Mathematical Statistics – Lecture notes

Luca Ganassali

Université Paris-Saclay

Last update: January 12, 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`.

-
- 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

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 ans 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 $(\Omega, \mathcal{F})$ 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* $\mathbb{R}^d$ is[2] 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 $X = (X_1, \ldots, X_d)$ 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), the so is X, and we define the *variance of* X as

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

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

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

From the above definition, its 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

[1] in this course, all random variables are real.

[2] in 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 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

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

that is, for all $1 \le i, j \le d$, $[\mathrm{Var}(X)]_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathrm{Cov}(X_i, X_j)$. Thus, $\mathrm{Var}(X)$ 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:*

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

Proof. Writing $Y = (Y_1, \dots, Y_d)$, it is readily seen that for all $1 \le i \le d$, $Y_i = \sum_{k=1}^d A_{i,k}X_k + b_i$, 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

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

using now linearity of (vectorial) expectation. □

Remark 1.2. With the above property, we have that for all $a \in \mathbb{R}^d$, $a^\top \Sigma a = \mathrm{Var}(a^\top X) \ge 0$. 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 manipulates 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 $X_n \xrightarrow[n\to\infty]{(d)} X$ and $Y_n \xrightarrow[n\to\infty]{(d)} c$ where c is a constant. Then, $(X_n, Y_n) \xrightarrow[n\to\infty]{(d)} (X, c)$.*

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

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

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

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. □

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

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

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

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

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 $v_n \underset{n\to\infty}{\longrightarrow} +\infty$. We assume that there exists $a \in \mathbb{R}^d$ and a random vector X of $\mathbb{R}^d$ such that*

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

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

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

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

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

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

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 $\varepsilon(x) \underset{x\to a}{\longrightarrow} 0$. We can then extend ε by continuity to a. Since $X_n \xrightarrow[n\to\infty]{(d)} a$ in distribution, then by continuity, $\varepsilon(X_n) \xrightarrow[n\to\infty]{(d)} \varepsilon(a) = 0$. Thus, we have for all n,

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

We get,

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

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

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. □

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* $t > 0$,

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

Example 1.1. If $S_n \sim \mathrm{Bin}(n, p)$, then $S_n/n \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[S_n/n] = p$ 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* $X_1, \ldots, X_n$ *be independent random variables such that for all* $1 \leq i \leq n$, $a_i \leq X_i \leq b_i$ *almost surely. Let* $S_n = X_1 + \ldots + X_n$. *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 $-b_i \leq -X_i \leq -a_i$. □

Proof of Lemma 1.1. Wlog we asusme 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 $\psi(u) = \psi(0) + u\psi'(0) + \frac{u^2}{2}\psi''(v) \leq \frac{u^2}{8}$.

□

Example 1.2. We continue our previous example, where $S_n \sim \mathrm{Bin}(n, p)$. 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 $X_1, \ldots, X_n$ be independent random variables such that for all $1 \leq i \leq n$, $|X_i - \mathbb{E}[X_i]| \leq M$ almost surely. Let $S_n = X_1 + \ldots + X_n$ and denote $V_n = \sum_{i=1}^n \mathrm{Var}(X_i)$. 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 $t > 0$,*

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

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

Proof. Expand the exponential in series and write

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

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 $t, s > 0$,

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

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 $S_n \sim \mathrm{Bin}(n, p)$. Here, each $X_i \in \{0, 1\}$, so $M = 1$ and $\mathrm{Var}(X_i) = p(1-p)$. Then

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

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 variabels are bounded. We can howefer 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 $B \in \mathcal{B}(\mathbb{R}^m)$,

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

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 heareafter the formal definition.

Definition 1.4 (Conditional distribution)**.** A function $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^m) \to [0, 1]$ 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 all[3] measurable all $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ 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 generated[4]. 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 proven that such conditional distribution 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.

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

[4] every 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 $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ 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* $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ *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 $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ such that $\mathbb{E}[f^2(X,Y)] < \infty$, the conditional variance of $f(X,Y)$ given $X = x$, denoted $\mathrm{Var}(f(X,Y) \,|\, X = x)$, is defined by

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

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

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

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

Proof.

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

□

1.4.3. Case where X, Y have a joint density

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 $\sigma-$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 $C \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^m)$ $(= \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^m))$,

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

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

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

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 $Q : \mathcal{X} \times \mathcal{B}(\mathbb{R}^m) \to [0,1]$ is a conditional distibution 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 measurablity of the density $p_{(X,Y)}$. To show (iii) we will even show that for all $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ 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 $p_{(X,Y)}(x,y)\mathbb{1}_E(x)$ (these two densities agree almost everywhere since $\mathbb{P}(X \in E) = 0$), we can assume that $p_{(X,Y)}(x,y) = 0$ if $x \notin E$. Then, for such an f,

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

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

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

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 $X \sim \mathrm{Bin}(k, y)$. 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

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

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 $X \,|\, Y = y \sim \mathrm{Bin}(k, y)$.

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 $Y \,|\, X = x \sim \mathrm{Beta}(x+1, k-x+1)$.

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.

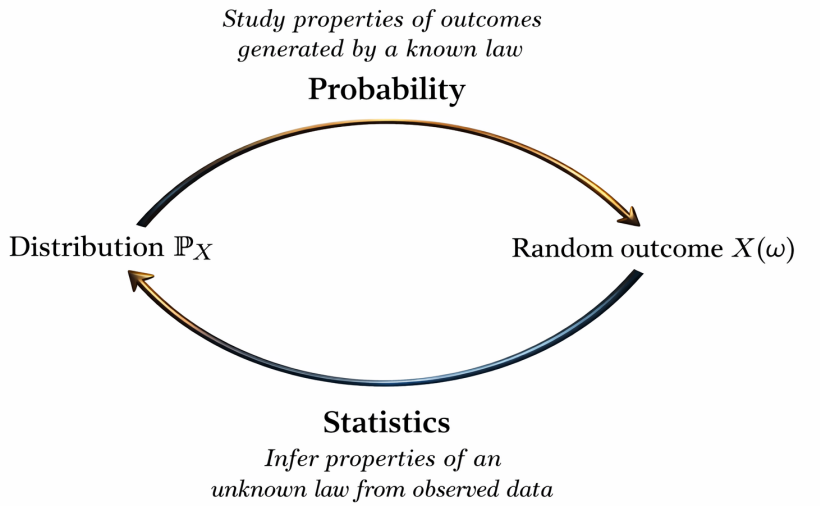

Figure 2.1 – *Probability and statistics viewed as inverse problems of each other.*

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 $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{X}, \mathcal{F})$. 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 be $(\mathcal{X}, \mathcal{F})$ 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 practicioner 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 or the first objective. But it is important to keep in mind that modelling reality as to less complex that 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 $\mathbb{P}_\theta$ (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 $\mathbb{P}_\theta$ 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

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

where $\mathrm{Ber}(\theta)$ 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 $2 \le i \le n$,

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

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 $1 \le i \le n$,

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

where $f : \mathbb{R} \to \mathbb{R}$, 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 model[1]. 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 $\mathrm{Var}_\theta(Y_1) = \sigma^2$.

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.

[1] when f is moreover assumed to be an affine function, this model will be the main focus of Chapter 6.

Definition 2.5 ($\nu-$dominated models)**.** A statistical model $\mathcal{M} = (\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ is *dominated by a measure ν (or $\nu-$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 infintely 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 $\sigma-$finite, by the Radon-Nykodym theorem (all $\mathbb{P}_\theta$ are finite thus $\sigma-$finite, see Appendix B for a reminder), every $\mathbb{P}_\theta$ has a density with respect to ν. These densities will be useful to the statistician. Since there are infintely 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 $\mathbb{P}_\theta$ 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 $\mathbb{P}_\theta$ 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 $\sigma-$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 $\sigma-$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 $\sigma-$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 $\sigma-$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. □

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

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

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 $(\theta_n^\star)_{n\in\mathbb{N}} \in \Theta^{\mathbb{N}}$. 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,

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

Thus the supremum in (2.3) is attained.

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

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

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

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

Consequently,

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

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

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

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

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

thus $\nu^\star = \nu(\{q > 0\} \cup \{p_\theta > 0\}) = \nu(\{q > 0\}) + \nu(\{q = 0\} \cup \{p_\theta > 0\}) = \nu^\star + \nu(\{q = 0\} \cup \{p_\theta > 0\})$. Thus $\mathbb{P}_\theta(\{q = 0\} \cup \{p_\theta > 0\}) = 0$ 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 $\mathbb{P}_{\theta_n^\star} \ll \nu'$ for all $n \geq 0$, then $Q \ll \nu'$. Thus Q is minimal. □

Remark 2.6. Note that the $\sigma-$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 $T : (\mathcal{X}, \mathcal{F}) \to (\mathcal{Y}, \mathcal{G})$ 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 $\mathbb{P}_\theta$ of X, or that reduces the data while preserving the amount of information about $\mathbb{P}_\theta$ 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 $\mathbb{P}_\theta$ 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 $\sigma-$finite measure. In this case, every $\mathbb{P}_\theta$ 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 $\sigma-$finite measure ν. For $\theta \in \Theta$, let p_θ be the density of $\mathbb{P}_\theta$ with respect to ν. Let $S : (\mathcal{X}, \mathcal{F}) \to (\mathcal{Y}, \mathcal{G})$ be a statistic. Then the following are equivalent:*

(i) $S(X)$ is a sufficient statistic;

(ii) there exists a measurable function $h : \mathcal{X} \to \mathbb{R}_+$ and for all $\theta \in \Theta$ a measurable function $g_\theta : \mathcal{Y} \to \mathbb{R}_+$ such that for all $x \in \mathcal{X}$,

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

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, $G = S\sharp\nu$ the distribution of $S(X)$.
- when $X \sim \mathbb{P}_\theta$, we denote $G_\theta = S\sharp\mathbb{P}_\theta$ the distribution of $S(X)$.

$\underline{(i) \implies (ii)}$ 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 $\sigma-$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 $S(X) = s$ when X follows any $\mathbb{P}_\theta$ 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 $\widetilde{\mathbb{P}}$ such that on any Borel set C

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

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

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

which shows that $\mathbb{P}_\theta$ has density $x \mapsto g_P(S(x))$ w.r.t. $\widetilde{\mathbb{P}}$.

Now, let us show $\widetilde{\mathbb{P}} \ll \nu$. 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 $\widetilde{\mathbb{P}}(N) = \int P_s(N) dG(s) = 0$.

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

$\underline{(ii) \implies (i)}$ 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,

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

$$= \int \left(\int f(s) g_\theta(s) dG(s) \right) h(x) dQ_s(x)$$
$$= \iint f(s) g_\theta(s) w(s) dG(s),$$

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 $\widetilde{Q}_s$ 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,

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

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 $\widetilde{Q}_s$ is the conditional distribution of X given $S(X) = s$ when $X \sim \mathbb{P}_\theta$. Since $\widetilde{Q}_s$ does not depend on θ, $S(X)$ is a sufficient statistic. □

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

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

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 $\sigma-$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). □

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 to much information. There must be a minimality notion associated with sufficience.

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 $\mathbb{P}_\theta-$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 $\nu-$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 $u : \mathbb{R}^k \to [0, 1]^k$, 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 $\mathbb{P}_\theta$), 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 $\theta \in \Theta$,

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

Thus, for all $\theta \in \Theta$, $\mathbb{E}_\theta[(\mathrm{id} - H_2)(S)] = 0$. 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 $\theta \in \Theta$,

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

The second equality is justified by step 1. This proves that $\mathbb{P}_\theta-$a.s., $\mathrm{Var}_\theta(H_2(S) \,|\, S') = 0$, that is, $\mathbb{P}_\theta-$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. □

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 $\mathbb{P}_\theta$, 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 $\mathbb{P}_\theta$ 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)**.** *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 $\mathbb{P}_\theta$,*

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

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,

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

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

which proves independence. □