# Mathematical Statistics – Lecture notes

Luca Ganassali

Université Paris-Saclay

Last update: January 5, 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

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

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

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

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

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

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

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

*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

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

using now linearity of (vectorial) expectation. □

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

For the interested reader, a reminder on Gaussian vectors can be found in Chapter 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

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

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

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

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 $2\varepsilon$. 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 $a$, $\varepsilon$ 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 $\varepsilon$ 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, 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 $\varepsilon$. 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 $\phi$ (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 $Z = (X,Y)$, which 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 $\mu$ and $\nu$ 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 $\mu$. 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. $\nu$.

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. $\nu$. 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 $\mu$ the counting measure on $\{0,\dots,k\}$ and $\nu$ 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 $Y \,|\, X = x \sim \mathrm{Beta}(x+1, k-x+1)$.