

MATHEMATICAL STATISTICS – LECTURE NOTES

Luca Ganassali

Université Paris-Saclay

Last update: January 26, 2026

Disclaimer

These lecture notes constitute a work in progress: it may contain (hopefully, merely minor) mistakes. I am grateful to anybody whose wisdom help improve the quality of the notes.

Acknowledgments

If you help me improve these notes, your name will probably end up here :)

BIBLIOGRAPHY

- [1] Patrick Billingsley, *Probability and measure* (Third edition), New York: Wiley, 1995.
 - [2] Jean-François Le Gall, *Measure Theory, Probability, and Stochastic Processes*, Graduate Texts in Mathematics, Springer (Volume 295, 2022). Lecture notes: <https://www.imo.universite-paris-saclay.fr/~jean-francois.le-gall/IPPA2.pdf>.
 - [3] Lecture notes on "Théorie de la mesure, Intégration, Probabilités" by Stéphane Nonnenmacher, Classe sino-française, USTC, 2024. https://www.imo.universite-paris-saclay.fr/~stephane.nonnenmacher/enseign/Cours_USTC_Integration+Probabilites_2024.pdf
 - [4] Robert W. Keener, *Theoretical Statistics: Topics for a Core Course*, Springer Texts in Statistics, Springer New York, 2010.
 - [5] Alexandre B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Series in Statistics, 2009.
 - [6] Lecture notes on "Statistique mathématique" by Arnaud Guyader, Université Pierre et Marie Curie. <https://perso.lpsm.paris/~aguyader/files/teaching/M1/PolycopiePartie1.pdf>.
 - [7] Lecture notes on "Statistics" by Zacharie Naulet, Université Paris-Saclay, 2023.
- - Zacharie Naulet, *Lecture notes on Statistics*, Université Paris-Saclay, 2023-2024.
 - Billingsley 1999
 - Le Gall, ...
 - Van der Vaart, Asymptotic statistics

Contents

Chapter 1 – Probabilistic tools for the statistician	7
1.1 Basics on random vectors	7
1.1.1 Real random variables, random vectors, expectation and variance . . .	7
1.2 Operations on limits	8
1.2.1 Slutsky’s Lemma	9
1.2.2 Delta method	9
1.3 Classical concentration inequalities	10
1.3.1 Markov’s and (Bienaymé-)Chebyshev’s inequalities	10
1.3.2 Hoeffding’s inequality	11
1.3.3 Bernstein’s inequality	12
1.3.4 Chernoff method	13
1.4 Conditional distributions, conditional expectation	14
1.4.1 Discrete case	14
1.4.2 General case	14
1.4.3 Case where X, Y have a joint density with respect to a product measure	16
Chapter 2 – Statistical models, sufficiency and completeness	19
2.1 Some definitions and vocabulary	19
2.2 Dominated models	21
2.3 Sufficient statistics	24
2.3.1 Some definitions	24
2.3.2 Neyman-Fisher’s factorization	25
2.3.3 Minimal sufficiency	28
2.4 Complete statistics	28
2.4.1 Definition and properties	28
2.4.2 Ancillary statistics, Basu’s Theorem	29
Chapter 3 – Parametric estimation	33
3.1 Oblivious parametric estimation	33
3.1.1 Bias and quadratic risk	33
3.1.2 Method of moments	35
3.1.3 Maximum likelihood estimation	36
3.1.4 Asymptotic properties of estimators	38
3.2 No one beats mother Nature: the Cramér–Rao lower bound	39
3.2.1 Fisher Information in regular models	39
3.2.2 Properties of Fisher information	41
3.2.3 The Cramér–Rao Theorem	43
3.2.4 Limitations of the Cramér–Rao lower bound	45
3.3 Sufficiency and Rao-Blackwell theorem	45
3.4 Uniformly minimum-variance unbiased estimators	46
3.4.1 Lehmann-Scheffé theorem	47
3.4.2 Sufficient and necessary conditions: a geometric point of view	48
3.4.3 Some examples: the good and the bad of unbiased estimation	49

Chapter A – Standard distributions	i
A.1 Discrete distributions	i
A.2 Continuous distributions	ii
A.3 Distributions from the Gaussian world	ii
Chapter B – Reminder on standard probability theory	iii
B.1 Reminder on measure theory	iii
B.1.1 Measures	iii
B.1.2 Absolute continuity, Radon-Nikodym derivative	iv
B.1.3 Real random variables, random vectors, expectation and variance . . .	iv
B.2 Convergence of random variables	vi
B.2.1 Convergence in probability	vi
B.2.2 Almost sure convergence	vi
B.2.3 Convergence in distribution	vii
B.2.4 A criterion for convergence in distribution in the real case	viii
B.2.5 A general criterion for convergence in distribution in the multidimen- sional case	ix
B.3 Classical convergence theorems	x
B.3.1 The strong Law of Large Numbers	x
B.3.2 The Central Limit Theorem	x
Chapter C – Reminder on Gaussian vectors	xiii
C.1 Gaussian variables and Gaussian vectors	xiii
C.2 Cochran’s Theorem and geometric properties of Gaussian vectors	xiv
C.3 Two other classical distributions: Student and Fisher distributions	xvi

CHAPTER 1

PROBABILISTIC TOOLS FOR THE STATISTICIAN

Before delving into the core course in statistics, this first chapter introduces or recalls specific tools from probability theory which will be useful for statistics. We assume that the reader is already familiar with basic measure theory, random variables, convergence of random variables and classical convergence theorems (law of large numbers and central limit theorem in the multidimensional case). A general reminder on these can be found in Appendix B.

Throughout, we consider a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, that is a measurable space (Ω, \mathcal{F}) with measure \mathbb{P} having total mass 1.

1.1. Basics on random vectors

1.1.1. Real random variables, random vectors, expectation and variance

Definition 1.1 (Random variable, random vector). A *random variable*¹ is a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A *random vector of \mathbb{R}^d* ² is a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The law (or distribution) \mathbb{P}_X of a random vector X is defined for all borelian set $B \in \mathcal{B}(\mathbb{R})$ by $\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$.

Remark 1.1. Note that since the projection on the k -th coordinate is continuous hence measurable, if $X = (X_1, \dots, X_d)$ is a random vector in \mathbb{R}^d , each of its coordinates are random variables.

Definition 1.2 (Expectation, variance, covariance). Let X be a random variable. If X is integrable, we define its *expectation* as

$$\mathbb{E}[X] := \int X(\omega) d\mathbb{P}(\omega) = \int x d\mathbb{P}_X(x).$$

If moreover X^2 is integrable (we say that X has finite second moment), then so is X , and we define the *variance of X* as

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

Moreover, if X, Y are two random variables with finite second moment, their *covariance* is defined by

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

From the above definition, it is easily seen that the expectation is linear over the real vector space of integrable random variables. The covariance is a bilinear operator on the real

¹in this course, all random variables are real.

²in this course, all random vectors take their values in \mathbb{R}^d .

vector space of random variables with finite second moment, and the variance is its associated quadratic form. The variance is a positive quadratic form

Definition 1.3 (Expectation, covariance matrix of a random vector). Let $X = (X_1, \dots, X_d)$ be a random vector in \mathbb{R}^d . If X_1, \dots, X_d are integrable, the *expectation* of X is defined as

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])^\top \in \mathbb{R}^d.$$

If $Y = (Y_1, \dots, Y_\ell)$ is another random vector in \mathbb{R}^ℓ , and if coordinates of X, Y have finite second moments (we say that the vectors X, Y have finite second moment), the *covariance matrix* of (X, Y) is defined as

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] \in \mathbb{R}^{d \times \ell},$$

that is, for all $1 \leq i \leq d, 1 \leq j \leq \ell$, $[\text{Cov}(X, Y)]_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])] = \text{Cov}(X_i, Y_j)$. We define the *covariance matrix* of X by

$$\text{Var}(X) := \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \in \mathbb{R}^{d \times d},$$

that is, for all $1 \leq i, j \leq d$, $[\text{Var}(X)]_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \text{Cov}(X_i, X_j)$. Thus, $\text{Var}(X)$ is a symmetric matrix. These definitions coincide with the usual expectation and variance of a random variable when $d = 1$.

In their vectorial forms, the expectation and covariance operators inherit from their properties in dimension 1.

Proposition 1.1. Let X be a random vector in \mathbb{R}^d with a finite second-order moment. Let $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. Then $Y = AX + b$ is a random vector in \mathbb{R}^m which also has a finite second-order moment, and we have:

$$\mathbb{E}[Y] = A\mathbb{E}[X] + b \quad \text{and} \quad \text{Var}(Y) = A\text{Var}(X)A^\top.$$

Proof. Writing $Y = (Y_1, \dots, Y_m)$, it is readily seen that for all $1 \leq i \leq m$, $Y_i = \sum_{k=1}^d A_{i,k}X_k + b_i$, and by linearity of expectation in dimension 1, $\mathbb{E}[Y_i]$ is finite and $\mathbb{E}[Y_i] = \sum_{k=1}^d A_{i,k}\mathbb{E}[X_k] + b_i = (A\mathbb{E}[X] + b)_i$. The coordinates of Y are affine transformations of coordinates of X , so they still all have finite second moment. A direct computation gives

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}[(AX + b - (A\mathbb{E}[X] + b))(AX + b - (A\mathbb{E}[X] + b))^\top] \\ &= \mathbb{E}[(AX - A\mathbb{E}[X])(AX - A\mathbb{E}[X])^\top] \\ &= \mathbb{E}[A(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top A^\top] \\ &= A\text{Var}(X)A^\top, \end{aligned}$$

using now linearity of (vectorial) expectation. □

Remark 1.2. With the above property, we have that for all $a \in \mathbb{R}^d$, $a^\top \Sigma a = \text{Var}(a^\top X) \geq 0$. A covariance matrix is therefore always symmetric positive semidefinite.

For the interested reader, a reminder on Gaussian vectors can be found in Appendix C.

1.2. Operations on limits

In this section, we introduce basic tools to manipulate limits in distribution, which are useful in many occasions in statistics.

1.2.1. Slutsky's Lemma

Can we go from convergence in distribution of the marginals to that of the joint? Usually, no, because the marginals do not determine the joint. But, if one of the coordinates converges to a constant, then the limit joint has no choice: it must be the product distribution. This is exactly the result stated by Slutsky's Lemma.

Proposition 1.2 (Slutsky's Lemma). *Let $(X_n)_{n \geq 1}$, $(Y_n)_{n \geq 1}$, X be random vectors such that $X_n \xrightarrow[n \rightarrow \infty]{(d)} X$ and $Y_n \xrightarrow[n \rightarrow \infty]{(d)} c$ where c is a constant. Then, $(X_n, Y_n) \xrightarrow[n \rightarrow \infty]{(d)} (X, c)$.*

Remark 1.3. In particular, since convergence in distribution is stable by applying continuous functions (see Remark B.10), we have $X_n + Y_n \xrightarrow[n \rightarrow \infty]{(d)} X + c$, and when $c \in \mathbb{R}$, $X_n Y_n \xrightarrow[n \rightarrow \infty]{(d)} cX$.

Proof of Proposition 1.2. Assume X_n, X belong to \mathbb{R}^d and Y belongs to \mathbb{R}^m . Since convergence in distribution is preserved by applying continuous transformations, we can assume $c = 0_m$ without loss of generality (replace Y_n by $Y_n - c$). We will use Lévy's theorem, hence establishing the simple convergence of $\Phi_{(X_n, Y_n)}(s, t)$ to $\Phi_{(X, 0)}(s, t) = \Phi_X(s)$, for all $(s, t) \in \mathbb{R}^d \times \mathbb{R}^m$. Let $(s, t) \in \mathbb{R}^d \times \mathbb{R}^m$. We have

$$\begin{aligned} |\Phi_{(X_n, Y_n)}(s, t) - \Phi_{(X, 0)}(s, t)| &\leq |\Phi_{(X_n, Y_n)}(s, t) - \Phi_{(X_n, 0)}(s, t)| + |\Phi_{(X_n, 0)}(s, t) - \Phi_{(X, 0)}(s, t)| \\ &= |\Phi_{(X_n, Y_n)}(s, t) - \Phi_{X_n}(s)| + |\Phi_{X_n}(s) - \Phi_X(s)|. \end{aligned}$$

The second term converges to 0 thanks to Lévy's Theorem (Theorem B.3). For the first term, note that

$$|\Phi_{(X_n, Y_n)}(s, t) - \Phi_{X_n}(s)| = |\mathbb{E}[e^{is^\top X_n + it^\top Y_n} - e^{is^\top X_n}]| \leq \mathbb{E}[|e^{it^\top Y_n} - 1|].$$

Now, let $\varepsilon > 0$. Since $y \mapsto e^{it^\top y}$ is continuous at $y = 0_m$, there exists $\delta > 0$ such that if $\|Y_n\| \leq \delta$ then $|e^{it^\top Y_n} - 1| \leq \varepsilon$. The previous bound becomes:

$$\begin{aligned} |\Phi_{(X_n, Y_n)}(s, t) - \Phi_{X_n}(s)| &\leq \mathbb{E}[|e^{it^\top Y_n} - 1| \mathbf{1}_{\|Y_n\| \leq \delta}] + \mathbb{E}[|e^{it^\top Y_n} - 1| \mathbf{1}_{\|Y_n\| > \delta}] \\ &\leq \varepsilon + 2\mathbb{P}(\|Y_n\| > \delta). \end{aligned}$$

Since $Y_n \xrightarrow[n \rightarrow \infty]{(d)} 0$, $\|Y_n\| \xrightarrow[n \rightarrow \infty]{(d)} 0$ in \mathbb{R} , and $\pm\delta$ is a continuity point of the c.d.f. of the r.v. 0 which is $\mathbf{1}_{\geq 0}$, we have by Theorem B.2:

$$\mathbb{P}(\|Y_n\| > \delta) \xrightarrow[n \rightarrow \infty]{} 1 - \mathbf{1}_{\delta > 0} + \mathbf{1}_{-\delta > 0} = 0.$$

Thus, for n large enough, the previous bound is less or equal to 2ε . This is true for all $\varepsilon > 0$, and concludes the proof. \square

1.2.2. Delta method

Suppose that, for a sequence of random variables X_n and a sequence of constants v_n , we have the convergence in distribution

$$v_n(X_n - a) \xrightarrow[n \rightarrow \infty]{(d)} X,$$

as in the classical central limit theorem. We are interested in the behavior of a transformed quantity $v_n(g(X_n) - g(a))$ when g is a sufficiently smooth function.

For example, if g is affine, i.e., $g(x) = \alpha x + \beta$, then it is immediate that

$$v_n(g(X_n) - g(a)) = v_n(\alpha X_n + \beta - \alpha a - \beta) = \alpha v_n(X_n - a) \xrightarrow[n \rightarrow \infty]{(d)} \alpha X.$$

For a more general (nonlinear) function g , the limiting distribution of $v_n(g(X_n) - g(a))$ can be obtained using the derivative (or differential) of g at a . This is the essence of the *Delta method*.

Proposition 1.3 (Delta method (multidimensional case)). *Let $(X_n)_{n \geq 1}$ be random vectors of \mathbb{R}^d and $(v_n)_{n \geq 1}$ a positive real sequence such that $v_n \xrightarrow{n \rightarrow \infty} +\infty$. We assume that there exists $a \in \mathbb{R}^d$ and a random vector X of \mathbb{R}^d such that*

$$v_n(X_n - a) \xrightarrow[n \rightarrow \infty]{(d)} X.$$

Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be differentiable at point a . Then,

$$v_n(g(X_n) - g(a)) \xrightarrow[n \rightarrow \infty]{(d)} dg_a(X).$$

Remark 1.4. In dimensions $d = m = 1$, this translates to $v_n(g(X_n) - g(a)) \xrightarrow[n \rightarrow \infty]{(d)} g'(a)X$.

Proof of Proposition 1.3. First off, note that since $v_n(X_n - a) \xrightarrow[n \rightarrow \infty]{(d)} X$, we have

$$X_n = a + v_n(X_n - a) \times \frac{1}{v_n} \xrightarrow[n \rightarrow \infty]{(d)} a + X \times 0 = a,$$

by Slutsky's Lemma. Now, since g is differentiable at point a , we can write a Taylor expansion of $g(x)$ at $x = a$:

$$g(x) = g(a) + dg_a(x - a) + \|x - a\|\varepsilon(x),$$

where dg_a denotes the differential of g at point a , ε is continuous from $\mathbb{R}^d \setminus \{a\}$ to \mathbb{R}^m , and $\varepsilon(x) \xrightarrow{x \rightarrow a} 0$. We can then extend ε by continuity to a . Since $X_n \xrightarrow[n \rightarrow \infty]{(d)} a$ in distribution, then by continuity, $\varepsilon(X_n) \xrightarrow[n \rightarrow \infty]{(d)} \varepsilon(a) = 0$. Thus, we have for all n ,

$$g(X_n) - g(a) = dg_a(X_n - a) + \|X_n - a\|\varepsilon(X_n).$$

We get,

$$\begin{aligned} v_n(g(X_n) - g(a)) &= v_n dg_a(X_n - a) + v_n \|X_n - a\| \varepsilon(X_n) \\ &= dg_a(v_n(X_n - a)) + \|v_n(X_n - a)\| \varepsilon(X_n) \\ &\xrightarrow[n \rightarrow \infty]{(d)} dg_a(X). \end{aligned}$$

The last convergence follows from the fact that dg_a is linear thus continuous (in finite dimension), and $\|v_n(X_n - a)\| \varepsilon(X_n) \xrightarrow[n \rightarrow \infty]{(d)} \|X\| \times 0 = 0$ by Slutsky's Lemma. Then, the sum of the two terms converges to $dg_a(X)$ again by Slutsky's Lemma. \square

1.3. Classical concentration inequalities

Concentration inequalities are a useful tool for statistics since they will help us prove convergence in probability, high probability guarantees, or derive asymptotic confidence intervals.

1.3.1. Markov's and (Bienaymé-)Chebyshev's inequalities

We start with basics.

Proposition 1.4 (Markov's inequality). *Let X be a non-negative random variable and $p \geq 1$ such that $\mathbb{E}[X^p] < \infty$. Then, for all $x > 0$,*

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[X^p]}{x^p}.$$

Proof. It simply consists in writing $X^p = X^p \mathbf{1}_{X \geq x} + X^p \mathbf{1}_{X < x}$ and take the expectation (finite by assumption), which gives $\mathbb{E}[X^p] \geq x^p \mathbb{P}(X \geq x) + 0$, and the desired result. \square

By applying Markov's inequality to $X - \mathbb{E}[X]$ with $p = 2$, one gets Bienaymé-Chebyshev's inequality:

Proposition 1.5 (Bienaymé-Chebyshev's inequality). *Let X be a random variable with finite variance (and mean). Then, for all $t > 0$,*

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

Example 1.1. If $S_n \sim \text{Bin}(n, p)$, then $S_n/n \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}[S_n/n] = p$ by the law of large numbers. To establish a first concentration inequality, we can apply Bienaymé-Chebyshev's inequality (B-C hereafter) to S_n/n : its variance is $\frac{p(1-p)}{n}$, and thus for all $\varepsilon > 0$,

$$\mathbb{P}(|S_n/n - p| \geq \varepsilon) \leq \frac{p(1-p)}{\varepsilon^2 n} \leq \frac{1}{4\varepsilon^2 n}.$$

This result is informative but not strong enough to recover almost sure convergence, since the harmonic series diverges. Next, we can somehow improve this concentration with *Hoeffding's inequality*.

1.3.2. Hoeffding's inequality

Proposition 1.6 (Hoeffding's inequality). *Let X_1, \dots, X_n be independent random variables such that for all $1 \leq i \leq n$, $a_i \leq X_i \leq b_i$ almost surely. Let $S_n = X_1 + \dots + X_n$. Then, for all $t > 0$*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof of Proposition 1.6. Let us start with a Lemma.

Lemma 1.1. *If $X \in [a, b]$ a.s., then for all $s \in \mathbb{R}$, $\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leq \exp(\frac{s^2(b-a)^2}{8})$.*

With the previous Lemma, for all $t, s > 0$,

$$\begin{aligned} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(\exp(s(S_n - \mathbb{E}[S_n])) \geq \exp(st)) \\ &\leq \exp(-st) \mathbb{E}[\exp(s(S_n - \mathbb{E}[S_n]))] \leq \exp(-st) \prod_{i=1}^n \mathbb{E}[\exp(s(X_i - \mathbb{E}[X_i]))] \\ &\leq \exp(-st) \prod_{i=1}^n \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) = \exp\left(-st + \frac{s^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right), \end{aligned}$$

which is minimal for $s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$, and gives the desired result. For the symmetric result, consider the $-X_i$, and apply Hoeffding's inequality to $-b_i \leq -X_i \leq -a_i$. \square

Proof of Lemma 1.1. Wlog we assume that $\mathbb{E}[X] = 0$ so that $a \leq 0 \leq b$. Then, by convexity of $x \mapsto e^{sx}$ for all $s \in \mathbb{R}$, we have for all $x \in [a, b]$, $e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$. Taking expectations yields

$$\mathbb{E}[e^{sX}] \leq \frac{b}{b-a}e^{sa} + \frac{-a}{b-a}e^{sb}$$

the last term is $e^{sa}(1 - p + pe^{s(b-a)})$, with $p = -\frac{a}{b-a} \in [0, 1]$. For $u = s(b-a)$, the log of the last term is equal to $\psi(u) := -pu + \ln(1 - p + pe^u)$. We see that $\psi(0) = 0$, $\psi'(0) = 0$ and $\psi''(u) = \frac{(1-p)pe^u}{(1-p+pe^u)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2} \leq \frac{1}{4}$ by the AM-GM inequality. Taylor's formula implies that for all $u > 0$, there exists $v \in [0, u]$ such that $\psi(u) = \psi(0) + u\psi'(0) + \frac{u^2}{2}\psi''(v) \leq \frac{u^2}{8}$. \square

Example 1.2. We continue our previous example, where $S_n \sim \text{Bin}(n, p)$. Now, we can apply Hoeffding's inequality with $a_i = 0$ and $b_i = 1$. This gives that for all $\varepsilon > 0$,

$$\mathbb{P}(|S_n/n - p| \geq \varepsilon) = \mathbb{P}(|S_n - np| \geq \varepsilon n) \leq 2 \exp(-2\varepsilon^2 n).$$

This result is much more powerful than B-C for a constant deviation ε . In particular, it is strong enough to recover almost sure convergence by Borel-Cantelli's Lemma.

1.3.3. Bernstein's inequality

In Hoeffding's inequality, the almost sure boundedness of the random variables (X_i) is used to obtain upper bounds on the Laplace transform $s \mapsto \mathbb{E}[e^{sX_i}]$ that do not depend on the variance of X_i . In this sense, the bound corresponds to a worst-case scenario. When additional information on the variances of the X_i 's is available, one can obtain sharper concentration results. A fundamental example of such an improvement is provided by *Bernstein's inequality*.

Proposition 1.7 (Bernstein's inequality). *Let X_1, \dots, X_n be independent random variables such that for all $1 \leq i \leq n$, $|X_i - \mathbb{E}[X_i]| \leq M$ almost surely. Let $S_n = X_1 + \dots + X_n$ and denote $V_n = \sum_{i=1}^n \text{Var}(X_i)$. Then, for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{t^2}{2(V_n + Mt/3)}\right),$$

and

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2(V_n + Mt/3)}\right).$$

Proof of Proposition 1.7.

Lemma 1.2. *Suppose that $|X| \leq c$ almost surely and $\mathbb{E}[X] = 0$. For any $t > 0$,*

$$\mathbb{E}[e^{tX}] \leq \exp\left(t^2 \sigma^2 \left(\frac{e^{tc} - 1 - tc}{(tc)^2}\right)\right),$$

where $\sigma^2 = \text{Var}(X)$.

Proof. Expand the exponential in series and write

$$\mathbb{E}[e^{tX}] = 1 + 0 + \sum_{r=2}^{\infty} \frac{t^r \mathbb{E}[X^r]}{r!} = 1 + t^2 \sigma^2 F \leq \exp(t^2 \sigma^2 F),$$

where $F := \sum_{r=2}^{\infty} \frac{t^{r-2} \mathbb{E}[X^r]}{r! \sigma^2}$. For $r \geq 2$, we have, using $|X| \leq c$, $\mathbb{E}[X^r] = \mathbb{E}[X^{r-2} X^2] \leq$

$c^{r-2}\sigma^2$, and therefore

$$F \leq \sum_{r=2}^{\infty} \frac{t^{r-2}c^{r-2}}{r!} = \frac{1}{(tc)^2} \sum_{r=2}^{\infty} \frac{t^r c^r}{r!} = \frac{e^{tc} - tc - 1}{(tc)^2}.$$

□

Now, back the proof of Bernstein's inequality, assume wlog that $\mathbb{E}[X_i] = 0$ for all $1 \leq i \leq n$. With the previous Lemma, for any $t, s > 0$,

$$\begin{aligned} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(S_n \geq t) = \mathbb{P}(e^{sS_n} \geq e^{st}) \leq e^{-st} \mathbb{E}[e^{sS_n}] \\ &\leq e^{-st} \exp\left(\sum_{i=1}^n s^2 \text{Var}(X_i) \left(\frac{e^{sc} - 1 - sc}{(sc)^2}\right)\right) = \exp\left(-st + \frac{e^{sc} - 1 - sc}{c^2} V_n\right) \end{aligned}$$

By taking the derivative, the previous right hand side is minimal when $s = \frac{1}{c} \log(1 + tc/V_n)$, and for this value of s , we get

$$\exp\left(-st + \frac{e^{sc} - 1 - sc}{c^2} V_n\right) = -\frac{V_n}{c^2} h(tc/V_n),$$

with $h : u \mapsto (1 + u) \log(1 + u) - u$. The proof is concluded by checking that, for all $u \geq 0$, $h(u) \geq \frac{u^2}{2+2u/3}$. For the symmetric result, consider again applying the one-side concentration bound to the $-X_i$. □

Example 1.3. We continue our previous example where $S_n \sim \text{Bin}(n, p)$. Here, each $X_i \in \{0, 1\}$, so $M = 1$ and $\text{Var}(X_i) = p(1 - p)$. Then

$$V_n = \sum_{i=1}^n \text{Var}(X_i) = np(1 - p).$$

Applying Bernstein's inequality, for all $\varepsilon > 0$,

$$\mathbb{P}(|S_n/n - p| \geq \varepsilon) = \mathbb{P}(|S_n - np| \geq n\varepsilon) \leq 2 \exp\left(-\frac{n\varepsilon^2}{2(p(1-p) + \varepsilon/3)}\right).$$

Notice that compared with Hoeffding's bound $2 \exp(-2\varepsilon^2 n)$, Bernstein's bound can be much tighter when $\varepsilon \leq p \ll 1$, because it uses the actual variance, $p(1 - p)$, rather than the maximal possible range, which is $1/4$.

1.3.4. Chernoff method

The fundamental assumption in Hoeffding's inequality is that the variabls are bounded. We can however obtain exponential concentration bounds in more generality, when 'merely' assuming that X has finite exponential moments, that is $\mathbb{E}[e^{\lambda X}] < \infty$ for all $\lambda > 0$. In this case, for all $c \in \mathbb{R}$ and all $\lambda > 0$, Markov's inequality yields

$$\mathbb{P}(X \geq c) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda c)) \leq \exp(-\lambda c) \mathbb{E}[e^{\lambda X}] =: \phi(\lambda)$$

and we conclude by minimising ϕ (or equivalently $\log \phi$), if we know how to do it. This simple yet powerful trick is called the *Chernoff method* and is at the heart of a myriad of concentration inequalities (including Hoeffding's and Bernstein's, as seen before).

1.4. Conditional distributions, conditional expectation

This part is largely inspired from [4], Section 6.

Consider X a random vector in \mathbb{R}^d , and Y a random vector in \mathbb{R}^m , defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The fundamental motivation for conditional distributions is the following. If X is observed and we learn that $X = x$, then the law of Y can be modified (or, updated) taking account of the new information given by the observation $X = x$.

1.4.1. Discrete case

When X is discrete, this update can be done by the standard formula for conditional probabilities. The set of possible values of X is $\mathcal{X}_0 := \{x \in \mathbb{R}^d, \mathbb{P}(X = x) > 0\}$. Define for all $x \in \mathcal{X}_0$, all Borel sets $B \in \mathcal{B}(\mathbb{R}^m)$,

$$Q_x(B) := \mathbb{P}(Y \in B \mid X = x) = \frac{\mathbb{P}(Y \in B, X = x)}{\mathbb{P}(X = x)}. \quad (1.1)$$

For all $x \in \mathcal{X}_0$, Q_x is a probability measure on \mathbb{R}^m called the *conditional distribution* for Y given $X = x$.

1.4.2. General case

Now, these conditional distributions should also exist more generally, in particular when X is a continuous random variable. However, defining them is not as direct as in the discrete case, since this would imply conditioning to a null probability event in (1.1) ($\mathbb{P}(X = x) = 0$ is x is not an atom of the law of X). We give hereafter the formal definition.

Definition 1.4 (Conditional distribution). A function $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^m) \rightarrow [0, 1]$ is a *conditional distribution of Y given X* if

- (i) for all $x \in \mathbb{R}^d$, $Q_x(\cdot) := Q(x, \cdot)$ is a probability measure on $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$,
- (ii) for all $B \in \mathcal{B}(\mathbb{R}^m)$, $x \mapsto Q_x(B)$ is measurable,
- (iii) for all³ measurable all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[|f(X, Y)|] < \infty$, then for all $x \in \mathbb{R}^d$, $y \mapsto f(x, y)$ is Q_x -integrable, for all $y \in \mathbb{R}^m$, $x \mapsto \int f(x, y) dQ_x(y)$ is \mathbb{P}_X -integrable, and

$$\mathbb{E}[f(X, Y)] = \iint f(x, y) dQ_x(y) d\mathbb{P}_X(x).$$

In particular, for all $A \in \mathcal{B}(\mathbb{R}^d)$, $B \in \mathcal{B}(\mathbb{R}^m)$,

$$\mathbb{P}(X \in A, Y \in B) = \int_A Q_x(B) d\mathbb{P}_X(x).$$

Remark 1.5. For all $B \in \mathcal{B}(\mathbb{R}^m)$, $Q_x(B)$ is unique \mathbb{P}_X -almost everywhere by point (iii) hereabove. Note however that the null sets depend on B , hence we cannot conclude directly that there exists a global null-measure set N such that $Q_x(B)$ is unique for all $B \in \mathcal{B}$, $x \in \mathbb{R}^d \setminus N$. In our setting, this technical issue is solved since $\mathcal{B}(\mathbb{R}^m)$ is countably generated⁴. Throughout, we will, by abuse of terminology, refer to Q as *the* conditional distribution for Y given X .

In our setting, X, Y are random vectors and it can be proven that such conditional distributions always exist (see [1], Theorem 33.3). This definition is non constructive, but conditional distributions can be obtained easily when X and Y have a joint density with respect to a product measure $\mu \times \nu$, see next Section.

³note that we need (ii) to define properly the integral in (iii)

⁴every open set in \mathbb{R}^m is a countable union of balls with rational radii and center in \mathbb{Q}^m .

Remark 1.6. If X and Y are independent, then $Q_x(\cdot) = \mathbb{P}(Y \in \cdot)$ is the conditional distribution for Y given X , that is, $Y|X \sim Y$.

When we have a conditional distribution, we can define *conditional expectations* as follows.

Definition 1.5 (Conditional expectation). Let Q be the conditional distribution for Y given X . For all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[|f(X, Y)|] < \infty$, the conditional expectation of $f(X, Y)$ given $X = x$, denoted $\mathbb{E}[f(X, Y) | X = x]$, is defined by

$$\mathbb{E}[f(X, Y) | X = x] := \int f(x, y) dQ_x(y).$$

Note that this quantity is well-defined by point (iii) of Definition 1.4. The *conditional expectation of $f(X, Y)$ given X* , denoted $\mathbb{E}[f(X, Y) | X]$, is the random variable $E \circ X$, where $E : x \mapsto \mathbb{E}[f(X, Y) | X = x]$.

Remark 1.7. Note that by the above definition, the conditional expectation is positive and linear.

Remark 1.8. Note that by Remark 1.6, if X and Y are independent, then for all integrable f , $\mathbb{E}[f(X, Y) | X = x] = f(x, Y)$. In particular, if X and Y are independent, $\mathbb{E}[Y | X] = Y$.

A fundamental result in statistics is the following:

Proposition 1.8 (Law of total expectation). For all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[|f(X, Y)|] < \infty$, we have

$$\mathbb{E}[f(X, Y)] = \mathbb{E}[\mathbb{E}[f(X, Y) | X]].$$

This is a consequence of point (iii) in the definition.

Definition 1.6 (Conditional variance). Let Q be a conditional distribution for Y given X . For all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[f^2(X, Y)] < \infty$, the conditional variance of $f(X, Y)$ given $X = x$, denoted $\text{Var}(f(X, Y) | X = x)$, is defined by

$$\text{Var}(f(X, Y) | X = x) = \mathbb{E}[f^2(X, Y) | X = x] - \mathbb{E}[f(X, Y) | X = x]^2.$$

We define the *conditional variance of $f(X, Y)$ given X* by $\text{Var}(f(X, Y) | X) = \mathbb{E}[f^2(X, Y) | X] - \mathbb{E}[f(X, Y) | X]^2$.

Proposition 1.9 (Law of total variance). For all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[f^2(X, Y)] < \infty$, we have

$$\text{Var}(f(X, Y)) = \mathbb{E}[\text{Var}(f(X, Y) | X)] + \text{Var}(\mathbb{E}[f(X, Y) | X]).$$

Proof.

$$\begin{aligned} \text{Var}(f(X, Y)) - \mathbb{E}[\text{Var}(f(X, Y) | X)] &= \\ &= \mathbb{E}[f^2(X, Y)] - \mathbb{E}[\mathbb{E}[f^2(X, Y) | X]] + \mathbb{E}[\mathbb{E}[f(X, Y) | X]^2] - \mathbb{E}[f(X, Y)]^2 \\ &= 0 + \mathbb{E}[\mathbb{E}[f(X, Y) | X]^2] - \mathbb{E}[\mathbb{E}[f(X, Y) | X]]^2 \\ &= \text{Var}(\mathbb{E}[f(X, Y) | X]). \end{aligned}$$

□

Remark 1.9. The definition of conditional expectation (resp. variance), as well as the law of total expectation (resp. variance) can easily be extended to the case where f has values in some \mathbb{R}^k , $k \geq 1$. This follows exactly the steps of Definition 1.3. To do so, we need to define the conditional covariance for two real functions f_1, f_2 of (X, Y) as

$$\text{Cov}(f(X, Y) | X = x) = \mathbb{E}[f_1(X, Y)f_2(X, Y) | X = x] - \mathbb{E}[f_1(X, Y) | X = x] \mathbb{E}[f_2(X, Y) | X = x].$$

1.4.3. Case where X, Y have a joint density with respect to a product measure

Let $Z = (X, Y)$ be a random vector in \mathbb{R}^{d+m} . Assume that the law of Z has a density $p_{(X,Y)}$ with respect to $\mu \times \nu$, where μ and ν are non-negative σ -finite measures on \mathbb{R}^d and \mathbb{R}^m . This density $p_{(X,Y)}$ is called the *joint density* of X and Y , and for all $C \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^m)$ ($= \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^m)$),

$$\mathbb{P}(Z \in C) = \iint \mathbf{1}_C(x, y) p_{(X,Y)}(x, y) d\mu(x) d\nu(y).$$

By Fubini's theorem, the order of integration can be inversed, hence for all $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}(Z \in A \times \mathbb{R}^m) = \iint \mathbf{1}_A(x) p_{(X,Y)}(x, y) d\mu(x) d\nu(y) \\ &= \int_A \left(\int p_{(X,Y)}(x, y) d\nu(y) \right) d\mu(x). \end{aligned}$$

This shows that X has a density $p_X : x \mapsto \int p_{(X,Y)}(x, y) d\nu(y)$ with respect to μ . This density is called the *marginal density* of X . Similarly, Y has marginal density $p_Y : y \mapsto \int p_{(X,Y)}(x, y) d\mu(x)$ w.r.t. ν .

Now, in our setting, there is a simple way to obtain conditional distributions, themselves with density.

Proposition 1.10. *Suppose X and Y have a joint density with respect to a product measure $\mu \times \nu$. Let p_X be the marginal density of X and let $E = \{x \in \mathbb{R}^d, p_X(x) > 0\}$. For $x \in E$, define*

$$p_{Y|X}(y|x) = \frac{p_{(X,Y)}(x, y)}{p_X(x)},$$

and Q_x the probability measure with density $y \mapsto p_{Y|X}(y|x)$ w.r.t. ν . When $x \notin E$, take $p_{Y|X}(y|x) = p_0$, where p_0 is a fixed density of an arbitrary probability distribution P_0 , and let $Q_x = P_0$. Then $Q : \mathcal{X} \times \mathcal{B}(\mathbb{R}^m) \rightarrow [0, 1]$ is a conditional distribution for Y given X .

Proof. Q_x is always a probability measure since for all $x \in E$,

$$\int p_{Y|X}(y|x) d\nu(y) = \frac{1}{p_X(x)} \int p_{(X,Y)}(x, y) d\nu(y) = 1.$$

Point (ii) follows from measurability of the density $p_{(X,Y)}$. To show (iii) we will even show that for all $f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}[|f(X, Y)|] < \infty$, then

$$\mathbb{E}[f(X, Y)] = \iint f(x, y) dQ_x(y) dP_X(x).$$

Note that up to changing $p_{(X,Y)}(x, y)$ to $p_{(X,Y)}(x, y) \mathbf{1}_E(x)$ (these two densities agree almost everywhere since $\mathbb{P}(X \in E^c) = 0$), we can assume that $p_{(X,Y)}(x, y) = 0$ if $x \notin E$. Then, for such an f ,

$$\begin{aligned} \mathbb{E}[f(X, Y)] &= \iint f(x, y) p_{(X,Y)}(x, y) d\mu(x) d\nu(y) \\ &= \iint f(x, y) p_{Y|X}(y|x) d\nu(y) p_X(x) d\mu(x) \\ &= \iint f(x, y) dQ_x(y) dP_X(x). \end{aligned}$$

Applying this to proper indicator functions gives (iii). □

Example 1.4. Consider μ the counting measure on $\{0, \dots, k\}$ and ν the Lebesgue measure on \mathbb{R} . Define

$$p_{(X,Y)}(x, y) = \binom{k}{x} y^x (1-y)^{k-x} \mathbf{1}_{x \in \{0, \dots, k\}, y \in]0, 1[}.$$

Let us see what happens in this model. First, one draws a uniform variable Y in $[0, 1]$, then conditionally on $Y = y$ we draw $X \sim \text{Bin}(k, y)$. Intuitively, it appears that the marginal distribution of X is a uniform distribution on $\{0, \dots, k\}$. Let us prove this. X has marginal density

$$p_X(x) = \int_0^1 \binom{k}{x} y^x (1-y)^{k-x} dy = \frac{1}{k+1},$$

for all $x \in \{0, \dots, k\}$. We used the result

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

This is, as one can expect, the uniform distribution on $\{0, \dots, k\}$. It is easy to check that the marginal density of Y is constant to 1.

Now, because $p_Y(y) = 1$, $p_{X|Y}(x|y) = \binom{k}{x} y^x (1-y)^{k-x}$, a binomial distribution, hence we denote $X|Y = y \sim \text{Bin}(k, y)$.

Similarly,

$$\begin{aligned} p_{Y|X}(y|x) &= (k+1) \binom{k}{x} y^x (1-y)^{k-x} \\ &= \frac{\Gamma(k+2)}{\Gamma(x+1)\Gamma(k-x+1)} y^{x+1-1} (1-y)^{k-x+1-1}, \end{aligned}$$

which the Beta distribution, and so $Y|X = x \sim \text{Beta}(x+1, k-x+1)$.

CHAPTER 2

STATISTICAL MODELS, SUFFICIENCY AND COMPLETENESS

This chapter gives an introduction to the main concept at the heart of mathematical statistics: statistical models. In these models, we develop the notions of statistical sufficiency and statistical completeness which hold intuitive roles, and prepare ourselves for a framework for parametric estimation, which is the object of the next chapter.

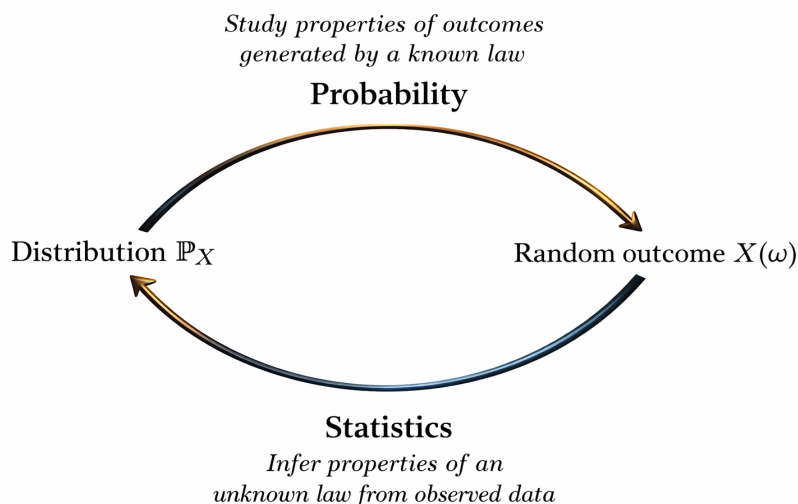


Figure 2.1 – *Probability and statistics viewed as inverse problems of each other.*

We give below the definitions of an observation and a statistical model.

2.1. Some definitions and vocabulary

Definition 2.1 (Observation). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $(\mathcal{X}, \mathcal{F})$ be a measurable space. An *observation* is a realization of a random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow (\mathcal{X}, \mathcal{F})$. In other words, $x \in \mathcal{X}$ is an observation if there is $\omega \in \Omega$ such that $x = X(\omega)$. Most often in this course, \mathcal{X} will be a subset of some \mathbb{R}^n .

Remark 2.1. The general abstract space $(\Omega, \mathcal{A}, \mathbb{P})$ can be arbitrarily complicated and is not uniquely defined, so there is no reason that we can say anything about \mathbb{P} based on observations. But we are now used to the fact that this general space is irrelevant to us: what matters to the statistician is to make statements about the law of X , $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$ looked at on the space $(\mathcal{X}, \mathcal{F})$. Our ambitious program is thus to learn information about \mathbb{P}_X based

on an observation $X(\omega)$. Note that, this can be viewed as the inverse problem of probability theory, where one study the properties of X knowing \mathbb{P}_X , as illustrated on Figure 2.1.

Definition 2.2 (Statistical model). A (statistical) model is a triplet

$$\mathcal{M} = (\mathcal{X}, \mathcal{F}, \mathcal{P})$$

where $(\mathcal{X}, \mathcal{F})$ is a measurable space (referred to as the *space of realizations*) and \mathcal{P} is a class of probability measures on $(\mathcal{X}, \mathcal{F})$. In this course, we will always denote

$$\mathcal{P} = (\mathbb{P}_\theta)_{\theta \in \Theta},$$

emphasizing that \mathcal{P} is indexed by the set Θ .

In line with Chapter 1, in this course we will consider only real random variables, and random vectors. Therefore, \mathcal{X} will always be a subset of some \mathbb{R}^d .

In frequentist statistics, the act of modeling consists in assuming that an observation x does not arise arbitrarily, but rather according to a statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$. In other words, there exists some $\theta \in \Theta$ such that

$$x = X(\omega), \quad \text{with } X \sim \mathbb{P}_\theta.$$

This fundamental assumption is often referred to as the *well-defined assumption*.

Remark 2.2 (On the consequence of the choice of a model). One has to be careful with modelling. Indeed, modelling is always schematizing, a choice of model says a lot about the assumptions we make. These assumptions are crucial and need to be explicit, discussed, motivated (by common sense, expert knowledge, literature). Moreover, there is no free lunch: a good model for the statistician is also, roughly speaking, *a model where computations are doable and yet something interesting happens*, whereas a good model for the practitioner is a model that renders every aspect of the studied phenomenon: these two objectives are by essence contradictory and need to be balanced. As mathematicians, we often care for the first objective. But it is important to keep in mind that modelling reality as to less complex than it is can lead to erroneous conclusions, and sometimes severe mistakes.

Under the well-defined assumption, the goals of a statistician typically involve addressing the following questions:

- (i) **Estimation.** Estimate a quantity of interest related to the distribution \mathbb{P}_θ (e.g., the mean, a parameter θ , the density function, etc.). This also involves quantifying the error of the estimation and designing estimators with desirable properties.
- (ii) **Hypothesis testing.** Decide whether a given assumption about \mathbb{P}_θ is consistent with the observed data (e.g., can we conclude that the data follows a particular distribution?). Such decisions inherently carry a risk of error, which must be carefully quantified.
- (iii) **Prediction.** In machine learning, when the observation has the form of $x = (z_i, y_i)_{1 \leq i \leq n}$, z_i are feature vectors and y_i are labels, coming from i.i.d. copies of a random pair (Z, Y) , we are interested in predicting the distribution of Y conditional on Z . This enables us to predict the value of a new label y_{new} given a new feature vector z_{new} .

Definition 2.3 (Parametric model). A statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ is called *parametric* if Θ is a subset of some \mathbb{R}^p space. Space Θ is then referred to as the *parameter space* and the surjective mapping $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is called a *parameterization* of the model.

In parametric models, the whole complexity of class $\mathcal{P} = (\mathbb{P}_\theta)_{\theta \in \Theta}$ is captured by at most p real numbers, which makes life easier for the statistician. In order to recover some information

on θ based on the observations, we want to ensure that a given distribution in \mathcal{P} corresponds to exactly one θ . This is exactly the definition of *identifiability*.

We conclude this section by giving examples of statistical models.

Definition 2.4 (Identifiability). A parametric statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ is called *identifiable* if its parameterization $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is also injective, that is, if different parameters yield different distributions.

Example 2.1 (Survey model). We run a survey on n individuals asking them whether they like pizza. Assuming all individuals' tastes are independent and identically distributed, a possible model is

$$\mathcal{M} = (\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), (\text{Ber}(\theta)^{\otimes n})_{\theta \in [0, 1]}), \quad (2.1)$$

where $\text{Ber}(\theta)$ is the Bernoulli distribution of parameter θ . Note that the fact that the laws in \mathcal{M} are product of the same distribution comes from the i.i.d. assumption. This model is parametric, and it is identifiable, since for any $\theta \in [0, 1]$, $X = (X_1, \dots, X_n) \sim \mathbb{P}_\theta$, $\mathbb{E}_\theta[X_1] = \theta$, thus $\theta \mapsto \mathbb{E}_\theta[X_1]$ is injective, and so is the parameterization $\theta \mapsto \mathbb{P}_\theta$.

Example 2.2 (A propagation model). We study the propagation of information among a chain of individuals. Individuals arrive sequentially, and the information received by individual i depends on the information of individual $i - 1$ and additive random noise, as follows: $X_1 \sim \mathcal{N}(x, 1)$, and for $2 \leq i \leq n$,

$$X_i = \rho X_{i-1} + \sqrt{1 - \rho^2} \xi_i, \quad (2.2)$$

where the ξ_i are i.i.d. $\mathcal{N}(0, 1)$ random variables, and $\rho \in [-1, 1]$ is a parameter controlling the strength and direction of influence. In this example, we define the statistical model by specifying the form of the distribution of X , with the parameterization made implicit (here, $\theta = (x, \rho)$). We have

$$\theta(x, \rho) \mapsto (\mathbb{E}_\theta[X_1], \mathbb{E}_\theta[X_1 X_2]) = (x, \rho),$$

thus the model is injective.

Example 2.3 (Regression model). We collect n measurements y_1, \dots, y_n of the energy consumption of a household at times t_1, \dots, t_n . We want to model the consumption variation with time. A possible way to do so is to write that for all $1 \leq i \leq n$,

$$Y_i = f(t_i) + \sigma \xi_i,$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$, the ξ_i are i.i.d. standard gaussians (noise of the measurements), and $\sigma > 0$. Here, the Y_i are random only through the noise variables ξ_i , not the t_i , which are deterministic. The model is non parametric. It is a regression model¹. Without no further assumption of function f , this model is not identifiable : two distinct functions f_1, f_2 which coincide on the set $\{t_1, \dots, t_n\}$ yield the same distribution of $Y = (Y_1, \dots, Y_n)$. The parameterization is however injective in σ , considering for instance $\text{Var}_\theta(Y_1) = \sigma^2$.

The remainder of this chapter is developed in a general setting where the model need not be parametric. However, following Definition 2.2, we will always index the class of distributions by $\theta \in \Theta$.

2.2. Dominated models

A wide range of statistical models contain a family of probability measures that are all absolutely continuous with respect to the same reference measure ν . Such models are called *dominated*. We refer to Definition B.2 for a reminder on absolute continuity.

¹when f is moreover assumed to be an affine function, this model will be the main focus of Chapter ??.

Definition 2.5 (ν -dominated models). A statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ is *dominated by a measure ν* (or *ν -dominated*) if every $P \in \mathcal{M}$ is absolutely continuous with respect to ν . When this is the case, we denote

$$\mathcal{M} \ll \nu.$$

Remark 2.3. A model \mathcal{M} is always dominated by (a multiple of) the counting measure on its space of realizations \mathcal{X} . In particular, there are always infinitely many dominating measures.

Remark 2.4. If the model is finite, that is Θ is finite, the measure

$$\nu = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{P}_\theta$$

is easily seen to be a dominating measure which is also a probability measure.

Remark 2.5. If $\mathcal{M} \ll \nu$ and ν is σ -finite, by the Radon-Nykodym theorem (all \mathbb{P}_θ are finite thus σ -finite, see Appendix B for a reminder), every \mathbb{P}_θ has a density with respect to ν . These densities will be useful to the statistician. Since there are infinitely many dominating measures, the reference measure should always be clearly stated.

Example 2.4 (Survey model, continued). In the survey model of Example 2.1, \mathcal{M} is dominated by $\nu^{\otimes n}$, where ν is the counting measure on \mathbb{N} . For all $\theta \in [0, 1]$, the density of \mathbb{P}_θ with respect to $\nu^{\otimes n}$ is

$$p_\theta : x \mapsto \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \mathbb{1}_{x_i \in \{0,1\}}.$$

We could also choose $(\delta_0 + \delta_1)^{\otimes n}$ as a dominating measure of \mathcal{M} . In this case, the density of \mathbb{P}_θ w.r.t. $(\delta_0 + \delta_1)^{\otimes n}$ is

$$p_\theta : x \mapsto \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

If $\mathcal{M} \ll \nu$ and $\nu \ll \nu'$, then $\mathcal{M} \ll \nu'$. A natural choice for a dominating measure could thus be a dominating measure which is minimal with respect to the preorder \ll , as defined hereafter.

Definition 2.6 (Minimal dominating measure). A measure ν_0 is a *minimal dominating measure* of a model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, \mathcal{P})$ if:

- $\mathcal{M} \ll \nu_0$,
- for all measure ν satisfying $\mathcal{M} \ll \nu$, then $\nu_0 \ll \nu$.

This definition implies that a minimal dominating measure is not unique, but that all minimal dominating measures are equivalent.

Example 2.5 (Survey model, continued). In the survey model of Example 2.1, $\mathcal{M} \ll (\delta_0 + \delta_1)^{\otimes n}$, which is minimally dominant. Indeed, if ν is a measure such that $\mathcal{M} \ll \nu$, and $N \in \mathcal{B}(\mathbb{R}^n)$ is such that $\nu(N) = 0$, then $\mathbb{P}_{1/2}(N) = 0$ thus N does not contain any element of $\{0, 1\}^n$. Consequently $(\delta_0 + \delta_1)^{\otimes n}(N) = 0$.

As seen before, any model is dominated by the counting measure on its space of realizations \mathcal{X} . We may wonder whether any statistical model admits a minimal dominating measure, and if such a measure can be a probability measure.

We end this chapter with a following answer result to this question: as soon as the model is dominated by a σ -finite measure, then a minimal dominating measure exists, and can be chosen to be a probability measure. It is a classical result that has been first established by Halmos and Savage in 1949.

Theorem 2.1 (Countable equivalent subset Theorem). *If a statistical model $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ is dominated by a σ -finite measure ν , then there exists a sequence $(\theta_n)_{n \in \mathbb{N}} \in \Theta^{\mathbb{N}}$ and non-negative weights $(\lambda_n)_{n \in \mathbb{N}}$ such that $Q = \sum_{n=0}^{\infty} \lambda_n \mathbb{P}_{\theta_n}$ is a minimal dominating probability measure.*

Proof of Theorem 2.1. We first start with the following Lemma.

Lemma 2.1. *Let ν be a σ -finite measure on $(\mathcal{X}, \mathcal{F})$. There exists ν' a probability measure on $(\mathcal{X}, \mathcal{F})$ such that $\nu' \sim \nu$ (that is $\nu \ll \nu'$ and $\nu' \ll \nu$).*

Proof of Lemma 2.1. Since ν is σ -finite, there exists a partition $(A_n)_{n \geq 1}$ of \mathcal{X} such that $\nu(A_n) < \infty$ for all $n \geq 1$. Choose a sequence $(c_n)_{n \geq 1}$ such that $c_n = 0$ if $\nu(A_n) = 0$, $c_n > 0$ if $\nu(A_n) > 0$, and $\sum_{n=1}^{\infty} c_n = 1$. Define the measure ν' for all $A \in \mathcal{F}$ by

$$\nu'(A) := \sum_{n \geq 1 : \nu(A_n) > 0} c_n \frac{\nu(A \cap A_n)}{\nu(A_n)}.$$

Then ν' is a probability measure and ν' is equivalent to ν . \square

In view of Lemma 2.1, without loss of generality, we can assume that ν is a probability measure.

We use the notation $\mathcal{P} = (\mathbb{P}_{\theta})_{\theta \in \Theta}$ and for $\theta \in \Theta$ we denote by p_{θ} the Radon–Nikodym derivative $\frac{d\mathbb{P}_{\theta}}{d\nu}$. We define

$$\nu^{\star} := \sup_{(\theta_n)_{n \in \mathbb{N}} \in \Theta^{\mathbb{N}}} \nu \left(\bigcup_{n \in \mathbb{N}} \{p_{\theta_n} > 0\} \right). \quad (2.3)$$

It is well defined and finite since ν is finite.

Step 1. We claim that the supremum in (2.3) is in fact a maximum, attained by some sequence $(\theta_n^*)_{n \in \mathbb{N}} \in \Theta^{\mathbb{N}}$. To see this, choose a doubly indexed sequence $(\theta_{m,n})_{m,n \in \mathbb{N}} \in \Theta^{\mathbb{N}}$ such that, for all $m \in \mathbb{N}$,

$$\nu \left(\bigcup_{n \in \mathbb{N}} \{p_{\theta_{m,n}} > 0\} \right) \geq \nu^{\star} - \frac{1}{2^m}.$$

The sequence

$$\left(\bigcup_{m \leq M} \bigcup_{n \in \mathbb{N}} \{p_{\theta_{m,n}} > 0\} \right)_{M \geq 0}$$

increases to

$$S := \bigcup_{m,n \in \mathbb{N}} \{p_{\theta_{m,n}} > 0\}.$$

By monotone continuity of measures,

$$\nu(S) = \lim_{M \rightarrow \infty} \nu \left(\bigcup_{m \leq M} \bigcup_{n \in \mathbb{N}} \{p_{\theta_{m,n}} > 0\} \right) = \nu^{\star}.$$

Thus the supremum in (2.3) is attained.

Step 2. Write $(\theta_n^*)_{n \in \mathbb{N}} \in \Theta^{\mathbb{N}}$ for the sequence attaining the maximum in (2.3). Define

$$Q := \sum_{n \in \mathbb{N}} 2^{-n} \mathbb{P}_{\theta_n^*}, \quad q := \sum_{n \in \mathbb{N}} 2^{-n} p_{\theta_n^*}.$$

Then q is a Radon–Nikodym derivative of Q with respect to ν , and by definition,

$$\{q > 0\} = \bigcup_{n \in \mathbb{N}} \{p_{\theta_n^*} > 0\}.$$

Consequently,

$$\nu(\{q > 0\}) = \nu^*.$$

Let us show that $\mathcal{M} \ll Q$. Take $N \in \mathcal{F}$ such that $Q(N) = 0$, and $\theta \in \Theta$. We write

$$\begin{aligned} \mathbb{P}_\theta(N) &= \mathbb{P}_\theta(N \cap \{p_\theta = 0\}) + \mathbb{P}_\theta(N \cap \{p_\theta > 0\} \cap \{q > 0\}) \\ &\quad + \mathbb{P}_\theta(N \cap \{p_\theta > 0\} \cap \{q = 0\}). \end{aligned}$$

The first term is zero since $\mathbb{P}_\theta(N \cap \{p_\theta = 0\}) = \int_{N \cap \{p_\theta = 0\}} p_\theta d\nu = 0$. For the second term, note that $Q(N) = 0$ implies $\int_N q d\nu = 0$. Then, $\int_{N \cap \{q > 0\}} q d\nu = 0$, $\mathbb{1}_{N \cap \{q > 0\}} q = 0$ ν -a.e., then $\mathbb{1}_{N \cap \{q > 0\}} = 0$ ν -a.e., then $\nu(N \cap \{q > 0\}) = 0$, and therefore $\mathbb{P}_\theta(N \cap \{q > 0\}) = 0$ since $\mathbb{P}_\theta \ll \nu$. For the third term, observe that because ν^* is optimal and $\{q > 0\} = \bigcup_{n \in \mathbb{N}} \{p_{\theta_n^*} > 0\}$, we have

$$\nu^* \geq \nu(\{q > 0\} \cup \{p_\theta > 0\}) \geq \nu(\{q > 0\}) = \nu^*,$$

thus $\nu^* = \nu(\{q > 0\} \cup \{p_\theta > 0\}) = \nu(\{q > 0\}) + \nu(\{q = 0\} \cap \{p_\theta > 0\}) = \nu^* + \nu(\{q = 0\} \cap \{p_\theta > 0\})$. Thus $\mathbb{P}_\theta(\{q = 0\} \cap \{p_\theta > 0\}) = 0$ which proves that the third term is 0. Thus $\mathbb{P}_\theta(N) = 0$. As θ was arbitrary, this shows $\mathcal{M} \ll Q$.

Step 3. It is immediate that Q is a probability measure. It remains to show that Q is minimal dominating. Let ν' be any measure such that $\mathcal{M} \ll \nu'$. Since $\mathbb{P}_{\theta_n^*} \ll \nu'$ for all $n \geq 0$, then $Q \ll \nu'$. Thus Q is minimal. \square

Remark 2.6. Note that the σ -finiteness of ν is crucial for the existence of a dominating probability measure. Consider for instance the model where $\mathcal{P} = (\delta_x)_{x \in \mathcal{X}}$, with \mathcal{X} uncountable. This model is dominated by the counting measure on \mathcal{X} , but not dominated by any probability measure. Indeed, such a probability measure ν would have to verify that $\nu(\{x\}) > 0$ for all $x \in \mathcal{X}$, which is impossible.

2.3. Sufficient statistics

2.3.1. Some definitions

In statistics, we often use functions $T(X)$ of the outcome X to infer properties of the underlying distribution of X . From the probabilistic point of view, such measurable functions of X are just other random variables or random vectors. Statisticians, however, often use another terminology and simply call such a function $T(X)$ a *statistic*.

Definition 2.7 (Statistic). For X a random variable in $(\mathcal{X}, \mathcal{F})$, and $T : (\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ a measurable function, the random vector $T(X)$ is called a *statistic*.

The goal of the statistician is to find statistics that estimate or test some properties of the law \mathbb{P}_θ of X , or that reduces the data while preserving the amount of information about \mathbb{P}_θ contained in it. Such statistics are called *sufficient statistics*.

Here again, in line with Chapter 1, in this course, \mathcal{Y} will be (a subset of) some \mathbb{R}^m .

Definition 2.8 (Sufficient statistic). Let $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model. A statistic $S(X)$ is *sufficient* if for any $\theta \in \Theta$, the conditional probability distribution under \mathbb{P}_θ of the data X given the statistic $S(X)$ does not depend of θ .

In words, if two observations x and x' have the same value $S(x) = S(x')$, where S is a sufficient statistic, the statistician cannot conclude that they come from different distributions in $(\mathbb{P}_\theta)_{\theta \in \Theta}$.

Remark 2.7. X is always a sufficient statistic, by definition. But it is not very interesting, of course, because the game is to compactify the information in a sufficient statistic.

Example 2.6. Let us continue on the survey model of Example 2.1. Consider the statistic $S(X) = \sum_{i=1}^n X_i$. Let us show that $S(X)$ is sufficient. For all $\theta \in [0, 1]$, for all $0 \leq s \leq n$,

$$\mathbb{P}_\theta(S(X) = s) = \binom{n}{s} \theta^s (1 - \theta)^{n-s},$$

and

$$\mathbb{P}_\theta(X = (x_1, \dots, x_n), S(X) = s) = \mathbb{1}_{\sum_{i=1}^n x_i = s} \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \mathbb{1}_{\sum_{i=1}^n x_i = s} \theta^s (1 - \theta)^{n-s},$$

thus the conditional distribution of X given $S(X)$ is

$$\mathbb{P}_\theta(X = (x_1, \dots, x_n) | S(X) = s) = \frac{\mathbb{1}_{\sum_{i=1}^n x_i = s} \theta^s (1 - \theta)^{n-s}}{\binom{n}{s} \theta^s (1 - \theta)^{n-s}} = \frac{\mathbb{1}_{\sum_{i=1}^n x_i = s}}{\binom{n}{s}},$$

which does not depend on θ . Therefore, $S(X)$ is a sufficient statistic in this model.

2.3.2. Neyman-Fisher's factorization

We give hereafter intuitive and useful characterization of sufficient measures in the case where the model is dominated by a σ -finite measure. In this case, every \mathbb{P}_θ has a density, in which sufficient statistics can be naturally read.

Theorem 2.2 (Neyman-Fisher's factorization Theorem). *Consider a statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ dominated by a σ -finite measure ν . For $\theta \in \Theta$, let p_θ be the density of \mathbb{P}_θ with respect to ν . Let $S : (\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ be a statistic. Then the following are equivalent:*

- (i) $S(X)$ is a sufficient statistic;
- (ii) there exists a measurable function $h : \mathcal{X} \rightarrow \mathbb{R}_+$ and for all $\theta \in \Theta$ a measurable function $g_\theta : \mathcal{Y} \rightarrow \mathbb{R}_+$ such that for all $x \in \mathcal{X}$,

$$p_\theta(x) = h(x)g_\theta(S(x)), \quad \nu\text{-a.e.}$$

Proof. By Lemma 2.1, we can assume that ν is a probability distribution, and a mixture of a countable number of elements in $(\mathbb{P}_\theta)_{\theta \in \Theta}$.

Introduce the following notations.

- when $X \sim \nu$, we denote by $\mathbb{P}_{X \sim \nu}$ the probability distribution and $\mathbb{E}_{X \sim \nu}$ the expectation, $G = S\# \nu$ the distribution of $S(X)$.
- when $X \sim \mathbb{P}_\theta$, we denote $G_\theta = S\# \mathbb{P}_\theta$ the distribution of $S(X)$.

(i) \implies (ii) First, note that $G_\theta \ll G$ for all $\theta \in \Theta$. Indeed, if N is such that $0 = G(N) = \nu(S^{-1}(N))$, then $0 = \mathbb{P}_\theta(S^{-1}(N)) = G_\theta(N)$. G being a probability distribution, it is σ -finite, and we can thus denote, for all $\theta \in \Theta$, g_θ the density of $S(X)$ w.r.t. G when $X \sim \mathbb{P}_\theta$.

Suppose that $S(X)$ is sufficient. Then, the conditional distribution of X given $S(X) = s$ under any \mathbb{P}_θ does not depend on θ : denote it by P_s . For any $\theta \in \Theta$, any Borel set B ,

$$\mathbb{P}_\theta(X \in B) = \mathbb{E}_\theta[\mathbb{1}_{X \in B}]$$

$$\begin{aligned}
 &= \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbb{1}_{X \in B} \mid S(X)]] \\
 &= \mathbb{E}_\theta[P_{S(X)}(B)] \\
 &= \int P_s(B) g_\theta(s) dG(s) \\
 &= \int \left(\int \mathbb{1}_B(x) dP_s(x) \right) g_\theta(s) dG(s) \\
 &= \int \left(\int \mathbb{1}_B(x) \mathbb{1}_{S(x)=s} dP_s(x) \right) g_\theta(s) dG(s) \\
 &= \int \left(\int \mathbb{1}_B(x) \mathbb{1}_{S(x)=s} dP_s(x) \right) g_\theta(S(x)) dG(s) \\
 &= \int \left(\int \mathbb{1}_B(x) dP_s(x) \right) g_\theta(S(x)) dG(s) \\
 &= \iint \mathbb{1}_B(x) g_\theta(S(x)) dP_s(x) dG(s).
 \end{aligned}$$

Define the distribution $\tilde{\mathbb{P}}$ such that on any Borel set C

$$\tilde{\mathbb{P}}(C) := \int P_s(C) dG(s) = \iint \mathbb{1}_C(x) dP_s(x) dG(s),$$

so that for any integrable f ,

$$\int f(x) d\hat{P}(x) = \iint f(x) dP_s(x) dG(s).$$

The expression of $\mathbb{P}_\theta(X \in B)$ now writes

$$\mathbb{P}_\theta(X \in B) = \int_B g_\theta(S(x)) d\tilde{\mathbb{P}}(x), \quad (2.4)$$

which shows that \mathbb{P}_θ has density $x \mapsto g_\theta(S(x))$ w.r.t. $\tilde{\mathbb{P}}$.

Now, let us show $\tilde{\mathbb{P}} \ll \nu$. Assume that N is such that $\nu(N) = 0$. Then for all $\theta \in \Theta$, $0 = \mathbb{P}_\theta(N) = \int P_s(N) dG_\theta(s)$, which means that $G_\theta(N') = 0$ where $N' = \{s : P_s(N) > 0\}$. Thus $\mathbb{P}_\theta(S(X) \in N') = 0$ for all $\theta \in \Theta$, and since ν is a mixture of a countable number of elements in $(\mathbb{P}_\theta)_{\theta \in \Theta}$, then $\nu(S(X) \in N') = 0 = G(N')$. This exactly means that G -almost every s does not belong to N' , that is $P_s(N) = 0$ for G -almost every s . Finally, this yields $\tilde{\mathbb{P}}(N) = \int P_s(N) dG(s) = 0$.

Now, we conclude by Radon-Nikodym theorem ($\tilde{\mathbb{P}}$ and ν are finite hence σ -finite). There exists a density $h = d\tilde{\mathbb{P}}/d\nu$, and (2.4) shows that \mathbb{P}_θ has density $g_\theta(S(x))h(x)$ with respect to ν .

(ii) \implies (i) Assume (ii) holds. First, we will show that (ii) implies that $S(X)$ has a particular density w.r.t. G when $X \sim \mathbb{P}_\theta$. To do this, we introduce and Q_s the conditional distribution of X given $S(X) = s$ when $X \sim \nu$. For any $\theta \in \Theta$, and any bounded continuous function f ,

$$\begin{aligned}
 \mathbb{E}_\theta[f(S(X))] &= \int f(S(x)) p_\theta(x) d\nu(x) \\
 &= \int f(S(x)) h(x) g_\theta(S(x)) d\nu(x) \\
 &= \mathbb{E}_{X \sim \nu}[f(S(X)) h(X) g_\theta(S(X))]
 \end{aligned}$$

$$\begin{aligned}
 &= \int \left(\int f(s) g_\theta(s) dG(s) \right) h(x) dQ_s(x) \\
 &= \iint f(s) g_\theta(s) w(s) dG(s),
 \end{aligned}$$

where $w(s) = \int h(x) dQ_s(x)$. This computation shows that $S(X)$ has density $s \mapsto g_\theta(s)w(s)$ with respect to G when $X \sim \mathbb{P}_\theta$.

Now, we show that the conditional distribution of X given $S(X)$ does not depend on θ . To do this, we postulate the following form. For all $s \in \mathcal{Y}$, define \tilde{Q}_s the distribution with density

$$x \mapsto \mathbb{1}_{w(s) \neq 0} h(x)/w(s) + \mathbb{1}_{w(s)=0} \eta_0(x)$$

with respect to Q_s , with η_0 arbitrary.

Let us show that $N = \{s \in \mathcal{Y} : w(s) = 0\}$ is of null G -measure. Recall that by definition $G(N) = \nu(S^{-1}(N))$. Now for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(S^{-1}(N)) = \mathbb{P}_\theta(S(X) \in N) = \int \mathbb{1}_{s \in N} g_\theta(s) w(s) dG(s) = 0.$$

Since ν is a mixture of a countable number of elements in $(\mathbb{P}_\theta)_{\theta \in \Theta}$, we have $\nu(S(X) \in N) = 0 = G(N)$.

Then for any $\theta \in \Theta$, any continuous bounded function f ,

$$\begin{aligned}
 \mathbb{E}_\theta[f(X, S(X))] &= \mathbb{E}_{X \sim \nu}[f(X, S(X)) g_\theta(S(X)) h(X)] \\
 &= \iint f(x, s) g_\theta(s) h(x) dQ_s(x) dG(s) \\
 &= \iint f(x, s) \underbrace{\frac{h(x)}{w(s)} dQ_s(x)}_{d\tilde{Q}_s(x)} \underbrace{g_\theta(s) w(s) dG(s)}_{dG_\theta(s)}.
 \end{aligned}$$

The third equality is justified $N = \{s \in \mathcal{Y} : w(s) = 0\}$ is of null G -measure.

We conclude as follows. The above computation shows that \tilde{Q}_s is the conditional distribution of X given $S(X) = s$ when $X \sim \mathbb{P}_\theta$. Since \tilde{Q}_s does not depend on θ , $S(X)$ is a sufficient statistic. \square

Example 2.7 (Uniform i.i.d. model). If the X_1, \dots, X_n are i.i.d. of distribution $\text{Unif}([0, \theta])$ for $\theta > 0$, the model is dominated, e.g. by the Lebesgue measure on \mathbb{R} . For all $\theta > 0$, the joint density of X_1, \dots, X_n writes

$$p_\theta(x_1, \dots, x_n) = \frac{1}{\theta^n} \mathbb{1}_{\max_i(x_i) < \theta} \mathbb{1}_{\min_i(x_i) > 0},$$

hence by Neyman-Fisher's Theorem, $S(X) := \max_i(X_i)$ is a sufficient statistic.

Corollary 2.1. Consider a statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ dominated by a σ -finite measure ν . Let S be a sufficient statistic, of the form $S = f(S')$ with f a measurable mapping and S' another statistic. Then, S' is also sufficient.

Proof. Just write the densities $p_\theta(x) = h(x)g_\theta(S(x)) = h(x)g_\theta(f(S'(x)))$ and apply Neyman-Fisher's Theorem (Theorem 2.2). \square

Remark 2.8. In particular, if g is a one-to-one mapping, then $g(S)$ is still sufficient, since $S = g^{-1}(g(S))$.

2.3.3. Minimal sufficiency

It is easy to see that one can always add extra information to a sufficient statistic to make it still sufficient. For instance, in Example 2.7, $\max_i X_i$ is sufficient, so $(\max_i X_i, \sum_i X_i)$ is also sufficient. But it contains too much information. There must be a minimality notion associated with sufficiency.

We first need the definition of \mathcal{M} -almost sure properties.

Definition 2.9 (\mathcal{M} -almost sure properties). On a model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$, a property is \mathcal{M} -almost sure or holds \mathcal{M} -almost surely if it is true \mathbb{P}_θ -a.s. for all $\theta \in \Theta$, that is if the event N on which the property is not verified satisfies $\mathbb{P}_\theta(N) = 0$ for all $\theta \in \Theta$.

Remark 2.9. In particular, any ν -a.s. property is \mathcal{M} -a.s. if $\mathcal{M} \ll \nu$.

Definition 2.10 (Minimal sufficiency). A statistic $S(X)$ is *minimal sufficient* on a model \mathcal{M} if

- $S(X)$ is sufficient,
- for any sufficient statistic S' , there exists a measurable f such that $S = f(S')$, \mathcal{M} -a.s.

Another class of statistics of interest are statistics which contain no superfluous information. These are exactly the *complete statistics*, defined below.

2.4. Complete statistics

2.4.1. Definition and properties

Definition 2.11 (Completeness). A statistic $S(X)$ is complete in a model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ if for any measurable function f ,

$$(\forall \theta \in \Theta, \mathbb{E}_\theta[f(S)] = 0) \implies f(S) = 0, \mathcal{M}\text{-a.s.}$$

Remark 2.10. Note that $S(X) = 1$ is always a complete statistic. $S(X)$ containing no superfluous information does not mean that it contains interesting information...

A key interest of this completeness notion lies in the following result.

Theorem 2.3. *If S is sufficient and complete for a model $(\mathcal{X}, \mathcal{F}, \mathcal{M})$, then it is minimal sufficient.*

Proof. Let S' be another sufficient statistic. Recall that S is \mathbb{R}^k valued, for some $k \geq 1$. Since we can find a measurable bijection $u : \mathbb{R}^k \rightarrow [0, 1]^k$, wlog we can assume that S takes its values in $[0, 1]^k$ (otherwise apply the proof to $u(S)$ which remains sufficient and complete, to find that $u(S)$ is of the form $u(S) = f(S')$ and then $S = u^{-1} \circ f \circ S'$).

Define $H_{1,\theta}(S') = \mathbb{E}_\theta[S | S']$ (well defined since S is bounded). Note that since S' is sufficient and S is a measurable function of X , $H_{1,\theta}(S')$ does not depend on θ , and we shall denote it $H_1(S')$. Our ultimate goal will be to show that $S = H_{1,\theta}(S')$, \mathcal{M} -a.s.

To do so, let us also define $H_{2,\theta}(S) = \mathbb{E}_\theta[H_1(S') | S]$ (well defined since by the law of total expectation $H_1(S')$ has a finite expectation under all \mathbb{P}_θ), which does not depend on θ for the same reasons. We denote it $H_2(S)$.

Step 1. We will first show that $S = H_2(S)$, \mathcal{M} -a.s. By the law of total expectation, for all $\theta \in \Theta$,

$$\mathbb{E}_\theta[S] = \mathbb{E}_\theta[\mathbb{E}_\theta[S | S']] = \mathbb{E}_\theta[H_1(S')] = \mathbb{E}_\theta[\mathbb{E}_\theta[H_1(S') | S]] = \mathbb{E}_\theta[H_2(S)].$$

Thus, for all $\theta \in \Theta$, $\mathbb{E}_\theta[(\text{id} - H_2)(S)] = 0$. By completeness of S , we can conclude that $S = H_2(S)$, \mathcal{M} -a.s.

Step 2. We are now going to show that $H_1(S') = H_2(S)$, \mathcal{M} -a.s., which will conclude the proof since $S = H_2(S) = H_1(S')$ \mathcal{M} -a.s., and H_1 is a measurable function. By the law of total variance (all variances exist because S is bounded), for all $\theta \in \Theta$,

$$\begin{aligned} \text{Var}_\theta(H_2(S)) &= \mathbb{E}_\theta[\text{Var}_\theta(H_2(S) | S')] + \text{Var}_\theta(\mathbb{E}_\theta[H_2(S) | S']) \\ &= \mathbb{E}_\theta[\text{Var}_\theta(H_2(S) | S')] + \text{Var}_\theta(\mathbb{E}_\theta[S | S']) \\ &= \mathbb{E}_\theta[\text{Var}_\theta(H_2(S) | S')] + \text{Var}_\theta(H_1(S')) \\ &= \mathbb{E}_\theta[\text{Var}_\theta(H_2(S) | S')] + \mathbb{E}_\theta[\text{Var}_\theta(H_1(S') | S)] + \text{Var}_\theta(\mathbb{E}_\theta[H_1(S') | S]) \\ &= \mathbb{E}_\theta[\text{Var}_\theta(H_2(S) | S')] + \mathbb{E}_\theta[\text{Var}_\theta(H_1(S') | S)] + \text{Var}_\theta(H_2(S)). \end{aligned}$$

The second equality is justified by step 1. This proves that \mathbb{P}_θ -a.s., $\text{Var}_\theta(H_2(S) | S') = 0$, that is, \mathbb{P}_θ -a.s., $H_2(S) = \mathbb{E}_\theta[H_2(S) | S'] = \mathbb{E}_\theta[S | S'] = H_1(S')$. Since this is true for all $\theta \in \Theta$, we conclude that $H_1(S') = H_2(S)$ holds \mathcal{M} -a.s. \square

Example 2.8. In the uniform i.i.d. model (Example 2.7), we saw that $S = \max_i(X_i)$ is a sufficient statistic. We can find the density of S under \mathbb{P}_θ , it is $s \mapsto ns^{n-1}/\theta^n \mathbb{1}_{[0,\theta]}(s)$. Let us now show that S is also complete. Let f be a function such that $\mathbb{E}_\theta[f(S)] = 0$ for all $\theta > 0$, that is $\int_0^\theta f(s)s^{n-1}ds = 0$ for all $\theta > 0$. Function $g : s \mapsto f(s)s^{n-1}$ as null integral on every interval on \mathbb{R}_+ , hence on all borelian sets by the monotone class Lemma, hence on the borelian $B = \{s : f(s) > 0\}$. This means that $f = 0$ Lebesgue-almost surely. By Theorem 2.3, S is minimal sufficient.

2.4.2. Ancillary statistics, Basu's Theorem

Definition 2.12 (Ancillary statistic). In a model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$, a statistic $A(X)$ is ancillary if its distribution under any \mathbb{P}_θ does not depend on θ .

For instance, if $S(X)$ is sufficient and X integrable, then $A(X) = \mathbb{E}[X | S(X)]$ is ancillary. Basu's theorem shows that the space of ancillary statistics is somewhat orthogonal to the space of sufficient complete statistics.

Theorem 2.4 (Basu's Theorem, 1955). *Let $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a model. Let $A(X), S(X)$ be statistics such that $A(X)$ is ancillary and $S(X)$ is complete and sufficient. Then, for all $\theta \in \Theta$, under \mathbb{P}_θ ,*

$$A(X) \perp\!\!\!\perp S(X).$$

In addition to its illustrative value, Theorem 2.4 can be used to prove the independence of two variables using purely statistical arguments. This may sometimes avoids complicated calculations, and always does constitute a more elegant proof.

Proof. We show that conditioning $A(X)$ to $S(X)$ never changes its distribution. Let $\theta \in \Theta$ and B a Borel set. Since A is ancillary, $\mathbb{P}_\theta(A(X) \in B)$ does not depend on θ : denote it $P_A(B)$. Since S is sufficient, $\mathbb{E}_\theta[\mathbb{1}_B(A(X)) | S(X)]$ does not depend on θ : denote it $P_A(B | S(X))$.

Besides, by the law of total expectation,

$$\mathbb{P}_\theta(A(X) \in B) = \mathbb{E}_\theta[\mathbb{1}_B(A(X))] = \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbb{1}_B(A(X)) | S(X)]] = \mathbb{E}_\theta[P_A(B | S(X))],$$

thus for all $\theta \in \Theta$, $\mathbb{E}_\theta[P_A(B | S(X)) - P_A(B)] = 0$. Since S is complete, this means that $P_A(B | S(X)) = P_A(B)$, \mathcal{M} -a.s. This is sufficient to conclude. Indeed, for any $\theta \in \Theta$, any Borel sets B, C ,

$$\begin{aligned} \mathbb{P}_\theta(A(X) \in B, S(X) \in C) &= \mathbb{E}_\theta[\mathbb{1}_B(A(X))\mathbb{1}_C(S(X))] \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbb{1}_B(A(X))\mathbb{1}_C(S(X)) | S(X)]] \\ &= \mathbb{E}_\theta[\mathbb{1}_C(S(X))\mathbb{E}_\theta[\mathbb{1}_B(A(X)) | S(X)]] \end{aligned}$$

$$\begin{aligned} &= \mathbb{E}_\theta[\mathbb{1}_C(S(X))\mathbb{E}_\theta[\mathbb{1}_B(A(X)) | S(X)]] \\ &= \mathbb{P}_\theta(A(X) \in B)\mathbb{P}_\theta(S(X) \in C), \end{aligned}$$

which proves independence. \square

A celebrated application of Basu's Theorem is a proof of the independence between the empirical mean and variance of an i.i.d. Gaussian sample.

Example 2.9. Consider the Gaussian i.i.d. model where X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ variables. Assume that the variance $\sigma^2 > 0$ is known while the mean $\mu \in \mathbb{R}$ is unknown. Define the empirical mean and variance:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Under \mathbb{P}_μ , the density of X with respect to the Lebesgue measure on \mathbb{R}^n writes

$$p_\mu(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

For any $x \in \mathbb{R}^n$, we use the decomposition

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2,$$

which yields the factorization

$$p_\mu(x) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2\right)}_{h(x)} \underbrace{\exp\left(-\frac{n}{2\sigma^2} (\bar{x}_n - \mu)^2\right)}_{g_\mu(\bar{x}_n)}.$$

By the Neyman–Fisher factorization theorem (Theorem 2.2), $S(X) = \bar{X}_n$ is a sufficient statistic. Let us show that it is complete. Using standard properties of Gaussian variables, under \mathbb{P}_μ , $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. Let f be a measurable function such that for all $\mu \in \mathbb{R}$, $\mathbb{E}_\mu[f(\bar{X}_n)] = 0$. This expectation can be written as

$$\int_{\mathbb{R}} f(x) \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2} (x - \mu)^2\right) dx = 0 \quad \text{for all } \mu.$$

Equivalently, for all $\mu \in \mathbb{R}$,

$$\int_{\mathbb{R}} f(x + \mu) \exp\left(-\frac{x^2}{2v^2}\right) dx = 0,$$

with $v = \sigma/\sqrt{n}$. This identity means that for $Z \sim \mathcal{N}(0, v)$, $\psi(\mu) := \mathbb{E}[f(\mu + Z)] = 0$ for all $\mu \in \mathbb{R}$. Multiply by $e^{-is\mu}$ and integrate with respect to μ :

$$\int_{\mathbb{R}} \mathbb{E}[f(\mu + Z)] e^{-is\mu} d\mu = 0.$$

By Fubini's theorem, we may exchange expectation and integration, obtaining

$$\mathbb{E}\left[\int_{\mathbb{R}} f(\mu + Z) e^{-is\mu} d\mu\right] = 0.$$

Using the change of variables $x = \mu + Z$, we get

$$\int_{\mathbb{R}} f(\mu + Z) e^{-is\mu} d\mu = e^{isZ} \int_{\mathbb{R}} f(x) e^{-isx} dx.$$

Therefore,

$$\mathbb{E} [e^{isZ}] \int_{\mathbb{R}} f(x) e^{-isx} dx = 0.$$

Since $Z \sim \mathcal{N}(0, v)$, its characteristic function is $\mathbb{E} [e^{isZ}] = \exp(-vs^2/2)$, which is positive for all $s \in \mathbb{R}$. Therefore,

$$\int_{\mathbb{R}} f(x) e^{-isx} dx = 0 \quad \text{for all } s \in \mathbb{R}.$$

By injectivity of the Fourier transform in $L^1(\mathbb{R})$, this implies that $f(x) = 0$ almost everywhere. Thus, \bar{X}_n is a complete sufficient statistic.

Now, remark that if $Y_i = X_i - \mu$, then under \mathbb{P}_μ the Y_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. But

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2,$$

In particular, the distribution of S_n^2 does not depend on μ , so S_n^2 is ancillary. By Basu's theorem, \bar{X}_n and S_n^2 are independent.

In ??, we will see another proof of the independence of Example 2.9, using geometric properties of Gaussian vectors.

CHAPTER 3

PARAMETRIC ESTIMATION

This chapter first presents the basic theory of parametric estimation: definitions, the moment and maximum likelihood methods. Then, we will be interested in optimality theory for estimation, namely the Cramer-Rao bound, and the Rao-Blackwell theorem for sufficient statistics.

3.1. Oblivious parametric estimation

We describe the general context of parametric estimation hereafter. Assume that we are given a parametric model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$, with $\Theta \subseteq \mathbb{R}^d$. Based on the observation $X(\omega)$ where $X \sim \mathbb{P}_\theta$ for some unknown $\theta \in \Theta$, the statistician wants to estimate some quantity $\varphi(\theta)$, where $\varphi : \Theta \rightarrow \mathbb{R}^\ell$. Most of the time, $\varphi(\theta)$ is simply θ , but it can also be a smaller part of θ or non-trivial transformations of θ , such as $\mathbb{P}_\theta(X \geq 0)$, for instance.

To estimate $\varphi(\theta)$ the game is to design a statistic $T(X)$ which will estimate $\varphi(\theta)$ properly. We call this statistic an *estimator of $\varphi(\theta)$* . In order to emphasize the fact that $T(X)$ is an estimator of $\varphi(\theta)$, we will sometime denote it by $T(X) = \widehat{\varphi(\theta)}$ or more simply $T(X) = \widehat{\varphi}$.

3.1.1. Bias and quadratic risk

To measure the quality of an estimator $T(X) = \widehat{\varphi}$, the most natural quantity to look at the average quadratic distance from $T(X)$ to the true value $\varphi(\theta)$. Hereafter, we denote by $\|\cdot\|$ the canonical Euclidean norm on (any) \mathbb{R}^ℓ .

Definition 3.1 (Quadratic risk). Assume $\text{Var}_\theta(T(X))$ is finite for all $\theta \in \Theta$. The *quadratic risk of $T(X)$* (when viewed as an estimator of $\varphi(\theta)$) is defined for all $\theta \in \Theta$ as

$$R_\theta(T(X)) = \mathbb{E}_\theta[\|T(X) - \varphi(\theta)\|^2].$$

This quadratic risk is very much related to the notion of *bias*, defined hereafter.

Definition 3.2 (Bias). Assume $T(X)$ is integrable for all $\theta \in \Theta$. The *bias of $T(X)$* (when viewed as an estimator of $\varphi(\theta)$) is defined for all $\theta \in \Theta$ as

$$b_\theta(T(X)) = \mathbb{E}_\theta[T(X)] - \varphi(\theta).$$

$T(X)$ is said to be *unbiased* if for all $\theta \in \Theta$, $b_\theta(T) = 0$ that is $\mathbb{E}_\theta[T(X)] = \varphi(\theta)$.

A fundamental identity, very simple yet very meaningful, is the following, which one can think of as the *statistician's Pythagorean Theorem*.

Proposition 3.1 (Statistician's Pythagorean Theorem). Assume $\text{Var}_\theta(T(X))$ is finite for all $\theta \in \Theta$. We have the identity, for all $\theta \in \Theta$,

$$R_\theta(T(X)) = \|b_\theta(T(X))\|^2 + \text{Tr}(\text{Var}_\theta(T(X))).$$

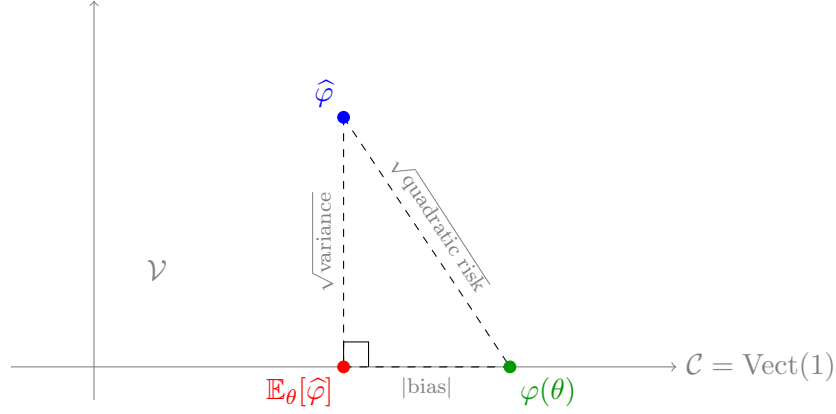


Figure 3.1 – Illustration of the Statistician's Pythagorean Theorem in dimension 1. Here, \mathcal{V} is the space of real random variables (defined on Ω), and $\mathcal{C} = \text{Vect}(1)$ is the space of constant random variables. The orthogonal projection of $\hat{\varphi}$ onto \mathcal{C} is $\mathbb{E}_\theta[\hat{\varphi}]$.

Proof. We use the shorthand notation T for $T(X)$. It suffices to write, for all $\theta \in \Theta$, that

$$\begin{aligned} R_\theta(T) &= \mathbb{E}_\theta[\|T - \mathbb{E}_\theta[T] + \mathbb{E}_\theta[T] - \varphi(\theta)\|^2] \\ &= \mathbb{E}_\theta[\|T - \mathbb{E}_\theta[T]\|^2] + \mathbb{E}_\theta[\|\mathbb{E}_\theta[T] - \varphi(\theta)\|^2] + 2(\mathbb{E}_\theta[T] - \varphi(\theta))^\top \mathbb{E}_\theta[T - \mathbb{E}_\theta[T]] \\ &= \mathbb{E}_\theta[\|T - \mathbb{E}_\theta[T]\|^2] + \|b_\theta(T)\|^2, \end{aligned}$$

and we have

$$\begin{aligned} \mathbb{E}_\theta[\|T - \mathbb{E}_\theta[T]\|^2] &= \mathbb{E}_\theta[\text{Tr}[(T - \mathbb{E}_\theta[T])^\top (T - \mathbb{E}_\theta[T])]] \\ &= \mathbb{E}_\theta[\text{Tr}[(T - \mathbb{E}_\theta[T])(T - \mathbb{E}_\theta[T])^\top]] \quad (\text{cyclicity of the trace}) \\ &= \text{Tr}[\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])(T - \mathbb{E}_\theta[T])^\top]] \quad (\text{linearity of Tr and } \mathbb{E}_\theta) \\ &= \text{Tr}(\text{Var}_\theta(T)). \end{aligned}$$

□

Example 3.1. Let's go back to our favorite survey model (2.1). A natural estimator for θ (here, $\varphi = \text{id}$) is given by the sample mean $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$. But is this estimator good? Let us compute its quadratic risk. For all $\theta \in [0, 1]$, $\mathbb{E}_\theta[\bar{X}_n] = \theta$, so it is unbiased. Now, since the X_i are i.i.d.,

$$\text{Var}_\theta(\bar{X}_n) = \frac{1}{n^2} \times n \text{Var}_\theta(X_1) = \frac{\theta(1-\theta)}{n}.$$

The quadratic risk of \bar{X}_n goes to 0 with speed $O(1/n)$, unless θ is degenerate, and can be uniformly bounded in θ by $\frac{1}{4n}$.

The main criterion to compare two estimators $\hat{\varphi}_1$ and $\hat{\varphi}_2$ of the same quantity $\varphi(\theta)$ is their quadratic risk. We say that $\hat{\varphi}_1$ is better than $\hat{\varphi}_2$ in the sense of quadratic risk (when viewed as an estimator of $\varphi(\theta)$) if for all $\theta \in \Theta$, $R_\theta(\hat{\varphi}_1) \leq R_\theta(\hat{\varphi}_2)$. This order is not total on the estimators of $\varphi(\theta)$ because the risk depends on θ .

In the case where $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are both unbiased, comparing their risks boils down to comparing $\text{Tr}(\text{Var}_\theta(\hat{\varphi}_1))$ and $\text{Tr}(\text{Var}_\theta(\hat{\varphi}_2))$ for all $\theta \in \Theta$ (as in the risk). One can also compare the whole covariance matrix globally, according to the following order on symmetric matrices.

Definition 3.3 (Loewner order). We define the following order relation for any A, B , square

symmetric matrices of same size:

$$A \preceq B \iff B - A \text{ is positive semidefinite.}$$

It is clear that if $A \preceq B$ then $\text{Tr}(A) \leq \text{Tr}(B)$. But the \preceq is strictly stronger, and is not a total order: if $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, we neither have $A \preceq B$ nor $B \preceq A$.

Proposition 3.2. *A matrix version of the Cauchy–Schwarz inequality holds in the Löwner order. Let U, V be random vectors in \mathbb{R}^d with finite second moments. Then*

$$\text{Cov}(U, V) \preceq \text{Var}(U)^{1/2} \text{Var}(V)^{1/2},$$

where $A^{1/2}$ denotes the (symmetric) positive semi-definite square root of a positive semi-definite matrix A .

Proof. Consider the joint random vector (U, V) and its covariance matrix

$$\Sigma = \begin{pmatrix} \text{Var}(U) & \text{Cov}(U, V) \\ \text{Cov}(V, U) & \text{Var}(V) \end{pmatrix}.$$

Since Σ is a covariance matrix, it is positive semi-definite. Assume first that $\text{Var}(U)$ and $\text{Var}(V)$ are invertible. Define

$$K := \text{Var}(U)^{-1/2} \text{Cov}(U, V) \text{Var}(V)^{-1/2}.$$

Then

$$\begin{pmatrix} I & K \\ K^\top & I \end{pmatrix} = \begin{pmatrix} \text{Var}(U)^{-1/2} & 0 \\ 0 & \text{Var}(V)^{-1/2} \end{pmatrix} \Sigma \begin{pmatrix} \text{Var}(U)^{-1/2} & 0 \\ 0 & \text{Var}(V)^{-1/2} \end{pmatrix} \succeq 0.$$

By the Schur complement, this implies $K^\top K \preceq I$, hence $\|K\|_{\text{op}} \leq 1$. Therefore,

$$\text{Cov}(U, V) = \text{Var}(U)^{1/2} K \text{Var}(V)^{1/2} \preceq \text{Var}(U)^{1/2} \text{Var}(V)^{1/2}.$$

If $\text{Var}(U)$ or $\text{Var}(V)$ is singular, the result follows by a standard regularization argument. \square

Now that the main definitions have been introduced, we turn to the core question: how to construct good estimators? The following sections are devoted to some generic and classical methods for building estimators, namely the method of moments and maximum likelihood estimation.

3.1.2. Method of moments

The method of moments is a very simple approach to parameter estimation in i.i.d. models, based on matching theoretical and empirical moments.

Assume that we work in an i.i.d. model, that is we observe i.i.d. random vectors X_1, \dots, X_n . Assume that the quantity of interest $\varphi(\theta)$ has the form

$$\varphi(\theta) = h(\mathbb{E}_\theta[f_1(X_1)], \mathbb{E}_\theta[f_2(X_1)], \dots, \mathbb{E}_\theta[f_m(X_1)]),$$

where h and f_1, \dots, f_m are known measurable functions such that the above expectations are well defined. The idea is simply to approximate these expectations using the empirical means. Define, for $1 \leq j \leq m$,

$$\hat{f}_j := \frac{1}{n} \sum_{i=1}^n f_j(X_i).$$

The *method of moments estimator* (MME) is then defined by

$$\hat{\varphi}_{\text{MM}} := h(\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m).$$

This method provides a simple, interpretable way to estimate parameters, though it may not always be optimal in terms of efficiency or variance, as we will see throughout the exercises.

Example 3.2. Back again to our survey model (Example 2.1), since $\theta = \mathbb{E}_\theta[X_1]$, that the moment method suggests to use the sample mean estimator $\hat{\theta}_{\text{MM}} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. As claimed before, this is a natural estimator: the empirical proportion of successes is used to estimate the true probability of success.

Remark 3.1. Note that the moment method does not provide a unique estimator for a given φ , since several moments of several functions of X may contain information about φ . Say that, in the survey model (Example 2.1), we now want to estimate $\varphi(\theta) = \theta(1 - \theta)$. Here there are at least two natural estimators that can be obtained using the method of moments. First, we may remark that

$$\varphi(\theta) = \text{Var}_\theta(X) = \mathbb{E}_\theta(X^2) - \mathbb{E}_\theta(X)^2$$

and use the estimator

$$\hat{\varphi}_{\text{MM},1} = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2.$$

Another possibility is to remark that

$$\varphi(\theta) = \theta(1 - \theta) = \mathbb{E}_\theta(X)(1 - \mathbb{E}_\theta(X))$$

and use $\hat{\varphi}_{\text{MM},2} = \bar{X}_n(1 - \bar{X}_n)$.

3.1.3. Maximum likelihood estimation

In this section, we assume that $\varphi(\theta) = \theta$, that is, we are directly interested in estimation of θ . The *maximum likelihood method* is a systematic approach to parameter estimation in dominated models, that selects the parameter maximizing the likelihood of the observed data, defined hereafter.

Definition 3.4 (Likelihood function). Assume $\mathcal{M} \ll \nu$ where ν is σ -finite, and denote by p_θ the density of \mathbb{P}_θ w.r.t. ν , for $\theta \in \Theta$. The ν -likelihood function is the mapping $L_\nu : \Theta \times \mathcal{X} \rightarrow \mathbb{R}_+$ such that

$$L_\nu(\theta, x) = p_\theta(x).$$

We write $L_\nu = L$ when the dominating measure ν is clear from context.

The numerical value of the likelihood coincides with that of the density. From a conceptual standpoint, however, the two functions differ: p_θ is viewed as a function of x , whereas L is regarded as a function of (θ, x) . Strictly speaking, the likelihood function is not unique, since it may depend on the choice of the dominating measure. When the choice of ν is clear from context, we will speak about *the* likelihood function.

Remark 3.2. In an i.i.d. model where $\mathbb{P}_\theta = P_\theta^{\otimes n}$, dominated by $\nu^{\otimes n}$, then the $\nu^{\otimes n}$ -likelihood satisfies \mathcal{M} -a.s.

$$L_{\nu^{\otimes n}}(\theta, x) = \prod_{i=1}^n L_\nu(\theta, x_i).$$

Definition 3.5 (Maximum likelihood estimator(s)). Let Θ be a measure space endowed with a σ -algebra \mathcal{G} . A *maximum likelihood estimator* (MLE) is any statistic $\hat{\theta}_{\text{MLE}} : (\mathcal{X}, \mathcal{F}) \rightarrow (\Theta, \mathcal{G})$

such that

$$L_\nu(\hat{\theta}_{\text{MLE}}(x), x) = \sup_{\theta \in \Theta} L_\nu(\theta, x) \quad \text{for } \nu - \text{almost every } x.$$

It is often convenient (especially in product models)¹ to work instead with the so called *log-likelihood* $\ell_\nu : \Theta \times \mathcal{X} \rightarrow [-\infty, +\infty[$ such that

$$\ell_\nu(\theta, x) := \begin{cases} \log L_\nu(\theta, x) & \text{if } 0 < L_\nu(\theta, x) < \infty, \\ -\infty & \text{if } L_\nu(\theta, x) = 0. \end{cases}$$

Remark 3.3. Why do we want measurability of $\hat{\theta}_{\text{MLE}}$? This is because we want to study probabilities that $\hat{\theta}_{\text{MLE}}$ is "close to" θ , which requires $\hat{\theta}_{\text{MLE}} : (\mathcal{X}, \mathcal{F}) \rightarrow (\Theta, \mathcal{G})$ to be measurable.

There is no reason for a MLE to exist, and when it does, it has no reason to be unique, nor to be a good estimator a priori.

But, under some additional assumptions, the following classical theorem justifies why maximizing the likelihood is a good idea and should work for our task.

Theorem 3.1. Fix $\theta_0 \in \Theta$. Note that by definition of the likelihood, $L_\nu(\theta_0, X) > 0$ \mathbb{P}_{θ_0} -a.s. Assume that for all $\theta \in \Theta$, $\log \frac{L_\nu(\theta, X)}{L_\nu(\theta_0, X)} \in L^1(\mathbb{P}_{\theta_0})$, such that the function

$$\Delta : \theta \mapsto \mathbb{E}_{\theta_0}[\ell_\nu(\theta, X) - \ell_\nu(\theta_0, X)] = \mathbb{E}_{\theta_0} \left[\log \frac{L_\nu(\theta, X)}{L_\nu(\theta_0, X)} \right]$$

is well-defined (finite on Θ). Then, Δ has a global maximum at $\theta = \theta_0$, and if the model is identifiable, $\theta = \theta_0$ is the unique global maximum.

Proof of Theorem 3.1. We have of course that $\Delta(\theta_0) = 0$. Since for any $x > 0$, $\log(x) = 2 \log \sqrt{x} \leq 2(\sqrt{x} - 1)$, it follows that

$$\Delta(\theta) \leq 2 \mathbb{E}_{\theta_0} \left[\sqrt{\frac{p_\theta(X)}{p_{\theta_0}(X)}} - 1 \right] = 2 \int \sqrt{p_\theta(x)p_{\theta_0}(x)} d\nu(x) - 2$$

and by Cauchy-Schwarz

$$\Delta(\theta) \leq 2 \int \sqrt{p_\theta(x)p_{\theta_0}(x)} d\nu(x) - 2 \leq 2 \left(\int p_\theta(x) d\nu(x) \right) \left(\int p_{\theta_0}(x) d\nu(x) \right) - 2 = 0,$$

with equality if and only if $p_\theta = p_{\theta_0}$ up to a null set of ν , which, if the model is identifiable, is the case only if $\theta = \theta_0$. \square

The heuristic behind the method of maximum likelihood estimation is simple. Suppose θ_0 is the true parameter. The maximizers of $\ell_\nu(\cdot, X)$ are also those of $\ell_\nu(\cdot, X) - \ell_\nu(\theta_0, X)$. There is a hope that this quantity is well approximated by its expectation, that is

$$\ell_\nu(\cdot, X) - \ell_\nu(\theta_0, X) \approx \mathbb{E}_{\theta_0}[\ell_\nu(\cdot, X) - \ell_\nu(\theta_0, X)]$$

as functions of θ , and then that their maximizers will also be close. Namely, by Theorem 3.1, if the planets align well, $\hat{\theta}$ is close to θ_0 .

Example 3.3 (Gaussian i.i.d. model with unknown mean and variance). Consider the Gaussian i.i.d. model (Example 2.9) where X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ variables, where we

¹this is because it is easier to differentiate a sum than a product.

estimate both μ and σ^2 . Take classically $\nu = (\text{Leb}_{\mathbb{R}})^{\otimes n}$ as the dominating measure. The log-likelihood is given by

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

We can optimize first in μ (ℓ is concave in μ) and then solve for σ , since ℓ is convex in $\alpha = 1/\sigma^2$. This is because²

$$\max_{\mu, \sigma} \ell(\mu, \sigma^2; X) = \max_{\sigma} (\max_{\mu} \ell(\mu, \sigma^2; X)) = \min_{\alpha} (\max_{\mu} \ell(\mu, \alpha; X)).$$

We have

$$\frac{\partial \ell}{\partial \nu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu), \quad \text{and} \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2,$$

Solving the first order conditions we obtain

$$\hat{\mu}_{\text{MLE}} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \text{and} \quad \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Note that in this case, for the estimation of μ , the method of moments and maximum likelihood estimation give the same result, namely the sample mean. Note that $\hat{\sigma}_{\text{MLE}}^2$ is slightly biased, it is usually called the *biased estimator of the variance*, as it underestimates the true variance. Indeed,

$$\begin{aligned} \hat{\sigma}_{\text{MLE}}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n} \sum_{i=1}^n [(X_i - \mu) - (\bar{X}_n - \mu)]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X}_n - \mu)^2, \end{aligned}$$

and taking expectations yields

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_{\text{MLE}}^2] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] - \mathbb{E}[(\bar{X}_n - \mu)^2] \\ &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2. \end{aligned}$$

The usual *unbiased* sample variance is obtained by replacing n with $n-1$ in the denominator:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

This example shows that maximum likelihood estimators are not always unbiased.

3.1.4. Asymptotic properties of estimators

We conclude this section with defining some asymptotic properties of estimators, namely consistency and asymptotic normality, which describe the behavior of estimators as the sam-

²we can always minimize a function by first minimizing over some of the variables – fixing the others – and then minimizing over the remaining ones. Minima (only) or maxima (only) are commutative.

ple size n increases. These properties are fundamental to understanding the efficiency and reliability of estimators in large samples.

Definition 3.6. An estimator $\hat{\varphi}_n = \hat{\varphi}_n(X_1, \dots, X_n)$ of $\varphi(\theta)$ is said to be *weakly consistent* if for all θ , under \mathbb{P}_θ ,

$$\hat{\varphi}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \varphi(\theta),$$

and $\hat{\varphi}_n$ is said to be *strongly consistent* for a parameter θ^* if for all θ , under \mathbb{P}_θ ,

$$\hat{\varphi}_n \xrightarrow[n \rightarrow \infty]{a.s.} \varphi(\theta).$$

Of course, strong consistency is a stronger condition than weak consistency. Most of the time, strong consistency is established via the strong law of large numbers, or by Borel–Cantelli’s Lemma.

Definition 3.7. Let $\hat{\varphi}_n = \hat{\varphi}_n(X_1, \dots, X_n)$ be an estimator of $\varphi(\theta)$ with values in \mathbb{R}^ℓ . $\hat{\varphi}$ is said to be *asymptotically normal* if for all $\theta \in \Theta$, there exists a positive semidefinite matrix $\Sigma = \Sigma(\theta) \in \mathbb{R}^{\ell \times \ell}$ such that under \mathbb{P}_θ ,

$$\sqrt{n}(\hat{\varphi}_n - \varphi(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Sigma(\theta)).$$

We emphasize that the asymptotic covariance matrix $\Sigma(\theta)$ depends on θ . Most of the time, asymptotic normality follows from the central limit theorem (potentially together with a Delta method), as illustrated hereafter.

Example 3.4. In the survey model (Example 2.1), the sample mean estimator $\hat{\theta}_n = \bar{X}_n$ is strongly consistent by the law of large numbers, and asymptotically normal via the CLT, since for all $\theta \in [0, 1]$, under \mathbb{P}_θ ,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \theta(1 - \theta)).$$

3.2. No one beats mother Nature: the Cramér–Rao lower bound

3.2.1. Fisher Information in regular models

We start by introducing Fisher information is an information-theoretic quantity which satisfies nice properties when the model is regular enough with respect to the parameter $\theta \in \Theta$. We will then see that Fisher information plays a crucial role in determining lower bounds on the variance of smooth estimators.

Definition 3.8 (Regular parametric model). A parametric model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ is called *regular* if:

- (i) Θ is an open set in some \mathbb{R}^d ,
- (ii) $\mathcal{M} \ll \nu$ where ν is σ -finite, and the ν -likelihood $L = L_\nu$ is positive everywhere on $\Theta \times \mathcal{X}$.
- (iii) The log-likelihood $\ell = \log L$ (well defined by (ii)) has a finite, continuous gradient in θ , everywhere on $\Theta \times \mathcal{X}$. The gradient in θ of ℓ , is called the *score function*, given by:

$$\dot{\ell}(\theta, x) := \nabla_\theta \ell(\theta, x) = \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1}(\theta, x) \\ \vdots \\ \frac{\partial \ell}{\partial \theta_d}(\theta, x) \end{bmatrix} \in \mathbb{R}^d. \quad (3.1)$$

- (iv) The derivatives $x \mapsto \frac{\partial}{\partial \theta_i} L(\theta, x)$ are locally dominated uniformly in (the neighborhood of) θ by functions in $L^1(\nu)$.
- (v) The function $\theta \mapsto \mathbb{E}_\theta(\|\dot{\ell}(\theta, X)\|^2)$ is well defined (finite) and continuous on Θ .

Remark 3.4 (Interpretation of the assumptions). Assumption (i) is suited for taking derivatives, (ii) enables to consider the log-likelihood everywhere and imposes that the support of \mathbb{P}_θ does not depend on θ . Note that this prevents models like $(\text{Unif}([0, \theta]))_{\theta > 0}$ to be regular. Assumption (iii) ensures that the score function (3.1) is well-defined and continuous in θ . Condition (iv) is needed to apply dominated convergence theorem, and implies for instance that the score $\ell(\theta, X)$ is of null expectation under every \mathbb{P}_θ . Now, (v) is key to define Fisher information hereafter, which will be continuous.

Remark 3.5. Every model studied in these lecture notes up to now is regular, except the uniform i.i.d. model (Example 2.7) $(\text{Unif}([0, \theta]))_{\theta > 0}$, for which point (ii) fails. Regularity will often be admitted, or proven by looking at the densities/likelihood and prove that they are regular enough to meet conditions (iii), (iv) and (v) in Definition 3.8.

Remark 3.6. Point (iv) in the regularity assumptions implies that in a regular model, the score is a centered random vector. Indeed, for all $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_\theta[\dot{\ell}_i(\theta, X)] &= \int \frac{\frac{\partial}{\partial \theta_i} L(\theta, x)}{L(\theta, x)} L(\theta, x) d\nu \\ &= \int \frac{\partial}{\partial \theta_i} L(\theta, x) d\nu \stackrel{(iv)}{=} \underbrace{\frac{\partial}{\partial \theta_i} \int L(\theta, x) d\nu}_{=1} = 0. \end{aligned}$$

We are now ready to introduce Fisher information, readily the covariance matrix of the score.

Definition 3.9 (Fisher information matrix). In a regular model, for all $\theta \in \Theta$, the *Fisher information matrix* is defined as:

$$I(\theta) = \text{Var}_\theta(\dot{\ell}(\theta, X)). \quad (3.2)$$

Remark 3.7 (Alternative expression of Fisher information). In a regular model, if moreover the likelihood is twice differentiable,

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta, x) = \frac{\partial}{\partial \theta_i} \frac{\frac{\partial}{\partial \theta_j} L(\theta, x)}{L(\theta, x)} = \frac{\frac{\partial^2}{\partial \theta_i \partial \theta_j} L(\theta, x)}{L(\theta, x)} - \frac{\frac{\partial}{\partial \theta_j} L(\theta, x) \frac{\partial}{\partial \theta_i} L(\theta, x)}{L(\theta, x)^2},$$

and taking the expectations gives

$$\begin{aligned} \mathbb{E}_\theta[\ddot{\ell}(\theta, X)]_{i,j} &= \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} L(\theta, x) \nu(dx) - \int \frac{\frac{\partial}{\partial \theta_j} L(\theta, x) \frac{\partial}{\partial \theta_i} L(\theta, x)}{L(\theta, x)^2} L(\theta, x) \nu(dx) \\ &\stackrel{(\star\star)}{=} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int L(\theta, x) \nu(dx) - \mathbb{E}_\theta[\dot{\ell}_i(\theta, X) \dot{\ell}_j(\theta, X)] \\ &= 0 - I(\theta)_{i,j}. \end{aligned}$$

If the second derivatives of the likelihood are locally dominated by integrable functions, then step $(\star\star)$ is legitimate, and we established the alternative expression:

$$I(\theta) = \text{Var}_\theta(\dot{\ell}(\theta, X)) = -\mathbb{E}_\theta[\ddot{\ell}(\theta, X)], \quad (3.3)$$

where $\ddot{\ell}$ is the Hessian matrix of ℓ .

3.2.2. Properties of Fisher information

We give a few intuitive properties of Fisher information. The first one is its additivity in product models.

Proposition 3.3 (Additivity of Fisher information). *Let $\mathcal{M}_1 = (\mathcal{X}_1, \mathcal{F}_1, (\mathbb{P}_\theta)_{\theta \in \Theta})$ and $\mathcal{M}_2 = (\mathcal{X}_2, \mathcal{F}_2, (\mathbb{Q}_\theta)_{\theta \in \Theta})$ be regular parametric models with Fisher information matrices $I_1(\theta)$ and $I_2(\theta)$ respectively. Consider the product model*

$$\mathcal{M}_1 \otimes \mathcal{M}_2 := (\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{F}_1 \otimes \mathcal{F}_2, (\mathbb{P}_\theta \otimes \mathbb{Q}_\theta)_{\theta \in \Theta}).$$

Then $\mathcal{M}_1 \otimes \mathcal{M}_2$ is regular and its Fisher information matrix $I(\theta)$ is given for all $\theta \in \Theta$ by

$$I(\theta) = I_1(\theta) + I_2(\theta).$$

Proof. The product model is dominated by a product of σ -finite measures, which is also σ -finite. The regularity is trivial, and the additivity of Fisher information follows from the fact that the log-likelihood for $x = (x_1, x_2)$ writes $\ell(\theta, x) = \ell_1(\theta, x_1) + \ell_2(\theta, x_2)$, where ℓ_1, ℓ_2 are the log-likelihoods in models \mathcal{M}_1 and \mathcal{M}_2 . Therefore, $\dot{\ell}(\theta, X) = \dot{\ell}_1(\theta, X_1) + \dot{\ell}_2(\theta, X_2)$, and $(X_1, X_2) \sim \mathbb{P}_\theta \otimes \mathbb{Q}_\theta$, so the terms in the sum are independent and their covariance matrix add up: if $\Theta \in \mathbb{R}^\ell$, for all $1 \leq i, j \leq \ell$,

$$\text{Cov}_\theta(\dot{\ell}_i(\theta, X), \dot{\ell}_j(\theta, X)) = \text{Cov}_\theta(\dot{\ell}_{1,i}(\theta, X_1), \dot{\ell}_{1,j}(\theta, X_1)) + \text{Cov}_\theta(\dot{\ell}_{2,i}(\theta, X_2), \dot{\ell}_{2,j}(\theta, X_2)).$$

□

Proposition 3.3 is particularly useful in i.i.d. models: if $\mathbb{P}_\theta = (P_\theta)^{\otimes n}$, where the total Fisher information $I(\theta)$ satisfies

$$I(\theta) = n \times I_1(\theta).$$

Example 3.5 (Fisher information in the Gaussian i.i.d. model). In the Gaussian i.i.d. model (Example 2.9) (which is regular), the log-likelihood of X_1 writes

$$\ell(\mu, \sigma^2, x) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2,$$

and the score is

$$\dot{\ell}(\mu, \sigma^2; x) = \nabla_{\mu, \sigma^2} \ell(\mu, \sigma^2, x) = \begin{pmatrix} \frac{1}{\sigma^2}(x - \mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x - \mu)^2 \end{pmatrix}.$$

The Fisher information matrix is here given by

$$I_1(\mu, \sigma^2) = \text{Var}_{\mu, \sigma^2}(\dot{\ell}(\mu, \sigma^2; X)) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

using $\text{Var}_{\mu, \sigma^2}((X - \mu)^2) = \mathbb{E}_{\mu, \sigma^2}[(X - \mu)^4] - \mathbb{E}_{\mu, \sigma^2}[(X - \mu)^2]^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$. The fact that $I_1(\mu, \sigma^2)$ is diagonal intuitively translates the fact that the quantity of information that X reveals on μ and σ^2 are uncorrelated. This is in line with the independence of the MLE for μ and σ^2 established before (see Example 2.9). Note that in view of Proposition 3.3, the total Fisher information satisfies $I = nI_1$.

We now state a simple yet fundamental result: any deterministic transformation $T : (\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ of the data can only deteriorate Fisher information. Moreover, Fisher information is preserved if and only if $T(X)$ is a sufficient statistic, thus directly linking sufficiency and Fisher information.

Theorem 3.2. Let $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a regular model dominated by a σ -finite measure ν . Denote by ℓ the log-likelihood and by $I(\theta)$ the Fisher information matrix. Let $T : (\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ be a statistic. Define \mathcal{M}_T the model induced by T as follows:

$$\mathcal{M}_T = (\mathcal{Y}, \mathcal{G}, (\mathbb{P}_\theta \circ T^{-1})_{\theta \in \Theta}).$$

This is the model where $T(X(\omega))$ is observed instead of $X(\omega)$. Assume that \mathcal{M}_T is regular and denote by ℓ_T its log-likelihood with respect to $\nu_T := \nu \circ T^{-1}$ and $I_T(\theta)$ its Fisher information matrix. Then the following hold:

- (i) For all $\theta \in \Theta$, $\mathbb{E}[\dot{\ell}(\theta, X) | T(X) = t] = \dot{\ell}_T(\theta, t)$.
- (ii) For all $\theta \in \Theta$, $I_T(\theta) \preceq I(\theta)$.
- (iii) $(I_T(\theta) = I(\theta) \text{ for all } \theta \in \Theta) \iff T(X) \text{ is sufficient in model } \mathcal{M}$.

Proof of Theorem 3.2. In this proof, we denote $L := \exp(\ell)$ and $L_T := \exp(\ell_T)$ the likelihoods. **Proof of (i).** We have to prove that the given conditional expectation exists and is equal to $\dot{\ell}_T(\theta, t)$. Let $A \in \mathcal{G}$. By definition,

$$\int_{T^{-1}(A)} L(\theta, x) \nu(dx) = \int_A L_T(\theta, t) \nu_T(dt).$$

Using the regularity assumptions, differentiating both sides with respect to θ , we get

$$\int_{T^{-1}(A)} \nabla_\theta L(\theta, x) \nu(dx) = \int_A \nabla_\theta L_T(\theta, t) \nu_T(dt),$$

and therefore, multiplying and dividing by L (left) and L_T (right):

$$\int_{T^{-1}(A)} \dot{\ell}(\theta, x) L(\theta, x) \nu(dx) = \int_A \dot{\ell}_T(\theta, t) L_T(\theta, t) \nu_T(dt).$$

Rewriting this as expectations, we have that for all $A \in \mathcal{Y}$,

$$\mathbb{E}_\theta[\dot{\ell}(\theta, X) \mathbb{1}_A(T(X))] = \mathbb{E}_\theta[\dot{\ell}_T(\theta, T(X)) \mathbb{1}_A(T(X))].$$

By the definition of conditional expectation, this readily implies $\mathbb{E}[\dot{\ell}(\theta, X) | T(X) = t] = \dot{\ell}_T(\theta, t)$.

Proof of (ii) and (iii). By point (i) and the law of total expectation:

$$\begin{aligned} \mathbb{E}[\dot{\ell}_T(\theta, T(X)) \dot{\ell}(\theta, X)^\top] &= \mathbb{E}[\dot{\ell}_T(\theta, T(X)) \mathbb{E}[\dot{\ell}(\theta, X) | T(X)]^\top] \\ &= \mathbb{E}[\dot{\ell}_T(\theta, T(X)) \dot{\ell}_T(\theta, T(X))^\top] \\ &= I_T(\theta), \end{aligned}$$

and thus

$$\begin{aligned} \text{Var}_\theta(\dot{\ell}(\theta, X) - \dot{\ell}_T(\theta, T(X))) &= \mathbb{E}[(\dot{\ell}(\theta, X) - \dot{\ell}_T(\theta, T(X)))(\dot{\ell}(\theta, X) - \dot{\ell}_T(\theta, T(X)))^\top] \\ &= I(\theta) + I_T(\theta) - 2\mathbb{E}[\dot{\ell}_T(\theta, T(X)) \dot{\ell}(\theta, X)^\top] \\ &= I(\theta) + I_T(\theta) - 2I_T(\theta) \\ &= I(\theta) - I_T(\theta), \end{aligned}$$

and the left hand term is positive semidefinite, which proves $I_T(\theta) \preceq I(\theta)$. Moreover,

$$(\forall \theta \in \Theta, I_T(\theta) = I(\theta)) \iff (\forall \theta \in \Theta, \dot{\ell}(\theta, X) = \dot{\ell}_T(\theta, T(X)), \quad \mathbb{P}_\theta - \text{a.s.})$$

Integrating over θ , this last condition is equivalent to the fact that the log-likelihood is of the form

$$\ell(\theta, x) = \ell_T(\theta, T(x)) + k(x), \quad \mathcal{M} - \text{a.s.}$$

Taking the exponential and applying Neyman–Fisher Theorem (Theorem 2.2) concludes the proof that $T(X)$ is sufficient. \square

3.2.3. The Cramér–Rao Theorem

We will now discuss the main result of this Section, the Cramér–Rao Theorem. This result states that in a regular model, any smooth unbiased estimator of $\varphi(\theta)$ has a variance which is always lower bounded by some quantity which depends on the smoothness of φ and on the Fisher information $I(\theta)$. When this lower bound is positive, this gives the first informational lower bound in statistics: no unbiased estimator can beat this bound. Recall that $\Theta \subseteq \mathbb{R}^d$ and $\varphi : \Theta \rightarrow \mathbb{R}^\ell$.

Remark 3.8. When the model is regular, the score function $\dot{\ell}$ satisfies the property that the operator $T(X) \mapsto \mathbb{E}_\theta[T(X)\dot{\ell}(\theta, X)^\top]$ acts as a differential operator as soon as the statistic $T(X)$ is smooth enough. Indeed, for a given statistic $T(X)$ taking values in \mathbb{R}^ℓ , we have, for all $\theta \in \Theta, 1 \leq i \leq d, 1 \leq j \leq \ell$,

$$\begin{aligned} \mathbb{E}_\theta[T_j(X)\dot{\ell}_i(\theta, X)] &= \int T_j(x) \left[\frac{\partial}{\partial \theta_i} \log L(\theta, x) \right] L(\theta, x) d\nu = \int T_j(x) \frac{\frac{\partial}{\partial \theta_i} L(\theta, x)}{L(\theta, x)} L(\theta, x) d\nu \\ &= \int T_j(x) \frac{\partial}{\partial \theta_i} L(\theta, x) d\nu \stackrel{(*)}{=} \frac{\partial}{\partial \theta_i} \int T_j(x) L(\theta, x) d\nu = \frac{\partial}{\partial \theta_i} \mathbb{E}_\theta[T_j(X)]. \end{aligned}$$

The estimators $T(X)$ for which the computation step $(*)$ is legitimate³ will be called *smooth*. We define them below. In the following, we suppose $\Theta \subseteq \mathbb{R}^d$.

Definition 3.10. In a regular model, $T(X)$ is a *smooth estimator* if $T(X)$ is integrable, and for all $\theta \in \Theta$, $T(X)\dot{\ell}(\theta, X)^\top$ is integrable, with

$$\mathbb{E}_\theta[T(X)\dot{\ell}(\theta, X)^\top] = J_\varphi(\theta), \quad (3.4)$$

where $J_\varphi(\theta)$ is the $\ell \times d$ Jacobian matrix of $\varphi : \theta \mapsto \mathbb{E}_\theta[T(X)]$.

We can show that for a regular model, T is smooth as soon as $\theta \mapsto \mathbb{E}_\theta[\|T(X)\|^2]$ is locally bounded (that is, bounded on a neighborhood of any $\theta \in \Theta$), but the proof is very cumbersome, and we refer to Theorem 11.2.5 of [7]. Since any continuous function is locally bounded, continuity of $\theta \mapsto \mathbb{E}_\theta[\|T(X)\|^2]$ (a pretty weak property to impose to an estimator) ensures that $T(X)$ is smooth.

We are now ready to state the Cramér–Rao Theorem.

Theorem 3.3 (Cramér–Rao Theorem). *Let $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a regular model. Suppose that $\varphi : \Theta \rightarrow \mathbb{R}^\ell$ is differentiable, denote by J_φ its Jacobian. Let $T : \mathcal{X} \rightarrow \mathbb{R}^\ell$ be an unbiased, smooth estimator of $\varphi(\theta)$. If $I(\theta)$ is non-singular (i.e. invertible), then:*

$$\text{Var}_\theta(T(X)) \succeq J_\varphi(\theta) I(\theta)^{-1} J_\varphi(\theta)^\top.$$

When $\ell = d = 1$, this writes

$$\text{Var}_\theta(T(X)) \geq \frac{\varphi'(\theta)^2}{I(\theta)}.$$

Proof of Theorem 3.3. We use the shorthand notation T for $T(X)$. The only main ingredient in this proof is the Cauchy–Schwarz inequality. We first consider the case $\ell = d = 1$. Then

³Note that the question on whether step $(*)$ is legitimate indeed depends on $T(X)$, because we need to dominate locally in θ , the product $T_j(x) \frac{\partial}{\partial \theta_i} L(\theta, x)$ by some function $T_j(x)g(x)$ which needs to be ν -integrable.

under the assumptions of the theorem, $\text{Cov}_\theta(T, \dot{\ell}(\theta, X)) = \mathbb{E}_\theta[T \dot{\ell}(\theta, X)] - \mathbb{E}_\theta[T] \mathbb{E}_\theta[\dot{\ell}(\theta, X)] = \mathbb{E}_\theta[T \dot{\ell}(\theta, X)] = \varphi'(\theta)$, and by the Cauchy-Schwarz inequality:

$$\text{Var}_\theta(T) \geq \frac{\text{Cov}_\theta(T, \dot{\ell}(\theta, X))^2}{\text{Var}_\theta(\dot{\ell}(\theta, X))} = \frac{\varphi'(\theta)^2}{I(\theta)},$$

since by assumption $I(\theta) > 0$.

To extend this to arbitrary d and ℓ , note that for any $x \in \mathbb{R}^\ell$ and $y \in \mathbb{R}^d$:

$$\mathbb{E}_\theta[\dot{\ell}(\theta, X)^\top y] = 0 \quad \text{and} \quad \text{Cov}_\theta(x^\top T, \dot{\ell}(\theta, X)^\top y) = x^\top J_\varphi(\theta) y.$$

Then, for $y \neq 0$, Cauchy-Schwarz gives, using $\text{Var}_\theta(\dot{\ell}(\theta, X)^\top y) = y^\top I(\theta) y$:

$$\text{Var}_\theta(x^\top T) \geq \frac{(x^\top J_\varphi(\theta) y)^2}{y^\top I(\theta) y},$$

which is true for any $y \neq 0$, we can thus take the supremum in the right hand side. Making the change variables $y = [I(\theta)]^{-1/2} u$, this gives

$$\text{Var}_\theta(x^\top T) \geq \sup_{u \neq 0} \frac{(x^\top J_\varphi(\theta) [I(\theta)]^{-1/2} u)^2}{\|u\|^2} = \sup_{\|u\|=1} (x^\top J_\varphi(\theta) [I(\theta)]^{-1/2} u)^2$$

The supremum is attained at $u = \frac{[I(\theta)]^{-1/2} J_\varphi(\theta)^\top x}{\|[I(\theta)]^{-1/2} J_\varphi(\theta)^\top x\|}$ and the inequality becomes

$$x^\top \text{Var}_\theta(T) x \geq \|[I(\theta)]^{-1/2} J_\varphi(\theta)^\top x\|^2 = x^\top J_\varphi(\theta) I(\theta)^{-1} J_\varphi(\theta)^\top x.$$

This proves that

$$\text{Var}_\theta(T) \succeq J_\varphi(\theta) I(\theta)^{-1} J_\varphi(\theta)^\top.$$

□

Example 3.6 (Exponential model). In the model where $X_1, \dots, X_n \sim \text{i.i.d. Exp}(\lambda)$, the log-likelihood writes

$$\ell(\lambda, x) = n \log \lambda - \lambda \sum_{i=1}^n x_i,$$

and

$$\dot{\ell}(\lambda, x) = n/\lambda - \sum_{i=1}^n x_i$$

which has variance $n \times \frac{1}{\lambda^2}$. The Fisher information is thus $I(\lambda) = n/\lambda^2$.

Assume we want to estimate $\varphi(\lambda) = 1/\lambda$. The Cramér–Rao lower bound for estimating λ is $\frac{\varphi'(\lambda)^2}{I(\lambda)} = \left(\frac{-1}{\lambda^2}\right)^2 \times \frac{\lambda^2}{n} = \frac{1}{\lambda^2 n}$. The sample mean estimator $\hat{\lambda}_n = \bar{X}_n$ is unbiased, and achieves exactly this variance. Such estimators are called efficient.

Definition 3.11 (Efficient estimators). An unbiased estimator $\hat{\varphi}$ is said to be *efficient* if it achieves the Cramér–Rao bound, i.e. if

$$\forall \theta \in \Theta, \text{Var}_\theta(\hat{\varphi}) = J_\varphi(\theta) I(\theta)^{-1} J_\varphi(\theta)^\top.$$

An estimator $\hat{\varphi} = \hat{\varphi}_n$ is said to be *asymptotically efficient* if its bias tends to 0 with n and if

$$\forall \theta \in \Theta, \text{Var}_\theta(\hat{\varphi}_n) \underset{n \rightarrow \infty}{\sim} J_\varphi(\theta) I(\theta)^{-1} J_\varphi(\theta)^\top.$$

3.2.4. Limitations of the Cramér–Rao lower bound

Fisher’s information is only defined for regular models, ruling out many interesting models. For example, the uniform i.i.d. model (Example 2.7) is not regular and lacks Fisher information. Even seemingly nice models can fail regularity, e.g., the Laplace i.i.d. model where $\mathbb{P}_\theta = P_\theta^{\otimes n}$ where P_θ has density w.r.t. $\text{Leb}_\mathbb{R}$ given by

$$x \mapsto \frac{1}{2}e^{-|\theta-x|},$$

is not regular: the likelihood fails to be differentiable at $x = \theta$. Moreover, the Cramér–Rao bound (Theorem 3.3) also requires regularity assumptions on the estimator T , which prevents its whole generality (how knows if any strange non-smooth estimator could beat the Cramér–Rao lower bound?).

Finally, the CRLB only applies to unbiased estimators. For biased estimators with bias $b_\theta(T) \neq 0$ a similar bound can be obtained. In dimension 1, this writes:

$$\mathbb{E}_\theta[(T - \varphi(\theta))^2] \geq b_\theta(T)^2 + \frac{(b'_\theta(T) + \varphi'(\theta))^2}{I(\theta)},$$

although the right-hand side depends on T and is therefore not universal.

Finally, the Cramér–Rao Theorem (Theorem 3.3), despite providing a useful benchmark, does not indicate whether this bound is attainable, nor does it explain how efficient estimators can be constructed. Adopting a more constructive perspective, the following section provides guidance in this direction.

3.3. Sufficiency and Rao-Blackwell theorem

Rao-Blackwell’s Theorem enables us to build a better estimator out of an existing one. The solution do to so is as follows: taking the conditional expectation of an estimator with respect to a sufficient statistic. This will always produce a “better” estimator, as shown below.

Proposition 3.4 (Rao-Blackwellized estimator). *We work on model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$. Let $S(X)$ be a sufficient statistic and $T(X)$ be any other statistic, \mathbb{P}_θ –integrable for all $\theta \in \Theta$. Then there exists a measurable function $k_T : S(\mathcal{X}) \rightarrow T(\mathcal{X})$ such that*

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta(T(X) \mid S(X)) = k_T(S(X)) \quad \mathbb{P}_\theta - a.s.$$

The estimator

$$T^*(X) := k_T(S(X))$$

defines an estimator which on X only through the sufficient statistic $S(X)$. This estimator $T^(X)$ is called the Rao-Blackwellized (RB) estimator of $T(X)$ by $S(X)$.*

Proof. We use the shorthand notations T for $T(X)$, S for $S(X)$. This proposition is the consequence of standard properties of conditional expectations, with the additional property that k_T does not depend on θ since S is sufficient. For the sake of completeness, we give the proof of this statement.

Under any \mathbb{P}_θ , since S is sufficient, the conditional distribution of $T(X)$ given $S(X) = s$ does not depend on θ : we denote it by Q_s . Let $\theta \in \Theta$ be arbitrary, and first assume that $T(x) = \sum_{i=1}^k t_i \mathbf{1}_{B_i}(x)$ with $B_i \in \mathcal{F}$, that it T is a simple function. Then by definition,

$$\mathbb{E}_\theta(T \mid S) = \sum_{i=1}^k t_i Q_S(B_i), \quad \mathbb{P}_\theta - a.s.$$

Thus, the proposition is proved by letting:

$$k_T(s) = \sum_{i=1}^k t_i Q_s(B_i).$$

Now, if T is not a simple function, its measurability guarantees that it is the pointwise limit of a monotonically increasing sequence of simple functions $(T_n)_{n \geq 0}$. Since $k_{T_n}(s) = \int T_n dQ_s$ is increasing in T_n , $k_{T_n}(\cdot)$ has a pointwise limit, which we denote $k_T(\cdot)$. The result is proved since by monotone convergence for conditional expectations: we have that $\mathbb{E}_\theta(T_n | S) \rightarrow \mathbb{E}_\theta(T | S)$, \mathbb{P}_θ -a.s. \square

The nice thing about Rao-Blackwellization is that if T is an estimator of some $\varphi(\theta)$, it turns out that T^* also estimates $\varphi(\theta)$ with a risk function no greater than that of T . We can be more precise with the following theorem.

Theorem 3.4. *If $T(X)$ has finite second moment under integrable under any \mathbb{P}_θ , the RB estimator $T^*(X)$ of $T(X)$ by $S(X)$ satisfies the following:*

- (i) For all $\theta \in \Theta$, $\mathbb{E}_\theta[T^*] = \mathbb{E}_\theta[T]$. (same bias)
- (ii) For all $\theta \in \Theta$, $\text{Var}_\theta(T^*(X)) \preceq \text{Var}_\theta(T(X))$. (improved variance)

Consequently, the quadratic risk functions of T^* and T satisfy:

$$\forall \theta \in \Theta, \quad R_\theta(T^*) \leq R_\theta(T).$$

Proof. (i) is simply obtained by the law of total expectation. For (ii), we use the law of total variance, for any $\theta \in \Theta$:

$$\begin{aligned} \text{Var}_\theta(T) &= \text{Var}_\theta(\mathbb{E}_\theta[T | S]) + \mathbb{E}_\theta(\text{Var}_\theta(T | S)) \\ &= \text{Var}_\theta(T^*) + \underbrace{\mathbb{E}_\theta(\text{Var}_\theta(T | S))}_{\succeq 0} \succeq \text{Var}_\theta(T^*). \end{aligned}$$

\square

Example 3.7. Consider the i.i.d. Uniform model on $[0, \theta]$ where $\theta > 0$ is unknown (Example 2.7). To illustrate the efficiency of Rao-Blackwellization, consider the naive estimator $T = 2X_1$. It is unbiased but very bad (why?). As seen before, a sufficient statistic for θ is $S = \max(X_1, \dots, X_n)$. Let us compute the conditional distribution of T given $S = s$. Intuitively, $X_1 = s$ with probability $1/n$ (this is when the maximum is attained by X_1), and X_1 is uniform on $[0, s]$ with probability $1 - 1/n$. If so, then

$$T^* = \mathbb{E}[T | S] = 2 \left(\frac{1}{n} S + \left(1 - \frac{1}{n} \right) \frac{S}{2} \right) = \frac{n+1}{n} S,$$

which is unbiased by Theorem 3.4 and can be proven to have variance $O(n^{-2})$ when $n \rightarrow \infty$.

3.4. Uniformly minimum-variance unbiased estimators

The Cramér–Rao bound (Theorem 3.3) is not necessarily attainable and, even when it is, its validity requires restricting attention to smooth estimators. A natural question in estimation theory, independent of these considerations, is therefore to identify estimators whose quadratic risk functions are uniformly smaller than those of any other competing estimator. In this approach, one compares an estimator against all other estimators, rather than against a lower bound that may not be optimal or even achievable.

Unfortunately, this is impossible without imposing a constraint on the estimator, because it is easily seen that for any parameter $\theta \in \Theta$,

$$\inf_T \mathbb{E}_\theta [\|T - \varphi(\theta)\|^2] = 0,$$

where the infimum is taken over all estimators. The previous equality holds because for all $\theta \in \Theta$, the constant estimator $T = \varphi(\theta)$ has quadratic risk zero.

However, this estimator performs very bad in general. One way (it is not the only one) to overcome the previous issue is to add a global constraint on the set of estimators over which we take the infimum. A natural constraint, already looked at in this chapter, is to restrict to the class of unbiased estimators. Then, the problem reduces to *finding an unbiased estimator with uniformly smaller variance* (in the matrix sense).

We can now define our optimization problem rigorously:

Definition 3.12 (Uniform Minimum Variance Unbiased Estimator (UMVUE)). Recall that $\varphi : \Theta \rightarrow \mathbb{R}^\ell$. Let

$$\mathcal{U}_\varphi := \left\{ T : \mathcal{X} \rightarrow \mathbb{R}^\ell, T \text{ measurable}, \forall \theta \in \Theta, \mathbb{E}_\theta(T(X)) = \varphi(\theta) \text{ and } \mathbb{E}_\theta(\|T(X)\|^2) < +\infty \right\}$$

denote the class of all unbiased estimators of $\varphi(\theta)$ with finite variance. Then, $T \in \mathcal{U}_\varphi$ is said to be a *Uniform Minimum Variance Unbiased Estimator (UMVUE)* for $\varphi(\theta)$ if and only if

$$\forall T' \in \mathcal{U}_\varphi, \forall \theta \in \Theta, \quad \text{Var}_\theta(T) \preceq \text{Var}_\theta(T').$$

It follows from Rao-Blackwell's Theorem (Theorem 3.4) that in order to find a UMVUE, it is sufficient to search within the subclass of \mathcal{U}_φ consisting of estimators depending only on a sufficient statistic S for the model \mathcal{M} .

3.4.1. Lehmann-Scheffé theorem

Rao-Blackwellized estimators have improved variance, but are they UMVUE? It may be that $T^* = \mathbb{E}[T|S]$ is of reduced variance, but not of optimal variance. The answer to this question is therefore negative without further assumptions. However, if the sufficient statistic S is also complete, then the Rao-Blackwellized estimator is UMVUE.

Theorem 3.5 (Lehmann-Scheffé theorem). *Suppose that $S(X)$ is both complete and sufficient and let $T \in \mathcal{U}_\varphi$. Then, the RB estimator $T^*(X)$ of $T(X)$ by $S(X)$ is UMVUE for $\varphi(\theta)$. Furthermore, if V is another UMVUE that is a function of S , then \mathcal{M} -a.s., $V = T^*$.*

Proof. Let $U \in \mathcal{U}_\varphi$ be an arbitrary unbiased estimator. Since S is sufficient, the Rao-Blackwellized estimator $U^* = \mathbb{E}[U|S]$ is also in \mathcal{U}_φ . Furthermore, by Rao-Blackwell's theorem:

$$\forall \theta \in \Theta, \quad \text{Var}_\theta(U^*) \preceq \text{Var}_\theta(U).$$

Since T^* and U^* are unbiased for φ , we have:

$$\mathbb{E}_\theta(T^*(S) - U^*(S)) = 0 \quad \forall \theta \in \Theta,$$

which by completeness of S , this implies that $U^* = T^*$, \mathcal{M} -a.s. But since U was arbitrary, this shows that T^* is UMVUE, and the uniqueness claim follows immediately. \square

In the proof above, we see that the main interest of completeness is in fact to guarantee that Rao-Blackwellization of any unbiased estimator will always produce the same estimator. This uniqueness is enough to conclude that T^* must be UMVUE.

3.4.2. Sufficient and necessary conditions: a geometric point of view

Lehmann-Scheffé's theorem provides sufficient conditions for being a UMVUE. The following theorem gives both sufficient and necessary conditions:

Theorem 3.6. *Let $T \in \mathcal{U}_\varphi$, and define:*

$$\mathcal{U}_0 := \{U : \mathcal{X} \rightarrow \mathbb{R}^\ell \text{ measurable}, \forall \theta \in \Theta, \mathbb{E}_\theta(U) = 0, \mathbb{E}_\theta(\|U\|^2) < \infty\}$$

the set of unbiased estimators of 0. The following are equivalent:

- (i) T is a UMVUE,
- (ii) $\text{Cov}_\theta(T, U) = 0$ for all $U \in \mathcal{U}_0$ and all $\theta \in \Theta$.

Proof. (i) \implies (ii). Choose $U \in \mathcal{U}_0$, $\lambda \in \mathbb{R}$, and $\theta \in \Theta$. Since T is UMVUE, we have

$$\begin{aligned} \forall \lambda \in \mathbb{R}, \text{Var}_\theta(T + \lambda U) &\succeq \text{Var}_\theta(T) \\ \iff \forall \lambda \in \mathbb{R}, \lambda^2 \text{Var}_\theta(U) + \lambda (\text{Cov}_\theta(T, U) + \text{Cov}_\theta(U, T)) &\succeq 0 \\ \iff \forall \lambda \in \mathbb{R}, \forall x \in \mathbb{R}^\ell, \lambda^2 x^\top \text{Var}_\theta(U)x + 2\lambda x^\top \text{Cov}_\theta(T, U)x &\geq 0. \end{aligned}$$

If $x \in \mathbb{R}^\ell$ is such that $x^\top \text{Var}_\theta(U)x = 0$, the right-hand side implies $x^\top \text{Cov}_\theta(T, U)x = 0$ since the only non negative linear function is the null. Assume now that x is such that $x^\top \text{Var}_\theta(U)x \neq 0$. Then, the above quadratic polynomial in λ with real roots $\lambda_1 = 0$ and

$$\lambda_2 = -2 \frac{x^\top \text{Cov}_\theta(T, U)x}{x^\top \text{Var}_\theta(U)x}$$

is non-negative on \mathbb{R} , which happens if and only if the two roots are the same, that is $x^\top \text{Cov}_\theta(T, U)x = 0$. We proved that $x^\top \text{Cov}_\theta(T, U)x = 0$ for all $x \in \mathbb{R}^\ell$, that is $\text{Cov}_\theta(T, U) = 0$.

(ii) \implies (i). Let $T' \in \mathcal{U}_\varphi$ be arbitrary. Clearly, $T - T' \in \mathcal{U}_0$. Take $\theta \in \Theta$. We have $\text{Cov}_\theta(T, T - T') = 0$ thus $\text{Var}_\theta(T) = \text{Cov}_\theta(T, T')$. This means that for any $x \in \mathbb{R}^\ell$,

$$x^\top \text{Var}_\theta(T)x = x^\top \text{Cov}_\theta(T, T')x = \mathbb{E}_\theta \left[x^\top (T - \varphi(\theta))(T' - \varphi(\theta))^\top x \right].$$

But by Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}_\theta \left[x^\top (T - \varphi(\theta))(T' - \varphi(\theta))^\top x \right] &= \text{Cov}_\theta(x^\top (T - \varphi(\theta)), x^\top (T' - \varphi(\theta))) \\ &\leq \sqrt{x^\top \text{Var}_\theta(T)x} \times \sqrt{x^\top \text{Var}_\theta(T')x}. \end{aligned}$$

In other words, $\text{Var}_\theta(T) \preceq \text{Var}_\theta(T')$. Since T' and $\theta \in \Theta$ are arbitrary, it follows that T is UMVUE. \square

This geometric point of view enables to show the uniqueness (\mathcal{M} -a.e.) of an UMVUE (but not the existence), without any appeal to Rao-Blackwell's (nor Lehman-Scheffé's) Theorem.

Corollary 3.1 (Uniqueness of UMVUE). *If $T_1, T_2 \in \mathcal{U}_\varphi$ are two UMVUEs of $\varphi : \Theta \rightarrow \mathbb{R}^\ell$, then $T_1 = T_2$, \mathcal{M} -a.e.*

Proof. Since T_1 and T_2 are UMVUEs, we have

$$\forall \theta \in \Theta, \text{Var}_\theta(T_1) = \text{Var}_\theta(T_2) =: V_\theta.$$

Note that \mathcal{U}_φ is an affine space, thus $T_3 := \frac{T_1+T_2}{2}$ also belongs to \mathcal{U}_φ . Moreover,

$$\begin{aligned} V_\theta &= \text{Var}_\theta(T_3) \\ &= \frac{1}{4}V_\theta + \frac{1}{4}V_\theta + \frac{1}{2}\text{Cov}_\theta(T_1, T_2) \\ &\preceq \frac{1}{4}V_\theta + \frac{1}{4}V_\theta + \frac{1}{2}V_\theta^{1/2}V_\theta^{1/2} = V_\theta, \end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. By definition of V_θ , T_3 is also UMVUE and one could use the equality case to conclude, or note that

$$\begin{aligned} V_\theta &= \text{Var}_\theta(T_3) \\ &= \frac{1}{4}V_\theta + \frac{1}{4}V_\theta + \frac{1}{2}\text{Cov}_\theta(T_1, T_2) \\ &= \frac{1}{4}\text{Var}_\theta(T_1 - T_2) + \text{Cov}_\theta(T_1, T_2) \\ &= \frac{1}{4}\text{Var}_\theta(T_1 - T_2) + \underbrace{\text{Cov}_\theta(T_1, T_2 - T_1)}_{=0} + V_\theta, \end{aligned}$$

which gives that for all $\theta \in \Theta$, $\text{Var}_\theta(T_2 - T_1) = 0$, that is since $T_2 - T_1$ has null expectation, $T_2 = T_1$, \mathcal{M} -a.e. \square

3.4.3. Some examples: the good and the bad of unbiased estimation

This section focuses on the optimality of unbiased estimators. However, restricting attention to unbiased estimators is not always appropriate, for instance because such estimators may simply not exist.

Example 3.8. Consider the model where $X \sim \text{Bin}(n, \theta)$. We would like to estimate $\varphi(\theta) = \frac{\theta}{1-\theta}$. Indeed, if $T : \mathcal{X} \rightarrow \mathbb{R}$ is an unbiased estimator of φ , then

$$\varphi(\theta) = \sum_{k=0}^n T(k) \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Hence, to be estimated without bias, φ must be a polynomial of degree at most n , which is not the case for $\theta/(1-\theta)$.

In the same vein, imposing a null bias can lead to absurd estimators.

Example 3.9 (Absurd UMVUE). For an unknown parameter $\theta > 0$, let X have distribution:

$$\mathbb{P}_\theta(X = k) = \frac{\theta^k e^{-\theta}}{k!(1 - e^{-\theta})}, \quad k \in \mathbb{N}^*.$$

Consider the estimation of $\varphi(\theta) = e^{-\theta}$. Using identifiability of power series, the unique unbiased estimator (hence also UVMUE) is $T(X) = (-1)^{X+1}$, which oscillates between 1 and -1 , an absurd estimator since $0 < e^{-\theta} < 1$.

Even when a UMVUE exists and has nice properties, it is more relevant to consider estimators with small quadratic risk. Simply minimizing variance while enforcing zero bias does not necessarily produce the estimator with the lowest overall risk.

Example 3.10. Consider the Gaussian linear model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_n),$$

where $X \in \mathbb{R}^{n \times p}$ is known, $\beta \in \mathbb{R}^p$ is unknown with $\|\beta\| = 1$. Suppose for simplicity that the design is orthonormal, so that $X^\top X = nI_p$. We will see in ?? that the ordinary least squares estimator

$$\hat{\beta}_{\text{ols}} = (X^\top X)^{-1} X^\top Y$$

is UMVUE with variance $\text{Var}_\beta(\hat{\beta}_{\text{ols}}) = (X^\top X)^{-1} = I_p/n$, so that its quadratic risk function equals $R_\beta(\hat{\beta}_{\text{ols}}) = p/n$.

For $\lambda \geq 0$, the ridge estimator

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I_p)^{-1} X^\top Y$$

is biased, since

$$\mathbb{E}_\beta[\hat{\beta}_{\text{ridge}}] = (X^\top X + \lambda I_p)^{-1} X^\top X \beta = \frac{n}{n + \lambda} \beta,$$

and its variance is

$$\text{Var}_\beta(\hat{\beta}_{\text{ridge}}) = (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1} = \frac{n}{(n + \lambda)^2} I_p,$$

Its quadratic risk is then

$$R_\beta(\hat{\beta}_{\text{ridge}}) = \frac{\lambda^2 p}{(n + \lambda)^2} + \frac{np}{(n + \lambda)^2}.$$

A direct computation shows that the risk is minimized by $\lambda = p$, giving

$$R_\beta(\hat{\beta}_{\text{ridge}}) = \frac{n + p}{(n + p)^2} p + \frac{np}{(n + p)^2} = \frac{p}{n + p},$$

which is strictly better than the risk p/n achieved by the UMVUE estimator.

APPENDIX A

STANDARD DISTRIBUTIONS

A.1. Discrete distributions

Distribution	Parameters	Support	Density (on the support)	Mean	Variance
Bernoulli, $\text{Ber}(p)$	$p \in (0, 1)$	$\{0, 1\}$	$k \mapsto p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial, $\text{Bin}(n, p)$	$n \in \mathbb{N}, p \in (0, 1)$	$\{0, \dots, n\}$	$k \mapsto \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Discrete Uniform, $\text{Unif}(\{a, \dots, b\})$	$a, b \in \mathbb{Z}, a < b$	$\{a, \dots, b\}$	$k \mapsto \frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Geometric, $\text{Geom}(p)$	$p \in (0, 1)$	\mathbb{N}^*	$k \mapsto p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson, $\text{Poi}(\lambda)$	$\lambda > 0$	\mathbb{N}	$k \mapsto e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ

A.2. Continuous distributions

Distribution	Parameters	Support	Density (w.r.t. Lebesgue measure)	Mean	Variance
Uniform, $\text{Unif}([a, b])$	$a < b$	$[a, b]$	$x \mapsto \frac{1}{b-a} \mathbb{1}_{x \in [a, b]}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential, $\text{Exp}(\lambda)$	$\lambda > 0$	$[0, +\infty[$	$x \mapsto \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal (multivariate), $\mathcal{N}(\mu, \Sigma)$	$\mu \in \mathbb{R}^d, \Sigma \succ 0$	\mathbb{R}^d	$x \mapsto \frac{1}{(2\pi)^{d/2} \sqrt{ \det(\Sigma) }} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$	μ	Σ
Gamma, $\text{Gamma}(\alpha, \beta)$	$\alpha > 0, \beta > 0$	$[0, +\infty[$	$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{x \geq 0}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta, $\text{Beta}(\alpha, \beta)$	$\alpha > 0, \beta > 0$	$[0, 1]$	$x \mapsto \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{1}_{x \in [0, 1]}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

A.3. Distributions from the Gaussian world

Distribution	Parameters	Support	Definition (it is the law of...)	Mean	Variance
Chi-square, $\chi^2(d)$	$d \in \mathbb{N}^*$	$[0, +\infty[$	$\ Z\ ^2 = Z_1^2 + \dots + Z_d^2$ where $Z \sim \mathcal{N}(0_d, I_d)$.	d	$2d$
Student, $\mathcal{T}(p)$	$p \in \mathbb{N}^*$	\mathbb{R}	$T = \frac{Z}{\sqrt{K/p}}$ where $Z \sim \mathcal{N}(0, 1)$ and $K \sim \chi^2(p)$ are independent	0, if $p \geq 2$	$\frac{p}{p-2}$, if $p \geq 3$
Fisher, $F(p_1, p_2)$	$p_1, p_2 \in \mathbb{N}^*$	$[0, +\infty[$	$F = \frac{K_1/p_1}{K_2/p_2}$, where $K_1 \sim \chi^2(p_1)$ and $K_2 \sim \chi^2(p_2)$ are independent	$\frac{p_2}{p_2-2}$	$\frac{2p_2^2(p_1+p_2-2)}{p_1(p_2-2)^2(p_2-4)}$

APPENDIX B

REMINDER ON STANDARD PROBABILITY THEORY

B.1. Reminder on measure theory

This section briefly recalls some key concepts of measure theory that will be useful in this course.

B.1.1. Measures

The general setting is as follows: we consider a measured space (E, \mathcal{E}, μ) equipped with a σ -algebra \mathcal{E} and a measure μ , that is an function $\mu : \mathcal{E} \rightarrow [0, +\infty]$ such that $\mu(\emptyset) = 0$ and for all sequence $(A_n)_{n \geq 1}$ of pairwise disjoint elements of \mathcal{E} , μ satisfies the σ -additivity property:

$$\mu \left(\bigcup_{n \geq 1} A_n \right) = \sum_{n \geq 1} \mu(A_n).$$

In this course, we will focus on two main cases of interest: discrete measures, and continuous real measures.

Example B.1.

- *Counting measure:* $(E, \mathcal{E}, \mu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$ where μ is the counting measure on \mathbb{N} defined for all $A \subset \mathbb{N}$ by $\mu(A) = |A|$, which can be infinite. This measure μ can be described as a sum of Dirac measures:

$$\mu = \sum_{k=0}^{+\infty} \delta_k,$$

where we recall $\delta_k(A) = \mathbb{1}_{k \in A}$. Seen like this, any function $\phi : E \rightarrow \mathbb{R}$ corresponds to a sequence $(\phi_k)_{k \geq 0}$, and if $\sum \phi_k$ converges absolutely, ϕ is μ -integrable and

$$\int_E \phi(x) \mu(dx) = \sum_{k \geq 0} \phi_k \mu(\{k\}) = \sum_{k \geq 0} \phi_k.$$

- *Lebesgue measure:* $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ where $\mathcal{B}(\mathbb{R})$ is the real Borelian σ -algebra (spanned by the open intervals), and λ is the Lebesgue measure on \mathbb{R} , which verifies $\lambda((a, b)) = b - a$ with by convention $+\infty - a = +\infty - (-\infty) = b - (-\infty) = +\infty$.

Note that these two measures are not finite, since $\mu(\mathbb{N}) = \lambda(\mathbb{R}) = +\infty$, but they are σ -finite.

Definition B.1 (σ -finite measures). Let (E, \mathcal{E}, μ) be a measured space. We say that μ is σ -finite if there exists a sequence $(E_n)_{n \geq 0}$ of measurable sets such that $\mu(E_n) < \infty$ for all $n \geq 0$, and $E = \bigcup_{n \geq 0} E_n$.

Remark B.1. The counting measure on \mathbb{N} is σ -finite: take $E_n = \{0, \dots, n\}$. The Lebesgue measure on \mathbb{R} is also σ -finite: take $E_n = [-n, n]$. However, the counting measure on \mathbb{R} is not σ -finite (proof left as an exercise).

B.1.2. Absolute continuity, Radon-Nikodym derivative

Absolute continuity is a pre-order over measures. It will be useful in Statistics to defined dominated models.

Definition B.2 (Absolute continuity). Let (E, \mathcal{E}) be a measurable space, μ and ν two measures on this space. We say that μ is *absolutely continuous* w.r.t. ν , which we denote $\mu \ll \nu$, if any null-measure set for ν also has a null μ -measure, that is

$$\forall A \in \mathcal{E}, \nu(A) = 0 \implies \mu(A) = 0.$$

If $\mu \ll \nu$ and $\nu \ll \mu$, the two measures are said to be *equivalent*.

Remark B.2. If $\mu \ll \nu$, then every property which holds ν -a.s. also holds μ -a.s.

When $\mu \ll \nu$ and both are σ -finite, we can define a density of μ with respect to ν .

Theorem B.1 (Radon-Nikodym theorem). Let (E, \mathcal{E}) be a measurable space, μ and ν two non-negative σ -finite measures on this space. If $\mu \ll \nu$, then there exists a non-negative measurable function $f : E \rightarrow \mathbb{R}_+$ which we denote $f = \frac{d\mu}{d\nu}$ such that for any μ -integrable function ϕ , we have

$$\int_E \phi(x) \mu(dx) = \int_E \phi(x) \frac{d\mu}{d\nu}(x) \nu(dx) = \int_E \phi(x) f(x) \nu(dx).$$

Moreover, this function f is unique up to equality ν -almost everywhere.

Remark B.3. This does not hold if the measures are no longer σ -finite. For example, note that on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ any measure is absolutely continuous w.r.t. the counting measure μ on this space (why?). This is the case of the Lebesgue measure λ . However, λ has no density w.r.t. to μ on \mathbb{R} . If it were the case, taking $\phi(x) = \mathbb{1}_{x=a}$ for some $a \in \mathbb{R}$ in the equation of Theorem B.1 would give

$$0 = \int_E \mathbb{1}_a(x) \lambda(dx) = \int_E \mathbb{1}_a(x) f(x) \mu(dx) = f(a),$$

hence $f = 0$, and for $\phi = \mathbb{1}_{[0,1]}$ we have

$$1 = \int_E \mathbb{1}_{[0,1]}(x) \lambda(dx) = \int_E \mathbb{1}_{[0,1]}(x) f(x) \mu(dx) = 0,$$

which is absurd.

B.1.3. Real random variables, random vectors, expectation and variance

Throughout, we consider a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, that is a measurable space with measure \mathbb{P} having total mass 1.

Definition B.3 (Random variable, random vector). A *random variable*¹ is a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A *random vector*² is a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The law (or distribution) \mathbb{P}_X of a random vector X is defined for all borelian set $B \in \mathcal{B}(\mathbb{R})$ by $\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$.

¹in this course, all random variables are real.

²in this course, all random vectors take their values in \mathbb{R}^d .

Remark B.4. Note that since the projection on the k -th coordinate is continuous hence measurable, if $X = (X_1, \dots, X_d)$ is a random vector in \mathbb{R}^d , each of its coordinates are random variables.

Definition B.4 (Expectation, variance, covariance). Let X be a random variable. If X is integrable, we define its *expectation* as

$$\mathbb{E}[X] := \int X(\omega) d\mathbb{P}(\omega) = \int x \mathbb{P}_X(x).$$

If moreover X^2 is integrable (we say that X has finite second moment), we define the *variance* of X as

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

Moreover, if X, Y are two random variables with finite second moment, their *covariance* is defined by

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

From the above definition, it is easily seen that the expectation is linear over the real vector space of integrable random variables. The covariance is a bilinear operator on the real vector space of random variables with finite second moment, and the variance is its associated quadratic form, which is positive.

Definition B.5 (Expectation, covariance matrix of a random vector). Let $X = (X_1, \dots, X_d)$ be a random vector in \mathbb{R}^d . If X_1, \dots, X_d are integrable, the *expectation* of X is defined as

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])^T \in \mathbb{R}^d.$$

If moreover X_1, \dots, X_d have finite second moments (we say that the vector X has finite second moment), the *covariance matrix* of X is defined as We define the *covariance matrix* of X by

$$\text{Var}(X) := \mathbb{E}[(X - \mu)(X - \mu)^T] \in \mathbb{R}^{d \times d},$$

that is, for all $1 \leq i, j \leq d$, $[\text{Var}(X)]_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \text{Cov}(X_i, X_j)$. Thus, $\text{Var}(X)$ is a symmetric matrix. These definitions coincide with the usual expectation and variance of a random variable when $d = 1$.

In their vectorial forms, the expectation and covariance operators inherit from their properties in dimension 1.

Proposition B.1. Let X be a random vector in \mathbb{R}^d with a finite second-order moment. Let $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. Then $Y = AX + b$ is a random vector in \mathbb{R}^m which also has a finite second-order moment, and we have:

$$\mathbb{E}[Y] = A\mathbb{E}[X] + b \quad \text{and} \quad \text{Var}(Y) = A\text{Var}(X)A^T.$$

Proof. Writing $Y = (Y_1, \dots, Y_d)$, it is readily seen that for all $1 \leq i \leq d$, $Y_i = \sum_{k=1}^d A_{i,k}X_k + b_i$, and by linearity of expectation in dimension 1, $\mathbb{E}[Y_i]$ is finite and $\mathbb{E}[Y_i] = \sum_{k=1}^d A_{i,k}\mathbb{E}[X_k] + b_i = (A\mathbb{E}[X] + b)_i$. The coordinates of Y are affine transformations of coordinates of X , so they still all have finite second moment. A direct computation gives

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}[(AX + b - (A\mathbb{E}[X] + b))(AX + b - (A\mathbb{E}[X] + b))^T] \\ &= \mathbb{E}[(AX - A\mathbb{E}[X])(AX - A\mathbb{E}[X])^T] = \mathbb{E}[A(X - \mathbb{E}[X])(X - \mathbb{E}[X])A^T] \\ &= A\text{Var}(X)A^T, \end{aligned}$$

using now linearity of (vectorial) expectation. □

Remark B.5. With the above property, we have that for all $a \in \mathbb{R}^d$, $a^T \Sigma a = \text{Var}(a^T X) \geq 0$. A covariance matrix is therefore always symmetric positive.

B.2. Convergence of random variables

Throughout, $\|\cdot\|$ denotes the canonical euclidean norm on \mathbb{R}^d .

B.2.1. Convergence in probability

Definition B.6 (Convergence in probability). A sequence $(X_n)_{n \geq 1}$ of random vectors in \mathbb{R}^d *converges in probability* to a random vector X in \mathbb{R}^d , and we denote $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} X$, if for all $\varepsilon > 0$,

$$\mathbb{P}(\|X_n - X\| \geq \varepsilon) \xrightarrow[n \rightarrow \infty]{} 0.$$

Lemma B.1 (Convergence in probability and continuous transformations). *Let $(X_n)_{n \geq 0}$ and X be random vectors such that $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} X$. If $f : \mathbb{R}^d \rightarrow \mathbb{R}^q$ is a continuous function, then $f(X_n) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} f(X)$.*

Proof of Lemma B.1. Let $\varepsilon > 0$ and $M > 0$. Since f is continuous, it is uniformly continuous on the compact ball $B = B(0, M)$. Let $\eta > 0$ be a ε -uniform continuity modulus for $f|_B$. Wlog we can assume that $\eta < 1$. If $x, y \in \mathbb{R}^q$ are such that $\|f(x) - f(y)\| \geq \varepsilon$, then either $\|y\| > M$, either $y \in B$ and we are in either of these cases: (i) $\|x\| > M$ or (ii) $\|x\| \leq M$ but $\|x - y\| > \eta$. In both cases (i) and (ii), we have $|x - y| > \eta$. This gives the following bound:

$$\mathbb{P}(\|f(X_n) - f(X)\| \geq \varepsilon) \leq \mathbb{P}(\|X\| > M) + \mathbb{P}(\|X_n - X\| > \eta).$$

Fix $\varepsilon' > 0$. The first member goes to 0 when $M \rightarrow \infty$ by the dominated convergence theorem, hence it is less than ε' for M large enough. Then, η being fixed by M the second term is also $\leq \varepsilon'$ for n large enough since $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} X$. \square

B.2.2. Almost sure convergence

Almost sure convergence can be seen as the stochastic version of simple convergence of functions.

Definition B.7 (Almost sure convergence). A sequence $(X_n)_{n \geq 0}$ of random vectors *converges almost surely* to a random vector X , and we denote $X_n \xrightarrow[n \rightarrow \infty]{a.s.} X$, if

$$\mathbb{P}(\lim_n X_n = X) = \mathbb{P}(\{\omega \in \Omega, \lim_n X_n(\omega) = X(\omega)\}) = 1.$$

Remark B.6. It is easy to see that if $f : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a continuous function, then $f(X_n) \xrightarrow[n \rightarrow \infty]{a.s.} f(X)$ as soon as $X_n \xrightarrow[n \rightarrow \infty]{a.s.} X$.

Lemma B.2. *Almost sure convergence implies convergence in probability.*

Proof. Apply Fatou's Lemma to $f_n := \mathbb{1}_{\|X_n - X\| < \varepsilon}$. By assumption, these functions satisfies $\liminf_n f_n = 1$ on a measurable set of measure 1, and we have

$$1 = \int \liminf_n f_n d\mathbb{P} \leq \liminf_n \int f_n d\mathbb{P} = \liminf_n \mathbb{P}(\|X_n - X\| < \varepsilon).$$

\square

Remark B.7. One of the (few) ways to prove almost sure convergence is appealing to Borel-Cantelli's Lemma. Namely, if for all $\varepsilon > 0$, $\sum_{n \geq 0} \mathbb{P}(\|X_n - X\| \geq \varepsilon) < \infty$, then $(X_n)_{n \geq 0}$ converges almost surely to X . The proof of this is left as an exercise.

B.2.3. Convergence in distribution

We now switch to convergence in distribution, which is a central concept in statistics. For the sake of completeness, we will define this convergence for *random vectors*, that is vectors of the form $Z = (X_1, \dots, X_d)$ where X_1, \dots, X_d are (real) random variables. Z hence takes values in some \mathbb{R}^d .

Definition B.8 (Convergence in distribution). Let $(Z_n)_{n \geq 1}$, Z be random vectors in \mathbb{R}^d . We say that $(Z_n)_{n \geq 1}$ *converges in distribution* to (the distribution of) Z , and we denote $Z_n \xrightarrow[n \rightarrow \infty]{(d)} Z$, if for all continuous bounded functions $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\mathbb{E}[\phi(Z_n)] \xrightarrow[n \rightarrow \infty]{} \mathbb{E}[\phi(Z)]. \quad (\text{B.1})$$

Remark B.8. Contrary to the first two convergence modes, we are not comparing the values of the random vectors $Z_n(\omega)$ and $Z(\omega)$ for all $\omega \in \Omega$ (like taking the difference), but only their distributions via the expectations in (B.1). In particular, these variables need not be defined on the same probability space! To emphasize that this is the sequence of distributions that converge, we often denote $Z_n \xrightarrow[n \rightarrow \infty]{(d)} D$, where D is a probability distribution (that of Z).

Remark B.9. In particular, as we will see below, convergence in distribution is weaker than convergence in probability. See a first simple example, if $X_n = X \sim \mathcal{N}(0, 1)$, $X_n \xrightarrow[n \rightarrow \infty]{(d)} -X$ by symmetry of the standard gaussian, but we do not have $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} -X$ since for instance $\mathbb{P}(|X_n - (-X)| > 1) = \mathbb{P}(2|X| > 1)$ is a positive constant.

Remark B.10. Here again, if $Z_n \xrightarrow[n \rightarrow \infty]{(d)} Z$ then $f(Z_n) \xrightarrow[n \rightarrow \infty]{(d)} f(Z)$ for any continuous (thus measurable) function $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$. Indeed, if ϕ is bounded and continuous, so is $\phi \circ f$.

Lemma B.3 (Convergence in probability implies converge in distribution (real case)). *If a sequence of random vectors $(Z_n)_{n \geq 1}$ in \mathbb{R}^d converges in probability to $Z \in \mathbb{R}^d$, then the convergence also happens in distribution. The reciprocal is not true, but it is true if X is a.s. constant.*

Proof of Lemma B.3. Take a continuous bounded function $f : \mathbb{R}^d \rightarrow \mathbb{R}$. By Lemma B.1, $f(Z_n) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} f(Z)$. Let $\varepsilon > 0$. We have

$$\begin{aligned} |\mathbb{E}[f(Z_n) - f(Z)]| &\leq \mathbb{E}[|f(Z_n) - f(Z)|] \\ &\leq \mathbb{E}[|f(Z_n) - f(Z)| \mathbf{1}_{|f(Z_n) - f(Z)| \leq \varepsilon}] + \mathbb{E}[|f(Z_n) - f(Z)| \mathbf{1}_{|f(Z_n) - f(Z)| > \varepsilon}] \\ &\leq \varepsilon + 2\|f\|_\infty \mathbb{P}(|f(Z_n) - f(Z)| > \varepsilon). \end{aligned}$$

By convergence in probability, the second term is less than ε for n large enough.

The reciprocal is not true in general (see Remark B.9) but we will show that it is true if $Z = (X_1, \dots, X_d)$ is a.s. constant. We first show this in the real case ($d = 1$). Assume that a sequence of random variables $(X_n)_{n \geq 1}$ converges in distribution to a constant $c \in \mathbb{R}^d$. Then, for all $\varepsilon > 0$, $c \pm \varepsilon$ is a continuity point of the c.d.f. of the r.v. c which is $\mathbf{1}_{\geq c}$, and we have by Theorem B.2:

$$\mathbb{P}(|X_n - c| \geq \varepsilon) = 1 - \mathbb{P}(X_n \in]c - \varepsilon, c + \varepsilon]) \xrightarrow[n \rightarrow \infty]{} 1 - \mathbb{P}(c \in]c - \varepsilon, c + \varepsilon]) = 0.$$

We conclude for the multi-dimensional case by noticing that, denoting $Z_n = (Z_{n,1}, \dots, Z_{n,d})$,

$$\mathbb{P}(\|Z_n - Z\| \geq \varepsilon) \leq \sum_{j=1}^d \mathbb{P}(|Z_{n,j} - Z_j| \geq \varepsilon/\sqrt{d}).$$

□

B.2.4. A criterion for convergence in distribution in the real case

In practice, condition (B.1) is cumbersome to check or establish. We hereafter start with a simple criterion characterization of convergence in distribution for the real-valued case. The law of a random variable X ($d = 1$) is characterized by a simpler object which is its *cumulative distribution function (c.d.f.)* F_X defined as follows:

$$\forall x \in \mathbb{R}, F_X(x) := \mathbb{P}(X \leq x) = \mathbb{P}_X([-\infty, x]).$$

Recall that this function is right-continuous, nondecreasing, and satisfies $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$. There is a discontinuity at x_0 iff X has a atom at x_0 .

Theorem B.2 (Convergence in distribution, simple convergence of c.d.f.s). *The following propositions are equivalent.*

- (i) the sequence of r.v.s. $(X_n)_{n \geq 1}$ converges in distribution to X ;
- (ii) for all continuity point x of F_X , $F_{X_n}(x) \xrightarrow{n \rightarrow \infty} F_X(x)$;
- (iii) There exists a common probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and r.v.s $(X'_n)_{n \geq 1}, X$ defined on this space such that
 - X' and X have same distribution,
 - for all $n \geq 0$, X_n and X'_n have same distribution,
 - $(X'_n)_{n \geq 1}$ converges almost surely to X' .

The implication (i) \implies (iii) is valid for more general case (r.v.s on Polish spaces) and constitutes the *Skorokod representation Lemma*.

Proof of Theorem B.2. (iii) \implies (i). Take a bounded, continuous function ϕ , we need to show that $\mathbb{E}[\phi(X_n)] \rightarrow \mathbb{E}[\phi(X)]$, that is $\mathbb{E}'[\phi(X'_n)] \rightarrow \mathbb{E}'[\phi(X')]$. But since f is continuous, $f(X'_n) \rightarrow f(X')$ almost surely. Since f is bounded, we obtain the convergence of the expectations (integrals) thanks to the dominated convergence theorem ($|f| \leq \|f\|_\infty$).

(i) \implies (ii). For this implication, we would like to apply the definition to function $g_x : t \mapsto \mathbb{1}_{t \leq x}$, but this is not continuous. Hence, we approximate it with continuous functions. For $a < b$, denote $\phi_{a,b}$ the only function such that $\phi_{a,b}(t) = 1$ when $x \leq a$, $\phi_{a,b}(t) = 0$ when $x > b$, and $\phi_{a,b}$ is linear between a and b . It is then clear that for all $x \in \mathbb{R}$, $n \geq 0$, $\varepsilon > 0$,

$$\phi_{x-\varepsilon, x}(X_n) \leq \mathbb{1}_{X_n \leq x} \leq \phi_{x, x+\varepsilon}(X_n)$$

and taking the expectations gives

$$\mathbb{E}[\phi_{x-\varepsilon, x}(X_n)] \leq F_{X_n}(x) \leq \mathbb{E}[\phi_{x, x+\varepsilon}(X_n)].$$

This gives in turn $\limsup F_{X_n}(x) \leq \limsup \mathbb{E}[\phi_{x, x+\varepsilon}(X_n)] = \mathbb{E}[\phi_{x, x+\varepsilon}(X)] \leq \mathbb{P}(X \leq x + \varepsilon) = F_X(x + \varepsilon)$, and on the other hand $\liminf F_{X_n}(x) \geq \liminf \mathbb{E}[\phi_{x-\varepsilon, x}(X_n)] = \mathbb{E}[\phi_{x-\varepsilon, x}(X)] \geq \mathbb{P}(X \leq x - \varepsilon) = F_X(x - \varepsilon)$. Then, letting $\varepsilon \rightarrow 0$ we get

$$F_X(x^-) \leq \liminf F_{X_n}(x) \leq \limsup F_{X_n}(x) \leq F_X(x),$$

and whenever x is a continuity point of F_X , then the above is an equality and (ii) is proven.

(ii) \implies (iii). We start by defining the generalized inverse. For any nondecreasing right-continuous real function f , its (left-continuous) generalized inverse is defined by

$$g : y \mapsto \inf \{x \in \mathbb{R}, f(x) \geq y\}$$

with $\inf \emptyset = +\infty$ by convention. Note that g is also nondecreasing. We have the following Lemma:

Lemma B.4. *Let $(f_n)_{n \geq 0}$, f be nondecreasing real functions. Assume that for all $x \in [a, b]$ that is a continuity point of f , $f_n(x) \rightarrow f(x)$. Let $(g_n)_{n \geq 0}$, g be the generalized inverses of $(f_n)_{n \geq 0}$, f . Then, for each point y in the interval $[f(a), f(b)]$ that is a continuity point of g , we have $g_n(y) \rightarrow g(y)$.*

Proof of Lemma B.4. Let y in the interval $[f(a), f(b)]$ be a continuity point of g . Let $\varepsilon > 0$. Since the continuity points of f is a dense set, we can choose $\varepsilon_1 \in]0, \varepsilon[$ such that $g(y) - \varepsilon_1$ is a continuity point of f . We have by definition $f(g(y) - \varepsilon_1) < y$ otherwise $g(y) \leq g(y) - \varepsilon_1$ which is contradictory. Choose $\delta \in]0, y - f(g(y) - \varepsilon_1)[$. Since $g(y) - \varepsilon_1$ is a continuity point of f , there exists $n_0 \geq 0$ such that for $n \geq n_0$, $f_n(g(y) - \varepsilon_1) \leq f(g(y) - \varepsilon_1) + \delta < y$. This implies that for $n \geq n_0$, $g_n(y) \geq g(y) - \varepsilon_1$. Another inequality of the form $g_n(y) \leq g(y) + \varepsilon_2$ is obtained similarly.

Take $\Omega' =]0, 1[$, \mathcal{F}' the Borelians of \mathbb{R} restricted to $]0, 1[$, and \mathbb{P}' the Lebesgue measure on $]0, 1[$. Denote by G , resp. $(G_n)_{n \geq 0}$ the generalized inverses of F_X , resp. $(F_{X_n})_{n \geq 0}$. Setting $X' = G$ and $X'_n = G_n$ satisfies the two first points, since we have the crucial equality $\{G(\omega) \leq x\} = \{\omega \leq F(x)\}$. Lemma B.4 implies that $(G_n)_n$ converges almost surely to G , since the non continuity points of G are at most countable, hence of measure 0. This concludes the proof of Theorem B.2. \square

Remark B.11. Theorem B.2 above gives us that in the real case, if $X_n \xrightarrow[n \rightarrow \infty]{(d)} X$ and a, b are continuity points of F_X , then

$$\mathbb{P}(X_n \in (a, b)) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(X \in (a, b)),$$

where (a, b) may be an open, semi-open, closed interval.

Remark B.12. Note that Theorem B.2 is specifically practical when dealing with minima/maxima of i.i.d. variables, because F_{X_n} can be computed easily.

B.2.5. A general criterion for convergence in distribution in the multidimensional case

Another method to check convergence in distribution in the multidimensional case is to use the characteristic function. We hereafter recall its definition and how it can be used for convergence in distribution.

Definition B.9 (Characteristic function). The *characteristic function* of a random vector Z taking values in \mathbb{R}^d is denoted Φ_Z and defined for each $t \in \mathbb{R}^d$ as follows:

$$\Phi_Z(t) := \mathbb{E}[e^{i\langle t, X \rangle}]. \quad (\text{B.2})$$

Note that this is nothing but a Fourier transform. A well-known result is that it characterizes the distribution of X . If X is real valued, the above simplifies to $\Phi_X(t) = \mathbb{E}[e^{itX}]$. The celebrated Lévy's Theorem establishes the equivalence between convergence in distribution and simple convergence of the characteristic function.

Theorem B.3 (Paul Lévy's theorem, 1922). *The sequence of random vectors $(Z_n)_{n \geq 1}$ converges in distribution to Z if and only if, for all $t \in \mathbb{R}^d$,*

$$\Phi_{Z_n}(t) \xrightarrow{n \rightarrow \infty} \Phi_Z(t).$$

The proof of this Theorem is beyond the scope of this course, and relies on Fourier analysis.

Remark B.13. Note that Theorem B.3 is specifically practical when dealing with sums of i.i.d. variables, because Φ_{Z_n} can be computed easily.

B.3. Classical convergence theorems

In many applications, convergence (almost sure, or in distribution) can be the consequence of these celebrated results.

B.3.1. The strong Law of Large Numbers

Theorem B.4 (Strong law of large numbers). *Let $(X_n)_{n \geq 1}$ be i.i.d. random vectors with finite mean $\mu \in \mathbb{R}^d$. Then, we have the almost sure convergence:*

$$\bar{X}_n := \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow[n \rightarrow \infty]{a.s.} \mu.$$

If the LLN can be thought of as a zero-order expansion of \bar{X}_n , the central limit theorem can be viewed as a (stochastic) first order expansion of \bar{X}_n , at the price of additional moment assumptions.

B.3.2. The Central Limit Theorem

In order to state and prove this theorem, we need a few notions on Gaussian vectors. A more extensive reminder is given in Chapter C.

Definition B.10 (Gaussian random variables). A r.v. X is a *Gaussian random variable* of mean $\mu \in \mathbb{R}$ and variance σ^2 with $\sigma > 0$, and we denote $X \sim \mathcal{N}(\mu, \sigma^2)$, if its distribution has the following density on \mathbb{R} , with respect to the Lebesgue measure:

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

If $\mu = 0$ and $\sigma = 1$, X is called *standard Gaussian*. If $\sigma = 0$, we extend the definition, X is said to be *degenerate Gaussian*: it is a r.v. a.s. equal to μ .

Definition B.11 (Gaussian random vectors). A random vector X is a *Gaussian random vector* if any linear combination of coefficient of X is a Gaussian random variable. A Gaussian vector X always has a finite covariance matrix. If X has mean μ and covariance matrix Σ , we denote $X \sim \mathcal{N}(\mu, \Sigma)$.

Lemma B.5 (Characteristic function of a Gaussian vector). *If $X \sim \mathcal{N}(\mu, \Sigma)$ in \mathbb{R}^d , then for all $s \in \mathbb{R}^d$,*

$$\Phi_X(s) = \exp\left(is^T \mu - \frac{1}{2}s^T \Sigma s\right).$$

Theorem B.5 (Central limit theorem, multidimensional case). *Let $(X_n)_{n \geq 1}$ be i.i.d. random vectors of finite mean $\mu \in \mathbb{R}^d$ and finite covariance matrix Σ . We denote $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$.*

Then, we have the following convergence in distribution:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Sigma).$$

A consequence of this result is that in the real case, if $Z_n := \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$, then for all $a, b \in \mathbb{R}$, $\mathbb{P}(a \leq Z_n \leq b) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(a \leq Z \leq b)$ where Z is a standard gaussian variable.

Proof of Theorem B.5 by Lévy's proof. Using Lévy's Theorem (Theorem B.3), the proof is very simple. Note that since the X_i are i.i.d. and have finite covariance matrix Σ , $\Phi_{X_i - \mu}(t) = 1 - \frac{1}{2}t^T \Sigma t + o(\|t\|^2)$. Let $Z_n := \sqrt{n}(\bar{X}_n - \mu)$. Its characteristic function Φ_{Z_n} writes for all $t \in \mathbb{R}$,

$$\begin{aligned} \Phi_{Z_n}(t) &:= \Phi_{\frac{(X_1 - \mu) + \dots + (X_n - \mu)}{\sqrt{n}}}(t) \\ &= \mathbb{E} \left[\exp \left(it^T \frac{(X_1 - \mu) + \dots + (X_n - \mu)}{\sqrt{n}} \right) \right] \\ &= \Phi_{X_1 - \mu} \left(\frac{t}{\sqrt{n}} \right)^n \\ &= \left(1 - \frac{1}{2n} t^T \Sigma t + o(\|t\|^2/n) \right)^n \xrightarrow[n \rightarrow \infty]{} \exp \left(-\frac{1}{2} t^T \Sigma t \right), \end{aligned}$$

which is the characteristic function of a random vector with distribution $\mathcal{N}(0, \Sigma)$ by Lemma B.5. \square

Remark B.14. Can we go further in the asymptotic stochastic expansion, namely going at order 3? The question is a priori not clear, because we do not know how to metrize convergence in distribution (the r.v.s of the sequence may not even be on the same probability space). In other terms, what is the speed of convergence towards a distribution? The Berry-Esseen Theorem (admitted) gives a partial answer to the question in the real case, by comparing the c.d.f.s of the two sides, uniformly on \mathbb{R} :

Theorem B.6 (Berry-Esseen theorem). *Let $(X_n)_{n \geq 1}$ be i.i.d. random variables, centered, such that $\mathbb{E}[X_1^2] = \sigma^2 < \infty$, and $\mathbb{E}[|X_1|^3] = \rho < \infty$. Define F_n the c.d.f. of $\sqrt{n} \frac{\bar{X}_n}{\sigma}$, and F the c.d.f. of a standard gaussian variable. Then, there exists a universal constant $C > 0$ such that for all $x \in \mathbb{R}$,*

$$|F_n(x) - F(x)| \leq \frac{C\rho}{\sigma^3 \sqrt{n}}.$$

APPENDIX C

REMINDER ON GAUSSIAN VECTORS

C.1. Gaussian variables and Gaussian vectors

Definition C.1 (Gaussian variables). A real random variable Z is said to be *Gaussian with mean $\mu \in \mathbb{R}$ and variance σ^2 with $\sigma > 0$* , and we write $Z \sim \mathcal{N}(\mu, \sigma^2)$, if its law admits on \mathbb{R} the following density, with respect to the Lebesgue measure:

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

If $\mu = 0$ and $\sigma = 1$, Z is said to be *standard Gaussian*. If $\sigma = 0$, Z is said to be a *degenerate Gaussian*; it is a random variable almost surely equal to μ .

Proposition C.1 (Fundamental properties of Gaussian variables). *We have the following properties:*

(i) Characteristic function (important) $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if for all $t \in \mathbb{R}$,

$$\mathbb{E}[e^{itX}] = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)$$

(ii) Sum of two independent Gaussian variables (important) if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with X_1, X_2 independent, then

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

The proof of the previous proposition is left as an exercise. For the proof of (i) in the direction \implies , we can set $\Phi_X(t) = \mathbb{E}[e^{itX}] = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2)$, define f_σ such that $\Phi_X(t) = e^{it\mu} f_\sigma(t)$ and find a differential equation on f_σ . For (i) in the other direction, as well as (ii), we use the well-known fact that the characteristic function characterizes the law, see Theorem 8.22 of [3] for the proof.

Definition C.2 (Gaussian vector). A random vector $X = (X_1, \dots, X_d)$ taking values in \mathbb{R}^d is said to be *Gaussian* if every linear combination of its components is Gaussian (real one-dimensional). We then write $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = \mathbb{E}[X]$ and $\Sigma = \text{Var}(X)$. These quantities are well defined since, by definition, all components are Gaussian and thus have a finite second moment.

Remark C.1. According to (ii) of the proposition above, any random vector whose coordinates are independent Gaussians is therefore a Gaussian vector (with diagonal covariance matrix).

Proposition C.2. *If $X \sim \mathcal{N}(\mu, \Sigma)$ and if Σ is invertible, then X admits on \mathbb{R}^d , with respect to the Lebesgue measure, the density*

$$f_{\mu, \Sigma} := x \mapsto \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

This proposition is admitted and can be established via the characteristic function, given in the following proposition.

Proposition C.3. *Let $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ non-invertible. Then $X - \mu$ belongs almost surely to a vector subspace of dimension $\text{rg}(\Sigma) < d$. In particular, X does not admit a density with respect to the Lebesgue measure on \mathbb{R}^d .*

Proof. Suppose Σ is non-invertible and denote by $r < d$ its rank. By the rank theorem, $\text{Ker } \Sigma$ has dimension $d - r > 0$. For all $a \in \text{Ker } \Sigma$,

$$\mathbb{E}[a^T(X - \mu)] = 0 \quad \text{and} \quad \text{Var}[a^T(X - \mu)] = a^T \text{Var}(X - \mu) a = a^T \Sigma a = 0,$$

which shows that $a^T(X - \mu) = 0$ almost surely. Taking a basis of $\text{Ker } \Sigma$, we deduce that almost surely,

$$X - \mu \in (\text{Ker } \Sigma)^\perp,$$

which has dimension $r < d$. □

Proposition C.4 (Fundamental properties of Gaussian vectors). *We have the following properties:*

(i) Characteristic function. $X \sim \mathcal{N}(\mu, \Sigma)$ in \mathbb{R}^d if and only if for all $s \in \mathbb{R}^d$,

$$\mathbb{E}[e^{is^T X}] = \exp \left(is^T \mu - \frac{1}{2} s^T \Sigma s \right)$$

(ii) Sum of two independent Gaussian vectors. if $Z_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Z_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ in \mathbb{R}^d , independent, then

$$Z_1 + Z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2).$$

(iii) Stability under linear transformation. Let $X \sim \mathcal{N}(\mu, \Sigma)$ in \mathbb{R}^d . Then for any matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, the vector $AX + b$ is still Gaussian in \mathbb{R}^m and

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T).$$

(iv) Reading independence from correlations. Let $X \sim \mathcal{N}(\mu, \Sigma)$ in \mathbb{R}^d . For any $J \subseteq \{1, \dots, d\}$,

$$(X_j)_{j \in J} \text{ are mutually independent} \iff \Sigma_{J,J} \text{ is diagonal,}$$

where $\Sigma_{J,J}$ is the submatrix of Σ obtained by keeping only the rows and columns indexed by J .

These results are admitted and essentially rely on proving (i), which again, builds on the fact that the characteristic function determines the law of a random vector.

C.2. Cochran's Theorem and geometric properties of Gaussian vectors

Gaussian vectors are deeply intertwined with Euclidean geometry: many of their properties can be interpreted through orthogonal projections and decompositions of \mathbb{R}^d . The

following Cochran's Theorem make this geometric perspective precise. To discuss this result, we first need to introduce the chi-square distribution.

Definition C.3 (Chi-square distribution). Let $d \geq 1$ and $Z \sim \mathcal{N}(0_d, I_d)$. The *chi-square distribution with d degrees of freedom*, denoted $\chi^2(d)$, is the distribution of the sum of squares of its components:

$$\|Z\|^2 = Z_1^2 + \dots + Z_d^2 \sim \chi^2(d).$$

Theorem C.1 (Cochran's Theorem). Let $Z \sim \mathcal{N}(\mu, \sigma^2 I_d)$ be a Gaussian vector in \mathbb{R}^d . Let $r \geq 1$ and let E_1, \dots, E_r be vector subspaces that are pairwise orthogonal and such that

$$E_1 \oplus \dots \oplus E_r = \mathbb{R}^d.$$

For $1 \leq j \leq r$, denote by Π_j the orthogonal projector onto E_j and by d_j its dimension. Then:

(i) the random vectors $\Pi_1 Z, \dots, \Pi_r Z$ are Gaussian, mutually independent, with respective laws $\mathcal{N}(\Pi_1 \mu, \sigma^2 \Pi_1), \dots, \mathcal{N}(\Pi_r \mu, \sigma^2 \Pi_r)$;

(ii) the random variables

$$\frac{\|\Pi_1(Z - \mu)\|^2}{\sigma^2}, \dots, \frac{\|\Pi_r(Z - \mu)\|^2}{\sigma^2}$$

are mutually independent and have respective distributions $\chi^2(d_1), \dots, \chi^2(d_r)$.

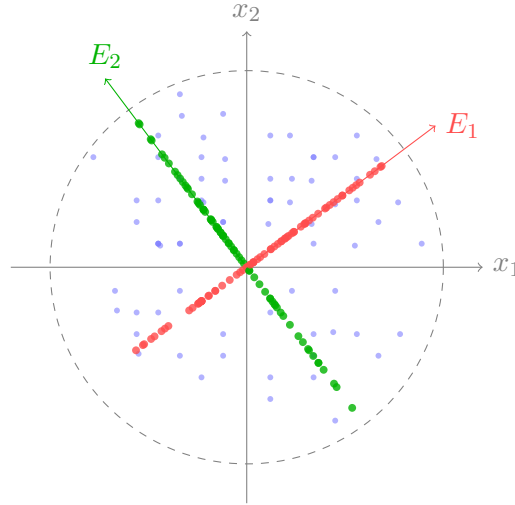


Figure C.1 – Illustration of Cochran's theorem in dimension 2, with $d_1 = d_2 = 1$. The green and red point clouds are independent.

Proof of Cochran's Theorem. Without loss of generality, we may assume that $\mu = 0$ and $\sigma^2 = 1$, that is, $Z \sim \mathcal{N}(0, I_d)$. Consider the concatenated vector

$$U := (\Pi_1 Z, \dots, \Pi_r Z) \in \mathbb{R}^{dr},$$

and define the block projection matrix

$$\Pi := \begin{pmatrix} \Pi_1 \\ \vdots \\ \Pi_r \end{pmatrix} \in \mathbb{R}^{dr \times d},$$

so that $U = \Pi Z$. To study the covariance of U , we compute

$$\text{Var}(U) = \Pi \text{Var}(Z) \Pi^T = \Pi \Pi^T = \begin{pmatrix} \Pi_1 \\ \vdots \\ \Pi_r \end{pmatrix} (\Pi_1, \dots, \Pi_r) = \begin{pmatrix} \Pi_1 \Pi_1 & \Pi_1 \Pi_2 & \cdots & \Pi_1 \Pi_r \\ \Pi_2 \Pi_1 & \Pi_2 \Pi_2 & \cdots & \Pi_2 \Pi_r \\ \vdots & & \ddots & \vdots \\ \Pi_r \Pi_1 & \cdots & \cdots & \Pi_r \Pi_r \end{pmatrix}.$$

Since the subspaces E_j are pairwise orthogonal, we have $\Pi_j \Pi_{j'} = 0$ for $j \neq j'$, and $\Pi_j^2 = \Pi_j$. Therefore, the matrix reduces to a block-diagonal matrix:

$$\text{Var}(U) = \Pi \Pi^T = \begin{pmatrix} \Pi_1 & 0 & \cdots & 0 \\ 0 & \Pi_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Pi_r \end{pmatrix}.$$

By Proposition C.4, the projected vectors $\Pi_j Z$ are mutually independent, and their marginals satisfy

$$\Pi_j Z \sim \mathcal{N}(0, \Pi_j \Pi_j^T) = \mathcal{N}(0, \Pi_j^2) = \mathcal{N}(0, \Pi_j),$$

which proves (i).

For (ii), for each j choose an orthonormal basis $(e_{j,1}, \dots, e_{j,d_j})$ of E_j . Then

$$\|\Pi_j Z\|^2 = \sum_{k=1}^{d_j} (e_{j,k}^T Z)^2,$$

where each $e_{j,k}^T Z \sim \mathcal{N}(0, 1)$ and the variables are independent. Hence,

$$\|\Pi_j Z\|^2 \sim \chi^2(d_j).$$

Since the vectors $\Pi_j Z$ are mutually independent, the random variables $\|\Pi_j Z\|^2$ are independent as well, which concludes (ii). \square

Figure C.2 illustrates the classical renormalization of a Gaussian vector, well known in dimension 1, and extended here to dimension d : if $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ invertible, then $\Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_d)$, where $\Sigma^{1/2}$ is the unique positive semi-definite symmetric square root of Σ .

Figure C.3 shows the typical shape of a Gaussian point cloud in \mathbb{R}^d . One may note that the smaller the eigenvalues of Σ , the closer one gets to the non-invertible (degenerate) case, and the more the ellipsoid tends to flatten: this is natural, since the random vector will eventually live in an effective space of dimension lower than d .

C.3. Two other classical distributions: Student and Fisher distributions

We will see the usefulness of these distributions for estimation in Gaussian models.

Definition C.4 (Student's distribution.). Let $Z \sim \mathcal{N}(0, 1)$ and $K \sim \chi^2(p)$ be *independent*. The *Student distribution with p degrees of freedom*, denoted $\mathcal{T}(p)$, is the distribution of

$$T = \frac{Z}{\sqrt{K/p}}.$$

Definition C.5 (Fisher's distribution.). Let $K_1 \sim \chi^2(p_1)$ and $K_2 \sim \chi^2(p_2)$ be *independent*. The *Fisher distribution with (p_1, p_2) degrees of freedom*, denoted $\mathcal{F}(p_1, p_2)$, is the distribution

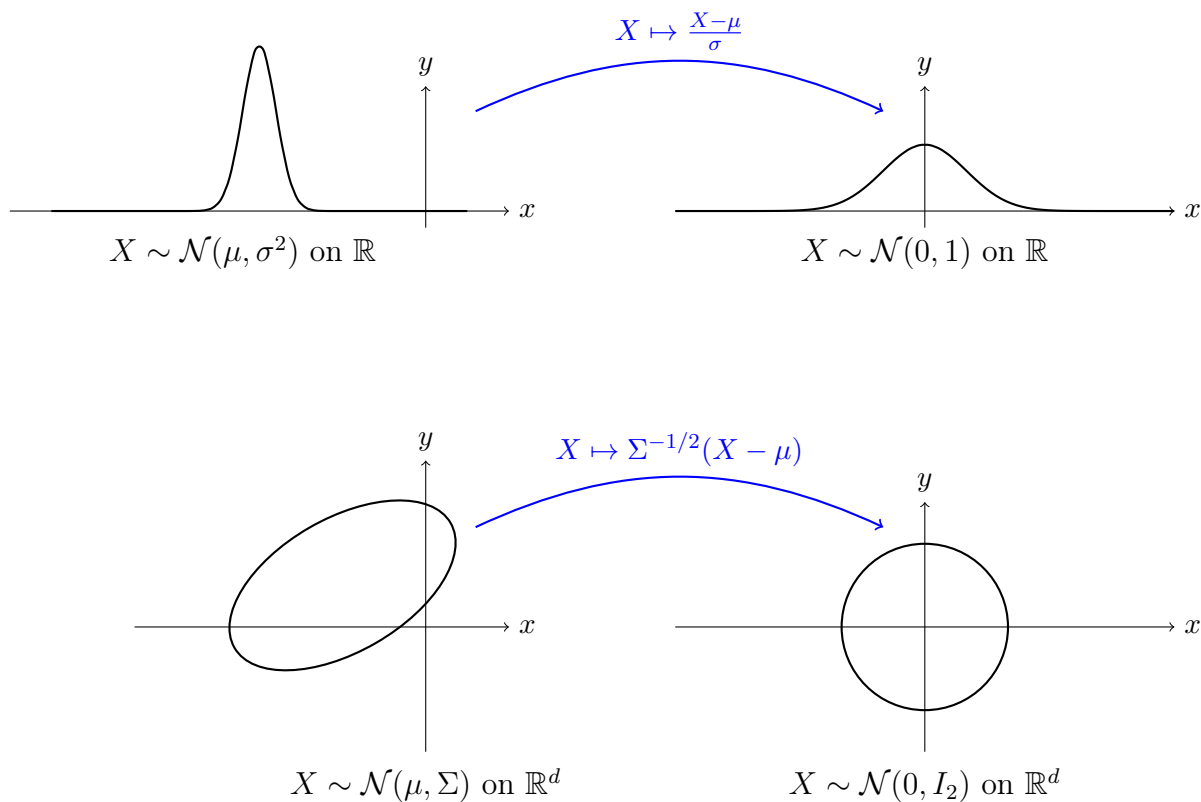


Figure C.2 – Illustration of Gaussian renormalization, in dimension 1 and in dimension d .

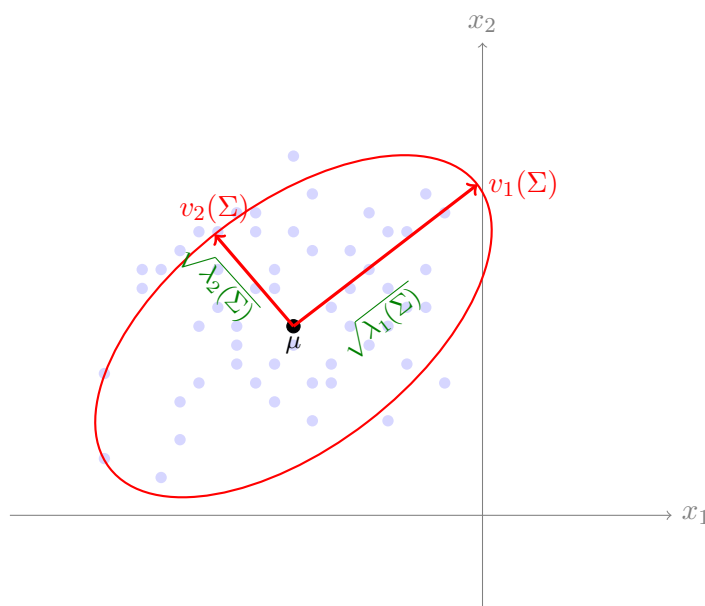


Figure C.3 – Typical shape of a Gaussian point cloud in \mathbb{R}^d . The ellipse corresponds to the geometric locus of points with constant density.

of

$$F = \frac{K_1/p_1}{K_2/p_2}.$$