xrightarrow{\mathbb{P}} X$, then there is a subsequence $(n_k)_{k\in\mathbb{N}}$ with $X_{n_k} \xrightarrow{a.s.} X$.*

(v) *If $X_n \xrightarrow{\mathbb{P}} X$ and $|X_n| \leq K$ for all n and a constant $K \geq 0$, then $X_n \xrightarrow{L^p} X$ for all $p \geq 1$.*

(vi) *If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{\text{d}} X$.*

Proof. For (i), the events $A_n := \{\sup_{m\geq n} |X_m - X| > \varepsilon\}$ decrease to a subset of the null event $\{X_m \not\to X\}$, so continuity from above (Proposition 3.1) gives $\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(A_n) \to 0$. Part (ii) follows from Proposition 4.4 (ii): $\mathbb{P}(|X_n - X| > \varepsilon) \leq \varepsilon^{-p}\mathbb{E}[|X_n - X|^p]$. For (iii), Jensen's inequality for the convex function $x \mapsto |x|^{q/p}$ gives $\mathbb{E}[|X_n - X|^p]^{q/p} \leq \mathbb{E}[|X_n - X|^q]$.

For (v), we first note that $|X| \le K$ almost surely by (iv). Hence, for every $\varepsilon \in (0,1)$, $\mathbb{E}[|X_n - X|^p] \le \varepsilon^p + (2K)^p\,\mathbb{P}(|X_n - X| > \varepsilon)$, and the right-hand side is smaller than $2\varepsilon^p$ for large n. For (iv) and (vi), see [Kle20]. □

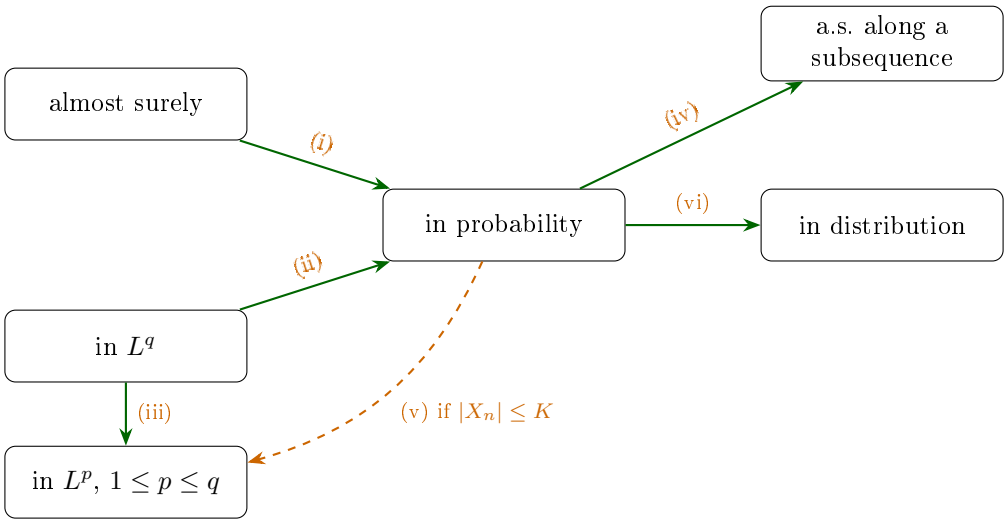

Figure 18: Relations between the modes of convergence from Theorem 5.2. Solid arrows hold in general, the dashed arrow under the stated boundedness condition. None of the solid arrows can be reversed in general (Example 5.5).

The following lemma is the main tool for proving almost sure convergence. For events $A_1, A_2, \dots$, the event $\limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$ consists of all outcomes that lie in infinitely many A_m.

Lemma 5.3 (Borel–Cantelli). *Let $A_1, A_2, \dots \in \mathcal{F}$.*

(i) *If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$.*

(ii) *If $A_1, A_2, \dots$ are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(\limsup_n A_n) = 1$.*

Proof. For (i), $\mathbb{P}(\limsup_n A_n) \le \mathbb{P}\big(\bigcup_{m\ge n} A_m\big) \le \sum_{m\ge n} \mathbb{P}(A_m)$ for all n, and the right-hand side tends to zero. For (ii), independence of the complements and $1 - x \le \mathrm{e}^{-x}$ give $\mathbb{P}\big(\bigcap_{m=n}^{N} A_m^{\mathrm{c}}\big) = \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) \le \exp\big(-\sum_{m=n}^{N} \mathbb{P}(A_m)\big) \to 0$ as $N \to \infty$. Hence $\mathbb{P}\big(\bigcup_{m\ge n} A_m\big) = 1$ for every n, and continuity from above gives the claim. □

Corollary 5.4 (Fast convergence in probability). *If $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $X_n \xrightarrow{a.s.} X$.*

Proof. For $\varepsilon = 1/k$, Lemma 5.3 (i) shows that the event $N_k := \limsup_n \{|X_n - X| > 1/k\}$ has probability zero. Outside the null set $\bigcup_k N_k$, we have $|X_n - X| \le 1/k$ for all large n and every k. □

Example 5.5 (The reverse implications fail).

(i) Let $X_1, X_2, \dots$ be independent with $\mathbb{P}(X_n = 1) = 1/n = 1 - \mathbb{P}(X_n = 0)$. Then $X_n \xrightarrow{\mathbb{P}} 0$, but $\sum_n \mathbb{P}(X_n = 1) = \infty$, so $X_n = 1$ for infinitely many n almost surely by Lemma 5.3 (ii). Hence X_n does not converge almost surely.

(ii) Let U be uniformly distributed on $[0,1]$ and $X_n := n\mathbf{1}\{U < 1/n\}$. Then $X_n \to 0$ almost surely, but $\mathbb{E}[|X_n|] = 1$ for all n, so there is no convergence in L^1.

(iii) Let X take the values ± 1 with probability $1/2$ each, and $X_n := -X$. All X_n have the distribution of X, so $X_n \xrightarrow{\mathrm{d}} X$, but $|X_n - X| = 2$.

(iv) Let U be as in (ii), $1 \le p < q$, and $X_n := n^{1/q}\mathbf{1}\{U < 1/n\}$. Then $\mathbb{E}[|X_n|^p] = n^{p/q-1} \to 0$, but $\mathbb{E}[|X_n|^q] = 1$, so $X_n \xrightarrow{L^p} 0$, but not in L^q.

(v) Let $X_n := 1$ for odd n and $X_n := 0$ for even n. The subsequence of even indices converges almost surely, but $(X_n)_{n\in\mathbb{N}}$ does not converge in probability.

5.2 Laws of large numbers

Definition 5.6 (Laws of large numbers)**.** A sequence $X_1, X_2, \dots$ of random variables with finite expectations satisfies the *weak law of large numbers* if

$$\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \xrightarrow{\mathbb{P}} 0,$$

and the *strong law of large numbers* if this convergence holds almost surely.

Theorem 5.7 (Weak law of large numbers)**.** *Let $X_1, X_2, \dots$ be pairwise uncorrelated random variables with finite variances such that*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}[X_i] = 0. \tag{5.1}$$

Then $X_1, X_2, \dots$ satisfy the weak law of large numbers. Condition (5.1) holds, in particular, if $\operatorname{Var}[X_i] \le C$ for all i and a constant $C \in (0,\infty)$, for instance if the X_i are identically distributed, and if $\sum_{i=1}^{\infty}\operatorname{Var}[X_i]/i^2 < \infty$.

Proof. By Chebyshev's inequality (Remark 1.17) and Bienaymé's formula (Remark 1.22), for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)\Big| \ge \varepsilon\Big) \le \frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n}\operatorname{Var}[X_i] \xrightarrow[n\to\infty]{} 0.$$

Bounded variances give the bound $C/(n\varepsilon^2)$. The last sufficient condition implies (5.1) by Kronecker's lemma; see [Kle20]. □

The proof also gives a rate: under bounded variances, the probability of a deviation larger than ε is at most $C/(n\varepsilon^2)$.

Theorem 5.8 (Strong law of large numbers)**.** *Let $X_1, X_2, \dots$ be independent random variables with finite expectations. They satisfy the strong law of large numbers if either they are identically distributed, or they have finite variances with $\sum_{n=1}^{\infty}\operatorname{Var}[X_n]/n^2 < \infty$.*

For proofs, see [Kle20, Dur19]. For independent and identically distributed (i.i.d.) random variables, Theorem 5.8 states that the empirical mean

$$M_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges almost surely to $\mathbb{E}[X_1]$. Applied to indicators $X_i = \mathbf{1}_{A_i}$ of independent events with the same probability, it shows that relative frequencies converge to probabilities, which justifies the frequency interpretation of probability.

Example 5.9 (Monte Carlo integration)**.** Let $g\colon [0,1]^d \to \mathbb{R}$ be a measurable function such that g^2 has a finite integral over $[0,1]^d$. We want to compute

$$\mu := \int_{[0,1]^d} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Let $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots$ be independent and uniformly distributed on $[0,1]^d$, that is, with density $\mathbf{1}\{\boldsymbol{x} \in [0,1]^d\}$. Then $\mathbb{E}[g(\boldsymbol{U}_1)] = \mu$, and the strong law of large numbers gives

$$\frac{1}{n}\sum_{k=1}^{n} g(\boldsymbol{U}_k) \xrightarrow{\text{a.s.}} \mu.$$

By Bienaymé's formula, the root mean squared error is

$$\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{k=1}^{n} g(\boldsymbol{U}_k) - \mu\Big)^2\Big]^{1/2} = \frac{\sigma}{\sqrt{n}}, \qquad \text{where } \sigma^2 := \mathrm{Var}[g(\boldsymbol{U}_1)] = \int_{[0,1]^d} g(\boldsymbol{x})^2\,\mathrm{d}\boldsymbol{x} - \mu^2.$$

The rate $n^{-1/2}$ does not depend on the dimension d, which is why Monte Carlo methods are used for high-dimensional integrals.

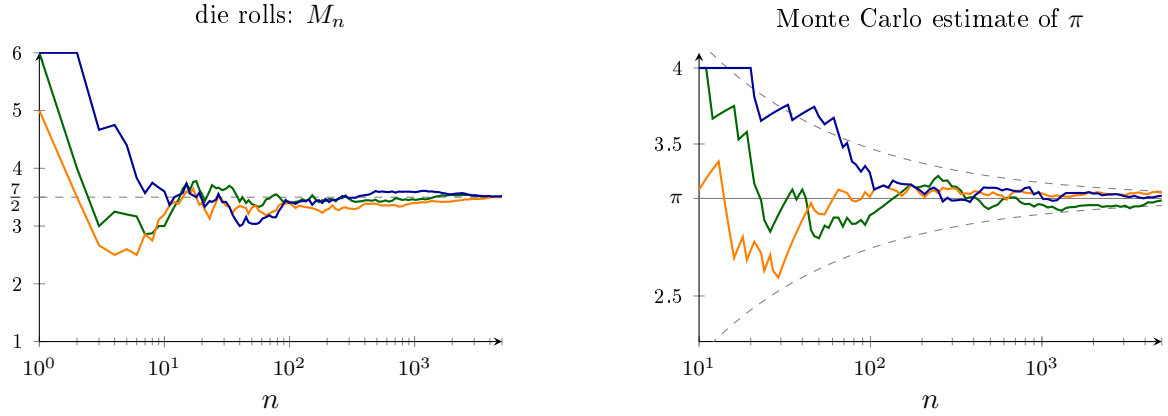

Figure 19: Laws of large numbers (simulated). Left: three paths of the empirical mean M_n of fair die rolls, converging to $\mathbb{E}[X_1] = 7/2$. Right: three paths of the Monte Carlo estimate $\frac{1}{n}\sum_{k=1}^{n} g(U_k)$ with $g(x) = 4 \cdot \mathbf{1}\{x_1^2 + x_2^2 \le 1\}$ on $[0,1]^2$; the dashed curves show $\pi \pm 2\sigma/\sqrt{n}$, see Example 5.9.

5.3 The central limit theorem

The law of large numbers says that $M_n - \mu$ tends to zero. The central limit theorem describes the fluctuations of order $n^{-1/2}$ and explains the prominent role of the normal distribution. We write

$$\Phi(y) := \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u, \qquad y \in \mathbb{R},$$

for the distribution function of the standard normal distribution $\mathcal{N}(0,1)$; see Section 6.

Theorem 5.10 (Central limit theorem of Lindeberg and Lévy)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathrm{Var}[X_1] \in (0,\infty)$. Then*

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

In particular, $\sqrt{n}(M_n - \mu)/\sigma \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

The standardised sum has expectation 0 and variance 1 by Bienaymé's formula; the theorem identifies its limiting distribution, whatever the distribution of X_1. For a proof, see [Kle20, Dur19]. Under the additional assumption $\mathbb{E}[|X_1|^3] < \infty$, the *Berry–Esseen theorem* bounds the supremum above by $C\,\mathbb{E}[|X_1 - \mu|^3]/(\sigma^3\sqrt{n})$ with a universal constant C; see [Dur19].

Corollary 5.11 (De Moivre–Laplace)**.** *Let $p \in (0,1)$ and $S_n \sim \mathrm{Bin}_{n,p}$. Then, for all $a < b$,*

$$\mathbb{P}\Big(a < \frac{S_n - np}{\sqrt{np(1-p)}} \le b\Big) \xrightarrow[n\to\infty]{} \Phi(b) - \Phi(a).$$

Proof. By Example 3.12, S_n has the distribution of a sum of n i.i.d. Bernoulli variables with expectation p and variance $p(1-p)$, and Theorem 5.10 applies. □

Example 5.12 (Coin tossing)**.** For 1000 tosses of a fair coin, the number S of heads satisfies, by Corollary 5.11,

$$\mathbb{P}(S \ge 530) = \mathbb{P}\Big(\frac{S - 500}{\sqrt{250}} \ge 1.897\Big) \approx 1 - \Phi(1.897) \approx 0.029.$$

The exact binomial probability is 0.0310. The approximation improves to $1-\Phi(29.5/\sqrt{250}) \approx 0.0310$ with the *continuity correction*, which replaces the threshold 530 by 529.5, the midpoint between the integers 529 and 530.

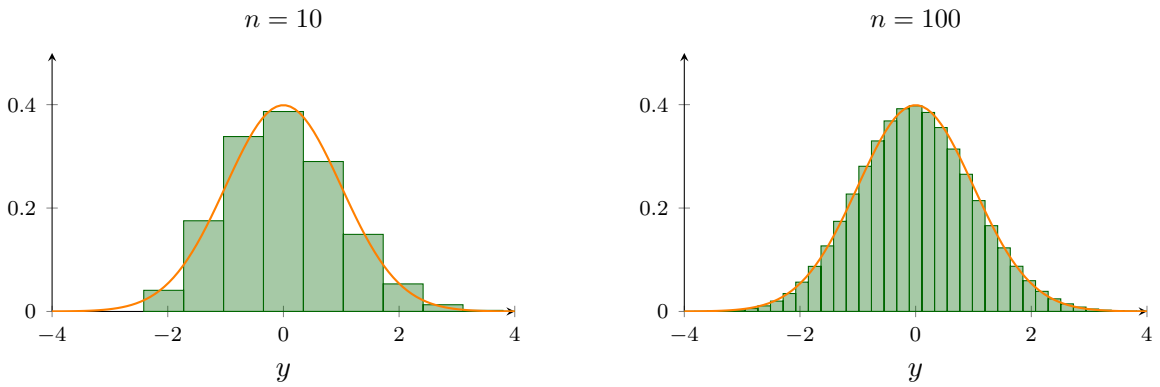

Figure 20: Central limit theorem for $S_n \sim \mathrm{Bin}_{n,0.3}$: histogram of the standardised variable $(S_n - np)/\sqrt{np(1-p)}$, with bars of width $1/\sqrt{np(1-p)}$ and area equal to the binomial probabilities, compared with the standard normal density (orange).

In statistical applications, unknown parameters in the standardisation are replaced by estimators. The following result shows that this does not change the limit.

Proposition 5.13 (Slutsky's theorem)**.** *Let $Y_n \xrightarrow{\mathrm{d}} Y$ and $Z_n \xrightarrow{\mathbb{P}} c$ for a constant $c \in \mathbb{R}$. Then $Y_n + Z_n \xrightarrow{\mathrm{d}} Y + c$ and $Y_n Z_n \xrightarrow{\mathrm{d}} cY$. Moreover, if $Z_n \xrightarrow{\mathbb{P}} c$ and $h\colon \mathbb{R} \to \mathbb{R}$ is continuous at c, then $h(Z_n) \xrightarrow{\mathbb{P}} h(c)$.*

For a proof, see [Kle20].

6 Normal distributions and their relatives

6.1 One-dimensional normal distributions

For $\mu \in \mathbb{R}$ and $\sigma > 0$, the *normal distribution* $\mathcal{N}(\mu, \sigma^2)$ has the density

$$f_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big), \qquad x \in \mathbb{R},$$

from Example 1.4; $\mathcal{N}(0,1)$ is called the *standard normal distribution*, with distribution function Φ and density $f_{0,1}$. That $f_{0,1}$ integrates to one follows from $\big(\int_{\mathbb{R}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x\big)^2 = \int_{\mathbb{R}^2} \mathrm{e}^{-(x^2+y^2)/2}\,\mathrm{d}(x,y) = 2\pi$, using polar coordinates. For $q \in (0,1)$, we write $z_q := \Phi^{-1}(q)$ for the q-quantile of $\mathcal{N}(0,1)$.

Proposition 6.1 (Properties of normal distributions)**.**

(i) *If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a \neq 0$, $b \in \mathbb{R}$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. In particular, $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $(X - \mu)/\sigma \sim \mathcal{N}(0,1)$.*

(ii) *$\Phi(-x) = 1 - \Phi(x)$ for all $x \in \mathbb{R}$, and $z_{1-q} = -z_q$ for all $q \in (0,1)$.*

(iii) *If $Z \sim \mathcal{N}(0,1)$ and $\alpha \in (0,1)$, then $\mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 1 - \alpha$.*

(iv) *If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X+Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.*

Proof. Part (i) follows from the substitution $y = ax+b$ in $\mathbb{P}(aX+b \leq z)$. Part (ii) follows from the symmetry $f_{0,1}(-x) = f_{0,1}(x)$, and (iii) from (ii), since $\mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 2\Phi(z_{1-\alpha/2}) - 1$. Part (iv) follows from Proposition 4.8 by completing the square in the exponent; it is also a special case of Theorem 6.5 (iii). □

The quantiles z_q cannot be computed in closed form. Frequently used values are

$$z_{0.9} \approx 1.282, \qquad z_{0.95} \approx 1.645, \qquad z_{0.975} \approx 1.960, \qquad z_{0.995} \approx 2.576.$$

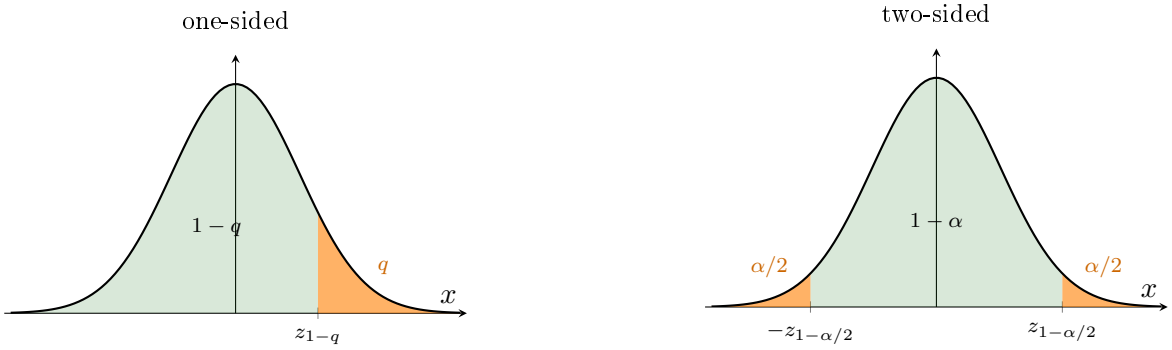

Figure 21: Quantiles of the standard normal distribution. Left: the area to the right of z_{1-q} under the density $f_{0,1}$ is q, shown for $q = 0.1$. Right: by symmetry, $\mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 1 - \alpha$ for $Z \sim \mathcal{N}(0,1)$, shown for $\alpha = 0.05$.

6.2 Multivariate normal distributions

Definition 6.2 (Multivariate normal distribution)**.** Let $d \in \mathbb{N}$, $\boldsymbol{\mu} \in \mathbb{R}^d$, and let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric and positive definite, that is, $\langle \boldsymbol{x}, \boldsymbol{\Sigma}\boldsymbol{x}\rangle > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$. A random vector $\boldsymbol{X}$ with values in $\mathbb{R}^d$ is called *normally distributed* with *mean vector* $\boldsymbol{\mu}$ and *covariance matrix* $\boldsymbol{\Sigma}$, written $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if it has the density

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) \coloneqq \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\Big(-\frac{1}{2}\big\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big\rangle\Big), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Positive definiteness ensures that $\boldsymbol{\Sigma}$ is invertible. One can check that $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\operatorname{Cov}[\boldsymbol{X}] = \boldsymbol{\Sigma}$, which justifies the names. For $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, the $d \times d$ identity matrix, the density factorises, so $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ is the distribution of a vector of d independent standard normal random variables (Proposition 4.5).

Proposition 6.3 (Construction from independent standard normal variables)**.** *Let $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, let $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ be invertible and $\boldsymbol{\mu} \in \mathbb{R}^d$. Then $\boldsymbol{B}\boldsymbol{Z} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{B}^\top)$. Conversely, every $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ arises in this way, for instance with the Cholesky factor $\boldsymbol{B}$ of $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$.*

Proof. By the change-of-variables formula, $\boldsymbol{Y} \coloneqq \boldsymbol{B}\boldsymbol{Z} + \boldsymbol{\mu}$ has the density

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{0},\boldsymbol{I}_d}\big(\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\big)}{|\det \boldsymbol{B}|} = \frac{1}{(2\pi)^{d/2}|\det \boldsymbol{B}|} \exp\Big(-\frac{1}{2}\big|\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\big|^2\Big).$$

Since $|\boldsymbol{B}^{-1}\boldsymbol{v}|^2 = \langle \boldsymbol{v}, (\boldsymbol{B}\boldsymbol{B}^\top)^{-1}\boldsymbol{v}\rangle$ and $\det(2\pi\boldsymbol{B}\boldsymbol{B}^\top) = (2\pi)^d(\det \boldsymbol{B})^2$, this is $f_{\boldsymbol{\mu},\boldsymbol{B}\boldsymbol{B}^\top}$. Every symmetric positive definite matrix has a Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$ with an invertible lower triangular matrix $\boldsymbol{B}$. □

Remark 6.4 (Degenerate normal distributions). For a symmetric positive semidefinite $\boldsymbol{\Sigma}$, we define $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the distribution of $\boldsymbol{B}\boldsymbol{Z} + \boldsymbol{\mu}$ for any $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ with $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Sigma}$; this distribution does not depend on the choice of $\boldsymbol{B}$. If $\boldsymbol{\Sigma}$ is singular, that is, $\langle \boldsymbol{x}, \boldsymbol{\Sigma}\boldsymbol{x}\rangle = 0$ for some $\boldsymbol{x} \neq \boldsymbol{0}$, then $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is concentrated on a proper affine subspace and has no density. For $d = 1$, we set $\mathcal{N}(\mu, 0)$ to be the point mass at μ.

Theorem 6.5 (Properties of multivariate normal distributions)**.** *Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with values in $\mathbb{R}^d$.*

(i) Affine images*: for $\boldsymbol{C} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{C}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top)$, possibly degenerate. In particular, $\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle \sim \mathcal{N}(\langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle, \langle \boldsymbol{\lambda}, \boldsymbol{\Sigma}\boldsymbol{\lambda}\rangle)$ for all $\boldsymbol{\lambda} \in \mathbb{R}^d$. Conversely, a random vector $\boldsymbol{Y}$ such that $\langle \boldsymbol{\lambda}, \boldsymbol{Y}\rangle \sim \mathcal{N}(\langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle, \langle \boldsymbol{\lambda}, \boldsymbol{\Sigma}\boldsymbol{\lambda}\rangle)$ for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ satisfies $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

(ii) Independence and uncorrelatedness*: the coordinates $X_1, \dots, X_d$ are independent if and only if $\boldsymbol{\Sigma}$ is diagonal; more generally, X_i and X_j are independent if and only if $\Sigma_{ij} = 0$.*

(iii) Sums*: if $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma}')$ is independent of $\boldsymbol{X}$, then $\boldsymbol{X} + \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\nu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}')$.*

For proofs, which are most easily given with characteristic functions, see [Geo13, Kle20]. Part (ii) needs the joint normality of $\boldsymbol{X}$: there are uncorrelated, dependent random variables X, Y that are each normally distributed. For instance, let $X \sim \mathcal{N}(0,1)$ and let V be independent of X with $\mathbb{P}(V = 1) = \mathbb{P}(V = -1) = 1/2$. Then $Y \coloneqq VX \sim \mathcal{N}(0,1)$ and $\operatorname{Cov}[X, Y] = \mathbb{E}[V]\mathbb{E}[X^2] = 0$, but $|Y| = |X|$, so X and Y are dependent.

The class of normal distributions is closed under conditioning, and the conditional expectation is an affine function.

Proposition 6.6 (Conditional distributions of bivariate normal vectors). *Let* $(X,Y)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\mu} = (\mu_X, \mu_Y)^\top$ *and positive definite*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & c \\ c & \sigma_Y^2 \end{pmatrix}.$$

Then, for every $y \in \mathbb{R}$, *the conditional density* $f(\cdot|y)$ *of* X *given* $Y = y$, *see Remark 1.24 (iv), is the density of*

$$\mathcal{N}\Big(\mu_X + \frac{c}{\sigma_Y^2}(y - \mu_Y),\ \sigma_X^2 - \frac{c^2}{\sigma_Y^2}\Big).$$

In particular, $\mathbb{E}[X|Y] = \mu_X + c\,\sigma_Y^{-2}(Y - \mu_Y)$.

Proof. Let $Z := X - c\,\sigma_Y^{-2}Y$. The vector $(Z,Y)^\top$ is the image of $(X,Y)^\top$ under an invertible linear map with determinant 1, so it is normally distributed by Theorem 6.5 (i), and its density is $f_{(Z,Y)}(z,y) = f_{(X,Y)}(z + c\,\sigma_Y^{-2}y,\, y)$. By Proposition 4.6, $\mathrm{Cov}[Z,Y] = c - c\,\sigma_Y^{-2}\sigma_Y^2 = 0$, so Z and Y are independent by Theorem 6.5 (ii). Moreover, $\mathbb{E}[Z] = \mu_X - c\,\sigma_Y^{-2}\mu_Y$ and $\mathrm{Var}[Z] = \sigma_X^2 - 2c^2\sigma_Y^{-2} + c^2\sigma_Y^{-2} = \sigma_X^2 - c^2\sigma_Y^{-2}$, which is positive since $\det \boldsymbol{\Sigma} > 0$. Hence

$$f(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{f_Z\big(x - c\,\sigma_Y^{-2}y\big)\,f_Y(y)}{f_Y(y)} = f_Z\big(x - c\,\sigma_Y^{-2}y\big),$$

which is the density of $\mathcal{N}\big(\mathbb{E}[Z] + c\,\sigma_Y^{-2}y, \mathrm{Var}[Z]\big)$, as claimed. □

The conditional variance does not depend on y and is smaller than σ_X^2 unless $c = 0$. The line $y \mapsto \mu_X + c\,\sigma_Y^{-2}(y - \mu_Y)$ is the *regression line* of X on Y; compare Theorem 2.17, where the empirical counterparts of these quantities appear.

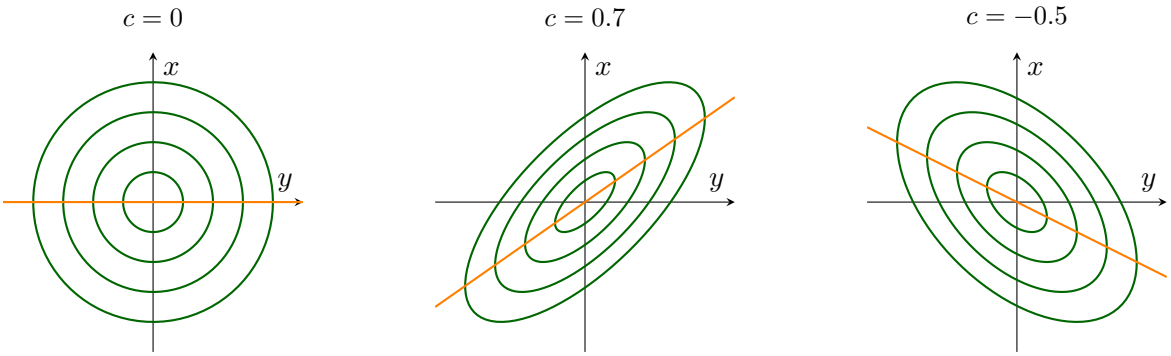

Figure 22: Level sets of the density of a bivariate normal vector $(X,Y)^\top$ with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ and covariance c. The orange line is the conditional expectation $y \mapsto \mathbb{E}[X|Y = y] = cy$ from Proposition 6.6; the horizontal axis shows y and the vertical axis x.

6.3 Chi-square and Student distributions

Estimating the variance of normal observations leads to two further families of distributions.

Definition 6.7 (Chi-square and Student distributions). Let $k \in \mathbb{N}$ and let $Z, Z_1, \ldots, Z_k$ be independent with distribution $\mathcal{N}(0,1)$.

(i) The distribution of $Z_1^2 + \cdots + Z_k^2$ is called the *chi-square distribution* with k degrees of freedom, written χ_k^2.

(ii) The distribution of $Z/\sqrt{(Z_1^2+\cdots+Z_k^2)/k}$ is called the *Student t-distribution* with k degrees of freedom, written t_k.

We write $\chi^2_{k;q}$ and $t_{k;q}$ for the q-quantiles of these distributions.

Proposition 6.8 (Densities). *The distribution χ^2_k has expectation k, variance $2k$ and the density*

$$x \mapsto \frac{x^{k/2-1}\mathrm{e}^{-x/2}}{2^{k/2}\,\Gamma(k/2)}\,\mathbf{1}\{x>0\}.$$

The distribution t_k is symmetric, so $t_{k;1-q} = -t_{k;q}$, and has the density

$$x \mapsto \frac{\Gamma\big((k+1)/2\big)}{\sqrt{k\pi}\,\Gamma(k/2)}\Big(1+\frac{x^2}{k}\Big)^{-(k+1)/2}, \qquad x\in\mathbb{R}.$$

Here $\Gamma(s)=\int_0^\infty x^{s-1}\mathrm{e}^{-x}\,\mathrm{d}x$ is the Gamma function. As $k\to\infty$, $t_k \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

For proofs, see [Geo13]. The density of t_1 is $1/(\pi(1+x^2))$, the *Cauchy distribution*, which has no expectation. The tails of t_k are heavier than those of $\mathcal{N}(0,1)$, and so are the quantiles; for instance $t_{8;0.95}\approx 1.860$, $t_{8;0.975}\approx 2.306$ and $t_{24;0.975}\approx 2.064$, compared with $z_{0.95}\approx 1.645$ and $z_{0.975}\approx 1.960$. The convergence $t_k \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ follows from the law of large numbers and Slutsky's theorem, since the denominator in Definition 6.7 tends to 1.

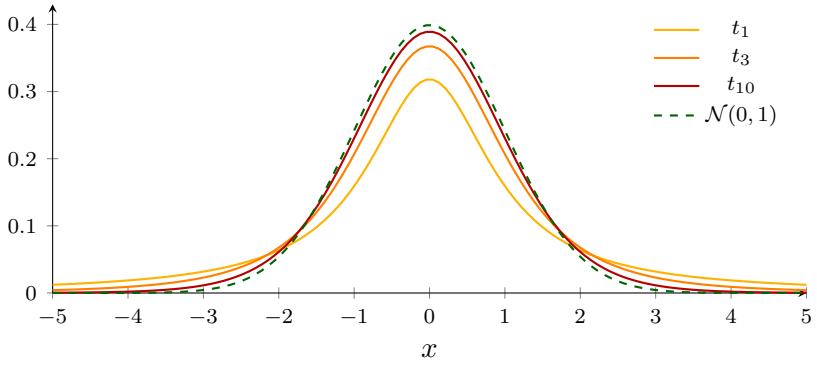

Figure 23: Densities of the Student t-distributions t_k for $k\in\{1,3,10\}$ and of the standard normal distribution. The t-distributions have heavier tails and converge to $\mathcal{N}(0,1)$ as $k\to\infty$ (Proposition 6.8).

The following theorem is the basis of all statistical procedures for normal observations with unknown variance. We recall the estimators $M=\frac{1}{n}\sum_{i=1}^n X_i$ and $V^*=\frac{1}{n-1}\sum_{i=1}^n (X_i-M)^2$ from Theorem 2.8.

Theorem 6.9 (Student's theorem). *Let $n\geq 2$ and let $X_1,\dots,X_n$ be independent with distribution $\mathcal{N}(\mu,\sigma^2)$.*

(i) $M\sim\mathcal{N}(\mu,\sigma^2/n)$.

(ii) $(n-1)V^*/\sigma^2\sim\chi^2_{n-1}$.

(iii) *M and V^* are independent.*

(iv) $\sqrt{n}\,(M-\mu)/\sqrt{V^*}\sim t_{n-1}$.

Sketch of proof. Part (i) follows from Proposition 6.1. For (ii) and (iii), one rotates $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ by an orthogonal matrix whose first row is $n^{-1/2}\mathbf{1}^\top$; by Theorem 6.5 (i), the rotated vector has independent coordinates with variance σ^2, its first coordinate is $\sqrt{n}M$, and the sum of squares of the remaining $n-1$ coordinates is $(n-1)V^*$. Part (iv) follows from (i)–(iii) and Definition 6.7, since

$$\sqrt{n}\,\frac{M-\mu}{\sqrt{V^*}} = \frac{\sqrt{n}(M-\mu)/\sigma}{\sqrt{\big((n-1)V^*/\sigma^2\big)/(n-1)}}.$$

For details, see [Geo13]. □

7 Point estimation

7.1 Statistical models and observations

In applications, a statistical model often arises as follows. An experiment is described by a family $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ of probability measures on a common space $(\Omega, \mathcal{F})$, and we observe a random vector $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ on it. Only the distributions $\mathbb{P}_\vartheta \circ \boldsymbol{X}^{-1}$ of the observation matter for statistical inference, so we may work with the model $(\mathcal{X}, \mathcal{F}_\mathcal{X}, \mathbb{P}_\vartheta \circ \boldsymbol{X}^{-1} \colon \vartheta \in \Theta)$ on the sample space $\mathcal{X}$ of the observations; this is the setting of Definition 2.2. For n independent repetitions, we obtain the product model of Remark 2.3 (iii), in which the coordinate projections $X_1, \dots, X_n$ are i.i.d. under every $\mathbb{P}_\vartheta$.

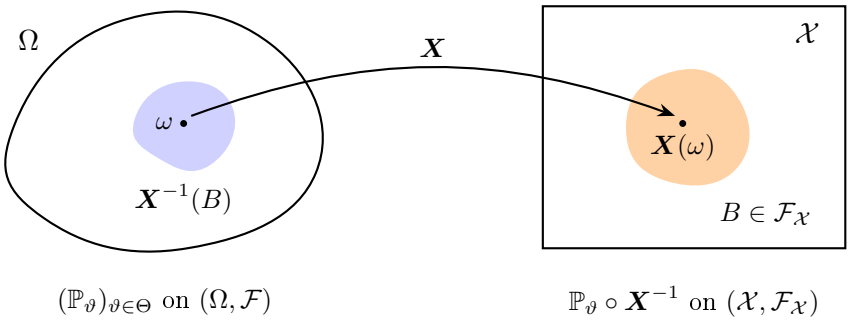

Figure 24: A statistical model as a family of distributions of an observation: the unknown measures $\mathbb{P}_\vartheta$ on $(\Omega, \mathcal{F})$ are pushed forward by the random vector $\boldsymbol{X}$ to the measures $\mathbb{P}_\vartheta \circ \boldsymbol{X}^{-1}$ on the sample space $\mathcal{X}$, for instance $\mathcal{X} = \mathbb{R}^n$; only the latter matter for inference.

7.2 Consistency and mean squared error

In Example 2.6, we compared two estimators by their behaviour as $n \to \infty$ and by their variances. We now formalise these criteria. Let $(\mathcal{X}_n, \mathcal{F}_n, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$, $n \in \mathbb{N}$, be a sequence of statistical models, for instance the n-fold product models of a fixed model, and let $\tau \colon \Theta \to \mathbb{R}$ be a characteristic.

Definition 7.1 (Consistency and mean squared error)**.**

(i) A sequence of estimators $T_n \colon \mathcal{X}_n \to \mathbb{R}$ for τ is called *consistent* if $T_n \to \tau(\vartheta)$ in $\mathbb{P}_\vartheta$-probability for every $\vartheta \in \Theta$, that is, $\mathbb{P}_\vartheta(|T_n - \tau(\vartheta)| > \varepsilon) \to 0$ for all $\varepsilon > 0$.

(ii) The *mean squared error* of an estimator T for τ with $\mathbb{E}_\vartheta[T^2] < \infty$ is $\mathrm{MSE}_T(\vartheta) := \mathbb{E}_\vartheta\big[(T - \tau(\vartheta))^2\big]$.

Proposition 7.2 (Bias–variance decomposition). *Let T be an estimator for τ with $\mathbb{E}_\vartheta[T^2] < \infty$ for all $\vartheta \in \Theta$. Then*

$$\mathrm{MSE}_T(\vartheta) = \mathrm{Var}_\vartheta[T] + \mathrm{B}_T(\vartheta)^2 \qquad \text{for all } \vartheta \in \Theta.$$

If T_n are estimators with $\mathrm{MSE}_{T_n}(\vartheta) \to 0$ for all ϑ, then $(T_n)_{n\in\mathbb{N}}$ is consistent.

Proof. Expanding $(T - \mathbb{E}_\vartheta[T] + \mathrm{B}_T(\vartheta))^2$, the mixed term has expectation zero. The second claim is Theorem 5.2 (ii) with $p = 2$. □

Example 7.3 (Estimators for uniform distributions, continued). In Example 2.6, the estimator $\tilde{T}_n = \max\{X_1, \dots, X_n\}$ has bias $-\vartheta/(n+1)$, so

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

This is larger than $\mathrm{MSE}_{\tilde{T}_n^*}(\vartheta) = \vartheta^2/(n(n+2))$ for $n \geq 2$, but much smaller than $\mathrm{MSE}_{T_n}(\vartheta) = \vartheta^2/(3n)$ for large n. All three estimators are consistent by Proposition 7.2.

Proposition 7.4 (Consistency of mean and variance estimators). *In the real n-fold product models of Theorem 2.8 with $\mathrm{Var}_\vartheta[X_1] < \infty$, the estimators $M_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $V_n^* = \frac{1}{n-1}\sum_{i=1}^n (X_i - M_n)^2$ are consistent for $m(\vartheta)$ and $v(\vartheta)$, respectively.*

Proof. Consistency of M_n is the weak law of large numbers (Theorem 5.7). Expanding the square gives $V_n^* = \frac{n}{n-1}\big(\frac{1}{n}\sum_{i=1}^n X_i^2 - M_n^2\big)$. By the strong law of large numbers (Theorem 5.8) applied to X_i and to X_i^2, and by Theorem 5.2 (i), this converges in probability to $\mathbb{E}_\vartheta[X_1^2] - m(\vartheta)^2 = v(\vartheta)$. □

7.3 Maximum-likelihood estimators in product models

In a product model, the likelihood function is a product, and the log-likelihood function is a sum.

Example 7.5 (Maximum-likelihood estimators). Consider the real n-fold product model with the following distributions Q_ϑ, and let $\boldsymbol{x} = (x_1, \dots, x_n)$ be an observation.

(i) *Normal distributions*, $Q_\vartheta = \mathcal{N}(\mu, \sigma^2)$ with $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. The log-likelihood function is

$$\log \rho_{\boldsymbol{x}}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

For fixed σ^2, it is maximal for $\mu = M(\boldsymbol{x})$, since $\sum_i (x_i - \mu)^2 = \sum_i (x_i - M(\boldsymbol{x}))^2 + n(M(\boldsymbol{x}) - \mu)^2$. If the x_i are not all equal, an event of probability one, differentiation in σ^2 then gives the maximum-likelihood estimators

$$\hat{\mu} = M = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - M)^2 = \frac{n-1}{n}\, V^*.$$

By Theorem 2.8, $\hat{\sigma}^2$ is biased, with bias $-\sigma^2/n$.

(ii) *Exponential distributions*, $Q_\lambda = \mathrm{Exp}_\lambda$ with $\lambda \in (0, \infty)$. If $\sum_i x_i > 0$, an event of probability one, then $\log \rho_{\boldsymbol{x}}(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$ is maximal for $\hat{\lambda} = 1/M$.

(iii) *Uniform distributions*, $Q_\vartheta = \mathrm{U}_{[0,\vartheta]}$ with $\vartheta \in (0,\infty)$, as in Example 2.6. For $\boldsymbol{x} \in [0,\infty)^n$, the likelihood function $\rho_{\boldsymbol{x}}(\vartheta) = \vartheta^{-n}\mathbf{1}\{\max_i x_i \le \vartheta\}$ vanishes for $\vartheta < \max_i x_i$ and is decreasing on $[\max_i x_i, \infty)$. Hence $\hat{\vartheta} = \tilde{T}_n = \max\{X_1, \dots, X_n\}$ is the maximum-likelihood estimator; it is biased, see Example 2.6. This example is not covered by differentiation, since the likelihood function jumps at its maximiser.

(iv) *Poisson distributions*, $Q_\lambda = \mathrm{Poi}_\lambda$ with $\lambda \in [0,\infty)$, where Poi_0 is the point mass at 0. The log-likelihood function $\big(\sum_i x_i\big)\log\lambda - n\lambda - \sum_i \log(x_i!)$ is maximal for $\hat{\lambda} = M$.

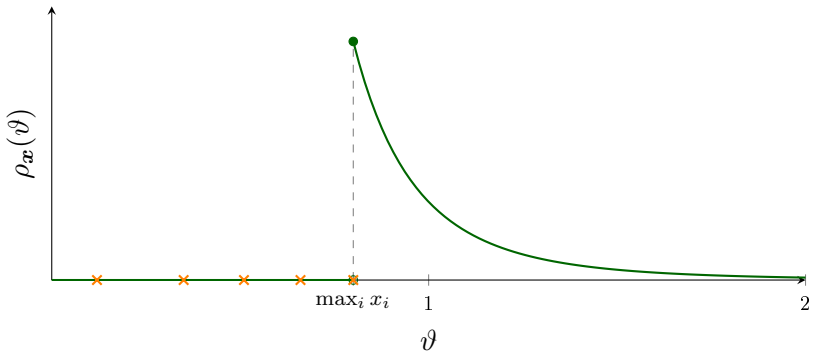

Figure 25: Likelihood function $\rho_{\boldsymbol{x}}(\vartheta) = \vartheta^{-n}\mathbf{1}\{\max_i x_i \le \vartheta\}$ of the uniform model in Example 7.5 (iii) for $n = 5$ observations (orange crosses). It vanishes to the left of $\max_i x_i$ and decreases to the right, so it is maximal at $\hat{\vartheta} = \max_i x_i$.

7.4 Best estimators

By Proposition 7.2, the mean squared error of an unbiased estimator is its variance, which motivates Definition 2.11. Best estimators can often be identified with the following tools; see [Geo13, LC98]. The *information inequality* of Cramér and Rao gives, under regularity conditions, a lower bound for the variance of all unbiased estimators. An unbiased estimator that attains this bound is a best estimator; under regularity conditions, this happens only in one-parameter exponential families and for the characteristic $\tau(\vartheta) = \mathbb{E}_\vartheta[T]$ of the canonical statistic T, see Remark 2.12. For example, in the Gaussian product model with known variance, M is a best estimator for the mean, and in the Poisson product model, M is a best estimator for λ.

8 Confidence intervals

Instead of a single value, a confidence region specifies a set of plausible parameter values, as in the second approach of Example 2.1.

Definition 8.1 (Confidence region)**.** Let $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\vartheta \colon \vartheta \in \Theta)$ be a statistical model, $\tau \colon \Theta \to \Sigma$ a characteristic and $\alpha \in (0,1)$. A map $C \colon \mathcal{X} \to \mathcal{P}(\Sigma)$ is called a *confidence region* for τ with *confidence level* $1-\alpha$ if

$$\mathbb{P}_\vartheta\big(\{\boldsymbol{x} \in \mathcal{X} \colon \tau(\vartheta) \in C(\boldsymbol{x})\}\big) \ge 1-\alpha \qquad \text{for all } \vartheta \in \Theta.$$

If $\Sigma = \mathbb{R}$ and every $C(\boldsymbol{x})$ is an interval, C is called a *confidence interval*.

The trivial region $C(\boldsymbol{x}) = \Sigma$ satisfies the definition for every α but carries no information, so one looks for small confidence regions. The probability statement concerns the random

set C for a fixed parameter: if the experiment is repeated independently, then, by the strong law of large numbers, the proportion of regions that contain $\tau(\vartheta)$ converges almost surely to a limit of at least $1-\alpha$ (Fig. 26).

Theorem 8.2 (Confidence intervals for the mean). *Let $\alpha \in (0,1)$, $n \geq 2$, and let M and V^* be as in Theorem 2.8.*

(i) *In the Gaussian product model with $Q_\mu = \mathcal{N}(\mu, \sigma_0^2)$, $\mu \in \mathbb{R}$, and known $\sigma_0 > 0$,*

$$C(\boldsymbol{X}) := \left[M - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ M + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$$

is a confidence interval for μ with confidence level $1-\alpha$.

(ii) *In the Gaussian product model with $Q_\vartheta = \mathcal{N}(\mu, \sigma^2)$, $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0,\infty)$,*

$$C(\boldsymbol{X}) := \left[M - t_{n-1;1-\alpha/2}\sqrt{V^*/n},\ M + t_{n-1;1-\alpha/2}\sqrt{V^*/n}\right]$$

is a confidence interval for μ with confidence level $1-\alpha$.

(iii) *In the real product model of any family Q_ϑ with mean $m(\vartheta)$ and variance $v(\vartheta) \in (0,\infty)$, the intervals*

$$C_n(\boldsymbol{X}) := \left[M - z_{1-\alpha/2}\sqrt{V^*/n},\ M + z_{1-\alpha/2}\sqrt{V^*/n}\right]$$

have asymptotic confidence level $1-\alpha$, *that is, $\mathbb{P}_\vartheta\big(m(\vartheta) \in C_n(\boldsymbol{X})\big) \to 1-\alpha$ as $n \to \infty$ for every ϑ.*

Proof. For (i), $\mu \in C(\boldsymbol{X})$ holds if and only if $|\sqrt{n}(M-\mu)/\sigma_0| \leq z_{1-\alpha/2}$. By Theorem 6.9 (i) and Proposition 6.1 (i), $\sqrt{n}(M-\mu)/\sigma_0 \sim \mathcal{N}(0,1)$ under $\mathbb{P}_\mu$, and Proposition 6.1 (iii) gives the probability $1-\alpha$. Part (ii) follows in the same way from Theorem 6.9 (iv) and the symmetry of t_{n-1}. For (iii), the central limit theorem gives $\sqrt{n}(M - m(\vartheta))/\sqrt{v(\vartheta)} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$, and Propositions 5.13 and 7.4 give $\sqrt{v(\vartheta)/V^*} \xrightarrow{\mathbb{P}} 1$. By Slutsky's theorem, $\sqrt{n}(M - m(\vartheta))/\sqrt{V^*} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$, and the claim follows since Φ is continuous. □

The length of these intervals decreases like $n^{-1/2}$ and increases as α decreases. The t-quantiles in (ii) are larger than the normal quantiles in (i), which compensates for the estimation of the variance.

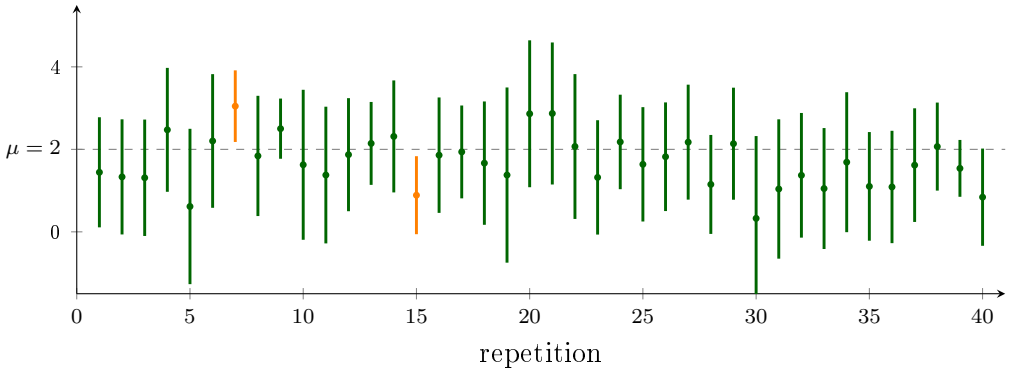

Figure 26: 40 simulated 95%-confidence intervals $M \pm t_{9;0.975}\sqrt{V^*/10}$ from Theorem 8.2 (ii), each based on $n = 10$ observations from $\mathcal{N}(2,4)$. In this simulation, 2 intervals (orange) do not contain the true mean; on average, 5% of the intervals miss it.

Example 8.3 (Emission measurements)**.** In $n = 9$ test drives of a vehicle, the following NO_x emissions are measured, in milligrams per kilometre:

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

The observed values are $M = 76.44$ and $\sqrt{V^*} = 5.59$. Assuming i.i.d. normally distributed measurements with unknown mean and variance, Theorem 8.2 (ii) with $t_{8;0.975} = 2.306$ gives the 95%-confidence interval

$$76.44 \pm 2.306 \cdot 5.59/3 = [72.15,\ 80.74]$$

for the mean emission μ.

For a success probability p in a Bernoulli product model, Theorem 8.2 (iii) with $\hat{p} := M$ yields the approximate confidence interval $\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/(n-1)}$, since $V^* = \frac{n}{n-1}\hat{p}(1-\hat{p})$ for 0-1-valued observations.

9 Statistical tests

9.1 Errors and their asymmetry

We use the notions of Definition 2.13 and Remark 2.14: a test $\varphi\colon \mathcal{X} \to [0,1]$ of Θ_0 against Θ_1, its power function $G_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi]$ and its level. The four possible situations are summarised as follows.

	decision for H_0	decision for H_1 (rejection of H_0)
H_0 is true	correct decision	type I error
H_1 is true	type II error	correct decision

A test of level α satisfies $\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta) \le \alpha$, so its type I error probability is controlled by a value α chosen in advance. No such control exists for the type II error probability $1 - G_\varphi(\vartheta)$, $\vartheta \in \Theta_1$, which can be as large as $1 - \alpha$ for parameters close to Θ_0 (Fig. 27). Hence only a rejection of H_0 is a reliable decision: it is either correct or a type I error, which happens with probability at most α. The statement that one wants to establish must therefore be chosen as the alternative H_1; see Example 9.6. Among all tests of level α, one seeks tests whose power $G_\varphi(\vartheta)$ on Θ_1 is as large as possible.

Example 9.1 (Testing problems)**.**

(i) To check whether a die is fair, let $X_1, \dots, X_n$ be i.i.d. Bernoulli variables indicating whether the i-th roll shows a six, with unknown success probability $p \in \Theta = (0,1)$. The hypothesis is $\Theta_0 = \{1/6\}$, and a natural test rejects it if $|M - 1/6| > c$ for a critical value c.

(ii) A vehicle type complies with an emission limit of 80 milligrams NO_x per kilometre if the mean emission μ of independent test drives satisfies $\mu \le 80$. Modelling the measurements as i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\vartheta = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$, the statement $\mu \le 80$ corresponds to $\Theta_0 = (-\infty, 80] \times (0, \infty)$.

In both cases, the key question is how to choose the critical value.

9.2 Gauss tests

Theorem 9.2 (One-sided Gauss test)**.** *Consider the Gaussian product model with* $Q_\mu = \mathcal{N}(\mu, \sigma_0^2)$, $\mu \in \Theta = \mathbb{R}$, *and known* $\sigma_0 > 0$. *Let* $\mu_0 \in \mathbb{R}$, $\alpha \in (0,1)$, *and test* $H_0 \colon \mu \le \mu_0$ *against* $H_1 \colon \mu > \mu_0$ *with*

$$\varphi(\boldsymbol{X}) \coloneqq \mathbf{1}\Big\{M > \mu_0 + z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}}\Big\}.$$

Then the power function is

$$G_\varphi(\mu) = \Phi\Big(\sqrt{n}\,\frac{\mu-\mu_0}{\sigma_0} - z_{1-\alpha}\Big), \qquad \mu \in \mathbb{R}.$$

It is increasing with $G_\varphi(\mu_0) = \alpha$, *so* φ *is a test of level* α, *and* $G_\varphi(\mu) \to 1$ *as* $n \to \infty$ *for every* $\mu > \mu_0$.

Proof. Under $\mathbb{P}_\mu$, the variable $Z \coloneqq \sqrt{n}(M-\mu)/\sigma_0$ is standard normal. With $\delta \coloneqq \sqrt{n}(\mu - \mu_0)/\sigma_0$, we obtain $G_\varphi(\mu) = \mathbb{P}_\mu(Z > z_{1-\alpha} - \delta) = 1 - \Phi(z_{1-\alpha} - \delta) = \Phi(\delta - z_{1-\alpha})$ by Proposition 6.1 (ii). The remaining claims follow since Φ is increasing with $\Phi(z_{1-\alpha}) = 1-\alpha$ and $\delta \to \infty$ for $\mu > \mu_0$. □

For $H_0 \colon \mu \ge \mu_0$ against $H_1 \colon \mu < \mu_0$, the test $\mathbf{1}\{M < \mu_0 - z_{1-\alpha}\sigma_0/\sqrt{n}\}$ has level α and power function $\Phi\big(\sqrt{n}(\mu_0-\mu)/\sigma_0 - z_{1-\alpha}\big)$, by the same argument.

Theorem 9.3 (Two-sided Gauss test)**.** *In the model of Theorem 9.2, the test*

$$\varphi(\boldsymbol{X}) \coloneqq \mathbf{1}\Big\{\sqrt{n}\,\frac{|M-\mu_0|}{\sigma_0} > z_{1-\alpha/2}\Big\}$$

of $H_0 \colon \mu = \mu_0$ *against* $H_1 \colon \mu \ne \mu_0$ *has the power function*

$$G_\varphi(\mu) = \Phi(\delta - z_{1-\alpha/2}) + \Phi(-\delta - z_{1-\alpha/2}), \qquad \text{where } \delta \coloneqq \sqrt{n}\,\frac{\mu-\mu_0}{\sigma_0}.$$

In particular, $G_\varphi(\mu_0) = \alpha$, *so* φ *is a test of level* α.

Proof. With Z as in the previous proof, $\sqrt{n}(M-\mu_0)/\sigma_0 = Z + \delta$, and $\mathbb{P}_\mu(|Z+\delta| > z_{1-\alpha/2}) = 1 - \Phi(z_{1-\alpha/2} - \delta) + \Phi(-z_{1-\alpha/2} - \delta)$. □

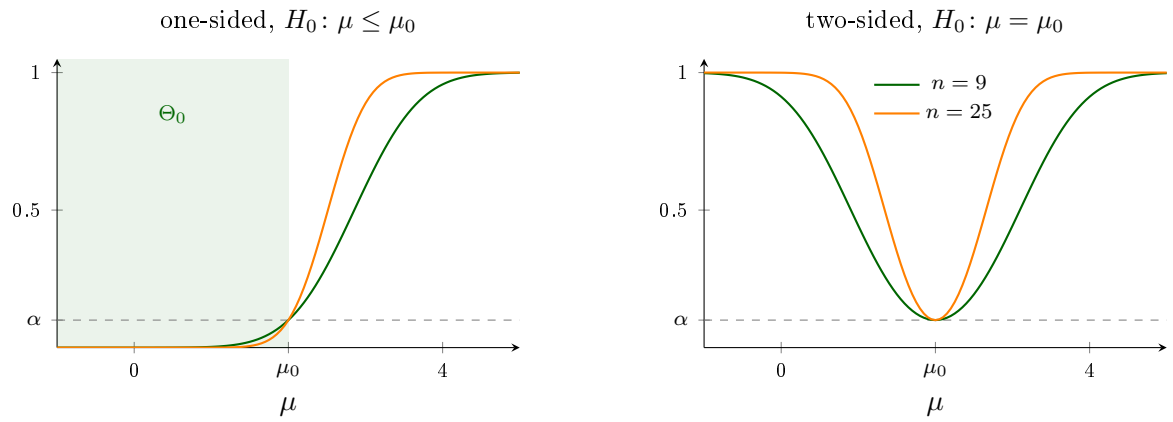

Figure 27: Power functions G_φ of the one-sided Gauss test (Theorem 9.2) and of the two-sided Gauss test (Theorem 9.3) for $\mu_0 = 2$, $\sigma_0 = 2$, $\alpha = 0.1$ and $n \in \{9, 25\}$. The type II error at $\mu \in \Theta_1$ is $1 - G_\varphi(\mu)$; it can be as large as $1-\alpha$ close to μ_0.

9.3 t-tests and p-values

If the variance is unknown, we replace σ_0 by $\sqrt{V^*}$ and the normal quantiles by quantiles of the t-distribution.

Theorem 9.4 (Student's t-tests)**.** *Consider the Gaussian product model with $Q_\vartheta = \mathcal{N}(\mu, \sigma^2)$, $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, let $n \geq 2$, $\mu_0 \in \mathbb{R}$, $\alpha \in (0,1)$, and*

$$T_{\mu_0} := \sqrt{n}\,\frac{M - \mu_0}{\sqrt{V^*}}.$$

Each of the following tests has level α:

(i) $\mathbf{1}\{T_{\mu_0} > t_{n-1;1-\alpha}\}$ *for* $H_0 \colon \mu \leq \mu_0$ *against* $H_1 \colon \mu > \mu_0$;

(ii) $\mathbf{1}\{T_{\mu_0} < -t_{n-1;1-\alpha}\}$ *for* $H_0 \colon \mu \geq \mu_0$ *against* $H_1 \colon \mu < \mu_0$;

(iii) $\mathbf{1}\{|T_{\mu_0}| > t_{n-1;1-\alpha/2}\}$ *for* $H_0 \colon \mu = \mu_0$ *against* $H_1 \colon \mu \neq \mu_0$.

Proof. Under $\mathbb{P}_{(\mu,\sigma^2)}$, $T_\mu \sim t_{n-1}$ by Theorem 6.9 (iv). For (i) and $\mu \leq \mu_0$, we have $T_{\mu_0} \leq T_\mu$, and hence $\mathbb{P}_{(\mu,\sigma^2)}(T_{\mu_0} > t_{n-1;1-\alpha}) \leq \mathbb{P}_{(\mu,\sigma^2)}(T_\mu > t_{n-1;1-\alpha}) = \alpha$. Part (ii) follows in the same way, and (iii) from the symmetry of t_{n-1}. □

The power functions of the t-tests can be computed from the noncentral t-distribution, and they tend to 1 as $n \to \infty$ for every parameter in the alternative; see [Geo13, LR22]. Figure 28 compares them with the Gauss tests.

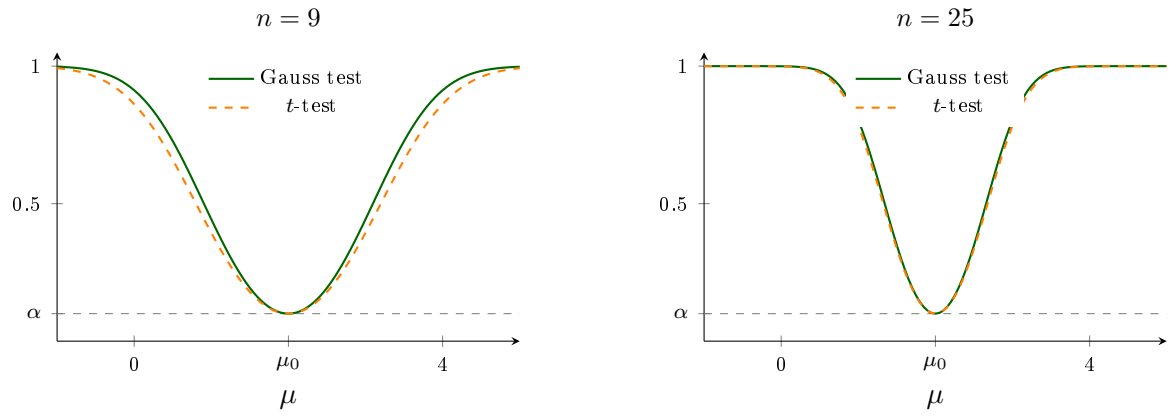

Figure 28: Power functions of the two-sided Gauss test (known $\sigma = 2$) and of the two-sided t-test (unknown σ) for $H_0 \colon \mu = \mu_0 = 2$ at level $\alpha = 0.1$. The power of the t-test is computed exactly from the noncentral t-distribution. Estimating σ costs power, and the loss shrinks as n grows.

Instead of fixing α in advance, one often reports the smallest level at which the observation leads to a rejection.

Definition 9.5 (p-value)**.** Let $T \colon \mathcal{X} \to \mathbb{R}$ be a statistic such that large values of T speak against H_0. For an observation $\boldsymbol{x}$, the number

$$\mathrm{pv}(\boldsymbol{x}) := \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta\big(T \geq T(\boldsymbol{x})\big)$$

is called the *p-value* of $\boldsymbol{x}$.

The test $\mathbf{1}\{\text{pv} \le \alpha\}$ has level α for every $\alpha \in (0,1)$; see [LR22]. A small p-value means that, under every parameter in the hypothesis, an observation at least as extreme as $\boldsymbol{x}$ would be unlikely. The p-value is not the probability that H_0 holds.

Example 9.6 (Emission measurements, continued)**.** We continue Example 8.3 and use the one-sided t-tests of Theorem 9.4 at level $\alpha = 0.05$ with $t_{8;0.95} = 1.860$. The observed test statistic is $T_{80} = 3 \cdot (76.44 - 80)/5.59 = -1.907$.

A manufacturer who wants to demonstrate compliance has to choose this statement as the alternative, that is, test $H_0 \colon \mu \ge 80$ against $H_1 \colon \mu < 80$. Since

$$M = 76.44 < 80 - 1.860 \cdot 5.59/3 = 76.53,$$

the hypothesis is rejected, and the p-value is $\mathbb{P}(t_8 \le -1.907) \approx 0.046$. The rejection is either correct or a type I error, whose probability is at most 5%.

A supervisory authority that wants to demonstrate a violation tests $H_0 \colon \mu \le 80$ against $H_1 \colon \mu > 80$. Since $M = 76.44$ is not larger than $80 + 1.860 \cdot 5.59/3 = 83.47$, the hypothesis cannot be rejected, and the p-value is about 0.954. This decision is either correct or a type II error, whose probability is not controlled, so it does not show that the vehicle complies with the limit.

10 Linear models

The regression model of Example 2.16 is a special case of the following class of models.

Definition 10.1 (Linear model)**.** Let $s, n \in \mathbb{N}$ with $s < n$, and let $\boldsymbol{A} \in \mathbb{R}^{n \times s}$ be a matrix of rank s, the *design matrix*. Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$ be a random vector of uncorrelated errors with $\mathbb{E}[\xi_k] = 0$ and $\operatorname{Var}[\xi_k] = 1$, whose distribution is fixed. The *linear model* consists of the observations

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\gamma} + \sigma\boldsymbol{\xi},$$

with unknown parameters $\boldsymbol{\gamma} \in \mathbb{R}^s$ and $\sigma > 0$, that is, of the statistical model

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_{\boldsymbol{\gamma},\sigma} \colon \boldsymbol{\gamma} \in \mathbb{R}^s,\ \sigma \in (0,\infty)\big),$$

where $\mathbb{P}_{\boldsymbol{\gamma},\sigma}$ is the distribution of $\boldsymbol{A}\boldsymbol{\gamma} + \sigma\boldsymbol{\xi}$. If $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, it is called a *Gaussian linear model*.

By Proposition 4.7, $\mathbb{E}_{\boldsymbol{\gamma},\sigma}[\boldsymbol{X}] = \boldsymbol{A}\boldsymbol{\gamma}$ and $\operatorname{Cov}_{\boldsymbol{\gamma},\sigma}[\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$.

Example 10.2 (Linear models)**.**

(i) *Gaussian product model*: $s = 1$, $\boldsymbol{A} = \boldsymbol{1}$ and $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, so that $\boldsymbol{X} \sim \mathcal{N}(\gamma\boldsymbol{1}, \sigma^2\boldsymbol{I}_n)$ consists of n i.i.d. $\mathcal{N}(\gamma, \sigma^2)$ observations.

(ii) *Simple linear regression* (Example 2.16): $s = 2$ and $\boldsymbol{A} = (\boldsymbol{1}\ \boldsymbol{t})$, the matrix with columns $\boldsymbol{1}$ and $\boldsymbol{t}$, which has rank 2 if and only if not all t_k are equal. Here we index the coefficients as $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\top$, with intercept γ_0 and slope γ_1, as in Part I.

(iii) *Polynomial regression*: $X_k = \gamma_1 + \gamma_2 t_k + \cdots + \gamma_s t_k^{s-1} + \sigma\xi_k$, that is, $A_{kj} = t_k^{j-1}$. This matrix has rank s if at least s of the t_k are distinct, since a non-zero polynomial of degree less than s has fewer than s zeros.

Theorem 10.3 (Least-squares estimators in linear models)**.** *In the linear model of Definition 10.1, let $L := \{\boldsymbol{A}\boldsymbol{y} \colon \boldsymbol{y} \in \mathbb{R}^s\}$ be the column space of $\boldsymbol{A}$ and $\boldsymbol{\Pi}_L$ the orthogonal projection onto L.*

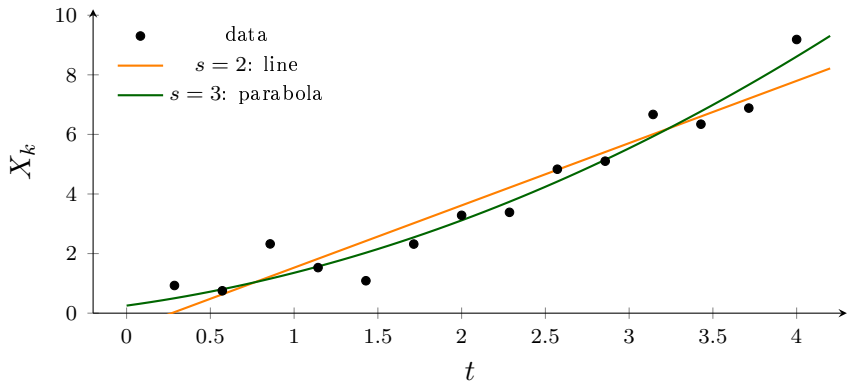

Figure 29: Least-squares fits in the polynomial regression model of Example 10.2 (iii) (simulated data). The quadratic model ($s = 3$) captures the curvature that the straight line ($s = 2$) misses.

(i) *The matrix $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible, and*

$$\hat{\boldsymbol{\gamma}} := (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{X}$$

is the unique minimiser of $\boldsymbol{y} \mapsto |\boldsymbol{X} - \boldsymbol{A}\boldsymbol{y}|^2$; moreover, $\boldsymbol{A}\hat{\boldsymbol{\gamma}} = \boldsymbol{\Pi}_L \boldsymbol{X}$.

(ii) *$\hat{\boldsymbol{\gamma}}$ is an unbiased estimator for $\boldsymbol{\gamma}$ with $\mathrm{Cov}_{\boldsymbol{\gamma},\sigma}[\hat{\boldsymbol{\gamma}}] = \sigma^2 (\boldsymbol{A}^\top \boldsymbol{A})^{-1}$.*

(iii) *$\hat{\sigma}^2 := |\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}|^2/(n-s)$ is an unbiased estimator for σ^2.*

Proof. For (i), $\langle \boldsymbol{y}, \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{y} \rangle = |\boldsymbol{A}\boldsymbol{y}|^2 > 0$ for $\boldsymbol{y} \neq \boldsymbol{0}$, since $\boldsymbol{A}$ has rank s; hence $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible. The vector $\hat{\boldsymbol{\gamma}}$ solves the *normal equations* $\boldsymbol{A}^\top(\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}) = \boldsymbol{0}$, so $\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}$ is orthogonal to L, and $\boldsymbol{A}\hat{\boldsymbol{\gamma}} = \boldsymbol{\Pi}_L \boldsymbol{X}$. By Pythagoras' theorem, $|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{y}|^2 = |\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}|^2 + |\boldsymbol{A}(\hat{\boldsymbol{\gamma}} - \boldsymbol{y})|^2$ for all $\boldsymbol{y} \in \mathbb{R}^s$, which is minimal if and only if $\boldsymbol{y} = \hat{\boldsymbol{\gamma}}$.

For (ii), $\mathbb{E}_{\boldsymbol{\gamma},\sigma}[\hat{\boldsymbol{\gamma}}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\gamma} = \boldsymbol{\gamma}$, and Proposition 4.7 gives

$$\mathrm{Cov}_{\boldsymbol{\gamma},\sigma}[\hat{\boldsymbol{\gamma}}] = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \, \sigma^2 \boldsymbol{I}_n \, \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1} = \sigma^2 (\boldsymbol{A}^\top \boldsymbol{A})^{-1}.$$

For (iii), $(\boldsymbol{I}_n - \boldsymbol{\Pi}_L)\boldsymbol{A}\boldsymbol{\gamma} = \boldsymbol{0}$ gives $\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}} = (\boldsymbol{I}_n - \boldsymbol{\Pi}_L)\boldsymbol{X} = \sigma(\boldsymbol{I}_n - \boldsymbol{\Pi}_L)\boldsymbol{\xi}$. Since $\boldsymbol{I}_n - \boldsymbol{\Pi}_L$ is the orthogonal projection onto the $(n-s)$-dimensional orthogonal complement of L, and $\mathrm{Cov}[\boldsymbol{\xi}] = \boldsymbol{I}_n$,

$$\mathbb{E}_{\boldsymbol{\gamma},\sigma}\big[|\boldsymbol{X} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}|^2\big] = \sigma^2 \, \mathbb{E}\big[\langle \boldsymbol{\xi}, (\boldsymbol{I}_n - \boldsymbol{\Pi}_L)\boldsymbol{\xi} \rangle\big] = \sigma^2 \, \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{\Pi}_L) = \sigma^2 (n - s). \qquad \square$$

In the Gaussian product model of Example 10.2 (i), Theorem 10.3 gives $\hat{\gamma} = M$ and $\hat{\sigma}^2 = V^*$, in accordance with Theorem 2.8.

Example 10.4 (Simple linear regression, continued)**.** For $\boldsymbol{A} = (\boldsymbol{1}\ \boldsymbol{t})$, we have

$$\boldsymbol{A}^\top \boldsymbol{A} = n \begin{pmatrix} 1 & M(\boldsymbol{t}) \\ M(\boldsymbol{t}) & \frac{1}{n}\sum_k t_k^2 \end{pmatrix}, \qquad (\boldsymbol{A}^\top \boldsymbol{A})^{-1} = \frac{1}{nV(\boldsymbol{t})} \begin{pmatrix} \frac{1}{n}\sum_k t_k^2 & -M(\boldsymbol{t}) \\ -M(\boldsymbol{t}) & 1 \end{pmatrix},$$

with $M(\boldsymbol{t})$ and $V(\boldsymbol{t})$ as in Section 2.5. The estimator $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1)^\top$ coincides with that of Theorem 2.17, and the diagonal entries of $\sigma^2(\boldsymbol{A}^\top \boldsymbol{A})^{-1}$ give

$$\mathrm{Var}[\hat{\gamma}_0] = \frac{\sigma^2 \sum_k t_k^2}{n^2 V(\boldsymbol{t})} \qquad \text{and} \qquad \mathrm{Var}[\hat{\gamma}_1] = \frac{\sigma^2}{nV(\boldsymbol{t})}.$$

The slope is thus estimated more precisely if the temperatures t_k are spread out.

Remark 10.5 (Further results for linear models). The *Gauss–Markov theorem* states that $\langle \boldsymbol{c}, \hat{\boldsymbol{\gamma}} \rangle$ has the smallest variance among all unbiased estimators of $\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle$ that are linear in $\boldsymbol{X}$, for every $\boldsymbol{c} \in \mathbb{R}^s$. In the Gaussian linear model, moreover, $\hat{\boldsymbol{\gamma}} \sim \mathcal{N}\big(\boldsymbol{\gamma}, \sigma^2 (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\big)$, $(n - s)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-s}$, and $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}^2$ are independent, which generalises Theorem 6.9. This yields confidence intervals and t-tests for the coefficients γ_j; see [Geo13, SL03].

References

[BD01] P. J. Bickel and K. A. Doksum, *Mathematical statistics: Basic ideas and selected topics. Vol. I*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001.

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

[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.